DISS. ETH NO. 27452

Identification and validation of novel human genomic safe harbor sites

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

ERIK AZNAURYAN

M.A. in Biotechnology, Columbia University

born on 11.12.1990

citizen of Russian Federation

accepted on the recommendation of

Prof. Sai Reddy, examiner

Prof. Randall Platt, co-examiner

Prof. Michele De Palma, co-examiner

2021

“A ship in harbor is safe, but that is not what ships are built for.”

John Augustus Shedd

Abstract

Numerous gene addition methods are gaining increasing popularity in the field of gene therapy, where replacement of the mutated copy of the gene is required, as well as in cell engineering, in which synthetic receptors can be introduced into a cell or a group of cells to create artificial gene circuits capable of eliciting therapeutic or tissue enhancing functions. Existing gene addition tools suffer from heterogeneity of transgene expression levels and may cause aberration to normal transcriptomic profile due to up- or down-regulation of both protein coding and non-protein coding genes. With the advent of targeted gene integration methods, the necessity for the identification of genomic loci, which would support durable and safe transgene expression – Genomic Safe Harbor (GSH) sites – became ever more urgent. In this dissertation I describe a pipeline for computational prediction and experimental validation of novel human GSH sites using existing as well as newly introduced genomic safety criteria.

In chapter 1 I explain the use of a rational approach to verify computationally predicted genomic sites by targeted integration of reporter as well as therapeutic genes into select computationally predicted locations. This approach yielded the identification of two candidate GSHs, which showed robust and durable expression in investigated cell lines and were later confirmed in primary human T cells and primary human dermal fibroblasts. The safety of transgene expression upon integration into these two sites was subsequently verified using bulk and single-cell transcriptomic analyses, which showed minimal changes in global RNA expression levels following transgene integration. Overall, these two newly identified GSH sites create a broad platform for safer and more reliable gene addition-based gene and cell therapies, facilitating their transition into clinical practice.

In chapter 2 I describe an attempt to implement a multiplexed experimental search of novel GSHs using high-throughput library-based approach. Specifically, described method would allow for a rapid screen of thousands GSH sites exploiting a library of guide RNAs targeting various computationally predicted GSH locations and a non-homologous end joining pathway to drive targeted insertion of a reporter transgene into a genomic locus determined by a guide RNA library member. Such pooled approach would allow to reveal a set of highly transcribed loci, allowing for their subsequent validation by individual transgene integration and transcriptomics assessment. This study, however, was associated with numerous experimental hurdles and was eventually discontinued with suggestions on further optimizations in the future.

To date, only three empirically validated sites in the human genome have been reported for durable expression in different cellular contexts. However, all three of them are located in gene dense regions surrounded by proven oncogenes, significantly increasing the risk of

integration-induced tumorigenesis. Furthermore, they do not support the rapid pace of innovation in synthetic biology that enables multiple transgene integration and genetic circuits to rewire and reprogram cellular function. Two novel, computationally and experimentally validated GSH sites described in this thesis open new opportunities for safer and more predictable genome engineering of human cells, expanding the toolkit for diverse cell therapy and synthetic biology applications, from the treatment of inherited disorders by replacing mutated genes with their functional copies, to creating synthetic networks in immune cells to drive multi-input response, to augmenting properties of cells and tissues by safe addition of enhancing transgenes. Finally, thanks to long-term high levels of transgene expression, identified GSH sites can be used for large-scale therapeutic protein manufacturing in human hosts.

Zusammenfassung

Zahlreiche Methoden der Genaddition erfreuen sich zunehmender Beliebtheit im Bereich der Gentherapie, bei der die mutierte Kopie des Gens ersetzt werden muss, sowie im Bereich des Cell Engineering, bei dem synthetische Rezeptoren in eine Zelle oder eine Gruppe von Zellen eingeführt werden können, um künstliche Genschaltkreise zu schaffen, die in der Lage sind, therapeutische oder gewebeverstärkende Funktionen hervorzurufen. Bestehende Werkzeuge zur Genaddition leiden unter der Heterogenität der Transgenexpressionsniveaus und können aufgrund der Hoch- oder Herunterregulierung sowohl von proteinkodierenden als auch von nicht-proteinkodierenden Genen eine Abweichung vom normalen transkriptomischen Profil verursachen. Mit dem Aufkommen von Methoden zur gezielten Genintegration wurde die Notwendigkeit zur Identifizierung von genomischen Loci, die eine dauerhafte und sichere Transgenexpression unterstützen würden - Genomic Safe Harbor (GSH) Sites - immer dringlicher. In dieser Dissertation beschreibe ich eine Pipeline zur rechnerischen Vorhersage und experimentellen Validierung neuartiger humaner GSH-Stellen unter Verwendung bestehender sowie neu eingeführter genomischer Sicherheitskriterien.

In Kapitel 1 erkläre ich die Anwendung eines rationalen Ansatzes zur Verifizierung rechnerisch vorhergesagter genomischer Stellen durch gezielte Integration von Reporter- sowie therapeutischen Genen in ausgewählte rechnerisch vorhergesagte Stellen. Dieser Ansatz führte zur Identifizierung von zwei GSH-Kandidaten, die eine robuste und dauerhafte Expression in untersuchten Zelllinien zeigten und später in primären menschlichen T-Zellen und primären menschlichen Hautfibroblasten bestätigt wurden. Die Sicherheit der Transgenexpression nach Integration an diesen beiden Stellen wurde anschließend mit Hilfe von Bulk- und Einzelzell-Transkriptomanalysen verifiziert, die minimale Änderungen der globalen RNA-Expressionsniveaus nach Transgenintegration zeigten. Insgesamt schaffen diese beiden neu identifizierten GSH-Stellen eine breite Plattform für sicherere und zuverlässigere Genadditions-basierte Gen- und Zelltherapien und erleichtern deren Übergang in die klinische Praxis.

In Kapitel 2 beschreibe ich einen Versuch, eine multiplexe experimentelle Suche nach neuartigen GSHs mit Hilfe eines bibliotheksbasierten Hochdurchsatzansatzes zu implementieren. Insbesondere würde die beschriebene Methode ein schnelles Screening von Tausenden von GSH-Stellen ermöglichen, indem eine Bibliothek von Leit-RNAs, die auf verschiedene rechnerisch vorhergesagte GSH-Stellen abzielen, und ein nicht-homologer Endverbindungsweg genutzt werden, um die gezielte Einfügung eines Reporter-Transgens in einen genomischen Locus, der durch ein Mitglied der Leit-RNA-Bibliothek bestimmt wurde, zu steuern. Ein solcher gepoolter Ansatz würde es ermöglichen, eine Reihe von hoch

transkribierten Loci aufzudecken, was deren anschließende Validierung durch individuelle Transgenintegration und Transkriptomik-Bewertung erlauben würde. Diese Studie war jedoch mit zahlreichen experimentellen Hürden verbunden und wurde schließlich mit Vorschlägen für weitere Optimierungen in der Zukunft abgebrochen.

Bislang wurden nur drei empirisch validierte Stellen im menschlichen Genom für eine dauerhafte Expression in verschiedenen zellulären Kontexten berichtet. Alle drei befinden sich jedoch in Gendichte-Regionen, die von nachgewiesenen Onkogenen umgeben sind, was das Risiko einer integrationsinduzierten Tumorgenese deutlich erhöht. Darüber hinaus unterstützen sie nicht das rasante Innovationstempo in der synthetischen Biologie, das die Integration mehrerer Transgene und genetischer Schaltkreise zur Neuverdrahtung und Umprogrammierung zellulärer Funktionen ermöglicht. Zwei neuartige, rechnerisch und experimentell validierte GSH-Stellen, die in dieser Arbeit beschrieben werden, eröffnen neue Möglichkeiten für ein sichereres und vorhersagbareres Genom-Engineering menschlicher Zellen und erweitern das Instrumentarium für verschiedene Anwendungen in der Zelltherapie und der synthetischen Biologie, von der Behandlung von Erbkrankheiten durch den Ersatz mutierter Gene durch ihre funktionalen Kopien über die Schaffung synthetischer Netzwerke in Immunzellen zur Steuerung von Multi-Input-Reaktionen bis hin zur Verbesserung der Eigenschaften von Zellen und Geweben durch die sichere Hinzufügung von verstärkenden Transgenen. Schließlich können die identifizierten GSH-Stellen dank der hohen Langzeitexpression von Transgenen für die Herstellung von therapeutischen Proteinen im großen Maßstab in menschlichen Wirten genutzt werden.

Acknowledgements

Conducting a graduate research resembles trying to navigate a gigantic palace: every time you open one door there are at least five or ten more doors behind it. In order to be able to find the desired room one should simultaneously search in multiple directions and have at least a basic initial map of the building. I am therefore extremely grateful to my PhD advisor, Dr Sai Reddy, for allowing me to seek different research and career routes throughout my 4.5 years in his lab, and for numerous resources he and the lab provides to everyone who is fortunate to work there. The initial vector for my research that he established put me on a picturesque road to numerous exciting discoveries and opportunities, for which again I am incredibly thankful. One such opportunity led me to Dr George Church's lab, to whom I am also very thankful for allowing me to develop my research in an exciting and diverse environment of his group and for a few chances to share thoughts on potential applications of my work. In this context, I would also like to thank all my teachers and mentors, each of whom gave me a tiny piece of the map to navigate the labyrinths of graduate research palace. Additionally, I am very grateful to my committee members, Dr Randall Platt and Dr Michele De Palma, who played a key role in productively changing my research approach in the middle of PhD, which is precisely why having engaged committee members is so important. And I want to thank Dr Yaakov Benenson for his interest in this study and agreeing to chair my defense.

Walking along the tangled road of graduate research, I was very happy to be surrounded by an intelligent and dedicated group of people. I want to particularly thank Lucia Csepregi, with whom we were interviewed on the same day and remained close friends ever since; Daniel Neumeier, who is one of the kindest, attentive and hardworking persons I have met and happy to call my friend; Dr Simon Friedensohn, who I was learning from at every conversation we shared; Roy Ehling, whose energy and positivity supported even during unsuccessful lab days; Bastian Wagner, whose breadth and depth of interests always motivated me to learn more; Edo Kapetanovic, whose pragmatic and dedicated approach encouraged me after each of our numerous conversations; Dr Alexander Yermanos, a great collaborator and a person, whose mere presence always made the atmosphere in the room more cheerful; and Dr Rodrigo Vazquez-Lombardi, my football mate and a great source of knowledge on science and babies. I would also like to thank Ann Waindok, Dr Derek Mason, Jakub Kucharczyk, Dr Raphaël Brisset Di Roberto, Florian Beiberich, Beichen Gao, Rocío Castellanos Rueda, Mason Minot and really every member of the lab, past and present, for creating an environment of curiosity and fun every day. Special thanks to Dr Ulrike Haessler for helping keep the workflow in the lab smooth and trying to accommodate to the needs of every lab member. Finally, I want to thank Anna Devaux, a brilliant master student who helped me in the early days of my PhD.

I would like to separately acknowledge Dr Denitsa Milanova, who I am exceptionally thankful for believing in my work and potential opportunities it might bring as well as for introducing me to Church lab and being a great mentor on science and non-science related topics. I also want to thank Isaac Han and Dr David Thompson for making my transition into Church lab smooth and for being inexhaustible sources of conceptual as well as technical knowledge. Finally, I want to thank Elvira Kinzina for being a great collaborator on various projects and a good friend, whose knowledge and ambitions motivated me even before starting PhD.

None of my current and future achievements could have been possible without daily input from my loving family. The visionary wisdom and leading by example that my father voluntarily and involuntarily shares with us every day are significantly broadening my view of life and helping me orient myself in circumstances unseen before. This thesis would definitely not exist without his role in my life. I want to thank my mom for the unconditional, selfless love and endless kindness she energizes and motivates me with on a daily basis. From the very first days, my uncle was supporting me every step of the way, with his wise yet pragmatic advices as well as emotionally and rationally balanced approach to all the issues I was eager to share with him, for which I can't thank him enough. My brother is my best friend and the closest person to me, who I want to thank for playing his unfathomably vital role in shaping me as a person. I thank my cousins for their love and support and for vividly demonstrating me the pace of change in the modern world. I thank my aunt for being a pure demonstration of love and caring for her family. And of course, I want to thank my grandparents, living and deceased, for all the efforts they dedicated to enriching my life with their love, wisdom, knowledge and example. My grandmother is a soul of our family, infusing her warmth in us every second we are fortunate to be in her company. Thanks to all my relatives and friends for playing part in my development.

Finally, I want to thank my wife for having the courage and agreeing to share this unpredictable and eventful journey with me and for amortizing occasionally bumpy road of our life with her love, caring and understanding, which she surrounds me with every day. You are my personal safe harbor. And I want to thank my newborn son, whose smile, even if lasting for a split second, is enough to fill me with immense and incomparable energy and pride, and whose daily achievements drive me to become better too. Your future is one of my main motivators and I am looking forward to more such motivators.

Contributions to this dissertation

Elvira Kinzina helped with bioinformatic search of novel human genomic safe harbor sites based on the existing as well as new criteria (Chapter 2.3.1).

Dr Alexander Yermanos helped with bulk and single-cell RNA sequencing analysis following transgene integration into two novel human genomic safe harbor sites (Chapter 2.3.3 and 2.3.5).

Dr Denitsa Milanova supervised and helped with experimental design of the skin cell therapy work (Chapter 2.3.4)

Anna Devaux was involved in performing experiments for the inducible endogenous gRNA platform development in Jurkat T cells (Chapter 3.4.2).

Table of contents

Abbreviations

List of abbreviations not defined in text

APC	antigen-presenting cell
bp	base pair
CAR T cells	chimeric antigen receptor T cells
cDNA	complimentary DNA
CRISPR	clustered regularly interspaced short palindromic repeats
dsDNA	double-stranded DNA
GFP	green fluorescent protein
kb	kilo bases
lncRNA	long non-coding DNA
SARS-CoV-2	sever acute respiratory syndrome coronavirus 2
sgRNA	single guide RNA
SNP	single nucleotide polymorphism
ssDNA	single-stranded DNA
tRNA	transport RNA
wt	wild type

1 General introduction

1.1 Gene and cell therapy

Millions of people around the world suffer from inherited as well as acquired pathological mutations in the genome, which could lead to a variety of life-threatening diseases, from rare hereditary conditions to oncological tissue transformations. As researchers and clinicians attempt to tackle these pathological conditions, new fields of gene and cell therapy have arisen. These relatively young therapeutic approaches encompass a wide spectrum of methods that rely on changing the malfunctioning gene utilizing various genome editing tools, as well as on the introduction of novel non-mutated copies of disease-causing genes and adding completely new genes carrying therapeutic functions, either in vivo in tissues where the damaged gene is expressed, or ex vivo in cells that are extracted from the body, undergo genetic change and then transferred back.

The history of successful clinical implementation of gene therapy begins in 1990 with a four-year-old girl Ashanthi DeSilva, who was suffering from severe combined immune deficiency (SCID) due to inherited mutation of Adenine Deaminase (*ADA*) gene[1]. An integrative ex vivo transfer of *ADA* gene through retroviral vector resulted in a remarkable increase in T cell count, allowing the child to lead a normal life. The success of this first trial therapy produced a spur of new indications and applications for gene transfer-based gene therapies, with many laboratories and clinical centers around the globe embracing the idea of providing the ultimate "real" treatment to patients with inherited debilitating disorders, as opposed to continuous protein replacement therapies. One of the trials, however, conducted at the turn of the millennia resulted in a significant setback to the entire field. In 1999 Jesse Gelsinger, diagnosed with an inherited ornithine transcarbamylase (OTC) deficiency, was treated with Adenovirus delivered *OTC* gene's cDNA and suffered a severe systemic inflammatory response due to overwhelming viral titers, resulting in death four days after the gene therapy[2]. This fatal outcome led to a major reorganization in the community, with stricter regulations and ethical controls introduced, diming the once bright light shining over clinic prospects of gene therapy[3]. The research in the field continued, however, and with advances in viral vector development gene therapy started to experience a renaissance in the second decade of the 21st century. One of the recent examples of successfully approved and marketed gene therapy drugs is Luxturna developed by Spark Therapeutics and marketed globally by Novartis[4]. Exploiting the AAV's DNA packaging capability, the company successfully manufactured a vector containing *RPE65* gene, mutations in which lead to Leber's congenital amaurosis. Subretinal injections of this vector results in restoration of enzymatic visual cycle and thus the

ability to distinguish light and shadows for patients, who were previously sentenced to complete blindness.

Some of the most exciting applications of gene and cell therapy have been introduced in the field of cancer immunotherapy by the development of CAR-T cell technology. CAR serves as a synthetic receptor capable of binding an extracellular antigen irrespective of MHC presentation. Such binding triggers an activation of CAR's intracellular domain leading to cytotoxic T cell response against the cell bearing the antigen[5]. CARs have demonstrated remarkable efficiency against liquid tumors such as acute lymphoblastic leukemia (ALL), however the effectiveness against solid tumors is yet to be achieved[6]. Currently approved CAR T cell therapies as well as the majority that are in clinical trials rely on random lentiviral integrations of CAR transgene, although targeted approaches are starting to be explored[7].

All of the discussed gene and cell therapy examples exploit the gene addition approach, supplementing cells with functional copies of malfunctioning gene or completely new genetic constructs providing therapeutic effect. A range of natural and synthetic methods exist for the introduction of such exogenous genes into cells.

1.2 Approaches to therapeutic gene addition

1.2.1 Non-integrative gene addition: Viral

One of the oldest and most commonly used methods to deliver genes of interest into target cells without genomic integration relies on natural ability of different viruses to carry nucleic acid payloads and possessing tropism to tissues within the human body.

Adenoviruses (AdV) are non-enveloped dsDNA-bearing viruses that comprise an entire family, members of which – serotypes – mainly differ in the composition of capsid proteins, which in turn determine viral tropism. The capacity for gene inserts of AdVs spans from 8 to 36 kb, and they benefit from a broad tropism to different tissues, with the exception of blood cells. These features allow adenovirus transduction method to be used for delivery of large gene elements into almost all desired target cells with extremely high efficiency[8]. Additionally, the episomal nature of transgene expression (no integration into the cell's genome) following Ad transduction allows for targeting dividing as well as non-dividing cells, albeit the expression in the latter will be diluted over time due to cellular division. The drawback of AdV transduction approach, however, is a particularly potent immune response to proteins encoded in the viral genome. Presentation of the viral epitopes on the surface of targeted cells may result in a potent adaptive cellular immune response, rapidly eliminating virally transduced cells and leading to diminishing of therapeutic effects and life-threatening toxicity. Furthermore, an

immune response to the virus itself may limit the subsequent use of this vector due to the development of adaptive neutralizing antibody response[9]. Efforts have been made to genetically modify adenoviral DNA to avoid this disadvantage. Viral genes have sequentially been removed until the most recent helper-dependent adenovirus (HD AdV), missing almost the entire viral genome has been developed, dramatically reducing the chances of vector toxicity while maintaining transduction and expression efficiency of the wt AdV[10].

Interestingly, one of the recent implementations of AdV vector for DNA delivery was in the realm of vaccines. SARS-CoV-2 virus outbreak resulted in a global pandemic in 2020 and forced a rapid development of vaccines against the new strain. Work on AdV-based vaccines was among the first ones to be initiated due to an overwhelming existing knowledge of this vector as well as ease of large-scale manufacturing. As with all AdV-based treatments, however, the issue of pre-existing immunity to the vector was a concern, so researchers focused on introducing genetic elements encoding SARS-CoV-2 epitopes into AdV serotypes that had low seroprevalence among the treated population[11].

A different example of a non-integrative viral approach uses Adeno-associated virus (AAV) – a ssDNA virus capable of bearing payloads of up to 5kb in length. AAVs are suitable for transduction into dividing as well as non-dividing cells, although AAV's mostly extrachromosomal expression pattern leads to a dilution of expression following cellular division. In rare cases, however, AAV's payload may be integrated into the genome, most frequently into the AAVS1 locus located in an intron of one of the phosphatase genes on chromosome 19[12]. Similar to Ad vector, AAVs possess a broad tropism to various tissues depending on the serotype of the AAV. Currently, more than twelve naturally existing AAV variants have been identified, and even more variants can be generated by directed evolution of capsid proteins, expanding AAV tissue-targeting potential. AAV is significantly less immunogenic and is known to be less toxic compared to Ad[13].

In an effort to increase the packaging size and reduce immunogenicity and toxicity of the AAV vector, researchers have developed a version of the virus devoid of almost all viral genomic sequence – a recombinant AAV (rAAV). The only remaining viral components are T-shaped inverted terminal repeats (ITR) that flank the gene of interest and are needed for the second strand synthesis following delivery into cells. rAAVs are unable to replicate and assemble into a viral particle on their own and require additional elements, that are co-transfected into vector-producing cell line during manufacturing process. Specifically, *rep* and *cap* genes encoding proteins responsible for replication and capsid formation, respectively, are provided on a separate *trans*-plasmid, while other genes involved in mRNA processing and translation and are naturally provided by AdV, are supplemented on a helper plasmid. The plasmid bearing

gene of interest flanked by ITRs, known as *cis*-plasmid, is co-transfected into manufacturing cell line, so that the DNA sequence of the gene of interest is incorporated into the assembling AAV vector[14].

Over 150 clinical trials have been initiated to-date using rAAV as the delivery vehicle for the functional copy of the malfunctioning gene. In addition to recent success in the field of ophthalmological disorders, applications of AAVs are swiftly transitioning into gene therapy of central nervous system, with recent studies showing efficient crossing of the blood-brain barrier by certain AAV serotypes following intravenous injection[15]. Overall, rapid developments in the field of AAV vectors create a landscape for more successful treatments of inherited monogenic diseases.

1.2.2 Non-integrative gene addition: Non-viral

Non-viral methods of clinically relevant gene addition most commonly rely on the use of lipid nanoparticles (LNPs), which are utilized as a delivery method for gene therapy as well as for nucleic acid vaccines. LNPs are typically comprised of self-assembled cationic lipids, which, when formulated with nucleic acids, form spherical nucleic acid bearing compartments in the aqueous environment due to hydrophilic and hydrophobic interactions between head and tail of lipid molecules. The chemical nature of primary head group (single or multiple positive ions) and the structure of the tail group (length and saturation) of the lipid molecules determine the transfection efficiency as well as toxicity of this nucleic acid delivery modality. Positively charged head group of LNP is essential for the interaction with lipid bilayer of cell's plasma membrane composed of negatively charged phospholipids, while chemical and physical properties of the tail group with alkyl bonds in the middle (e.g. oleoyl lipids) allows for increased fluidity, which supports robust fusion between LNP and cell membranes[16].

The toxicity associated with the cationic nature of the head group of the LNPs is a significant concern for therapeutic applications. Cytokine release syndrome is one of the most concerning manifestations of this cytotoxicity as an increased release of inflammokines, such as IL-6 and interferon-γ can lead to generalized inflammation and even death. Furthermore, compounds bearing large net positive charge are known to accumulate in the liver, lung and spleen when administered systemically, and might lead to inflammation in these organs[17]. Numerous optimization steps are being employed by researchers and clinicians in an effort to mitigate undesired toxic side-effects of LNP delivered gene therapy. One of the major advances was the development of LNPs using ionizable cationic lipids, which were shown to be less immunogenic than charged lipids while maintaining the same nucleic acid encapsulation and delivery properties[18].

Currently, LNPs are being extensively used as mRNA vaccine delivery modalities to tackle SARS-CoV-2 pandemic. Specifically, an ionizable cationic LNP encapsulates and delivers spike-protein encoding nucleoside-modified mRNA molecule. Once administered, LNPs target membranes of both APCs as well as other cells, leading to the presentation of Spike epitopes and the development of adaptive cellular and humoral immunity through formation of neutralizing antibodies as well as antigen-specific $CD4^+$ and $CD8^+$ T cells responses[19,20].

All non-integrative gene addition methods showed remarkable potential in cases when targeted cells are not dividing or when transient expression of introduced gene is sufficient for desired outcome (Fig. 1.1A,B). In cases when gene expression needs to be maintained in the dividing tissues, an integrative gene addition is required.

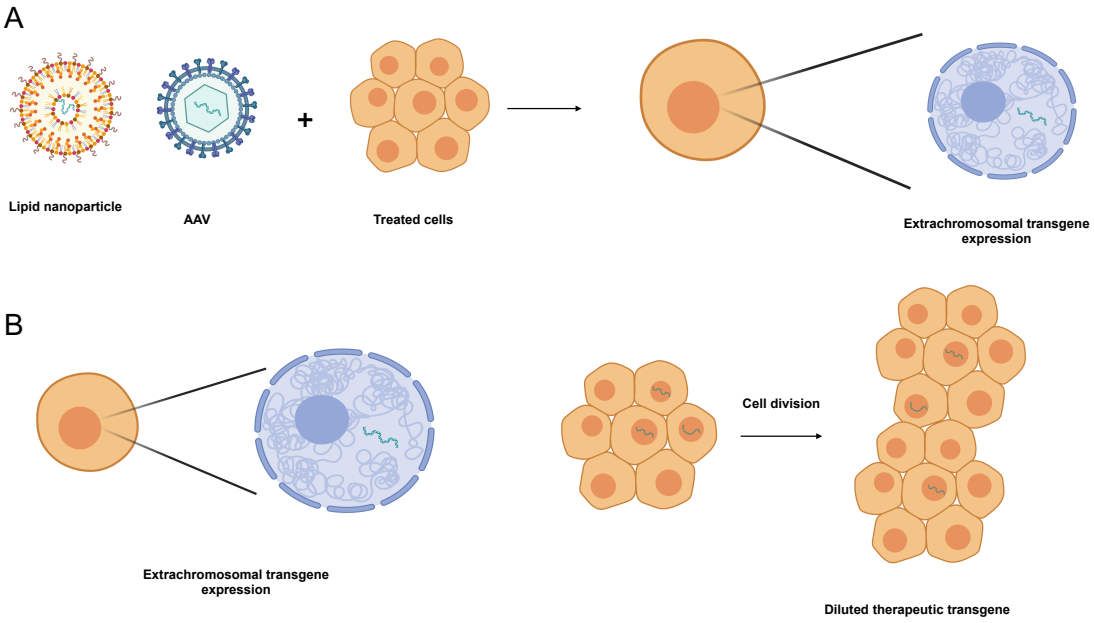


Figure 1.1. Non-integrative gene addition methods result in transient transgene expression. A) A schematic rendering of the use LNP or AAV as a delivery tool for non-integrative gene transfer into desired cells, resulting in an episomal transgene expression. B) The transient nature of such expression eventually leads to the dilution of transgene levels in dividing cells, diminishing the therapeutic effect of this gene addition approach.

1.2.3 Integrative gene addition: Non-targeted

A common approach for gene addition into dividing cells, such as HSCs or T cells, is via γ-retroviral or lentiviral delivery methods. Both of these viral genera are part of *Retroviridae* family that are generally characterized by RNA genome, which upon conversion to DNA is integrated into the genome of targeted cells. Specifically, once bound to the surface of target cells, mature virus fuses with cellular membrane or enters through endocytosis. In the cytoplasm viral RNA is eventually converted into dsDNA through a multi-step reverse transcription (RT) process, and produced DNA is then imported into the nucleus and integrated

into the cell's genome, helped by viral as well as host proteins[21,22]. Retroviruses exhibit semi-random mode of genomic integration, with γ-retroviruses preferentially inserting its reverse transcribed genetic material close to transcriptional start site, CpG islands and DNaseI hypersensitive sites, while lentiviruses integrate inside whole transcriptional units[22]. Possessing the ability to actively translocate across nuclear pores, lentiviral DNA integration step doesn't rely on cell division and can occur in genomic regions located in spatial proximity to nuclear pores in non-dividing cells as well as in cells that yet not entered the cell cycle[23]. On the other hand, γ-retroviruses lack active nuclear import and rely on the breakdown of nuclear envelope during mitosis to integrate into the genome[24].

The most up-to-date approach for retroviral vector generation utilizes three separate plasmids encoding different proteins needed for viral assembly and function. *Gag* gene, responsible for structural properties of the virus, and *pol* gene, necessary for reverse transcription and integration of the viral genome, are encoded on one of the plasmids. The second plasmid bears *rev* gene which regulates viral protein expression, while the third plasmid contains *env* encoding glycoprotein of the viral envelop. These three plasmids are transfected together with the vector plasmid carrying the gene of interest surrounded by long terminal repeats (LTR), which are essential for viral transcription, RT and genomic integration, into a HEK293 cell line which serves as the most common producer of complete retroviral vectors for gene therapy[25].

The main advantage of the retroviral vectors is the size of the desired gene that can be introduced into targeted cells, which can reach 13kb. This enormous capacity places retroviral vectors in a unique position to be used for integrating full cDNA of very large genes driven by long promoters and containing enhancer sequences in dividing cells[26]. The downside of these vectors is related to the unpredictability of the insertion and expression patterns of genomically integrated transgene. In a considerable number of cases integration events happen in the vicinity of oncogenes or tumor suppressor genes, expression of which might be altered leading to malignant transformations of targeted cells (Fig. 1.2A,B). Known cases of such insertional oncogenesis have been described in X-linked severe combined immunodeficiency (SCID-X1) patients undergoing γ-retroviral gene therapy on $CD34^+$ cells. Insertion of gene encoding the IL-2 receptor γ chain 35kb upstream of *LMO2* protooncogene resulted in the development of clonal T cell acute lymphoblastic leukemia[27,28]. Unintended consequences of retroviral integrations also include the development of genomic instability in targeted cells due to aberration in centrosome duplication[29,30]. Although newer versions of retroviral gene therapy seem to reduce the risk of this major side effect, insertional oncogenesis is still a significant concern for the field[21].

