Targeting the transposable elements of the genome to enable large-scale genome editing and bio-containment technologies.

Oscar Castanon Velasco

NNT : 2019SACLX006

Targeting the transposable elements of the genome to enable large-scale genome editing and biocontainment technologies

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Ecole Polytechnique

École doctorale interfaces n°573 : approches interdisciplinaires, fondements, applications et innovation (INTERFACES)
Spécialité de doctorat: Biologie

Thèse présentée et soutenue à Palaiseau, le 14 mars 2019, par

Oscar Castanon-Velasco

Composition du Jury :

M. Hannu Myllykallio Directeur de Recherche, LOB, Ecole Polytechnique CNRS UMR7645 and INSERM U1182	Directeur de thèse
M. George M. Church Professeur de Génétique, Blavatnik Institute, Harvard Medical School Broad Institute & Wyss Inst. of Biologically Inspired Engineering	Co-Directeur de thèse
Mme Yad Ghavi-Helm Chargée de recherche, Institut de Génomique Fonctionnelle de Lyon Univ Lyon, CNRS UMR5241 ENS de Lyon et Université Claude Bernard Lyon 1	Rapporteur
François Képès Co-fondateur de la société Synovance, Membre de l'Académie des Technologies ; Antérieurement, Directeur de Recherche CNRS à l'université Paris-Saclay (Evry)	Rapporteur
M. Ariel B. Lindner Directeur de Recherche (INSERM), Centre de Recherches Interdisciplinaires, Université Paris Descartes	Examinateur
M. Herman van Tilbeurgh Professeur de biologie structurale, Université Paris-Sud CNRS UMR 9198	Président
M. Marc Güell Professeur Assistant « Tenure-track » de Biologie Synthétique Head of the translational synthetic biology group	Examinateur

Thèse de doctorat

Dans les champs de l'observation, le hasard ne favorise que les esprits préparés.

In the fields of observation, chance favors only the prepared minds.

Louis Pasteur

Acknowledgements

This research work would not have been possible without the members of the Laboratory of Optics and Biosciences (LOB). I want to thank in particular my supervisors Hannu Myllykallio and Hubert Becker for their invaluable guidance to design the research project as well as the LOB's Director François Hache for his continuous and flawless support. Also, I would like to express my gratitude to the expert members of the jury – François Képès, Ariel Lindner, Yad Ghavi-Helm, Marc Güell and Herman Tilbeurgh – that have graciously accepted and taken the time to be part of the doctoral committee despite their many responsibilities. Thank you for their kindness and availability.

This PhD work has been realized in collaboration with the Church Lab at Harvard Medical School, where most experiments have been performed. I would like to thank my co-supervisor George M. Church for providing such a thrilling and exceptional environment in the laboratory, as well as the team I have been closely working with daily. This includes Cory Smith, my brilliant, happy mentor and partner in crime in the lab; Marc Güell and Luhan Yang, my early mentors and talented researchers, whose teachings early in my PhD were critical for me to reach experimental maturity. Also, it has been a delight to work closely with Parastoo Khoshakhlagh, Khaled Saïd, Chun-Ting, Amanda Hornick and more recently with Raphael Ferreira. Working with all these amazing people led this PhD project to move forward, leading to three publications.

In addition, I would like to acknowledge my mother Iiris, my brother Pablo as well as my late father Manolo – may he rest in peace – for their unflinching moral and intellectual support despite the thousands of miles separating us. Finally, my friends on both sides of the Atlantic Ocean – Mihan, George Romeo, Aleks, Ingred, Benedikt, Mehdi, Kosar, Antoine, Sébastien, Alain, Yacine, Pierre, Germain, Christelle et Geoffrey to name a few – have been of utmost importance for me to stay healthy in mind and body during my PhD adventures.

Résumé en français

Le ciblage des éléments transposables du génome humain pour développer des technologies permettant son remaniement à grande échelle et des technologies de bio-confinement

Mots clés : Génie génétique, CRISPR-Cas9, Base Editing, édition génomique à grande échelle

Les nucléases programmables comme CRISPR-Cas9 sont des signes avant-coureurs d'une nouvelle révolution en génie génétique et portent en germe un espoir de modification radicale de la santé humaine. Le « multiplexing » ou la capacité d'introduire plusieurs modifications simultanées dans le génome sera particulièrement utile en recherche tant fondamentale qu'appliquée. Ce nouvel outil sera susceptible de sonder les fonctions physiopathologiques de circuits génétiques complexes et de développer de meilleures thérapies cellulaires ou traitements antiviraux. En repoussant les limites du génie génétique, il sera possible d'envisager la réécriture et la conception de génomes mammifères. Le développement de notre capacité à modifier profondément le génome pourrait permettre la création de cellules résistantes aux cancers, aux virus ou même au vieillissement ; le développement de cellules ou tissus transplantables compatibles entre donneurs et receveurs ; et pourrait même rendre possible la résurrection d'espèces animales éteintes.

Dans ce projet de recherche doctoral, nous présentons l'état de l'art du génie génétique « multiplex », les limites actuelles et les perspectives d'améliorations. La découverte du système immunitaire adaptatif chez la bactérie a permis le développement de l'outil d'édition génétique CRISPR-Cas9. Sa modularité dû en particulier à la petite taille, le faible coût et la production rapide des ARN guides ou « gRNAs » qui le constituent permet pour la première fois de « multiplexer » l'édition génétique chez les cellules et organismes mammifères. Cette technologie ouvre enfin la voie au criblage parallèle à haut-débit de séquences d'intérêt ainsi que la modification simultanée de plusieurs cibles génomiques. Par ailleurs, CRISPR-Cas9 peut également réguler l'expression de plusieurs gènes en les activant ou les inhibant de manière simultanée. Malgré les promesses d'un tel outil, il est important d'en souligner les limitations : la nucléase Cas9 doit parfois être intégrée et exprimée au cours du temps au sein du génome, causant potentiellement des translocations et autres anomalies dans les cellules dans lesquelles elle est présente ; aussi, la génération de cassures double-brins dans le noyau peut provoquer les réponses apoptotiques qui réduisent de manières drastiques le nombre de cellules pouvant survivre le processus. Par conséquent, les futures applications, qu'elles soient académiques, thérapeutiques ou industrielles, requérons des améliorations substantielles en termes d'efficacité de l'outil et de viabilité des cellules afin de déverrouiller le remaniement du génome à grande échelle. La technologie du « base editing » – qui permet la conversion d'une cytosine (C) en Thymine (T) (ou Guanine (G) en Adénine (A) sur le brin complémentaire) ou A → G (ou T → C sur le brin complémentaire) – apporte de la précision tout en préservant son efficacité et réduit la toxicité et les aberrations chromosomiques par rapport aux nucléases générant des cassures double-brins. Les « base editors » sont par conséquent de solides candidats pour permettre les modifications génétiques à l'échelle du génome.

Nous tirons profit de la technologie du « base editing » ainsi que de la multitude d'éléments transposables présents dans notre ADN pour construire une plateforme d'optimisation et développer de nouveau outils permettant le remaniement du génome à grande échelle. La difficulté d'introduire plusieurs gRNAs dans une même cellule ainsi que la génotoxicité des éditeurs génétiques actuels sont les deux obstacles majeurs afin de rendre possible le « multiplexing » génétique. Nous essayons ici de surmonter le deuxième obstacle pour développer des éditeurs génétiques plus sûrs que nous testons via le ciblage des éléments transposables du génomes tels que les virus endogènes humains de type W (HERV-W), les « Long-interspersed nuclear éléments-1 » (LINE-1) et Alu. Ces transposons constituent des cibles attrayantes puisque leur haut niveau de conservation nous permet de concevoir un nombre limité de gRNAs ciblant simultanément un nombre de séquences allant de plusieurs dizaines de milliers à plusieurs centaines de milliers. Initialement, le ciblage d'éléments transposables par le système CRISPR-Cas9 générant des cassures double-brins et l'identification de cellules hautement modifiées n'ont pas eu de succès. Nous avons alors conçu et testé des gRNAs ciblant les séquences LINE-1 en utilisant les « nicking Base Editors (nBEs) » – ne générant qu'une cassure simple brin. Cette stratégie a permis la génération de clones cellulaires HEK 293T contenant jusqu'à 781 modifications. Des modifications génétiques de cette ampleur n'avaient jamais été démontrées jusqu'alors.

Afin de continuer à améliorer ces outils d'édition génétique, nous avons décidé d'inactiver le site enzymatique catalysant la cassure simple brin restante des nBEs, générant deux nouvelles enzymes – dCBE et dABE – en se basant sur l'hypothèse que les cassures simple brin au niveau des éléments répétés du génome contribuaient fortement à la toxicité cellulaire observée. Cette stratégie a permis l'identification de clones cellulaires présentant un nombre de modifications deux ordres de magnitudes supplémentaires par rapport aux cellules editées via les nBEs, avec plus de 13200 modifications dans la lignée cellulaires HEK 293T et 2600 modifications dans des cellules souches pluripotentes induites (ou iPSCs), soit le plus grand nombre de modifications génétiques simultanées jamais observé, par conséquent établissant un socle pour l'écriture des génomes de mammifères.

En outre, l'observation de la toxicité engendrée par la multitude de coupures double-brins dans le génome nous a amené à développer un bio-interrupteur susceptible d'éviter les effets secondaires des thérapies cellulaires. Cette technologie s'appuie sur la toxicité générée par le ciblage des éléments Alu par un système Cas9 conditionnel induit par la doxycycline. Expérimentalement, ce système « suicide » permet l'élimination de 99.98% des cellules. Par ailleurs, nous démontrons que la doxycycline n'est pas toxique pour les cellules et que le système ne s'active pas spontanément en l'absence de la molécule inductrice. Ces deux caractéristiques sont en effet absolument requises pour un potentiel développement en clinique.

En conclusion, nous exposons les considérations éthiques qu'apporte le domaine du génie génétiques et apportons des pistes de réflexions pour diminuer les risques identifiés.

Table of Contents

Abbreviations

DNA DeoxyriboNucleic Acid

ssDNA single-stranded DNA

dsDNA double-stranded DNA

RNA RiboNucleic Acid

mRNA messenger RNA

tRNA transfer RNA

A Adenosine

T Thymine

C Cytosine

G Guanine

U Uracil

CRISPR Clustered Regularly Interspaced Short Palindromic Repeats

Cas CRISPR-associated proteins

spCas9 *Saccharomyces pyogenes* Cas9

saCas9 *Staphylococcus aureus* Cas9

nCas9 nicking Cas9

dCas9 dead Cas9

BE Base Editor

nBE nicking BE

dBE dead BE

ABE A to G BE

TRAC TCR Alpha subunit Constant

CBE C to T BE

gRNA or sgRNA single-stranded guide RNA

crRNA CRISPR RNA

trRNA trans-activating RNA

PAM Protospacer Adjacent Motif

TALEN Transcription Activator-Like Effector Nucleases

ZFN zinc finger nuclease

LINE-1 Long-interspersed Element-1

SINE Short-interspersed Elements

HERV Human Endogenous Retro-Viruses

PERV Porcin Endogenous Retro-Viruses

SRP Signal Particle Recognition

DSB Double-Stranded Break

SSB Single-Stranded Break

RNP RiboNucleic Protein

SSAP Single-Strand Annealing Protein

MAGE Multiplexed Automated Genome Engineering

CAGE Conjugative Assembly Genome Engineering

CAR Chimeric Antigen Receptor

POLE2 DNA Polymerase Epsilon subunit 2

HLA-I human leukocyte antigen class I

PD-1 programmed death-1

GP-write Genome Project Write **ENCODE** ENCyclopedia Of DNA Elements

RGE Repetitive Genetic Element

TE Transposable Elements

HDR Homology-Directed Repair

NHEJ Non-Homologous End Joining

ssODN Single-Stranded Oligo DeoxyNucleotide

AAV adeno-associated virus

HSV Herpes Simplex Virus

HAC Human Artificial Chromosomes

MMCT Microcell-Mediated Chromosome Transfer

HEK 293T Human Embryonic Kidney 293 T-antigen

hiPSCs human induced Pluripotent Stem Cells

PGP1 Personal Genome Project 1

RNAseq high-throughput RNA-sequencing

NGS Next Generation Sequencing

AcrIIA Anti-crispr protein type IIA

JAK2 JAnus Kinase 2

GTFIIB General Transcription Factor IIB

PB Piggybac

DOX Doxycyline

PI propidium Iodide

CRISPR-DS CRISPR-Defense System

RAC Recombinant DNA Advisory Board

FDA Food and Drug Administration

EMA the European Medicines Agency

1. General Introduction

The recent development of programmable and site-specific endonucleases, such as Zinc-finger enzymes in 1985 and the CRISPR-Cas9 system in 2013, has opened the way to genome engineering. The field can be seen as both a revolution and a threat to society. On the bright side, genome engineering holds the hope to cure genetic diseases such as cystic fibrosis or Huntington without the fear that they will be transmitted to descendants. It may allow us to, one day, eradicate cancer, increase our life spans, bring back extinct species, or even adapt to the radically different environments, that may be required to the realization of our deepest dreams of spatial conquests. On the dark side, genetic engineering brings to surface our fears of eugenics and the possibility of increasing the inequalities of the world by only enhancing individuals that can afford it. Ethics will be of utmost importance to responsibly identify both the opportunities and dangers brought by the ability to edit genomes. However, will our capacity to design genomes will be critical to achieve this human revolution as compared to existing modalities? In that regard, understanding the central dogma of biology can provide us with answers.

1.1. The central dogma of biology: DNA, RNA and proteins

The central dogma of molecular biology (Fig. 1.1) relies on three basic building blocks: 1) Long strings of Deoxyribonucleotides (DNA) encodes the information needed for cells to function and is transmitted from generation to generation through the replication process; 2) The coding segments of the DNA are then transcribed into strings of Ribonucleotides (RNA), which are temporary copies of the DNA that it originated from, through a process called transcription; and finally, 3) the RNA transcripts are either directly functional (non-coding RNA such as ribosomal, regulatory RNAs or

tRNAs) or translated into proteins that have a variety of functions, such as stabilizers or catalyzers of essential biochemical reactions. All three processes—replication, transcription and translation—are heavily regulated and allow cells to respond differently according to internal and external stimuli.

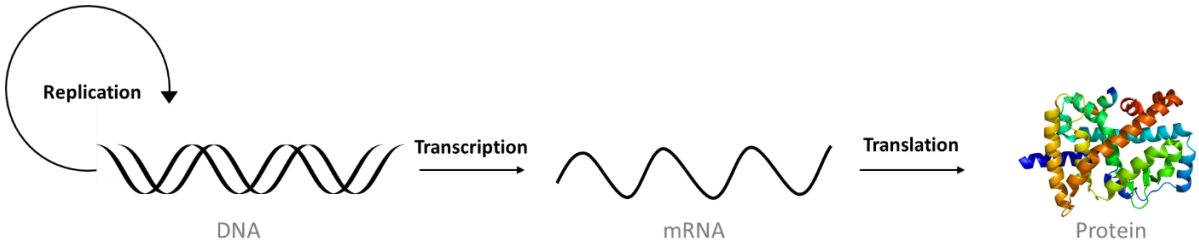


Figure 1.1 | Replication, transcription, translation: The dogma of Biology

The human genome is composed of about three billion base pairs distributed into 23 pairs of chromosomes. It includes about 20,000 genes, each one coding for a protein that can lead to the generation of several different variants when modifications occur downstream at the RNA or protein levels. However, the genome is not only composed of genes, but also sequences that do not translate into proteins. In fact, this non-coding DNA accounts for most of our genome and corresponds, for instance, to sequences involved in the regulation of genes. This includes introns, which are segments within genes that are removed a process called splicing, and transposable elements that account for more than 50% of the genome and whose functions are not well understood.

For a cell to pass its information on to the next generation, its DNA must undergo replication, which is divided into three phases. The first step is called initiation. In this process, DNA helicases begin to separate the double-stranded DNA at distinct loci in the genome, termed the origins of replication. From there, the elongation process takes place. In this phase, other protein complexes, including DNA polymerases, are synthesizing a complementary copy of the DNA starting from replication forks. Integration of nucleotides does not occur randomly but follows the so-called Watson-Crick base

pairing, where Adenosine (A) interacts with Thymine (T) and Guanine (G) with Cytosine (C). Finally, the termination happens when adjacent replication forks fuse to generate two independent and identical DNA segments.

DNA is a form of permanent information storage that remains in the nucleus of the cell. Transcription is the process by which this information is turned into a temporary string of RNA that will lead to the generation of proteins during translation. DNA and RNA are not identical copies since they are composed of slightly different units: deoxyribonucleotides versus ribonucleotides. The transcription process involves transcription factors that bind to the promoter and enhancers, sequences upstream of the gene to be transcribed. These transcription factors recruit RNA polymerases and together, synthesize the string of RNA in the 5' to 3' orientation using the strand of the gene's DNA sequence as a template. Similar to replication, three phases describe this process: initiation, in which the RNA polymerase binds to the promoter and separates the DNA strands; elongation, which corresponds to the progressive synthesis of RNA by addition of ribonucleotides based on the DNA template; and finally, the termination step ends with the release of the RNA polymerase and the transcript thanks to the transcription of sequences called terminators.

Once the transcript is generated, it is exported out of the nucleus and charged into the ribosome, the master unit of translation by which proteins are synthesized. The ribosome is a large double-unit-complex made of proteins and RNA that reads through the transcripts using a set of rules, the genetic code, which matches the three-nucleotide units of the transcript, called codons, with one of 20 specific amino-acids. Translation is initiated when the small subunit of the ribosome binds to the start of the RNA sequence. Then, a transfer RNA, or tRNA, which carries an amino-acid (usually a methionine) interacts with the first codon of the transcript through its complementary anti-codon sequence and is followed by the recruitment of the large ribosomal subunit. In the elongation step, the ribosome interacts with tRNAs that successively load their specific amino-acid into the growing polypeptidic

chain. The ribosome catalyzes the generation of a bond in between each incorporated amino-acid. The process is terminated when the translation complex interacts with a stop codon, which triggers the release of the newly synthesized protein as well as the ribosomal complex.

Replication, transcription and translation are highly regulated processes, allowing cells to respond adequately to their environment. For instance, gene expression can be lowered or increased by repressors or activators, RNA can be modified post-transcription through splicing, and proteins can be modified by addition of polysaccharide chains. Such regulations can also explain how we are able to generate a large diversity of proteins from a relatively low gene number.

Since the genome holds the permanent information necessary for life, is transmitted to future generations and comes upstream of the RNA and protein modalities, it constitutes the keystone that can lead to a radical transformation of human health, influence our own evolution or even wipe out species carrying diseases such as malaria. However, with such power, comes great responsibilities, and this necessitates regulation and oversight, which will be discussed later in this dissertation. For now, I will describe more thoroughly CRISPR-Cas9, the programmable nuclease that revolutionized the field of gene editing and that could lead the way towards the customized design of whole genomes.

1.2. CRISPR-Cas9: From bacterial immune system to genome editing tool

To develop therapeutic compounds, or study biological phenomena, researchers in academia and industry were widely successful at modulating gene expression by the disruption of RNA and protein levels through the targeted modulation of transcription and translation pathways. Now, purposely engineering the DNA—which holds the permanent form of information needed for life—could

revolutionize fundamental and applied biological research. Unfortunately, this task has proven much more challenging in the past decades when scientists relied on the spontaneous or uncontrolled (through chemicals or transposons) modification of genes (or mutants) to further investigate a phenotype. With the advent of site-specific nucleases such as Zinc-Fingers, TALEN and particularly CRISPR-Cas9 since 2013, modifying the genome has never been more easy or efficient, opening new avenues for fundamental and applied biological research.

CRISPR-Cas9 was initially discovered to be an adaptive immune system for bacteria to protect against viruses and plasmids by targeting and degrading exogenous DNA specifically. This molecular system is composed of "Clustered Regularly Interspaced Short Palindromic Repeats" or CRISPR[1] as well as CRISPR-associated (Cas) genes. CRISPR is a segment of DNA with a succession of repeats and spacers. The spacers are sequences from viruses that have been incorporated into the host DNA and act as a memory to fend off future infections. The Cas9 protein possesses an endonuclease activity that makes DNA double-strand breaks at the location it binds to.

Basically, when a new viral invasion occurs, the corresponding spacer and the adjacent repeat sequence are transcribed and processed into a single stranded CRISPR-RNA (crRNA). The crRNA then complexes with a trans-activating RNA (trRNA). The Cas9 protein interacts with the RNA complex to form a ribonucleoprotein. The ribonucleotide sequence recognizes and binds the complementary strand of the viral DNA, guiding and activating the Cas9 endonuclease specifically against the virus. The target sequence always includes a Protospacer Adjacent Motif (PAM), which is necessary to elicit Cas9 enzymatic activity. This serves as a safety mechanism to prevent targeting of the bacterial CRISPR array where the crRNA is transcribed from and which lacks the PAM.

Since the discovery of this bacterial immune system, researchers have repurposed CRISPR-Cas9 as a genome editing tool. The crRNA and trRNA have been combined into one single-stranded guide RNA (sgRNA), and the Cas9 nuclease has been engineered to function in eukaryotic organisms such as

plants and animals. Today scientists can conveniently design and synthesize sgRNAs to target their locus of choice for a wide range of purposes.

1.3. The CRISPR-Cas9 revolution: From single edits to whole-genome recoding

CRISPR-Cas9 bears the hopes of curing virtually any monogenic disease, such as beta thalassemia and sickle-cell disease, treatments for which are currently in clinical trials. However, what if curing or understanding a disease meant modifying more than one or ten or even greater than a hundred genes? Would this multiplexing be feasible with current genome-editing technologies?

The ease of use, modularity, efficiency and low cost of CRISPR-Cas9 system are the features that hold the potential of the technology and led to its massive adoption in the scientific community. CRISPR-Cas9 is modular since the endonuclease does not need to be designed for each target and, as mentioned previously, it is the sgRNA that gives the specificity to the system. Designing and synthesizing one sgRNA is fast and costs only a few dollars. This modularity enables what is called multiplexing, which is the ability 1) to highly parallelize gene-editing experiments for drug screening, for instance; and 2) to easily target more than one locus per genome. For instance, a research group broke the record for the number of edits per cell and generated a viable swine cell line with 62 inactivated porcine retroviruses. However, the field of genome editing needs to overcome significant challenges to progress towards the full recoding of mammalian genomes. Notably, the stress induced by DNA double-stranded breaks leads to substantial cell toxicity and genome-wide editing requires the challenging delivery of multiple sgRNAs.

Overcoming these difficulties and enabling fully multiplexed editing could help us understand the poorly characterized portion of the genome that does not include genes, known as “dark matter” most

of which consists of transposable elements. These elements account for more than 50% of our DNA and include Long-interspersed Element-1 (LINE-1), Alu or Human Endogenous Retro-Viruses (HERVs), whose involvement in our physio-pathology mostly escapes scientific knowledge. These elements, initially described as “junk DNA”, are mobile and have actively shaped our evolution. They can duplicate and integrate themselves back into the genome, potentially disrupting gene expression and impacting our physiology. However, the copy number of such elements varies from tens to tens of thousands, which makes their inactivation at the genomic level impracticable with current tools.

In addition, safe large-scale multiplexing of genome editing could enable the recoding of mammalian genomes with a wide range of applications. The past decades have seen the exponential development of DNA sequencing. With the advent of high-throughput technologies we can now sequence a whole human genome for a few hundred dollars in a few hours, while it took three billion dollars and more than 20 years to construct the first draft of the human genome assembly. However, technologies to write DNA in a high-throughput manner—including large-scale editing and synthesis—have not yet reached the level of DNA sequencers. Therefore, in 2016 the Genome Project Write or GP-write initiative was launched in order to drive the scientific community together towards exponentially improving the design, editing, synthesis and assembly of genomes within the next decade. Genome recoding and writing could lead to revolutionary advances in human health, such as generating virus and cancer resistant cell lines, enabling the development of universal donor cell therapies, and engineering xeno-compatible and synthetic organs. However, genome-wide recoding of mammalian DNA will remain inaccessible in the current state of editing technologies since targeting more than a few sequences has already been shown to be highly toxic to eukaryotic cells.

Would it be possible to leverage programmable nucleases such as the CRISPR-Cas9 system to develop tools to massively and safely edit transposable elements or recode mammalian genomes? Why is genome editing at multiple elements toxic to the cell? What are the cellular and molecular

mechanisms in place that prevent heavy DNA modifications? Can we get around these putative pathways to enable massive genome engineering? Such are the questions that are driving this research project.

1.4. Experimental strategy and research outline

We chose to use the mobile elements of the genome as a tool and as an interesting biological problem. On one hand, the copy number range of these ubiquitous sequences of the genome will allow us to determine the hard limit for the number of edits that an individual cell can handle, and it will guide us on how to improve and fine-tune these editing tools to break these limits through iterations. On the other hand, as I previously mentioned, inactivating transposable elements will help us better understand their poorly known pathophysiology, since, even though these sequences have been regarded as both opportunistic and junk DNA, recent research has shown that they may be involved in neurological and embryonic development, as well as in cancer and neurodegenerative diseases.

As the large-scale editing project moves forward, the observation that the massive modification of mammalian cells can be highly toxic will lead us to explore the potential development of bio-containment technologies. CRISPR-Cas9 is an exciting technology that brings opportunities to the field of biomedical research. Nevertheless, beyond the positive impact lies the potential for the non-ethical use of the editing tool (e.g. bioterrorism, designer babies, etc.) that motivates the development of countermeasures. Following-up with such concerns, we sought to leverage the observed genotoxic feature of CRISPR-Cas9 targeted to repetitive elements in order to develop a CRISPR Defense System (CRISPR-DS) that ensures the introduction or activation of Cas9 to trigger cell death, rendering cell populations in which the system is active effectively non-editable by Cas9.

Finally, we further leveraged this genotoxic mechanism of action to design a surveillance system aimed at making cell therapies safer. The field of regenerative medicine has boomed in the past few years—partly thanks to the genome editing revolution— and has shown great promises to cure diseases such as acute lymphoblastic leukemia, for which the standards of care are very low. Although the transplantation of engineered cells, tissues or organs have showed impressive results, several risks have been identified, notably, the potential to become oncogenic or trigger the cytokine release syndrome by CAR-T therapies, to name a few. In order to mitigate these risks, the development of safety switches provides the option to selectively eliminate transplanted therapeutic cells in case of adverse effects.

This PhD dissertation will be structured as follows. First, the transposable elements of the genome will be defined and the state-of-the-art of eukaryotic multiplexed genome editing will be outlined. From there, I will describe the research that led to successfully achieving large-scale genome editing at repetitive elements. Finally, I will detail the CRISPR-mediated biocontainment technologies—a project that originated from the observation that the original CRISPR-Cas9 genome-editing tool is highly toxic when targeting high copy-number sequences.

2. Transposable Elements of the Genome

2.1. Introduction

Transposable elements are wide-spread repetitive sequences in many organisms' genomes that can duplicate and/or move to a new locus, either autonomously or dependent on another mobile element. Repetitive elements have actively shaped the evolutionary history of eukaryotic genomes and continue to generate novel variants when they duplicate into new loci or are the sites of recombination or translocation. The initial publication of the human genome[2,3] showed that up to two-thirds of the nuclear DNA was repetitive[3] and is largely transposon derived (Fig 1.2). This large fraction of the genome was initially labeled as junk, nevertheless they are correlated with some of our most important physiological processes and can act as a potential cause of disease when duplicating and affecting the expression of key genes.

The most widespread transposable elements are the Long-Interspersed Elements-1 or LINE-1, an autonomous non-LTR class I retrotransposon that constitutes about 17% of the whole human genome; the Alu elements, non-autonomous Short-Interspersed Elements (SINE) dependent on the LINE-1 machinery to transpose itself, that make about 10% of our genome; and finally the Human Endogenous Retro-Viruses (HERVs), LTR retrotransposons which compose about 9% of our DNA.

In the following paragraphs I am going to describe LINE-1 and Alu elements, the two most prevalent transposable elements of the genome, which will be the main objects of our focus in the experimental work of this PhD project.

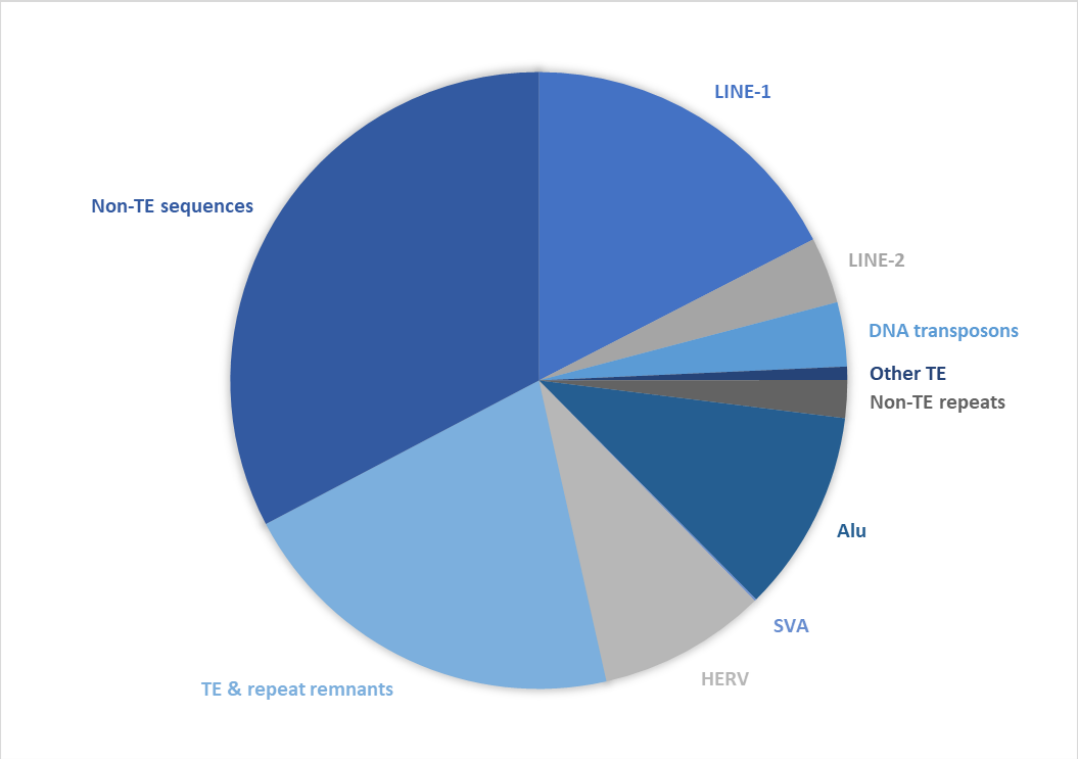


Figure 1.2 | Distribution of Transposable Elements (TEs) of the genome

2.2. LINE-1

Long interspersed elements-1 (LINE-1) are class I retrotransposons that can duplicate and integrate randomly in a genomic location, impacting chromatin structure and gene expression in their nearby environment[5–7].

LINE-1 is a repetitive element of 6kb, representing about 17% of mammalian genomes[3,8] even though most of them are inactive today. LINE-1 elements capable of retrotransposition activity are composed of a promoter, two open reading frames coding for two proteins responsible for the stabilization of the LINE-1 transcript (ORF1) and its reintegration into the genome (ORF2), and finally a poly-A tail[6,9,10]. When a LINE-1 element integrates itself back into an intragenic region, it can modify the expression

of a gene through many mechanisms, such as alternative splicing, creating a new transcription start site, or even enhancing an existing promoter[11,12]. In addition, the retrotransposon can also have epigenetic consequences when altering chromatin structure near its integration site[13,14]. These independent retrotransposons have mostly been seen as opportunistic elements that would try to multiply whenever they would get a chance, sometimes affecting the genome. However, the genetic diversity brought by those elements is highly regulated in every cell in which they are active, indicating that they may have a more important role in biological processes of the host than originally expected. Interestingly, such repetitive elements are inactive in most somatic cells but for the exception of neurons. Indeed, a recent discovery showed that expression of LINE-1 elements increases as a consequence of the reduction of Sox2 expression, a factor expressed in embryonic stem cells and neural progenitor populations.

Surprisingly, LINE-1 becomes one of the most expressed sequences when a neural stem cell commits to differentiation. Retrotransposons thus generate a mosaicism in neuronal precursor cells[15] as a consequence of LINE-1-mediated gene expression disruptions as described previously. However, due to the high copy number of those retrotransposons and the resulting difficulty of detecting new integration events that could impact gene expression, it is still unclear whether the strong expression of the LINE-1 elements in neuronal precursors is directly involved in the differentiation into a variety of neuronal cells. Today, this remains a hotly debated question in the field.

2.3. Alu

Alu elements are about 300 base pair sequences which replicative success has led it to compose about 10% of the human genome with more than one million copies. The canonical Alu element is structured with a left monomer which possesses two RNA III polymerase promoters, separated by a poly-A spacer

from a right monomer that includes an additional 31-nucleotide insertion. The 3' end of the element has a poly-A terminal sequence. The Alu sequence is likely to have emanated from a partially truncated 7SL RNA gene – which encodes the RNA of the Signal Particle Recognition (SRP) – before the divergence of primates and rodents, followed by its duplication in the primate lineage where it gained its efficient duplication feature[16].

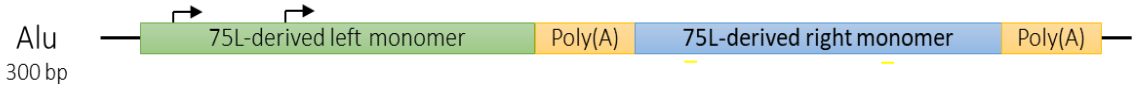


Figure 2.1 | Structure of a canonical Alu element

Alu is a non-autonomous short-interspersed element (SINE) that relies on the LINE-1 ORF2 protein to duplicate and insert itself in different locations of the genome, also called "jumping". Alu elements are divided into several subfamilies that are mostly inactive today. Nevertheless, the youngest sub-groups, AluYa5 and AluYb8 are still capable of "jumping" and can influence the expression and the architecture of the human genome[17].

The "jumping" of Alu elements, similarly to LINE-1, can act as negative and positive drivers of our evolution. If an Alu sequence inserts itself in the neighborhood of an essential gene and disrupts its expression, it can have a deleterious effect on our physiology, such is the case in several neurological syndromes and cancers[18,19]. On the other hand, *de novo* Alu integrations can lead to diversified gene expression within cells or individuals. By recombination events, it can also cause deletions or duplications and contribute to the evolution of our species.

Finally, with the improvement of sequencing technologies and the biocomputing analysis pipelines, research groups have shown that Alu transcripts can regulate several steps of gene expression[20,21] such as splicing or polyadenylation.

Now that we are acquainted with the transposable elements that will be the main objects of this PhD work, both as a tool and as biological elements to be investigated, I am going to review the DNA editing technology that are and will be available to us to pursue the large-scale modification of mammalian genomes.

3. Multiplex Genome Editing Technologies

Reprinted with permission from ACS Chem. Biol., 2018, 13 (2), pp 313–325
DOI: 10.1021/acschembio.7b00842; Publication Date (Web): December 14, 2017

3.1. Introduction

The facile programmability of CRISPR systems has enabled rapid and widespread adoption, leading to the current revolution in nearly every aspect of genome editing, including multiplex editing. The evolved function of CRISPR systems as multi pathogen defenses requires a system that is naturally multiplexable and readily adaptable to arbitrary target sequences. CRISPR RNAs functionally combine with CRISPR associated proteins (Cas) to provide antiviral immunity by targeting foreign nucleic sequences through a complementarity driven mechanism[22,23]. As a molecular tool, the ability to reprogram a single common effector protein through the use of small, *trans*-acting modular guide RNA (gRNA) sequences that target DNA via nucleobase pairing logic[24,25,25,26] is an elegant and near ideal solution to the problem of multiplexing, offering affordable and scalable sequence targeting. Older editing technologies required protein-based targeted factors that were large in coding size, and comparatively slow to generate[27,28]. With the emergence and refinement of gRNA-targeted CRISPR technologies, we now have the prospect of multiplexing on a scale previously impossible to consider with pre-CRISPR editing technologies.