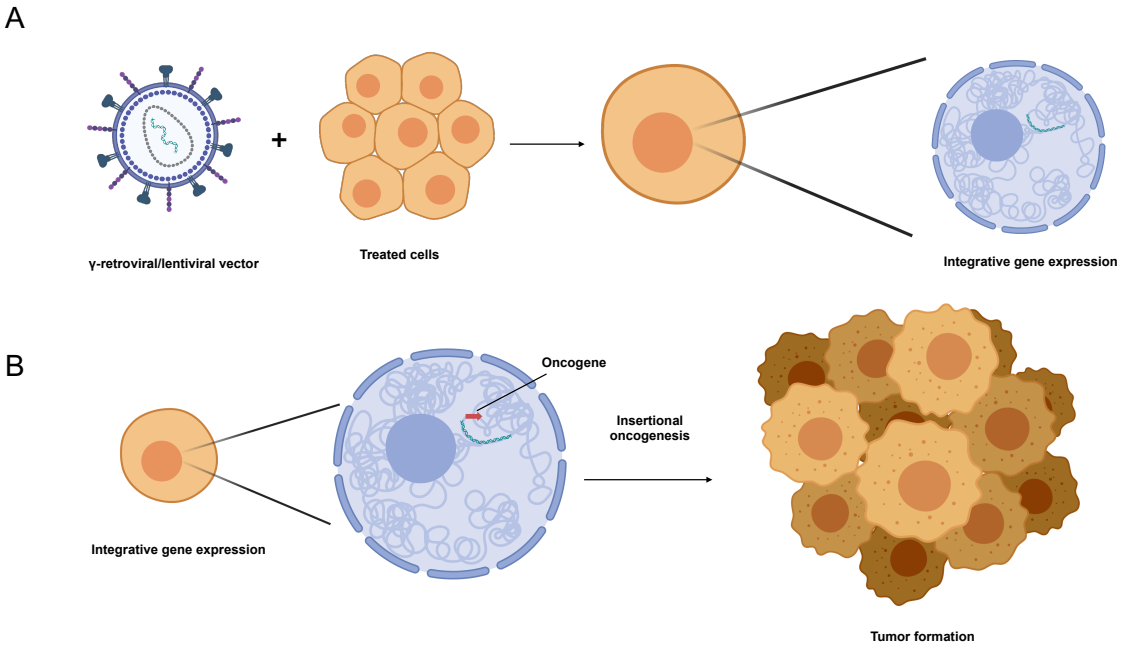


Figure 1.2. Retroviral vectors induce integrative gene addition into a random genomic locus. A) A schematic rendering of the use lentiviral vectors as a delivery tool for integrative gene transfer in desired cells, resulting in a genomic insertion of the gene of interest. B) Potential side effects of such genomic insertions involve upregulation of oncogenes or downregulation of tumor suppressor genes, all of which may result in tumorigenesis.

1.2.4 Integrative gene addition: Targeted

Last two decades has seen a spur of new technologies aimed at precise targeting of desired sites in the genome for generating genetic knockouts, modifying SNPs or integrating large transgenes for research as well as clinical applications.

Meganucleases are one of the earliest precise genome engineering tools initially identified in phages, bacteria, and certain eukaryotes. This large family of enzymes recognize specific DNA targets of 12-45 bp in length and are capable of eliciting a double-strand break (DSB) in the target sequence. Despite possessing high efficiency in producing targeted DSBs, this approach has a significant drawback: desired genomic target site needs to contain the precise cleavage sequence recognized by a given maganuclease. Thus, an initial introduction of such cleavage sequence in target genomic site of interest is required, making genome engineering using meganucleases a long and cumbersome process[31]. Various meganuclease engineering efforts attempted to expand the target sequence, but the challenge of targeting diverse range of genomic sites remained. A major breakthrough occurred with the development of zinc finger nucleases (ZFNs) – synthetic proteins possessing separate DNA-binding and DNA-cleavage domains. The DNA-binding domain is composed of modular zinc fingers – 30 amino acid units

that dictate specificity to target DNA region, with each zinc finger recognizing a triplet of nucleotides.[32] Through protein engineering techniques researchers have developed zinc fingers for most triplet sequences, allowing to target a large portion of the genome. The DNA-cleavage domain is most commonly composed of *Fok*I – a natural type IIS restriction enzyme with no apparent sequence specificity. *Fok*I enzyme has to dimerize in order to cleave DNA sequence determined by zinc fingers. This dimerization requirement implies a double DNA binding event by ZFNs, thus improving specificity of targeted DNA break.[33] The disadvantage of genome engineering using ZFNs is associated with a long and labor-intensive process of creating new ZFN proteins for each genomic target site. Additionally, some triplets cannot be recognized by zinc fingers limiting their use for certain genomic regions, and encouraging the development of target-specific nucleases with single nucleotide resolution[34]. The latter concern was precisely addressed by the development of transcription activator-like effector nucleases (TALENs). Similar to ZFNs this genome engineering tool is composed of TAL effector domain responsible for binding to specific DNA sequence and a *Fok*I nuclease, which elicits sequence-unspecific DNA break. TAL effector domain possesses a repetitive region of 13-29 tandem repeats, each comprised of 34 amino acids[35]. These repeats differ by only 2 amino acid residues located at position 12 and 13, referred to as repeat variable di-residues (RVD). Each RVD encodes binding to specific base on the DNA sequence, thus allowing for one repeat-one nucleotide recognition and hence making TALENs more versatile in targeting virtually any genomic region[36]. Despite being a significant improvement over ZFNs, the engineering of TALENs remains cumbersome and lengthy, with each new genomic target site requiring a new protein production step[34].

The most recent and by far the most widely adopted of the targeted genome engineering technologies is CRISPR – a method based on bacterial adaptive immunity system evolutionarily developed against invading bacteriophages and capable of incorporating parts of phage genome that penetrated bacterial cell into a specific locus of the bacterial chromosome. These incorporated pieces of DNA – protospacers – are composed of 20 bp and can then be transcribed together with repetitive sequences known as direct repeats, producing an array of phage derived protospacers and direct repeat sequences. The array is then processed by auxiliary CRISPR proteins producing crisprRNAs (crRNA, determined by the protospacer), which, when bound by separately transcribed tracerRNA (trRNA), associate with Cas9 enzyme and direct the latter to phage DNA upon subsequent invasions. Complementarity-based binding of cr-trRNA to the phage genome brings Cas9 – most commonly used nuclease – into close physical proximity to the invading DNA, and the enzyme then carries out a blunt-ended double-strand DNA break in a position 3-4 nucleotides upstream of the protospacer adjacent motif (PAM) – sequence of nucleotides immediately

following the protospacer sequence – which most frequently for Cas9 is NGG[37]. Thus, change of the specificity of the CRISPR system is dictated by a change of 20 nucleotides in the crRNA sequence, which can be cheaply and rapidly chemically synthesized. Absence of protein engineering and purification steps needed for each new ZFN and TALEN experiment, make CRISPR-based genome engineering approaches easier and faster compared to the previous techniques[1].

Researchers learned to utilize double-strand break induced by described genome engineering tools for targeted genome editing and gene addition in human cells. Scientists realized that repair pathways that exist in the human cells to mitigate different types of DNA damage, including double-strand DNA break, can be exploited to elicit precise genome engineering events. One of these repair pathways which is most relevant for large gene additions is homology directed repair (HDR). This mechanism involves the use of the homologous DNA sequence on the second intact chromosome as a template for the repair of double-strand breaks. Specifically, after the break 5'-ended strands are resected, the broken sister chromatids form heteroduplexes, which are eventually resolved by synthesis of a complementary strand, thus filling the resected gap and nicking the double-strand break. If an exogenous DNA sequence bearing gene of interest flanked by homology arms to the target integration site is supplied during the generation of Cas9-induced break, this sequence can serve as a substrate for the HDR and result in targeted insertion of gene of interest into the desired locus[38–40].

Despite incredible versatility, ease of use and low cost of the CRISPR system for targeted gene insertion, it is still associated with a few drawbacks. First, the HDR pathway only occurs in the mitotically active cells, since the sister chromatid may only be present in S and G2 phases. Thus, this approach might not be suitable for non-dividing cells, such as neurons. Methods to mitigate this exist in the genome engineering toolkit, e.g., using alternative repair pathways that operate in G0 and G1, such as non-homologous end-joining (NHEJ) or microhomology-mediated end joining (MMEJ), which unlike HDR are not scarless and may leave insertions, deletions or duplications at the target site (Fig 1.3)[41]. Furthermore, CRISPR-based cutting event do not always happen exclusively at the desired genomic targets, but frequently results in off-target events. These can be predicted with modern computational tools, albeit still frequently unavoidable[42].

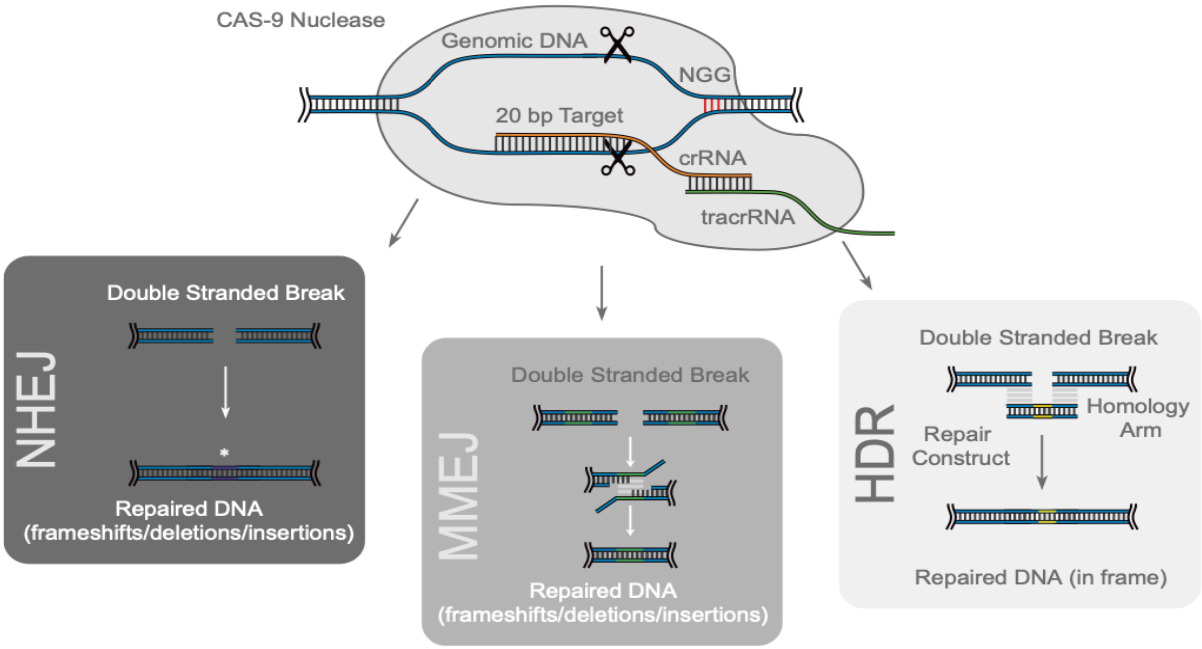


Figure 1.3. Overview of CRISPR/Cas9 genome engineering. After complexing with a gRNA molecule (crRNA:tracrRNA), Cas9 nuclease is guided to the DNA target site, where it creates a double-stranded DNA break. The cell reacts with endogenous DNA repair mechanisms (NHEJ, MMEJ, HDR), that can either generate indels for knockouts or the precise sequence substitution in the presence of a homologous sequence enabling targeted modifications or insertions of specific DNA sequences. Figure adapted from Kelton, Pesch et al[43].

1.3 Genomic safe harbor sites

Despite the advent of precise genome engineering methods, the target human genomic sites for safe and stable transgene integration and expression – genomic safe harbors (GSH) – are yet to be identified. Currently, three loci are used in the research setting for stable and robust transgene expression in a variety of cell types – AAVS1, CCR5 and human ROSA26 (Fig.1.4A,B,C)[44]. Despite being able to support long-term expression of genes of interest without apparent detrimental effect to host cells, all three sites are located inside introns of functional genes, physiological roles of which are not clearly understood yet. Furthermore, these sites locate in close proximity with oncogenes, thus insertion of exogenous promoter sequences and active transcription of transgenes from all three sites may lead to oncogenesis[45,46]. Recently, several research groups have attempted to identify novel GSH sites, focusing the search exclusively on sites predetermined by various genomic integration tools (e.g. lentiviral deliveries or cre-recombinase sites)[47,48]. These methods, however, omit a vast number of potential GSH sites that are not targeted by these tools, which significantly

limits the search.

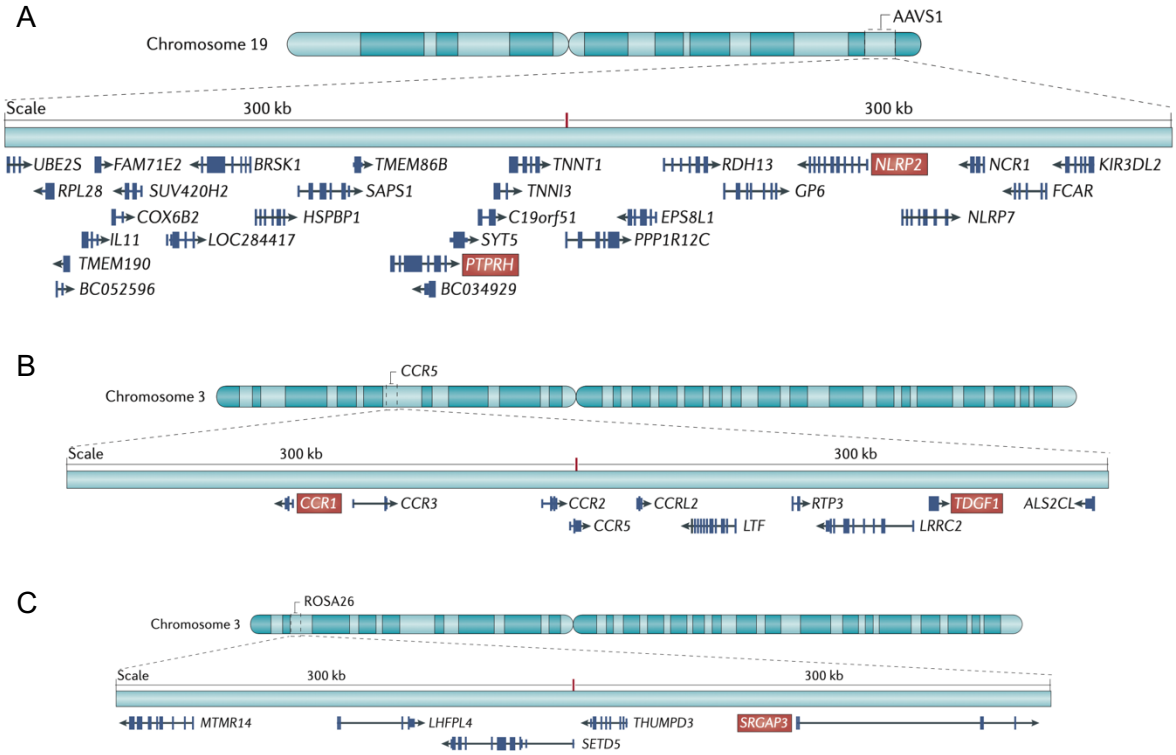


Figure 1.4. Currently used genomic sites for transgene integration. A) AAVS1 site is located in the first intron of *PPP1R12C* gene on the q arm of chromosome 19. It serves as the most common integration site for rare genomic integration events of AAV vector and is located in a gene dense region surrounded by genes (in black) and proto-oncogenes (in red). B) CCR5 site is located in the first intron of *CCR5* gene on the p arm of chromosome 3. CCR5 chemokine is thought to be dispensable for cells, although its full function has not been fully investigated yet. This site is also surrounded by genes and proto-oncogenes in a close linear proximity. C) ROSA26 site is located in the intron of *THUMPD3* gene, also on the p arm of chromosome 3. The function of *THUMPD3* gene is not well-understood yet. ROSA26, similar to the other two sites, is surrounded by genes and proto-oncogenes. Figure adapted from Sadelain et al[44].

Safety evaluation of the original three as well as newly identified genomic safe harbor sites revolved around the assessment of the transcription of genes located within short linear distance from the insertion site (around 300 kb). Despite absence of aberrations in the expression of the nearby genes, detrimental changes in gene expression following integration into these sites couldn't be excluded since the whole transcriptome profile of integrated cells was not assessed[49]. RNA-sequencing of entire transcriptomes in engineered cells would provide a more complete picture of changes that occur in cells following transgene integrations in the genome. Additionally, recent developments in single-cell RNA-sequencing allow to infer gene expression perturbations on the level of individual cells, providing a deeper understanding of transcriptomic changes happening in cells[50]. Utilizing this tool to verify safety of the investigated GSH site would thus be of particular relevance, as it would provide the most comprehensive assessment of the consequences of such genomic integrations,

especially in the context of a genome-wide threat of insertional oncogenesis.

For a more technical discussion of GSH that are being investigated in the research and clinical gene therapy as well as recent approaches to GSH discovery, the reader should refer to Chapter 2 Introduction section.

1.4 Cell types for GSH-based genome engineering

A range of different cell types can be envisioned as primary beneficiaries of targeted gene integrations into GSH sites. Such approaches could be particularly advantageous for mitotically active cells, for which episomal transgene expression will lead to dilution following cellular division, while use of integrative viral vectors may lead to insertional oncogenesis[27,51]. One of the examples here is primary human T cell engineering, which can be conducted ex vivo and infused back into the patients to trigger targeted T cell response. Several T cell engineering approaches have been described and used in the clinical practice. The most successful of them are CAR-T cell therapies, which rely on genomic integration of synthetic antigen receptor for targeted triggering of cytotoxic response against malignant cells[52]. This technology relies on random viral-mediated transgene knock-ins – a method that can be substituted by targeted gene insertion into a verified genomic safe harbor site to avoid complications of insertional oncogenesis. Similarly, T cells bearing other engineered immune receptors, such as synNotch receptor capable of eliciting customizable T cell response against target antigens[53], will benefit significantly from stable and safe expression from a validated GSH, enabling its faster transition into clinics. In this context, studying Jurkat T cell line as a proxy for T cell engineering could be particularly useful[54]. Experimentally validating novel genomic safe harbor sites for safety and durability of transgene expression in this cell line could be an important initial step before transitioning to primary T cells.

Engineering of other blood cells can also exploit the use of genomic safe harbors for targeted ex vivo gene knock-ins. Lentiviral based transgene integrations into hematopoietic stem cells (HSCs) for treatment of inherited immune deficiencies, hemoglobinopathies and metabolic disorders have seen particular success over the last years[55]. However, these therapies can also be subject to side effects of semi-random transgene insertions. For instance, lentivirus-driven ex vivo integration of hemoglobin subunit beta gene into HSCs of patients suffering from transfusion-dependent β thalassemia has advanced into clinical trials, which were subsequently halted due to the development of acute myeloid leukemia (AML) and myelodysplastic syndrome (MDS) in two patients (clinicaltrials.gov – NCT04628585). Although other mechanisms beyond lentiviral integration may have caused this side effect, the risk of

insertional oncogenesis is still preoccupying scientists and clinicians utilizing lentiviral cell therapies, and transition to GSH-based therapeutic gene integrations would be beneficial in alleviating these concerns.

Gene therapy of inherited skin disorders could be another exciting application for GSH sites. Mutations in keratinocytes and fibroblasts, which undergo continuous proliferation throughout life and during wound healing, respectively, result in severe pathologies that could eventually lead to patient death. Among such genetic diseases are over 30 types of epidemolysis bullosa – a debilitating condition manifested in blister formation due to mutations in genes responsible for structural integrity of skin layers[56]. AAV-based gene therapy approaches to this group of diseases require repetitive administration due to constantly dividing nature of skin cells[57]. On the other hand, targeted in vivo integration of functional copies of mutated genes into safe harbor sites in fibroblasts or epidermal stem cells could provide a lasting treatment. Additionally, ex vivo safe harbor-based engineering of skin grafts could be beneficial for people with burns, augmenting transplanted skin with regenerative factors[58].

Furthermore, gene expression from genomic safe harbor sites can be used for industry relevant protein manufacturing in human host cells, such as HEK293. Currently, plasmid based episomal expression is used by the majority of academic laboratories for the recombinant protein production purposes[59]. Big biopharmaceutical companies however rely on proprietary target sites for gene insertion and cell banking. Identification of novel safe harbor sites that provide durable high-level transgene expression may be of significant utility for academic and industrial protein production.

Hypothesis and Objectives

Currently employed techniques in gene and cell therapies are either not suitable for mitotically active cells due to episomal nature of current non-integrative transgene expression, or may lead to a serious complication of insertional oncogenesis due to random integration of the transgene into the human genome using current integrative approaches. Possessing genomic sites that would allow for safe, durable and predictable transgene expression in a variety of cellular context could significantly enhance mammalian synthetic biology and serve as a useful tool for clinical cell therapy applications as well as industrial protein manufacturing. Several attempts to identify suitable genomic regions have been made by researchers, however all of them were limited to only those sites that were predetermined by the nature of the transgene insertion as well as lacked a comprehensive assessment of global transcriptomic changes following the integration and expression of desired transgenes.

In this thesis I attempt to identify regions in the human genome that would be suitable for safe and stable expression of genes of interest in various cell types – Genomic Safe Harbor sites. In order to identify such sites, existing safety criteria are used based on linear distance from functional genes and oncogenes described in the literature and additional criteria are added to avoid various regulatory elements, such as lncRNAs, tRNAs, and structural units of chromosomes – centromeres and telomeres. A bioinformatic search is conducted based on these criteria to predict putative GSH sites. A handful of the predicted sites are then selected and experimentally tested for the durability of reporter and therapeutic protein expression in different cell types. The safety of the sites that support durable transgene expression is validated by bulk and single-cell transcriptomics assays. Finally, a high-throughput approach is attempted to screen large number of putative GSHs using a library of sgRNAs targeting bioinformatically predicted genomic loci.

2 Discovery and validation of novel human genomic safe harbor sites for gene and cell therapies

This is an author-produced version of an article submitted for publication in *Cell Genomics*: [Erik Aznauryan, Alexander Yermanos, Elvira Kinzina, Anna Devaux, Edo Kapetanovic, Denitsa Milanova, George M. Church, Sai T. Reddy]

E.A. and S.T.R designed the study; E.A., A.Y., E.Kap., D.M., G.M.C and S.T.R. contributed to experimental design; E.A. and A.D. performed experiments; E.Kin. developed the bioinformatic pipeline for GSH identification; E.A. and A.Y analyzed data; G.M.C. and S.T.R. supervised the work; E.A, A.Y. and S.T.R wrote the manuscript with input from authors.

ETH Zürich and Harvard University have filed for patent protection on the technology described herein, and E.A., D.M., S.T.R. and G.M.C are named as co-inventors on the patent.

2.1 Summary

Existing approaches for the integration and expression of genes of interest in a desired human cellular context are marred by the safety concerns related to either the random nature of viral-mediated integration or unpredictable pattern of gene expression in currently employed targeted genomic integration sites. Disadvantages of these methods lead to their limited use in clinical practice, thus encouraging future research in identifying novel human genomic sites that allow for predictable and safe expression of genes of interest. We conducted a bioinformatic search followed by experimental validation of novel genomic sites and identified two that demonstrated stable expression of integrated reporter and therapeutic genes without detrimental changes to cellular transcriptome. The cell-type agnostic criteria used in our bioinformatic search suggest wide-scale applicability of our sites for engineering of a diverse range of tissues for therapeutic as well as enhancement purposes, including modified T-cells for cancer therapy and engineered skin to ameliorate inherited diseases and aging. Additionally, the stable and robust levels of gene expression from identified sites allow for their use in industry-scale biomanufacturing of desired proteins in human cells.

2.2 Introduction

Development of technologies for predictable, durable and safe expression of desired genetic constructs (i.e., transgenes) in human cells will contribute significantly to the improvement of gene and cell therapies[60,61], as well as for protein manufacturing[62]. One prominent beneficiary of such technologies are genetically engineered T-cell therapies, which requires genomic

integration of transgenes encoding novel immune receptors[63,64]; another example are gene therapies for highly proliferating tissues, such as inherited skin disorders, in which entire wild-type gene copies have to be integrated into epidermal stem cells[65,66]. Advances in genome editing using targeted integration tools[67] already allow precise genomic delivery and sustained expression of transgenes in certain cellular contexts, such as chimeric antigen receptors (CARs) integrated into the T cell receptor alpha chain locus in T-cells[7], and coagulation factors delivered to hepatocytes using recombinant adeno-associated viral (rAAV) vectors[68]. These applications, however, are limited to specific cell types and cause disruption to the endogenous genes, limiting the diversity of cellular engineering applications. Specific loci in the human genome that support stable and efficient transgene expression, without detrimentally altering cellular functions are known as Genomic Safe Harbor (GSH) sites. Thus, precise integration of functional genetic constructs into GSH sites greatly enhances genome engineering safety and efficacy for clinical and biotechnology applications.

Empirical studies have identified three sites that support long-term expression of transgenes: AAVS1, CCR5 and hRosa26 – all of which were established without any a-priori safety assessment of the genomic loci they reside in[46]. The AAVS1 site, located in an intron of *PPP1R12C* gene region, has been observed to be a region for rare genomic integration events of the Adeno-associated virus's payload[69]. Despite being successfully implemented for durable transgene expression in numerous cell types[70], the AAVS1 site location is in a gene-dense region, suggesting potential disruption of expression profiles of genes located in the vicinity of this loci[44]. Additionally, studies indicated frequent transgene silencing and decrease in growth rate following transgene integration into AAVS1[45,71], which represents a liability for clinical gene therapy. The second site lies within the *CCR5* gene, which encodes a protein involved in chemotaxis and also serves as co-receptor for HIV cellular entry in T cells[72]. Serendipitously, researchers identified that the naturally occurring CCR5-delta-32 mutation present in people of Scandanavian-origin results in an HIV-resistant phenotype[73]. This finding suggested disposability of this gene and applicability of CCR5 locus for targeted genome engineering, especially for T cell therapies[74,75]. However, similar to AAVS1, the CCR5 locus is located in a gene-rich region, surrounded by tumor associated genes[44], thus severely limiting its safe use for therapeutic purposes. Additionally, *CCR5* expression has been associated with promoting functional recovery following stroke[76], thus disrupting *CCR5* may be undesirable in clinical practice. The third site, human Rosa26 (hRosa26) locus, was computationally predicted by searching the human genome for orthologous sequences of mouse Rosa26 (mRosa26) locus[77]. The mRosa26 was originally identified in mouse embryonic stem cells by using random integration by lentiviral-mediated delivery of gene trapping constructs consisting of promotorless transgenes (β-galactosidase and neomycin phosphotransferase), resulting in

sustainable expression of these transgenes throughout embryonic development[78,79]. Similar to the other two currently employed GSH sites, hRosa26 is located in an intron of a coding gene *THUMPD3*[77], the function of which is still not fully characterized. This site is also surrounded by proto-oncogenes in its immediate vicinity[44], which may be upregulated following transgene insertion, thus potentially limiting the use of hRosa26 in clinical settings.

Attempts have been made to identify new human GSH sites that would satisfy various safety criteria, thus avoiding the disadvantages of existing sites. One approach developed by Sadelain and colleagues used lentiviral transfection of beta-globin and green fluorescence protein (GFP) genes into induced pluripotent stem cells (iPSCs), followed by the assessment of the integration sites in terms of their linear distance from various coding and regulatory elements in the genome, such as cancer genes, miRNAs and ultraconserved regions[49]. They discovered one lentiviral integration site that satisfied all of the proposed criteria, demonstrating sustainable expression upon erythroid differentiation of iPSCs. However, global transcriptome profile alterations of cells with transgenes integrated into this site were not assessed. A similar approach by Weiss and colleagues used lentiviral integration in Chinese hamster ovary (CHO) cells to identify sites supporting long-term protein expression for biotechnological applications (e.g., recombinant monoclonal antibody production)[48]. Although this study led to the evaluation of multiple sites for durable, high-level transgene expression in CHO cells, no extrapolation to human genomic sites was determined. Another study aimed at identifying novel GSHs through bioinformatic search of mCreI sites residing in loci that satisfy GSH criteria[47]. Similarly, to previous work, several stably expressing sites were identified and proposed for synthetic biology applications in humans. However, local and global gene expression profiling following integration events in these sites have not been carried out.

All of the potential new GSH sites possess a shared limitation of being narrowed by lentiviral- or Cre-based integration mechanisms. Additionally, safety assessments of some of these newly identified sites, as well as previously established AAVS1, CCR5 and Rosa26, were carried out by evaluating the differential gene expression of genes located solely in the vicinity of these integration sites, without observing global transcriptomic changes following integration. A more comprehensive bioinformatic-guided and genome-wide search of GSH sites based on established criteria, followed by experimental assessment of transgene expression durability in various cell types and safety assessment using global transcriptome profiling would, thus, lead to the identification of a more reliable and clinically useful genomic region.

In this study, we used bioinformatic screening to rationally identify multiple sites that satisfy established as well as newly introduced GSH criteria. We then used CRISPR/Cas9 targeted genome editing to individually integrate a reporter gene into these sites to monitor long-term

expression of the transgene in HEK293T and Jurkat cells. This experimental evaluation in cell lines was followed by testing of two promising candidate sites in primary human T-cells and human dermal fibroblasts using reporter and therapeutic transgenes, respectively. Finally, bulk and single-cell RNA-sequencing experiments were performed to analyze the transcriptomic effects of such integrations into these two newly established GSH sites.