The field of genome editing will need to overcome several limitations that currently prevent highly multiplexed, genome-wide editing of eukaryotic cells (Fig. 3.1). First, new methods that mitigate or avoid the genotoxic stress of multiplex DNA cleavage will need to be developed. Parallel advances will be needed to increase the efficiency of multiplexed delivery while avoiding the cytotoxicity of current multiplex delivery strategies. To edit large portions of a genome, hybrid methods utilizing bacterial DNA assembly methods to produce large CRISPR-targeted donor templates will need to be

established, and the delivery of such large DNAs will need to be dramatically improved to see application in large-scale genome editing projects.

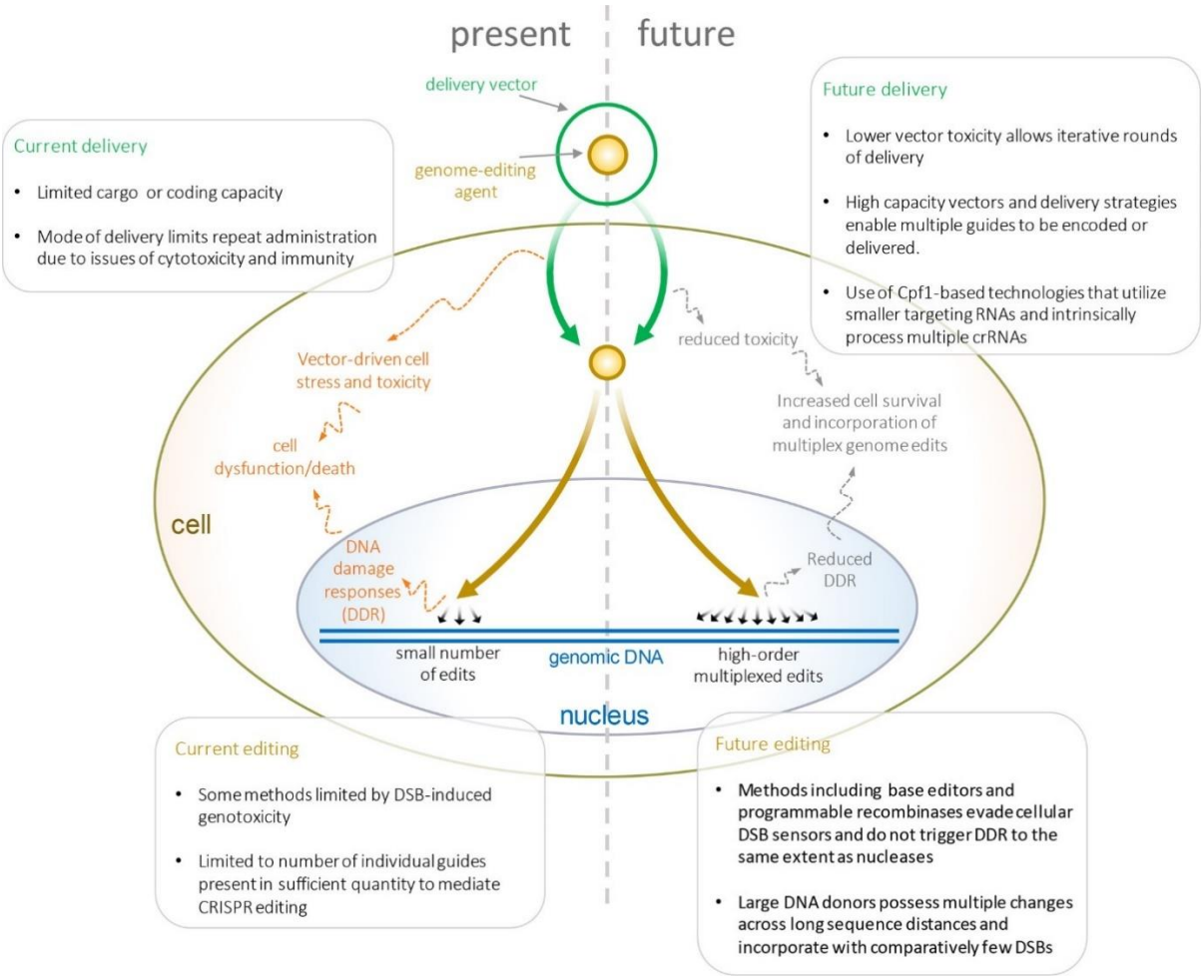


Figure 3.1 | Process of multiplex editing, current limitations, and future improvements

3.2. The State of Multiplex Genome Editing Technologies

3.2.1. Multiplex Editing in Eukaryotic Genomes

The key advantage of CRISPR-Cas based genome editors over previous approaches is the capacity to retarget modifications through the use of distinct gRNAs[29]. The modularity of the CRISPR-Cas9 system and the small size of gRNAs enable a comparatively scalable and rapid production of multiple, distinct

genome editing agents. This has not only allowed genome-wide libraries of gRNAs to be screened in parallel as a pooled population[30–32] but also enables multiplexing by simply delivering or encoding multiple distinct gRNAs per cell. CRISPR-Cas systems thus overcome many design, synthesis, and delivery difficulties when compared to zinc finger nucleases (ZFNs) or transcription activator-like effector nucleases (TALEN).

The intrinsically low barrier to multiplex editing with CRISPR-Cas was recognized from the first reports of its use as an editing tool[25,26,33–35]. Indeed, demonstration of multiplex capability is a common feature of nearly all CRISPR-Cas methodological reports, whether demonstrating novel CRISPR-targeted activities or modes of CRISPR delivery. In addition, the use of CRISPR as a gene expression regulatory tool has also demonstrated the feasibility of a multiplexed CRISPR system through multiplexed activation and/or repression of genes[36–38]. However, “multiplex” as used with regard to genome editing thus far describes a very small number of simultaneous edits. Published protocols only demonstrate modification of as many as seven[39] distinct genomic targets for SpCas9, and up to four with Cas9 orthologues such as Cpf1[40]. Even considering this small number of multiplex edits, with an increasing number of targets, the efficiency at each site decreases when compared to single target editing rates. Moreover, cells can only tolerate a relatively small number of simultaneous double strand breaks (DSBs) due to native DNA damage responses and apoptotic signaling. These factors enforce both procedural and scientific constraints on any effort that invokes multiplex editing.

The largest multiplex CRISPR-based editing effort was recently reported by our group, where targets within 62 porcine endogenous retroviral (PERV) sequences were modified to ablate PERV expression and production, a major barrier to adoption of pig-based organ transplant therapies[41]. However, this feat was enabled by the high sequence identity between distinct PERV elements, which allowed the use of just two distinct guide sequences with a single guide directing the majority of modifications. While the 62 PERV knockout effort was successful, issues of limited editing efficiency and genotoxicity

enforced practical constraints on the development of this multiplexed genome editing protocol. First, acceptable modification rates were only achieved by long-term expression of editing agents from transposon-integrated expression units, as standard transfection was ineffective, and lentiviral integrants were silenced. Cas9 and gRNA expression units were integrated into the genome, with expression proceeding over 14 days. Standard approaches such as multiplex transfection or lentiviral integration, and shorter editing windows, failed to achieve a significant level of editing across multiple PERV loci. Second, the presence of multiple DSBs triggered apoptotic responses and limited the number of surviving, completely modified clones. Cells experiencing the most edits were likely depleted from the population via apoptosis. Thus, while a small number of clones with all PERV knockouts were isolated (8% of cells showed 60–100% PERV knockout rates), the overwhelming majority of surviving cells had less than 10% of PERV sequences edited. This genotoxicity-driven selective process raises concerns over the functionality of edited clones. Given the high expected toxicity of multiple DSBs, surviving clones might be expected to carry mutations that enable evasion of genotoxicity-driven apoptotic death, including in p53. Indeed, in cells known to be more sensitive to DSBs, such as pluripotent stem cells where even single DSBs can lead to apoptosis, some CRISPR-editing protocols call for treatment with p53 inhibitory agents[42]. Even in more robust cell lines, CRISPR nuclease-induced apoptosis may follow introduction of as a few as 4–12 DSBs[43,44]. Finally, the presence of multiple simultaneous DSBs dramatically increases the chance of nonlethal but undesirable translocations[45]. These factors necessitate careful functional, karyotypic, or whole genome sequencing-based screening of clones, which further limits the scale and rapidity of multiplex editing efforts.

Though far from ideal, protocols similar to that used in the PERV knockout effort could be adapted for distinct editing efforts of a similar scope. However, current editing approaches simply do not scale beyond a small number of distinct target sequences and are only practical in scenarios where project

goals can tolerate a small number of surviving clones. Hypothetical future applications, whether academic, therapeutic, or industrial, will require substantially higher efficiency and survivability, with genomic modifications multiple orders of magnitude more numerous.

3.2.2. Strategies for Multiplex Guide RNA Expression

When considering edits across multiple distinct loci, multiplex genome editing using CRISPR requires the simultaneous presence of multiple guides inside the cell, which presents a major obstacle to successful multiplex editing in mammalian cells. Though several groups have developed methods that offer CRISPR-based multiplex editing, no single method currently exists to effectively deliver or express multiple guides with the efficiency and scale needed for massively multiplexed genome editing goals.

Early reports demonstrated that guide RNAs driven by independent RNA polymerase III transcriptional promoters could be functionally expressed in mammalian cells in a multiplex fashion[25,26]. Advances in the molecular biology toolbox have greatly simplified the assembly of complex expression constructs with techniques such as golden gate cloning, which enabled the construction of Cas9 with 7 pol-III regulated gRNAs in a single construct[39]. Cotransfection of gRNA-encoding material has also shown some success, with an early report demonstrating disruption of as many as five genes following cotransfection of gRNA-encoding PCR products[34]. Other work has combined the use of multiple distinct promoters with cotransfection techniques to simultaneously express as many as 12 gRNAs[46]. However, this method encounters multiplex scalability issues, as with increasing numbers of genomic targets, the repetitive nature of each transcriptional unit results in genetically unstable constructs prone to recombination in E. coli during propagation of plasmid constructs. Moreover, the use of separate promoters imposes an unwanted sequence size burden, further complicating delivery of multiplex expression constructs whether by methods of transfection or viral transduction. An

alternative approach drives multiplex gRNA expression from a single transcriptional unit, freeing potentially large amounts of vector capacity. Such methods depend upon enzymatic processing at sites internal to the multiguide primary transcript to release individual gRNA units. This approach was demonstrated first through coexpression of the Csy4 enzyme, which natively processes CRISPR RNA transcripts[47–49]. More recently, Cpf1—a class II CRISPR system orthologue of Cas9[50] —has shown potential as a candidate for multiplex genome editing due to its ability to directly process gRNAs through a DNase-independent RNase domain[40]. Extensive efforts are being put into expanding the utility of Cpf1 via mutagenesis to alter and expand PAM motif recognition[51]. As a two-component system, Cpf1 provides an advantage over other single-construct methods, as it does not depend on separate expression of RNA endonucleases, though potentially removing a layer of processing-dependent system control. Other processing methods that can provide higher regulatory ability rely on endogenous processing mechanisms that function in trans, such as a gRNA adapted tRNA-processing system[52], or in cis through self-processing gRNAs cleaved and catalyzed by ribozymes flanking the gRNA[49,53]. Finally, as discussed below, editing strategies that utilize purified CRISPR ribonucleic protein (RNP) complexes offer a simple mode of multiplexing guides through simple mixing of expressed gRNA material[54–56]. The potential limits of multiplex RNP delivery are currently unknown, though it is likely constrained by the stability of the RNP complex, the intracellular half-life of gRNA, and the number of molecules delivered per cell.

As the demand for CRISPR-based genome editing is increasingly extending to the areas of basic biology, biological engineering, and therapeutics, new methods to enable multiplex editing will be needed. Though all large multiplex efforts will require methods with increased levels of efficiency, scalability, and straightforward experimental implementation, researchers will have to consider the applicability of a given approach to their own system of interest.

3.2.3. Lessons from Bacterial Genome Engineering

To date, the most expansive examples of multiplex editing of intact genomes have come from bacterial systems and were performed without the CRISPR components. Our lab and others have pushed the development of Multiplexed Automated Genome Engineering, or MAGE[57], in multiple prokaryotic organisms. MAGE relies on the introduction of short, single strand DNA (ssDNA) oligos into cells expressing a single-strand annealing protein (SSAP), such as the lambda phage beta protein. The SSAP acts at the replication fork to load the ssDNA oligos onto lagging-strand DNA, leading to their sequence incorporation concomitant with replication. The ability to mix a large number of distinct ssDNA species into a single round of MAGE, and to select or enrich for edited clones, combine with the high growth rate of many prokaryotic models to enable rapid iteration.

However, in higher eukaryotic cells, particularly those that suffer from comparatively slow growth rates and whose DNA replication and repair processes diverge from those of prokaryotes, MAGE may not be a directly transferrable approach. Even if every aspect of MAGE were applicable to a human genetic system, the scale of the editing task would be daunting. Considering coding sequence alone, an effort analogous to the ongoing multiyear “RE. coli” effort[58–60] if performed in human cells would require a roughly 5-fold greater number of edits, and this in a cellular system which propagates an order of magnitude more slowly. More generally, editing in a higher eukaryotic system must contend with a roughly 1000-fold larger genome in comparison to E. coli. While more modest multiplex editing would have a profound impact on basic research and therapeutics, the future of genome writing, recoding, and de-extinction research requires molecular tools that can handle gigabases (gb) of genetic material.

Despite these points of departure, the factors that enable MAGE to scale so effectively to bacterial genome engineering are worth considering. Future eukaryotic multiplex editing methods should

possess (i) high per target modification efficiency, (ii) high-order, per round multiplexability, (iii) ease of programmability (preferably via base pairing), and (iv) short between-round recovery, enabling iterative modification over tractable experimental time scales. Finally, as discussed below, potential future methods of multiplex mammalian genome editing and genome writing may benefit directly from MAGE, with established pipelines of MAGE-based DNA editing serving as a front-end to downstream eukaryotic genome edits.

3.3. Application of Multiplex Genome Editing

New methods of multiplex editing will permit novel applications ranging from basic biological research to genetic therapeutics, industrial applications in metabolic engineering, and the synthetic biological aims of large-scale genome writing and even de-extinction efforts. Practical approaches to safely and efficiently introducing arbitrarily large numbers of edits will be necessary for all of these.

3.3.1. Combinatorial Functional Genomic Methods

The ability to introduce combinations of polymorphic alleles into a genome, whether for specific clonal studies or in library fashion, would enable new methods of studying the genetics of complex traits, with applications in evolutionary biology, population genetics, and the basic biological study of human diseases. Haplotypes could be manipulated experimentally. Ancestral sequences could be reconstructed, and the functional impact of such changes could be evaluated in cell culture or even in animal models bearing homologous sequence changes across multiple variable loci. Sequencing data emerging from the field of cancer genomics has identified an astounding number of mutations arising in tumors, but the functional impact of any given sequence variant, and the interaction between such

variants, has been difficult to ascertain. The ability to deconvolute the functional impact of any given set of mutations via editing would greatly enable the field of cancer biology.

3.3.2. Therapeutic Application of Multiplex Editing

Near-term therapeutic applications of multiplex editing could be seen even with relatively few edits. The deletion of unwanted exons in diseases arising from splicing defects[61] can be achieved with as few as two multiplex DSBs. The engineering of T-cells for immunotherapeutic applications has been hotly pursued[62], with a recent demonstration simultaneously disrupting three target genes whose activities confound the current generation of chimeric antigen receptor (CAR) therapies[63]. To prevent off-target graft-versus-host responses, reduce host-versus-graft immunity, and block immunosuppressive signaling, the researchers targeted TCR alpha subunit constant (TRAC), human leukocyte antigen class I (HLA-Is), and programmed death-1 (PD-1) genes for Cas9 nuclease disruption, respectively. Future applications may require many more. In the case of CAR T-cells, editing additional factors that complicate CAR therapeutic potential through non-PD-1 suppression pathways such as TIM-3, CTLA-4, or Lag-3, T-cell exhaustion, or suppressive cytokines like IL-10 may all serve to augment cancer immunotherapies. Receptors that mediate graft-versus-host responses, and host-versus-graft antigens could also be targeted. A fully mature T-cell immunotherapeutic technology could potentially require modification of dozens of sites. A distinct cancer application of multiplex editing technology, as discussed above, may come from the study of mutations arising from cancer genomics. Applied clinically, a pipeline of functional interrogation via multiplex editing of sequence variants found in patient tumor samples, if executed rapidly, could be a powerful diagnostic and predictive tool.

The emerging field of CRISPR-based eukaryotic antiviral therapy[64] is another area where advances in multiplex editing provide a clear, near-term benefit. DNA viruses and retroviruses can be inactivated or destroyed via targeted viral genome modification, and this approach has been demonstrated for a

number of viral classes, including HIV, HBV, and multiple Herpesviruses[65–67]. However, as shown with Cas9-targeting of HIV proviral genomes, the ability of viruses to rapidly evolve allows evasion of single-target approaches via mutations conferring resistance to cleavage[65,68]. However, multiplex antiviral targeting can negate evasion at any single target site. This approach has proven to be effective in cell culture models of HIV infection[65]; HCMV, HSV-1, and EBV infection[67]; and HBV infection[69]. Antiviral activity can be further augmented by a combination of multiplex editing with simultaneous CRISPR-based transcriptional activation of native viral defenses[70,71]. This strategy may benefit from methods that allow cleavage or transcriptional regulation from a single protein effector[72]. Advances in multiplex delivery and the safety of multiplex editing will further enable this emerging mode of antiviral therapy.

Finally, modifying host-versus-graft antigens in human-sourced donor tissues is an area with a clear need for more advanced multiplex editing technologies. This need is even more pronounced in efforts to “humanize” nonhuman donor tissues, potentially requiring many more genome modifications than have been achieved thus far. Depending on donor material sourcing, the process of editing may be required to turn around edited cells within a short and therapeutically relevant time scale. And because donor tissues may persist in a recipient for decades, the fidelity of the editing process and the resulting functionality of edited cells will be of paramount importance.

3.3.3. Genome Writing

The announcement of the Genome Project Write (GP-write) consortium in 2016 formally ushered in a new era of genomics that moves beyond sequencing genomes and into ground-up writing of genomes[73]. The ambition of GP-write is to understand, design, and test living systems through large, truly genome-scale engineering. Pilot projects include engineering cancer- and virus-resistant mammalian cell lines for the production of biologics, immuno-compatible xenotransplantation, and genomes with new functionalities like biocontainment. The complete synthesis of the small

bacterium *Mycoplasma genitalium*, the engineering and synthesis of large recoded *Escherichia coli* fragments, and the ongoing global efforts to synthesize the full genome of the yeast *Saccharomyces cerevisiae* have set groundbreaking precedents that lay the groundwork for the ambitions of GP-write[60,74–76]. However, projects like GP-write, and so-called "de-extinction" genome-writing efforts, require, in addition to advances in DNA synthesis output and cost-effectiveness, a large degree of multiplexability in gene editing, assembly, and delivery tools.

The nascent field of de-extinction seeks to revive extinct organisms through genome-writing technologies to engineer viable organisms from the ground up, beginning from limited genome sequence data. A range of motivations drives these efforts, from understanding the evolutionary history of extinct lineages and the genetic mutations and bottlenecks which accelerate extinction[77] to ecological restoration *via* reintroduction of extinct keystone species to mitigating the challenges brought about by the ongoing Anthropocene extinction wave. Our lab, in collaboration with Revive and Restore, a nonprofit partnership for the genetic rescue of endangered and extinct species, is working on the de-extinction of one of the most iconic Pleistocene animals, *Mammuthus primigenius*, or the woolly mammoth, an important species in maintaining the Pleistocene steppe ecosystem. Ongoing de-extinction efforts around the world now include the Passenger Pigeon and Health Hen, and preservation efforts for endangered species including the Northern White Rhino and Black-footed Ferret. The development of de-extinction technology would, for the first time in history, mean that bodily extinction of a species is no longer a terminal fate.

Species de-extinction without multiplexing is not feasible. Sequencing data from mammoths reveals a 99.78% identity to the modern *Loxodonta africana* (African elephant) genome at the level of protein sequence, and 0.6% different at the level of nucleic acid sequence[78]. Assuming a 3.3 gb size of the elephant genome, between 7.3 and 19.8 million nucleotide changes must be made to achieve full de-extinction, intraspecies variation notwithstanding. Recent sequencing of several more mammoth

specimens placed the mammoth phylogenetically closer to the Asian, rather than the African, elephant, likely decreasing the expected editing burden[79]. Still the scale of such an editing goal dwarfs any genome engineering effort to date.

From basic biological and evolutionary studies, to ecological engineering and conservation projects, and nascent therapeutic modalities, the future application of genome editing technologies may impact every aspect of our world. To see full application, robust multiplex editing capabilities must be developed. The ability to massively multiplex modifications on a genomic scale will require fundamental improvements to methods of editing, delivery vehicles, and donor DNA construction.

3.3.4. Repetitive Genetic Elements

One field with an intrinsically low barrier to multiplex study is that of repetitive genetic elements (RGEs). Though early application of genome editing focused on protein coding genes, there is an increasing interest in developing methods to interrogate the noncoding complement of the genome. Projects such as the encyclopedia of DNA elements[80] (ENCODE) presented many intriguing observations linking chromatin structure, gene expression, and developmental timing to noncoding loci, including RGEs. Repetitive elements such as Alu, LINE-1 retrotransposons, SVA, and human endogenous retroviruses (HERVs) may occur with ~2700 to 1 million copies per genome[81], representing a nontrivial portion of the total genomic sequence, and are a major source of sequence variation. Expression from such elements appears to be highly regulated, indicating that they may play important roles in biological processes. Such elements are suspected of roles in neurological diseases such as ataxia telangiectasia[13], Rett syndrome[82], and human cancers[83]. However, the study of these sequences is hampered by the inability to distinguish the effects of individual repeats, and to manipulate them at the level of DNA sequence. Targeting multiple RGEs with a small number of gRNAs may be relatively simple given high sequence conservation.

However, such editing with current nuclease-based CRISPR protocols would result in an exceedingly large number of DSBs, and therefore high genotoxicity. Indeed, one study reported extreme toxicity while attempting to modify a repetitive sequence present in 151 copies in a cancer cell line[43]. Multiplex editing protocols that are intrinsically less cytotoxic must be developed if RGEs are to be studied thoroughly. Improved tools to alter the structure and expression of these elements is required to properly interrogate and assess the function of this major component of eukaryotic genomic structure.

3.4. Methods of Multiplex Genome Editing

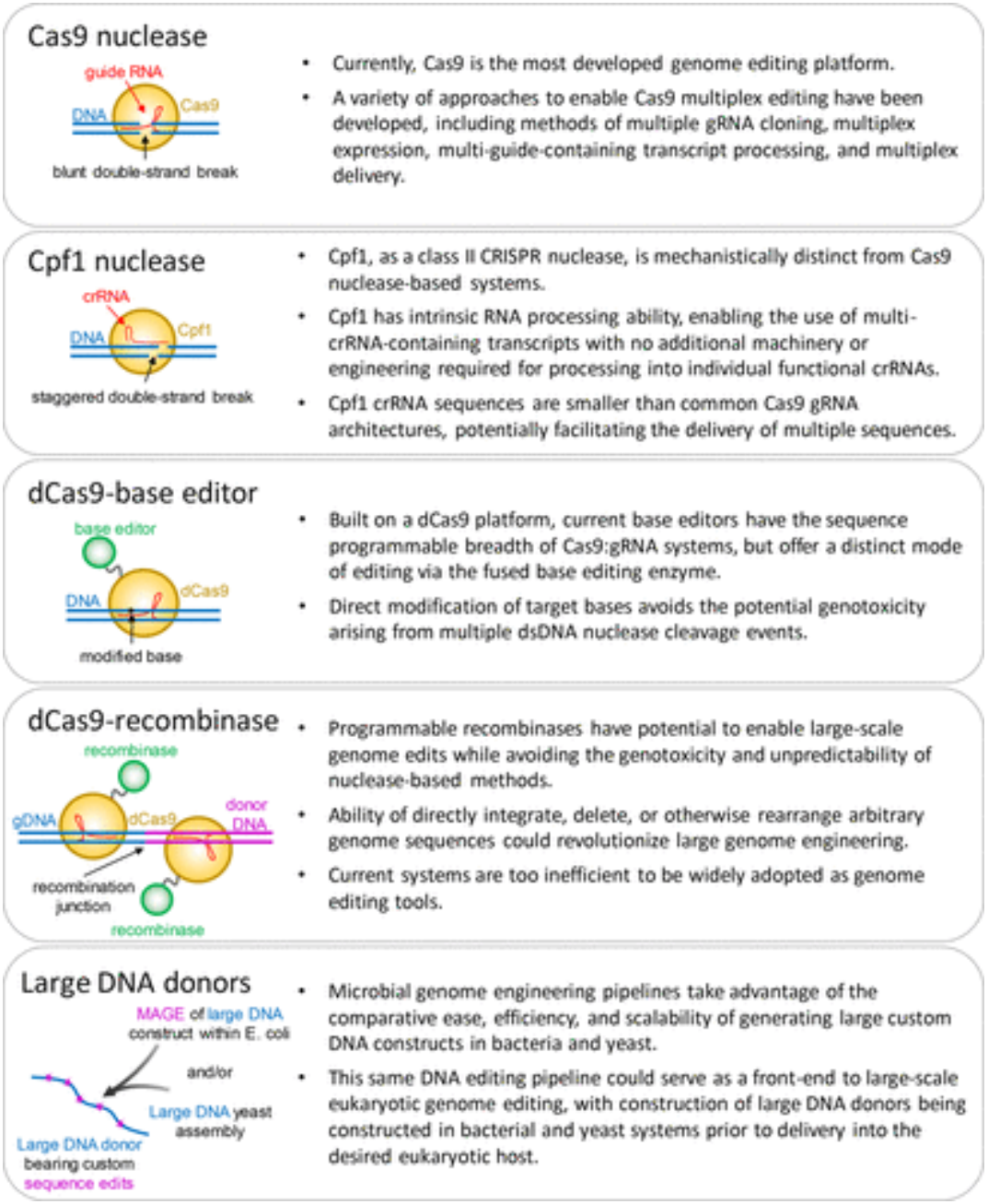


Figure 3.2 | Technologies for introducing multiplex genome edits

A host of different genome editing technologies based primarily around engineered Cas9 systems, but also including other CRISPR components such as Cpf1 and non-CRISPR microbial genome engineering methods, have seen rapid development in recent years (Fig. 3.2). Each technology brings specific capabilities and limitations with regard to multiplex editing, which will be explored in detail below.

3.4.1. Base Editing

One potentially powerful method is the recent development of base-editing technologies built upon the CRISPR platform[84–86], which rely on deaminases to target specific bases for conversion. Deamination of a target cytosine to uracil leads to conversion to a thymidine by cellular DNA repair processes, enabling C → T (or G → A in the complementary strand) transitions of bases within the target window. And with the development of adenine base editors, evolved from a tRNA adenosine deaminase, enabling A → G transitions (or T → C in the complementary strand), it is now possible to target all single base transitions located at an appropriate distance from PAM sequences. As this nascent technology improves, increases in efficiency and target base specificity are expected.

Base editing offers a number of advantages, including the ability to generate sequence changes in a more defined manner than NHEJ, and with potentially greater efficiency than current methods of HDR. One especially intriguing property of base editor technology is that genomic edits are introduced without the generation of DSBs. This mechanism of editing not only reduces the genotoxicity that arises from cellular sensing of multiple DSBs but also mitigates the possibility of genomic translocations caused by improper relegation of mispaired ends during DNA repair. Base editors may thus offer a higher ceiling to multiplex editing as the process becomes limited more by delivery and editing efficiency than by toxicity.

3.4.2. Programmable Recombinases

Research into methods that increase nuclease-targeted homology-directed repair (HDR) in higher eukaryotic cells has been aggressively pursued since the very beginnings of genome editing. As an alternative cellular process to NHEJ, which introduces small, random, insertions or deletions following error-prone repair, HDR enables specified sequence changes through incorporation of an exogenous donor DNA template[87]. The ability to specify arbitrary sequence changes with high efficiency in the DNA of living cells would present a major improvement on the already revolutionary capabilities enabled by CRISPR technology. However, in practice, HDR rates are generally low in higher eukaryotic cells, precluding many applications that demand defined sequence changes. Despite the potential power of high-efficiency HDR, and the great effort put into solving this problem, there is as yet no breakthrough solution to high-efficiency multiplexable HDR. As discussed above, targeted base-editing technology holds great potential in generating defined sequence changes, but its multiplex efficiency and scalability are currently unknown.

An alternative approach to this problem would use site-specific recombinases or integrases, functioning in a manner analogous to the widely used Cre recombinase[88] or piggybac transposase[89], but with user-defined sequence specificity. Through the action of paired recombinase sites, potentially any sequence can be enzymatically deleted, inverted, inserted, or exchanged with base pair resolution. This approach has a number of theoretical advantages.

Whereas nuclease-based editing is mediated by stochastic endogenous DNA repair mechanisms, a recombinase-based approach directly performs targeted strand cleavage, exchange, and religation. Where HDR must compete with NHEJ at the site of repair, modification *via* a recombinase provides an inherently more predictable sequence outcome. Furthermore, natural rates of HDR vary greatly between genomic loci, are altered during the course of the cell cycle, and differ between cell types.

Consistently high rates of homologous recombination are generally only achieved in cells artificially stalled during key points of the cell cycle, or in mutant cell lines that exhibit high basal rates of homologous recombination[90,91]. A recombinase, as the sole mediator of editing, could potentially normalize modification rates independent of these variables. Additionally, as the majority of cells in the body are postmitotic and thus exhibit low HDR, donor-based gene conversion approaches may not be applicable in vivo in noncycling cells[91,92]. A recombinase approach, or at least one that is fundamentally different from invocation of HDR process—especially when enacted across multiple targets—may be required for *in vivo* modification of postmitotic tissue cells.

Finally, whereas multiple DSBs are known to trigger apoptotic cellular responses, the catalytic mechanism of many recombinases appears less prone to induction of DNA damage responses and associated cellular toxicity. During recombination, the recombinase catalytic intermediate is covalently linked to the DNA backbone, with unlinked free ends held within the tetrameric recombinase complex, preventing detection by DNA damage surveillance proteins[93]. While prolonged expression of certain recombinases can negatively affect genome stability[94], this may be avoided if expressed or delivered transiently as with nuclease-based editors.

Despite the potential of recombinase-based genome editing tools, and a two-decade history of recombinase engineering, even the best demonstrations of programmable recombinases to date have yet to achieve wider adoption as genome editing tools. Ideally, a recombinase technology should be (i) fully programmable to arbitrary DNA sequences, (ii) offer mechanistic control over recombination directionality, and (iii) operate with high efficiency. To date, no single recombinase technology satisfies these criteria.

Simple translation fusions of DNA-binding domains to integrases and transposases have been used to tether integration complexes to a target sequence, thereby increasing integration around that locus. Other efforts have used highly engineered chimeric recombinase fusions that attempt to replicate

some aspects of native recombinase functionality, with fusion to zinc fingers[95], TALE arrays[96], and dCas9[97] having been demonstrated. However, all of these suffer from a host of defects that prevent their wider adoption, including low efficiency, catalytic domain-constrained sequence preferences, and lack of control over the directionality of the recombined product. As an obligate tetrameric enzymatic process, the orchestration of a recombinase reaction may prove technically challenging[93]. Not only must catalytic domains be active as chimera, they need to be appropriately spaced to interact in the context of the dimeric recombinase full-site and must also interact as tetramers in a manner that provides directional control over recombinase resolution. Additionally, the catalytic domain itself must impart minimal sequence preferences on the action of the chimeric enzyme. Further insights into catalytic domain sequence preferences, and basic enzymatic mechanisms, should inform future efforts to generate truly programmable recombinases.

In the absence of programmable recombinases, existing site-specific enzymes may still have utility in large-scale editing efforts. Recombinases, perhaps best exemplified by the now ubiquitous Cre/loxP and Flp/FRT tyrosine recombinase systems, allow efficient exchange between donor DNA and the genome[88]. The difficulty in engineering the sequence recognition of recombinases generally limits their utility to cases of large integration or exchange events and necessitates pre-engineering of the target genome to contain the recombinase target site. For certain genome engineering goals, however, this may not be limiting. Through strategic use of expanded panels of orthogonal recombinases, such as the diverse set of characterized serine recombinases[98], a pipeline based on orthogonal recombinase exchange events could enable certain bottom-up genome engineering goals.

3.4.3. Large Donor DNAs

An approach distinct from triggering multiplex editing events may be to manifest an identical sequence outcome by converting multiple positions simultaneously *via* large donor DNAs and

homology-directed repair processes, triggered by a comparatively small number of DSB events. Alternatively, chromosomal regions could be altered *via* recombinase-mediated cassette exchange, inserting donor material through the action of engineered recombinases. Regardless of the means of introduction, the engineering of appropriate donor material must proceed on a scale relevant to multilocus editing goals.

DNA synthesis services today can generate sequences kilobases (kb) in length. However, donors of this size may only be of utility in cases where desired sequence edits are closely spaced, as direct synthesis beyond 1 kb can become costly. Thus, methods that can rapidly engineer DNAs on the order of tens to hundreds of kilobases are needed to address sequence variation in higher eukaryotic genomes. When considering the introduction of hundreds to millions of genomic modifications into a eukaryotic genome, experimental pipelines initially developed for prokaryotic genome-engineering efforts take on new relevance[60].

Several technologies with a high degree of multiplexing must be combined to streamline whole genome engineering. Thankfully, integrated protocols for introducing thousands of widely spaced edits into bacterial DNA have already been developed. MAGE (as described above) provides a scalable pipeline for engineering of sequence variants in E. *coli.* Combined with the availability of BAC libraries for a number of mammalian genomes, a process of high throughput MAGE editing in bacteria, followed by CAGE (conjugative assembly genome engineering) assembly of multiple MAGE-edited fragments, could rapidly build up edited donors of hundreds of kilobases prior to delivery into mammalian cells[58].

Multiple BAC-sized constructs could even be combined into donor material that approaches megabase (Mb) scale. The exo, beta, and gam genes from the λ-Red phage responsible for bacterial recombineering can process a double stranded DNA fragment to single stranded DNA, which can then be incorporated into the lagging strand during replication. Recently, a 100 kb DNA was cut out of the

episomal vector via CRISPR/Cas9 and incorporated into the E. coli genome viaλ-Red recombineering[99]. If many of these large fragments can be incorporated in close proximity, then excised via recombinases in a circular form, they can potentially be delivered to mammalian cells. Alternatively, DNA assembly in yeast can generate a complete 580 kb mycobacterium genome from 35 kb subgenomic fragments[74]. Yeast assembly may thus be extraordinarily useful for large-scale genome engineering.

3.4.4. Generating Large ssDNA

Irrespective of how large donors are constructed, the efficiency of integration must be maximized. In the absence of efficient programmable recombinases, or strategic use of their natural recombinase counterparts, HDR rates will need to be augmented. One approach to augmenting DSB-triggered HDR is the use of single-stranded DNA. Single-stranded DNA confers higher modification rates than double-strand DNA (dsDNA) and also minimizes the required length of sequence homology arms[100,101].

Short ssDNA oligo deoxynucleotide (ssODN) donors are synthetically accessible at low cost, making their use popular in HDR editing[25,26,102] experiments. But given the still low and variable HDR rates, and direct competition of each event with error-prone NHEJ, multiplex HDR with ssODNs is practically unlikely to scale for even modest multiplex editing goals. Larger insertions require DNA donors that may not be as cost-effective to synthesize as ssODNs, or donors that are beyond current DNA synthesis length capabilities. Thus, bacterially produced plasmid or BAC DNA is typically used, and often in dsDNA form. This is despite the higher efficiency of ssDNA as a donor. The difficulty in isolating sufficient quantities of large, high quality ssDNA has limited the adoption of large ssDNA donor generation methods.