2.3 Results

2.3.1 Bioinformatic search of novel GSH sites

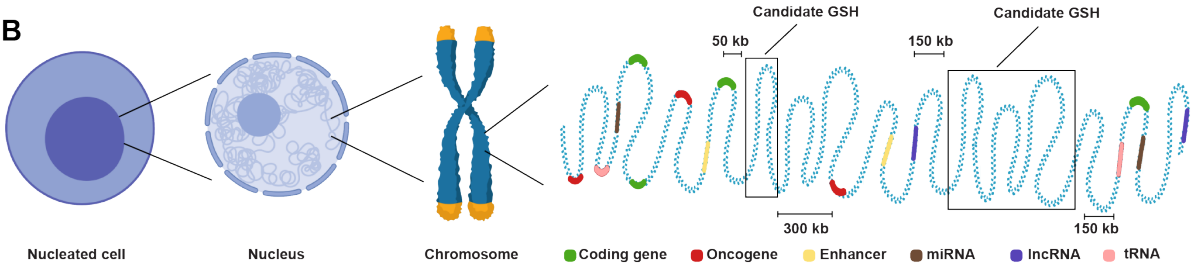
To identify novel sites that could serve as potential GSHs, we first conducted a genome-wide bioinformatic search based on previously established and widely accepted[44] as well as newly introduced criteria that would satisfy safe and stable gene expression (Fig. 2.1A,B). We started by eliminating gene-encoding sequences and their flanking regions of 50 kb to thus avoid disruption of functional regions of gene expression. We then identified oncogenes and eliminated regions of 300 kb upstream and downstream to prevent insertional oncogenesis, a common complication of lentiviral integrations that may arise through unintended upregulation of an oncogene in the vicinity of the integration site[28]. We used oncogenes from both tier 1 (extensive evidence of association with cancer available) and tier 2 (strong indications of the association exist) to decrease the likelihood of oncogene activation upon integration. Additionally, genes can be substantially regulated by mircoRNAs, which cleave and decay mature transcripts as well as inhibit translation machinery, thus modulating protein abundance[80]. We, therefore, excluded miRNA-encoding regions and 300 kb long regions around them. Apart from promoters and microRNAs, gene expression may depend on the presence of enhancers that could be located kilobases away[81,82]. We therefore excluded enhancers as well 20 kb regions around them, which provides an overall distance of up to 70 kb from gene-enhancer units, decreasing the chance of altering physiological gene expression. Additionally, we excluded regions surrounding long non-coding RNAs and tRNAs as well as 150 kb around them as they are involved in differentiation and development programs determining cell fate and are essential for normal protein translation, respectively[83–85]. Finally, we excluded centromeric and telomeric regions to prevent alterations in DNA replication, cellular division and normal aging[86].

Based on our bioinformatic screening, we identified close to two thousand sites that satisfied all of our criteria (Sup. table 1). We chose five sites that varied significantly in size (GSH1, 2, 7, 8, GSH31), designed guide RNAs (gRNA) targeting these sites and possessing high on- and off-target scores (high on-target and low off-target activities), and characterized the durability and safety of transgene expression at these sites experimentally (Fig. 2.1C,D).

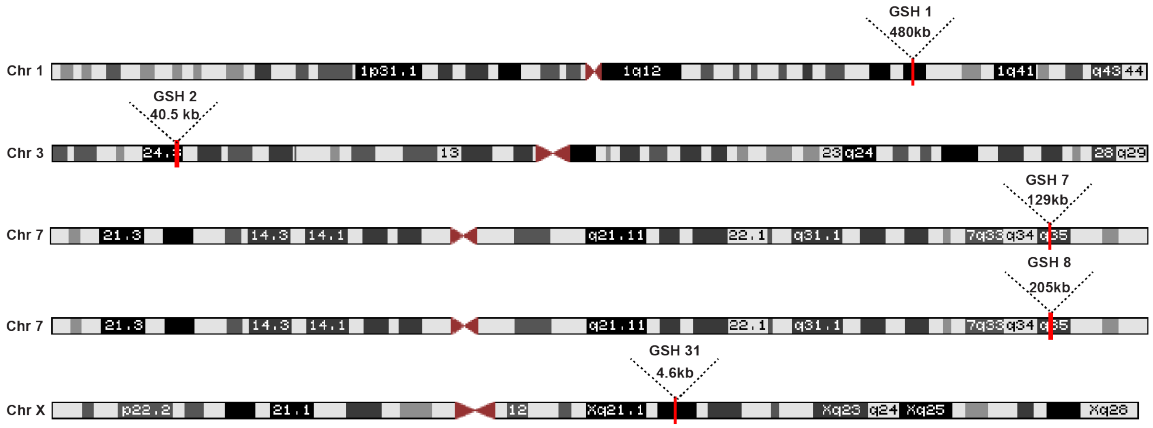
A

GSH criteria	Rationale	Database
50 kb away from known genes	To avoid perturbing endogenous gene expression	GENCODE gene annotation
300 kb away from known oncogenes	To prevent insertional oncogenesis	Cancer Gene Census, GENCODE gene annotation
300 kb away from miRNAs; 150 kb away from lncRNAs, tRNAs	To preserve regulation of gene expresion and cellulat development	MirGeneDB, ENCODE, GENCODE gene annotation
300 kb away from telomeres and centromers	To prevent dysregualtion of cellular division	UCSC Genome Browser GRCh38
20 kb away from known enhancer regions	To prevent interferece with enhancer-gene interactions	EnhancerAtlas 2.0

B



C



D

GSH ID	Chromosome	Coordinates (GRCh38)	gRNA sequence
GSH1	1 (q31.3)	195,338,589-195,818,588	TTAGTCCTAGTGCCATGAAG TGG
GSH2	3 (p24.3)	22,720,711-22,761,389	CATCAGACTTGATAGCACTG AGG
GSH7	7 (q35)	145,090,941-145,219,513	AGGTGCCTCCAATAAAGCAA GGG
GSH8	7 (q35)	145,320,384-145,525,881	TGTGGAACCATGAATCCGAA GGG
GSH31	X (q21.31)	89,174,426-89,179,074	ATAGGCTGTCCATAACCCGG TGG

Figure 2.1. Bioinformatic identification of novel genomic safe harbor sites. A) Table shows GSH criteria, rationale and databases used to computationally predict GSH sites in the human genome. B) Schematic representation of candidate GSH sites, showing linear distances from different encoding and regulatory elements in the genome according to the established and newly introduced criteria. C) Chromosomal locations and lengths of five candidate GSH sites, which were subsequently experimentally tested. D) Chromosomal coordinates of five candidate GSH sites and the gRNA sequences used for subsequent CRISPR/Cas9 genome editing. See also Supplementary table 1 for the list of all computationally predicted sites.

2.3.2 Experimental validation of bioinformatically identified GSH sites by targeted transgene integration in human cell lines

In order to experimentally assess transgene expression from the five predicted novel GSH sites, we performed targeted integration of a gene construct encoding a red fluorescence reporter protein (mRuby) into two common human cell lines – HEK293T and Jurkat cells. HEK293 are commonly used for medium- to large-scale production of recombinant proteins[59], thus identifying GSH in HEK293 may be relevant for protein manufacturing. The Jurkat cell line was derived from T-cells of a pediatric patient with acute lymphoblastic leukemia[54] and has been used extensively used for assessing the functionality of engineered immune receptors, thus discovery of GSH in this cell line supports applications in T cell therapies[87,88]. For integration of mRuby, we employed a CRISPR/Cas9-based genome editing strategy that uses the Precise Integration into Target Chromosome (PITCh) method[89,90], assisted by microhomology-mediated end-joining (MMEJ)[91]. This approach utilizes a reporter-bearing plasmid possessing short microhomology sequences flanked by gRNA binding sites. Once inside the cell the reporter gene together with microhomologies directed against the candidate GSH site are liberated from the plasmid by Cas9-generated double-stranded breaks (DSB) at gRNA binding sites on the PITCh donor plasmid. A different gRNA-Cas9 pair generates DSBs at the candidate GSH locus, and the freed reporter gene with flanking micro-homologies is integrated by exploiting the MMEJ repair pathway (Fig. 2.2A,B). This PITCh MMEJ approach allowed us to rapidly generate donor plasmids targeted against different predicted safe harbor sites, in contrast to the more elaborate process of cloning long homology arms (i.e., >300 bp) required for homology-directed repair (HDR). The error-prone mechanism of MMEJ-mediated integration did not represent a substantial concern since the targeted sites are distanced from any identified coding or regulatory element and thus mutations arising following integration are unlikely to cause any detrimental changes.

Using the PITCh approach, we transfected mRuby transgene into the five candidate GSH sites using the best predicted gRNA sequence for each site (see Methods). We then conducted a pooled selection of mRuby-expressing HEK293T and Jurkat cells by fluorescence-activated cell sorting (FACS), followed by expansion for one week and single-cell sorting to produce monoclonal populations of mRuby-expressing cells. In order to determine sites that support long-term stable transgene expression, we monitored clones with homogenous and high mRuby expression levels by performing flow cytometry at day 30, 45, 60 and 90 after integration.

Out of five candidate GSH sites, four sites in HEK293T cells – GSH1, 2, 7 and 31 (Fig. 2.2C,G) – and two sites in Jurkat cells – GSH1 and 2 (Fig. 2.2D,H) – demonstrated stable mRuby

expression levels 90 days after integration. Interestingly, two sites in HEK293T cells – GSH1, GSH2 – allowed for over an order of magnitude higher transgene expression levels as compared to the commonly used AAVS1 site throughout the 90-day duration of cell culture (Fig. 2.2G). Transgene integration into these sites was confirmed by genotyping using primer pairs amplifying the junction between tested GSH and the transgene (Fig. 2.2E,F).

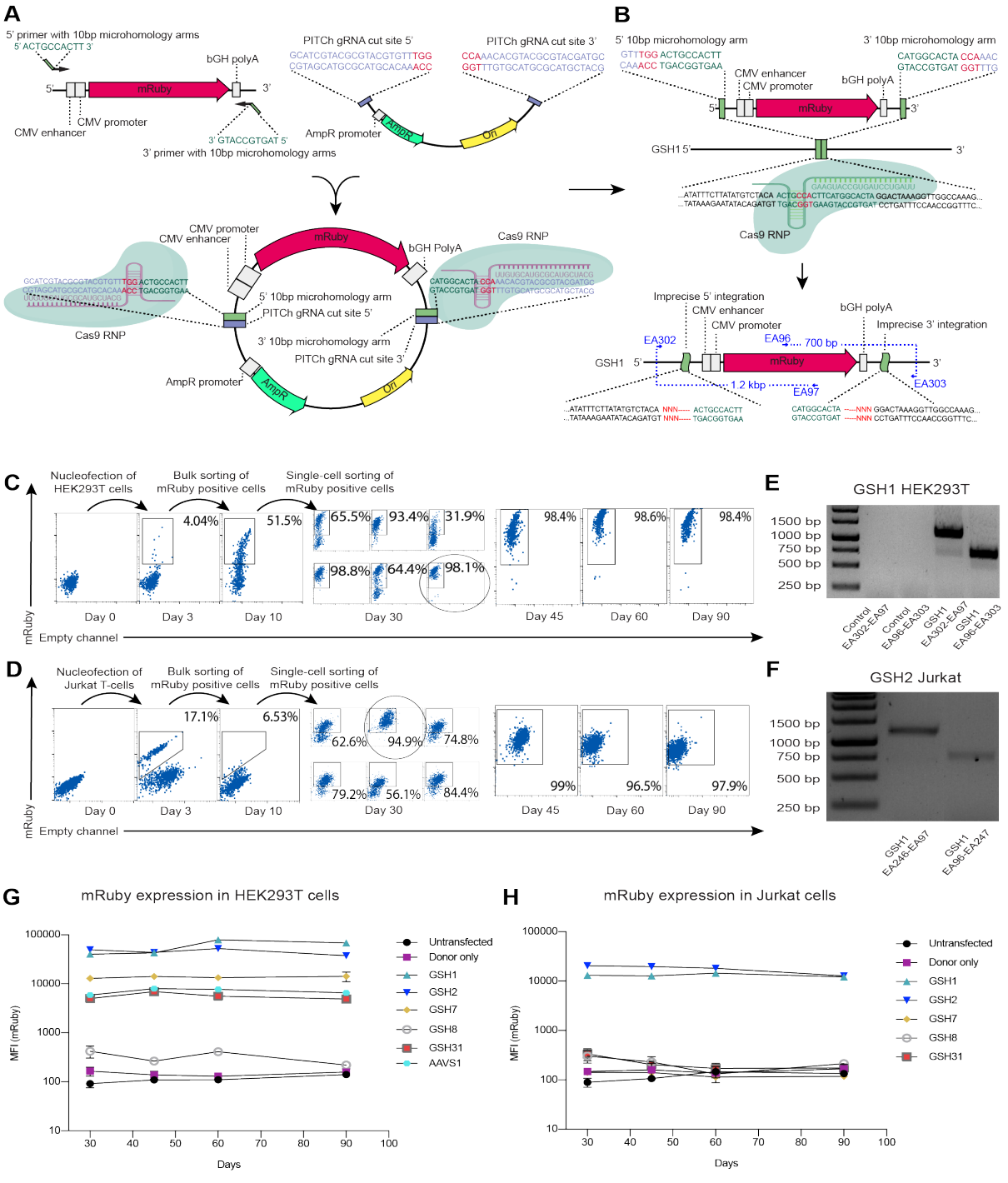


Figure 2.2. Experimental validation of candidate GSH sites by targeted genome editing in HEK293T and Jurkat cells. A) PITCh plasmid is generated by cloning an mRuby-bearing insert with micro-homologies against specific GSH into a backbone possessing PITCh gRNA target sites, needed for the liberation of the insert inside the engineered cell by Cas9. B) Once inside the cell, the mRuby insert is integrated into a desired site by the

MMEJpathway following a Cas9-induced double-stranded break of the targeted site. C, D) Flow cytometry demonstrating the isolation of clonal populations expressing the mRuby transgene from GSH1 locus in HEK293T cells and GSH2 locus in Jurkat cells using pooled and single-cell flow cytometry mediated sortings. The highest expressing GSH1-HEK293T clone and GSH2-Jurkat clone was expanded in cell culture and flow cytometry measurements at day 45, 60 and 90 demonstrated stable levels of transgene expression. E, F) Genotyping of the GSH1 site in HEK293T cells and GSH2 site in Jurkat cells using primers spanning the junction between integration site and the trangene show mRuby integration into the predicted locus. G) mRuby transgene integration into each of the tested GSH sites in HEK293T show stable expression from GSH1, GSH2, GSH7 and GSH 31. Data are represented as mean ± SEM, N=2. H) mRuby transgene integration into each of the tested GSH sites in Jurkat show stable expression from GSH1 and GSH2. Data are represented as mean ± SEM, N=2.

2.3.3 Transcriptome profiling of cell lines following targeted integration in GSH sites

In order to assess whether targeted integration into the candidate GSH sites resulted in aberration of the global transcriptome profiles, we performed a bulk RNA-sequencing and analysis. Following ninety days in culture the clone showing the highest GSH2-integrated mRuby levels was compared with untreated cells from the same culture for both HEK293T and Jurkat cells (Fig. 2.3A). Paired-end sequencing on Ilumina NextSeq500 with an average read length of 100 base-pairs and 30 million reads per sample was employed on two biological replicates of untreated and GSH2-mRuby cultures of HEK293T and Jurkat cells. We first performed a principal component analysis and visualized each sample in two-dimensions using the first two principal components. This immediately revealed transcriptional similarity within the integrated and wild-type samples of the same biological replicate for both cell lines (Fig. 2.3B). While biological variation was observed between the HEK293T samples, the Jurkat samples, both treated and untreated, maintained conserved transcriptional profiles. Performing differential gene expression analysis revealed minor differences between integrated and unintegrated samples for both cell lines relative to the differences between the two cell types (Fig. 2.3C). It was additionally promising that the most differentially expressed genes were not shared between Jurkat and HEK293T cell lines, further suggesting integration in GSH2 does not systematically alter gene expression. Interestingly, differentially expressed genes were scattered across different chromosomes, as opposed to being concentrated within the integrated chromosome where more local contacts exist, again pointing at biological variation (Figure 2.3D). Furthermore, performing gene ontology analysis revealed no significant enrichment of cancer associated genes or pathways in both HEK and Jurkat cells (Fig. S1, S2), again supporting the potential safety of the GSH2 site. We lastly quantified the differences in gene expression for both cell lines either across biological replicates without GSH2 integration versus within a biological replicate with or without GSH2 integration (Fig. 2.3E). Mirroring our principal component analysis (Fig. 2.3B), this analysis again supports that the differences in gene expression we observe arise from biological variation between clones

and not due to integration at GSH2.

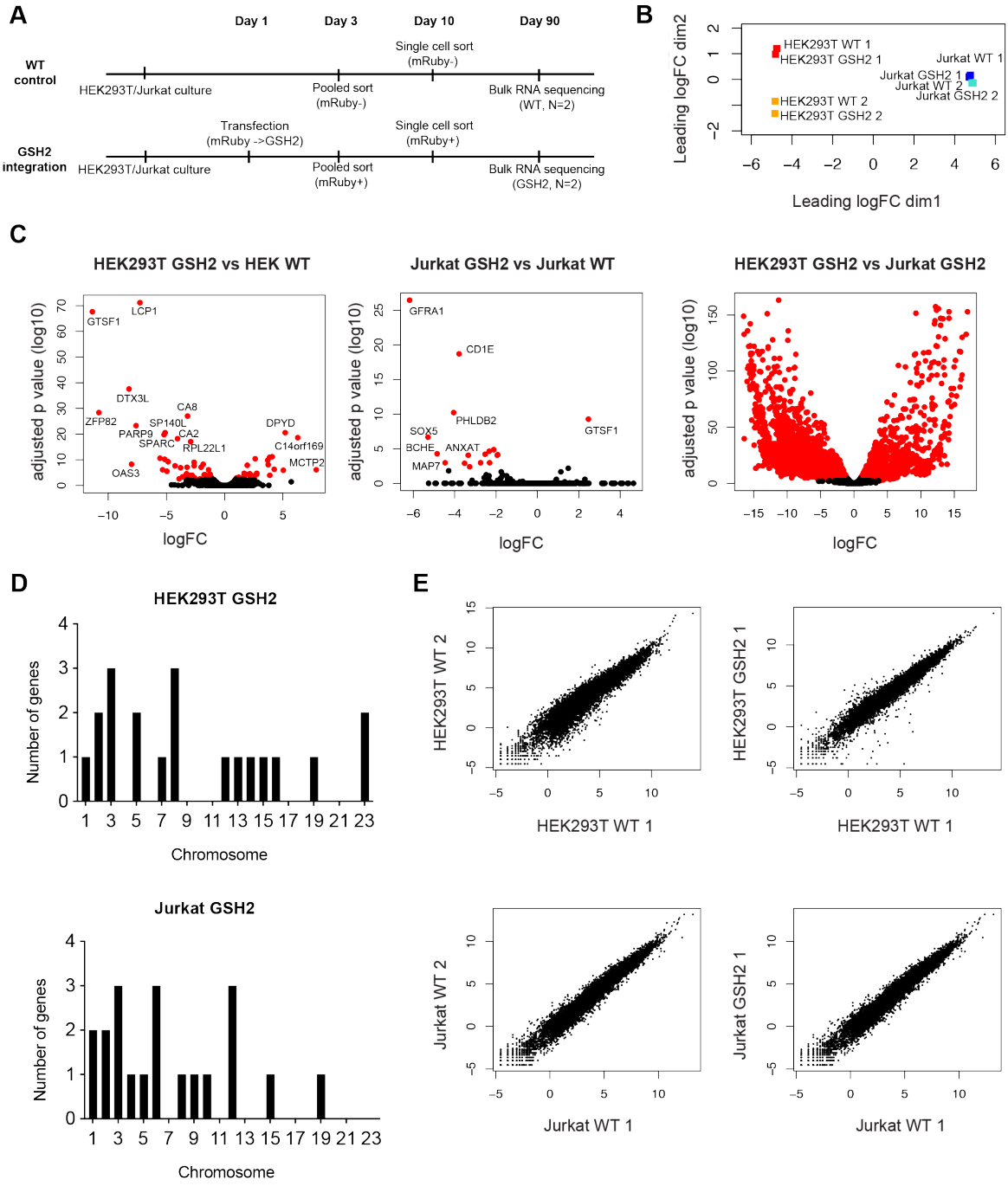


Figure 2.3. RNA sequencing and transcriptome analysis of HEK293T and Jurkat cells following mRuby integration into GSH2. A) Pipeline of bulk RNA-seq experiment on GSH2 integrated and non-integrated HEK293T and Jurkat cells. B) PCA of two biological replicates of HEK293T and Jurkat cells with and without mRuby integration into GSH2. C) Differential expression of genes following GSH2 integration in HEK293T and Jurkat and comparison of HEK293T and Jurkat non-integrated cells. D) Chromosomal distribution of differentially expressed genes in HEK293T and Jurkat cells. Genes with an adjusted p-value of less than 0.05 were considered differentially expressed. E) Correlation of gene expression either between biological replicates without GSH2 integration or within a biological replicate with or without integration in GSH2. See also S1 and S2 for the functional classification

of differentially expressed genes in HEK and Jurkat, respectively.

2.3.4 Targeted integration in novel GSH sites in primary human T-cells and primary human dermal fibroblasts

We next sought to characterize targeted integration into GSH1 and GSH2 sites in primary human cells. One of the potential applications of targeted integration into novel GSH sites is for the ex-vivo engineering of human T-cells, which are being extensively explored for adoptive cell therapies in cancer and autoimmune disease. Thus, we first tested GSH1 and GSH2 in primary human T-cells isolated from peripheral blood of a healthy donor. This time we targeted these sites by employing an HDR-based integration approach using a linear double-stranded DNA donor template, which contained the mRuby transgene driven by a CMV promoter and with 300bp homology arms (Fig. 2.4A). Phosphorothioate bonds and biotin groups were also added to 5' and 3' ends of the HDR template to increase its stability and prevent concatemerization, respectively[92]. Nucleofection of Cas9-gRNA ribonucleoprotein (RNP) complexes and HDR templates into primary T-cells resulted in mRuby-positive expression in 1.3% of cells for GSH1 and 1.24% of cells for GSH2. These mRuby-expressing cells were isolated by FACS on day four, cultured for another seven days; a second round of sorting was performed on the mRuby-positive populations. Following these two rounds of pooled sorting, a highly enriched population of T cells stably expressing the mRuby transgene was isolated and cultured for the duration of T cell ex-vivo culture (up to day 20), with mRuby expression from GSH1 and GSH2 in 94.7% and 91.8% of cells, respectively (Fig. 2.4B). Correct integration into GSH1 and GSH2 was confirmed by genotyping and Sanger-sequencing using primers amplifying the junction between GSH1/GSH2 loci and the mRuby donor (Fig. 2.4C).

Another possible ex-vivo application of identified GSH sites includes engineering dermal fibroblasts and keratinocytes for autologous skin grafting in people with burns or inherited skin disorders. A group of genetic skin disorders named junctional epidermolysis bullosa (JEB) is associated primarily with mutations in a family of multi-subunit laminin proteins, which are involved in anchoring the epidermis layer of the skin to derma[56]. Certain variants of JEB are specifically related to mutations in a beta subunit of laminin-5 protein, encoded by the *LAMB3* gene[93]. Using a similar dsDNA HDR donor with 300bp homology arms possessing phosphorothioate bond and biotin, we used Cas9 HDR to integrate the *LAMB3* gene tagged with GFP (total insert size 5,409 bp) into GSH1 and GSH2 sites in primary human dermal fibroblasts isolated from neonatal skin (Fig. 2.4D). After lipofection of fibroblasts with Cas9 and HDR templates, expression of GFP, which is indicative of LAMB3 expression, was observed in 7.23% (GSH1) and 10.5% (GSH2) of cells. These cells were sorted at day three, cultured for seven days and the GFP-positive population – 3.45% for GSH1 and 1.19% for

GSH2 – was sorted again. Similar to T-cells, two rounds of pooled sorting led to over 92% enrichment of GFP-positive cells, with the expression of *LAMB3*-GFP transgene maintained for the duration of cell culture (over 25 days) (Fig. 2.4E). Genotyping and Sanger-sequencing confirmed successful integration into both loci by using primers amplifying the junction between GSH1/GSH2 and the LAMB3-GFP donor (Fig. 2.4F).

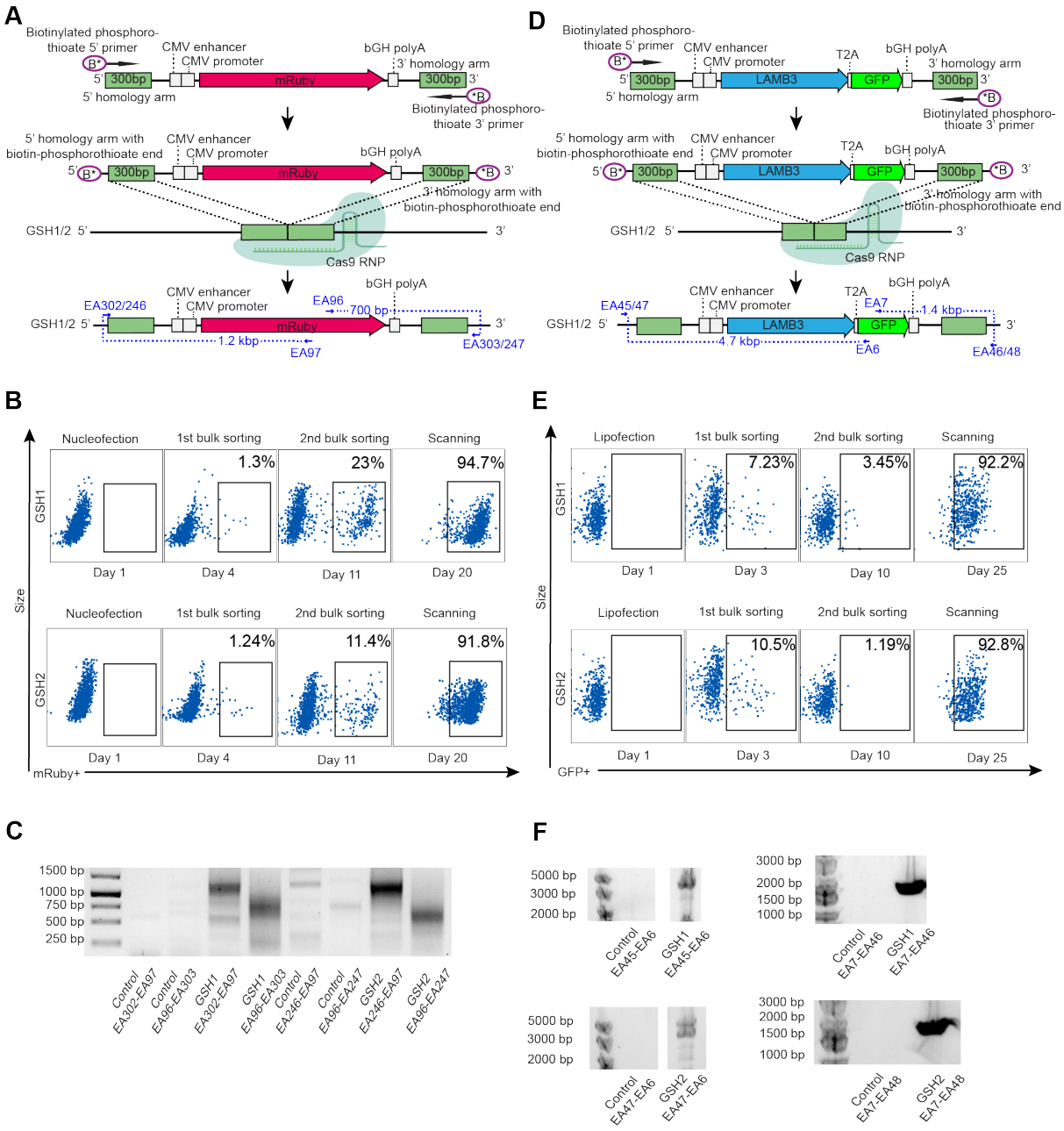


Figure 2.4. Targeted transgene integration into GSH1 and GSH2 in primary human cells. A) Targeted integration of mRuby into GSH1 and GSH2 in primary human T cells by Cas9 HDR. B) Flow cytometry plots demonstrating mRuby expression in both GSH1 and GSH2 in primary human T cells following two rounds of pooled sorting. C) PCR-based genotyping of GSH1 and GSH2 sites by using primers spanning the junction of targeted site and the inserted transgene indicate correct integration of mRuby in primary human T cells. D) Targeted integration of LAMB3-T2A-GFP into GSH1 and GSH2 in primary human dermal fibroblasts by Cas9 HDR. E) Flow cytometry plots demonstrating GFP expression in both GSH1 and GSH2 in primary human dermal fibroblasts

following two rounds of pooled sorting. F) PCR-based genotyping of GSH1 and GSH2 sites by using primers spanning the junction of targeted site and the inserted transgene indicate correct integration of LAMB3-T2A-GFP in primary human dermal fibroblasts. See also Supplementary table 2 for precise sequences of donor constructs.

2.3.5 Single-cell RNA sequencing and analysis of primary human T cells following transgene integration into a novel GSH site

Lastly, we assessed transcriptome-wide effects on a single-cell level following transgene integration into GSH1 in primary T-cells. We performed single-cell RNA sequencing using the 10X Genomics protocol, which consists of encapsulating cells in gel beads bearing reverse transcription (RT) reaction mix with unique cell primers. Following the RT reaction, the cDNA is pooled, and the library is amplified for subsequent next-generation sequencing.

This single-cell sequencing workflow was applied to human T cells expressing mRuby in GSH1 after 25 days in culture, wildtype (non-transfected) cells were used as a control. We also compared these cells with wild-type controls from a different donor to again compare whether GSH integration resulted in more variability in gene expression relative to a biological replicate (Fig. 2.5A). Performing differential gene expression analysis across the three samples revealed fewer up- or downregulated genes following GSH1 integration relative to the untreated, second patient sample (Fig. 2.5B). We performed uniform manifold approximation projection (UMAP) paired with an unbiased clustering based on global gene expression, which resulted in 13 distinct clusters (Fig. 2.5C). Many genes defining these clusters corresponded to typical T cell markers such as IL7R, ICOS, CD28, CCL5, CD74, and NKG7 (Fig. 2.5D). We subsequently quantified the proportion of cells per cluster for each sample, again demonstrating congruent gene expression signatures from cells arising from a single patient, regardless of whether integration in GSH1 occurred or not (Fig. 2.5E). Furthermore, similar to bulk RNA-sequencing results on cell lines, none of the most differentially expressed genes that were upregulated in cells with GSH1 transgene integration were associated with any cancer-related pathways (Fig. 2.5F). Interestingly, the expression of the Jun gene encoding the oncogenic c-Jun transcription factor is decreased in cells bearing transgene integration into GSH1. Taken together, both our single-cell and bulk RNA-sequencing data suggest that the computationally determined and experimentally validated GSHs have minimal influences on global gene expression.