Current methods of producing large ssDNA donors rely on either production of phagemid constructs, primer extension and linear amplification of circular bacterial episome templates, or gel electrophoretic separation of large DNA strands[100,103,104]. Phagemid constructs are attractive, as standard plasmid cloning, arraying, and library generation methods can be applied directly to phagemid construction. However, excessively long M13 filamentous phage particles result in low production yields, especially if phagemid preparations are to be massively parallelized, with a practical limit to ssDNA phagemid length of roughly 10 kb. Linear ssDNA polymerization methods are similarly flexible and potentially scalable to multiple donor species, but here again donor length is constrained by the processivity of existing polymerases. Finally, though gel electrophoretic separation does allow isolation of large, multikilobase ssDNA donors, the procedure is laborious, low-yielding, and practically difficult to parallelize to genome-scale donor library production.

Emerging DNA synthesis technologies continue to push the envelope of synthetic DNA length and cost. If appropriately developed with novel *in vitro* assembly methods, synthetic DNA inputs could yield large, multikilobase ssDNA donors. Alternatively, novel recombinant DNA techniques that allow input of large dsDNA constructs such as BACs or YACs, and enzymatically manipulate the products to isolate specific strands, could be especially powerful.

3.4.5. Programmed Genome Rearrangement

Looking further afield, there is a range of natural processes that appear to generate large-scale genomic rearrangements and reductions in a programmed manner. Large-scale genome rearrangements occur in a number of organisms, and the mechanisms underlying these processes are only beginning to be understood. If the specific factors responsible for orchestrating programmed genome rearrangement can be identified and abstracted for use as molecular tools, then they may enable future genome-scale engineering efforts.

Millions of base pairs are eliminated from somatic tissues of the lamprey following programmed rearrangement[105]. And unicellular eukaryotic organisms, including the protozoans *tetrahymena* and *oxytricha*, undergo perhaps the most dramatic examples of genome-scale editing. The transcriptionally active "macronucleus" of these organisms is rearranged on a massive scale in comparison to the germline "micronucleus," with upward of 225 000 fragments being rearranged during macronuclear formation[106]. The reproduction and survival of the organism depends upon faithful execution of this program in every generation. Despite the thousands of fragments in play, the correct genomic products are rearranged with base pair resolution, genome wide. There is potential for the processes that mediate natural genome rearrangement to be adapted as genome engineering tools.

Orchestration of this process involves a multitude of cellular factors, which may converge noncoding RNAs and RNAi related pathways[107,108]. This suggests a potential mechanism driven ultimately by the rules of nucleic acid base pairing and is consistent with the seeming sequence flexibility of the rearrangement process. Regardless of the molecular mechanistic details, the existence of such natural genome rearrangement processes is encouraging evidence that multiplex genetic alterations can occur on a truly genomic scale and demonstrates one potential route toward that goal. Further study into this mechanism is ongoing at laboratories around the world, and ultimate elucidation of such protozoan genome-rearrangement pathways may someday lead to a new class of genome engineering tools.

3.4.6. Multiplex Delivery

As emerging methods maximize the per guide efficiency of modifications, and minimize the toxicity of editing itself, our ability to multiplex may be constrained by our ability to deliver a requisitely large number of guides to target cells. Regardless of the specific mechanisms effecting genomic modifications, highly multiplexed editing goals will require methods of delivering more complex cargoes than those established for single or duplex editing experiments. Moreover, as any given method will carry theoretical and practical limits to the number of distinct gRNAs and donors that can be accommodated, large-scale genome editing efforts will likely require approaches to multiplex delivery that are sufficiently fast, cost-effective, and of low cytotoxicity such that they can be iterated over successive rounds of modification within experimentally tractable time scales.

One approach that offers simple multiplexing is delivery of expressed Cas9:gRNA ribonucleoprotein (RNP) complexes, whether by lipid nanoparticles[54,56] or electroporation[55]. The direct delivery of editing material avoids potentially troublesome combinatorial, repeat DNA cloning steps for multiple gRNAs and allows simple premixing of in vitro transcribed sequences. Another advantage of this approach is the short half-life of the delivered material; cells experience nuclease activity within a short temporal window, which in addition to measurably reducing off-target effects[54–56], may also reduce genotoxicity, potentially allowing frequent, repeat rounds of modification. Though only a small quantity of material is delivered, and is only present for a short time, the irreversible nature of error-prone NHEJ events makes RNP delivery a powerful approach to targeted gene knockout. RNP delivery has also been demonstrated for the Cas9-deaminase BE3[109]. The absence of DSB-triggered genotoxicity from this strategy may provide additional reductions in multiplex delivery toxicity, further enabling the development of iterative, RNP-based multiplex editing protocols.

Current approaches to RNP delivery rely on lipid nanoparticle transfection or electroporation. While effective for a given instance of modification, the cumulative toxicity of such methods may limit repeated rounds of modification over experimentally short times. Recent developments in *ex*

vivo delivery may further enable multiplex editing. These include microfluidic approaches where passing cells at high speed through constrictions smaller than cell diameter results in transient disruption of cell membranes, allowing cargo in solution to pass through[110]. This approach has been further augmented by application of an electric field that disrupts nuclear membranes, permitting both cytoplasmic and nuclear delivery[111]. In addition, advances in nanomaterials are providing novel approaches to cell delivery. A substrate-only system includes the use of nanowires coated with molecules that are released when cells are penetrated during culture on the nanowire substrate, allowing codelivery of proteins and siRNAs[112]. Another system utilizes a nanofabricated substrate combined with laser pulse illumination, generating controlled microcavitation bubbles to transiently permeate cell membranes in close proximity to the substrate, allowing delivery of RNP-sized cargo[113]. In combination with improvements in cell viability compared to traditional methods (*e.g.*, electroporation), these approaches may permit delivery of multiplex gene editing cargos at high efficiency into cells *ex vivo*, though large DNA cargos remain a delivery challenge.

Viral vectors offer distinct approaches to multiplex *ex vivo* delivery. Transduction with multiple low-capacity, nonintegrating viral vectors at once is one potential route. Choice of viral vector would be key, as production methods must not only scale to either parallel or pooled library production, but high multiplicity of infection must be both achievable (necessitating high production titers) and induce low innate intracellular antiviral responses. This would allow multiple viral genomes to transduce a given cell, and the transduction process might be iterated rapidly. A viral vector such as recombinant adeno-associated virus (rAAV) is one candidate for this, though new methods to augment coincident transduction events, and to reduce vector immune signaling, may be necessary to apply this approach within a pipeline of genome-scale editing. A distinct viral approach would utilize large DNA viral vectors like herpes simplex virus (HSV) amplicon vectors. The ~150-kb packaging capacity of HSV replicons potentially offers the ability to deliver a large number of gRNAs on a single vector. However,

if applied to genome-scale editing, upstream assembly of the vector, and the appropriate replication and packaging of such a large repetitive construct, may prove practically difficult.

Finally, where current cell culture and *ex vivo* approaches to delivery offer higher multiplex capacity, *in vivo* multiplex delivery capabilities are comparatively limited. General avenues for *in vivo* delivery of multiple simultaneous gene editors include nanoparticle (lipid, polymer), viral (AAV, lentivirus), and even whole tissue electroporation[114] . However, nanoparticles are constrained by the bioavailability of individual components across formulation methods; viral vectors are restricted by DNA cargo capacity (AAV) and unpredictable effects of genome integration (lentivirus), and electroporation is limited by physical accessibility of target tissues. Though the inability to introduce a very large number of changes *in vivo* is unlikely to be a barrier to any near therapeutic application, application of multiplex antiviral defense *in vivo*, though requiring relatively few edits, will face distinct delivery challenges and necessitate new multiplex delivery methods.

3.4.7. Delivery of Large DNAs

As discussed above, the use of large donor DNAs could enable effectively multiplexed higher eukaryotic genome editing. However, no matter how efficient edited donor generation pipelines become, all donor material must ultimately be delivered to mammalian cells with sufficient efficiency to recover modified clones. Genome-scale engineering may require repeated delivery of DNAs ranging from many hundreds of kilobases to megabases in size. However, established methods of introducing such large DNAs such as microinjection are very low throughput, toxic to recipients, and may subject the delivered material to mechanical shearing. Cell–cell fusion-based approaches may stabilize large DNAs during delivery but are themselves extremely inefficient. Barriers to fusion-based delivery include complications at the level of initial fusion with recipient cell bodies, and subsequent import or incorporation of DNA into the nucleus.

Large human artificial chromosomes (HACs) like chromosome 14 and 21 have been built and are essential to generating humanized animal models or to study phenotypes in the context of different haplotypes[115]. Traditionally, they have been delivered by microcell-mediated chromosome transfer (MMCT). MMCT is an arduous procedure that first requires transfer and manipulation of DNA into a recombinogenic line, followed by a series of culture treatments that result in condensation of chromatin and envelopment of chromosomal material in membrane-bound cell fragments, which are then fused to recipient cells via fusogens like polyethylene-glycol or virus-mediated agglutinization[116]. Following this demanding process, the efficiency of incorporation is extremely low, on the order of 1×10^{-6}, precluding routine use as part of a genome-scale engineering effort.

Another delivery alternative recently reported involves the fusion of yeast spheroplasts (or cell-wall free yeast) with cultured mammalian cells[117]. The advantages of such a delivery system are attractive, as it may interface seamlessly with upstream MAGE and yeast-assembly methods. Though 1000-fold more efficient than MMCT, this yeast-based DNA delivery protocol is currently limited to roughly 0.1% in cultured cells. Higher efficiencies may be needed to cost-effectively apply this approach to genome engineering on the gigabase scale found in higher eukaryotic organisms.

One major bottleneck in this process may be the postfusion breakdown of the yeast nucleus in recipient cells. Yeast natively have a closed mitosis, and the release of yeast nucleus-borne genetic material, by as-yet undefined processes, is likely extremely inefficient. Moreover, the presence of large DNAs in the cytoplasmic space likely triggers cellular antiviral responses. Finally, the entire yeast nuclear content is delivered to recipient cells; unwanted yeast genomic DNA is incorporated into host cells alongside the desired material.

Future development of this system to enable yeast nuclear breakdown, or nucleus–nucleus fusion, as occurs naturally during yeast sexual reproduction, within the recipient cell, could dramatically increase transfer efficiency. Nucleus–nucleus fusion may also effectively evade certain antiviral responses.

Finally, the development of methods to degrade, exclude, or otherwise negate the yeast genomic material in favor of the desired donor material would greatly increase the utility of this approach.

As with bacterial MAGE-based genome engineering, a mammalian genome writing campaign would use a tiered program of parallel engineering efforts, with modified genomic regions being built up in separate lineages, ultimately requiring hybridization of complementarily editing lines. To avoid the inefficiencies of MMCT at this stage, improvements in cell–cell fusion are especially attractive. Recently, the discovery of fusogenic peptide Myomixer and paired receptor Myomaker has been shown to mediate surprisingly efficient fusion between myoblasts, fibroblasts, or myoblast–fibroblast heterotypic cell fusions[118]. Strategies that augment target cell fusion, when applied to the delivery of large DNAs, could facilitate future genome-scale engineering.

3.5. Conclusion

Multiplexed genome editing as enabled by CRISPR-based tools has the potential to transform our ability to study complex biological problems and enable sophisticated therapeutic modalities. Extending bacterial and yeast genome engineering protocols to the generation of edited mammalian donors may enable actual genome-scale engineering when combined with CRISPR-guided genomic integration. Multiplexing DNA synthesis, editing, assembly, and delivery technologies are at the core of streamlining large genome engineering such as the projects envisioned by GP-write and required for de-extinction efforts. These future applications require fundamental improvements and new developments to the effectors of editing, the production of donor material, and the delivery of both. In the coming years, progress on these fronts will foster a new era of genome biology, where researchers gain the ability to systematically alter genomes on a massive scale.

4. Developing Large-Scale Genome Editing Technologies

4.1. Introduction

The Human Genome Project completed the first draft of the human genome sequence in 2004. Since the initiation of this effort, DNA sequencing technologies quality and cost have improved exponentially: a whole human genome can now be sequenced in a few hours for a few hundred dollars while it took more than 20 years and about 3 billion dollars to complete the first human genome sequence. Now, even though the capacity to write DNA at the genome scale – including both large-scale DNA editing and synthesis – has greatly improved in recent years, it is still outpaced by the fulgurant development of high-throughput DNA sequencing. In this context, similar to the Human Genome Project, initiatives such as Genome Project Write (GP-Write), launched in 2016 aim to reduce drastically the cost of designing, synthesizing, assembling and testing genomes[119]. Magnifying our ability to write DNA could transform the field of human health by making possible the engineering of virus, cancer or aging-resistant cell lines, enabling the development of universal donor cell therapies[120], and generating xeno-compatible or synthetic organs[41] among other countless applications waiting to be tested. In addition, DNA writing could help the scientific community probe the physiological and pathological relevance of the "dark matter of the genome" – the non-coding sequences which include Transposable Elements (TE) - whose functions are still widely unknown but often associated with diseases[121,122].

In an early proof of concept demonstrating genome-wide recoding of an entire living organism by multiplex automated genome engineering (MAGE), all 321 occurrences of the UAG stop codon in *Escherichia coli* MG1655 were replaced with UAA stop codons[59]. The design, synthesis, and testing of an ongoing large-scale genome recoding project to remove a total of seven codons out of the 64 possible 3-letter codes, involving the alteration of ~62,214 codons is currently underway[60] and in

theory will provide pan-virus resistance by altering the highly conserved genetic code. Nonstandard amino acids could also be introduced along with synthetic derivatives aimed toward new functionality and control over synthetic circuits and biological systems[59]. To achieve a similar goal in human cells would require an estimated modifying 4438 to 9811 loci to recode all instances of one of the three stop codons[123]. As current DNA-editing technologies are unable to successfully re-write genomes at hundreds to thousands of loci, the recoding of mammalian organisms at such a scale poses a great challenge. Developing DNA editing tools capable of large-scale modifications could set forth a clear technological path towards achieving genome-wide recoding.

The discovery and widespread-implementation of the CRISPR/Cas system[24–26] has dramatically expanded the toolbox for genome engineering and has revolutionized the future prospects of basic biological research, data storage in living systems[124], agricultural science[125], and medicine[126]. One of the initial advantages of CRISPR/Cas based genome editors over previous approaches was the capacity to multiplex by simply using several gRNAs. This not only allowed libraries of guides to be screened in a single cell population but also allowed for the targeting of up to six independent loci at once[34], although the efficiency at each site decreased when compared to that of a single guide transfection. Recently a team reported the genome-wide knock-out of PERVs in a swine cell line at all their loci, representing ~62 genetic modifications in a transformed pig cell line[41]. Two years later a live pig was born with genome-wide KO of all 25 PERVs[127]. This advance widens the scope for large-scale editing of mammalian genomes, opening new possibilities for the study of higher copy number biological elements.

Developing genome editing tools capable of large-scale modifications could also lead to an improved understanding of the physio-pathology of TEs such as Alu[128], Long Interspersed Elements-1 (LINE-1)[129–131] or Human Endogenous RetroViruses (HERV)[132] by enabling causal investigation of their function. These DNA sequences are highly abundant and have homology to 45% of the human genome[133]. While

originally characterized as "junk DNA," these TEs have dynamically shaped the evolution of our genome and continue to replicate and insert themselves throughout the human genome today, activities which have been linked to human physiology and disease. LINE-1 sequences – which constitute about 17% of the genome – contains two open reading frames (ORFs), ORF-1 which binds the LINE-1 RNA and shuttles it back to the nucleus for retrotransposition, and ORF-2 which functions as an endonuclease and reverse transcriptase. LINE-1 expression is largely suppressed in most somatic cells[134], but can be highly active in neurons[135] and disrupt gene expression[131]. The hypothesis that they may have a role in neuronal diversity, brain development[129,136] and neurological diseases has thus been explored. Even though the co-expression of LINE-1 elements and neural differentiation factors has been described, it is still unclear whether such retrotransposons take advantage of a specific cell environment to duplicate themselves or whether LINE-1 is directly involved in these phenotypes. Creating knockouts using classical approaches to DNA editing to study such high copy number targets in mammalian genomes, however, is not feasible due to the high toxicity of DSBs[44] making their study challenging[137].

Before large-scale genome editors can be used to recode eukaryotic organisms or study high copy number TEs they will need to overcome two main hurdles: 1) the delivery of multiple gRNAs in a single large batch or over iterative treatments with subsets of targets; and 2) the cytotoxicity associated with genome wide DNA modifications[43]. While this study does not address the myriad challenges of gRNA delivery, we aim to tackle editing-associated cytotoxicity due to DSBs and SSBs generated by current DNA editors.

The recent development of DNA base editors by fusion of a deaminase to Cas9 enables gRNA targeted single nucleotide deamination for C:G base pair conversion to T:A using cytidine base editors[84] (CBEs) or A:T base pair conversion to G:C using adenine base editors (ABEs) within a specific target window[86]. Base editing has been broadly demonstrated with high efficiency in a range of species including human

zygotes[138]. Using properly designed gRNAs, C->T conversions may be used to generate stop codons to knock-out protein coding genes of interest[44]. Additional improvements in base editing purity – the frequency of desired base conversion within a target window – have been achieved by fusing bacterial mu-gam protein to the base editor to generate nCBE4-gam[139]. The first generation of CBEs used dead Cas9 (dCas9) as the targeting system, but low efficiencies caused a shift to nick-Cas9 (nCas9) in all generations beyond dCBE2 (*table 4.S1*). Here we propose the development of new base editors that retain mu-gam and the improved linker sequence and distribution but utilize dCas9 to prevent any toxicity from SSBs. We hypothesize this will improve the survival of highly-edited clones and allow us to push the upper limit of simultaneously-edited loci within a single cell.

This study aims to improve survival of human cells after large-scale genome editing and gauge the upper limit to their genetic amenability. We outline the difficulties of using CRISPR/Cas9 to edit TEs, primarily due to the highly toxic nature of cutting hundreds to millions of loci genome-wide. To stress-test the safety of our new DNA editors compared to exiting editing tools, we compared the tolerance of human induced pluripotent stem cells (iPSCs), and 293T after DNA editing with high copy number-targeting gRNAs. Samples were screened for targeted deamination, random indel mutagenesis and their capacity to form stable edited cell lines. We added a "survival cocktail" of small molecules and growth factors including bFGF and Pifithrin-alpha, an inhibitor of p53 in combination with currently-available and newly-developed DNA editors. Finally, we combined the best DNA editor and survival conditions to probe the feasibility of large-scale editing in human iPSCs.

4.2. Methods

4.2.1. Transposable element gRNA design

gRNAs targeting Alu were designed by downloading the consensus sequence from repeatmasker (http://www.repeatmasker.org/species/hg.html). LINE-1 gRNAs were designed based on the consensus of

146 “Human Full-Length, Intact LINE-1 Elements” available from the L1base 2[140]. HL1gR 1-6 were designed to generate stop codons from C->T deamination mutations. EN, RT and ENRT pairs of gRNAs were designed to create moderate size deletions (200-800bp) easily distinguishable from their wt full-length forms by gel visualization. HERV-W gRNAs were designed based on the consensus sequence of the 26 sequences identified by Grandi et al.[141] that can lead to the translation of putative proteins.

4.2.2. qPCR evaluation of copy number across repetitive element targeting gRNAs -

The qPCR reactions were generated using the KAPA SYBR FAST Universal 2X qPCR Master Mix (Catalog #KK4602) according to the manufacturer’s instructions. The LightCycler 96 machine from Roche was used to perform the qPCRs and the results were extracted using the LightCycler 96 SW 1.1 software. The following thermocycling conditions were used: "preincubation" stage = 95°C for 180 sec; "2-step cycling" stage: annealing = 95°C for 3 sec and elongation = 60°C for 20 sec; "Melting" stage was kept standard. The following primers were used to perform the qPCRs.

Primer name	**Sequence**	**Target**
ZY-JAK2-F	AGCAAGTATGATGAGCAAGC	JAK2
SB-JAK2-R	AAAACAGATGCTCTGAGAAAGGC	
P1(b)_REBE_F	TAGGAACAGCTCCGGTCTACA	LINE-1 promoter
P1_REBE-ilu_R	AATGCCTCGCCCTGCTTCGG	
P5_REBE-ilu_F	CCAATACAGAGAAGTGCTTAAAGG	LINE-1 ORF1
P5_REBE-ilu_R	CTTGGAGGCTTTGCTCATTTCT	
P7_REBE-ilu_F	CCCATCAGTGTGCTGTATTCAGG	LINE-1 ORF2
P7_REBE-ilu_R	GGCCTTCTTTGTCTCTTTTG	
P13_REBE-ilu_F	AACAGGCTCTGAAATTGTGGC	LINE-1 ORF2
P13_REBE-ilu_R	GCTGGCCTCATAAAATGAGTTAG	

P15_REBE-ilu_F	GTTCTGGCCAGGGCAATCAG	LINE-1 ORF2
P15_REBE-ilu_R	CCTGAGACTTTGCTGAAGTTGC	
P3_HERVWenv_F	AATACCACCCTCACTGGGCT	HERV-W env
P3_HERVWenv_R	CAGATTGGAAACAAGAGGTCC	

4.2.3. SpCas9 and gRNA plasmids used for genome editing

The following Cas9 plasmids were used: pCas9_GFP (Addgene #44719), hCas9 (Addgene #41815). Base editing plasmids used: pCMV_BE3 (Addgene #73021), pCMV_BE4 (Addgene #100802), pCMV_BE4-gam (Addgene #100806), ABE 7.10 (Addgene #102909). The gRNAs used in this study were synthesized and cloned as previously described[142]. Briefly, two 24mer oligos with sticky ends compatible for ligation were synthesized from IDT for cloning into the pSB700 plasmid (Addgene Plasmid #64046).

4.2.4. SaCas9 and gRNA plasmids used for genome editing

Cas9 plasmid: pX600-AAV-CMV::NLS-SaCas9-NLS-3xHA-bGHpA (Addgene #61592). Base editing plasmid: SaBE4-gam (Addgene #100809). The gRNAs used in this study were synthesized and cloned as previously described[143]. Briefly, two 24mer oligos with sticky ends compatible for ligation were synthesized from IDT for cloning into the BPK2660 plasmid (Addgene Plasmid #70709).

4.2.5. Maintenance and transfection of HEK 293T cells

HEK293T cells were obtained from ATCC with verification of cell line identification and mycoplasma negative results. They were expanded using 10% fetal bovine serum (FBS) in high-glucose DMEM with glutamax passaging at a typical rate of 1:100 and maintained at 37°C with 5% CO_2. Transfection was conducted using Lipofectamine 2000 (Thermofisher Catalogue # 11668019) using the protocol

recommended by the manufacturer with slight modifications outlined below. 24 hours before transfection ~1.0 x 10^5 cells were seeded per well in a 12-well plate along with 1 mL of media. A total of 2 μg of DNA was transfected using 2 μL of Lipofectamine 2000 per well. For Cas9 plasmids, the DNA content per well contained 1 μg of pCas9_GFP mixed with 1 μg of gRNA-expressing plasmid. For BE plasmids, 1.5 μg of BE was mixed with 0.5 μg of gRNA plasmid. In the dBE vs nBE comparison, Pifithrin-α (10 ng/μl) from Sigma-Aldrich P4359 (source # 063M4741V, Batch # 0000003019) was added to the media 30 minutes before transfection and maintained in the first day media change.

4.2.6. FACS Single cell direct NGS preparation

To quantify early genetic editing in cells transfected with Cas9/BE and gRNA expression plasmids, single cells were sorted and prepared as follows. Two days post-transfection, single cells were FACS-sorted into 96-well PCR plates containing 10 μL of QuickExtract™ DNA Extraction Solution (Epicentre Cat. # QE09050) per well and genomic DNA (gDNA) was extracted using the manufacturer's protocol. Briefly, the sorted plates were sealed, vortexed and heated at 65°C for 6 minutes then 98°C for 2 minutes. The NGS library was prepared as described later below.

4.2.7. Single cell clonal isolation and sequence verification

Single cells were FACS-sorted into flat bottom 96-well plates containing 100 μL of DMEM with 10% FBS and 1% Penicillin/Streptomycin per well. Sorted plates were incubated for ~14 days until well-characterized grown colonies were visible, with periodic media changes performed as necessary. To extract gDNA, the cells were first detached using 30 μL TrypLE™ Express (Thermofisher Cat. # 12604021), neutralized with 30 μL growth media, and then 4 μL of the resulting cell suspension was transferred to 10 μL of QE. Genomic DNA was extracted according to manufacturer's protocol, as described previously.

4.2.8. Nested PCR Illumina MiSeq library preparation and sequencing

Library preparation was conducted as previously described[144]. Briefly, genomic DNA was amplified using locus-specific primers attached to part of the Illumina adapter sequence. A second round of PCR included the index sequence and the full Illumina adapter. All PCRs were carried out using KAPA HiFi HotStart ReadyMix (KAPA Biosystems KK2602) according to the manufacturer's thermocycler conditions. Libraries were purified using gel extraction (Qiagen Cat. # 28706), quantified using Nanodrop and pooled together for deep sequencing on the MiSeq using 150 paired end (PE) reads.

4.2.9. NGS indel analysis

Raw Illumina sequencing data was demultiplexed using bcl2fastq. All paired end reads were aligned to the reference genome using bowtie2[145] and the resulting alignment files were parsed for their cigar string to determine the position and size of all indels within each read using a custom perl script. All indels that were sequenced in both forward and reverse reads were summed across all reads and reported for each sample along with total reads. Indels within a 30bp window from the 5' start of the gRNA proceeding through the PAM and extending an additional seven bp's (for a 20bp gRNA) were counted and summed for each sample.

4.2.10. Dual gRNA deletion frequency NGS analysis

Reads were analyzed for dual gRNA large deletions by detecting sequences in between the gRNAs to indicate full length unedited (at least not dual gRNA-edited) and sequences beyond the normal wild type amplicon that only appear when the deletion has occurred to identify deletion reads. The custom perl script used for analysis is available (sup. X)

4.2.11. NGS base editing deamination analysis

All paired end reads were aligned to the reference genome using bowtie2, and the resulting alignment files were converted to bam, sorted, indexed, and variant called using samtools[146]. All SNV data within a 30bp window from the 5' start of the gRNA proceeding through the PAM and extending an additional seven bp's (for a 20bp gRNA) is reported to analyze the editing window and purity of editing. A custom perl script used for analysis.

4.2.12. Automated CRISPR and Base Editing pipeline

Here we describe the steps followed in the automated CRISPR base editing pipeline. The input consists of a set of reference genomes R, a set of gRNA G, type of editors, E, with their window-specific details, and a set of samples S. The output is the comprehensive analysis of base editing in the window specific to a base editor for all samples. To achieve this, we align gRNA set G and samples S to reference genomes R using bwa. Furthermore, we sort the alignment outputs using Picard. To calculate the window specific to base editor, we consider the starting positions of G and add offsets and window sizes from E. We now precisely look into these specific windows and report the number of reads supporting different alleles. For indel analysis, we compute the reads with indels in these windows and report into the final analysis.

4.2.13. Site directed mutagenesis to remove remaining nick from base editors

We deactivated the remaining nuclease domain of Cas9 from (C)BE4-gam (Addgene #100806), pCMV-ABE7.10 (Addgene #102919), and SaCas9-BE4-gam (Addgene #?). Agilent QuikChange XL Site-Directed Mutagenesis Kit (catalogue # 200517) was used with the following primer sequences:

SaCas9-fwd – ATAACAAAGTTCTGGTTAAACAGGAGGAAGCCTCTAAAAAAGGGAACCGGACC

SaCas9-rev – GGTCCGGTTCCCTTTTTTAGAGGCTTCCTCCTGTTTAACCAGAACTTTGTTAT

SpCas9-fwd – TTTATCTGATTACGACGTCGATGCCATTGTACCCCAATCCTTTTTG

SpCas9-rev – CAAAAAGGATTGGGGTACAATGGCATCGACGTCGTAATCAGATAAA

4.2.14. Propidium Iodide and Annexin V staining and FACS analysis

Cells were dissociated with TrypLE, diluted in an equal volume of PBS, and then centrifuged at ~300g for 5 minutes at room temperature. We resuspended samples into 500µl PBS and half of the cells were pelleted for later gDNA analysis. The remainder was centrifuged and resuspended into 100µl of Annexin V Binding Buffer (ref #V13246) diluted into ultrapure water at a 1:5 ratio. Subsequently, we added 5µl of Alexa 647 Annexin V dye (ref #A23204) and incubated samples in the dark for 15 minutes. We then added 100µl of Annexin V Binding Buffer and added 4µl of Propidium Iodide (ref #P3566) diluted into the Annexin V Binding Buffer at a 1:10 ratio. Samples were incubated in the dark for another 15 minutes. Cells were washed with 500µl of Annexin V Binding Buffer and centrifuged again to be finally resuspended into 400µl of Annexin V Binding Buffer. All samples were filtered using a cell strainer and were run on the LSR 11 using a 70-µm nozzle. Analysis was conducted using FlowJo software.

4.2.15. Karyotype analysis of LINE-1 dBE-edited 293T single cell clones

Stable HEK 293T edited isolated cell lines (BE4-gam, dBE4-gam, ABE and dABE) were expanded and were karyotypically compared with the control groups and the wild type HEK 293T. Actively growing cells were passaged 1-2 days prior to sending to BWH CytoGenomics Core Laboratory. The cells were received by the core at 60-80% confluency. Chromosomal count, variances and abnormalities were investigated.

4.2.16. Maintenance and expansion of human iPSCs

Human iPSCs were cultured with mTeSR medium on tissue culture plates coated with Matrigel (BD Biosciences). For routine passaging, iPSCs were digested with TrypLE (Thermofisher # 12604013) for 5 minutes and washed with an equal volume PBS by centrifugation at 300g for 5 minutes. Digested iPSC pellets were physically broken down to form a single cell suspension and then plated onto Matrigel coated plates at a density of $3x10^4$ per cm^2 with mTeSR™ medium supplemented with 10µM Y-27632 ROCK inhibitor (R_i) (Millipore, 688001) for the first 24 hours.

4.2.17. Nucleofection in PGP-1 iPSCs

30 minutes prior to transfection media was changed to mTeSR supplemented with Pifithrin-α (10 ng/µl) from Sigma-Aldrich P4359 (source # 063M4741V, Batch # 0000003019); a notable spiky edge colony morphology was observed similar to when Ri is added. Human iPSCs were digested with TrypLE for 5 minutes and the single cells were washed once with PBS. (CS: 4 x 10^6, PK: 1 x 10^6) iPSCs were then re-suspended in 100 µl of P3 Primary Cell Solution (Lonza) supplemented with (CS: 13.5 µg, PK: 6.75 µg) of dABE plasmid, (CS: 4.5 µg, PK: 2.25 µg) of gRNA plasmid, and (CS: 2 µg, PK: 1 µg) pMax. The combined cells and DNA were then nucleofected in 4D-Nucleofector (Lonza) using the hES H9 program (CB150). The nucleofected iPSCs were then plated onto a single well of a 6-well Matrigel-coated plate in mTeSR medium supplemented with 10 µM R_i and Pifithrin-α (10 ng/µl).

4.2.18. Clonal isolation of PGP-1 iPSCs

96-well plates were coated with Matrigel (BD Biosciences) at a concentration of 50 µl/well. A cloning medium solution of 10% CloneR™ (StemCell Technologies #05888) and Pifithrin- α (10 ng/µl) in mTeSR™ was prepared and was added to the coated wells. Cells were digested using TrypLE, which was neutralized by an equal amount of cloning medium. The cell solution was then centrifuged at 300 x g for 5 minutes, the supernatant was aspirated, and the cell pellet was resuspended in the cloning medium. The cells were then

passed through a 40-µm cell strainer and were FACS-sorted into 1) individual wells containing warm cloning medium at a density of 1 cell/well and 2) 2 x 96-well PCR plates for direct NGS analysis. To prevent disturbance, there was no media change during the first 48 hours, and the plates were not removed from the incubator during this period. A half-medium change was performed on days 3 and 4 with cloning medium. The growing colonies were monitored and a mTeSR™ medium change was done daily for the following days until extracting the DNA using QuickExtract™ and proceeding with library preparation and sequencing.

4.3. Results

4.3.1. gRNA design and copy number estimation of transposable elements

To probe the limits of current editing technologies in a single transient round of modification for both efficiency and survival of edited clones, we designed and tested gRNAs against the TEs Alu, LINE-1, and HERV targeting a range of copy numbers from 30 to greater than 100,000 across the human genome (Fig. 4.1A). Alu and LINE-1 gRNAs were respectively designed on the consensus sequences obtained from repeatmasker[34] and on the consensus of the 146 full-length sequence that encodes both functional ORF1 and ORF2 proteins. Finally, gRNAs against HERV-W envelop particles were design on the consensus of putatively active retro-viruses[147].

In order to evaluate the total number of sites that we were targeting using the CRISPR-Cas9 systems, we performed PCRs on both HEK 293T and PGP1 genomic DNA (gDNA). The copy number of HERV-W, LINE-1, and Alu elements at the edited sites were respectively estimated at 36, 26,100 and 161,000 in HEK 293T; and 32, 19,000 and 124,000 in PGP1 iPSCs (Fig 4.1B). It is expected that the HEK 293T cell line contains even more relative transposable elements per cell since its genome is triploid (Fig. 4.S1) when PGP1 has a diploid karyotype. A complementary bioinformatic approach was taken to assess the targets' copy number by aligning the designed gRNAs to the human reference genome. We showed an example of one gRNA targeting LINE-1 ORF2 (HL1gR4) and plotted its distribution throughout every chromosome (Fig 4.1C). The total number of matches for HL1gR4 allowing two base pair mismatches is 12,657, about half the number than the qPCR estimate, with the vast majority having an intact PAM (Fig. 4.1D). The difficulty to sequence loci such as tandem repeats, centromeres or ribosomal repeat sequences may explain why the reference human genome underestimates the transposable elements copy number, therefore explaining the discrepancy between the two approaches. In the rest of the manuscript we will base our editing numbers based on the qPCR copy number estimate.

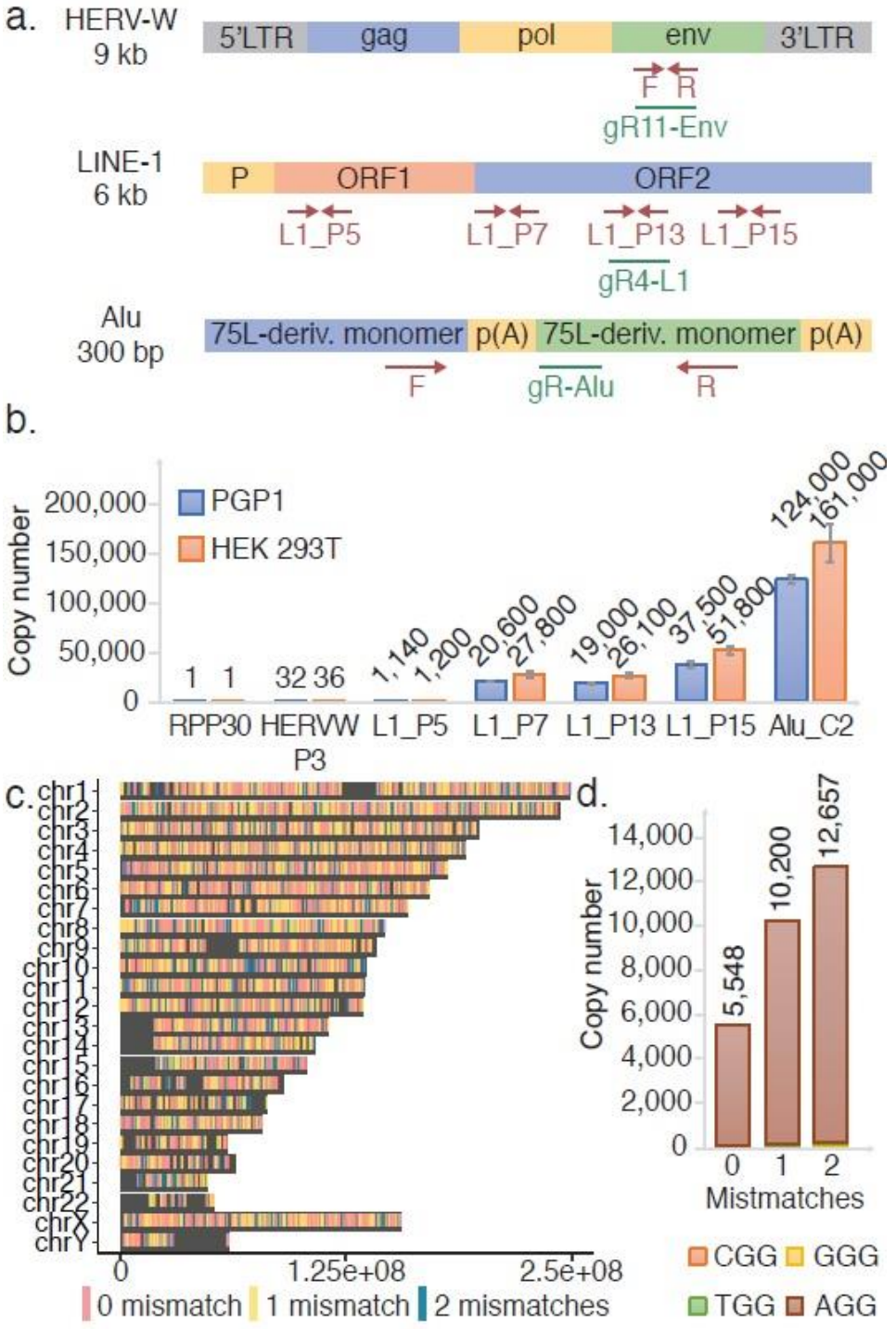


Figure 4.1 | Utilizing high copy repetitive elements for the development of an extremely safe DNA editor. **(A)** A summary of HERV, LINE-1, and Alu. Representation of TEs with qPCR primer sites shown in red and gRNAs shown in green. **(B)** qPCR estimation of LINE-1 copy number across the element compared to single copy number controls in PGP1 and HEK 293T. Errors bars display standard deviation, n=3. (**C**) Genome wide distribution of HL1gR4. **(D)** HL1gR4 copy number and PAM distribution.