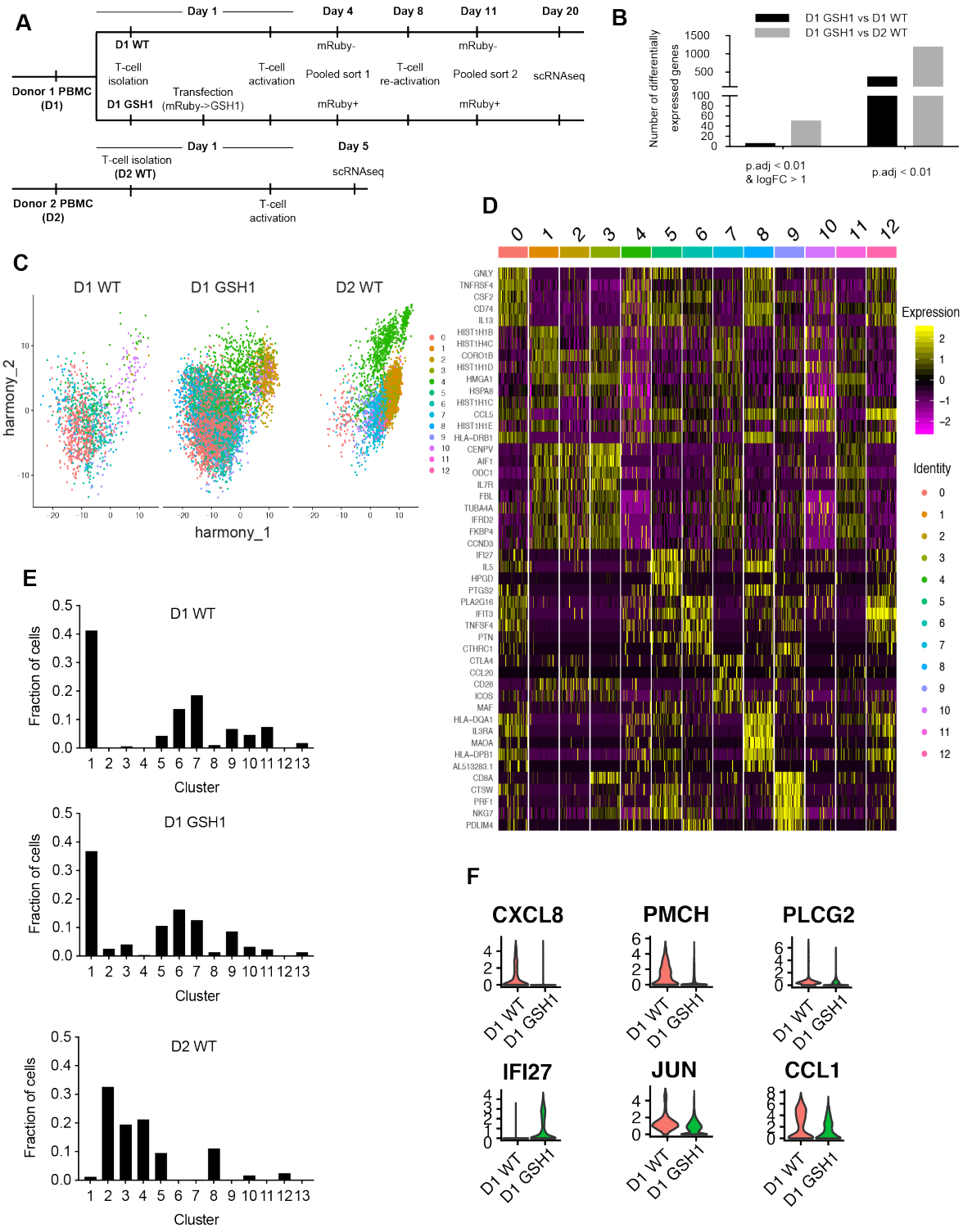


Figure 2.5. Single-cell RNA-seq of primary human T-cells following targeted transgene integration into GSH1 site. A) Pipeline of the RNA-seq experiment following Cas9 HDR targeted integration of mRuby into GSH1 (GSH1-mRuby cells) and T-cell activation. B) Number of differentially expressed genes GSH1-mRuby T-cells and WT T-cells (non-integrated) from donor 1 and GSH1-mRuby T-cells from donor 1 and WT T-cells from donor 2. C) UMAP analysis comparing transcriptional clusters of GSH1-mRuby and WT T-cells from donor 1 and WT T-cells from donor 2. Each point represents a unique cell barcode, and each color corresponds to cluster identity. D) Expression of genes determining the seven largest clusters. Intensity corresponds to normalized gene expression. E) Distribution of GSH1-mRuby-and WT T-cells from donor 1 and WT T-cells from donor 2 across different clusters.

F) Normalized expression for selected differentially expressed genes between GSH1mRuby and WT T-cells from donor 1.

2.4 Discussion

In this study we used bioinformatic screening to identify novel GSH sites and performed phenotypic validation by targeted transgene integration in human cell lines and primary cells, resulting in durable and stable transgene expression. The potential safety of GSH sites was confirmed by observing minimal changes in transcriptomic profiles following transgene integration. None of the upregulated transcripts following transgene integration were associated with any of the known cancer pathways. These findings make the newly identified sites potentially preferable to currently used AAVS1, CCR5 and hRosa26, which have the drawbacks of being located within functional genes, in gene-dense regions and surrounded by oncogenes[44]. Although previous studies have also resulted in the discovery of sites capable of long-term expression of transgenes, they were limited by the integration mechanism researchers employed and changes to the entire transcriptome following integration events were not evaluated, as they were focused on differential expression of a handful of genes in the vicinity of the discovered site[49]. Finally, generalizability of the criteria used to establish our new GSH sites suggests their possible applicability to different cell types, expanding the genome engineering toolkit for diverse cell therapy and synthetic biology applications[94].

The most immediate use of identified GSH sites may involve safe and predictable engineering of human T-cells for adoptive cell therapy applications[95]. Copious endeavors to design, modify and augment functions of T-cells ex-vivo have been successfully initiated in research labs[7,96]. However, most strategies have relied on viral-mediated delivery, which results in random transgene integration and is thus associated with the risk of insertional oncogenesis, potentially leading to cancerous transformations of engineered cells, and unpredictability of transgene expression levels associated with the nature of the integration locus and frequent silencing of the integrated construct. Performing targeted integration into GSH sites would enable long-term transgene expression in a safe manner and would support advanced efforts in engineered T cell therapies such as armored CAR-T cells, capable of overcoming hostile tumor microenviroments[97] as well as T cells bearing synthetic receptors that introduce logic gates into cell's behavior, allowing for safer and more effective treatments[53]. Additionally, given the demonstrated efficiency in dermal fibroblasts, we envision a rapid application of the discovered sites to skin engineering, particularly in the context of treatment of the inherited skin disorders, wound healing as well as skin rejuvenation.

Another exciting aspect of the identified GSH sites is the level of transgene expression

observed, especially in HEK293T cells, which are known to be suitable for large-scale production of therapeutic proteins. We observed high levels of reporter gene expression from GSH1 and GSH2 in HEK293T that were sustained for over three months and exceeded expression levels from the AAVS1 site. This high expression level can theoretically be enhanced further by multiple biallelic integration events into identified loci and thus be exploited for durable large-scale production of commercially valuable proteins. Direct long-term comparison of the transgene expression from the identified sites and the expression following lentiviral integration will be needed to confirm the advantage of GSH expression for protein manufacturing.

In summary, two novel human genomic safe harbor sites identified and validated in this study may serve as a robust and safe platform for a variety of clinically and industrially relevant cell engineering approaches, culminating in safer and more reliable gene and cell therapies.

2.5 Methods

Computational search for GSH sites

Previously established criteria[44] as well as newly introduced ones were used to predict genomic locations of novel GSHs. Specifically, coordinates of all known genes were extracted from GENCODE gene annotation (Release 24). A set of tier 1 and tier 2 oncogenes was obtained from Cancer Gene Census. The miRNA coordinates were obtained from MirGeneDB[98]. Enhancer regions were obtained from the EnhancerAtlas 2.0 database[99], coordinates were transposed into GRCh38/hg38 genome and union of enhancer sites was used. Genomic locations of sequences of tRNA and lncRNA were extracted from GENCODE gene annotation (Release 24). UCSC genome browser GRCh38/hg38 was used to get coordinates of telomeres and centromeres as well as unannotated regions. BEDTools[100] were used to determine flanking regions of each element of the criteria as well as to obtain union or difference between sets of coordinates. The custom source code developed for the computational identification of novel human genomic safe harbor sites is available at https://github.com/elvirakinzina/GSH.

Plasmids, guide RNA design and HDR donor generation

PITCh plasmids were generated through standard cloning methods. CMV-mRuby-bGH insert was amplified from pcDNA3-mRuby2 plasmid (Addgene, Plasmid #40260) with primers containing mircohomology sequences against specific GSHs and AAVS1 site with 10bp of overlapping ends for the pcDNA3 backbone. The pcDNA3 backbone was amplified with

primers containing sequences of PITCh gRNA cut site (GCATCGTACGCGTACGTGTTTGG) on both 5' and 3' ends of the backbone. The insert and the backbone were assembled using Gibson Assembly Master Mix (New England Biolabs, #E2611L).

Guide RNA sequences for five tested GSH sites were predicted using Geneious gRNA design tool. Briefly, coordinates of the predicted GSH sites were pasted into UCSC Genome Browser GRCh38/hg38 and DNA sequences were extracted and transferred into Geneious. An internal gRNA design tool was used to identify gRNA sequences located in the predicted GSHs against the entire human genome. The evaluation of the efficacy of double-stranded break generation (on-target activity) was based on Doench et al., 2016[101], while the specificity of the gRNA-induced break (off-target activity) was assessed based on Hsu et al., 2013[102]. Guide RNAs with high on-target and off-target scores were used to target predicted GSHs.

Plasmids encoding CMV-mRuby-bGH flanked by GSH1/GSH2 300bp homology arms were ordered from Twist Biosciences in pENTR vector. HDR donors were amplified from these plasmids using biotinylated primers with phosphorothioate bonds between the first 5 nucleotides on both 5' and 3' ends. Plasmid encoding CMV-LAMB3-T2A-GFP-bGH was generated by overlap extension PCR of LAMB3 cDNA, purchased from Genscript (NM_000228.3) and GFP-bGH sequence from Addgene (Plasmid #11154). T2A sequence was added to 5'primer of GFP-bGH. Produced insert was cloned into the abovementioned pENTR vector from Twist Biosciences bearing GSH1 and GSH2 300bp homology arms as well as CMV promoter sequence using Gibson Assembly Master Mix (NEB, #E2611L). HDR donors were amplified from these plasmids using biotinylated primers with phosphorothioate bonds between the first 5 nucleotides on both 5' and 3' ends. HDR donors were then purified from PCR mix using SPRI beads (Beckman Coulter, #B23318) at 0.4X beads to PCR mix ratio.

HEK293T and Jurkat cell culture, transfection and sorting

HEK293T cells were obtained from the American Type Culture Collection (ATCC) (#CRL-3216); the Jurkat leukemia E6-1 T cell line was obtained from ATCC (#TIB152). HEK cells were cultured in Dulbecco's Modified Eagle's Medium (DMEM) (ATCC 30-2002) supplemented with 2mM L-glutamine (ATCC 30-2214). Jurkat cells were cultured in ATCC-modified RPMI-1640 (Thermo Fisher, #A1049101). All media were supplemented with 10% FBS, 50 U ml-1penicillin and 50 µg ml^{-1} streptomycin. Detachment of HEK cells for passaging was performed using the TrypLE reagent (Thermo Fisher, #12605010). All cell lines were cultured at 37°C, 5% CO2 in a humidified atmosphere.

Prior to transfection of HEK293T and Jurkat gRNA molecules were assembled by mixing 4 µl

of custom Alt-R crRNA (200 µM, IDT) with 4 µL of Alt-R tracrRNA (200 µM, IDT, #1072534), incubating the mix at 95°C for 5 min and cooling it to room temperature. 2 µL of assembled gRNA molecules were mixed with 2 µL of recombinant SpCas9 (61 µM, IDT, #1081059) and incubated for > 10 min at room temperature to generate Cas9 RNP complexes.

For transfection of HEK cells 100 µL format SF Cell line kit (Lonza, V4XC-2012) and electroporation program CM-130 was used on the 4D-Nucleofector. $1x10^6$ HEK cells were transfected with 2 µg of PITCh donor, 2 µl of Cas9 RNP complex against specific GSH and 2 µl of Cas9 RNP complex against PITCh plasmid to liberate MMEJ insert.

For transfection of Jurkat cells 100 µL format SE Cell line kit (Lonza, V4XC-1012) and electroporation program CL-120 was used on the 4D-Nucleofector. $1x10^6$ Jurkat cells were transfected with 2 µg of PITCh donor, 2 µl of Cas9 RNP complex against specific GSH and 2 µl of Cas9 RNP complex against PITCh plasmid to liberate MMEJ insert.

Transfected HEK and Jurkat cells were bulk sorted on day 3 and single-cell sorted on day 10 following transfection using Sony SH800S sorter. Best expressing clone was selected on day 30, split into two wells and cultured for another 2 months. mRuby expression of the best expressing clone was analyzed on BD LSRFortessa Flow Cytometer on day 45, 60 and 90 following transfection.

Human T-cells culture, transfection and sorting

Human peripheral blood mononuclear cells were purchased from Stemcell Technologies (#70025) and T cells isolated using the EasySep Human T Cell Isolation kit (Stemcell Technologies, #17951). Primary human T cells were cultured for up to 20 days in ATCC-modified RPMI (Thermo Fisher, #A1049101) supplemented with 10% FBS, 10 mM non-essential amino acids, 50 µM 2-mercaptoethanol, 50 U ml-1penicillin, 50 µg ml^{-6} streptomycin and freshly added 20 ng ml^{-1} recombinant human IL-2, (Peprotech, #200-02). T cells were cultured at 37°C, 5% CO2 in a humidified atmosphere. On day 1 of culture, transfection of primary T cells with Cas9 RNP complexes and GSH1/GSH2-mRuby HDR templates was performed using the 4D-Nucleofector and a 20 uL format P3 Primary Cell kit (Lonza, V4XP-3032). Briefly, gRNA molecules were assembled by mixing 4 µl of custom Alt-R crRNA (200 µM, IDT) with 4 µL of Alt-R tracrRNA (200 µM, IDT, #1072534), incubating the mix at 95°C for 5 min and cooling it to room temperature. 2 µL of assembled gRNA molecules were mixed with 2 µL of recombinant SpCas9 (61 µM, IDT, #1081059) and incubated for > 10 min at room temperature to generate Cas9 RNP complexes. $1x10^6$ primary T cells were transfected with 1 µg of HDR template, 1 µl of GHS1/GSH2 Cas9 RNP complex using the EO115 electroporation program. T cells were activated with Dynabeads™ Human T-Activator CD3/CD28 (Thermo

Fischer, #11161D) 3-4 hours following transfection. mRuby-positive T-cells were bulk sorted on day 4 using Sony SH800S sorter, re-activated with the new beads on day 8, sorted again on day 11 and analyzed on BD LSRFortessa Flow Cytometer on day 20.

Human dermal fibroblasts culture, transfection and sorting

Neonatal human dermal fibroblasts were purchased from Coriell Institute (Catalog ID GM03377). Primary fibroblasts were cultured for up to 25 days in Prime Fibroblast media (CELLNTEC, CnT-PR-F). Cells were passaged at 70% confluency using Accutase (CELLNTEC, CnT-Accutase-100). Detached cells were centrifuged for 5 min, 200 x g at room temperature and seeded at seeded at 2,000 cells per cm^2. Fibroblasts were cultured at 37°C, 5% CO2 in a humidified atmosphere. Fibroblasts were transfected using Lipofectamine™ CRISPRMAX™ Cas9 Transfection Reagent (ThermoFisher Scientific, CMAX00001). Briefly, cells were transfected at 50% confluency with 1:1 ratio of custom sgRNA (40 pmoles, Synthego) and SpCas9 (40pmoles, Synthego) and 2.5 µg of GSH1/GSH2 LAMB3-T2A-GFP HDR template. GFP-positive fibroblasts were bulk sorted on day 3 and 10 using Sony SH800S sorter and analyzed on BD LSRFortessa Flow Cytometer on day 25.

Genotypic analysis of GSH integration

Genomic DNA was extracted from $1x10^6$ cells using PureLink Genomic DNA extraction kit (ThermoFischer Scientific, #K1820-01). 5 µL of genomic DNA extract were then used as templates for 25 µL PCR reactions using a primer with one primer residing outside of the homology arm of the integrated sequence and the other primer inside the integrated sequence. Obtained bands were gel extracted using Zymoclean Gel DNA Recovery Kit (Zymo Research, #D4001), 4ul of eluted DNA was cloned into a TOPO-vector using Zero-blunt TOPO PCR Cloning Kit (ThermoFischer Scientific, #450245), incubated for 1 hour, transformed into NEB 5-alpha Competent E. coli cells (New England Biolabs, C2987H) and plated on agar plates containing kanamycin at 50 µg/ml. Produced clones were picked and inoculated for overnight culture in 5ml of liquid broth supplemented with kanamycin at 50 µg/ml. Liquid cultures were mini-prepped the following morning using ZR Plasmid Miniprep - Classic kit (Zymo Research, #D4015) and Sanger sequenced by Microsynth using M13-forward and M13-reverse standard primers.

Bulk RNA-sequencing of HEK293T and Jurkat cells GSH2 and WT

Following single-cell sort, the best expressing clone (GSH2) and wild-type (WT) of HEK293T

and Jurkat cells were split into 2 wells (1 and 2) and cultured for 80 days, after which total RNA was extracted using PureLink RNA Mini Kit (ThermoFischer Scientific, #12183018A). Extracted total RNA was depleted of rRNA using RiboCop rRNA Depletion Kit (Lexogen, #144), first and second strands of cDNA were generated with SuperScript Double-Stranded cDNA Synthesis Kit (ThermoFischer Scientific, #11917010) using random hexamers and flow cell adapters were ligated to the produced double-stranded cDNA. DNA fragments were enriched by PCR using Q5 High-Fidelity 2X Master Mix (New England Biolabs, #M0492S) and sequenced by the Illumina NextSeq 500 system in the Genomics Facility Basel. Sequencing reads were aligned to the human reference genome (GRCh38) using Subread (v1.6.2) using unique mapping[103]. Expression levels were quantified using the featureCounts function in the Rpackage Rsubread at gene-level[104]. Normalization across the samples was performed using default parameters in the Rpackage edgeR[105]. Differential expression analysis was performed using the exactTest function in the edgeR package. Gene ontology was performed by supplying those differentially expressed genes (adjusted p value < 0.05) to the goana function[106].

Single-cell RNA sequencing of human T-cells

Single-cell RNA sequencing was conducted on day 20 of culture for Donor 1 WT (D1 WT) and Donor 1 GSH1 (D1 GSH1) and on day 5 for Donor 2 WT (D2 WT). Single cell 10X libraries were constructed from the isolated single cells following the Chromium Single Cell 3' GEM, Library & Gel Bead Kit v3 (10X Genomics, PN-1000075). Briefly, single cells were co-encapsulated with gel beads (10X Genomics, 2000059) in droplets using Chromium Single Cell B Chip (10X Genomics, 1000074). Final D1 WT, D1 GSH1 and D2 WT libraries were pooled and sequenced on the Illumina NovaSeq platform (26/8/0/93 cycles). Raw sequencing files were supplied to cellranger (v3.1.0) using the count argument under default parameters and the human reference genome (GRCh38-3.0.0). Filtering, normalization and transcriptome analysis was performed using a previously described pipeline in the R package Platypus[107]. Briefly, filtered gene expression matrices from cellranger were supplied as input into the Read10x function in the R package Seurat[108]. Cells containing more than 5% mitochondrial genes, or less than 150 unique genes detected were filtered out before using the RunPCA function and subsequent normalization using the function RunHarmony from the Harmony package under default parameters[109]. Uniform manifold approximation projection was performed with Seurat's RunUMAP function using the first 20 dimensions and the previously computed Harmony reduction. Clustering was performed by the Seurat functions FindNeighbors and FindClusters using the Harmony reduction and first 20 principal components and the default cluster resolution of 0.5, respectively[108]. Cluster-specific genes

were determined by Seurat's FindMarkers function for those genes expressed in at least 25% of cells in one of the two groups. Differential genes between samples were calculated using the FindMarkers function from Seurat using the default Wilcoxon Rank Sum Test with Bonferroni multiple hypothesis correction. The source code for the analysis of scRNA-seq data is available at https://github.com/alexyermanos/Platypus.

3. High throughput screening of novel genomic safe harbors sites using libraries of inducible endogenous guide RNAs

3.1 Summary

Use of gRNA libraries to simultaneously interrogate multiple genomic regions has found its numerous applications in gene knock out studies as well as in the investigation of the non-coding segments of the genome. We attempted to exploit endogenous expression of gRNA libraries from a specified genomic locus to generate a double-strand break in various positions across the human genome in a high-throughput manner. Induced DNA breaks would be followed by reporter transgene knock in via NHEJ repair pathway, thus allowing to assess the ability of targeted sites to support transgene expression. Cells positive for the reporter gene would be isolated and the gRNA locus would be sequenced to derive the integration site of the reporter gene, and hence identify potential safe harbor. This approach would thus link a fluorescent phenotypic readout with specific genomically encoded gRNA sequence, allowing for rapid multiplexed GSH identification.

3.2 Introduction

3.2.1 Genomic safe harbors

Targeted gene addition mandates safe and efficient expression of gene of interest from specific loci in the genome, referred to as Genomic Safe Harbor (GSH) sites. In the previous chapter we discussed how researchers have empirically identified three genomic sites capable of stable expression of introduced genes in a variety of cellular contexts. All three of these sites are located in the introns of protein-coding genes, functions of which are not fully investigated yet. These sites are also surrounded by oncogenes located in linear proximity to insertion locus. Thus, despite providing durable expression, integration into currently used genomic sites may lead to the disruption of functional genes as well as to insertional oncogenesis, preventing the use of these sites beyond research setting[44]. In Chapter 2 we also described a rational validation approach for a handful of computationally predicted GSHs. In contrast to it, a large, multiplexed screening of all bioinformatically identified sites scattered across different positions in the human genome would allow to rapidly validate novel GSH loci for their capability to express inserted transgene. Successfully validated target sites can then undergo further characterization to verify safety and durability of transgene expression, similar to the one described in Chapter 2.

3.2.2 Criteria for novel genomic safe harbor sites

In order to support safe expression of the genes of interest in a range of cell types, an ideal GSH should satisfy a set of criteria. The aim of these criteria is to spatially separate the genomically introduced promoter and the transgene sequences from endogenous coding and regulatory elements of the genome in a cell type agnostic manner. The first genetic element to be avoided are annotated protein-coding sequences. There are around 20,000 currently identified protein-coding genes, transgene integration in the vicinity of which may lead to aberrant and misregulated endogenous gene expression[110]. Thus, newly identified GSH sites need to exclude any gene-encoding sequences as well as regions surrounding them. Annotated oncogenes and the regions around them also need to be avoided to minimize the risk of upregulation of oncogene expression following transgene insertion, a side-effect associated with lentiviral gene delivery method[27]. MicroRNAs, lncRNA, tRNAs and other non-protein encoding RNA molecules are responsible for a wide range of regulatory and developmental roles in the cell[84,111]. Changes in their expression may lead to perturbations in cellular function and even cell death[112]. For instance, *Xist* lncRNA is responsible for X chromosome inactivation in female cells by initiating the recruitment of the X chromosome to the periphery of the nucleus, where it is anchored to the nuclear lamina and deacetylated. In male cells *Xist* is silenced allowing for transcription of the genes from the only X chromosome[83]. However, if transgene addition results in an unintended upregulation of *Xist* expression in male cells, the inactivation of the entire X chromosome will be initiated, silencing all of its single copied genes, leading to cell death[112]. To minimize the possibility of such events, all regulatory RNAs should be avoided, thus maintaining physiological regulatory machinery of the cells. Additionally, enhancer elements are essential for activating gene expression through the recruitment of transcription factors. These genomic elements can be located in close proximity to genes as well as tens and hundreds kilobases away and need to be avoided to prevent misregulation of gene transcription[81]. Finally, structural components of chromosomes – centromeres and telomeres – play critical role in successful cellular division. Centromeres serve as attachment points for kinetochores and are required for normal chromosomal segregation during mitosis. Abnormal separation of sister chromosomes and unequal chromosomal distribution following cellular division can be associated with cancer development[86]. Telomeres are six nucleotide repeats spanning the ends of the chromosomes and protecting the coding portions of chromosomal ends from progressive degradation due to “end replication problem”. Interfering with physiological regulation of telomere length may lead to premature cellular senescence[113]. Thus, both telomeric and centromeric regions on the chromosomes need to be excluded from an ideal GSH site.

3.2.3 Inducible endogenous guide RNA expression

In addition to efficient functioning when provided as plasmid encoded sequences or as exogenous in vitro generated molecules, recent research has demonstrated that guide RNA encoding sequences can be integrated into the host genome, endogenously expressed from an accessible genomic locus and capable of complexing with Cas9 enzyme to elicit a double-strand break (Fig. 3.1A)[114]. Additionally, scientists have demonstrated the ability to introduce a library of protospacers targeted at different genomic sites into a locus containing U6 promoter, which is required for the RNA polymerase III dependent short RNA transcription, and sgRNA hairpin sequence, which once transcribed is essential for the Cas9 complexation (Fig. 3.1B)[114]. Furthermore, the ability to control short non-coding RNA expression through a Tet-inducible U6 promoter system has also been described. Specifically, tet repressor (tetR) is capable of binding to tet operator (TetO) sequences surrounding the TATA-box of RNA polymerase III promoters, repressing the expression of short RNA. This repression can then be alleviated by the addition of tetracycline or doxycycline drugs, which bind to tetR and allow the polymerase to progress with the transcription (Fig. 3.1C)[115,116]. These studies collectively suggest the opportunity to create an inducible library-based gRNA expression platform transcribed from an accessible genomic locus to interrogate numerous genomic sites in a multiplexed format (Fig. 3.1D).

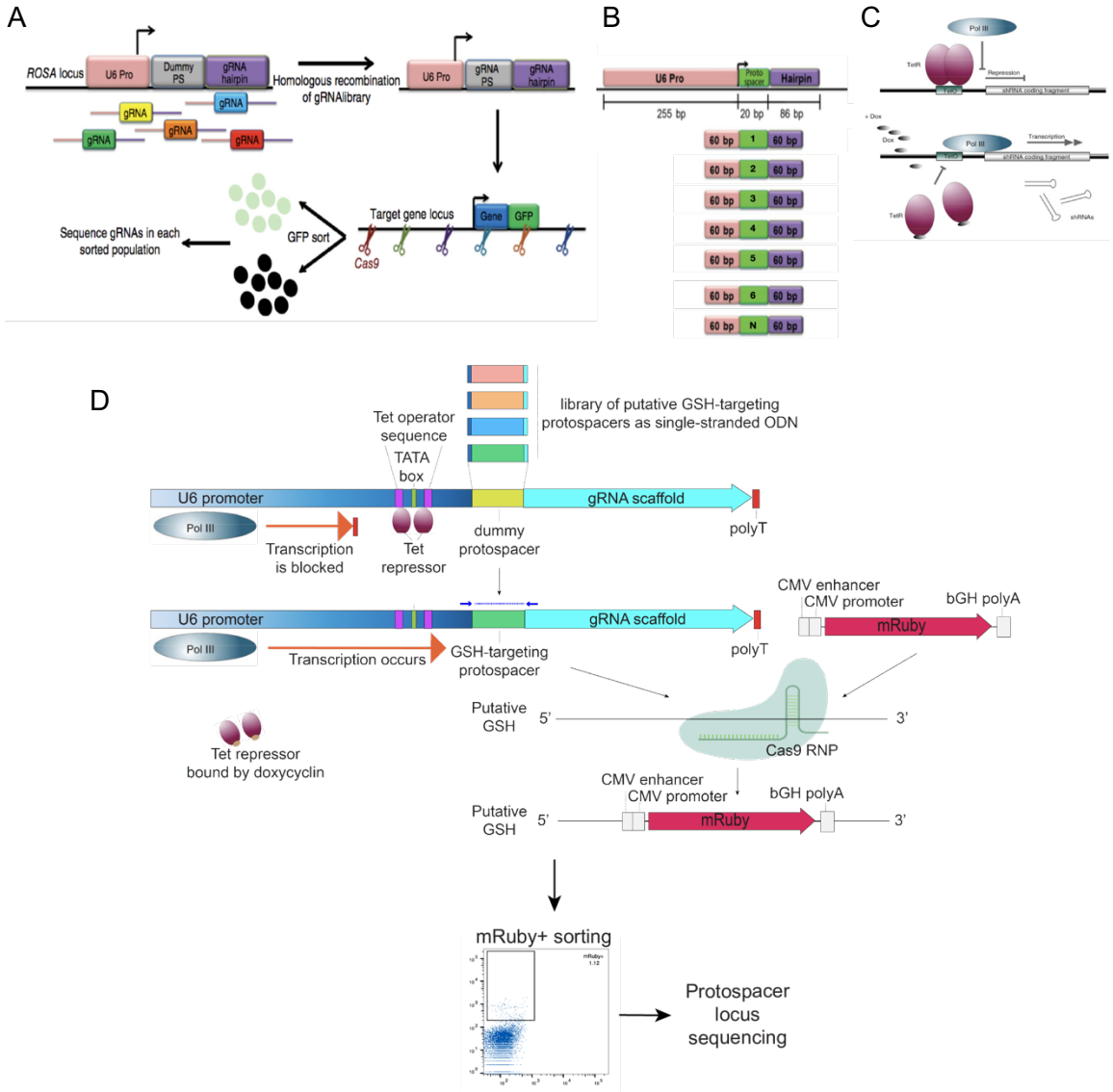


Figure 3.1. Generation endogenous inducible sgRNA expression using a library of protospacer sequences. A) Rajagopal et al. have demonstrated the possibility to use mouse ROSA26 locus to support endogenous sgRNA expression. B) They have also showed that a library of protospacers can be incorporated into this locus, allowing to target different regions of the cell's genome in a multiplexed format. Adapter from Rajagopal et al.[114] C) Kappel et al.[115] and Henriksen et al.[116] have described an inducible expression of short hairpin RNA (shRNA) using a TetR system, in which TetO surround a TATA box of RNA polymerase III promoter, allowing the repressor to prevent sgRNA transcription. Once doxycycline is added and interacts with the repressor, the latter disassociates from the promoter and allows for the progression of sgRNA transcription. Adapted from Kappel et al.[115] D) In our design, we envisioned to use genomically integrated inducible sgRNA expression platform complemented with a library of protospacer sequences, which after doxycycline induction should result in a functional sgRNA transcription. This temporally controlled sgRNA expression will be accompanied by transfection of NHEJ mRuby donor, which will be expressed in case of integration into a true GSH. Cells durably expressing mRuby will be sorted and protospacer locus will be sequenced to identify mRuby integration site.