4.3.2. CRISPR/Cas9 editing at a range of high copy number targeting gRNA does not allow the isolation of stably edited clones

HEK 293T cells were transfected with plasmids expressing pCas9_GFP and LINE-1 targeting gRNAs to disrupt the two key enzymatic domains of ORF-2: endonuclease (EN) and Reverse transcriptase (RT) (Fig. 4.2A). Three days after transfection, we observed indel frequencies at the LINE-1 expected targets ranging from 1.3% to 8.7%, corresponding to an average of respectively 339 and 2271 edits per haploid genome in the population (Fig. 4.2B). This degree of genetic alteration has previously been reported to be toxic[29] which we confirmed with a Propidium Iodide cell death assay: Cas9 targeting of LINE-1 showed a 7-fold increase in apoptosis as compared to the control (fig. S16). We then conducted a time course experiment to determine the long-term survival and stability of these cells first observed to have edits at hundreds of loci genome-wide.

In this experiment, we transfected pairs of LINE-1 gRNAs targeting the EN, RT or both (ENRT) domains. The use of dual gRNAs causes large deletions (~300-800 bp) that can be detected through gel visualization since the deletions shift the products' electrophoretic motility (fig. 4.S2). While samples from day two through five show clear editing with the expected deletion band sizes (Fig. 4.2C), they were not detectable anymore at day 9 and 14 indicating that mutated cells either died out as suggested by our previous cell death assay or were overgrown by wild type cells. The deep sequencing of expected dual gRNA deletion bands confirmed the LINE-1 gRNA breakpoints. While there were no visible bands at day nine and 14, we thought rare clones may be surviving and decided to repeat the experiment followed with clonal isolation. After early indications of editing no clones had detectable mutations at day 12 and beyond (data not shown) indicating that any significant level of indel activity at LINE-1 is toxic or limits growth and clonal isolation.

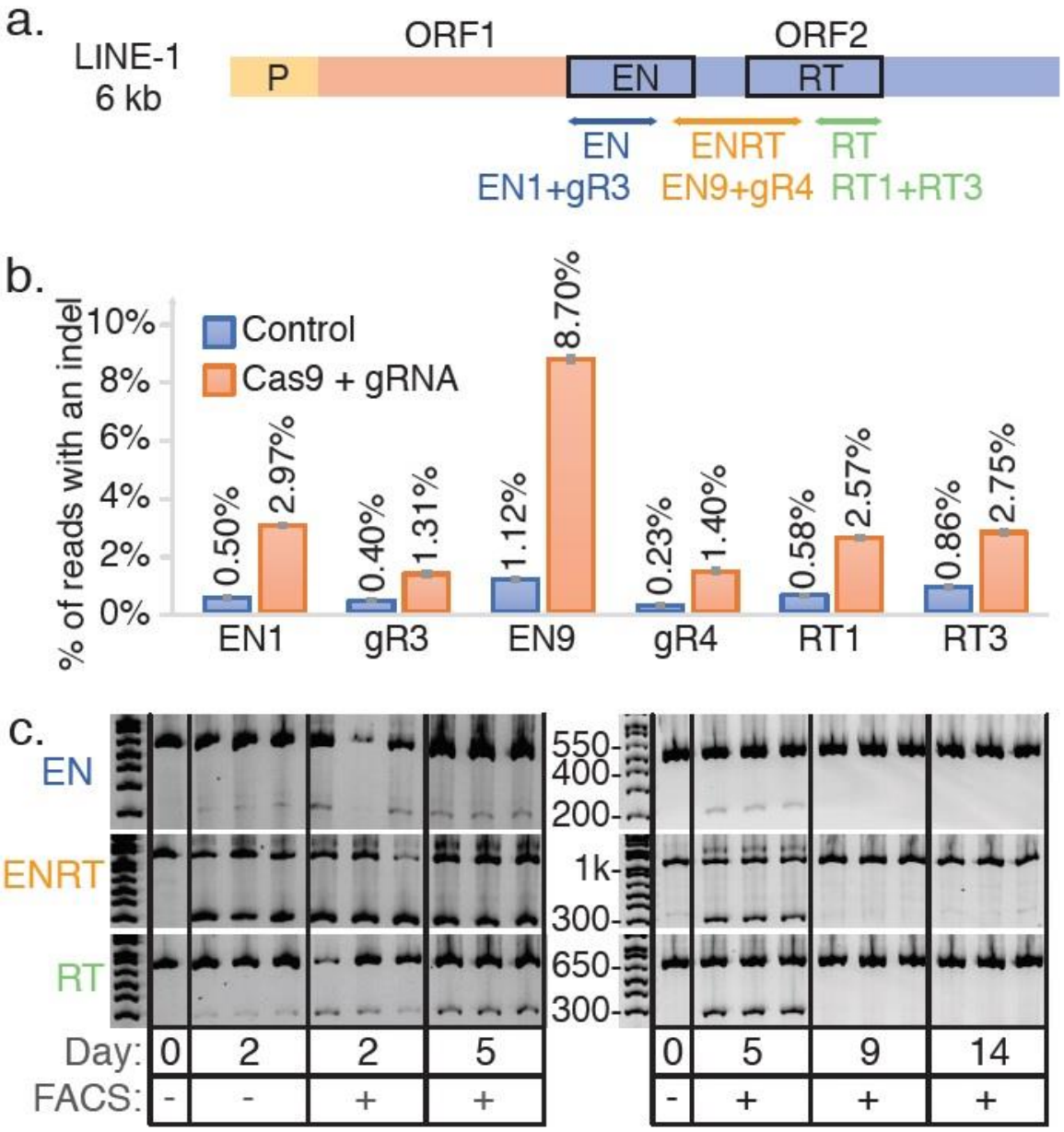
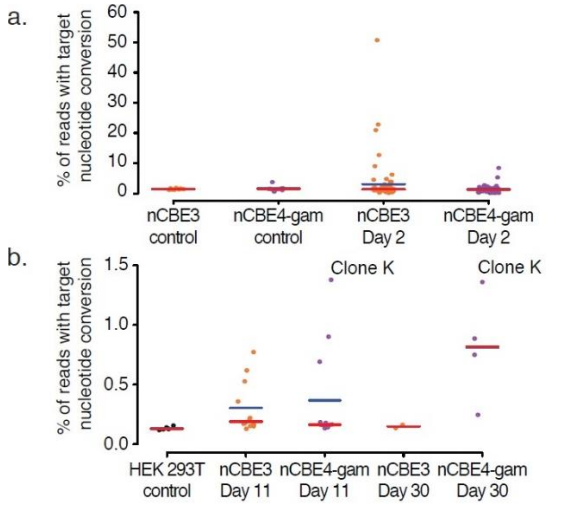


Figure 4.2 | CRISPR-Cas9 based genome editing at high copy number repetitive elements is detectable but ultimately lethal. (A) Schematic of LINE-1 including the two protein coding genes ORF-1 and ORF-2. Three dual gRNA deletions were designed to disrupt the EN and RT domains of ORF-2. **(B)** LINE-1 gRNAs transfected with Cas9. Displayed are single transfections with 95% confidence intervals for a proportion as the error bars. **(C)** Gel image visualizing dual gRNA deletion bands compared to wild type control bands.

4.3.3. nCBE and nABE activities confirmed at LINE-1

We next hypothesized that current nicking base editor technologies (nBEs) may help to overcome the viability of LINE-1 edited clones. We therefore designed and tested LINE-1 targeting gRNAs (HL1gR1-6 [Fig. 4.S3A]) that generate a STOP codon early in ORF-2 using C->T deamination. HEK 293T cells were transfected with nCBE3 and each designed gRNAs. Deamination events were detected at each of the six gRNA target loci above background levels (~0.05% – 0.67%) determined by a population of cells that underwent a mock transfection (fig. 4.S3A). These same CBE gRNAs were also compatible with ABEs as they contain at least one adenine within their deamination window. Base editing with nCBE4-gam and nABE was detected from genomic DNA in 4/5 gRNAs for CBE (fig. 4.S3B) and 4/5 gRNAs for ABE (fig. 4.S3C). nABE had the highest editing efficiency using HL1gR6 at 4.94% or ~1290 loci genome wide 3 days after transfection. Yet, HL1gR4 was chosen as the best target for future studies as its signal to background error ratio was the lowest its efficiency among the highest of all the LINE-1 amplicons/gRNAs tested. The HL1gR4 target also contains three C's within its target window that are all efficiently co-edited as a clear watermark signal of directed mutation. An Alu targeting gRNA showed increased cell survival when using nCBE3 compared to Cas9 (Fig 4.S10).



c.

Control

	Day 2							Day 30						
	T_3	C_4	T_5	A_6	C_7	C_8	A_9	T_3	C_4	T_5	A_6	C_7	C_8	A_9
G	0.3	0.5	0.3	0.6	0.5	0.5	0.7	0.1	0.2	0,0	0.6	0.4	0.3	0.4
A	0.3	0.6	0.4	98.4	1.1	1.1	98.6	0.1	0.1	0.1	98.9	0.9	0.3	99.2
T	98.9	1.1	97.9	0.4	4.4	1.2	0.5	99.6	0.4	99.3	0.2	3.7	0.2	0.3
C	0.4	97.6	1.3	0.4	93.7	97.3	0.2	0.2	99.2	0.5	0.2	95	99.2	0.1

nCBE3

	Day 2							Day 30						
	T_3	C_4	T_5	A_6	C_7	C_8	A_9	T_3	C_4	T_5	A_6	C_7	C_8	A_9
G	0.3	1.1	0.1	0.8	3.0	1.2	0,7	0.1	0.3	0.1	0.5	0.3	0.3	0.3
A	0.4	0.7	0.5	98	1.5	1.5	98.1	0.2	0.2	0.1	98.8	0.9	0.4	99.2
T	98.9	53.0	98.4	0.7	51.2	50.7	0.7	99.5	0.5	99.2	0.2	3.7	0.3	0.3
C	0.3	43.9	0.8	0.2	44	47	0.1	0.2	99	0.6	0.3	97	98.9	0.1

nCBE4 gam

	Day 2							Day 30						
	T_3	C_4	T_5	A_6	C_7	C_8	A_9	T_3	C_4	T_5	A_6	C_7	C_8	A_9
G	0.3	0.5	0.2	0.9	0.7	0.9	0.6	0.1	0.2	0.1	0.6	0.4	0.3	0.3
A	0.6	0.7	0.4	98.1	1.2	0.6	98.5	0.1	0.1	0.2	98.9	0.8	0.4	99.3
T	98.5	9.9	98.3	0.4	10.8	8.5	0.5	99.5	1.9	99.3	0.2	4.8	1.3	0.3
C	0.6	88.9	1	0.5	87.2	90.1	0.2	0.2	97.7	0.5	0.2	93.9	98	0.1

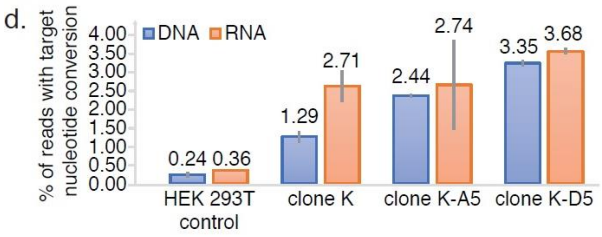


Figure 4.3 | nBEs targeting LINE-1 enables survival of stable cell lines with hundreds of edits. (A) Base editing in HEK 293Ts two days after transfection comparing nCBE3 vs nCBE4-gam. FACS single cells are plotted as individual points representing targeted base editing nucleotide deamination. Red line indicates the median and the blue line the mean. **(B)** Single cell live culture growth and stable cell line generation at day 11 and 30. **(C)** Base editing activity across the CBE target window of ~3-9. Comparing day two and 30 for analysis of initial editing activity in most highly edited clones. **(D)** LINE-1 deamination analyzed from either RNA or genomic DNA. SEM's are displayed as error bars, n=2.

4.3.4. nCBEs enable isolation of stable cell lines with hundreds of edits

Using the confirmed nCBE activity at LINE-1 we aimed to derive stable cell lines with as many edits as possible to probe the boundaries of multi-locus DNA editing, understanding the limitations and expected toxicity of nicking the genome at potentially tens of thousands of loci along every chromosome. 293Ts were transfected with HL1gR4 and either CBE3 or CBE4-gam with control samples receiving a non-targeting gRNA. We hypothesized that the bacterial mu-gam protein that binds DNA after DSBs and has been reported to increase purity after base editing may play a beneficial role in survival after large-scale genome engineering. Two days post-transfection, single cells were FACS sorted into 96-well plates for direct NGS analysis resulting in a high editing efficiency of up to 53.9% C to T conversion, or an estimated 14,000 loci (Fig. 4.3A), in the most highly edited single cell. CBE3 had a significantly higher mean deamination frequency than CBE4-gam at this early timepoint. A parallel plate containing growth media was sorted to assess viable colony formation and the edited 293Ts' capacity to form a stable cell line. While both CBE3 and CBE4-gam had edited cells at day 11, all cell lines with CBE3 edits died before analysis could be conducted at day 30. Four surviving cell lines were isolated with deamination frequencies up to ~1.37 % of LINE-1 or an estimated ~356 sites (Fig 4.3B). Data presented in Fig. 4.3C shows both the purity of the desired deamination products and the editing window. Clone K was the most highly edited single cell isolated and was stable in terms of target C to T mutation frequency from day 11 to 30 across multiple independent PCR replicates at each time point.

By subjecting the top edited single cell isolate clone K to another round of CBE4-gam editing (fig. 4.S4A) we detected cells with up to 36.26 % C to T deamination were detected on day 2, and four living clones with deamination frequencies ranging from 2.43% to 5.04% – corresponding to about 643 to 1315 edits – were isolated (fig. 4.S4B). While the clone with the highest number of deaminated sites did not grow after a freezing and thawing cycle, the three other cell lines were stable in culture for a period

longer than 30 days, and were termed “clone K-A5”, “clone K-A2” and “clone K-D5”, with respectively 643, 749, and 781 edits. This observation of the highest edited clone dying off after initial detection was observed for all types of editors. We confirmed nBE activity at the lower copy number target HERV (fig. 4.S5). Due to the difficulty amplifying and analyzing the Alu target likely because of high subfamily polymorphism and short repeat sequence (290bp) we proceeded exclusively with LINE-1 targeting gRNAs for the rest of the study.

To confirm that LINE-1 editing at the genome level had a repercussion on the corresponding transcripts we performed RNA-seq on clone K, clone K-D5, and cloneK-A5 and analyzed the percentage of C to T conversion resulting in a stop codon in ORF2 in the RNA reads (Fig. 4.3D). Theoretically, since most of the active LINE-1 subsets should generate transcripts, the presence of the expected STOP codon at the messenger RNA level may indicate the inactivation of these elements. The results showed that a higher number of edits in the clones was correlated with a higher number of STOP codons at the RNA level, suggesting that potentially active LINE-1 were impacted by the multiplexed editing.

4.3.5. Nick-less dBE confirmation at a single locus

Suspecting that generating single-stranded nicks genome-wide could lead to cytotoxicity, we decided to inactivate the remaining HNH nuclease domain of Cas9 by an H840A mutation in the Cas9 backbone and generate a set of dCas9-BEs including dCas9-CBE4-gam (dCBE4-gam), dCas9-CBE4 (dCBE4), and dCas9-ABE (dABE). Nick-less dCas9-BEs were tested on single-locus targets to confirm their deamination activity and compare them to their nBEs equivalents and the existing dCas9-CBE2 (dCBE2). dCBE4 and dCBE4-gam showed a 2.38- and 2.29-fold improvement in editing efficiency over CBE2 in 293Ts at day five respectively (fig. 4.S6A). Compared to their nicking counterparts this was a 34.7% or 53.2% reduction in efficiency but indel activity was reduced to background levels (fig. 4.S6A).

dABE had no previous dead counterparts to compare to but retained 40.2% of nABE's deamination efficiency at a single locus control while reducing indel levels to background (fig. 4.S6B).

4.3.6. Nick-less dBE targeting of LINE-1 in 293T

We then transfected 293T cells with HL1gR4 and either nCBE4-gam, dCBE4-gam, nABE, or dABE that were individually sorted and analyzed for target nucleotide deamination 2 days after transfection. Single edited cells resulted in high editing efficiency of up to 54.9% with nCBE4-gam, or 14,300 loci, when we observed significant reductions to mean target nucleotide deamination frequency with dCBE and dABE when compared to their nBE equivalents (Fig. 4.4A). In parallel, single cells were grown to determine whether viable highly edited clones could be isolated. The editing efficiency trend reversed in live cells: dBE showed a significantly increased deamination frequency over nBE (Fig. 4.4B). dABE produced the mostly highly edited clone with 50.61% targeted nucleotide deamination frequency or an estimated 13,200 loci. Base editors that retain nicking activity only generated a few rare cells with an editing frequency consistent with our prior experiments in Fig. 4.4B. Results were replicated using another LINE-1 targeting gRNA and similar trends were observed (fig. 4.S7).

The nucleotide composition of all bases in the gRNA and PAM are displayed for the most highly edited clone and parental 293T control for each BE condition used, indicating some non-specific nucleotide conversions for both nCBE and dCBE but not nABE or dABE (fig. 4.S8). The mean single cell deamination frequency was reduced from 5.32% using nABE to 1.45% using dABE, indicating that retaining the nick and using nABE resulted in a 3.67-fold decrease in editing efficiency at the early timepoint (Fig. 4.4B). The tables are turned in terms of cell viability at day 14, where dBEs gain a marked advantage in the total number of live cells, editing frequency of single cells, and mean target deamination frequency. There was a 14.8-fold increase in mean editing frequency among surviving live clones when using dABE compared to nABE (Fig. 4.4B). A 2.38-fold increase was also observed for dCBE4-gam compared to

nCBE4-gam. The deamination editing window for the most highly edited clones are displayed in fig. 4.S8. These demonstrate a high base editing purity and no detectable nucleotide conversion beyond the expected range. Similar purity and base editing window results were observed in bulk transfected cells though day ten (fig. 4.S9). During the first three days of editing the dBEs have lower editing frequency when compared to nBEs but after day seven and ten dABE gains a significant edge over nABE. (Fig. 4.4C).

Chromosomal integrity analysis was performed for clones edited at LINE-1 with nABE, dABE, nCBE4-gam, and dCBE4-gam. The karyotype results are shown in table 4.S2 and show that the top edited clones are not significantly different than control groups in terms of total number of aberrations. Further analysis in a karyotypically normal and stable cell line is required to fully assess chromosomal stability after large-scale genome editing.

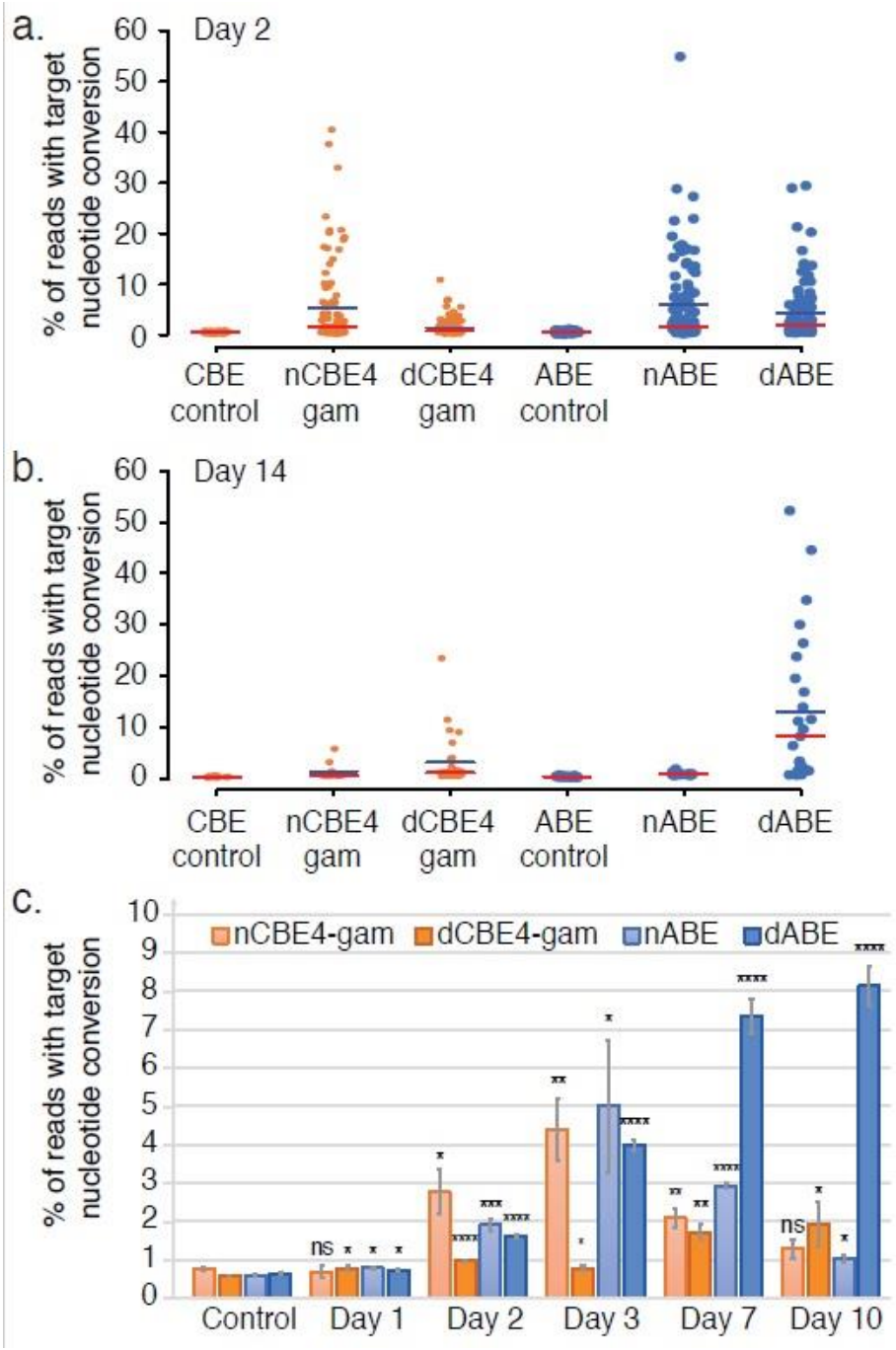


Figure 4.4 | dBEs improve survival of highly edited cells with thousands of edits genome wide. (A) nBE compared to dBE in 293T single cells each represented as a single data point. Base editing is displayed as either target C->T or A->G conversion for CBE and ABE respectively. The red line indicates the median and the blue line the mean. **(B)** Live single cell analysis at day 14 of the same experiment. **(C)** Deamination frequency over time comparing dBE to nBE from day one to ten. Error bars represent SEM, n=3.

4.3.7. dABE activity in PGP1 iPSCs

We next attempted the large-scale genome editing of PGP1 induced pluripotent stem cells (iPSCs). The survival cocktail and single cell isolation time line is shown in Fig. 4.5A. The same experiment was conducted with two slight variations of the electroporation protocol differed in terms of total cells transfected and the total amount of DNA used (CS and PK conditions). Single cells were sorted and analyzed for target nucleotide deamination frequency 18 hours post electroporation. The highest edited single cell had ~6.96% target A to G conversion or ~1320 sites (Fig. 4.5B). In parallel live single cells were isolated after stable cell lines formed at 11 days after transfection. Colonies were analyzed for targeted LINE-1 A to G deamination with a 1.30% and 0.96% editing frequency for CS and PK conditions respectively (Fig. 4.5C). The median editing efficiency of the CS live clones was higher than that of PK live clones in contrast to the value observed at the earlier time point, suggesting that lower editing efficiency in earlier time points may increase the viability of stably edited cell lines. The most highly edited clone had a deamination frequency of 13.75% which corresponds to 2600 sites genome wide, exceeding by three order of magnitude the number of simultaneous edits previously recorded in iPSCs[35]. The increased background that occurs in single cell direct analysis Fig. 4.5B compared to isolation from an expanded colony Fig. 4.5C is likely due to the necessary over-amplification required to get enough genomic material from a single cell. Similar observations were made in previous experiments using 293T cells. All other previously tested DNA editors failed to produce any detectable edits at the LINE-1 locus in human iPSCs which are prone to apoptosis after even minor DNA damage and rapidly deplete cells transfected with Cas9 and TE gRNAs (fig. 4.S15).

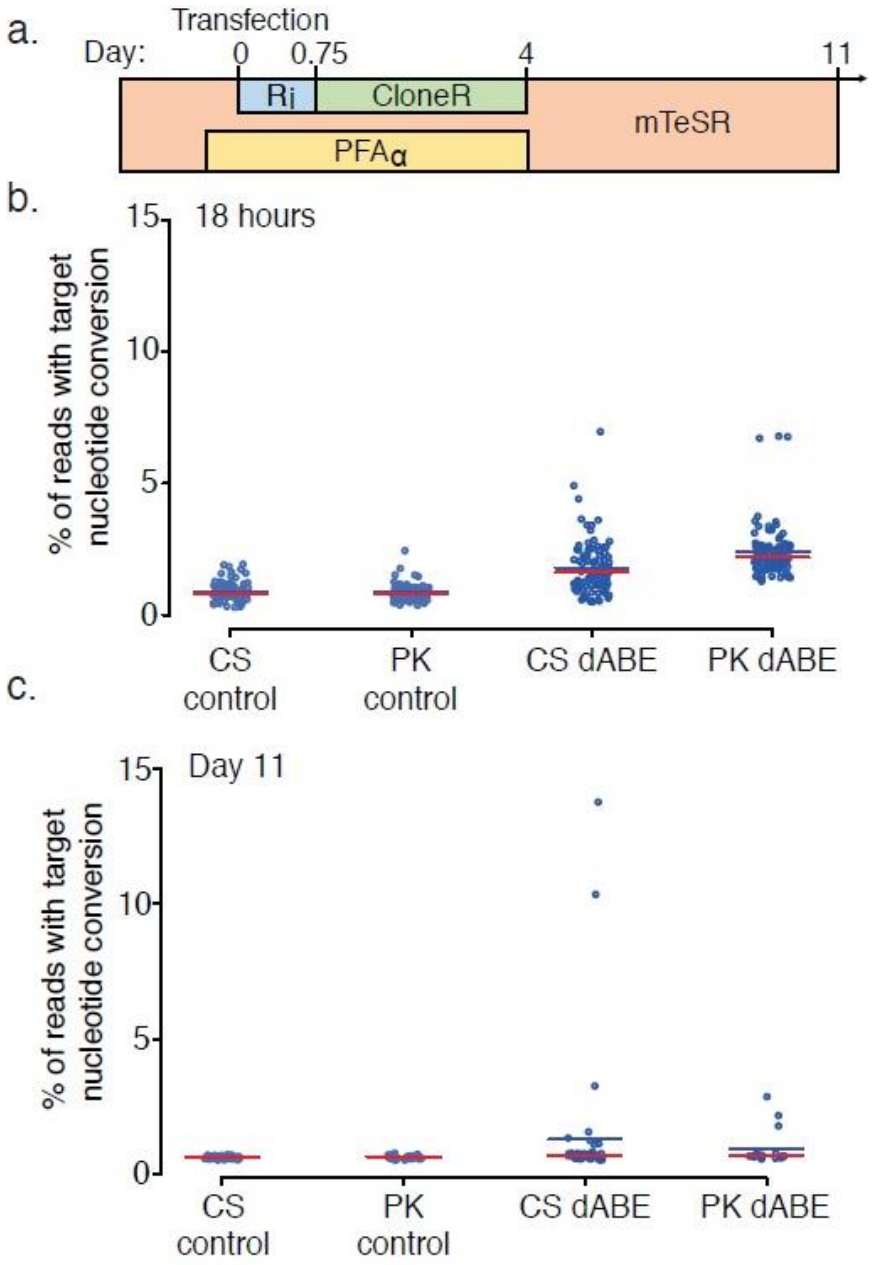


Figure 4.5 | Survival cocktail and conditions for clonal derivation of iPSCs after large-scale genome engineering. (A) Human iPSC transfection timeline and survival cocktail conditions. **(B)** Eighteen-hour single cell direct NGS analysis of dABE targeting LINE-1. CS and PK indicate the two researchers who conducted the experiments. The red line indicates the median and the blue line the mean. **(C)** Live cell colony analysis of surviving iPSCs at day 11 post transfection.

4.4. Discussion

CRISPR has recently brought a radical transformation in the basic and applied biological sciences, leading to commercial applications, a multitude of clinical trials[36], and even the controversial tests of human germline modification[37–41]. While the use of CRISPR and its myriad derivatives has greatly reduced the activation energy and technical skill required to perform genome editing several barriers need to be overcome before its full potential can be properly realized: 1) the need for custom RNA, and perhaps DNA for each target, 2) difficult delivery, 3) inefficiencies once delivered, 4) off-target errors, 5) on-target errors, 6) the toxicity of DNA damage, 7) the challenge of multiplexing beyond 62 loci[3], 8) the limitation of insertion sizes below 7.4kb[42], 9) immune reactions to Cas, gRNA and vector. This study aims to develop tools that address weaknesses 5, 6, and 7.

Improving the actual multiplexed eukaryotic genome editing capabilities by several orders of magnitude holds the potential of revolutionizing human health. Combinatorial functional genomic assays would enable the study of complex genetic traits with applications in evolutionary biology, population genetics, and human disease pathology. In addition, analyzing the functional significance of any generated set of mutations through editing would empower the field of cancer biology. Multiplex editing has also permitted the development of successful engineered cell treatments such as the chimeric antigen receptor (CAR) therapies, which require the simultaneous editing of three target genes. Future treatments may require many more modifications to augment cancer immunotherapies, slow down oncogenic growth, and reduce adverse effects such as graft versus host disease. Furthermore, customizing host-versus-graft antigens in human- or nonhuman- donor tissues may require more modifications than have been done so far, for which the development of genome-wide editing technologies is needed. Special attention will be required to the safety of the editing and its impact on the functional activity of the transplants, since donor tissues may persist in the patient for decades.

To complete genome-wide recoding and enable projects such as GP-write ultra-safe cells1, the de-extinction efforts to regain the lost biodiversity, or the codon reduction to confer pan-virus resistance, safe DNA editors must be developed to increase the number of genetic modifications to several orders of magnitude without triggering overwhelming DNA damage, as well as overcoming the delivery of multiple distinct gRNAs per cell, the latter of which we do not address is this study. C321.ΔA is a massively modified strain of E. coli MG1655 has all instances of the Amber stop codon replaced and has shown to be resistant to a range of viruses[6]. To attempt such a feat on the human genome, 4438 Amber codons[8] will require to be modified according to a simple analysis of the human genome reference. We have shown that gene editors that do not cause double- or single-stranded DNA breaks can generate a number of edits sufficient to theoretically achieve this genome recoding and pave the way towards making pan-virus resistant human cells. This could have commercial application towards cell-based production of monoclonal antibodies, recombinant protein therapeutics, and synthetic meat production.

As our study demonstrates, genome wide disruption of high copy number repetitive elements is now possible and opens new opportunities to study the “dark matter” of the human genome. CBEs that allow the generation of STOP codons within an open reading frame will be a great tool to probe at the functions of transposable elements, potentially turning observed associations with physio-pathological phenotypes into causations. For instance, large-scale inactivation of HERV-W and LINE-1 elements could help investigate their respective role in multiple sclerosis and neurological processes.

However more in-depth studies will be necessary to assess the impact of this massive editing on normal cell processes, since collateral damage may occur. We expect the thorough on- and off- target analysis at repetitive elements to remain a difficult task to accomplish due to their high level of polymorphism, therefore, strong biological controls as well as new experimental and bioinformatics pipelines will be needed to overcome such a challenge.

In our study, we observed that dABE increases the viability of highly edited clones as compared to dCBE. This difference may be explained by two factors: First, when using HL1gR4, CBE has three target nucleotides within its deamination window as compared to one for ABE, and as a consequence, CBE converts three times more nucleotide than ABE, potentially causing additional cytotoxicity. Second, when using CBE, the uracil N-glycosylase (UNG) actively catalyzes the removal of the deaminated cytosine, generating several nicks genome-wide that promotes DNA damage and potential cell death. The conversion of adenosine into inosine using ABE may not be detected as efficiently by the DNA repair machinery therefore increasing the viability of large-scale editing. That is why, we anticipate the conditional modulation of DNA repair processes such as mismatch repair (MMR) or base excision repair (BER) – that trigger downstream single- and double-stranded breaks in the genome – to further improve the extent of dBEs performance.

Finally, since dBEs do not generate direct breaks into the genome, they decrease indel frequency to background and may not trigger DNA sensors such as p53, while retaining about 34% to 53% deamination frequencies as compared to their nBE counterparts. As a consequence, successful genetic modifications with dBEs may not enrich for pro-oncogenic cells that have disrupted DNA-damage guardians as it has been reported for Cas9[43]. Even at low level of multiplexing, this feature may promote dBEs as an essential tool for therapeutic applications such as gene therapies.

In summary, this work optimized large-scale genome editing to enable cell viability after the simultaneous editing of thousands of loci per single human cell. The ability to safely edit many loci genome wide may facilitate the true potential of personalized medicine as we further develop our understanding of gene interactions and epistasis. We envision these new safe DNA editors to be combined with further improvements in multiplex delivery of gRNAs to usher in a new phase of synthetic biology where it is possible to imagine recoding whole mammalian genomes. When combined with further modulation of DNA repair and pro-survival factors there may be no practical

limit to the number of bases that can be altered in a single round of editing, opening up new possibilities that were previously not thought possible. We have overcome the toxicity limitation that prevented large-scale genome editing in human iPSCs by expanding its boundary by three orders of magnitude. The continued development of multiplex delivery along with non-toxic, high-efficiency DNA editors without DSBs or SSBs is paramount to the success of genome-wide recoding efforts to probe the inner workings of life itself, ultimately leading to the radical redesign of nature and ourselves.