3.2.4 NHEJ-based transgene knock-in

One of the approaches to introduce a desired transgene into specific genomic locus following a CRISPR-induced double-strand break is via a non-homologous end joining (NHEJ) repair pathway[41]. As opposed to HDR the NHEJ pathway doesn't rely on homologous recombination and thus the donor DNA doesn't need to be pre-designed to contain homology arms determined by the integration site. Additionally, the mechanism of NHEJ pathway allows it to occur in the absence of cellular division. NHEJ repair is initiated by the recognition of the double-strand break via a Ku protein complex, which then recruits enzymes necessary for DNA end processing and ligation. Due to its untemplated nature, NHEJ repair often results in small insertion and/or deletions at the site of the DNA break[117–119]. Researchers managed to devise an approach that supports a robust and efficient transgene integration via this pathway – homology-independent targeted integration (HITI) – which relies on CRISPR-induced cutting of a circular donor inside the transfected cell followed by the genomic integration of the donor into the target region determined by a separate gRNA-Cas9 pair[120].

3.3 Project rationale

Identification of novel human genomic sites capable of durable and safe expression of genes of interest will significantly improve existing gene and cell therapies, reducing the risk associated with them. To augment computational search discussed in previous chapter and to rapidly validate large portions of the human genome for such suitable sites, a high-throughput method involving interrogation of hundreds of sites in a multiplexed manner needs to be devised. This chapter will describe an attempt to use CRISPR-based NHEJ-facilitated knock in of a reporter gene into various genomic locations guided by library of inducible endogenously expressed guide RNAs in human immortalized leukemic T-cell line. Jurkat cell were chosen due to an envisioned use of identified GSHs in T-cell engineering.

Firstly, we aimed to establish a cellular platform that would allow for such multiplexed GSH evaluation approach. This platform would consist of constitutively expressed Cas9 nuclease integrated into CCR5 locus. Cas9 encoding donor will also contain a GFP reporter gene, allowing to isolate cells that have undergone successful transgene integration. This initial genome engineering step will be followed by AAVS1-targeted insertion of a landing pad bearing Tet-inducible RNA polymerase III promoter, a dummy protospacer and a gRNA hairpin sequence – all required for an inducible guide RNA expression. TetR and mRuby will also be introduced into the same genomic location to keep sgRNA expression initially repressed and allow for isolation of successfully engineered cells, respectively. A library of protospacer sequences predicted by our computational search and targeted against potential GSH loci will

be transfected in the form of single-stranded oligodeoxynucleotides into AAVS1 site of these engineered Jurkat cells, replacing the dummy protospacer. Finally, once the library is introduced, cells will be transfected with a DNA donor encoding a reporter gene and supplemented with doxycycline, allowing for an inducible expression of a full gRNA. This in presence of a constitutive Cas9 enzyme results in a targeted double-strand DNA break in a genomic location determined by a specific protospacer library member expressed in a given cell, allowing for a reporter gene on the DNA donor to be integrated into the protospacer-determined site. If this reporter is integrated into a region that demonstrates high expression levels, the cell bearing it will be sorted using flow cytometry, resulting in a pool of cells possessing reporter sequences in potential GSHs. The AAVS1 site of these pooled cells will be genotyped using next-generation sequencing to determine their protospacer library members and, hence the sites, where the integration of the reporter gene occurred. These newly discovered sites will be additionally genotyped to verify the match between the protospacer in the inducible landing pad in AAVS1 locus and the actual integration site. Finally, the most robustly expressing sites will be evaluated separately for their safety and long-term stability of transgene expression.

3.4 Results

3.4.1 Generation of constitutive Cas9 expressing Jurkat T cells

In the first genome engineering step we generated a donor construct containing Cas9 enzyme expressed by a strong constitutive CMV promoter, followed by a GFP gene expressed by a separate promoter. This transgenic construct was flanked by homology arms sequences of 1000 bp against CCR5 locus to facilitate HDR-driven knock in following the double-strand break (Fig. 3.2A). This donor was transfected in the form of circular plasmid into Jurkat T cells together with CRISPR/Cas9 ribonucleoprotein complex (RNP) targeting CCR5 site. Transfected cells were pool-sorted using flow cytometry based on GFP expression, hence transgene integration, and then single-cell sorted to establish a clonal population of Jurkat cells constitutively expressing Cas9 nuclease (Fig. 3.2B). Precise integration of the Cas9-GFP transgene was confirmed by PCR genotyping using primers spanning the junction between CCR5 locus and the transgene, followed by Sanger sequencing (Fig. 3.2C). To test the activity of genomically integrated Cas9, we transfected Cas9 expressing clone with a synthetic sgRNA targeting GFP gene and observed a robust reduction of GFP expressing cells by flow cytometry, demonstrating the expected Cas9 function of cutting the GFP gene, repair of which results in frameshift and knock out (Fig. 3.2D). The constitutive expression of Cas9 in Jurkat cells was not a concern, as previous studies demonstrated absence of any pathological effects associated with such long-term nuclease expression.

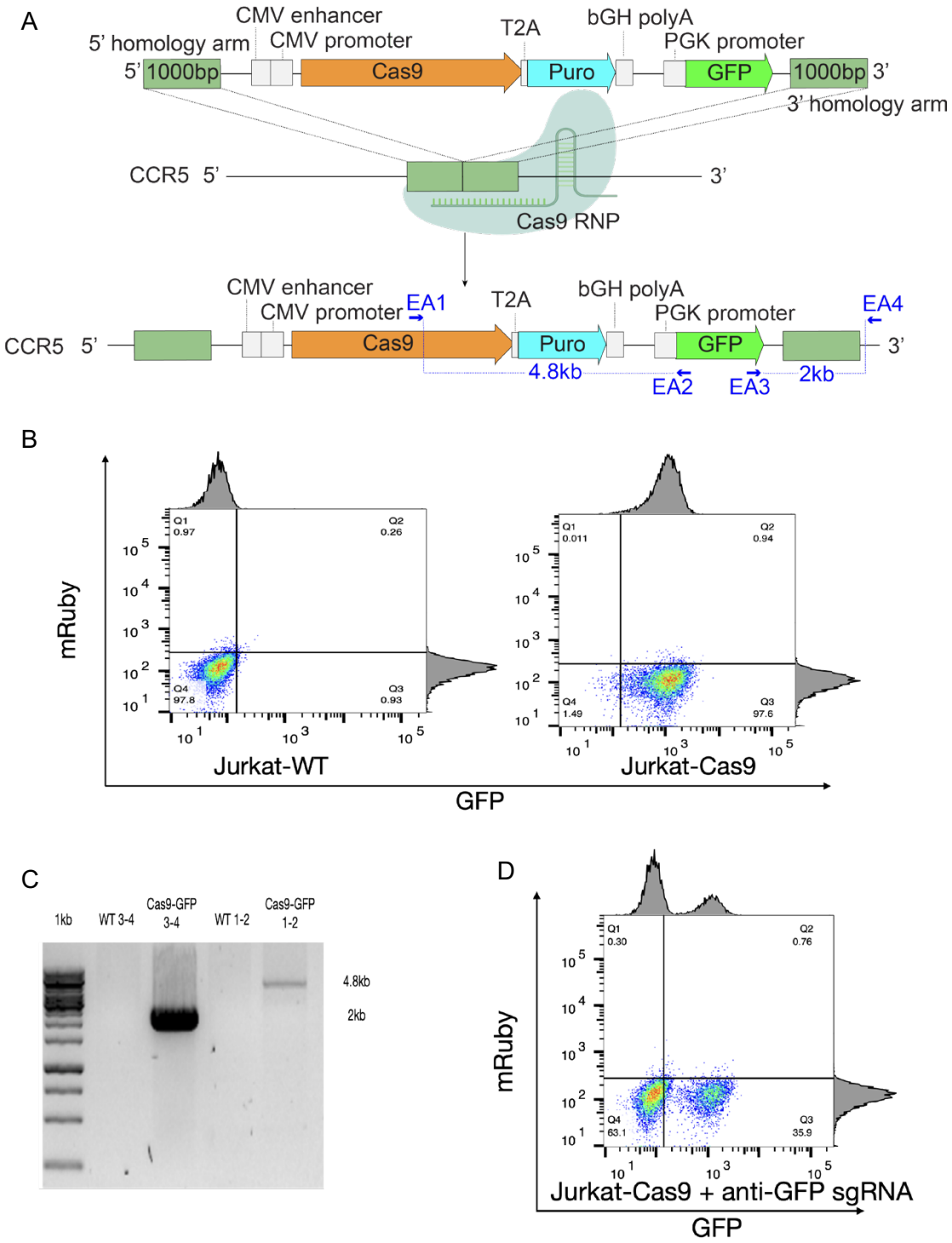


Figure 3.2. Generation of constitutive Cas9-GFP Jurkat cell line. A) Schematic representation of targeted integration of Cas9 and GFP genes into CCR5 locus using HDR-based CRISPR knock in. B) The expression of Cas9-GFP cassette from CCR5 is verified by flow cytometry. C) PCR genotyping of the integrated Cas9-GFP cassette shows CCR5 specific integration. D) Functional activity of Cas9 is verified by supplementing Cas9 expressing clone with anti-GFP sgRNA leading to GFP knockout, confirmed by flow cytometry.

3.4.2 Generation of inducible endogenous gRNA expressing platform in Jurkat T cells

The second cellular platform engineering step involved creation of inducible endogenous gRNA landing pad for subsequent transfection of protospacer library. We tested several inducible gRNA expression transgenes that differed in the composition of the inducible RNA polymerase III promoter. We generated a truncated and a full version of a U6 promoter with its TATA box surrounded by tet operator (TetO) sequences needed for binding of TetR. We also constructed H1 promoter with the same TetO/TATA box layout. All of the constructed gRNA landing pads contained GFP targeting protospacer to test whether the GFP gene inserted into CCR5 locus is knocked out upon addition of doxycycline. Additionally, each of the tested transgenes contained 1000 bp homology arms targeting AAVS1 locus as well as a gene trap system bearing a splice acceptor and T2A self-cleaving peptide sequence followed by mRuby reporter gene. Finally, these AAVS1 targeting constructs encoded constitutively TetR gene to repress anti-GFP guide expression (Fig. 3.3A).

Similar to Cas9-GFP transgene, these inducible gRNA transgenes were transfected in the form of circular plasmid together with AAVS1 targeting synthetic sgRNA complex in Jurkat clones already bearing constitutive Cas9 gene. Clones with successful integration of gRNA landing pads were isolated based on mRuby expression using pooled and then single cell flow cytometry sorts. Integration was also verified by PCR genotyping using primers spanning the junction between AAVS1 locus and the transgene, followed by Sanger sequencing.

To test whether GFP knock out activity of these sgRNA platforms was only occurring after the induction with doxycycline, we measured mean fluorescence intensity (MFI) of GFP with and without the addition of this drug in our two stage-engineered cells. We observed a similar decrease in GFP MFI under both conditions suggesting the leakiness of the expression of sgRNA even when doxycycline is not added (Fig. 3.3B). We additionally confirmed this result using a T7E1 assay, which demonstrated a clear GFP cutting and NHEJ-based repair under both conditions (Fig. 3.3C).

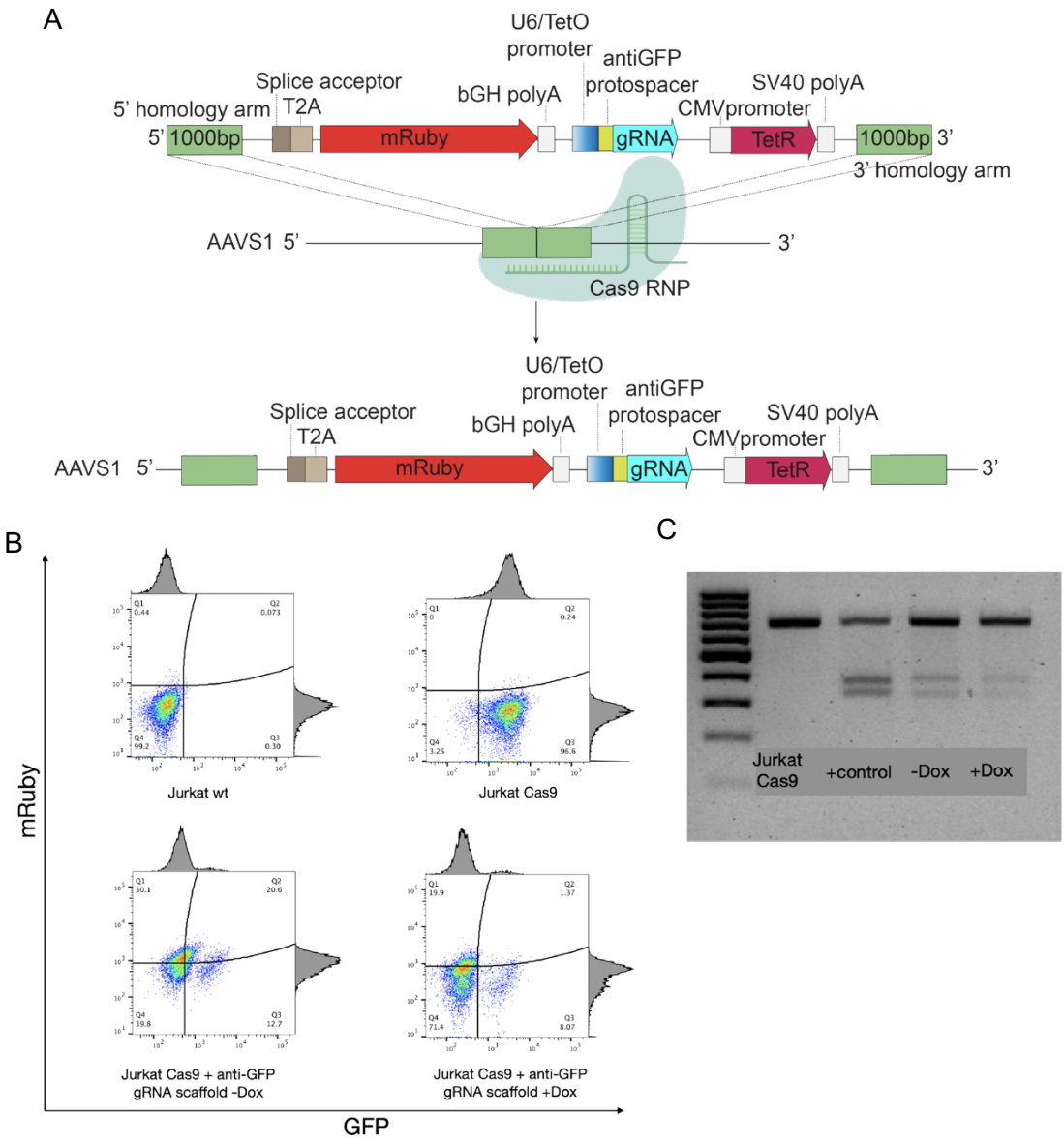


Figure 3.3. Generation of inducible endogenous sgRNA expression platform. A) Schematic representation of targeted integration of inducible sgRNA expression platform bearing anti-GFP sgRNA and TetR using gene trap technology with mRuby into the AAVS1 locus. B) Flow cytometry plots showing successful GFP knockout with and without the addition of doxycycline. C) T7E1 assay demonstrating indel formation in GFP gene when targeted with in vitro transcribed sgRNA (+control) as well as via the inducible endogenous sgRNA platform with and without doxycycline.

3.5 Discussion

High-throughput novel GSH screen attempted in this project aimed at rapid interrogation of thousands of computationally predicted genetic sites in a multiplexed fashion. There are three limiting factors that can impede the GSH screen described in this project. The first is the efficiency of protospacer library incorporation, which is directly dependent on the HDR efficiency in investigated cells, Jurkat immortalized T cell in our case. Being able to augment

HDR rates will significantly increase the diversity of the protospacer library members integrated into the endogenous genomic locus.

Secondly, the leakiness of the tet-inducible RNA Polymerase III promoter may result in the premature expression of gRNA and thus cutting of the target site leading to an indel-prone NHEJ repair and a loss of potential GSH site prior to transfection of the reporter gene. This was indeed observed in our study, when presence and absence of doxycycline inducer resulted in the same level of GFP cutting. One approach to mitigate this effect is to test other inducible RNA Polymerase III systems, not based on tetR. Specifically, Cre-recombinase has recently been shown as an efficient tool to elicit tightly inducible gRNA expression, with loxP sites surrounding the poly-T stretch located in the middle of sgRNA sequences, leading to premature transcription termination and degradation of the short transcript in the absence of cre-recombinase[121]. Alternatively, a tighter tetR system can be designed and tested by changing the positions of tet operators along the U6 or H1 promoter sequences, beyond TATA box, or experimenting with directed evolution approaches to identify sequences supporting tighter gRNA expression, using same flow cytometry-based readout.

Finally, NHEJ-based transgene knock in needs to be optimized to improve the efficiency of homology-independent reporter gene integration into potential GSH sites, pre-determined by the protospacer library member incorporated into the sgRNA landing pad. HITI could be one of the approaches, optimizations of which may result in an increased knock in efficiency. Specifically, testing various circular donor types, that are linearized upon transfection by donor-targeting sgRNAs complexed with Cas9, such as single-cut minicircle plasmids, single- or double-cut plasmids, may result in the identification of knock in strategies suitable for Jurkat transfections. Linear donors, such as ds or ssDNAs can also be tried to augment the efficiency of transgene integration.

Addressing all three of these limiting factors will allow to generate a robust cellular platform for multiplexed screening and identification of novel human GSH sites.

3.6 Methods

Cell culture

Jurkat cells were cultured in ATCC-modified RPMI-1640 (Thermo Fisher, #A1049101). All media were supplemented with 10% FBS, 50 U ml-1penicillin and 50 µg ml^{-1} streptomycin. Detachment of HEK cells for passaging was performed using the TrypLE reagent (Thermo Fisher, #12605010). Cells were cultured at 37°C, 5% CO2 in a humidified atmosphere.

Jurkat T-cell transfection

Prior to transfection of Jurkat gRNA molecules were assembled by mixing 4 µl of custom Alt-R crRNA (200 µM, IDT) with 4 µL of Alt-R tracrRNA (200 µM, IDT, #1072534), incubating the mix at 95°C for 5 min and cooling it to room temperature. 2 µL of assembled gRNA molecules were mixed with 2 µL of recombinant SpCas9 (61 µM, IDT, #1081059) and incubated for > 10 min at room temperature to generate Cas9 RNP complexes.

CCR5-targeting gRNA sequence: TGACATCAATTATTATACAT.

AAVS1-targeting gRNA sequence: GGGGCCACTAGGGACAGGAT.

For transfection of Jurkat cells 100 µL format SE Cell line kit (Lonza, V4XC-1012) and electroporation program CL-120 was used on the 4D-Nucleofector. $1x10^6$ Jurkat cells were transfected with 2 µg of donor plasmid, 2 µl of Cas9 RNP complex against target genomic site.

Inducible sgRNA expression

Plasmid encoding gene trap sequence for AAVS1 integration was purchased from Addgene (#22075) and engineered to contain mRuby sequence and full U6 promoter sequence with two Tet Operator elements surrounding the TATA-box followed by GFP targeting sgRNA sequence as well as Tet repressor gene. 2 µg of this donor plasmid were transfected together with AAVS1-targeting sgRNA into the AAVS1 locus of Cas9-GFP expressing Jurkat cells. Clonal $mRuby^+$ GFP^+ expressing population was subjected to 10 ng/mL of doxycycline. Cells were analyzed using Sony SH800S to observe the inducibility of endogenous sgRNA expression.

Flow cytometry

Transfected Jurkat cells were bulk sorted on day 3 and single-cell sorted on day 10 following transfection using Sony SH800S sorter. For selection of Cas9 transgene expressing clones, cells were isolated based on GFP expression. For selection of sgRNA landing pad transgene expressing clones, cells were isolated based on mRuby expression.

Genotyping

Genomic DNA was extracted from $1x10^6$ cells using PureLink Genomic DNA extraction kit (ThermoFischer Scientific, #K1820-01). 5 µL of genomic DNA extract were then used as templates for 25 µL PCR reactions using a primer with one primer residing outside of the homology arm of the integrated sequence and the other primer inside the integrated sequence. Obtained bands were gel extracted using Zymoclean Gel DNA Recovery Kit (Zymo Research, #D4001), 4ul of eluted DNA was cloned into a TOPO-vector using Zero-blunt TOPO PCR Cloning Kit (ThermoFischer Scientific, #450245), incubated for 1 hour, transformed into NEB 5-alpha Competent E. coli cells (New England Biolabs, C2987H) and plated on agar plates containing kanamycin at 50 µg/ml. Produced clones were picked and inoculated for overnight

culture in 5ml of liquid broth supplemented with kanamycin at 50 µg/ml. Liquid cultures were mini-prepped the following morning using ZR Plasmid Miniprep - Classic kit (Zymo Research, #D4015) and Sanger sequenced by Microsynth using M13-forward and M13-reverse standard primers.

T7E1 assay

Region in the GFP gene was PCR amplified in using the following primer set: Forward GGGCGAGGAGCTGTTCA, Reverse CAGCTCGTCCATGCCGAGAG. After running on the gel bands were extracted and mixed with NEBuffer2 (New England Biolabs) and annealed for 10 minutes, with temperature reduced at 0.3C increments per minute. PCR products were then digested with T7E1 enzyme for 40 minutes at 37C, followed by gel electrophoresis.

4. General discussion

Therapeutic integrative gene addition possesses significant value for gene and cell therapies, particularly in the context of mitotically active cells. Being able to sustain gene expression through genomic insertion in target cells and all of its progenies allows to prevent transgene diluting effects due to cellular division. Despite showing clinical success for certain disease indications, current approaches to integrative gene addition bear a significant disadvantage of random nature of transgene insertion. Development of novel genome engineering tools supporting precise genomic modifications now permits a targeted insertion of therapeutic transgenes in different cell types, dividing and non-dividing. However, absence of well-characterized genomic sites capable of sustaining safe and durable transgene expression still holds targeted transgene integration approaches from enteric into clinical practice. Additionally, various mammalian synthetic biology approaches rely on creation of genetic circuits containing large sets of coding and non-coding genes, expression of which needs to be predictable and not perturbing other cellular functions. Identification of novel GSH sites would thus introduce novel opportunities to the fields of cell engineering, gene therapy and synthetic biology improving safety, durability and predictability of their research and therapeutic applications.

In this dissertation I described a computational search based on existing as well new GSH criteria to predict sites safe for genomic insertion and transgene expression. This bioinformatic pipeline generated close to 2,000 integration sites in the human genome, each of which satisfied initial safety screen. I then discussed how these predicted sites can be investigated for their ability to provide high and stable transgene expression levels. First, a rational approach targeting several predicted sites individually was applied using CRISPR/Cas9 method and MMEJ-based integration. Best expressing sites were then investigated for their safety using single cell and bulk transcriptomics. This approach identified two genomic sites capable of stable expression of reporter and therapeutic genes of interest in cell lines as well as in primary human cells, paving way for their inclusion in future gene and cell therapy clinical trials. I also described an attempted alternative high-throughput approach, in which we sought to validate thousands of GSH sites simultaneously using libraries of inducible gRNAs targeting computationally predicted sites. This methodology, however, faced an experimental hurdle of leaky gRNA expression and was ultimately put on hold. Recently introduced technological advances in tight inducible gRNA expression could allow to restart this high-throughput approach, leading to identification of more GSH sites supporting even more robust transgene expression levels.

Although the safety of virally integrated gene therapies has markedly improved over the past

years and more regulatory approvals has been witnessed for lentivirus delivered transgenes[21], the concern associated with perturbation of normal cellular function following semi-random lentiviral integration events still remains. For instance, initially successful clinical trials of Lentiglobin autologous cell therapy product, composed of ex vivo lentivirally engineered HSCs bearing a functional copy of hemoglobin subunit beta in patients with sickle cell disease, had to be halted due to oncogenic transformation of engineered cells. The causes of this event are currently unknown, and the insertional oncogenesis might not be the culprit in this particular case, however the unpredictable nature of transgene integration using lentiviruses remains a concern and other cell therapy trials may potentially result in similar complications.

Gene and cell therapies are the main beneficiaries of the newly identified GSH sites. New targets for safe and stable transgene integration can play a critical role in treatment of inherited genetic disorders in proliferating tissues, such as skin, blood cells, stem cells residing across the body, etc. With groundbreaking research occurring in the field of immunoengineering, exploiting safe harbors to introduce new receptors and response elements into immune cells can be particularly attractive. Stem cell engineering could also be an exciting application for the identified GSHs. Being able to safely introduce therapeutic genes into iPSCs followed by differentiation into desired tissues could be very valuable for allogeneic cell therapies[70]. Additionally, recombinant protein production field is likely to benefit from having a long-term predictable expression from a GSH, allowing to create better stable cell banks of human protein producing cell lines. Finally, one particularly exciting application for the validated GSHs is integration of enhancing genes, providing novel beneficial properties to cells and tissues. An example of such approach includes aging reversal through integration of transcription factors associated with early developmental phenotype[122]. In this case, despite the epigenetic remodeling that the cell undergoes during development and aging, the expression of the rejuvenating transgene from GSH will be maintained. Another example could be knocking in of neurotrophic factors into the GSHs of adult neurons to promote neural regeneration[123]. The opportunities for augmenting human cells are limitless and with the safety provided by safe harbor expression, they may start transitioning from research to therapies.

Creation of complex synthetic biology gene circuits in mammalian cells requires a predictable expression of multiple coding or regulatory genetic elements, which would have minimal interference with the function of the endogenous genes. This is particularly relevant for ex vivo engineered human cells that will be introduced into the patient to elicit a multi-component response to a range of different inputs for therapeutic and diagnostic applications, such as genetically reprogrammed T cells or hematopoietic stem cells. Multiple human genomic safe harbor sites could satisfy this necessity for predictable and safe expression of the components of synthetic gene networks. One of the most exciting applications of such technology could be

for tissue specific reprogramming of engineered pluripotent stem cells with different transcription factors integrated in GSHs and expressed in response to tissue specific stimuli resulting in desired differentiation patterns. This approach could be particularly relevant for the reconstitution of tissue residing stem cells that undergo aging-associated deficiencies.

Skin is one of the mitotically active organs which is easily accessible for genetic manipulation, making it a good a candidate for GSH-based gene integration for therapeutic and tissue enhancing purposes. One of the skin diseases that can potentially be treated with targeted knock-in of genes of interest is psoriasis – a common autoimmune condition associated with an immune cascade inducing faster keratinocyte proliferation[124]. One approach to mitigating this recurring disease is to introduce an "antipsoriatic cytokine converter" described by Shukur et al. – a gene network that responds to an increased presence of proinflammatory cytokines TNF and IL22 resulting in production of anti-inflammatory cytokines IL4 and IL10[125]. This is achieved by sequential activation of endogenous TNF alpha, followed by NF-kB driven expression of IL22 receptor, which when activated by IL22 cytokine induces STAT3-driven expression of secreted IL4 and IL10 capable of preventing or attenuating acute manifestation of psoriatic disease. Both of the genetic components used to create described gene network can be expressed from safe harbor regions in dermal fibroblasts to make this treatment modality safer. In addition, numerous inherited disorders of the skin can be addressed by targeted integration of therapeutic transgenes into safe harbor regions. Epidermolysis bullosa alone can be caused by mutations in around 30 different genes involved in structural integrity of the skin. Replacing this malfunctional genes with their wild-type copies expressed from a GSH site will result in durable and safe treatment[56]. Finally, additional enhancing properties can be introduced into skin tissue using GSH-based genome engineering: wound or burn healing can be augmented by the introduction of platelet derived growth factor or epidermal growth factor into skin cells, damage-induced expression of which will result in faster fibroblast proliferation and angiogenesis[58] (Fig. 4.1).

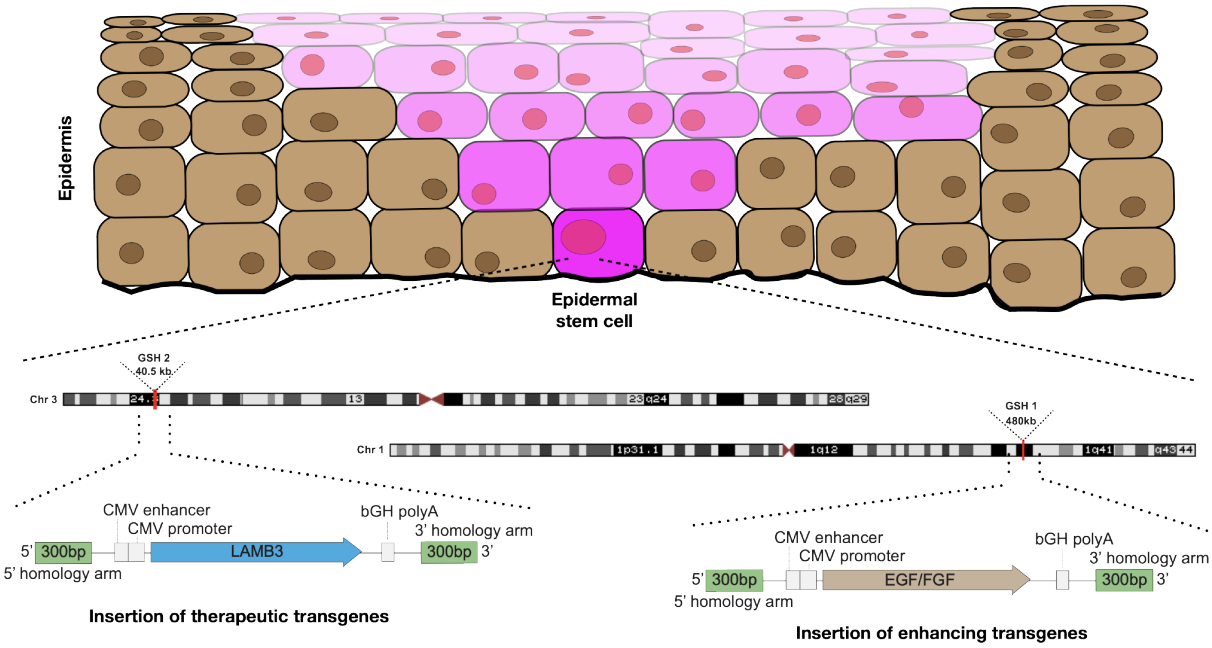


Figure 4.1. GSH application to in vivo skin cell engineering. A schematic of GSH-based gene therapy applications to inherited skin disorders and wound healing. Targeting of therapeutic and enhancing genes to safe harbor site in epidermal stem cells may allow for continuous safe gene expression in proliferating cells of epidermis.

Integrations into novel GSH sites would have a significant impact on a rapidly evolving field of T cell engineering. Specifically, gene cassettes encoding synthetic receptors which target cancer cells can be integrated into novel GSHs ex vivo for safer expression once introduced into patients. As an example, synthetic Notch (synNotch) receptor described by Roybal et al., is capable of sensing a variety of orthogonal molecular inputs and drive the expression of customizable cytokines in primary human T cells, sidestepping the natural T cell activation[87]. Cancer therapy could be one of the applications of this approach, when engagement of synNotch receptor by cancer cell antigens results in activation of introduced Tbet transcription factor, responsible for T_{h1} lineage skewing of $CD4^+$ T cells. T_{h1} cells produce IFNγ essential for the induction of innate immune mechanisms against tumor cells. Engineering of T cell bearing donors that encode such synthetic receptors as well as their response elements could be done in genomic safe harbor locus for safer clinical applications. In addition, GSH-based insertion of CARs as well as knock-out of endogenous TCR in iPSCs followed by the differentiation of the latter into T cells may provide a safer route for allogeneic off-the-shelf T cell therapies without inducing graft-versus-host disease[126].

Scientists can also employ other tools to make the expression from a genomic safe harbor safer. One such approach is the use of insulator elements, a relatively short DNA sequence operating as a binding site for insulating proteins. These proteins are isolating genomic sequences between them allowing for a separate regulation within this isolated region[127].

Adding such insulator sequences to transgene bearing HDR donor would further increase the safety profile of GSH engineered cells. Additionally, methods to increase integration efficiency in dividing and non-dividing cell should be explored, helping to achieve larger pools of GSH transgene integrated cells. In Chapter 2 I described the approach we used to GSH knock in transgenes into mitotically active primary cells using dsDNA donors and nucleofection and transfection approaches. Despite being able to isolate positive clones, the integration efficiency using such methods was low. One way to improve this is by using alternative donor types, for example HDR donors packaged into a AAV vector, or ssDNA donor delivered in an LNP.

Furthermore, additional methods to verify safety of transgene expression from identified loci can be used. One could envision assessing changes in chromosomal architecture following transgene insertion into an identified GSHs, which could be done by comparing HiC profiles of cells with and without GSH insertion[128,129]. Additionally, using modern high-throughput mass spectrometry approaches, cellular metabolome could also be investigated in GSH integrated versus non-integrated cells to confirm stability of metabolic pathway following GSH integration[130].

Finally, the experimental validation for novel genomic safe harbor sites can be expanded to all other computationally predicted regions that satisfy introduced safe harbor criteria using optimized high-throughput screening approaches for simultaneous interrogation of multiple putative GSHs. In chapter 3 I discussed an attempt to use inducible sgRNA expression for temporal regulation of DNA cutting in the presence of NHEJ donor. This inducible expression was based on Tet-repressor system, which demonstrated leakiness of the RNA polymerase III driven sgRNA expression. An alternative strategy for inducible sgRNA expression has been proposed by Chylinski et al., which is based on Cre-mediated recombination event resulting in the excision of polyT stretch placed in the middle of the sgRNA sequence[121] (Fig. 4.2A). This event allows for the production of functional sgRNA molecule in the presence of Cre-recombinase, thus capable of a very tight temporal control of the DNA cutting as compared to leaky U6/TetO system (Fig. 4.2B). Together with the endogenous sgRNA expression platform using libraries of protospacers described by Rajagopal et al., this approach would address the limitation of the study in chapter 3 and allow for inducible expression of libraries of sgRNAs, and thus interrogating multiple genomic sites in a high-throughput manner (Fig. 4.2C), eventually allowing to assess all computationally predicted safe harbor sites in a high-throughput manner.

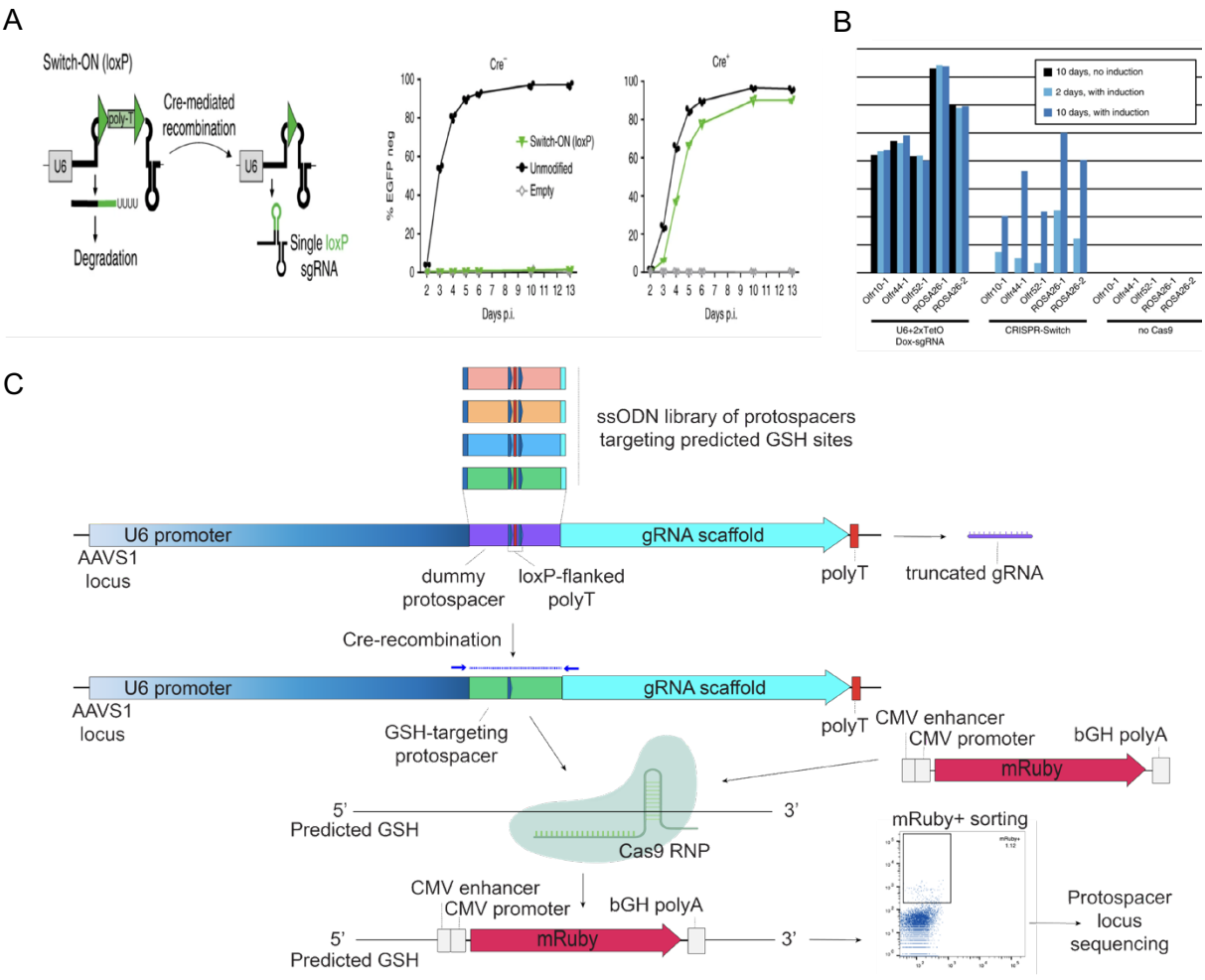


Figure 4.2. Cre-recombinase inducible endogenous sgRNA expression platform. A) Chylinski et al. have reported a novel approach for temporally controlled sgRNA expression using Cre-mediated excision of polyT sequence located in the middle of sgRNA. This event prevents the premature sgRNA degradation and allows for the production of functional sgRNA. B) This method was compared with U6/TetO inducible sgRNA expression an showed very tight expression exclusively in the presence of Cre. As observed in Chapter 3, U6/TetO system demonstrated leaky sgRNA expression, unsuitable for on demand expression of sgRNAs. C) Similar to the design in chapter 3, genomically integrated inducible sgRNA expression platform can be used to express a library of sgRNA sequences targeting bioinformatically predicted GSHs following Cre-mediated recombination. This temporally controlled sgRNA expression will be accompanied by transfection of NHEJ mRuby donor, which will be expressed in case of integration into a true GSH. Cells durably expressing mRuby will be sorted and protospacer locus will be sequenced to identify mRuby integration site and hence the newly validated GSH.

Conclusion

Prior to this research, targeted insertion of therapeutic genes into human cells was limited to three serendipitously identified sites, which due to the unknown function of the genes they are located in as well as being surrounded by oncogenes, did not experience transition into clinical practice. In this dissertation I described computational prediction of novel genomic safe harbor

sites based on existing and newly introduced criteria. From these predicted sites, two were validated for safe and stable expression of genes of interest in a range of different cell types. This discovery will help improve gene and cell therapies relying on gene addition as well as support the rapid pace of innovation in synthetic biology that enables multiple transgene integration and genetic circuits to rewire, reprogram and augment cellular function.

5. Appendices

S1. Gene ontology analysis of HEK293T cells following GSH2 mRuby integration

A

Term	Ont	N	Up	Down	P.Up	P.Down
response to external biotic stimulus	BP	443	4	16	0.6143769180	0.0002243617
response to other organism	BP	443	4	16	0.6143769180	0.0002243617
ketone body catabolic process	BP	3	2	0	0.0002689309	1.0000000000
negative regulation of viral entry into host cell	BP	11	0	3	1.0000000000	0.0003344033
positive regulation of B cell differentiation	BP	11	0	3	1.0000000000	0.0003344033
response to biotic stimulus	BP	468	4	16	0.6574891374	0.0004134829
positive regulation of gamma-delta T cell activation	BP	3	0	2	1.0000000000	0.0005045283
positive regulation of gamma-delta T cell differentiation	BP	3	0	2	1.0000000000	0.0005045283
positive regulation of synapse structural plasticity	BP	4	2	0	0.0005345013	1.0000000000
granulocyte chemotaxis	BP	34	0	4	1.0000000000	0.0009591469
regulation of gamma-delta T cell activation	BP	4	0	2	1.0000000000	0.0010003837
regulation of gamma-delta T cell differentiation	BP	4	0	2	1.0000000000	0.0010003837
negative regulation of viral life cycle	BP	59	1	5	0.4326283413	0.0010142029
regulation of lamellipodium assembly	BP	22	3	0	0.0011412228	1.0000000000
regulation of synapse structural plasticity	BP	6	2	0	0.0013196258	1.0000000000
epithelial cell proliferation	BP	227	8	4	0.0014733018	0.3448881229
regulation of B cell differentiation	BP	18	0	3	1.0000000000	0.0015462581
regulation of viral entry into host cell	BP	18	0	3	1.0000000000	0.0015462581
granulocyte migration	BP	39	0	4	1.0000000000	0.0016171357
response to wounding	BP	404	11	9	0.0016423665	0.0829742818

B

Term	Ont	N	Up	Down	P.Up	P.Down
catalytic activity	MF	4393	60	63	0.0004425268	0.198861176
nucleotide binding	MF	1938	33	25	0.0004592407	0.561260414
nucleoside phosphate binding	MF	1939	33	25	0.0004636010	0.562378980
small molecule binding	MF	2042	34	26	0.0005547028	0.590411411
glucosyltransferase activity	MF	16	0	3	1.0000000000	0.001081704
histone demethylase activity (H3-K4 specific)	MF	6	2	0	0.0013196258	1.000000000
hydro-lyase activity	MF	40	1	4	0.3188250968	0.001778738
purine ribonucleotide binding	MF	1521	26	19	0.0020434120	0.619167125
ribonucleotide binding	MF	1533	26	19	0.0022831866	0.633343282
purine nucleotide binding	MF	1534	26	19	0.0023042196	0.634513905
GTPase activity	MF	191	7	0	0.0023371550	1.000000000
cysteine-type endopeptidase activity involved in apoptotic process	MF	8	2	0	0.0024326946	1.000000000
carbonate dehydratase activity	MF	6	0	2	1.0000000000	0.002458218
purine ribonucleoside triphosphate binding	MF	1496	25	19	0.0034322371	0.588921634
purine ribonucleoside binding	MF	1502	25	19	0.0036195709	0.596262893
purine nucleoside binding	MF	1504	25	19	0.0036839179	0.598698854
ribonucleoside binding	MF	1505	25	19	0.0037164545	0.599914705
nucleoside binding	MF	1510	25	19	0.0038828266	0.605972281
carbohydrate derivative binding	MF	1711	27	22	0.0051205153	0.567117557
structural molecule activity	MF	493	4	14	0.6973007445	0.005219669

C

Term	Ont	N	Up	Down	P.Up	P.Down
nucleosome	CC	75	0	6	1.000000000	0.0004338150
protein-DNA complex	CC	141	0	8	1.000000000	0.0005198931
DNA packaging complex	CC	81	0	6	1.000000000	0.0006556767
endoplasmic reticulum lumen	CC	131	2	7	0.356037244	0.0016441876
extracellular region part	CC	2423	21	47	0.723227120	0.0022964254
extracellular region	CC	2662	25	50	0.571146871	0.0031648623
neuron projection	CC	651	14	10	0.003578165	0.3453728522
banded collagen fibril	CC	8	0	2	1.000000000	0.0045104152
fibrillar collagen trimer	CC	8	0	2	1.000000000	0.0045104152
membrane part	CC	3827	50	44	0.005563888	0.8672688832
polymeric cytoskeletal fiber	CC	445	3	13	0.802157386	0.0055882831
supramolecular complex	CC	445	3	13	0.802157386	0.0055882831
supramolecular fiber	CC	445	3	13	0.802157386	0.0055882831
supramolecular polymer	CC	445	3	13	0.802157386	0.0055882831
nuclear nucleosome	CC	28	0	3	1.000000000	0.0056421780
Cul4A-RING E3 ubiquitin ligase complex	CC	9	0	2	1.000000000	0.0057495189
extracellular matrix component	CC	90	0	5	1.000000000	0.0064458127
intrinsic component of membrane	CC	2956	40	34	0.008817007	0.8283156717
matrix side of mitochondrial inner membrane	CC	1	1	0	0.009537549	1.0000000000
integral component of membrane	CC	2904	39	34	0.011201088	0.7947113334

S2. Gene ontology analysis of Jurkat cells following GSH2 mRuby integration

A

Term	Ont	N	Up	Down	P.Up	P.Down
regulation of chondrocyte differentiation	BP	32	0	2	1	0.0008461026
wound healing	BP	334	0	4	1	0.0009106351
cocaine metabolic process	BP	1	0	1	1	0.0013625070
negative regulation of phospholipase A2 activity	BP	1	0	1	1	0.0013625070
neuroblast differentiation	BP	1	0	1	1	0.0013625070
positive regulation of neutrophil apoptotic process	BP	1	0	1	1	0.0013625070
tropane alkaloid metabolic process	BP	1	0	1	1	0.0013625070
regulation of cartilage development	BP	45	0	2	1	0.0016713049
response to wounding	BP	404	0	4	1	0.0018433431
DNA rewinding	BP	2	0	1	1	0.0027232667
fast-twitch skeletal muscle fiber contraction	BP	2	0	1	1	0.0027232667
myoblast migration involved in skeletal muscle regeneration	BP	2	0	1	1	0.0027232667
neutrophil apoptotic process	BP	2	0	1	1	0.0027232667
neutrophil clearance	BP	2	0	1	1	0.0027232667
positive regulation of prostaglandin biosynthetic process	BP	2	0	1	1	0.0027232667
positive regulation of unsaturated fatty acid biosynthetic process	BP	2	0	1	1	0.0027232667
regulation of neutrophil apoptotic process	BP	2	0	1	1	0.0027232667
chondrocyte differentiation	BP	68	0	2	1	0.0037755636
regulation of wound healing	BP	70	0	2	1	0.0039962310
negative regulation of T-helper 2 cell differentiation	BP	3	0	1	1	0.0040822810

B

Term	Ont	N	Up	Down	P.Up	P.Down
phospholipase inhibitor activity	MF	10	0	2	1.000000000	7.812861e-05
lipase inhibitor activity	MF	11	0	2	1.000000000	9.541400e-05
calcium-dependent phospholipid binding	MF	28	0	2	1.000000000	6.468814e-04
acetylcholinesterase activity	MF	1	0	1	1.000000000	1.362507e-03
double-stranded DNA-dependent ATPase activity	MF	1	0	1	1.000000000	1.362507e-03
glial cell-derived neurotrophic factor receptor activity	MF	1	0	1	1.000000000	1.362507e-03
enzyme inhibitor activity	MF	209	0	3	1.000000000	2.648948e-03
choline binding	MF	2	0	1	1.000000000	2.723267e-03
cholinesterase activity	MF	2	0	1	1.000000000	2.723267e-03
epidermal growth factor-activated receptor activity	MF	2	0	1	1.000000000	2.723267e-03
polypeptide N-acetylgalactosaminyltransferase activity	MF	14	1	0	0.003362687	1.000000e+00
structural molecule activity	MF	493	0	4	1.000000000	3.801635e-03
phospholipase A2 inhibitor activity	MF	3	0	1	1.000000000	4.082281e-03
cadherin binding involved in cell-cell adhesion	MF	272	0	3	1.000000000	5.555523e-03
protein binding involved in cell-cell adhesion	MF	279	0	3	1.000000000	5.961947e-03
acetylgalactosaminyltransferase activity	MF	25	1	0	0.005999504	1.000000e+00
cadherin binding	MF	282	0	3	1.000000000	6.141495e-03
protein binding involved in cell adhesion	MF	283	0	3	1.000000000	6.202064e-03
endogenous lipid antigen binding	MF	5	0	1	1.000000000	6.795082e-03
exogenous lipid antigen binding	MF	5	0	1	1.000000000	6.795082e-03

C

Term	Ont	N	Up	Down	P.Up	P.Down
adherens junction	CC	629	0	6	1	0.0001229812
anchoring junction	CC	639	0	6	1	0.0001341854
sarcolemma	CC	80	0	3	1	0.0001617804
plasma membrane	CC	2747	0	10	1	0.0010935994
focal adhesion	CC	362	0	4	1	0.0012289977
cell-substrate adherens junction	CC	363	0	4	1	0.0012416438
blood microparticle	CC	39	0	2	1	0.0012569701
cell-substrate junction	CC	365	0	4	1	0.0012672117
cell periphery	CC	2820	0	10	1	0.0013583354
cell junction	CC	1055	0	6	1	0.0019737821
endothelial microparticle	CC	2	0	1	1	0.0027232667
neurofilament cytoskeleton	CC	2	0	1	1	0.0027232667
extrinsic component of external side of plasma membrane	CC	3	0	1	1	0.0040822810
intermediate filament cytoskeleton	CC	80	0	2	1	0.0051873829
extracellular space	CC	566	0	4	1	0.0062184195
basal cortex	CC	5	0	1	1	0.0067950824
cornified envelope	CC	5	0	1	1	0.0067950824
extrinsic component of endosome membrane	CC	5	0	1	1	0.0067950824
cell-cell adherens junction	CC	302	0	3	1	0.0074221620
external side of plasma membrane	CC	106	0	2	1	0.0089475814

Table 1. Computationally predicted GSH sites.

Chr.	Start	End	Size	ID
chr1	195338589	195818588	479999	GSH1
chr3	22720711	22761389	40678	GSH2
chrX	89174426	89179074	4648	GSH31
chr7	145090941	145219513	128572	GSH7
chr7	145320384	145525881	205497	GSH8
chr1	4105262	4125527	20265	
chr1	4225899	4262026	36127	
chr1	5240899	5342977	102078	
chr1	14541575	14548703	7128	
chr1	34327292	34369582	42290	
chr1	38646034	38658930	12896	
chr1	60299679	60353058	53379	
chr1	61512793	61535520	22727	
chr1	61576366	61579030	2664	
chr1	64321297	64334397	13100	
chr1	65691559	65705302	13743	
chr1	66424579	66431399	6820	
chr1	66472315	66483382	11067	
chr1	68688627	68713014	24387	
chr1	68753786	68905897	152111	
chr1	72362970	72598920	235950	
chr1	73933822	73976014	42192	
chr1	78800659	78839763	39104	
chr1	79574181	79843196	269015	
chr1	79883946	79964942	80996	
chr1	80796788	81029014	232226	
chr1	81149005	81158567	9562	
chr1	82042436	82065219	22783	
chr1	82411133	82547687	136554	
chr1	82588475	82610029	21554	
chr1	82710119	82753182	43063	
chr1	86206943	86233686	26743	
chr1	86274496	86296823	22327	
chr1	87521655	87655285	133630	
chr1	88055714	88109103	53389	
chr1	88149895	88263152	113257	
chr1	88363302	88367432	4130	
chr1	90120300	90203455	83155	
chr1	90303559	90360909	57350	
chr1	91074804	91123611	48807	
chr1	91164419	91210765	46346	
chr1	96524125	96533208	9083	
chr1	98423499	98475475	51976	
chr1	102047030	102049738	2708	
chr1	102539630	102572439	32809	
chr1	102613189	102613321	132	
chr1	103158496	103197336	38840	
chr1	103238162	103264878	26716	
chr1	103841108	103876566	35458	
chr1	103979438	104022982	43544	
chr1	104301393	104593886	292493	
chr1	104634632	104744867	110235	
chr1	104793981	104826208	32227	
chr1	104866968	104965109	98141	
chr1	105005885	105088762	82877	
chr1	105129498	105197411	67913	
chr1	105238169	105383993	145824	
chr1	105768958	105777623	8665	
chr1	106274564	106317447	42883	
chr1	106358207	106494341	136134	
chr1	106594744	106668238	73494	
chr1	106988167	107006678	18511	
chr1	113249258	113278135	28877	
chr1	113318971	113340748	21777	
chr1	118314489	118524655	210166	
chr1	118763678	118801408	37730	
chr1	163559595	163719338	159743	
chr1	163825577	163873669	48092	
chr1	163973873	164256545	282672	
chr1	164438361	164488657	50296	
chr1	165033992	165060626	26634	
chr1	187793860	187894353	100493	
chr1	187935143	188017297	82154	
chr1	189187262	189616148	428886	
chr1	190043097	190047661	4564	
chr1	190951658	191001509	49851	
chr1	191378467	191382046	3579	
chr1	191422804	191708706	285902	
chr1	193877035	193937569	60534	
chr1	193978345	193999472	21127	
chr1	194538205	194625045	86840	
chr1	194665799	194668794	2995	
chr1	194769236	195076680	307444	
chr1	195176776	195297813	121037	
chr1	195859392	195865261	5869	
chr1	197985478	198020764	35286	
chr1	198061596	198106962	45366	
chr1	199543307	199565187	21880	
chr1	199665280	199702490	37210	
chr1	199802890	199826977	24087	
chr1	208305376	208457099	151723	
chr1	213324773	213342300	17527	
chr1	214180901	214194171	13270	
chr1	214714588	214812154	97566	
chr1	217192388	217293778	101390	
chr1	217334568	217376991	42423	
chr1	218675978	218742429	66451	
chr1	221057985	221083864	25879	
chr1	232335643	232347964	12321	
chr1	233517478	233564003	46525	
chr1	233722512	233765758	43246	
chr1	238688305	238727048	38743	
chr1	238829200	238933993	104793	
chr1	238974783	239002747	27964	
chr1	242574696	242621139	46443	
chr1	242721576	242722619	1043	
chr1	242823103	242832065	8962	
chr10	1787476	1853593	66117	
chr10	2717239	2739038	21799	
chr10	2839125	2860530	21405	
chr10	9123468	9125832	2364	
chr10	9437057	9495470	58413	
chr10	9536210	9608782	72572	
chr10	15421062	15424596	3534	
chr10	15910520	15955807	45287	
chr10	16121809	16128700	6891	
chr10	18990953	18998770	7817	
chr10	20438691	20497170	58479	
chr10	20602046	20615848	13802	

chr10	20656646	20721318	64672	chr10	83487129	83522405	35276	chr11	25218770	25301913	83143
chr10	22053769	22068073	14304	chr10	84611263	84617278	6015	chr11	25342649	25487277	144628
chr10	22844639	22878023	33384	chr10	84658092	84813264	155172	chr11	26773427	26837648	64221
chr10	23513123	23595999	82876	chr10	84913362	84954690	41328	chr11	26878458	26897185	18727
chr10	23636767	23644745	7978	chr10	84995456	85043420	47964	chr11	36773176	37228321	455145
chr10	26371378	26388202	16824	chr10	90213960	90252520	38560	chr11	37328417	37652124	323707
chr10	29031898	29062443	30545	chr10	92783115	92784712	1597	chr11	37776456	37788600	12144
chr10	29103311	29239060	135749	chr10	100609998	100616694	6696	chr11	38262799	38273729	10930
chr10	29786781	29841031	54250	chr10	105315235	105320229	4994	chr11	38314471	38348994	34523
chr10	29941467	29962799	21332	chr10	105401249	105457043	55794	chr11	38856655	38930053	73398
chr10	36239389	36288615	49226	chr10	105584765	105602777	18012	chr11	38970797	39111452	140655
chr10	36676138	36790108	113970	chr10	105969197	105990263	21066	chr11	39311213	39523985	212772
chr10	44094466	44108413	13947	chr10	106338722	106500105	161383	chr11	39564727	39680811	116084
chr10	47094274	47162503	68229	chr10	107214534	107251341	36807	chr11	39780908	39811505	30597
chr10	47203419	47250385	46966	chr10	107292077	107411559	119482	chr11	39924388	39957243	32855
chr10	53125815	53141071	15256	chr10	107511669	107544972	33303	chr11	42412427	42541371	128944
chr10	53508288	53552028	43740	chr10	108347849	108397974	50125	chr11	42582099	42726050	143951
chr10	53592810	53716399	123589	chr10	108991310	109134933	143623	chr11	42766770	42837854	71084
chr10	55796908	56001781	204873	chr10	109289746	109403875	114129	chr11	42878598	42889774	11176
chr10	56042531	56307227	264696	chr10	109444653	109478118	33465	chr11	42995111	43013636	18525
chr10	56425945	56545962	120017	chr10	109584990	109611813	26823	chr11	48678493	48702710	24217
chr10	56646031	56682644	36613	chr10	109731603	109757650	26047	chr11	48767083	48777549	10466
chr10	56723410	57004458	281048	chr10	111155248	111199938	44690	chr11	48818309	48829318	11009
chr10	57614587	57845133	230546	chr10	111548941	111594994	46053	chr11	49501793	49503211	1418
chr10	57885939	57905294	19355	chr10	111635790	111908711	272921	chr11	50483245	50521348	38103
chr10	58032993	58056422	23429	chr10	118515103	118522870	7767	chr11	79762300	79816804	54504
chr10	61176420	61196196	19776	chr10	120174226	120304700	130474	chr11	80118708	80171619	52911
chr10	61237038	61302638	65600	chr10	125312971	125423685	110714	chr11	80477911	80477959	48
chr10	64220850	64453296	232446	chr10	126829774	126838094	8320	chr11	80578066	80601199	23133
chr10	64494060	64570373	76313	chr10	128508017	128704118	196101	chr11	81007634	81072518	64884
chr10	64670922	64712882	41960	chr10	128744864	128762878	18014	chr11	81172586	81326649	154063
chr10	76710994	76738043	27049	chr10	129110062	129128250	18188	chr11	81367365	81502225	134860
chr10	80828088	80836317	8229	chr10	130633154	130840784	207630	chr11	81606425	81719315	112890
chr10	80877055	80890664	13609	chr10	130881570	130912587	31017	chr11	91533289	91644361	111072
chr10	80991201	81086539	95338	chr11	15314207	15328530	14323	chr11	93900531	93915469	14938
chr10	81194731	81212450	17719	chr11	20285465	20313684	28219	chr11	96866364	96950273	83909
chr10	81253198	81544430	291232	chr11	21629550	21695849	66299	chr11	97409987	97604938	194951
chr10	81585166	81676009	90843	chr11	21736599	21780742	44143	chr11	97707582	97728780	21198
chr10	81716779	81728607	11828	chr11	21821466	21963447	141981	chr11	98181198	98490106	308908
chr10	81804364	81818896	14532	chr11	23571202	23580587	9385	chr11	100408885	100512911	104026
chr10	83118071	83261627	143556	chr11	23954598	24011392	56794	chr11	101359591	101401563	41972
chr10	83364475	83387033	22558	chr11	24052158	24085476	33318	chr11	104214379	104264148	49769