4.5. Supplementary Figures and Tables

	C-deaminase	A-deaminase	UGI	Nick	Mu gam
BE1[1]	X				
BE2[1]	X		X		
BE3[1]	X		X	X	
BE4[2]	X		X2	X	
BE4-gam[2]	X		X2	X	X
dBE4*	X		X2		
dBE4-gam*	X		X2		X
ABE[3]		X		X	
dABE*		X			

*Synthesized and tested in this study

[1]Komor et al. (2016) *Nature* **533**(7603):420-4

[2]Komor et al. (2017) *Sci Adv* **3**(8):eaao4774

[3]Guadelli et al. (2017) *Nature* **551**:464-71

Table 4.S1 | Evolution of Base Editors variants

	BE4_2_A11	BE4_C2_A7	dBE4_3_C6	dBE4_C1_B2	ABE_2_A4	ABE_C2_B9	dABE_2_E7	dABE_C1_E2	293T9(CYG-18-PK-0040)
-X				x	x		x	x	
add(X)(q28)	x	x	x		x	x			x
der(X)add(X)(p11.2)add(X)(q28)								x	
add(1)(p36.1)	x	x	x	x	x	x	x		x
add(1)(q42)	xx	xx	xx	xx	xx	xx	xx		xx
del(1)(q31)	x	x	x	x		x	x		x
i(1)(p10)								x	
add(1)(q21)								x	
-2									
add(3)(p13)								x	
add(3)(p24)			xx						x
del(3)(p22)	x				x				
add(3)(q12)			x	x					x
del(3)(q22)	x				x	x		x	
add(4)(p15)	x								
del(4)(q31)	x								
-4		x			x	x	x		x
add(8)(p21)	x	x	x	x	x	x	x		x
-9						x			
add(10)(p11)									
add(10)(p13)	x	x	x	x	x	x	x		x
add(11)(p15)			x		x				
add(13)(p11)	xx	xx	xx	xx	xx	xx	xx		xx
add(13)(q34)	x	x	x	x	x	x	x		x
-13									
add(14)(p11.2)			x	x				x	
-15	x	x	x	x	x	x	x		x
add(15)(p11.2)		x							
-18	x	x	x	x	x	x	x		x
-21	x	x		x	x	x	x		x
-22							x		
i(21)(q10)			x						
mar	x-xx	x-xx	x-xxx	x-xx	xx-xxx	x-xx	x-xxx	x-xxx	x-xxxx

Table 4.S2 | Karyotype chromosomal abnormality list

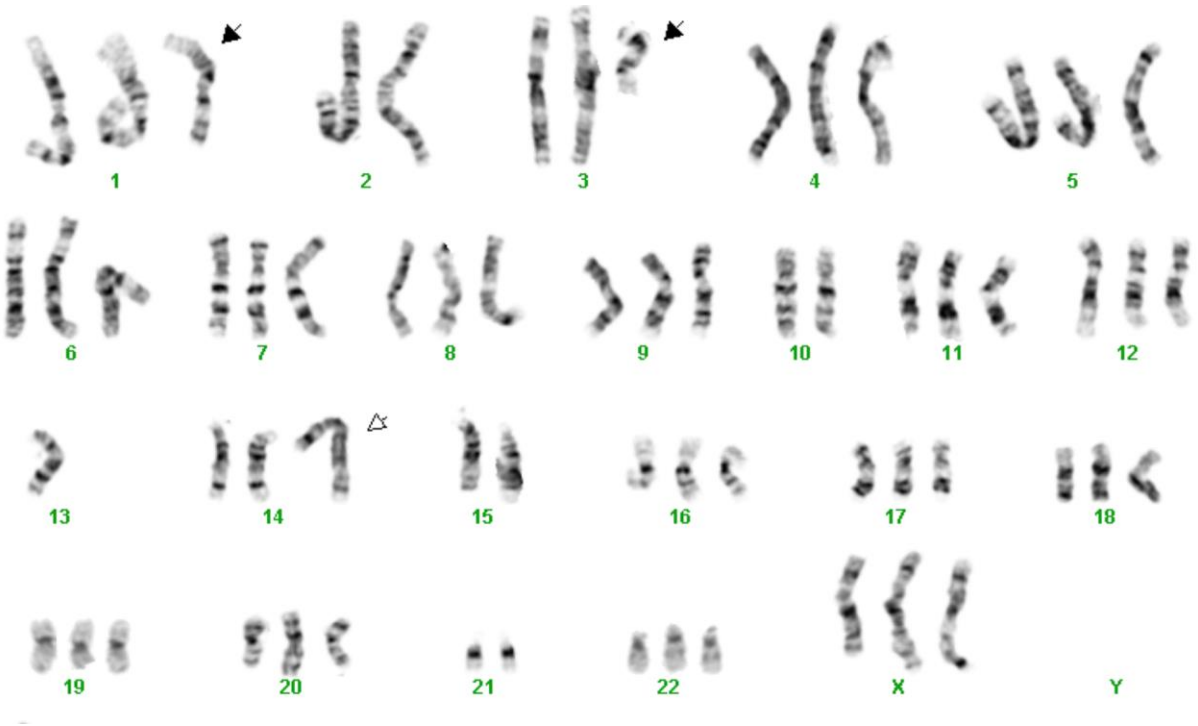


Figure 4.S1 | Karyotype of HEK 293T

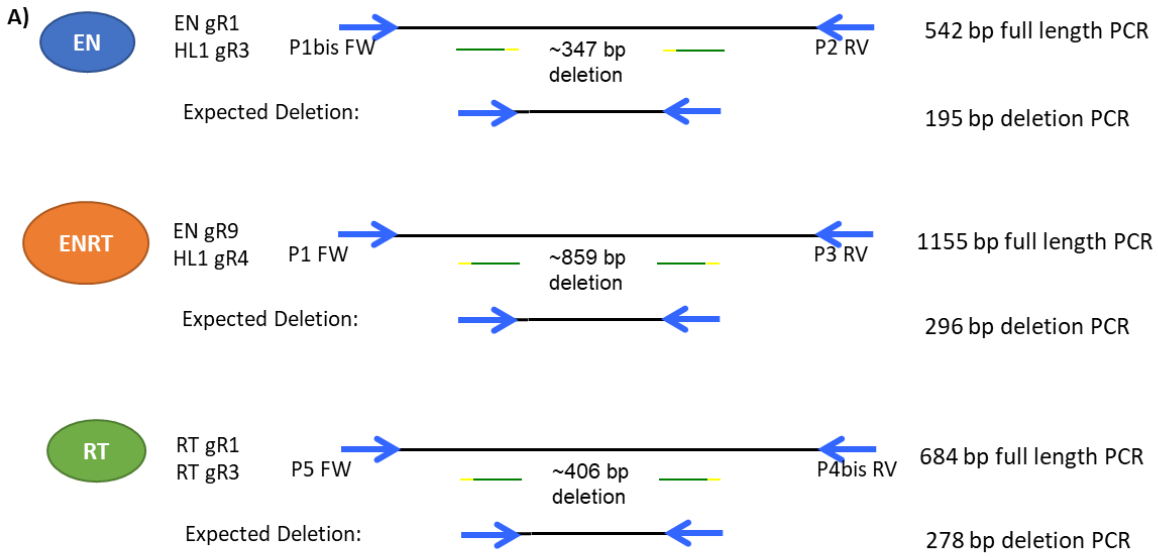


Figure 4.S2 | dual gRNA LINE-1 deletions

A)

HL1 gR1

5′-AACGAGACAGAAAGTCAACAAGG-3′
ABE 5′-AACGGGGCAGAAAGTCAACAAGG-3
CBE 5′-AACGAGATAGAAAGTCAACAAGG-3′

HL1 gR2

5′-CCGCTCAACTACATGGAAACTGA-3′
ABE 5′-CCGCTCAACTACATGGAAACCGA-3
CBE 5′-CCGCTCAACTACATAAAAACTGA-3′

HL1 gR3

5′-CCTGAATGACTACTGGGTACATA-3′
ABE 5′-CCTGAATGACTACTGGGCACATA-3
CBE 5′-CCTGAATGACTACTAAATACATA-3′

HL1 gR4

5′-ATTCTACCAGAGGTACAAGGAGG-3′
ABE 5′-ATTCTGCCGGAGGTACAAGGAGG-3
CBE 5′-ATTTTATTAGAGGTACAAGGAGG-3′

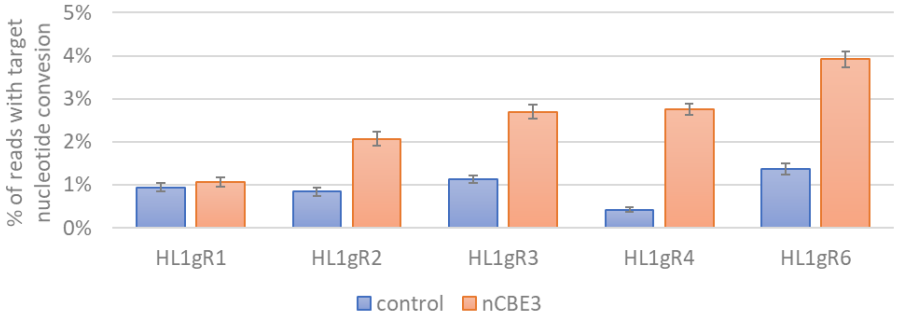
HL1 gR5

5′-CCTGGGATGCAAGGCTGGTTCAA-3′
ABE 5′-CCTGGGATGCAAGGCCGGCCCAA-3
CBE 5′-CCTGGGATGCAAGGCTAATTCAA-3′

HL1 gR6

5′-GGGTATTCAATTAGGAAAAGAGG-3′
ABE 5′-GGGTGTTCAATTAGGAAAAGAGG-3
CBE 5′-GGGTATTTAATTAGGAAAAGAGG-3′

B)



C)

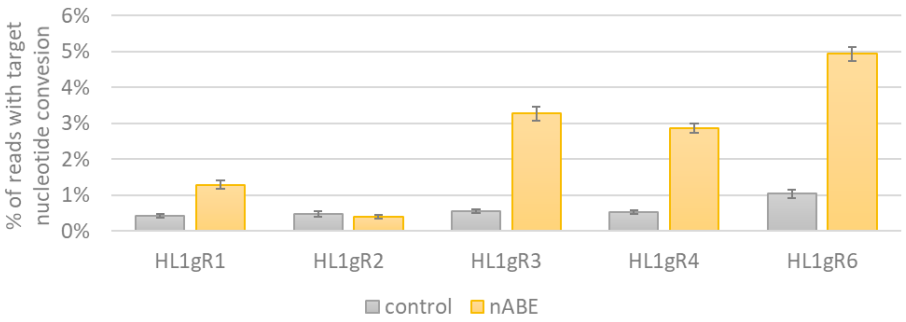


Figure 4.S3 | nBE targeting LINE-1

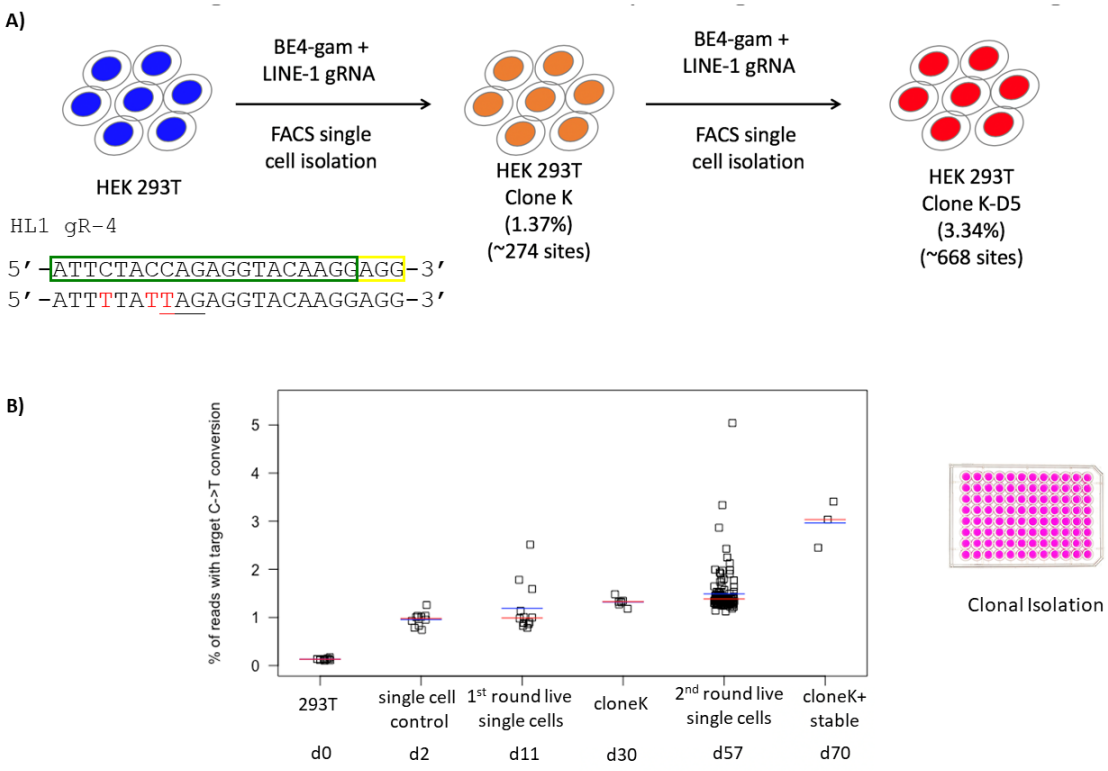


Figure 4.S4 | Utilizing high copy repetitive elements for the testing nBEs

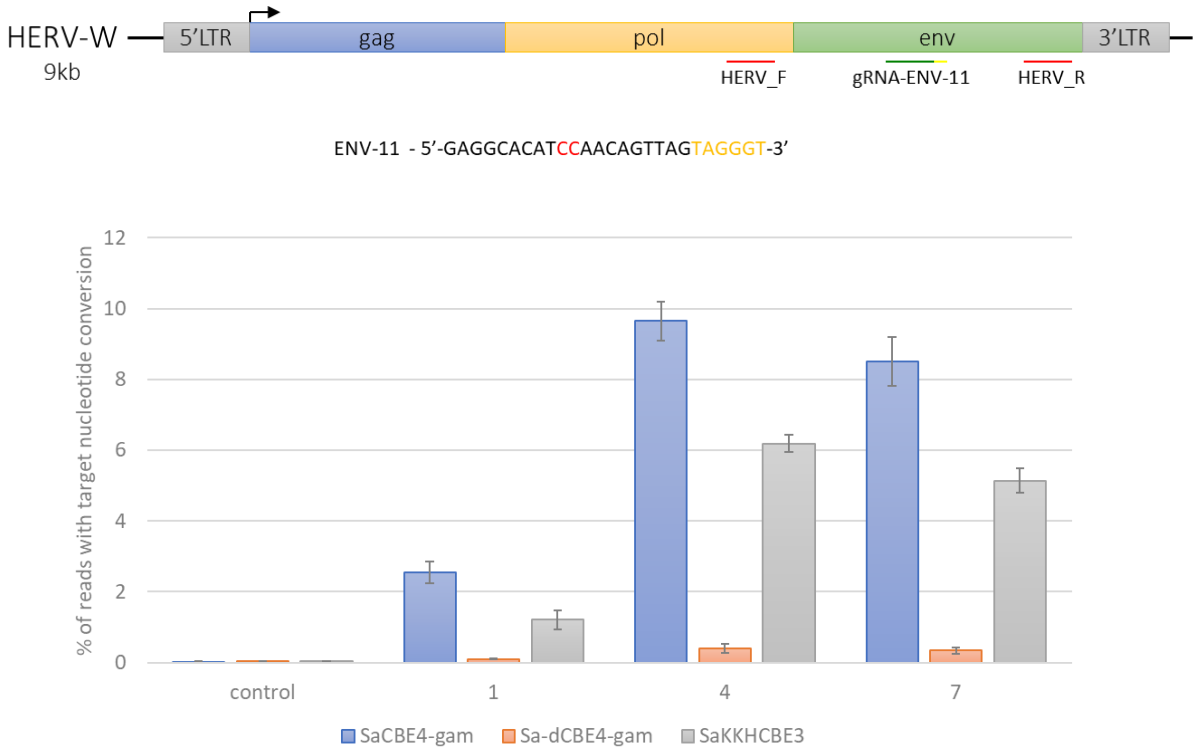


Figure 4.S5 | Targeting HERV-W using nBEs and dBEs

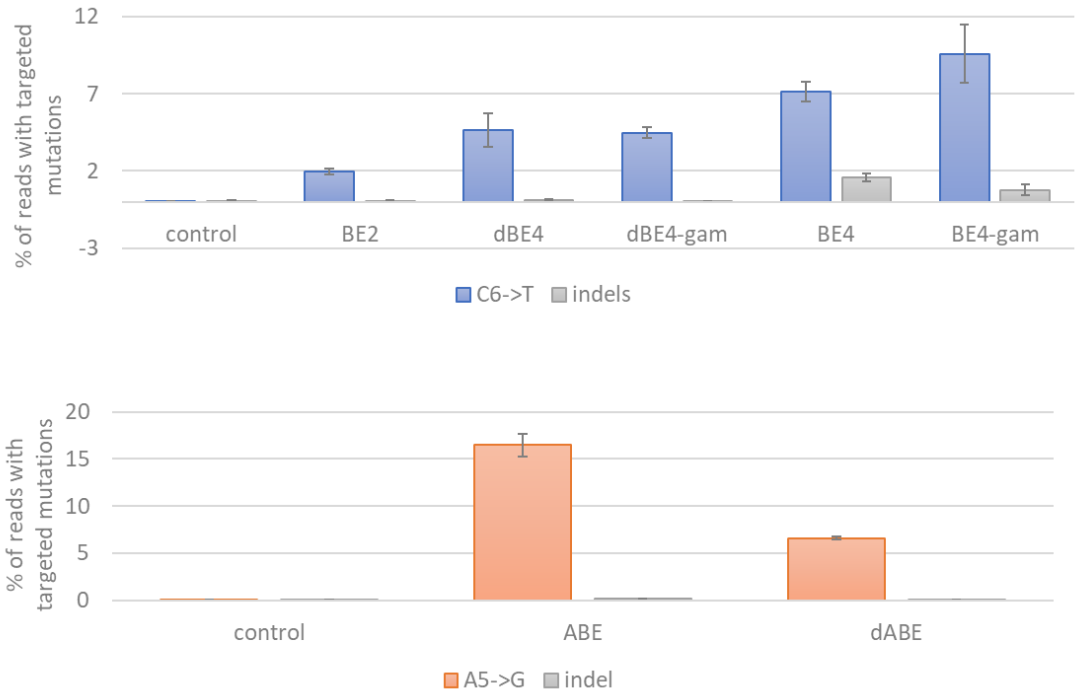


Figure 4.S6 | dBE vs nBE at a single locus target

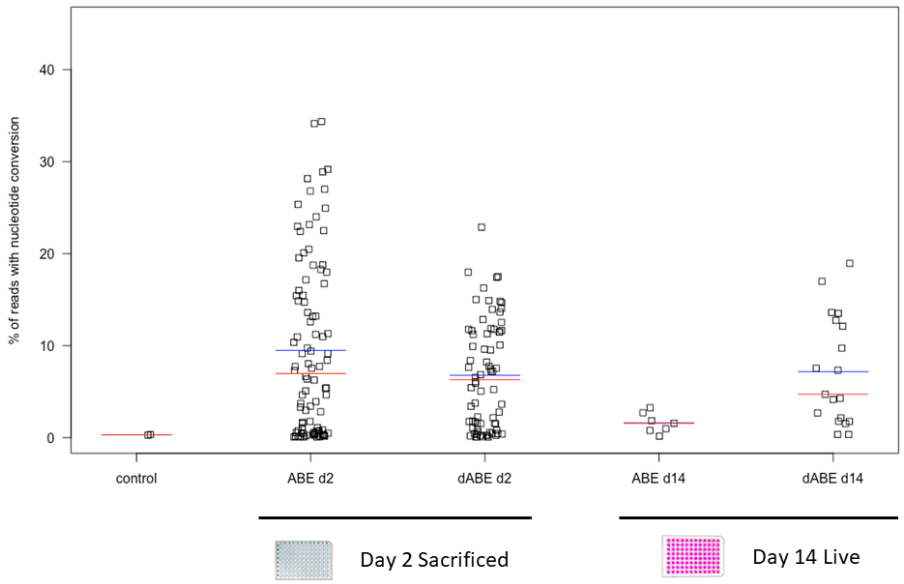


Figure 4.S7 | dABE targeting LINE-1 single cell analysis with HL1gR46

control

	A_1	T_2	T_3	C_4	T_5	A_6	C_7	C_8	A_9	G_{10}	A_{11}	G_{12}	G_{13}	T_{14}	A_{15}	C_{16}	A_{17}	A_{18}	G_{19}	G_{20}	PAM A	PAM G	Indel % G
G	0.3	0.1	0.2	0.2	0.1	0.6	0.3	0.4	0.3	98.1	0.3	98.3	97.7	0.1	1.0	1.0	0.7	0.6	96.9	97.7	0.4	97.7	98.6
A	99.2	0.2	0.2	0.2	0.1	98.6	0.9	0.4	99.2	0.5	99.1	0.9	1.2	0.1	98.6	1.2	98.8	97.8	2.8	1.4	99.4	1.8	0.7
T	0.3	99.7	99.3	0.6	99.0	0.2	3.5	0.4	0.3	0.2	0.3	0.4	0.7	99.1	0.2	1.9	0.2	1.5	0.3	0.4	0.1	0.4	0.4
C	0.1	0.3	0.3	98.9	0.7	0.4	95.1	98.7	0.1	1.0	0.2	0.6	0.4	0.6	0.1	96.1	0.3	0.2	0.1	0.5	0.1	0.2	0.3

nCBE4-gam

	A_1	T_2	T_3	C_4	T_5	A_6	C_7	C_8	A_9	G_{10}	A_{11}	G_{12}	G_{13}	T_{14}	A_{15}	C_{16}	A_{17}	A_{18}	G_{19}	G_{20}	PAM A	PAM G	Indel % G
G	0.3	0.0	0.1	0.5	0.1	0.6	0.4	0.5	0.3	98.1	0.3	98.4	97.9	0.1	1.1	0.7	0.6	0.5	96.8	98.2	0.3	97.9	98.2
A	99.3	0.1	0.1	0.2	0.1	98.7	0.9	0.4	99.2	0.6	99.3	0.8	1.2	0.1	98.5	1.2	99.0	97.9	2.9	1.1	99.4	1.7	0.9
T	0.2	99.7	99.5	7.4	99.3	0.2	9.9	6.9	0.3	0.1	0.3	0.3	0.6	99.2	0.3	2.1	0.1	1.4	0.2	0.2	0.2	0.3	0.5
C	0.2	0.2	0.2	91.8	0.4	0.4	88.7	92.2	0.1	1.1	0.1	0.3	0.2	0.6	0.1	96.0	0.2	0.2	0.1	0.5	0.1	0.2	0.4

dCBE4-gam

	A_1	T_2	T_3	C_4	T_5	A_6	C_7	C_8	A_9	G_{10}	A_{11}	G_{12}	G_{13}	T_{14}	A_{15}	C_{16}	A_{17}	A_{18}	G_{19}	G_{20}	PAM A	PAM G	Indel % G
G	0.3	0.0	0.0	1.1	0.0	0.3	0.6	0.2	0.2	97.7	0.1	98.6	98.3	0.0	0.7	1.4	0.6	0.6	96.3	97.8	0.4	97.6	98.6
A	99.5	0.0	0.1	0.2	0.1	98.5	1.3	0.4	99.1	0.7	99.4	0.6	1.1	0.1	98.7	1.3	99.3	97.3	3.3	1.7	99.4	1.9	0.9
T	0.1	99.7	99.5	23.9	99.0	0.2	25.8	24.2	0.3	0.3	0.1	0.3	0.4	99.2	0.3	2.2	0.2	1.8	0.3	0.1	0.2	0.5	0.2
C	0.0	0.2	0.1	74.3	0.5	0.4	71.8	74.9	0.0	0.9	0.1	0.3	0.2	0.4	0.1	96.4	0.1	0.3	0.0	0.3	0.0	0.1	0.3

nABE

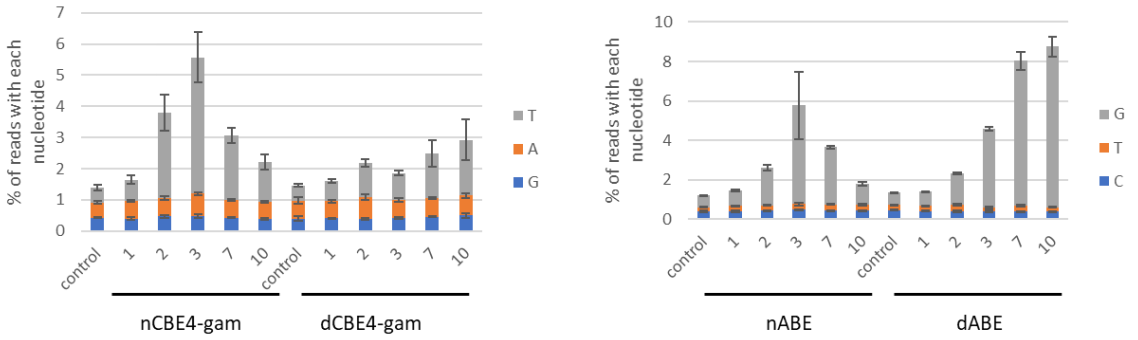
	A_1	T_2	T_3	C_4	T_5	A_6	C_7	C_8	A_9	G_{10}	A_{11}	G_{12}	G_{13}	T_{14}	A_{15}	C_{16}	A_{17}	A_{18}	G_{19}	G_{20}	PAM A	PAM G	Indel % G
G	0.3	0.1	0.1	0.2	0.3	1.3	0.3	0.4	0.2	98.4	0.3	98.4	97.6	0.1	1.0	0.9	0.8	0.6	96.4	97.9	0.4	97.4	98.5
A	99.2	0.1	0.2	0.2	0.1	97.9	0.8	0.5	99.3	0.4	99.3	0.9	1.3	0.1	98.6	1.4	98.7	97.4	3.2	1.4	99.2	1.9	0.9
T	0.3	99.7	99.4	0.5	98.8	0.2	3.6	0.4	0.3	0.2	0.2	0.6	0.6	99.0	0.2	1.9	0.3	1.7	0.3	0.2	0.2	0.5	0.3
C	0.1	0.2	0.2	99.0	0.8	0.5	95.1	98.6	0.1	0.9	0.1	0.5	0.4	0.7	0.1	96.2	0.3	0.3	0.0	0.5	0.0	0.2	0.3

dABE

	A_1	T_2	T_3	C_4	T_5	A_6	C_7	C_8	A_9	G_{10}	A_{11}	G_{12}	G_{13}	T_{14}	A_{15}	C_{16}	A_{17}	A_{18}	G_{19}	G_{20}	PAM A	PAM G	Indel % G
G	0.4	0.1	0.1	0.2	0.1	49.6	0.4	0.4	2.3	98.4	0.3	98.6	97.9	0.1	0.8	0.7	0.7	0.5	96.9	97.9	0.3	97.8	98.7
A	99.3	0.1	0.2	0.1	0.1	49.7	0.7	0.4	97.3	0.5	99.2	0.6	1.2	0.0	98.7	1.1	98.9	97.9	2.8	1.2	99.5	1.7	0.7
T	0.2	99.7	99.4	0.5	99.2	0.2	3.3	0.4	0.3	0.1	0.3	0.6	0.6	99.3	0.4	1.9	0.1	1.5	0.3	0.4	0.2	0.3	0.3
C	0.1	0.2	0.3	99.0	0.6	0.5	95.5	98.8	0.1	0.9	0.2	0.5	0.4	0.5	0.1	96.4	0.2	0.1	0.0	0.5	0.0	0.2	0.3

Figure 4.S8 | Deamination frequencies for highest edited clones per editor at each position of the gRNA

A)



B)

HL1gR4 population control

	A_1	T_2	T_3	C_4	T_5	A_6	C_7	C_8	A_9	G_{10}	A_{11}	G_{12}	G_{13}	T_{14}	A_{15}	C_{16}	A_{17}	A_{18}	G_{19}	G_{20}	PAM A	PAM G	PAM G	Indel %
G	0.2	0.1	0.1	0.3	0.1	0.6	0.4	0.3	0.3	98.3	0.2	98.2	97.8	0.0	1.1	1.0	0.8	0.4	95.9	97.7	0.4	97.6	98.7	1.22
A	99.5	0.1	0.3	0.1	0.1	98.7	0.9	0.7	99.3	0.5	99.2	1.1	1.2	0.0	98.7	1.4	98.8	98.1	3.8	1.6	99.4	1.9	0.9	
T	0.2	99.7	99.4	0.4	99.3	0.2	3.6	0.5	0.3	0.2	0.3	0.3	0.6	99.2	0.1	1.8	0.2	1.2	0.2	0.3	0.1	0.5	0.3	
C	0.0	0.2	0.2	99.2	0.5	0.4	95.0	98.5	0.1	1.0	0.2	0.5	0.5	0.7	0.1	95.9	0.2	0.3	0.1	0.4	0.1	0.2	0.2	

nCBE4-gam

	A_1	T_2	T_3	C_4	T_5	A_6	C_7	C_8	A_9	G_{10}	A_{11}	G_{12}	G_{13}	T_{14}	A_{15}	C_{16}	A_{17}	A_{18}	G_{19}	G_{20}	PAM A	PAM G	PAM G	Indel %
G	0.2	0.0	0.2	0.4	0.1	0.5	0.4	0.4	0.4	98.1	0.3	98.4	97.6	0.2	0.9	1.0	0.7	0.5	95.1	98.0	0.4	97.5	98.4	1.41
A	99.4	0.2	0.2	0.4	0.1	98.9	0.9	0.6	99.2	0.6	99.1	0.9	1.3	0.1	98.8	1.2	98.8	98.1	4.4	1.4	99.5	2.1	1.0	
T	0.3	99.6	99.3	2.9	99.1	0.1	5.5	2.5	0.3	0.2	0.3	0.3	0.6	99.2	0.2	1.6	0.1	1.1	0.3	0.3	0.1	0.3	0.2	
C	0.1	0.2	0.3	96.2	0.7	0.4	93.0	96.5	0.1	1.0	0.2	0.3	0.5	0.5	0.1	96.3	0.3	0.3	0.1	0.4	0.1	0.2	0.4	

dCBE4-gam

	A_1	T_2	T_3	C_4	T_5	A_6	C_7	C_8	A_9	G_{10}	A_{11}	G_{12}	G_{13}	T_{14}	A_{15}	C_{16}	A_{17}	A_{18}	G_{19}	G_{20}	PAM A	PAM G	PAM G	Indel %
G	0.3	0.1	0.2	0.4	0.1	0.7	0.5	0.4	0.3	97.8	0.3	96.8	96.7	0.2	1.1	1.0	0.9	0.5	91.5	97.0	0.5	95.9	98.0	1.43
A	99.2	0.1	0.3	0.3	0.2	98.6	0.9	0.6	99.3	0.8	99.1	1.2	1.8	0.1	98.4	1.5	98.6	98.0	8.2	1.8	99.3	3.5	1.4	
T	0.3	99.5	99.2	1.7	99.0	0.2	4.6	1.5	0.3	0.2	0.4	1.4	1.0	98.9	0.4	2.3	0.2	1.2	0.3	0.4	0.2	0.4	0.3	
C	0.1	0.4	0.3	97.6	0.7	0.5	93.9	97.4	0.1	1.1	0.1	0.6	0.5	0.8	0.1	95.2	0.3	0.2	0.1	0.9	0.1	0.2	0.3	

nABE

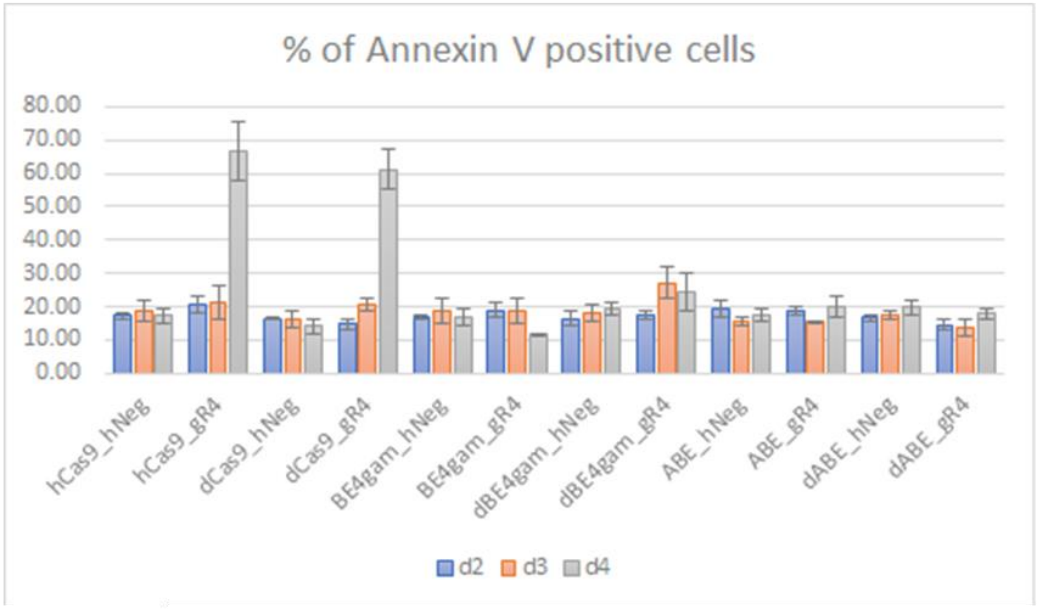
	A_1	T_2	T_3	C_4	T_5	A_6	C_7	C_8	A_9	G_{10}	A_{11}	G_{12}	G_{13}	T_{14}	A_{15}	C_{16}	A_{17}	A_{18}	G_{19}	G_{20}	PAM A	PAM G	PAM G	Indel %
G	0.5	0.1	0.2	0.5	0.1	3.0	0.4	0.4	0.5	97.9	0.3	97.3	96.8	0.1	1.0	1.2	1.0	0.6	90.1	96.8	0.6	96.7	97.9	1.49
A	99.1	0.3	0.2	0.4	0.2	96.2	0.9	0.6	98.9	0.9	98.9	1.4	2.0	0.2	98.5	1.5	98.6	98.0	9.5	2.1	99.1	2.7	1.5	
T	0.3	99.4	99.3	0.7	98.7	0.3	3.4	0.7	0.4	0.2	0.4	0.7	0.8	98.5	0.4	2.2	0.3	1.1	0.4	0.4	0.2	0.4	0.3	
C	0.1	0.4	0.3	98.3	0.9	0.5	95.1	98.3	0.1	1.1	0.2	0.5	0.5	1.2	0.1	95.3	0.2	0.3	0.1	0.7	0.1	0.2	0.4	

dABE

	A_1	T_2	T_3	C_4	T_5	A_6	C_7	C_8	A_9	G_{10}	A_{11}	G_{12}	G_{13}	T_{14}	A_{15}	C_{16}	A_{17}	A_{18}	G_{19}	G_{20}	PAM A	PAM G	PAM G	Indel %
G	0.4	0.1	0.2	0.4	0.1	8.2	0.5	0.5	0.9	97.7	0.3	96.8	96.3	0.1	1.0	1.2	0.9	0.5	89.0	96.8	0.4	96.2	97.7	1.41
A	99.2	0.3	0.3	0.3	0.2	91.1	0.9	0.5	98.6	0.8	99.0	1.5	2.0	0.1	98.5	1.6	98.7	98.2	10.7	1.8	99.3	3.2	1.5	
T	0.3	99.4	99.1	0.8	98.8	0.3	3.4	0.7	0.3	0.2	0.4	1.1	1.0	98.8	0.4	2.3	0.3	1.0	0.4	0.3	0.2	0.4	0.3	
C	0.1	0.4	0.3	98.4	0.9	0.4	95.1	98.3	0.1	1.2	0.2	0.6	0.6	1.0	0.1	95.0	0.2	0.3	0.1	1.0	0.1	0.2	0.4	

Figure 4.S9 | Base editing purity in HEK 293T targeting LINE-1

A)



B)

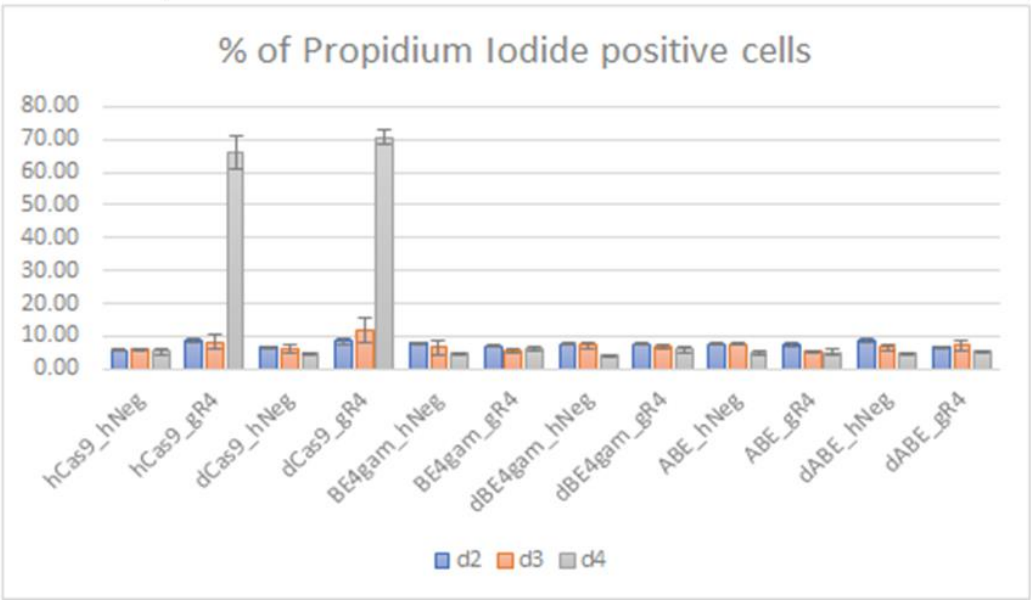


Figure 4.S10 | Annexin V and propidium iodide assays for cytotoxicity

5. CRISPR-mediated biocontainment technologies

5.1. Introduction

The discovery and wide-scale implementation of the CRISPR/Cas system[24,26,25] has dramatically expanded the toolbox for genome engineering and has revolutionized the future prospects of basic biological research, data storage in living systems[124], agricultural science[148], and medicine[149]. Beyond the initial excitement and optimism surrounding CRISPR/Cas for its ease of use and positive applications lie the potential for dual use that demands awareness and motivates the development of protections and countermeasures. Following up with such concerns, upon our previous observation that targeting transposable elements such as Alu or LINE-1 using CRISPR-Cas9 induces a high level of genotoxicity we designed CRISPR-DS that ensures that introduction or activation of Cas9 triggers cell death, rendering cell populations in which the system is active effectively non-editable by Cas9.