chr11	104759321	104819533	60212	chr12	61650843	61658258	7415	chr13	42658013	42660365	2352
chr11	106361791	106449170	87379	chr12	72878844	72965956	87112	chr13	47006299	47257081	250782
chr11	106489976	106624011	134035	chr12	73358317	73524928	166611	chr13	47357301	47408940	51639
chr11	107068524	107126285	57761	chr12	73565680	73598665	32985	chr13	47509234	47675329	166095
chr11	110927177	110939869	12692	chr12	74778851	74870968	92117	chr13	52898641	52918249	19608
chr11	113562117	113563491	1374	chr12	80153333	80159452	6119	chr13	52959099	52978758	19659
chr11	113604341	113620392	16051	chr12	82143133	82152287	9154	chr13	53562629	53665418	102789
chr11	114756933	114876250	119317	chr12	83322740	83470569	147829	chr13	54282866	54390703	107837
chr11	114917014	115078217	161203	chr12	83837208	84119904	282696	chr13	54491315	54755866	264551
chr11	116092995	116247623	154628	chr12	84499878	84569465	69587	chr13	55311490	55385696	74206
chr11	116288463	116346898	58435	chr12	84610257	84621597	11340	chr13	55733524	55949199	215675
chr11	121683693	121799738	116045	chr12	84722026	84809487	87461	chr13	56050448	56164372	113924
chr11	127487033	127520739	33706	chr12	85492912	85517877	24965	chr13	56266942	56269506	2564
chr11	127561471	127607567	46096	chr12	85697320	85731891	34571	chr13	56310276	56697084	386808
chr11	127722973	127817842	94869	chr12	87220429	87239902	19473	chr13	56737866	56835283	97417
chr11	127858606	127890912	32306	chr12	87280706	87538342	257636	chr13	56935602	57044266	108664
chr11	127991654	128030426	38772	chr12	87579094	87645732	66638	chr13	57085040	57090917	5877
chr11	133582519	133633670	51151	chr12	88249887	88325888	76001	chr13	57441518	57482758	41240
chr12	14031957	14066589	34632	chr12	88366768	88380441	13673	chr13	58427774	58477654	49880
chr12	16094359	16129188	34829	chr12	88752195	88836752	84557	chr13	58581026	58701046	120020
chr12	16937845	16939125	1280	chr12	90450340	90467758	17418	chr13	58802233	58887272	85039
chr12	17172460	17308983	136523	chr12	92642091	92652842	10751	chr13	59121803	59212679	90876
chr12	17765463	17794760	29297	chr12	94700562	94722001	21439	chr13	59313391	59430069	116678
chr12	17835492	17950268	114776	chr12	94762887	94784146	21259	chr13	59546803	59615582	68779
chr12	17991018	18030868	39850	chr12	102613491	102631581	18090	chr13	61091768	61149797	58029
chr12	18887878	18946395	58517	chr12	105477017	105514885	37868	chr13	61249880	61274688	24808
chr12	18987229	18997073	9844	chr12	108014562	108015750	1188	chr13	61577946	61672979	95033
chr12	23401499	23479499	78000	chr12	108056548	108079470	22922	chr13	62481515	62522284	40769
chr12	24037334	24063255	25921	chr13	20346005	20353666	7661	chr13	63478094	63517680	39586
chr12	28631511	28671974	40463	chr13	22426524	22566319	139795	chr13	64226011	64380109	154098
chr12	29854842	30050982	196140	chr13	22607091	22613952	6861	chr13	64420871	64662007	241136
chr12	30542302	30578987	36685	chr13	26455238	26456543	1305	chr13	64762072	64908096	146024
chr12	38379728	38394176	14448	chr13	26810785	26855361	44576	chr13	65009396	65080215	70819
chr12	39242874	39243227	353	chr13	29645688	29683397	37709	chr13	65120981	65259854	138873
chr12	58294865	58344595	49730	chr13	31587999	31648309	60310	chr13	65360953	65716173	355220
chr12	59862576	59942863	80287	chr13	31689151	31689541	390	chr13	66028219	66153870	125651
chr12	60266384	60291823	25439	chr13	36540441	36558115	17674	chr13	67529976	67694296	164320
chr12	60502439	60797346	294907	chr13	36598889	36623911	25022	chr13	67735064	67740406	5342
chr12	60838088	61016401	178313	chr13	37296718	37296818	100	chr13	67961951	68029024	67073
chr12	61057153	61181158	124005	chr13	39907227	39929105	21878	chr13	68069790	68239678	169888
chr12	61282937	61450148	167211	chr13	42419657	42419778	121	chr13	68381613	68584189	202576
chr12	61490900	61550066	59166	chr13	42460658	42512735	52077	chr13	69043573	69051831	8258

chr13	69472101	69490952	18851	chr13	86552979	86563494	10515	chr14	42858112	42939200	81088
chr13	69531728	69650593	118865	chr13	86604252	86675499	71247	chr14	43039483	43224823	185340
chr13	70289429	70409463	120034	chr13	87087108	87169003	81895	chr14	43383632	43490882	107250
chr13	70614437	70851150	236713	chr13	87269269	87277199	7930	chr14	43646683	43845780	199097
chr13	71347235	71387965	40730	chr13	87981545	87992866	11321	chr14	46701821	46727786	25965
chr13	72028206	72050896	22690	chr13	88386082	88390863	4781	chr14	46768530	46789628	21098
chr13	72091692	72403901	312209	chr13	88780397	88843424	63027	chr14	47945092	47974327	29235
chr13	72510601	72517212	6611	chr13	88884182	88928446	44264	chr14	48096216	48246507	150291
chr13	72636810	72653692	16882	chr13	89028785	89064846	36061	chr14	48641767	48769359	127592
chr13	74715445	74778552	63107	chr13	89719230	89733882	14652	chr14	48869593	48907712	38119
chr13	74880080	75040075	159995	chr13	89887237	89910246	23009	chr14	49227505	49323965	96460
chr13	76185225	76192521	7296	chr13	90281682	90343286	61604	chr14	49424383	49436578	12195
chr13	76233313	76488512	255199	chr13	90685080	90731765	46685	chr14	54041995	54156477	114482
chr13	76529316	76570483	41167	chr13	103197632	103275199	77567	chr14	54197355	54320708	123353
chr13	76636499	76728497	91998	chr13	103577685	103793592	215907	chr14	61133733	61137558	3825
chr13	77377050	77410884	33834	chr13	103834366	104546463	712097	chr14	62289973	62405414	115441
chr13	78960542	78960600	58	chr13	104587247	104764326	177079	chr14	62446172	62464871	18699
chr13	79001360	79040782	39422	chr13	104895642	105060434	164792	chr14	65794121	65807296	13175
chr13	79238190	79256308	18118	chr13	105101234	105301794	200560	chr14	77985012	78009941	24929
chr13	80178283	80266345	88062	chr13	105947643	106078370	130727	chr14	82180349	82353254	172905
chr13	80419534	80569097	149563	chr13	106205490	106226562	21072	chr14	82480156	82492619	12463
chr13	80707027	80838347	131320	chr13	108023372	108081768	58396	chr14	83189602	83360867	171265
chr13	81452407	81457182	4775	chr13	108417734	108429428	11694	chr14	83401625	83700383	298758
chr13	81497874	81639910	142036	chr13	108531817	108546151	14334	chr14	84065050	84122840	57790
chr13	81741072	81746649	5577	chr13	109551641	109578281	26640	chr14	84227088	84321709	94621
chr13	81787399	82108887	321488	chr13	111394249	111403855	9606	chr14	84362479	84689941	327462
chr13	82149609	82310917	161308	chr14	25482234	25526429	44195	chr14	84790052	84853130	63078
chr13	82351667	82704995	353328	chr14	25567183	25670365	103182	chr14	84893892	84901273	7381
chr13	82745745	82773130	27385	chr14	26293305	26352198	58893	chr14	85049296	85052848	3552
chr13	82874136	83095539	221403	chr14	26392978	26393092	114	chr14	85704426	85717126	12700
chr13	83206604	83248070	41466	chr14	27895286	27940004	44718	chr14	85757898	85784709	26811
chr13	83348167	83503704	155537	chr14	28040115	28055168	15053	chr14	86551285	86755777	204492
chr13	83544448	83718958	174510	chr14	28095910	28168140	72230	chr14	97308736	97308815	79
chr13	83759700	83821010	61310	chr14	38461930	38504387	42457	chr15	37640808	37800591	159783
chr13	83957781	83970326	12545	chr14	38545177	38555254	10077	chr15	46150392	46291353	140961
chr13	84070762	84126859	56097	chr14	39708732	39901479	192747	chr15	46401746	46543287	141541
chr13	84167605	84175563	7958	chr14	39942237	40236251	294014	chr15	47007617	47078073	70456
chr13	84216307	84220027	3720	chr14	40552786	40624044	71258	chr15	49706232	49729603	23371
chr13	84318209	84412363	94154	chr14	40664806	40761335	96529	chr15	53279698	53296254	16556
chr13	84713236	84721753	8517	chr14	40802083	40804897	2814	chr15	53337014	53381959	44945
chr13	84762517	84915086	152569	chr14	41125877	41246285	120408	chr15	54683414	54731040	47626
chr13	85702170	85742789	40619	chr14	42002560	42096519	93959	chr15	54771768	54880721	108953

chr15	60268320	60297133	28813	chr17	55660938	55669626	8688	chr18	51793939	51795006	1067
chr15	68482162	68485076	2914	chr17	56645590	56673190	27600	chr18	51895628	51929966	34338
chr15	68525972	68528968	2996	chr17	56714138	56738879	24741	chr18	51970718	52040171	69453
chr15	72950260	73001709	51449	chr17	70518916	70578916	60000	chr18	53835903	53854401	18498
chr15	87197161	87234809	37648	chr17	70929230	70947774	18544	chr18	53895159	54036585	141426
chr15	87275603	87282057	6454	chr17	71352177	71387531	35354	chr18	56341262	56477100	135838
chr15	96601517	96622004	20487	chr17	71428351	71446337	17986	chr18	56517888	56547207	29319
chr15	96933312	97041261	107949	chr17	71747279	71794893	47614	chr18	61052726	61283581	230855
chr16	8044395	8126262	81867	chr17	71835735	71871850	36115	chr18	64589723	64627795	38072
chr16	15307020	15308977	1957	chr18	2390847	2411171	20324	chr18	64732203	64747816	15613
chr16	19981202	19981484	282	chr18	4509458	4625053	115595	chr18	64788524	64955872	167348
chr16	25574169	25642025	67856	chr18	7351153	7452767	101614	chr18	66119647	66293770	174123
chr16	26879126	26916706	37580	chr18	7493531	7516781	23250	chr18	66334502	66451082	116580
chr16	48900015	48951036	51021	chr18	22499352	22536464	37112	chr18	66703501	66822483	118982
chr16	48991794	49009507	17713	chr18	25652190	25666386	14196	chr18	66863227	66896115	32888
chr16	51521232	51595755	74523	chr18	25707176	25716252	9076	chr18	66996222	67031583	35361
chr16	52837005	52884812	47807	chr18	27745164	27804518	59354	chr18	67187751	67246281	58530
chr16	54520699	54584757	64058	chr18	28321109	28601731	280622	chr18	70800744	70850476	49732
chr16	54625569	54697246	71677	chr18	28953393	28998619	45226	chr18	72094577	72315145	220568
chr16	55104665	55106644	1979	chr18	29698241	29850144	151903	chr18	72355877	72486679	130802
chr16	59258974	59355627	96653	chr18	29978708	30288909	310201	chr18	73499878	73537294	37416
chr16	59396383	59490125	93742	chr18	30329649	30563274	233625	chr18	77774515	77813463	38948
chr16	59590266	59605601	15335	chr18	34378791	34443289	64498	chr18	78328332	78330548	2216
chr16	60203973	60209841	5868	chr18	37723176	37752988	29812	chr19	56998556	57015333	16777
chr16	60706649	60719978	13329	chr18	37793742	37838050	44308	chr2	2477110	2491988	14878
chr16	60822775	60875732	52957	chr18	37878792	38087046	208254	chr2	4318190	4464203	146013
chr16	60978541	60994187	15646	chr18	38127786	38361247	233461	chr2	4877084	4895276	18192
chr16	61118418	61281959	163541	chr18	38401995	38505208	103213	chr2	4995383	5263739	268356
chr16	61322713	61483039	160326	chr18	38837998	38966046	128048	chr2	5364134	5399779	35645
chr16	61523777	61541755	17978	chr18	40249233	40262826	13593	chr2	6151275	6156750	5475
chr16	62090697	62137539	46842	chr18	40303596	40484833	181237	chr2	13157013	13251173	94160
chr16	62876048	62906785	30737	chr18	40525563	40791328	265765	chr2	13291919	13387672	95753
chr16	66013784	66251808	238024	chr18	40832070	41066819	234749	chr2	15057664	15116908	59244
chr17	14571026	14575728	4702	chr18	41107571	41215390	107819	chr2	17173140	17234291	61151
chr17	52210017	52240514	30497	chr18	41256132	41315782	59650	chr2	17436417	17460583	24166
chr17	52685701	52712120	26419	chr18	41869421	41902704	33283	chr2	22688288	22811551	123263
chr17	53192892	53303426	110534	chr18	43165691	43184864	19173	chr2	35335689	35421404	85715
chr17	53587604	53631629	44025	chr18	43327650	43348851	21201	chr2	35521841	35674758	152917
chr17	53875213	53902935	27722	chr18	43649836	43782158	132322	chr2	35775142	35859017	83875
chr17	54005324	54188536	183212	chr18	43882265	43989470	107205	chr2	35899783	36032833	133050
chr17	54229310	54427795	198485	chr18	44121701	44173435	51734	chr2	36073581	36127654	54073
chr17	54528168	54624908	96740	chr18	50508795	50510077	1282	chr2	40917452	41093779	176327

chr2	41207555	41409362	201807	chr2	88236636	88244591	7955	chr2	137798620	137882679	84059
chr2	41450144	41546385	96241	chr2	101636822	101646849	10027	chr2	138952747	139216164	263417
chr2	41646492	41727073	80581	chr2	103326260	103444240	117980	chr2	142431701	142493417	61716
chr2	49745122	49787429	42307	chr2	103673830	103706658	32828	chr2	142593975	142651522	57547
chr2	49828195	49828622	427	chr2	105538880	105574913	36033	chr2	142692282	142782218	89936
chr2	53060020	53372418	312398	chr2	105626192	105657411	31219	chr2	146244821	146427426	182605
chr2	56536171	56620202	84031	chr2	107013674	107104690	91016	chr2	146639310	146687748	48438
chr2	56660940	56700299	39359	chr2	114311226	114370124	58898	chr2	147048218	147141228	93010
chr2	56800563	56825336	24773	chr2	115895752	115943905	48153	chr2	147182024	147358341	176317
chr2	56866094	56998349	132255	chr2	115984649	116017095	32446	chr2	150009191	150021422	12231
chr2	57129606	57142735	13129	chr2	116057819	116326486	268667	chr2	150785348	150832505	47157
chr2	57277535	57379547	102012	chr2	116367222	116637217	269995	chr2	152927767	153021986	94219
chr2	57480693	57494534	13841	chr2	116868761	116873840	5079	chr2	153812646	153821912	9266
chr2	57597597	57658431	60834	chr2	117231562	117263174	31612	chr2	155139538	155213457	73919
chr2	57699219	57705427	6208	chr2	117303924	117512754	208830	chr2	155380527	155419025	38498
chr2	57856010	57857650	1640	chr2	117553498	117607562	54064	chr2	155459783	155861616	401833
chr2	71736768	71831032	94264	chr2	118336386	118387391	51005	chr2	156663735	156727705	63970
chr2	71871940	71934181	62241	chr2	118428207	118581111	152904	chr2	156828372	156882167	53795
chr2	72062103	72077774	15671	chr2	118634952	118683699	48747	chr2	157065257	157160846	95589
chr2	75870018	75945202	75184	chr2	121946702	121990535	43833	chr2	160543794	160699312	155518
chr2	75985958	76056015	70057	chr2	122031355	122316710	285355	chr2	162947972	163102250	154278
chr2	76157324	76208033	50709	chr2	122357448	122524046	166598	chr2	163203307	163205272	1965
chr2	76311661	76395078	83417	chr2	122564794	122789586	224792	chr2	163339096	163543395	204299
chr2	76495410	76545412	50002	chr2	122830346	122915320	84974	chr2	166574772	166632704	57932
chr2	76645678	76697718	52040	chr2	123223481	123293265	69784	chr2	166673504	166766100	92596
chr2	78277806	78362792	84986	chr2	123594445	123618804	24359	chr2	166806854	166838486	31632
chr2	78749406	78832446	83040	chr2	125039325	125063312	23987	chr2	168298141	168331519	33378
chr2	78932734	78975685	42951	chr2	125177154	125285389	108235	chr2	168372321	168401323	29002
chr2	81020190	81110775	90585	chr2	125326117	125340095	13978	chr2	175604640	175685068	80428
chr2	81251184	81278137	26953	chr2	125409658	125511662	102004	chr2	175725878	175747335	21457
chr2	81716728	81824150	107422	chr2	125552384	125560968	8584	chr2	180057113	180329187	272074
chr2	82155700	82218426	62726	chr2	125915842	125960098	44256	chr2	180369993	180421711	51718
chr2	82688711	82764983	76272	chr2	128731576	128733431	1855	chr2	182573192	182599021	25829
chr2	82907384	82943946	36562	chr2	128774259	128844756	70497	chr2	182639825	182666040	26215
chr2	82984704	83168889	184185	chr2	128887115	128913845	26730	chr2	183395652	183557441	161789
chr2	83269105	83367962	98857	chr2	129423848	129446296	22448	chr2	183658119	183754528	96409
chr2	83722912	83741214	18302	chr2	129572394	129605793	33399	chr2	184126147	184328974	202827
chr2	83791812	83860186	68374	chr2	129646593	129719192	72599	chr2	184429044	184548365	119321
chr2	83900940	83981139	80199	chr2	133708227	133904696	196469	chr2	184989492	185243498	254006
chr2	84090720	84117065	26345	chr2	133945522	134069982	124460	chr2	185284264	185497134	212870
chr2	84157829	84165107	7278	chr2	136440669	136494370	53701	chr2	189361967	189391432	29465
chr2	87876489	87878271	1782	chr2	136695192	136715544	20352	chr2	191501092	191528807	27715

chr2	191569695	191628067	58372	chr20	55125140	55258897	133757	chr3	26863364	26889059	25695
chr2	192245709	192387721	142012	chr20	55655215	55671852	16637	chr3	26929753	27040591	110838
chr2	192428475	192479918	51443	chr20	55712620	55933702	221082	chr3	28178604	28191583	12979
chr2	192926899	193038100	111201	chr20	56055472	56106737	51265	chr3	28908340	28961241	52901
chr2	193078850	193206482	127632	chr20	56881460	56883613	2153	chr3	30084156	30088247	4091
chr2	193327614	193454889	127275	chr20	56924497	56950086	25589	chr3	30189033	30254319	65286
chr2	193495631	193680554	184923	chr21	17023691	17160468	136777	chr3	30994142	31111703	117561
chr2	194078164	194079353	1189	chr21	19453739	19543840	90101	chr3	31319406	31402831	83425
chr2	195176370	195298531	122161	chr21	19677962	19689708	11746	chr3	33947681	34009333	61652
chr2	200061159	200118358	57199	chr21	20049755	20106751	56996	chr3	34712877	34822640	109763
chr2	200159196	200255880	96684	chr21	20480762	20547952	67190	chr3	34937040	34948862	11822
chr2	208970664	209022596	51932	chr21	22670058	22832581	162523	chr3	34989574	35125135	135561
chr2	209232018	209249794	17776	chr21	23640724	23702283	61559	chr3	35326882	35329835	2953
chr2	209349862	209374057	24195	chr21	26727697	26739375	11678	chr3	35370571	35444459	73888
chr2	210800493	211054933	254440	chr21	27089742	27093343	3601	chr3	36044577	36102073	57496
chr2	220947893	221031427	83534	chr21	27193949	27208884	14935	chr3	36220448	36330343	109895
chr2	221213053	221326149	113096	chr21	30931555	30941663	10108	chr3	39579479	39593637	14158
chr2	221366897	221368026	1129	chr21	30982419	30988158	5739	chr3	39694705	39758913	64208
chr2	221791813	221868162	76349	chr21	40914473	40976395	61922	chr3	45305620	45321503	15883
chr2	224117706	224119663	1957	chr22	34359262	34432605	73343	chr3	66687412	66782220	94808
chr2	225131858	225249709	117851	chr22	34473363	34491178	17815	chr3	66822992	66900619	77627
chr2	225912112	225930297	18185	chr22	49048386	49114200	65814	chr3	66941383	66948306	6923
chr2	226335371	226357843	22472	chr3	1454217	1500781	46564	chr3	67100510	67146299	45789
chr2	226398699	226595983	197284	chr3	1541537	1545776	4239	chr3	70462726	70610236	147510
chr2	228266700	228302119	35419	chr3	1655855	1680069	24214	chr3	70650982	70654692	3710
chr2	235328905	235355750	26845	chr3	1780474	1812380	31906	chr3	71883989	71885299	1310
chr20	6278948	6296722	17774	chr3	5406761	5741809	335048	chr3	73291896	73332432	40536
chr20	6886136	6919613	33477	chr3	5782557	5812841	30284	chr3	74186892	74190829	3937
chr20	7517632	7583561	65929	chr3	7791533	7802804	11271	chr3	74584871	74806797	221926
chr20	7746977	7781797	34820	chr3	13123117	13176688	53571	chr3	74847533	75034154	186621
chr20	12068677	12093894	25217	chr3	13217582	13249130	31548	chr3	75265636	75278548	12912
chr20	12466485	12521563	55078	chr3	15929571	15975750	46179	chr3	78444731	78475100	30369
chr20	12562313	12699417	137104	chr3	16016640	16119549	102909	chr3	79817815	79904288	86473
chr20	16121643	16173749	52106	chr3	19070401	19098453	28052	chr3	80008142	80038727	30585
chr20	17033033	17137509	104476	chr3	19585646	19700692	115046	chr3	80079455	80113326	33871
chr20	20788731	20804120	15389	chr3	19741444	19829471	88027	chr3	80267234	80392080	124846
chr20	39403023	39594216	191193	chr3	22432929	22606803	173874	chr3	80492305	80574000	81695
chr20	39635016	39635720	704	chr3	23041582	23045069	3487	chr3	82613675	82622867	9192
chr20	40276357	40322813	46456	chr3	26066369	26147819	81450	chr3	82663605	82756514	92909
chr20	54300327	54324810	24483	chr3	26266877	26296680	29803	chr3	82858021	83313380	455359
chr20	54701171	54809687	108516	chr3	26420318	26426470	6152	chr3	83354112	83887044	532932
chr20	54909812	55014224	104412	chr3	26467190	26469329	2139	chr3	83988236	84024941	36705

chr3	84065691	84240111	174420	chr3	145082088	145473537	391449	chr4	26019550	26061864	42314
chr3	84340253	84421413	81160	chr3	145574415	145634571	60156	chr4	27432225	27535144	102919
chr3	86646996	86690034	43038	chr3	145735092	145774256	39164	chr4	27639900	27817722	177822
chr3	86730806	86766739	35933	chr3	146740325	146759684	19359	chr4	28135063	28212278	77215
chr3	86867036	86887968	20932	chr3	147641835	147694122	52287	chr4	28761476	28761499	23
chr3	87394015	87449540	55525	chr3	152969796	152977250	7454	chr4	29517850	29625988	108138
chr3	94300000	94417255	117255	chr3	157727749	157826560	98811	chr4	29666718	29698756	32038
chr3	94557620	94788171	230551	chr3	161607442	161666908	59466	chr4	29800310	29857658	57348
chr3	95343238	95604422	261184	chr3	161971908	162059899	87991	chr4	30058316	30445181	386865
chr3	95733193	95997862	264669	chr3	162100645	162436540	335895	chr4	30485913	30626256	140343
chr3	96038594	96116812	78218	chr3	162536949	162551626	14677	chr4	32503220	32533016	29796
chr3	96295335	96299553	4218	chr3	162651792	162684758	32966	chr4	33139016	33188238	49222
chr3	96410039	96469868	59829	chr3	162725552	162757340	31788	chr4	33587877	33626346	38469
chr3	96510618	96516356	5738	chr3	162798092	162803261	5169	chr4	34452621	34507605	54984
chr3	99266266	99276376	10110	chr3	163629365	163675565	46200	chr4	34819432	34916240	96808
chr3	101048966	101090109	41143	chr3	163716313	163951605	235292	chr4	35020772	35225900	205128
chr3	102823913	102836123	12210	chr3	164052465	164121470	69005	chr4	35266638	35279259	12621
chr3	102876879	103109959	233080	chr3	164221556	164510902	289346	chr4	35319967	35437811	117844
chr3	103291230	103474032	182802	chr3	165887472	165923051	35579	chr4	35546003	35559041	13038
chr3	103682337	104004707	322370	chr3	165963821	166110020	146199	chr4	35599773	35898220	298447
chr3	104310751	104452699	141948	chr3	166231820	166243755	11935	chr4	36791900	36851815	59915
chr3	104553266	104692179	138913	chr3	166344353	166403764	59411	chr4	43648844	43791974	143130
chr3	104817388	105057658	240270	chr3	167127968	167190286	62318	chr4	45201652	45273252	71600
chr3	105157756	105316908	159152	chr3	176164537	176430381	265844	chr4	45530136	45734030	203894
chr3	106169552	106275736	106184	chr3	180087053	180111885	24832	chr4	45774754	45945118	170364
chr3	106376293	106465896	89603	chr3	180263108	180264173	1065	chr4	46585213	46673829	88616
chr3	106565998	106622905	56907	chr3	190942429	191022620	80191	chr4	52221075	52390493	169418
chr3	110027767	110261912	234145	chr3	191740356	191844429	104073	chr4	57874655	58005165	130510
chr3	110302684	110377481	74797	chr4	10937739	11233990	296251	chr4	58045913	58051571	5658
chr3	115435347	115508532	73185	chr4	11274736	11318826	44090	chr4	58168070	58222809	54739
chr3	117189389	117195940	6551	chr4	12069688	12073450	3762	chr4	58263557	58351258	87701
chr3	117236680	117255611	18931	chr4	12401286	12487904	86618	chr4	59326146	59401141	74995
chr3	117296405	117494204	197799	chr4	12528612	12589085	60473	chr4	59942114	59988268	46154
chr3	118707668	118793072	85404	chr4	12691009	12709100	18091	chr4	60088586	60423724	335138
chr3	135589888	135715523	125635	chr4	13098670	13283413	184743	chr4	60464476	60600585	136109
chr3	135756305	135875890	119585	chr4	17370882	17377695	6813	chr4	60972684	60993655	20971
chr3	137132196	137455880	323684	chr4	18071876	18267885	196009	chr4	62342337	62348324	5987
chr3	137586142	137621948	35806	chr4	18308625	18338061	29436	chr4	62389068	62439494	50426
chr3	144098719	144136904	38185	chr4	18639508	18943516	304008	chr4	62645462	62766825	121363
chr3	144237020	144292210	55190	chr4	20087972	20177548	89576	chr4	62868794	62978310	109516
chr3	144595844	144848538	252694	chr4	23954089	24114582	160493	chr4	63293881	63300113	6232
chr3	144889286	145041354	152068	chr4	24155388	24371175	215787	chr4	63644278	63801790	157512

chr4	63842526	64115500	272974	chr4	126222004	126513508	291504	chr4	176470458	176487450	16992
chr4	64156254	64225256	69002	chr4	126614776	126682714	67938	chr4	177030605	177030882	277
chr4	64483215	64556417	73202	chr4	126723508	126893429	169921	chr4	177071686	177092538	20852
chr4	64695525	64717129	21604	chr4	129455022	129574170	119148	chr4	178196844	178236694	39850
chr4	65267433	65269562	2129	chr4	130105368	130226228	120860	chr4	178337285	178351454	14169
chr4	66576941	66847261	270320	chr4	130554344	130658506	104162	chr4	178456830	178556086	99256
chr4	66947371	67053523	106152	chr4	131035641	131055836	20195	chr4	178596846	178635885	39039
chr4	67094295	67267304	173009	chr4	131171976	131223241	51265	chr4	178770774	179019436	248662
chr4	71109321	71137285	27964	chr4	133299116	133510855	211739	chr4	179119594	179239133	119539
chr4	71622087	71636498	14411	chr4	133625208	133658225	33017	chr4	179615305	179806852	191547
chr4	71973980	71975889	1909	chr4	134607852	134609398	1546	chr4	179907581	180488011	580430
chr4	72618799	72719551	100752	chr4	135273779	135321176	47397	chr4	180909038	180914088	5050
chr4	72858192	72915177	56985	chr4	135421924	135515116	93192	chr4	181415029	181472665	57636
chr4	85161277	85196156	34879	chr4	135617157	135642383	25226	chr4	181573001	181670018	97017
chr4	85296912	85425113	128201	chr4	137362799	137375425	12626	chr4	183802847	183803430	583
chr4	89358010	89401355	43345	chr4	141787842	141894067	106225	chr4	187854609	187920272	65663
chr4	90004629	90018098	13469	chr4	141934845	141973159	38314	chr5	2331966	2334507	2541
chr4	90058878	90077534	18656	chr4	144277425	144355899	78474	chr5	2375323	2393165	17842
chr4	92054335	92110366	56031	chr4	147180072	147255486	75414	chr5	2433931	2507656	73725
chr4	93880964	93934000	53036	chr4	153923451	153992121	68670	chr5	2548446	2562590	14144
chr4	95715032	95790018	74986	chr4	154931873	154995845	63972	chr5	3769789	3862707	92918
chr4	95918764	96069139	150375	chr4	156056988	156120966	63978	chr5	5566044	5869028	302984
chr4	96968864	96970700	1836	chr4	156161746	156246560	84814	chr5	12204043	12225343	21300
chr4	103774534	103811615	37081	chr4	156346646	156435978	89332	chr5	12347504	12392204	44700
chr4	106603448	106760090	156642	chr4	157416075	157422489	6414	chr5	13356366	13454346	97980
chr4	107485220	107508869	23649	chr4	159955124	160082155	127031	chr5	14064557	14093701	29144
chr4	110995799	111013582	17783	chr4	160122911	160188736	65825	chr5	18105857	18128629	22772
chr4	111114227	111252705	138478	chr4	160229510	160389253	159743	chr5	18169379	18186122	16743
chr4	111381518	111490494	108976	chr4	160736817	160768421	31604	chr5	18286196	18497311	211115
chr4	114154225	114356742	202517	chr4	160868869	160953876	85007	chr5	19292346	19348800	56454
chr4	115193303	115578811	385508	chr4	161053958	161193979	140021	chr5	19389544	19422950	33406
chr4	115709755	115714098	4343	chr4	162223684	162272187	48503	chr5	21087691	21142474	54783
chr4	116665146	116671495	6349	chr4	162373907	162471681	97774	chr5	23164527	23212158	47631
chr4	116712251	116789278	77027	chr4	162572041	162590667	18626	chr5	23386237	23393585	7348
chr4	116908973	117033553	124580	chr4	162908393	162934794	26401	chr5	23578597	23610626	32029
chr4	117135576	117164596	29020	chr4	164464722	164488905	24183	chr5	23739386	23801347	61961
chr4	121501967	121506685	4718	chr4	164529711	164547718	18007	chr5	27171150	27235023	63873
chr4	121547465	121614584	67119	chr4	166153895	166177764	23869	chr5	27646401	27680442	34041
chr4	122022848	122027839	4991	chr4	166676119	166683383	7264	chr5	27721204	27936174	214970
chr4	124762732	124896927	134195	chr4	167458789	167598429	139640	chr5	28036285	28136388	100103
chr4	124937733	125016398	78665	chr4	167792287	167796673	4386	chr5	28437665	28574656	136991
chr4	125902793	125904779	1986	chr4	175129381	175196357	66976	chr5	29650868	29745937	95069