Based on this same mechanism of action we also designed a safety switch, a “suicide” system to make human cell therapies safer in case adverse effects based on CRISPR-DS mechanism of action. Regenerative medicine holds promises to cure a wide range of diseases by transplantation of cells, tissues or organs in order to replace or complement the malfunctioning ones. Particularly, the development of autologous T-cell therapies in the past decade has shown impressive results[150,151,152]. Even though cell therapies have shown great promises through their efficacy against a variety of diseases, several risks have been identified. The transplantation of hiPSCs for therapeutic purposes that have been reprogrammed genetically and/or epigenetically to become pluripotent can potentially become a vector of oncogenic development[153]. Concerning T-cell based therapies, there are risks of on- and off- target effects, where the tumor antigens recognized by the modified T-cells are also present in other non-malignant tissues[150,154] or when the T-cell receptors are less specific than intended and recognize a different antigen [155]. Finally, several CD19 CAR T-cell therapies have shown to provoke the cytokine release syndrome in which the massive activation of T-cells and subsequent

"cytokine storm" were deadly to patients[156]. Therefore, in order to mitigate the potential risks associated with cell therapies, the development and implementation of safety switches programmed within the therapeutic cells are important to give the option of selectively eliminating all transplanted cells in case of undesired secondary effects.

CRISPR-DS relies on the expression of a class of sgRNAs in cells that will cause Cas9 to generate DNA double-strand breaks (DSBs) cuts of high-lethality if Cas9 is expressed. CRISPR-DS may either be activated in cells that have never been otherwise edited, rendering them safe from alteration, or after they have already gone through earlier rounds of editing, establishing "tamper-proof" information storage within a biological system. The CRISPR defense system would prevent edits to populations of cells by removing those that encounter Cas9, preventing them from passing on their genetic modification.

For the CRISPR-DS to reliably trigger cell death, cleavage of the genomic targets must not only be lethal, but because even low doses of Cas9 can affect genome edits over time, also be sensitive on a per cell basis to rapidly respond to Cas9 nuclease activity. For instance, when facing undesired edits, Cas9 variants with protein modifications may exhibit reduced interactions with standard sgRNA scaffolds. To achieve high Cas9 sensitivity, we explored the targeting of repetitive genetic elements naturally found within eukaryotic genomes. Others have reported the deleterious and deadly effects of targeting sites in moderately repetitive regions (4-62 copies per cell)[43,44,41]. While Kuscu et al. reports elevated apoptosis at ~12 copies per cell, this amount of DNA damage can be tolerated in at least some cells as demonstrated by the isolation of multiple independent clones with the knock-out of 62 Porcine Endogenous Retroviruses (PERV) elements resulting in viable cells and ultimately the birth of healthy pigs without PERV expression or transmission. Therefore, we tested sgRNAs targeting two high copy number Transposable Elements (TEs): Long Interspersed Element 1 (LINE-1) (~1 x 10^4

copies[140]) and Alu (~1 x 10^6 copies[157]). Our hypothesis was that Cas9 would trigger a massive number of DSBs at these targets, overwhelming the cell's ability to repair this damage.

As alternative approaches, we tested the CRISPR-DS with sgRNAs that target the essential genes *POLE2* (sgRNA-*POLE2*), a subunit of the DNA polymerase, and *GTFIIB* (sgRNA-*GTFIIB*), the General Transcription factor IIB. Additionally, we compared CRISPR-DS to natural anti-CRISPR proteins, AcrIIA2 and AcrIIA4, that phages have evolved in an ongoing arms race with bacteria[158,159,160] as a means to prevent undesired edits. These anti-CRISPR proteins are small peptides that act by binding the PAM recognition domain of Cas9 to prevent its interaction with DNA, as revealed by crystallography[161]. Anti-CRISPR proteins provides genome editing resistance by inhibiting Cas9 function as compared to CRISPR-DS whereby cells receiving Cas9 are cleared, resulting in an uneditable cell population. By including these proteins our experiments thus provide an initial basis for comparing the effectiveness of these two strategies.

5.2. Methods

5.2.1. Cas9, sgRNA and anti-CRISPR plasmids used for genome editing

Expression vector encoding humanized pCas9_GFP protein was obtained from Addgene.org (Plasmid #44719). The traditional route of using online sgRNA design tools does not work for repetitive elements as they are designed to avoid these sequences for concerns of deleterious off-target effects. The consensus sequence for Alu and LINE-1 were obtained from previous publications and guides were selected to target the most critical elements of these TEs (promoter for Alu and ORF-2 for LINE1). The sgRNAs used in this study were synthesized and cloned as previously described[142], briefly two 24mer oligos with sticky ends compatible for ligation were synthesized from IDT for cloning into the pSB700_mCherry plasmid (Addgene Plasmid #64046) after cutting with the BsmbI restriction enzyme.

After sequence confirmation using the humanU6 primer, plasmids were prepared using the Qiagen Plasmid Plus Midi Kit (Cat # 12943). Expression vectors encoding Anti-CRISPR proteins were obtained from Addgene: AcrIIA2 (pJH373 plasmid, ID# 86840) and AcrIIA4 (pJH376 plasmid, ID# 86840).

5.2.2. Human iPSCs cell culture

The non-integrated PGP1 hiPSC line was generated and obtained from the Church Lab and were cultured with mTeSR medium on tissue culture plates coated with Matrigel (BD Biosciences). For routine passaging, iPSCs were digested with TrypLE (cat #12604039) for 5 minutes and washed with an equal volume PBS by centrifugation at 300g for 5 minutes. Digested iPSCs were then plated onto Matrigel coated plates at a density of $3x10^4$ per cm^2 with mTeSR medium supplemented with 10 µM ROCK Inhibitor Y-27632 (Stemgent) for the first 24 hours.

5.2.3. Transfection of human iPSCs

Human iPSCs were digested with TrypLE for 5 minutes and the single cells were washed once with PBS. $0.8x10^6$ iPSCs were then re-suspended in 100µl of P3 Primary Cell Solution (Lonza Cat# V4XP-3024) supplemented with plasmid and then nucleofected using the 4D-Nucleofector (Lonza) with the hES H9 program (CB150). For PB integration 8µg of PB-gRNA was transfected with 2µg of PB supertransposase. For gene editing experiments 7.5µg of pCas9_GFP plasmid and 5µg of sgRNA plasmid were used. After electroporation, the iPSCs were then plated onto Matrigel-coated plates in mTeSR medium supplemented with 10µM Y-27632. Cells were harvested at days 1, 2, 3, 5 and 10 after transfection and the genomic DNAs were isolated using DNeasy Blood and Tissue Kit (Qiagen).

5.2.4. Synthesis and genomic integration of the CRISPR-DS into HEK 293T cells

sgRNAs targeting Alu, LINE-1, and a non-human control were amplified from the pSB700 plasmid and cloned into PB-TRE-dCas9_VPR (Addgene #63800) using the following primers: U6-NheI-F: GCAGCTAGCGAGGGCCTATTTCCCATGATT, and sgRNA-*Bam*HI-R: TGCGGATCCAACGCGGAACTCCATATATGG. dCas9_VPR was removed during the cloning process using the restriction enzymes *Nhe*I and *Bam*HI to integrate the U6-gRNA construct into the PiggyBac transposon sequences. Colonies were Sanger sequence verified and prepped using the Qiagen plasmid plus midi kit. HEK 293T cells were then lipofected with PB-gRNAs (Alu, AluYa5, LINE-1, and non-human) and PB-transposase. Cells were selected with puromycin (1µg/ml) beginning at day two until day nine. Populations of puromycin resistant cells were used for the initial Cas9 genome editing trials. Individual cells from the puro resistance population were grown after single-cell sorting into 96-well plates and isolated for further testing in Cas9 genome editing experiments.

5.2.5. Propidium Iodide and Annexin V staining and FACS analysis

Cells were dissociated with TrypLE, diluted in an equal volume of PBS then centrifuged at ~300g for 5 minutes at room temperature. We resuspended samples into 500µl PBS and half of the cells were pelleted for later gDNA analysis. The reminder was centrifuged and resuspended into 100µl of Annexin V Binding Buffer (ref #V13246) diluted into ultrapure water at a 1:5 ratio. Subsequently, we added 5µl of Alexa 647 Annexin V dye (ref #A23204) and incubated samples in the dark for 15 minutes. We then added 100µl of Annexin V Binding Buffer and added 4µl of Propidium Iodide (ref #P3566) diluted into the Annexin V Binding Buffer at a 1:10 ratio. Samples were incubated in the dark for another 15 minutes. Cells were washed with 500µl of Annexin V Binding Buffer and centrifuged again to be finally

resuspended into 400µl of Annexin V Binding Buffer. All samples were filtered using a cell strainer and were run on the LSR 11 using a 70-µm nozzle. Analysis was conducted using FlowJo software.

5.2.6. Antibody staining and fluorescent microscopy

Cells were fixed with 4% formaldehyde for 10 minutes at room temperature and blocked with PBS containing 10% normal donkey serum, 0.3 M glycine, 1% BSA and 0.1% tween for 2h at room temperature. Staining of the treated cells with Anti-γH2AX antibody (10µg/ml) was performed overnight at 4°C in PBS containing 1% BSA and 0.1% tween. The cells were washed three time (5 minute intervals) with PBS followed by secondary staining. Cells were imaged using a Zeiss AxioObserver.Z1 microscope equipped with a Plan-Apochromat 20×/0.8 objective, an EM-CCD digital camera system (Hamamatsu) and a four-channel LED light source (Colibri), and Zeiss TIRF/ LSM 710 confocal (ZeissTIRF-confocal), 63×.

5.2.7. Transfection of HEK 293T

HEK293T cells were cultured in Dulbecco's modified Eagle's medium supplemented with 10% FBS (Gibco) at 37 °C with 5% CO2 incubation. Transfection was conducted using Lipofectamine 2000 (Cat# 11668027) using the suppliers recommended protocol. 1µg SpCas9_GFP, 1µg sgRNA-JAK2, and 1µg sgRNA-test was used per 80k cells in a 12-well plate. Cell pellets were collected three days after transfection for genomic DNA extraction and sequencing analysis.

5.2.8. Preparation of HEK 293T samples for Insertions and Deletions (indels) analysis

Following the genomic DNA extraction of HEK293T samples using DNeasy Blood & tissue Kit from Qiagen (Cat# 69506) according to the supplier's protocol, we amplified 586 bp of the JAK2 locus using the following primers: P2F-JAK2: CGTTGATGGCAGTTGCAGGTC. and P3R-JAK2: GTACTGAAAAGGCCAGTTATTCC. Amplicons were obtained after PCR amplification using Kapa HiFi HotStart Readymix kit from Kapa Biosystems (Cat# KK2602) according to the supplier's protocol. PCR products were then run on E-Gels EX 2% Agarose (Cat# G402002) from Invitrogen and amplicons of about 586 bp were extracted using the Qiagen QIAquick Gel Extraction Kit (Cat# 28706). Gel extracted PCR products were then submitted to Genewiz for Sanger DNA sequencing using the P3R primer.

5.2.9. Insertions and deletions (indels) analysis

Indels analysis of all samples from Fig. 5.3a was executed using TIDE web tool[162]. The experiment was performed with 3 replicates as described in "Preparation of HEK293T samples for Insertions and Deletions (indels) analysis". Sequencing trace files provided through Genewiz services were then analyzed using TIDE web tool that assesses genome editing by CRISPR/Cas9 using a decomposition algorithm that identifies and quantifies insertions and deletions in the expected editing site. The following advanced settings were used: Alignment window: left boundary = 100; Decomposition window = 268 bp to 350 bp; indels size range = 0 to 10 bp; P-value threshold = 0.001. For each sample, the Control Sample Chromatogram file uploaded comes from a HEK 293T sample transfected only with sgRNA-non-human. Data was plotted using Excel displaying the mean of three biological replicates with the error bars representing the standard error. Statistical analysis was conducted using the student's t test

5.2.10. Illumina MiSeq library preparation and sequencing

Library preparation was conducted as previously described[144]. Briefly, genomic DNA was amplified using locus specific primers attached to part of the Illumina adapter sequence. A second round of PCR included the index sequence and the full Illumina adapter. Libraries were purified using gel extraction (Qiagen #28706), quantified using the NanoDrop and pooled together for deep sequencing on the MiSeq using 150 paired end (PE) reads.

5.2.11. NGS data analysis

The activity of Cas9 was measured by the number of reads containing insertions or deletions around the sgRNA target site. FastQC was initially used to confirm sequence quality, length and diversity. CRISPR RGEN tools[163] was used to quantify indel disruption at the targeted site by submitting the fastq file, reference, and sgRNA sequence. Data was plotted using Excel displaying the mean of three biological replicates with the error bars representing the standard error. Statistical analysis was conducted using the student's t test.

5.3. Results

5.3.1. Design of the sgRNAs targeting repetitive elements

We designed sgRNAs to target human repetitive host genomic sequences to generate numerous double-strand DNA breaks and initiate apoptosis or otherwise render the cell or tissue non-viable. Such lethal sgRNAs may be present in a cell prophylactically, being entirely benign to the cell or tissue until it encounters a genome editing agent (Fig. 5.1a).

Figure 5.1 | CRISPR Defense System prevents the formation of populations harboring DNA edits.

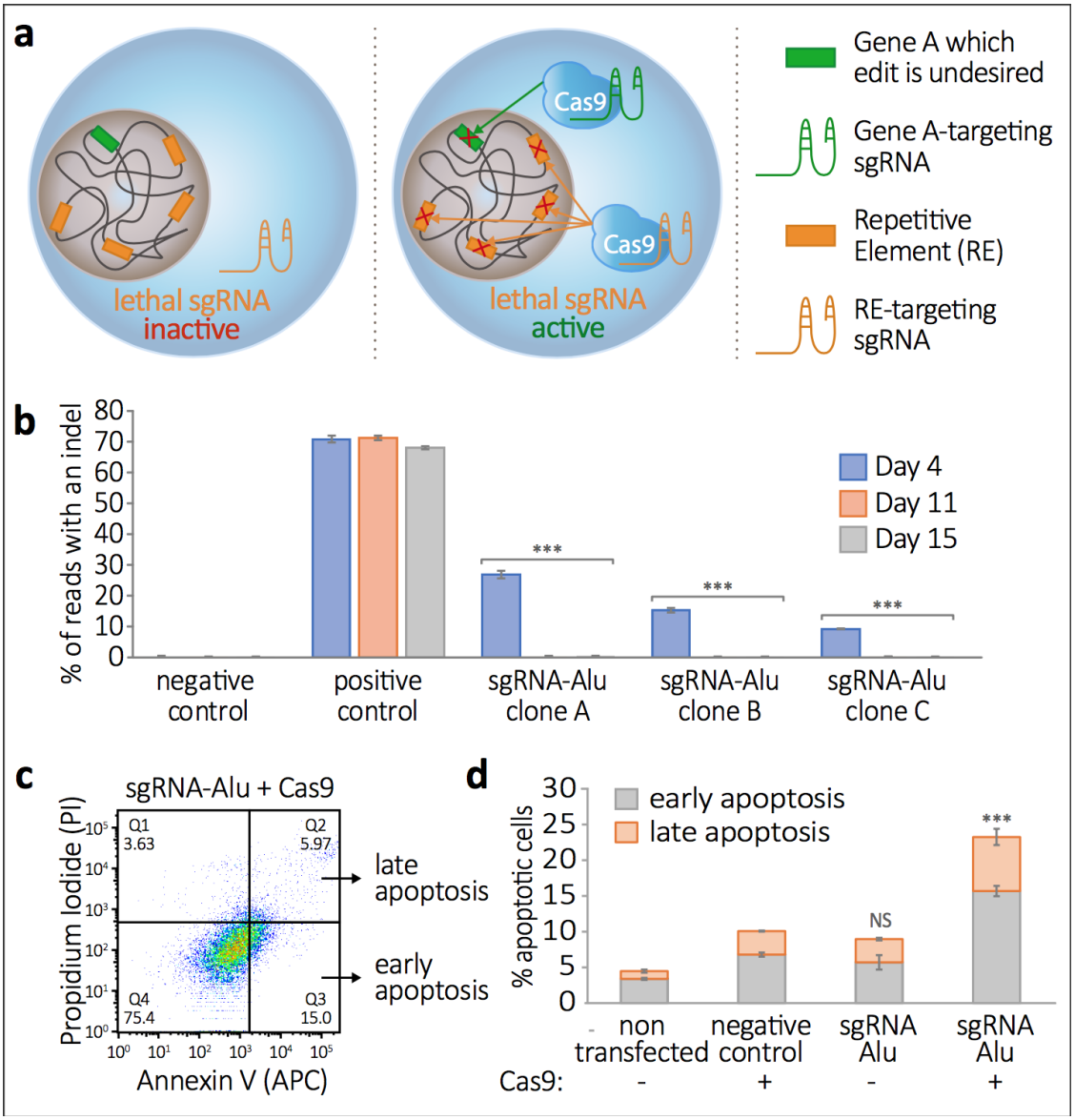


Figure 5.1 | CRISPR Defense System prevents the formation of populations harboring DNA edits. (**a**) Schematic of typical CRISPR/Cas editing in the presence of the CRISPR-DS before transfection (left) and after transfection (right). Cas9 pairs with sgRNA and disrupts its target site creating indels represented by a red X. When the CRISPR-DS is expressed in a host cell the RE sgRNA is benign in the absence of Cas9 but acts as a surveillance system to clear the cell in its presence. (**b**) Prevention of DNA modification at the JAK2 locus by CRISPR-DS. The graph represents the mean of three biological replicates for indel mutation rate at the JAK2 locus at days 4, 11, and 15 after transfection in blue, orange, and grey respectively. (**c**) Apoptosis assay FACS plot of the sample treated with sgRNA-Alu. The plot separates early apoptotic HEK 293T cells (Annexin V+ propidium iodide-) from late apoptotic HEK 293T cells (Annexin V+ propidium iodide+). (**d**) Early (gray) and late apoptosis (light orange) percentages of the samples with (+) or without (-) Cas9 three days after transfection. In all histograms, error bars represent standard error, n=3. Student's t-test was performed and marked NS, not significant ($P > 0.05$); *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$ as compared to the positive control in (b) and to the negative control in (d).

A functional CRISPR-DS should decrease genome editing at a known high efficiency locus, such as *JAK2*. To test this, we transiently transfected sgRNA-Alu and sgRNA-LINE-1 separately, along with *S. pyogenes* Cas9 (SpCas9) and a sgRNA targeting *JAK2* – our test gene – in Human Embryonic Kidney 293T (HEK 293T) cells. We confirmed with Next Generation Sequencing (NGS) that sgRNA-Alu and sgRNA-LINE-1 decreased *JAK2* insertions and deletions (indels) by about 3.6-fold two days after transfection as compared to our positive control (Fig. 5.S1a). With similar efficiencies between our two RE-targeting sgRNAs, we decided to proceed with sgRNA-Alu in designing the CRISPR-DS as it targets significantly more sequences in the genome than sgRNA-LINE-1 (a 100-fold difference).

5.3.2. CRISPR Defense System prevents the formation of populations harboring DNA edits

For our next phase of testing, we integrated constitutively expressed repetitive element-targeting sgRNAs into the genome using PiggyBac (PB) transposition and expanded our experiments to include a second cell type: PGP1 human induced Pluripotent Stem Cells (hiPSC) in addition to HEK 293T cells. While HEK 293T are highly tolerant to DSBs, PGP1 hiPSC are very sensitive to them[164]. We assayed DNA editing efficiency by transfection with SpCas9 and a sgRNA targeting our test gene *JAK2* (sgRNA-*JAK2*). Clonal cell populations stably expressing repetitive element targeting sgRNAs were isolated and exposed to this genome editing challenge, and subsequently analyzed with NGS at several different time points to quantify the presence of edits at the *JAK2* locus (Fig. 5.1a). With exposure to SpCas9, any successfully *JAK2*-modified cell will also be cut at the high-copy repetitive element, resulting in rapid and near complete cell clearance. In HEK 293T cells, while non-homologous end joining (NHEJ) levels of up to 72% (Fig. 5.1b) were observed in control samples, the three sgRNA-Alu-expressing clonal cell populations A, B and C, displayed 27.6%, 15.6% and 9.4% indels by day 4 and background

levels of 0.10%, 0.09% and 0.04% by day 15, respectively, representing a reduction of DNA editing ≥99.9% in clones expressing sgRNA-Alu as compared to control cells stably expressing a sgRNA targeting no sequences in the human genome (sgRNA-non-human) (Fig. 5.1b). Similar results were observed in PGP1 hiPSCs (Fig. 5.S2). These results show that, as expected, the CRISPR-DS system is effective at preventing cell populations from being genetically altered by clearing out cells in which SpCas9 is expressed.

5.3.3. CRISPR-Cas9 targeting high-copy number loci rapidly causes DNA damage

Evidence that cells expressing the CRISPR-DS system in these experiments are removed following Cas9 expression is supported by the decrease over time of fractions of cells exhibiting *JAK2* mutations; however, these data do not identify the mechanism of cell death. We hypothesized that CRISPR-DS bearing cells undergo apoptosis triggered by the massive DNA damage caused by expression of SpCas9 and tested this by undertaking standard cell death and apoptosis assays followed by flow cytometry analysis, and immunostaining cell samples for γH2AX, a known marker of DSBs. In HEK 293T cells stably expressing sgRNA-Alu, expression of SpCas9 significantly increased the percentage of early apoptotic and late apoptotic cell populations, as measured by Annexin V and propidium iodide, exceeding by about 2.3-fold the apoptosis triggered in HEK 293T control cells stably expressing a sgRNA- non-human (Fig. 5.1c). On the contrary, when HEK 293T cell lines expressing sgRNA-Alu did not receive SpCas9, they displayed similar cell death levels to cells expressing sgRNA-non-human (Fig. 5.1d) showing that sgRNA-Alu did not display any abnormal toxicity on its own. With respect to the γH2AX staining associated with DSB induced DNA damage, we observed a clear increase in γH2AX foci along with the abnormal formation of fused cells in the sgRNA-Alu expressing cells that was not observed in the non-human targeting control (Fig. 5.2a, 5.2b, 5.2c, Fig. 5.S1b, and 5.S1c). These results support the

hypothesis that CRISPR-DS induces apoptotic death from massive simultaneous double-stranded DNA cleavage, while remaining non-toxic in the absence of a foreign Cas9-based DNA editor.

Figure 5.2: CRISPR-Cas9 targeting high-copy number loci rapidly causes DNA damage.

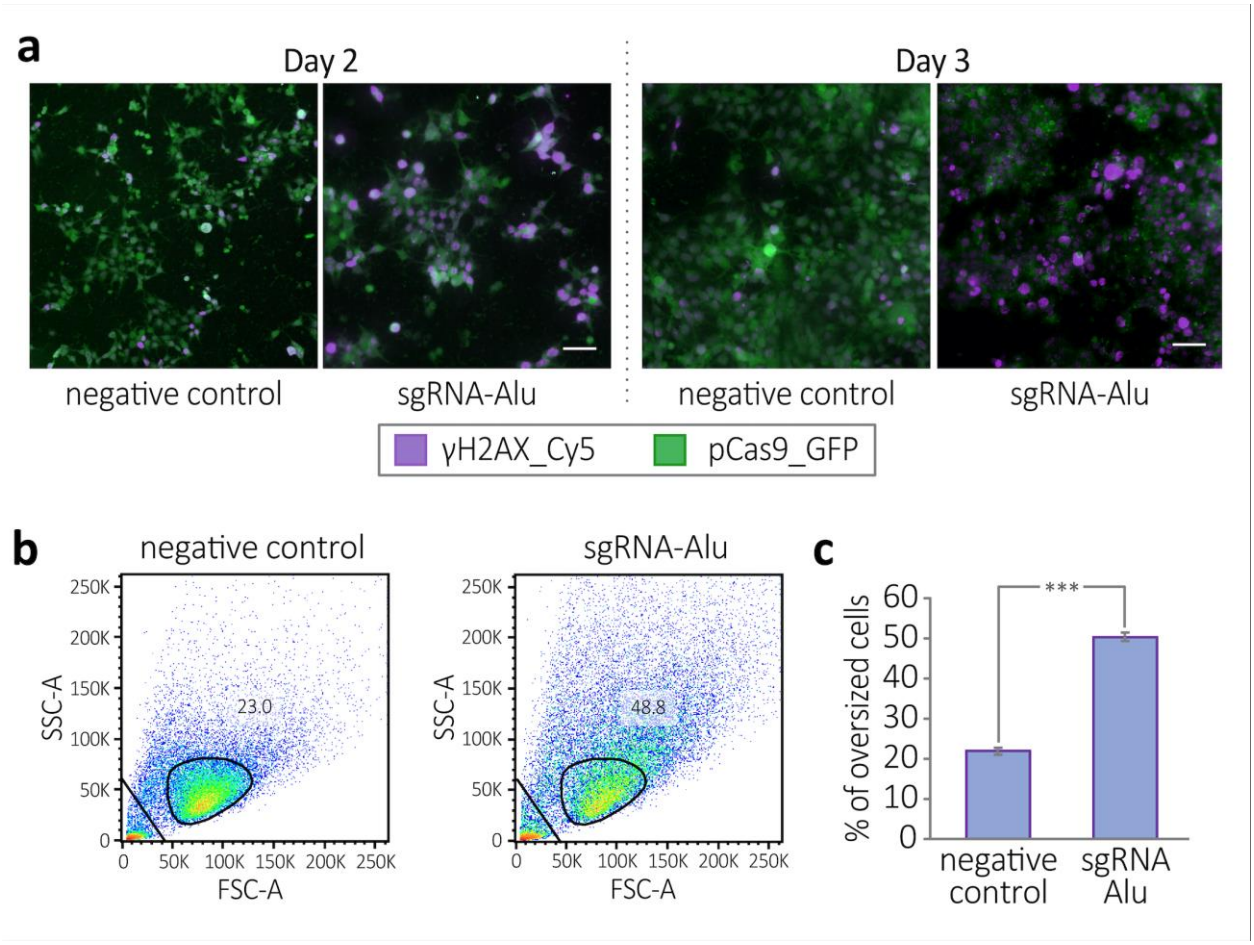


Figure 5.2 | CRISPR-Cas9 targeting high-copy number loci rapidly causes DNA damage. (**a**) 20x magnification microscope images of γH2AX immunostained cells that were transfected with SpCas9 and either sgRNA-Alu or sgRNA-non-human (negative control) 2 or 3 days after transfection. Transfected cells appear green due to the GFP marked SpCas9 (SpCas9_GFP) and γH2AX foci appear purple as antibodies are stained with a Cy5 fluorophore (γH2AX _Cy5). Scale bar = 50 µm. (**b**) FACS plot showing "oversized" cells using their forward scatter (FSC) on the x-axis and their side scatter (SSC) on the y-axis. Cells were analyzed three days after being transfected with sgRNA-Alu or sgRNA-non-human. Cells are considered "oversized" when they fall outside the normal range shown by the circular gate. Debris are excluded from the analysis (triangular gate). (**c**) Percentage of oversized cells three days after being transfected with sgRNA-Alu or sgRNA-non-human as measured by FACS. Error bars represent standard error, n=3. Student's t-test was performed and marked NS, not significant ($P > 0.05$); *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$ throughout the figure.

5.3.4. CRISPR-DS compared to systems targeting essential genes or using anti-CRISPR proteins

We next sought to compare the efficiency of CRISPR-DS in preventing DNA edits against other approaches described above: such as using sgRNAs targeting known essential genes (e.g. DNA polymerase subunits) and using anti-CRISPR proteins (acrIIA2 and acrIIA4) to inhibit Cas9 activity in human cells. To test these systems, we transiently transfected HEK 293T cells with SpCas9 and *JAK2*-targeting sgRNA, along with either the repetitive element targeting guides sgRNA-Alu (Fig. 5.3a) and sgRNA-LINE-1 (Fig. 5.S1a); the essential genes *POLE2* (Fig. 3a) and *GTFIIB* (Fig. 5.S2a); the anti-CRISPR proteins AcrIIA4 (Fig. 5.S3a) and AcrIIA2 (Fig. 5.S2a). Cells transfected with SpCas9 and our sgRNA-non-human alone constituted our negative control and when transfected in addition with sgRNA-*JAK2*, our positive control. The samples transfected with sgRNA-Alu showed a significant drop in the percentage of indels at *JAK2,* from 34.4% in the positive control to a background level of 0.9% by day nine after transfection (Fig. 5.3a). The anti-CRISPR protein AcrIIA4 was also able to decrease *JAK2* edits down to 2.6%, which was above background levels and did not display any toxicity as compared to the negative control when assayed for cell death (Fig. 5.3b and 5.3c). On the contrary, AcrIIA2 didn't have any effect in our hands and couldn't prevent DNA edits when compared to the positive control (Fig. 5.S2a). The sgRNA-*POLE2* decreased genome editing down to 8.6% and did not show increased cell death as compared to sgRNA-non-human three days after transfection (Fig. 5.3b and 5.3c). Overall, the CRISPR-DS using repetitive element targeting showed the highest efficiency and resulted in frequencies of edits indistinguishable from background levels. Even though essential-gene-targeting sgRNAs and anti-CRISPR protein AcrIIA4 did not lower the frequency of edits in human cells down to background levels of detection, such strategies in combination with repetitive elements targeting sgRNAs could be used to generate a multi-layered security system to safeguard the genome from DNA edits.

Figure 5.3 | CRISPR-DS compared to systems targeting essential genes or using anti-CRISPR proteins.

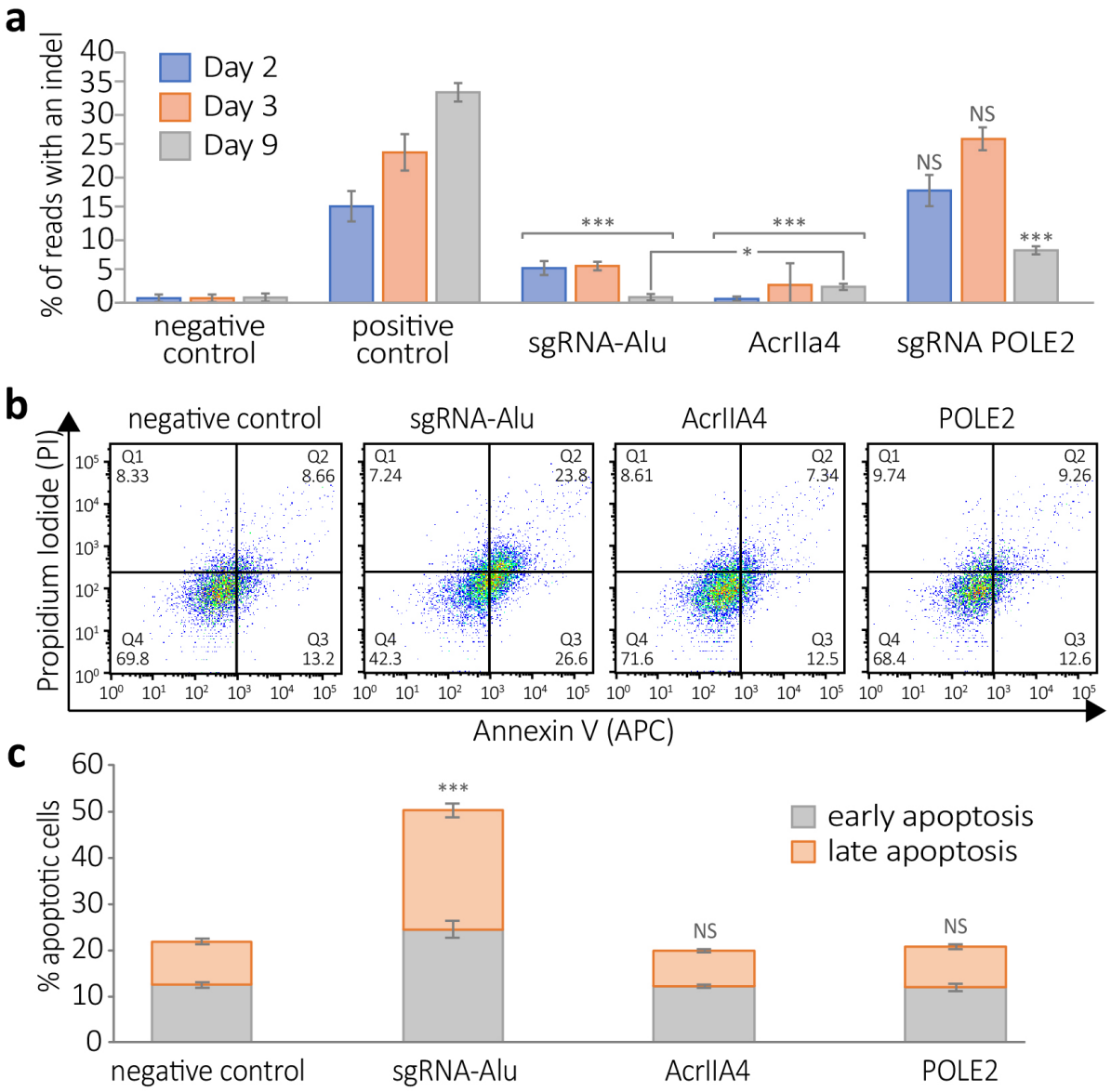


Figure 5.3 | CRISPR-DS compared to systems targeting essential genes or using anti-CRISPR proteins. (**a**) Prevention of DNA edits in HEK 293T cells at the JAK2 locus by sgRNA-Alu, anti-CRISPR protein AcrIIa4, and sgRNA targeting essential gene POLE2. Percentage of reads with an indel is plotted on the y-axis and represents mean of three biological replicates. (**b**) Three days after transfection, apoptosis was measured by FACS using Annexin V and propidium iodide staining in HEK 293T cells transiently expressing Cas9 and sgRNA-Alu, AcrIIa4, or sgRNA-POLE2. FACS data displaying Annexin V on the x-axis and propidium iodide on the y-axis in HEK 293T cells expressing sgRNA-non-human, sgRNA-Alu, anti-CRISPR protein AcrIIa4, and sgRNA-POLE2. (**c**) Percentage of apoptotic cells is plotted on the y-axis with early apoptosis in gray as measured by Annexin V+ propidium iodide-, and late apoptosis in light orange as measured by Annexin V+ Propidium iodide+ FACS populations. In all histograms, error bars represent standard error, n=3. Student t tests were performed and marked NS, not

significant (P > 0.05); *P < 0.05; **P < 0.01; ***P < 0.001 as compared to the positive control in (a) and to the negative control in (c).

5.3.5. Towards the development of a safety switch for cell therapies based on the Cas9-targeting of repetitive elements

Generation of the Cas9-based safety switch

Leveraging the efficiency of the CRISPR-DS mechanism to eliminate human cells, we sought to build a conditional circuit to activate the cell killing mechanism on demand. To do so, using the PiggyBac (PB) transposition system, we permanently integrated a doxycycline (DOX) inducible Cas9 endonuclease plasmid containing a hygromycin resistance cassette into the HEK 293T clonal cell line stably expressing the gRNA-Alu that showed the best cell-clearance efficiency (clone A). We expect the addition of DOX into the cell culture media of the resulting cell line to trigger the Cas9 expression, enabling the targeting of the Alu elements of the genome and therefore activating the elimination of the cells containing this safety switch.

Following the described PB integration and the subsequent hygromycin selection, we obtained a cell population stably expressing both the inducible Cas9 and gRNA-Alu. Single cells of this heterogeneous population were then sorted using flow cytometry which resulted in the growth of 24 clones. Each clonal population was duplicated and treated either with or without DOX for 10 days. We selected, expanded and further analyzed the 2 clones (A' and B') that displayed the most cell elimination under the microscope (Fig. 5.4).

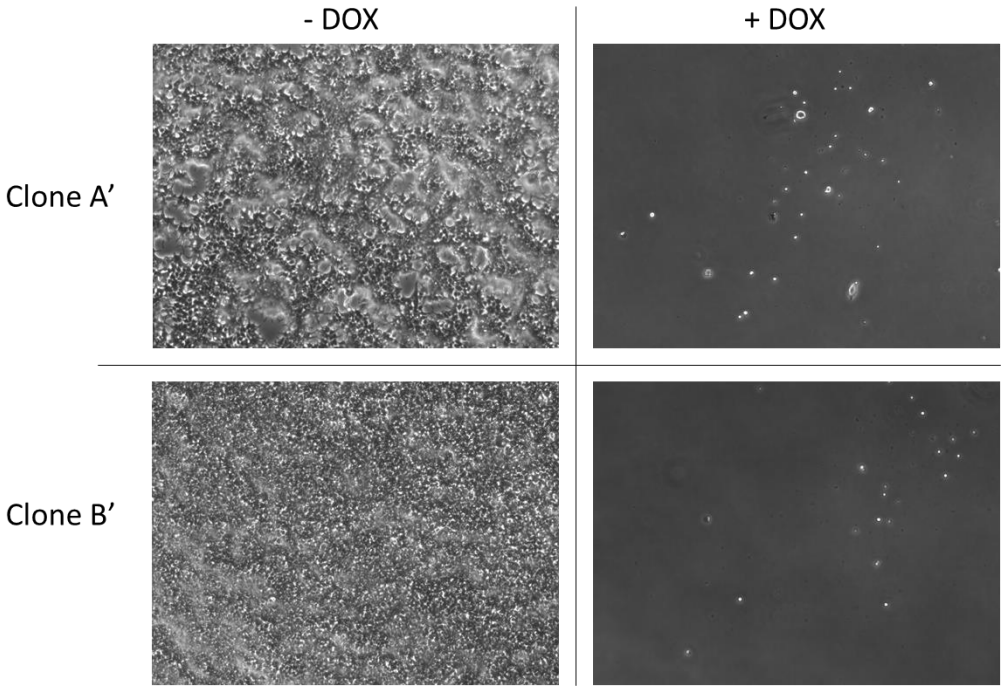


Figure 5.4 | Biosafety switch engineered HEK 293T treated with or without DOX.

10X microscope images.

Test of the safety and efficiency of the CRISPR-DS based switch

Under the microscope, clones A' and B' appeared to show almost complete clearance after 10 days of treatment with DOX. We next sought to make a quantitative analysis of the safety switch efficiency by counting the number of viable cells by flow cytometry with or without DOX treatment. To do so, we treated the clone B' with 0, 1X or 5X of DOX and counted the number of live cells (propidium iodide positive cells) after 3, 6 or 9 days of treatments. After 3 days of treatment with 5X of DOX, the clonal population displayed a 99.7% reduction of cells as compared to the same population treated without DOX. The observed cell clearance went up to 99.98% after 9 days of treatment (Fig. 5.5).

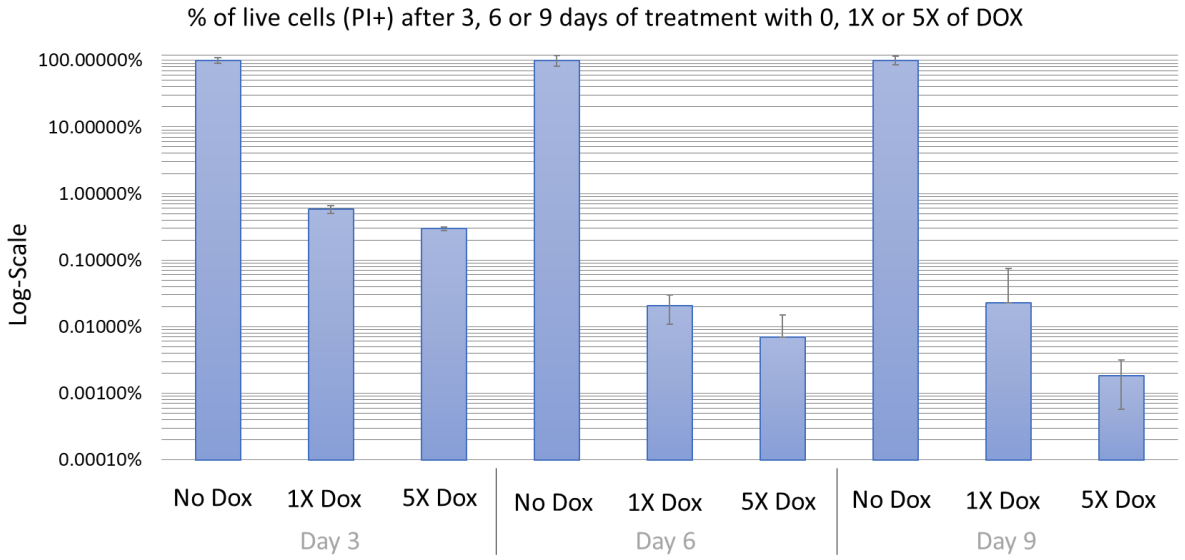


Figure 5.5 | Cell elimination sensitivity of the biosafety switch

We have shown that the activation of our inducible circuit targeting repetitive elements is effective and results in the almost complete clearance of the cells. However, in order to consider implementing a safety switch in a clinical setting, the system, on top of being efficient would ideally have 1) little spontaneous action so that the transplanted cells stay viable and keep their therapeutic benefits; and 2) the activating molecule should be inert and non-toxic. Therefore, we next investigated and quantified the self-activation or "leakiness" of our system when no DOX has been added and we assessed the cell toxicity triggered by DOX. We cultured and treated the B' clonal population as well as control non-modified HEK 293T with 0, 1X or 5X of DOX for 3 days and performed an Annexin V – Propidium Iodide assay to quantify cell death in the different conditions (Fig. 5.6). The control HEK 293T cells displayed basal apoptosis levels when treated with 1X or 5X of DOX, suggesting that both concentration of the activating molecule are not toxic to the cells. Similarly, our safety switch engineered clone B' treated with DOX did not display any abnormal apoptosis as compared to the control, when 1X or 5X or DOX triggered respectively and 6- and 7-fold increase in the apoptosis level as compared to the control cells.

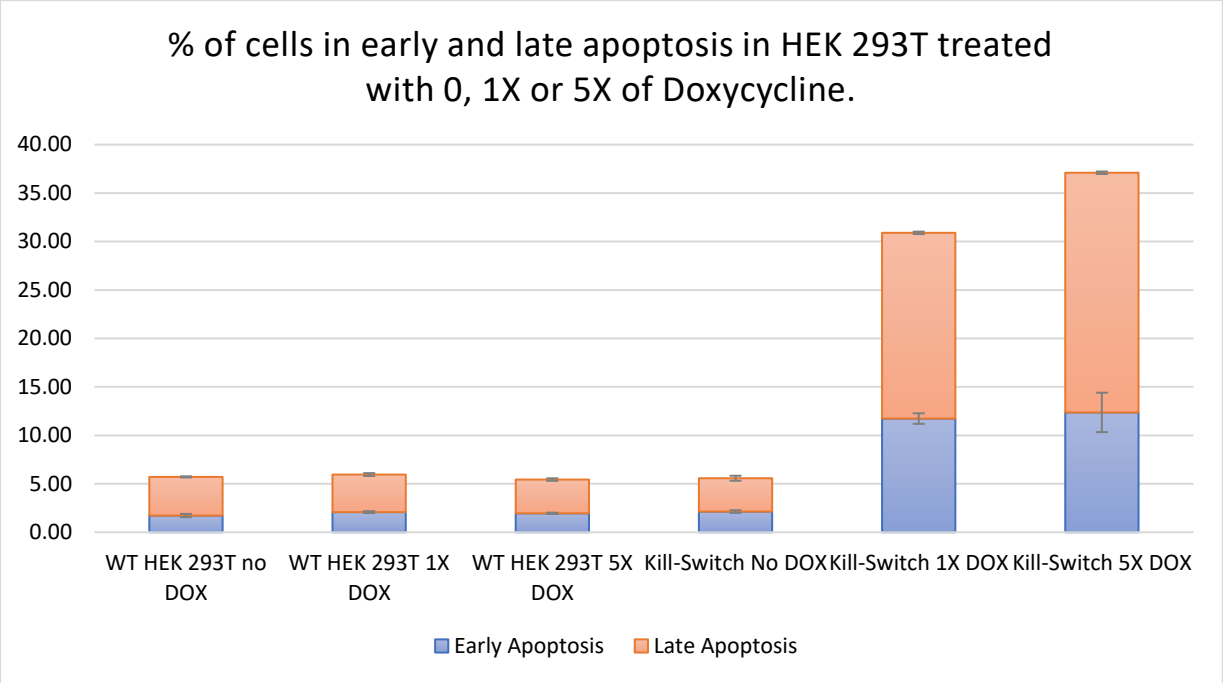


Figure 5.6 | Evaluation of doxycycline toxicity and spontaneous activation of the biosafety switch

Together these results suggest that 1) the safety switch is effective at eliminating the engineered cells to up to 99.98%, 2) DOX is not toxic to the cells and 3) the system does not activate itself spontaneously in the absence of the activating molecule but only when it is added to the medium of culture.

5.4. Discussion

Here we present a molecular CRISPR-DS strategy using lethal sgRNAs that induce rapid and robust cell death upon encountering a CRISPR/Cas9-based genome editor, but will be otherwise inert, as a prophylactic defense system against undesired genetic modification. Protein based anti-CRISPR approaches dramatically reduce Cas9 activity although they are still leaky while CRISPR-DS provides more stringent and persistent protection in such cases where prevention of undesired DNA editing is paramount. Like “death by a thousand cuts,” these high copy number repetitive element targeting sgRNAs act as a rapid sensor of Cas9, inducing cell death before any unintended edits may be passed

on to viable daughter cells or progeny. Additionally, the cleavage of genomic repeats offers enhanced sensitivity to low-level genome editing enzymatic activity due to increased target copy number and is also robust to sporadic natural host genomic mutations that might evade cell death triggered by cleavage of a single-copy essential gene. We need to anticipate and discuss well in advance, potential development of systems based on CRISPR-DS: 1) to create a population of CRISPR resistant organisms aimed at preventing unwanted modifications of industrial cell lines, plants or tissues protected with intellectual property, in cases of potential Dual Use Research of Concern (DURC); or 2) to generate a "tamper resistant" system for protecting encoded DNA in living cells used for data storage[124,165].

Here we note that CRISPR-DS can be activated in cells in which Cas9 has previously been used to make DNA modification, so long as Cas9 has been eliminated from those cells. As the use of Cas9 technologies for gene therapy is becoming more common, many therapeutic applications involve the use of *ex vivo* delivery of Cas9 to disrupt a target allele[166] or precise correction by homologous recombination[167]. CRISPR-DS may be applied after an initial round of modification to negatively select cells that still contain gene editing reagents and thus had the opportunity to generate undesired off-target mutations and secondary effects downstream.

In clinical settings, we have shown that the CRISPR-DS mechanism could potentially be leveraged as a highly escape-resistant biocontainment switch that would be activated if the host experiences complications such as host vs graft disease, cytokine storm, cancer or other unanticipated reaction from the modified cells. This bio-safety switch has shown to be highly effective at eliminating the cells in which the system is activated with undetectable spontaneous activity and activated by a non-toxic molecule. However, since the system includes Cas9, an exogenous protein coming from *Streptococcus pyogenes*, and even though its expression is repressed when not activated, we cannot exclude the potential immunogenicity of the biosafety switch in human. Therefore, this concern should be thoroughly addressed before considering testing the technology in a clinical setting.

Finally, the escape-resistant feature combined with a highly discriminatory method of delivery or activation in cancer cells may overcome common cancer treatment limitations such as the development of resistance in some cells that ultimately leads to recurrence.

To further enhance the utility of CRISPR-DS as a gene editing countermeasure, the design of the lethal sgRNAs would ideally account for a wide range of potential genome editing effectors. In the case of known CRISPR/Cas9-based systems, sgRNA scaffolds specific to these systems are all that is required to protect against each category of enzyme. To adapt our genome editing prevention system to additional orthologues beyond SpCas9, new sgRNA targets with compatible PAMs could be designed for *S. aureus* Cas9, Cpf1 or future Cas variants rapidly in response to their release (**Table 5.S1**). Utilizing evolutionarily conserved repetitive elements, a broad set of species may be covered by a relatively small number of sgRNA targets, while multiple orthologs of CRISPR/Cas9 may be included in the CRISPR Defense System to keep pace with the continuously expanding toolbox for genome editing. CRISPR-DS and other systems with the capacity to temporally, spatially, and conditionally control CRISPR/Cas activity in mammalian cells could play a central role in the safe and responsible implementation of genome editing technologies as they proceed towards an unforeseeable singularity.

5.5. Supplementary figures

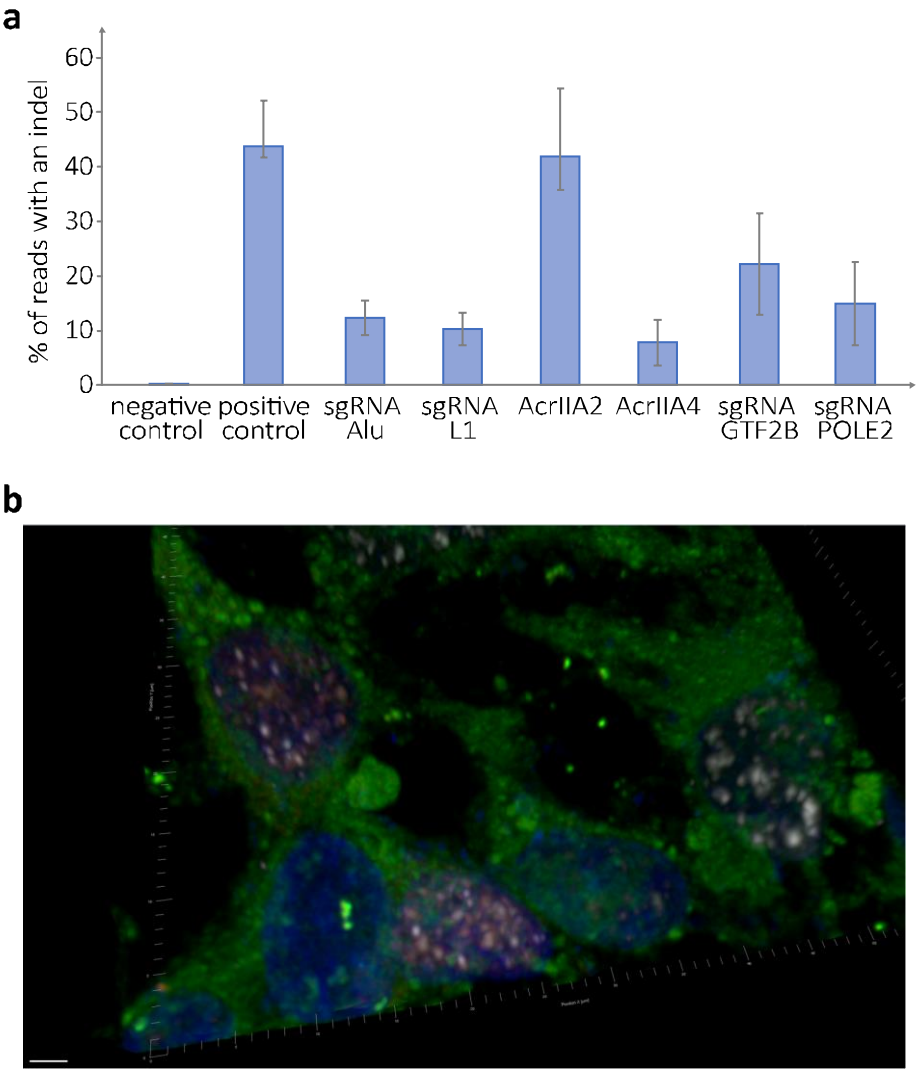


Figure 5.S1 | CRISPR-DS compared to systems targeting essential genes or using anti-CRISPR proteins

(a) Prevention of DNA modification at the JAK2 locus by CRISPR-DS in HEK 293T cell line transfected with either repetitive element sgRNA, anti-CRISPR plasmids, or an essential gene sgRNA. The graph represents the mean of three biological replicates for indel mutation rate at the JAK2 locus three days after transfection, which is plotted on the y-axis. Error bars represent standard error measurement. (b) Representative confocal microscopy (Zeiss TIRF/ LSM 710) images of HEK 293T cells transfected with SpCas9_GFP and sgRNA-Alu three days after transfection. The double-strand DNA breaks are indirectly shown utilizing foci of phosphorylated histones via γH2AX immunostaining. Transfected cells are shown in green due to the pCas9_GFP plasmid used and γH2AX foci appear white and are stained with a secondary Cy5 fluorophore. (Scale bar = 2 um, magnification: 63X).

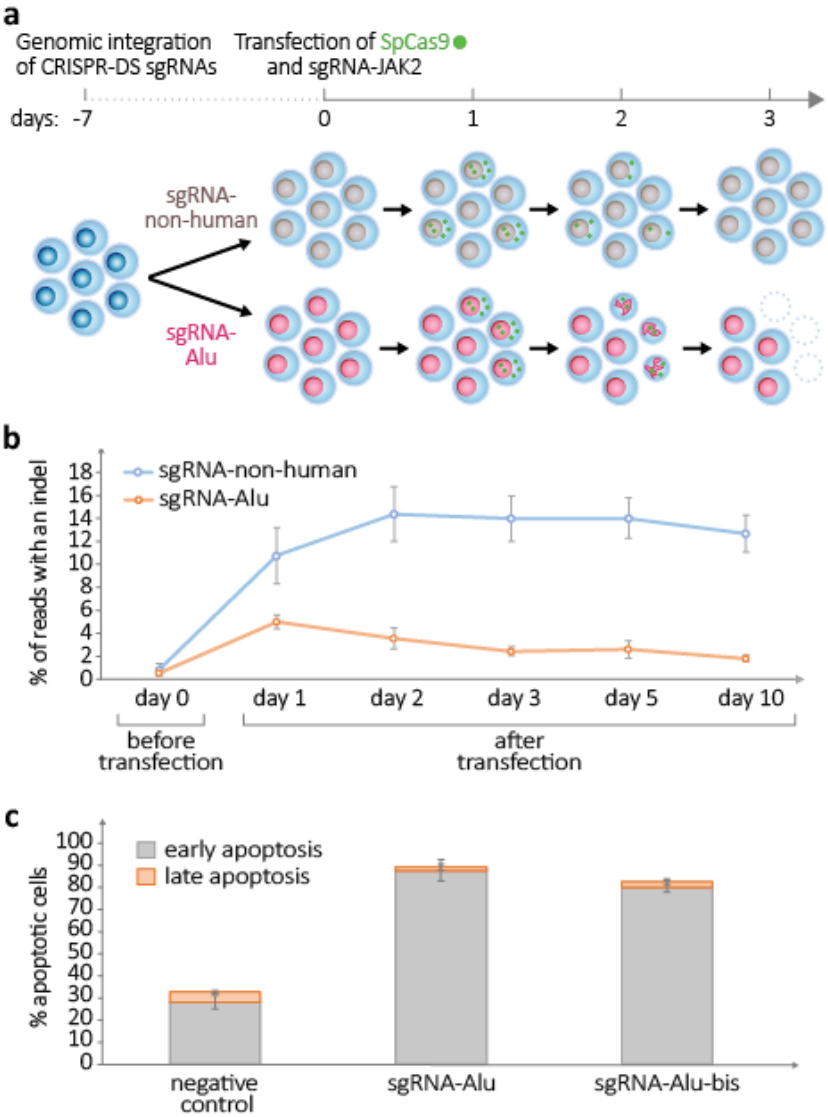


Figure 5.S2 | Prevention of CRISPR-Cas based DNA edits at an endogenous locus in human stem cells.

(a) Visual depiction of the genomic integration of the CRISPR-DS system and experimental overview. Active CRISPR-DS expressing sgRNA-Alu is shown in red while the control (non-functional) CRISPR-DS is shown in light brown expressing sgRNA-non-human. (b) Genome editing efficiency plotted as percentage of insertions or deletions on the y-axis by day on the x-axis. The negative control is a piggybac integrated sgRNA-non-human cell line that was transfected with a SpCas9 expressing plasmid that did not include the sgRNA-JAK2 while the positive control was conducted in the same cell line but included sgRNA-JAK2. (c) Percentage of apoptotic cells is plotted on the y-axis with early apoptosis in gray as measured by Annexin V+ propidium iodide-, and late apoptosis in light orange as measured by Annexin V+ Propidium iodide+ FACS populations.

Enzyme	PAM	Bacterial species	Publication date	# of publications
SpCas9	NGG	*Streptococcus pyogenes*	2012	5068
NmCas9	NNNNGATT	*Neisseria meningitidis*	2013	10
StCas9	NNAGAAW	*Streptococcus thermophilus*	2013	27
SaCas9	NNGRRT	*Staphylococcus aureus*	2015	43
Cpf1	TTTN	*Acidaminococcus*	2015	134
FnCas9	NGG	*Francisella novicida*	2016	41
BlatCas9	NNNNCNDD	*Brevibacillus laterosporus*	2016	1
CjCas9	NNNVRYAC	*Campylobacter jejuni*	2017	5

Table 5.S1 | Most referenced Cas9 orthologs

6. Ethical and regulatory considerations of genome editing

Despite all the promise held by the field of genome engineering I described in this dissertation, it is important to be aware of its potential threats, and to identify the dangers that may lie ahead. Indeed, since the genome is 1) permanent, 2) transmitted to all future generations and 3) encodes the necessary information to generate and regulate the two other pillars of biology – RNA and protein – the ability to become the architect of the genome brings massive responsibilities. Therefore, as our genome designer's capabilities develop, the ethical supervision of its applications will become indispensable.

Since the discovery of CRISPR-Cas9 programmable nuclease in 2013, several research groups have already tried to edit the human germline. In 2015, a zygote was edited to try to modify the human β-globin gene involved in pathologies such as beta thalassemia or sickle cell disease, but embryos were not continued. More recently, researcher He Jankui announced the birth of the first CRISPR-Cas9 edited babies as an attempt to prevent AIDS by disrupting the CCR5 gene known as the HIV entry door in T-cells. Even though programmable nucleases have shown stellar development in past years, these events might have happened prematurely. First, the scientific and general public communities have not reached an ethical consensus on what should or should not be allowed in full knowledge of the implications for future generations and, in a climate of increasing fear of going down a path towards designer babies. Other concerns regard the consent for the procedure: Who will be responsible for the fate of the genetically modified babies? In addition, should we accept the engineering of our genomes to enhance our features? Finally, our understanding and assessment of potential adverse effects of these recent technologies remains limited. Altogether, these reasons require clear and rigorous thinking along with legitimate oversight.

Modifying the germline focuses most of the attention on ethical discussions, since any change would be transmitted down to future generations. This debate, even though it is not new, has regained importance due to the ease of use of CRISPR-Cas9 related technologies and the potential for its quick translation in humans. The scientific community agrees that genome modifications in the germline should not be attempted yet, but that research on the safety of gene editing technologies and delivery methods should be pursued.

Besides the philosophical and societal debate on the purposes for which genome editing should be used, the safety of the technologies is another major limitation to its application. The unintended editing of the genome also known as off-targets, as well as the generation of mosaicism, when not all cells carry the intended modifications, are risks that cannot be ignored. Another limitation of germline editing is the lack of capacity to genetically and functionally characterize the zygote editing after the procedure. If we cannot assess the potential collateral damage, such therapies may result in devastating outcomes for future generations such as malformations and cancers. The scientific community generally acknowledges that germline editing should not be attempted until those concerns are addressed, since the risks will outweigh the benefits of such therapies.

In addition, in some cases, preimplantation genetic diagnosis or in-vitro fertilization technologies may be better suited than genome editing to prevent diseases. Nonetheless, scientists agree that in cases where both parents carry homozygous disease-causing alleles or in cases of polygenic pathologies, genome editing might be a better answer.

In the future, if gene editing technologies become totally safe, allowing the cure for genetic diseases may become a moral imperative; however, people may start to leverage such tools for more controversial purposes such as enhancement, which should be heavily overseen, publicly debated, and regulated accordingly. In a more distant future, the potentially limited accessibility of genetic enhancements to the higher social classes of the population raises concerns about increasing

inequalities within the society and the fear that individuals may only be valued according to their genetic features. In any case, while our knowledge about the genome remains limited, the notion of improving it seems irrelevant or at least extremely biased and ignores long-term unintended and unknown consequences.

Finally, one may wonder about the legitimacy of consent, since the patients affected by the genetic modifications are not born yet: the embryos and future generations. However, in the context of therapies such as preimplantation genetic diagnosis or in-vitro fertilization, parents are already deciding the fate of their future progeny. In addition, ethicists worry about the value of informed consent given to parents since the risks of editing for reproductive purposes remain poorly known.

Even though germline editing should not be attempted before safety has been reached, genome editing in somatic cells could be beneficial and rapidly translated into the clinic. The major driver for adoption will be the identification of the risk-benefit ratio and whether it favors the patient's health outcomes. This will depend on the cell type that needs to be edited (differentiated cells vs progenitors), the nature of the therapy (gene addition, single or multiplexed editing, etc.) and therapeutic application, whether it is for life-threatening diseases (e.g. acute lymphoblastic leukemia) or cosmetic procedures.

Safety and efficacy evaluations may help determine the risk-benefit ratio. Indeed, the major identified risk of gene editing nucleases comes from the generation of double-stranded breaks (DSBs) in the genome. It can simply lead to cell toxicity or create cells that have insertions of deletions in unintended locations in the genome and disrupt gene expression or provoke aberrant chromosomal rearrangements. Therefore, assays need to be designed carefully to prevent such events and their negative consequences. Such a toolbox could include the analysis of off- and on- targets, the genotoxicity of the programmable nucleases, or the impact of other molecules used in combination to improve different editing modalities such as homologous recombination.

Unfortunately, evaluating the genotoxicity of genomic modifications carries a few limitations. First, the toxicity risks linked to genome editing need to be analyzed in a context where cells are naturally undergoing spontaneous modifications. Second, the actual sequencing technologies used to assess genome integrity have a detection threshold that becomes problematic when analyzing whole genomes that are billions of base pairs long, since off-targets and rearrangements may not be detected. Moreover, the sequencing assays are more adapted to therapies using cells which originated from clonal populations. Therefore, as a complement to genetic assays, functional analysis methods such as apoptosis, proliferation, or differentiation tests need to be deployed.

The design of a regulatory framework to oversee genome editing therapies will have to evaluate the medical benefits as compared to the toxicity risks and should be tailored depending on the applications, whether the modifications are in somatic reproductive cell lines, ex- or in- vivo. An additional risk is the potential immunogenicity and the duration of action of the editing agents. To limit such concerns, scientists can deliver the nucleases in protein form that have a more transient action as compared to the use of plasmids, therefore limiting the exposure of the genome to the editing tool and the potential of an immune reaction against it.

The advent of genome editing tools brings hope to the treatment of inherited genetic diseases but holds ethical and safety concerns that could negatively impact future generations. Translating such technologies to the clinic will require the identification and evaluation of the risk-benefit ratio, to provide informed consent under strict regulatory oversight entities such as the NIH Recombinant DNA Advisory Board (RAC) the Food and Drug Administration (FDA), or the European Medicines Agency (EMA).

7. Research overview & perspectives

Engineering whole genomes has the potential to radically transform human health and shape our evolution. In this PhD thesis I described the state-of-the-art of the genome editing tools that may enable such revolution. To enhance the current low-scale multiplexed editing capabilities we decided to undertake the development of new editing tools by leveraging the targeting of transposable elements of the genome. As the project moved along, the observation that targeting repetitive elements using CRISPR-Cas9 triggered massive genotoxicity led us to use this feature to develop bio-containment technologies. The results of this research project are outlined below.

7.1. Multiplexed genome editing: today's limits and tomorrow's promises

The modularity of the CRISPR-Cas9, along with the small size, low cost and rapid production of gRNAs, enables eukaryotic multiplexed genome editing for the first time. This multiplexity allows the high-throughput parallel screening of genomic targets as well as the delivery of multiple distinct gRNAs to modify several loci per cell. In addition, research groups have shown that CRISPR can regulate gene expression by activating or repressing genes in a multiplexed manner. For instance, CRISPR-Cas9 led to the inactivation of the 62 porcine endogenous retroviruses, a major barrier to adoption of pig-based organ transplant therapies. However, it is important to highlight the limitations of such a feat: Cas9 had to be integrated in the genome and expressed over time, potentially causing translocations or other abnormalities in the edited cells; also, the generation of multiple double-stranded breaks in the genome triggered apoptotic responses that drastically reduced the number of surviving cells. Therefore, potential future applications, whether academic, therapeutic, or industrial, will require

substantially higher efficiency and survivability, with genomic modifications multiple orders of magnitude more numerous.

Successful CRISPR multiplex genome editing needs to overcome the challenge of multiple gRNA delivery per cell. Currently, the advances in molecular biology have enabled the generation of more complex constructs containing up to 12 gRNAs and corresponding edits per cell. The transfection of several gRNA plasmids, together with a Cas9 plasmids or the introduction of the Cas9 protein along the gRNA transcripts have also proved efficient at a few targets. However, these approaches will be greatly challenged as the number of genomic targets increases and new technologies will need to be developed.

Improving the actual multiplexed eukaryotic genome editing capabilities by several orders of magnitude holds the potential of revolutionizing human health. Combinatorial functional genomic assays would enable the study of complex genetic traits with applications in evolutionary biology, population genetics, and human disease pathology. In addition, analyzing the functional significance of any generated set of mutations through editing would empower the field of cancer biology. Multiplex editing has also permitted the development of successful engineered cell treatments such as the chimeric antigen receptor (CAR) therapies, which require the simultaneous editing of three target genes. Future treatments may require many more modifications to augment cancer immunotherapies, slow down oncogenic growth, and reduce adverse effects such as graft versus host disease.

CRISPR-based antiviral therapies represent another area that could benefit from larger-scale genome editing. Even though research groups have shown promising results against a number of viral classes such as HIV, HBV and herpes viruses, the ability of viruses to rapidly evolve and evade single-site targeting requires the disruption of several targets at once to prevent viral resistance. Furthermore, the combination of multiplex editing using CRISPR-based transcriptional activation or repression to

improve native viral defenses could further enhance antiviral therapies. Finally, customizing host-versus-graft antigens in human- or nonhuman- donor tissues may require more modifications than have been done so far, for which the development of genome-wide editing technologies is needed. Special attention will be required to the safety of the editing and its impact on the functional activity of the transplants, since donor tissues may persist in the patient for decades.

The ability to write and recode whole mammalian genomes could lead to the generation of cancer-, virus-, and aging-resistant cell lines for the production of biologics or immuno-compatible xenotransplantation. However, genome recoding requires the multiplexed editing to improve to many orders of magnitude. For instance, de-extinction efforts to bring back the wooly mammoth would require the modification of 0.6% of the modern of elephant genome, corresponding from 7.3 to 19.8 million modifications. To achieve such a daunting task, massive multiplex modifications on a genomic scale would require fundamental improvements to methods of editing, delivery vehicles, and donor DNA construction.

Research groups have presented many intriguing observations linking chromatin structure, gene expression, developmental timing and human disease to transposable elements such as Alu, LINE-1 or human endogenous retroviruses (HERVs). The high copy number and identity of these sequences makes their study challenging and establishing causation with their physio-pathological phenotypes requires truly large-scale genome editing.

Multiplexed genome engineering relies on the development of novel technologies to reach its full potential and allow the emergence of the above-mentioned applications. Base editing – which allows the conversion of C → T (or G → A in the complementary strand) or A → G (or T → C in the complementary strand) – brings accuracy, keeps current editing efficiencies, and reduces toxicity and abnormal karyotypes as compared to double-strand break nucleases. Base editors are therefore strong candidates to enable full genome-scale editing. The development of programmable

recombinases in combination with the generation of large donor dsDNA or ssDNA, as well as programmed rearrangement or MAGE technologies, would represent a crucial advance to enable whole-genome engineering without inducing multiple distinct editing events that may be more prone to inducing apoptotic cellular responses.

Finally, multiplexed delivery of editors and vectors is a major limiting factor to genome-wide engineering. Current delivery protocols include the use of lipid nanoparticles, electroporation, microfluidic approaches or microcavitation, to name a few. These technologies permeate the cellular and/or nuclear membrane in order to introduce cargos in the form of plasmids, transcripts or proteins within single cells or tissues. A range of engineered viral vectors with different features are an alternative delivery option that apply to in-vivo delivery as well. For instance, modified adeno-associated viruses are particularly adapted for in-vivo delivery but rely on a relatively small cargo capacity of a few thousands bases, while the herpes simplex virus can incorporate DNA segments to up to 150 kilobases. As for the delivery of large DNA donors or synthetic chromosomes within mammalian cells, it may require technologies enabling cell-cell or nucleus-nucleus fusion.

In the following section I will explain why we selected the Base Editors as potential candidates to enable large-scale genome editing and how we customized them to overcome double-strand break induced cytotoxicity.

7.2. Large-scale genome editing at repetitive elements

As discussed previously, the delivery of multiple gRNAs and the genotoxicity of current editors are two major hurdles to making whole genome engineering possible. In this research project, we focus on overcoming the latter by developing safer editing tools, which we stress-tested by targeting

transposable elements of the genome such as HERV-W, LINE-1 and Alu. These sequences constitute attractive targets since their high level of identity allow us to design a few gRNAs that can target a range going from tens to hundreds of thousands at once.

While the first generation of Base Editors used a nickless Cas9, the low efficiencies led to shift towards the nicking Cas9. We demonstrated that keeping features from more recent BE versions such as the Mu-gam and the improved linker sequence and distribution but converting back to dCas9 reduced their toxicity while retaining their activity. In addition, the use of Pifithrin-alpha, a p53 inhibitor, as well as growth factor bFGF, in combination with the modified editors dCas9-CBE4-gam (dCBE4-gam) and dCas9-ABE (dABE) allowed the successful isolation of highest edited HEK 293T and induced-pluripotent stem cell clones.

At the early phases of the project, as expected, the targeting of the transposable elements using the standard double-strand break Cas9 did not yield results and no viable clones could be detected. We then designed and tested LINE-1 targeting gRNAs to generate a STOP codon using the already existing nicking CBEs (nCBEs) that catalyze C to T nucleotide conversions. While CBE3 showed the highest efficiency 2 days after transfection, CBE4-gam provided the highest edited and viable HEK 293T clones with up to about 781 edits after a second round of transient transfection. Multiplexed editing in a single cell at this level has never been previously reported.

To keep improving multiplexed editing capabilities we decided to inactivate the remaining nickase of the nCas9 to create dCBE4-gam and dABE, hypothesizing that nicking the genome at repetitive elements was contributing to the observed cytotoxicity. dCBE4-gam and dABE activity were first confirmed at a single locus, retaining respectively 53.2% and 40.2% activity as compared to their nicking counterparts. When targeting LINE-1, while nBEs displayed the highest editing efficiencies 2 days after transfection, dBEs resulted in the isolation of the highest edited clones by two more orders of magnitude as compared to nBEs, with up to 6000 edits for dCBE4-gam and 13 200 edits for dABE.

We then undertook to repeat these experiments in induced-pluripotent stem cells that are known to be more sensitive to DNA damage and genome editing. Even though the base conversion percentage was overall lower, we observed the generation of viable clones with up to 3481 sites disrupted genome wide with dABE, while BEs and nBEs failed to produce any detectable edits. This is about three orders of magnitude more than any previously reported multiplexed editing in iPSCs.

To confirm that our large-scale genome editing at LINE-1 elements using CBEs to generate STOP codons were impacting LINE-1 expression, we decided to perform RNA-seq on the clones K-A5 and K-D5, displaying respectively 2.4% and 3.9% edits, corresponding to about 626 and 1018 edits. Theoretically, since most of the active subsets of LINE-1 elements should generate transcripts, the presence of the expected STOP codon at the messenger RNA level may indicate the inactivation of these elements. The results showed that the clone with higher edits at the DNA level displayed higher STOP codons at the RNA level, indicating that potentially active LINE-1 were impacted by the multiplexed editing.

Genome wide disruption of high copy number repetitive elements is now possible and opens new opportunities to study the “dark matter” of the genome. CBEs that allow the generation of STOP codons within an open reading frame will be a great tool to probe at the functions of transposable elements, potentially turning observed associations with physio-pathological phenotypes into causations. However more in-depth studies will be necessary to assess the impact of this massive editing on normal cell processes, since collateral damage may occur. We expect the thorough on- and off- target analysis at repetitive elements to remain a difficult task to accomplish due to their high level of polymorphism, therefore, strong biological controls as well as new experimental and bioinformatics pipelines will be needed to overcome such a challenge.

We envision that these new safe DNA editors, in combination with further improvements in multiplex delivery of gRNAs as well as the modulation of DNA repair and pro-survival pathways, may usher in a new phase of synthetic biology where the recoding of whole mammalian genomes becomes possible.

7.3. CRISPR-mediated bio-containment technologies

In an attempt to develop a protective technology that would prevent the use of CRISPR-Cas9 genome editing for undesirable applications, we created the CRISPR-Defense System or CRISPR-DS. The technology relies on the genotoxicity triggered by the targeting of Alu elements and ensures that introduction or activation of Cas9 triggers cell death, rendering cell populations in which the system is active effectively non-editable by Cas9.

gH2AX staining of double-stranded DNA breaks and standard Annexin V – Propidium Iodide assays confirmed our hypothesis that CRISPR-DS induces apoptosis from massive simultaneous double-stranded DNA cleavage, while remaining non-toxic in the absence of a foreign Cas9-based DNA editor.

Next, we leveraged the potency of the CRISPR-DS mechanism of action to eliminate cells in order to design a bio-safety switch for cell therapies. Indeed, the transplantation of engineered cells can lead to the development of cancers, in the case of stem cells, or the cytokine release syndrome, in the case of CAR therapies. The development of conditional safety switches encoded in the therapeutic cells can serve as a tool to mitigate these potential risks.

We therefore built a CRISPR-DS safety switch in which a doxycycline (DOX) inducible-Cas9 as well as a gRNA targeting Alu were both stably integrated into the cells thanks to the Piggybac transposition system, expecting to trigger the Cas9 expression and subsequent apoptosis by addition of DOX. Experimentally, the “suicide” system worked by eliminating up to 99.98% of engineered cells.

Furthermore, we showed that DOX was not toxic to the cells and that the system does not trigger itself spontaneously in the absence of the activating molecule. These two features are indeed absolutely required for potential clinical translation of the technology.

Programmable site-specific nucleases have already strongly impacted the realm of life sciences thanks to their modularity, versatility, affordability and multiplexability. In this PhD research project, we have successfully increased the multiplexing power of genomic modifications by three orders of magnitude when compared to previously reported records, leading to the generation of highly edited and viable clones and potentially paving the way to genome recoding. We also have harnessed the genotoxicity induced by the targeting of transposable elements in order to develop switches to make current and future cell therapies safer. Along with other research groups optimizing DNA editors to better image cell processes, modulate gene expression or epigenetics, the potential of curing a wide-range of diseases seem to become a realistic goal. However, as our genome designer's capability increases, the ethical safety and sociological risks – some of which we have mentioned – need to be proactively discussed at all layers of society, identified and monitored closely by neutral and legitimate regulators.

List of figures and tables

Main figures

Supplemental figures and tables

Bibliography

1. Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. & Nakata, A. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. *J. Bacteriol.* **169**, 5429–5433 (1987).
2. Venter, J. C. *et al.* The Sequence of the Human Genome. *Science* **291**, 1304–1351 (2001).
3. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
4. de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A. & Pollock, D. D. Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLoS Genet.* **7**, (2011).
5. Kazazian, H. H. *et al.* Haemophilia A resulting from de novo insertion of L 1 sequences represents a novel mechanism for mutation in man. *Nature* **332**, 164 (1988).
6. Singer, M. F., Thayer, R. E., Grimaldi, G., Lerman, M. I. & Fanning, T. G. Homology between the KpnI primate and BamH1 (M1F-1) rodent families of long interspersed repeated sequences. *Nucleic Acids Res.* **11**, 5739–5745 (1983).
7. Skowronski, J. & Singer, M. F. Expression of a cytoplasmic LINE-1 transcript is regulated in a human teratocarcinoma cell line. *Proc. Natl. Acad. Sci. U. S. A.* **82**, 6050–6054 (1985).
8. Xing, J., Witherspoon, D. J. & Jorde, L. B. Mobile element biology: new possibilities with high-throughput sequencing. *Trends Genet.* **29**, 280–289 (2013).
9. Martin, S. L. & Bushman, F. D. Nucleic Acid Chaperone Activity of the ORF1 Protein from the Mouse LINE-1 Retrotransposon. *Mol. Cell. Biol.* **21**, 467–475 (2001).
10. Dombroski, B. A. *et al.* An in vivo assay for the reverse transcriptase of human retrotransposon L1 in Saccharomyces cerevisiae. *Mol. Cell. Biol.* **14**, 4485–4492 (1994).
11. Speek, M. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. *Mol. Cell. Biol.* **21**, 1973–1985 (2001).

12. Wolff, E. M. *et al.* Hypomethylation of a LINE-1 Promoter Activates an Alternate Transcript of the MET Oncogene in Bladders with Cancer. *PLoS Genet.* **6**, (2010).
13. Coufal, N. G. *et al.* L1 retrotransposition in human neural progenitor cells. *Nature* **460**, 1127–1131 (2009).
14. Yu, F., Zingler, N., Schumann, G. & Strätling, W. H. Methyl-CpG-binding protein 2 represses LINE-1 expression and retrotransposition but not Alu transcription. *Nucleic Acids Res.* **29**, 4493–4501 (2001).
15. Muotri, A. R. *et al.* Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* **435**, 903–910 (2005).
16. Kriegs, J. O., Churakov, G., Jurka, J., Brosius, J. & Schmitz, J. Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. *Trends Genet.* **23**, 158–161 (2007).
17. Hormozdiari, F. *et al.* Alu repeat discovery and characterization within human genomes. *Genome Res.* **21**, 840–849 (2011).
18. Polesskaya, O. *et al.* The role of Alu-derived RNAs in Alzheimer's and other neurodegenerative conditions. *Med. Hypotheses* **115**, 29–34 (2018).
19. Kim, S., Cho, C.-S., Han, K. & Lee, J. Structural Variation of Alu Element and Human Disease. *Genomics Inform.* **14**, 70–77 (2016).
20. Cowley, M. & Oakey, R. J. Transposable Elements Re-Wire and Fine-Tune the Transcriptome. *PLoS Genet.* **9**, (2013).
21. Daniel, C., Behm, M. & Öhman, M. The role of Alu elements in the cis-regulation of RNA processing. *Cell. Mol. Life Sci.* **72**, 4063–4076 (2015).
22. Mojica, F. J., Díez-Villaseñor, C., Soria, E. & Juez, G. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Mol. Microbiol.* **36**, 244–246 (2000).

23. Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709–1712 (2007).
24. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
25. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
26. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
27. Beerli, R. R. & Barbas, C. F. Engineering polydactyl zinc-finger transcription factors. *Nat. Biotechnol.* **20**, 135–141 (2002).
28. Reyon, D. *et al.* FLASH assembly of TALENs for high-throughput genome editing. *Nat. Biotechnol.* **30**, 460–465 (2012).
29. Gaj, T., Gersbach, C. A. & Barbas, C. F. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol.* **31**, 397–405 (2013).
30. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80–84 (2014).
31. Shalem, O. *et al.* Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. *Science* **343**, 84–87 (2014).
32. Adamson, B. *et al.* A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* **167**, 1867–1882.e21 (2016).
33. Jinek, M. *et al.* RNA-programmed genome editing in human cells. *eLife* **2**, e00471 (2013).
34. Wang, H. *et al.* One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* **153**, 910–918 (2013).
35. Li, J.-F. *et al.* Multiplex and homologous recombination-mediated genome editing in Arabidopsis and Nicotiana benthamiana using guide RNA and Cas9. *Nat. Biotechnol.* **31**, 688–691 (2013).

36. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173–1183 (2013).
37. Gilbert, L. A. *et al.* CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442–451 (2013).
38. Dominguez, A. A., Lim, W. A. & Qi, L. S. Beyond editing: repurposing CRISPR-Cas9 for precision genome regulation and interrogation. *Nat. Rev. Mol. Cell Biol.* **17**, 5–15 (2016).
39. Sakuma, T., Nishikawa, A., Kume, S., Chayama, K. & Yamamoto, T. Multiplex genome engineering in human cells using all-in-one CRISPR/Cas9 vector system. *Sci. Rep.* **4**, 5400 (2014).
40. Zetsche, B. *et al.* Multiplex gene editing by CRISPR-Cpf1 using a single crRNA array. *Nat. Biotechnol.* **35**, 31–34 (2017).
41. Yang, L. *et al.* Genome-wide inactivation of porcine endogenous retroviruses (PERVs). *Science* **350**, 1101–1104 (2015).
42. Niu, D. *et al.* Inactivation of porcine endogenous retrovirus in pigs using CRISPR-Cas9. *Science* **357**, 1303–1307 (2017).
43. Aguirre, A. J. *et al.* Genomic Copy Number Dictates a Gene-Independent Cell Response to CRISPR/Cas9 Targeting. *Cancer Discov.* **6**, 914–929 (2016).
44. Kuscu, C. *et al.* CRISPR-STOP: gene silencing through base-editing-induced nonsense mutations. *Nat. Methods* (2017). doi:10.1038/nmeth.4327
45. Lin, Y. *et al.* CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. *Nucleic Acids Res.* **42**, 7473–7485 (2014).
46. Black, J. B. *et al.* Targeted Epigenetic Remodeling of Endogenous Loci by CRISPR/Cas9-Based Transcriptional Activators Directly Converts Fibroblasts to Neuronal Cells. *Cell Stem Cell* **19**, 406–414 (2016).

47. Haurwitz, R. E., Jinek, M., Wiedenheft, B., Zhou, K. & Doudna, J. A. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science* **329**, 1355–1358 (2010).

48. Tsai, S. Q. *et al.* Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. *Nat. Biotechnol.* **32**, 569–576 (2014).

49. Nissim, L., Perli, S. D., Fridkin, A., Perez-Pinera, P. & Lu, T. K. Multiplexed and programmable regulation of gene networks with an integrated RNA and CRISPR/Cas toolkit in human cells. *Mol. Cell* **54**, 698–710 (2014).

50. Zetsche, B. *et al.* Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* **163**, 759–771 (2015).

51. Gao, L. *et al.* Engineered Cpf1 variants with altered PAM specificities increase genome targeting range. *Nat. Biotechnol.* **35**, 789–792 (2017).

52. Xie, K., Minkenberg, B. & Yang, Y. Boosting CRISPR/Cas9 multiplex editing capability with the endogenous tRNA-processing system. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 3570–3575 (2015).

53. Gao, Y. & Zhao, Y. Self-processing of ribozyme-flanked RNAs into guide RNAs in vitro and in vivo for CRISPR-mediated genome editing. *J. Integr. Plant Biol.* **56**, 343–349 (2014).

54. Zuris, J. A. *et al.* Cationic lipid-mediated delivery of proteins enables efficient protein-based genome editing *in vitro* and *in vivo*. *Nat. Biotechnol.* **33**, 73–80 (2015).

55. Kim, S., Kim, D., Cho, S. W., Kim, J. & Kim, J.-S. Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Res.* **24**, 1012–1019 (2014).

56. Liang, X. *et al.* Rapid and highly efficient mammalian cell engineering via Cas9 protein transfection. *J. Biotechnol.* **208**, 44–53 (2015).

57. Wang, H. H. *et al.* Programming cells by multiplex genome engineering and accelerated evolution. *Nature* **460**, 894–898 (2009).

58. Isaacs, F. J. *et al.* Precise manipulation of chromosomes in vivo enables genome-wide codon replacement. *Science* **333**, 348–353 (2011).

59. Lajoie, M. J. *et al.* Genomically recoded organisms expand biological functions. *Science* **342**, 357–360 (2013).

60. Ostrov, N. *et al.* Design, synthesis, and testing toward a 57-codon genome. *Science* **353**, 819–822 (2016).

61. Ousterout, D. G. *et al.* Multiplex CRISPR/Cas9-based genome editing for correction of dystrophin mutations that cause Duchenne muscular dystrophy. *Nat. Commun.* **6**, 6244 (2015).

62. Lim, W. A. & June, C. H. The Principles of Engineering Immune Cells to Treat Cancer. *Cell* **168**, 724–740 (2017).

63. Liu, X. *et al.* CRISPR-Cas9-mediated multiplex gene editing in CAR-T cells. *Cell Res.* **27**, 154–157 (2017).

64. Soppe, J. A. & Lebbink, R. J. Antiviral Goes Viral: Harnessing CRISPR/Cas9 to Combat Viruses in Humans. *Trends Microbiol.* **25**, 833–850 (2017).

65. Wang, G., Zhao, N., Berkhout, B. & Das, A. T. CRISPR-Cas9 Can Inhibit HIV-1 Replication but NHEJ Repair Facilitates Virus Escape. *Mol. Ther. J. Am. Soc. Gene Ther.* **24**, 522–526 (2016).

66. Lin, S.-R. *et al.* The CRISPR/Cas9 System Facilitates Clearance of the Intrahepatic HBV Templates In Vivo. *Mol. Ther. Nucleic Acids* **3**, e186 (2014).

67. van Diemen, F. R. *et al.* CRISPR/Cas9-Mediated Genome Editing of Herpesviruses Limits Productive and Latent Infections. *PLoS Pathog.* **12**, e1005701 (2016).

68. Wang, Z. *et al.* CRISPR/Cas9-Derived Mutations Both Inhibit HIV-1 Replication and Accelerate Viral Escape. *Cell Rep.* **15**, 481–489 (2016).

69. Sakuma, T. *et al.* Highly multiplexed CRISPR-Cas9-nuclease and Cas9-nickase vectors for inactivation of hepatitis B virus. *Genes Cells Devoted Mol. Cell. Mech.* **21**, 1253–1262 (2016).

70. Bialek, J. K. *et al.* Targeted HIV-1 Latency Reversal Using CRISPR/Cas9-Derived Transcriptional Activator Systems. *PloS One* **11**, e0158294 (2016).
71. Bogerd, H. P., Kornepati, A. V. R., Marshall, J. B., Kennedy, E. M. & Cullen, B. R. Specific induction of endogenous viral restriction factors using CRISPR/Cas-derived transcriptional activators. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E7249-7256 (2015).
72. Kiani, S. *et al.* Cas9 gRNA engineering for genome editing, activation and repression. *Nat. Methods* **12**, 1051–1054 (2015).
73. Boeke, J. D. *et al.* GENOME ENGINEERING. The Genome Project-Write. *Science* **353**, 126–127 (2016).
74. Gibson, D. G. *et al.* Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome. *Science* **319**, 1215–1220 (2008).
75. Richardson, S. M. *et al.* Design of a synthetic yeast genome. *Science* **355**, 1040–1044 (2017).
76. Mitchell, L. A. *et al.* Synthesis, debugging, and effects of synthetic chromosome consolidation: synVI and beyond. *Science* **355**, (2017).
77. Rogers, R. L. & Slatkin, M. Excess of genomic defects in a woolly mammoth on Wrangel island. *PLoS Genet.* **13**, e1006601 (2017).
78. Miller, W. *et al.* Sequencing the nuclear genome of the extinct woolly mammoth. *Nature* **456**, 387–390 (2008).
79. Rohland, N. *et al.* Genomic DNA sequences from mastodon and woolly mammoth reveal deep speciation of forest and savanna elephants. *PLoS Biol.* **8**, e1000564 (2010).
80. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
81. Wang, H. *et al.* SVA elements: a hominid-specific retroposon family. *J. Mol. Biol.* **354**, 994–1007 (2005).

82. Muotri, A. R. *et al.* L1 retrotransposition in neurons is modulated by MeCP2. *Nature* **468**, 443–446 (2010).

83. Kemp, J. R. & Longworth, M. S. Crossing the LINE Toward Genomic Instability: LINE-1 Retrotransposition in Cancer. *Front. Chem.* **3**, (2015).

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

85. Kim, Y. B. *et al.* Increasing the genome-targeting scope and precision of base editing with engineered Cas9-cytidine deaminase fusions. *Nat. Biotechnol.* **advance online publication**, (2017).

86. Gaudelli, N. M. *et al.* Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* **551**, 464–471 (2017).

87. Moore, J. K. & Haber, J. E. Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in Saccharomyces cerevisiae. *Mol. Cell. Biol.* **16**, 2164–2173 (1996).

88. Meinke, G., Bohm, A., Hauber, J., Pisabarro, M. T. & Buchholz, F. Cre Recombinase and Other Tyrosine Recombinases. *Chem. Rev.* **116**, 12785–12820 (2016).

89. Di Matteo, M. *et al.* PiggyBac toolbox. *Methods Mol. Biol. Clifton NJ* **859**, 241–254 (2012).

90. Hartlerode, A. J. & Scully, R. Mechanisms of double-strand break repair in somatic mammalian cells. *Biochem. J.* **423**, 157–168 (2009).

91. Mao, Z., Bozzella, M., Seluanov, A. & Gorbunova, V. DNA repair by nonhomologous end joining and homologous recombination during cell cycle in human cells. *Cell Cycle Georget. Tex* **7**, 2902–2906 (2008).

92. Orii, K. E., Lee, Y., Kondo, N. & McKinnon, P. J. Selective utilization of nonhomologous end-joining and homologous recombination DNA repair pathways during nervous system development. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 10017–10022 (2006).

93. Grindley, N. D. F., Whiteson, K. L. & Rice, P. A. Mechanisms of site-specific recombination. *Annu. Rev. Biochem.* **75**, 567–605 (2006).

94. Janbandhu, V. C., Moik, D. & Fässler, R. Cre recombinase induces DNA damage and tetraploidy in the absence of loxP sites. *Cell Cycle Georget. Tex* **13**, 462–470 (2014).

95. Akopian, A., He, J., Boocock, M. R. & Stark, W. M. Chimeric recombinases with designed DNA sequence recognition. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 8688–8691 (2003).

96. Mercer, A. C., Gaj, T., Fuller, R. P. & Barbas, C. F. Chimeric TALE recombinases with programmable DNA sequence specificity. *Nucleic Acids Res.* **40**, 11163–11172 (2012).

97. Chaikind, B., Bessen, J. L., Thompson, D. B., Hu, J. H. & Liu, D. R. A programmable Cas9-serine recombinase fusion protein that operates on DNA sequences in mammalian cells. *Nucleic Acids Res.* **44**, 9758–9770 (2016).

98. Xu, Z. *et al.* Accuracy and efficiency define Bxb1 integrase as the best of fifteen candidate serine recombinases for the integration of DNA into the human genome. *BMC Biotechnol.* **13**, 87 (2013).

99. Wang, K. *et al.* Defining synonymous codon compression schemes by genome recoding. *Nature* **539**, 59–64 (2016).

100. Quadros, R. M. *et al.* Easi-CRISPR: a robust method for one-step generation of mice carrying conditional and insertion alleles using long ssDNA donors and CRISPR ribonucleoproteins. *Genome Biol.* **18**, 92 (2017).

101. Paix, A. *et al.* Precision genome editing using synthesis-dependent repair of Cas9-induced DNA breaks. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E10745–E10754 (2017).

102. Papaioannou, I., Simons, J. P. & Owen, J. S. Oligonucleotide-directed gene-editing technology: mechanisms and future prospects. *Expert Opin. Biol. Ther.* **12**, 329–342 (2012).

103. Wrenbeck, E. E. *et al.* Plasmid-based one-pot saturation mutagenesis. *Nat. Methods* **13**, 928–930 (2016).

104. Hegedüs, E., Kókai, E., Kotlyar, A., Dombrádi, V. & Szabó, G. Separation of 1-23-kb complementary DNA strands by urea-agarose gel electrophoresis. *Nucleic Acids Res.* **37**, e112 (2009).

105. Smith, J. J., Antonacci, F., Eichler, E. E. & Amemiya, C. T. Programmed loss of millions of base pairs from a vertebrate genome. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 11212–11217 (2009).

106. Chen, X. *et al.* The architecture of a scrambled genome reveals massive levels of genomic rearrangement during development. *Cell* **158**, 1187–1198 (2014).

107. Mochizuki, K. Developmentally programmed, RNA-directed genome rearrangement in Tetrahymena. *Dev. Growth Differ.* **54**, 108–119 (2012).

108. Yerlici, V. T. & Landweber, L. F. Programmed Genome Rearrangements in the Ciliate Oxytricha. *Microbiol. Spectr.* **2**, (2014).

109. Rees, H. A. *et al.* Improving the DNA specificity and applicability of base editing through protein engineering and protein delivery. *Nat. Commun.* **8**, 15790 (2017).

110. Sharei, A. *et al.* A vector-free microfluidic platform for intracellular delivery. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 2082–2087 (2013).

111. Ding, X. *et al.* High-throughput Nuclear Delivery and Rapid Expression of DNA via Mechanical and Electrical Cell-Membrane Disruption. *Nat. Biomed. Eng.* **1**, (2017).

112. Shalek, A. K. *et al.* Vertical silicon nanowires as a universal platform for delivering biomolecules into living cells. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 1870–1875 (2010).

113. Saklayen, N. *et al.* Intracellular Delivery Using Nanosecond-Laser Excitation of Large-Area Plasmonic Substrates. *ACS Nano* **11**, 3671–3680 (2017).

114. Maresch, R. *et al.* Multiplexed pancreatic genome engineering and cancer induction by transfection-based CRISPR/Cas9 delivery in mice. *Nat. Commun.* **7**, 10770 (2016).

115. Kazuki, Y. & Oshimura, M. Human artificial chromosomes for gene delivery and the development of animal models. *Mol. Ther. J. Am. Soc. Gene Ther.* **19**, 1591–1601 (2011).

116. Doherty, A. M. O. & Fisher, E. M. C. Microcell-mediated chromosome transfer (MMCT): small cells with huge potential. *Mamm. Genome Off. J. Int. Mamm. Genome Soc.* **14**, 583–592 (2003).

117. Brown, D. M. *et al.* Efficient size-independent chromosome delivery from yeast to cultured cell lines. *Nucleic Acids Res.* **45**, e50 (2017).

118. Bi, P. *et al.* Control of muscle formation by the fusogenic micropeptide myomixer. *Science* **356**, 323–327 (2017).

119. Boeke, J. D. *et al.* The Genome Project-Write. *Science* **353**, 126–127 (2016).

120. Ruella, M. & Kenderian, S. S. Next Generation Chimeric Antigen Receptor T Cell Therapy: Going off the Shelf. *BioDrugs Clin. Immunother. Biopharm. Gene Ther.* **31**, 473–481 (2017).

121. Kazazian, H. H. & Moran, J. V. Mobile DNA in Health and Disease. *N. Engl. J. Med.* **377**, 361–370 (2017).

122. Chenais, B. Transposable Elements in Cancer and Other Human Diseases. (2015). Available at: https://www.ingentaconnect.com/content/ben/ccdt/2015/00000015/00000003/art00010. (Accessed: 14th January 2019)

123. Sun, J., Chen, M., Xu, J. & Luo, J. Relationships among stop codon usage bias, its context, isochores, and gene expression level in various eukaryotes. *J. Mol. Evol.* **61**, 437–444 (2005).

124. Shipman, S. L., Nivala, J., Macklis, J. D. & Church, G. M. Molecular recordings by directed CRISPR spacer acquisition. *Science* **353**, aaf1175 (2016).

125. Waltz, E. Gene-edited CRISPR mushroom escapes US regulation. *Nat. News* **532**, 293 (2016).

126. Boyiadzis, M. M. *et al.* Chimeric antigen receptor (CAR) T therapies for the treatment of hematologic malignancies: clinical perspective and significance. *J. Immunother. Cancer* **6**, (2018).

127. Niu, D. *et al.* Inactivation of porcine endogenous retrovirus in pigs using CRISPR-Cas9. *Science* (2017). doi:10.1126/science.aan4187

128. Wang, J. *et al.* Inhibition of activated pericentromeric SINE/Alu repeat transcription in senescent human adult stem cells reinstates self-renewal. *Cell Cycle Georget. Tex* **10**, 3016–3030 (2011).

129. Coufal, N. G. *et al.* L1 retrotransposition in human neural progenitor cells. *Nature* **460**, 1127–1131 (2009).

130. Coufal, N. G. *et al.* Ataxia telangiectasia mutated (ATM) modulates long interspersed element-1 (L1) retrotransposition in human neural stem cells. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 20382–20387 (2011).

131. Kazazian, H. H. *et al.* Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* **332**, 164–166 (1988).

132. Göttle, P. *et al.* Rescuing the negative impact of human endogenous retrovirus envelope protein on oligodendroglial differentiation and myelination. *Glia* (2018). doi:10.1002/glia.23535

133. Burns, K. H. & Boeke, J. D. Human transposon tectonics. *Cell* **149**, 740–752 (2012).

134. Ostertag, E. M. *et al.* A mouse model of human L1 retrotransposition. *Nat. Genet.* **32**, 655–660 (2002).

135. Bodea, G. O., McKelvey, E. G. Z. & Faulkner, G. J. Retrotransposon-induced mosaicism in the neural genome. *Open Biol.* **8**, (2018).

136. Muotri, A. R. *et al.* Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* **435**, 903–910 (2005).

137. Thompson, D. B. *et al.* The Future of Multiplexed Eukaryotic Genome Engineering. *ACS Chem. Biol.* **13**, 313–325 (2018).

138. Zhou, C. *et al.* Highly efficient base editing in human tripronuclear zygotes. *Protein Cell* **8**, 772–775 (2017).

139. Komor, A. C. *et al.* Improved base excision repair inhibition and bacteriophage Mu Gam protein yields C:G-to-T:A base editors with higher efficiency and product purity. *Sci. Adv.* **3**, (2017).

140. Penzkofer, T. *et al.* L1Base 2: more retrotransposition-active LINE-1s, more mammalian genomes. *Nucleic Acids Res.* **45**, D68–D73 (2017).

141. Grandi, N., Cadeddu, M., Blomberg, J. & Tramontano, E. Contribution of type W human endogenous retroviruses to the human genome: characterization of HERV-W proviral insertions and processed pseudogenes. *Retrovirology* **13**, 67 (2016).

142. Chavez, A. *et al.* Highly efficient Cas9-mediated transcriptional programming. *Nat. Methods* **12**, 326–328 (2015).

143. Kleinstiver, B. P. *et al.* Broadening the targeting range of Staphylococcus aureus CRISPR-Cas9 by modifying PAM recognition. *Nat. Biotechnol.* **33**, 1293–1298 (2015).

144. Byrne, S. M. & Church, G. M. Crispr-mediated Gene Targeting of Human Induced Pluripotent Stem Cells. *Curr. Protoc. Stem Cell Biol.* **35**, 5A.8.1-22 (2015).

145. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

146. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

147. Grandi, N., Cadeddu, M., Blomberg, J. & Tramontano, E. Contribution of type W human endogenous retroviruses to the human genome: characterization of HERV-W proviral insertions and processed pseudogenes. *Retrovirology* **13**, 67 (2016).

148. Li, J.-F., Zhang, D. & Sheen, J. Targeted plant genome editing via the CRISPR/Cas9 technology. *Methods Mol. Biol. Clifton NJ* **1284**, 239–255 (2015).

149. Ma, H. *et al.* Correction of a pathogenic gene mutation in human embryos. *Nature* **advance online publication**, (2017).

150. Morgan, R. A. *et al.* Cancer Regression in Patients After Transfer of Genetically Engineered Lymphocytes. *Science* **314**, 126–129 (2006).

151. Rosenberg, S. A. *et al.* Treatment of patients with metastatic melanoma with autologous tumor-infiltrating lymphocytes and interleukin 2. *J. Natl. Cancer Inst.* **86**, 1159–1166 (1994).

152. Grupp, S. A. *et al.* Chimeric Antigen Receptor–Modified T Cells for Acute Lymphoid Leukemia. *N. Engl. J. Med.* **368**, 1509–1518 (2013).

153. Yagyu, S., Hoyos, V., Del Bufalo, F. & Brenner, M. K. An Inducible Caspase-9 Suicide Gene to Improve the Safety of Therapy Using Human Induced Pluripotent Stem Cells. *Mol. Ther.* **23**, 1475–1485 (2015).

154. Yee, C. *et al.* Melanocyte Destruction after Antigen-Specific Immunotherapy of Melanoma. *J. Exp. Med.* **192**, 1637–1644 (2000).

155. Linette, G. P. *et al.* Cardiovascular toxicity and titin cross-reactivity of affinity-enhanced T cells in myeloma and melanoma. *Blood* **122**, 863–871 (2013).

156. Kochenderfer, J. N. *et al.* B-cell depletion and remissions of malignancy along with cytokine-associated toxicity in a clinical trial of anti-CD19 chimeric-antigen-receptor–transduced T cells. *Blood* **119**, 2709–2720 (2012).

157. Batzer, M. A. & Deininger, P. L. Alu repeats and human genomic diversity. *Nat. Rev. Genet.* **3**, 370–379 (2002).

158. Pawluk, A. *et al.* Naturally Occurring Off-Switches for CRISPR-Cas9. *Cell* **167**, 1829–1838.e9 (2016).

159. Rauch, B. J. *et al.* Inhibition of CRISPR-Cas9 with Bacteriophage Proteins. *Cell* **168**, 150–158.e10 (2017).

160. Harrington, L. B. *et al.* A Broad-Spectrum Inhibitor of CRISPR-Cas9. *Cell* **170**, 1224–1233.e15 (2017).

161. Dong, D. *et al.* Structural basis of CRISPR–SpyCas9 inhibition by an anti-CRISPR protein. *Nature* **546**, 436–439 (2017).

162. Brinkman, E. K., Chen, T., Amendola, M. & van Steensel, B. Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Res.* **42**, e168 (2014).

163. Park, J., Lim, K., Kim, J.-S. & Bae, S. Cas-analyzer: an online tool for assessing genome editing results using NGS data. *Bioinformatics* **33**, 286–288 (2017).

164. García, C. P. *et al.* Topoisomerase I inhibitor, camptothecin, induces apoptogenic signaling in human embryonic stem cells. *Stem Cell Res.* **12**, 400–414 (2014).

165. Shipman, S. L., Nivala, J., Macklis, J. D. & Church, G. M. CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria. *Nature* **547**, 345–349 (2017).

166. Zhang, Y. *et al.* CRISPR-Cas9 mediated LAG-3 disruption in CAR-T cells. *Front. Med.* (2017). doi:10.1007/s11684-017-0543-6

167. Zhu, P. *et al.* CRISPR/Cas9-Mediated Genome Editing Corrects Dystrophin Mutation in Skeletal Muscle Stem Cells in a Mouse Model of Muscle Dystrophy. *Mol. Ther. Nucleic Acids* **7**, 31–41 (2017).

Le ciblage des éléments transposables du génome humain pour développer des technologies permettant son remaniement à grande échelle et des technologies de bio-confinement

Mots clés : Génie génétique, CRISPR-Cas9, Base Editing, édition génomique à grande échelle

Les nucléases programmables comme CRISPR-Cas9 sont des signes avant-coureurs d'une nouvelle révolution en génie génétique et portent en germe un espoir de modification radicale de la santé humaine. Le « multiplexing » ou la capacité d'introduire plusieurs modifications simultanées dans le génome sera particulièrement utile en recherche tant fondamentale qu'appliquée. Ce nouvel outil sera susceptible de sonder les fonctions physiopathologiques de circuits génétiques complexes et de développer de meilleures thérapies cellulaires ou traitements antiviraux. En repoussant les limites du génie génétique, il sera possible d'envisager la réécriture et la conception de génomes mammifères. Le développement de notre capacité à modifier profondément le génome pourrait permettre la création de cellules résistantes aux cancers, aux virus ou même au vieillissement ; le développement de cellules ou tissus transplantables compatibles entre donneurs et receveurs ; et pourrait même rendre possible la résurrection d'espèces animales éteintes. Dans ce projet de recherche doctoral, nous présentons l'état de l'art du génie génétique « multiplex », les limites actuelles et les perspectives d'améliorations. Nous tirons profit de ces connaissances ainsi que de l'abondance des éléments transposables de notre ADN afin de construire une plateforme d'optimisation et de développement de nouveaux outils de génie génétique qui autorisent l'édition génomique à grande échelle. En outre, l'observation de la toxicité engendrée par la multitude de coupures double-brins dans le génome nous a amenés à développer un bio-interrupteur susceptible d'éviter les effets secondaires des thérapies cellulaires actuelles ou futures. Enfin, en conclusion, nous exposons les potentielles inquiétudes et menaces qu'apporte le domaine génie génétiques et apportons des pistes de réflexions pour diminuer les risques identifiés.

Targeting the transposable elements of the human genome to enable large-scale genome editing and biocontainment technologies

Keywords: Genome engineering, CRISPR-Cas9, Base Editing, large-scale multiplex editing

Programmable and site-specific nucleases such as CRISPR-Cas9 have started a genome editing revolution, holding hopes to transform human health. Multiplexing or the ability to simultaneously introduce many distinct modifications in the genome will be required for basic and applied research. It will help to probe the physio-pathological functions of complex genetic circuits and to develop improved cell therapies or anti-viral treatments. By pushing the boundaries of genome engineering, we may reach a point where writing whole mammalian genomes will be possible. Such a feat may lead to the generation of virus-, cancer- or aging- free cell lines, universal donor cell therapies or may even open the way to de-extinction. In this doctoral research project, I outline the current state-of-the-art of multiplexed genome editing, the current limits and where such technologies could be headed in the future. We leveraged this knowledge as well as the abundant transposable elements present in our DNA to build an optimization pipeline and develop a new set of tools that enable large-scale genome editing. We achieved a high level of genome modifications to up to 13 000 in HEK 293T cells and 2600 in induced-pluripotent stem cells, about three orders of magnitude greater than previously recorded, therefore paving the way to mammalian genome writing. In addition, through the observation of the cytotoxicity generated by multiple double-strand breaks within the genome, we developed a bio-safety switch that could potentially prevent the adverse effects of current and future cell therapies. Finally, I lay out the potential concerns and threats that such an advance in genome editing technology may be bringing and point out possible solutions to mitigate the risks.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France