chr5	30298893	30315404	16511	chr5	99727957	99784671	56714	chr5	144838336	144934707	96371
chr5	30415684	30450614	34930	chr5	99825429	99898294	72865	chr5	145055789	145072609	16820
chr5	30491390	30943976	452586	chr5	100203779	100251440	47661	chr5	153373543	153410582	37039
chr5	34308767	34315742	6975	chr5	100953266	100988158	34892	chr5	155162324	155347290	184966
chr5	34423850	34476153	52303	chr5	101028876	101238711	209835	chr5	161109279	161180049	70770
chr5	35280589	35281708	1119	chr5	101338821	101513433	174612	chr5	162205536	162274377	68841
chr5	35519379	35528483	9104	chr5	101632128	101766474	134346	chr5	162812313	162862264	49951
chr5	39771513	39781461	9948	chr5	101866780	101926295	59515	chr5	163644058	163659209	15151
chr5	39822291	39838575	16284	chr5	103691985	103929910	237925	chr5	163937827	163963507	25680
chr5	40417657	40604962	187305	chr5	105542970	105661169	118199	chr5	165955608	165978497	22889
chr5	41643224	41680064	36840	chr5	105701919	105872993	171074	chr5	166305439	166329123	23684
chr5	42022241	42028383	6142	chr5	105973108	106057621	84513	chr5	166432599	166737416	304817
chr5	42341523	42373776	32253	chr5	106098375	106312893	214518	chr5	175145513	175200265	54752
chr5	44116731	44121372	4641	chr5	106353625	106365575	11950	chr5	175341252	175390038	48786
chr5	44162148	44238731	76583	chr5	106467241	106493065	25824	chr6	8152578	8179654	27076
chr5	51700000	51712176	12176	chr5	106596803	106665196	68393	chr6	8935445	8974220	38775
chr5	51752936	51813991	61055	chr5	107245332	107326888	81556	chr6	9311777	9546109	234332
chr5	52230987	52232566	1579	chr5	108432098	108537055	104957	chr6	15713058	15844943	131885
chr5	52332859	52406378	73519	chr5	110034751	110132157	97406	chr6	16952302	17004584	52282
chr5	52447134	52525192	78058	chr5	110172953	110239232	66279	chr6	17181372	17183461	2089
chr5	53265126	53352487	87361	chr5	110815161	110817992	2831	chr6	17224271	17231345	7074
chr5	58277953	58368299	90346	chr5	113645285	113906005	260720	chr6	18621880	18686170	64290
chr5	62036129	62075014	38885	chr5	114303754	114310944	7190	chr6	18986242	19057280	71038
chr5	62196946	62198958	2012	chr5	114818410	114881272	62862	chr6	19991329	19992454	1125
chr5	62239748	62256161	16413	chr5	117188569	117251103	62534	chr6	23181780	23187710	5930
chr5	62678582	62726876	48294	chr5	119754048	120095447	341399	chr6	23496560	23599495	102935
chr5	62827263	62965255	137992	chr5	120970837	120990089	19252	chr6	23699774	23739576	39802
chr5	63073489	63245463	171974	chr5	121515314	121525847	10533	chr6	23780340	23804443	24103
chr5	63352056	63678541	326485	chr5	121721872	121801954	80082	chr6	23906818	23921878	15060
chr5	63719267	63807892	88625	chr5	123286867	123294884	8017	chr6	24052433	24076121	23688
chr5	71881410	71998538	117128	chr5	123716806	123897334	180528	chr6	40066342	40121565	55223
chr5	72039390	72057233	17843	chr5	125752227	125812745	60518	chr6	40923167	40942754	19587
chr5	84679015	84949811	270796	chr5	127229900	127240830	10930	chr6	44703281	44759316	56035
chr5	85999199	86037525	38326	chr5	129164028	129234960	70932	chr6	45745212	45848450	103238
chr5	86137847	86155232	17385	chr5	129275718	129309558	33840	chr6	47092363	47118964	26601
chr5	86531426	86567903	36477	chr5	130327693	130336620	8927	chr6	47159904	47181531	21627
chr5	87902908	88118894	215986	chr5	130436680	130440049	3369	chr6	48264794	48368502	103708
chr5	91589085	91793686	204601	chr5	130648779	130714443	65664	chr6	48409266	48651490	242224
chr5	91894529	91932596	38067	chr5	130764499	130935678	171179	chr6	48751872	48804320	52448
chr5	98461792	98526340	64548	chr5	131058742	131109026	50284	chr6	48845116	48902069	56953
chr5	99145013	99209933	64920	chr5	144009884	144090878	80994	chr6	49002781	49027711	24930
chr5	99250735	99399431	148696	chr5	144585670	144797592	211922	chr6	49195664	49223599	27935

chr6	50249282	50364034	114752	chr6	86542993	86679707	136714	chr6	119424062	119435325	11263
chr6	50963256	51262582	299326	chr6	86818932	86844392	25460	chr6	119502017	119595033	93016
chr6	51303354	51335281	31927	chr6	87963067	87987725	24658	chr6	119635837	119715159	79322
chr6	54535850	54552486	16636	chr6	90637045	90928534	291489	chr6	120315277	120334808	19531
chr6	54675488	54690117	14629	chr6	90969312	91240312	271000	chr6	120461606	120476293	14687
chr6	63033170	63143071	109901	chr6	91866428	91948745	82317	chr6	120576400	120636620	60220
chr6	65888039	66043430	155391	chr6	92064734	92069422	4688	chr6	120831512	120930059	98547
chr6	66144909	66327791	182882	chr6	92169633	92237840	68207	chr6	120970831	121029493	58662
chr6	66368553	66660241	291688	chr6	92538827	92573002	34175	chr6	121734272	121808178	73906
chr6	66778904	66785752	6848	chr6	94587706	94720790	133084	chr6	121908768	122161647	252879
chr6	66826542	67099508	272966	chr6	95370790	95425182	54392	chr6	122261811	122286788	24977
chr6	67260480	67292395	31915	chr6	95727450	95865096	137646	chr6	122920704	122945970	25266
chr6	67333171	67406345	73174	chr6	95905848	95941106	35258	chr6	123122927	123124309	1382
chr6	67675464	67737645	62181	chr6	98549872	98679966	130094	chr6	123165051	123166338	1287
chr6	69439511	69478654	39143	chr6	99002922	99032448	29526	chr6	126411618	126434630	23012
chr6	69519436	69571380	51944	chr6	99739898	99747293	7395	chr6	126733847	126816611	82764
chr6	71007038	71024679	17641	chr6	99788061	99843943	55882	chr6	128820674	128833140	12466
chr6	71708398	71794120	85722	chr6	101032987	101114808	81821	chr6	129911547	129919526	7979
chr6	71834886	71836702	1816	chr6	101155590	101348787	193197	chr6	129960338	129963698	3360
chr6	72453143	72472640	19497	chr6	102129555	102378087	248532	chr6	130493063	130523995	30932
chr6	74907925	75034325	126400	chr6	102503792	102778885	275093	chr6	131333535	131405418	71883
chr6	76142940	76213667	70727	chr6	102878976	102891013	12037	chr6	131519536	131523143	3607
chr6	76254447	76372450	118003	chr6	102931773	102952513	20740	chr6	137269449	137337878	68429
chr6	76925860	77191704	265844	chr6	103053897	103180016	126119	chr6	137378748	137442201	63453
chr6	77232498	77293556	61058	chr6	103220766	103287807	67041	chr6	140255326	140280210	24884
chr6	77393654	77412204	18550	chr6	103333588	103435917	102329	chr6	140321006	140477012	156006
chr6	77546671	77583732	37061	chr6	103476661	103533134	56473	chr6	140517896	140610622	92726
chr6	77624514	77640657	16143	chr6	103633268	103967632	334364	chr6	141070026	141132812	62786
chr6	77976974	78285060	308086	chr6	104076763	104572363	495600	chr6	141173644	141297010	123366
chr6	78325830	78454466	128636	chr6	104613203	104637240	24037	chr6	141705976	141819882	113906
chr6	80649370	80908496	259126	chr6	105453084	105462666	9582	chr6	141860694	141938232	77538
chr6	80949286	80972631	23345	chr6	105782196	105786319	4123	chr6	144903034	145032722	129688
chr6	81013419	81317680	304261	chr6	112688473	112775938	87465	chr6	153497488	153615706	118218
chr6	81358472	81377101	18629	chr6	112876470	112921492	45022	chr6	153731553	153775767	44214
chr6	82420828	82480745	59917	chr6	113062876	113172165	109289	chr6	153816521	153888432	71911
chr6	82521567	82720910	199343	chr6	114685875	114728508	42633	chr6	154655954	154656642	688
chr6	82761692	82782600	20908	chr6	114916976	115308497	391521	chr6	155531183	155572809	41626
chr6	84859578	84916396	56818	chr6	115408752	115483541	74789	chr6	155705026	155760736	55710
chr6	85016625	85086501	69876	chr6	115784191	115830509	46318	chr6	155801496	155896796	95300
chr6	85919464	85943020	23556	chr6	115871255	115881148	9893	chr6	156000910	156162866	161956
chr6	86264576	86382243	117667	chr6	116982163	116988299	6136	chr6	156203616	156328749	125133
chr6	86486138	86502227	16089	chr6	118367676	118368695	1019	chr6	163909841	163934022	24181

chr6	164589220	164677749	88529	chr7	52828972	52841836	12864	chr7	113601402	113698317	96915
chr6	164981630	165056634	75004	chr7	52947807	52950610	2803	chr7	113739083	113826776	87693
chr6	165167524	165215613	48089	chr7	53086924	53098029	11105	chr7	114782603	114853577	70974
chr6	169960102	170012516	52414	chr7	53238938	53266148	27210	chr7	115381355	115453366	72011
chr7	4319000	4412941	93941	chr7	57705392	57717855	12463	chr7	118546171	118811399	265228
chr7	4453707	4461958	8251	chr7	62806779	62894826	88047	chr7	119001259	119104011	102752
chr7	4562033	4617022	54989	chr7	67490024	67541057	51033	chr7	119229149	119336389	107240
chr7	6989603	7016610	27007	chr7	67847029	67870252	23223	chr7	119377091	119389543	12452
chr7	8802963	8804896	1933	chr7	68455723	68530628	74905	chr7	119430265	119434324	4059
chr7	8845616	8883074	37458	chr7	68571368	68590797	19429	chr7	119534409	119554555	20146
chr7	9339857	9400122	60265	chr7	68774048	68838357	64309	chr7	121491261	121493333	2072
chr7	9440840	9544616	103776	chr7	68879113	68890151	11038	chr7	123290461	123306628	16167
chr7	10167780	10172796	5016	chr7	70932540	70939928	7388	chr7	125616401	125680869	64468
chr7	10273011	10299819	26808	chr7	70980660	71082168	101508	chr7	125721613	125767870	46257
chr7	11882198	12001492	119294	chr7	79740199	79751344	11145	chr7	126090300	126166625	76325
chr7	12042260	12161240	118980	chr7	79795117	79862103	66986	chr7	126207357	126228969	21612
chr7	12804985	12826683	21698	chr7	79962208	80046599	84391	chr7	144900084	144901840	1756
chr7	15992360	16037526	45166	chr7	81825833	81859176	33343	chr7	144942556	144955057	12501
chr7	17110743	17129833	19090	chr7	82493798	82539847	46049	chr7	152911921	152983400	71479
chr7	20481747	20492669	10922	chr7	83714625	83752307	37682	chr7	153194588	153223089	28501
chr7	21271428	21325391	53963	chr7	83793063	83869601	76538	chr7	153578218	153698593	120375
chr7	21996084	22014374	18290	chr7	84734322	84769513	35191	chr7	153791306	153815801	24495
chr7	22055190	22068237	13047	chr7	85236855	85271121	34266	chr7	156116343	156136121	19778
chr7	24454328	24487331	33003	chr7	85686344	85754923	68579	chr8	2264124	2343875	79751
chr7	25034020	25059908	25888	chr7	85795677	85848647	52970	chr8	5192707	5201508	8801
chr7	31169572	31204131	34559	chr7	85889379	85955083	65704	chr8	5242246	5313370	71124
chr7	31244885	31264644	19759	chr7	85995813	86166784	170971	chr8	21600084	21640402	40318
chr7	32349329	32360362	11033	chr7	86267733	86330304	62571	chr8	23939487	23940050	563
chr7	32401114	32406962	5848	chr7	89644262	89692942	48680	chr8	23980894	24019956	39062
chr7	37499249	37535499	36250	chr7	94227157	94244340	17183	chr8	26188738	26204882	16144
chr7	41283507	41327447	43940	chr7	94285226	94344560	59334	chr8	26917273	26936075	18802
chr7	41368251	41517167	148916	chr7	96759891	96805140	45249	chr8	34496731	34572496	75765
chr7	42287870	42314453	26583	chr7	97341596	97378543	36947	chr8	34613222	34634051	20829
chr7	42355249	42511725	156476	chr7	97487896	97548932	61036	chr8	35039784	35104343	64559
chr7	45236302	45253619	17317	chr7	97649260	97651481	2221	chr8	36371404	36387390	15986
chr7	49080068	49080136	68	chr7	98529622	98567296	37674	chr8	41111472	41125114	13642
chr7	49404816	49610896	206080	chr7	104039516	104040064	548	chr8	48114708	48143849	29141
chr7	51439114	51464250	25136	chr7	109102666	109172319	69653	chr8	48849490	48863915	14425
chr7	51780870	51820334	39464	chr7	109814357	109909217	94860	chr8	48988934	49004310	15376
chr7	51920402	52015234	94832	chr7	110049624	110165890	116266	chr8	50910062	50965020	54958
chr7	52342913	52473400	130487	chr7	110266380	110282238	15858	chr8	51077250	51107687	30437
chr7	52514124	52788252	274128	chr7	111612517	111626459	13942	chr8	54919903	54985202	65299

chr8	56826874	56867083	40209	chr8	131596765	131662511	65746	chr9	29876711	30234997	358286
chr8	59271346	59296961	25615	chr8	131762980	131854087	91107	chr9	30882225	30885484	3259
chr8	59337783	59405884	68101	chr8	134052407	134061864	9457	chr9	31040572	31115617	75045
chr8	62194769	62198590	3821	chr8	135792495	135897020	104525	chr9	31556615	31594575	37960
chr8	70840371	70954781	114410	chr8	136316122	136339447	23325	chr9	31697139	32028225	331086
chr8	70995603	71005456	9853	chr8	137155458	137300849	145391	chr9	32068965	32172507	103542
chr8	75616843	75665409	48566	chr8	137341653	137374903	33250	chr9	32213243	32243557	30314
chr8	75706199	75736364	30165	chr8	137475021	137646723	171702	chr9	38197561	38210426	12865
chr8	75836460	76082780	246320	chr8	139280344	139310061	29717	chr9	71496904	71530659	33755
chr8	76123548	76215542	91994	chr8	141668737	141719409	50672	chr9	73220393	73241119	20726
chr8	77064028	77206049	142021	chr8	141760153	141792644	32491	chr9	73341227	73411490	70263
chr8	77246793	77249071	2278	chr8	141833534	141857872	24338	chr9	73801077	73820895	19818
chr8	77687290	77998100	310810	chr9	1214867	1278663	63796	chr9	73923297	74014495	91198
chr8	78990522	79047362	56840	chr9	1383953	1524026	140073	chr9	74055255	74213492	158237
chr8	79147453	79147716	263	chr9	1564758	1891899	327141	chr9	75248177	75304983	56806
chr8	82362904	82364567	1663	chr9	7246243	7326529	80286	chr9	75345775	75351888	6113
chr8	83150846	83150951	105	chr9	7367325	7427044	59719	chr9	75804738	75840643	35905
chr8	83558900	83645348	86448	chr9	7528320	7547822	19502	chr9	78380093	78539157	159064
chr8	83686086	83760726	74640	chr9	11115685	11125313	9628	chr9	78639539	78691719	52180
chr8	88958908	89100764	141856	chr9	11426314	11568495	142181	chr9	78792744	78985059	192315
chr8	89141544	89193400	51856	chr9	11668621	11948659	280038	chr9	79295674	79316958	21284
chr8	89293793	89362620	68827	chr9	12350451	12550099	199648	chr9	80184555	80307028	122473
chr8	97408478	97412034	3556	chr9	13637511	13659215	21704	chr9	80446964	80513570	66606
chr8	97452960	97525044	72084	chr9	13700045	13777970	77925	chr9	80618618	80684241	65623
chr8	97565882	97574202	8320	chr9	16117879	16153934	36055	chr9	81059553	81326120	266567
chr8	106820245	106930966	110721	chr9	17873829	17907452	33623	chr9	82794792	82929584	134792
chr8	107032796	107123199	90403	chr9	17948188	18057806	109618	chr9	84440131	84455045	14914
chr8	107548055	107597278	49223	chr9	18098548	18210596	112048	chr9	84495803	84554110	58307
chr8	110155964	110184742	28778	chr9	22364672	22396840	32168	chr9	85077070	85109681	32611
chr8	111906826	111922927	16101	chr9	22989185	23116051	126866	chr9	85209950	85266406	56456
chr8	113766958	113900885	133927	chr9	23156761	23332812	176051	chr9	85307188	85442578	135390
chr8	114000998	114064535	63537	chr9	23373558	23485191	111633	chr9	87346392	87373633	27241
chr8	114105327	114132066	26739	chr9	24048054	24235985	187931	chr9	88802160	88821637	19477
chr8	114445839	114449618	3779	chr9	24276735	24395951	119216	chr9	100637906	100681239	43333
chr8	114490410	114741652	251242	chr9	24741993	24822711	80718	chr9	100782355	100785716	3361
chr8	114841929	114938688	96759	chr9	24955735	25038810	83075	chr9	100826476	100862319	35843
chr8	114979496	115358495	378999	chr9	25155409	25605610	450201	chr9	101843756	102005251	161495
chr8	117366630	117470712	104082	chr9	26268408	26459107	190699	chr9	102105741	102108735	2994
chr8	121269754	121330136	60382	chr9	27660745	27679275	18530	chr9	102149501	102206603	57102
chr8	121430465	121489292	58827	chr9	29188976	29204074	15098	chr9	102345464	102369635	24171
chr8	128714679	128754772	40093	chr9	29368952	29586485	217533	chr9	102807893	102872573	64680
chr8	131482869	131555927	73058	chr9	29706885	29774741	67856	chr9	103579892	103639472	59580

chr9	103774248	103927566	153318	chrX	20609244	20767129	157885	chrX	66835072	66874970	39898
chr9	105147061	105194621	47560	chrX	20807865	20951123	143258	chrX	66915726	67039890	124164
chr9	105901425	105911353	9928	chrX	20991875	21158043	166168	chrX	67080636	67166512	85876
chr9	108408929	108551369	142440	chrX	21264743	21324417	59674	chrX	67207204	67236530	29326
chr9	109589950	109590787	837	chrX	23454372	23503771	49399	chrX	68892147	68980952	88805
chr9	115452644	115550741	98097	chrX	23603896	23614261	10365	chrX	69021872	69029556	7684
chr9	118133382	118321903	188521	chrX	25118625	25144380	25755	chrX	69359924	69389685	29761
chr9	118362681	118537751	175070	chrX	25279932	25327705	47773	chrX	73325515	73326380	865
chr9	118895312	119024380	129068	chrX	25368447	25541352	172905	chrX	74981824	74989402	7578
chr9	119065154	119103457	38303	chrX	25696327	25710003	13676	chrX	75635506	75657016	21510
chr9	119674125	119754190	80065	chrX	26043998	26065188	21190	chrX	76118667	76122935	4268
chr9	120143969	120194978	51009	chrX	26396242	26408336	12094	chrX	77313145	77324211	11066
chrX	3434653	3439804	5151	chrX	26840286	26866150	25864	chrX	78377691	78445173	67482
chrX	3480584	3482116	1532	chrX	28031355	28055741	24386	chrX	78604751	78606068	1317
chrX	4087855	4156772	68917	chrX	28096511	28161437	64926	chrX	78849296	78895331	46035
chrX	4197510	4268236	70726	chrX	28202199	28326962	124763	chrX	79012877	79034935	22058
chrX	4308946	4335637	26691	chrX	30067881	30093955	26074	chrX	79227668	79239922	12254
chrX	4376357	4419412	43055	chrX	30134699	30159422	24723	chrX	79420165	79439960	19795
chrX	4460160	4477199	17039	chrX	30365178	30407074	41896	chrX	79480702	79599475	118773
chrX	4783572	4852125	68553	chrX	30447830	30508823	60993	chrX	79640261	79713916	73655
chrX	4892895	4939845	46950	chrX	33482051	33530301	48250	chrX	79754718	79823682	68964
chrX	5000311	5004274	3963	chrX	34567358	34569667	2309	chrX	79864430	79904567	40137
chrX	5045054	5062351	17297	chrX	34855042	34869301	14259	chrX	80151308	80178488	27180
chrX	5234661	5249106	14445	chrX	35177750	35215260	37510	chrX	80986540	81011215	24675
chrX	5359974	5360121	147	chrX	35256014	35319974	63960	chrX	81051949	81063700	11751
chrX	6311412	6315122	3710	chrX	35360718	35392230	31512	chrX	81477276	81550071	72795
chrX	6906167	6939328	33161	chrX	35433016	35502183	69167	chrX	81675056	81746388	71332
chrX	7641199	7661548	20349	chrX	35542937	35544172	1235	chrX	81787154	81872539	85385
chrX	7716699	7750997	34298	chrX	36627491	36669917	42426	chrX	81913283	81951219	37936
chrX	7791721	7792261	540	chrX	38981436	39031179	49743	chrX	82051327	82306196	254869
chrX	11907718	11973846	66128	chrX	39584824	39611215	26391	chrX	82346938	82447668	100730
chrX	14157623	14194381	36758	chrX	40985864	41026998	41134	chrX	82618001	82625637	7636
chrX	14294928	14312114	17186	chrX	42882117	42899595	17478	chrX	82745216	82791682	46466
chrX	14352888	14396441	43553	chrX	43473492	43544224	70732	chrX	82931349	82948698	17349
chrX	14437215	14465263	28048	chrX	43585042	43591682	6640	chrX	83051035	83122532	71497
chrX	15033374	15062268	28894	chrX	46123935	46131267	7332	chrX	83235442	83311449	76007
chrX	15167768	15172192	4424	chrX	46172047	46175923	3876	chrX	83352227	83356022	3795
chrX	16351560	16385627	34067	chrX	54618560	54621984	3424	chrX	83659214	83698824	39610
chrX	17212514	17232636	20122	chrX	55368682	55401474	32792	chrX	83799939	83811125	11186
chrX	19172637	19212872	40235	chrX	56431127	56432514	1387	chrX	83936699	83953549	16850
chrX	19253642	19293892	40250	chrX	64784665	64837781	53116	chrX	83994305	84008345	14040
chrX	20502431	20568496	66065	chrX	66728578	66763120	34542	chrX	84582423	84635891	53468

chrX	84676643	84705135	28492	chrX	104304933	104324061	19128	chrX	127846733	127901293	54560
chrX	84827377	84884150	56773	chrX	104364867	104422693	57826	chrX	127942051	128000961	58910
chrX	85429743	85433273	3530	chrX	104478931	104490217	11286	chrX	128765790	128779194	13404
chrX	85474021	85551606	77585	chrX	109104590	109141428	36838	chrX	128888834	128910786	21952
chrX	85592358	85603608	11250	chrX	109182192	109256172	73980	chrX	128951532	128991535	40003
chrX	86938816	86986177	47361	chrX	110573021	110578181	5160	chrX	129093183	129108134	14951
chrX	87026953	87037216	10263	chrX	112148233	112166608	18375	chrX	129148912	129187703	38791
chrX	87196804	87236546	39742	chrX	112207348	112209492	2144	chrX	129247490	129306540	59050
chrX	87277298	87375582	98284	chrX	112564290	112580647	16357	chrX	129347278	129358381	11103
chrX	87907341	87988148	80807	chrX	113133867	113252899	119032	chrX	137081674	137274971	193297
chrX	88238653	88319342	80689	chrX	113293661	113297412	3751	chrX	137375049	137391257	16208
chrX	88439520	88480487	40967	chrX	113397619	113453959	56340	chrX	137737994	137758849	20855
chrX	88521267	88617358	96091	chrX	115357556	115368181	10625	chrX	137799621	138047924	248303
chrX	88658074	88697224	39150	chrX	116232207	116257779	25572	chrX	138088688	138299890	211202
chrX	89057485	89133724	76239	chrX	116900945	116913048	12103	chrX	138340672	138347350	6678
chrX	93162476	93163627	1151	chrX	117013676	117051406	37730	chrX	139287990	139312039	24049
chrX	93479686	93485123	5437	chrX	117092152	117290569	198417	chrX	141330360	141330822	462
chrX	93827618	93953814	126196	chrX	117331317	117347937	16620	chrX	141974140	141981244	7104
chrX	94053922	94128004	74082	chrX	117468882	117543007	74125	chrX	142255290	142314186	58896
chrX	94228106	94260440	32334	chrX	117583711	117589751	6040	chrX	142490110	142513047	22937
chrX	94370025	94586662	216637	chrX	117681258	117727215	45957	chrX	142553777	142556859	3082
chrX	94691794	94701178	9384	chrX	121255014	121282553	27539	chrX	142656945	142659373	2428
chrX	94969916	95013140	43224	chrX	121421053	121438696	17643	chrX	142700129	142876752	176623
chrX	95206004	95324289	118285	chrX	121525902	121538667	12765	chrX	142917488	142975917	58429
chrX	95365041	95403368	38327	chrX	121579437	121599376	19939	chrX	143978133	143997554	19421
chrX	95771144	95904266	133122	chrX	121640216	121692289	52073	chrX	144102222	144104085	1863
chrX	96260261	96263560	3299	chrX	122061612	122077130	15518	chrX	144775606	144779469	3863
chrX	96460782	96564210	103428	chrX	122392616	122421724	29108	chrX	144891660	145007107	115447
chrX	97792589	97797262	4673	chrX	122589460	122790425	200965	chrX	145383619	145396623	13004
chrX	98028641	98039208	10567	chrX	122894170	122967818	73648	chrX	145437377	145475008	37631
chrX	98079946	98169691	89745	chrX	125029921	125076835	46914	chrX	145534396	145538627	4231
chrX	99044930	99059389	14459	chrX	125117577	125149150	31573	chrX	146376553	146377455	902
chrX	99100133	99104084	3951	chrX	125408655	125513133	104478	chrX	147641909	147698519	56610
chrX	99144824	99207232	62408	chrX	125553867	125612903	59036	chrX	147739265	147750177	10912
chrX	99247960	99335194	87234	chrX	125712994	125821962	108968	chrX	148102746	148137129	34383
chrX	99435276	99510511	75235	chrX	126602851	126647208	44357	chrX	148270860	148292128	21268
chrX	99551261	99581991	30730	chrX	126924001	126943052	19051	chrX	149107166	149262232	155066
chrX	99937691	100036638	98947	chrX	126983804	127041465	57661	chrX	151332855	151346514	13659
chrX	100460273	100460386	113	chrX	127082211	127101032	18821				
chrX	100501134	100523329	22195	chrX	127174729	127225372	50643				

Table 2. Donor constructs targeting GSH1 and GSH2 sites

Donor name	Donor sequence (GSH-targeting homology arms in green)
GSH1 CMV-mRuby	CTGCATTTAAGTAGGATTCAATAATTTTAAAGTGCAGGGACAAAATTTCCTCATATGGCTCACTAGCTACATTGCAAA TTTCTTGAAATCAGAACACAGAAGTGCAGTCCTGTGCTCGCAATGCAGACTTGCAGGGTGTAGAGGCATAAATGGCT CCAGAGCCAGGGACATGGGTCCAGAGGGGGGTAGTCTCCAGAAGACTCCTTTCGGGCCTATTACCATGCCTCAGA GGTCCAAGTGGGGCATGGTGAATATATTATCCTTTATATTATATTTCTTATATGTCTACAACTGCCACTTGACATTGAT TATTGACTAGTTATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCCGCGTTACATAACT TACGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCAATAATGACGTATGTTCCCATA GTAACGCCAATAGGGACTTTCCATTGACGTCAATGGGTGGACTATTTACGGTAAACTGCCCACTTGGCAGTACATCA AGTGTATCATATGCCAAGTACGCCCCCTATTGACGTCAATGACGGTAAATGGCCCGCCTGGCATTATGCCCAGTACA TGACCTTATGGGACTTTCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGGTGATGCGGTTTTGGC AGTACATCAATGGGCGTGGATAGCGGTTTGACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGTCAATGGGAG TTTGTTTTGGCACCAAAATCAACGGGACTTTCCAAAATGTCGTAACAACTCCGCCCCATTGACGCAAATGGGCGGTA GGCGTGTACGGTGGGAGGTCTATATAAGCAGAGCTCTCTGGCTAACTAGAGAACCCACTGCTTACTGGCTTATCGA AATTAATACGACTCACTATAGGGAGACCCAAGCTTGCGGCCGCCACCATGGTGCGGGGTTCTCATCATCATCATCAT CATGGTATGGCTAGCATGACTGGTGGACAGCAAATGGGTCGGGATCTGTACGACGATGACGATAAGGATCCGATG GTGTCTAAGGGCGAAGAGCTGATCAAGGAAAATATGCGTATGAAGGTGGTCATGGAAGGTTCGGTCAACGGCCACC AATTCAAATGCACAGGTGAAGGAGAAGGCAATCCGTACATGGGAACTCAAACCATGAGGATCAAAGTCATCGAGGG AGGACCCCTGCCATTTGCCTTTGACATTCTTGCCACGTCGTTCATGTATGGCAGCCGTACTTTTATCAAGTACCCGA AAGGCATTCCTGATTTCTTTAAACAGTCCTTTCCTGAGGGTTTTACTTGGGAAAGAGTTACGAGATACGAAGATGGT GGAGTCGTCACCGTCATGCAGGACACCAGCCTTGAGGATGGCTGTCTCGTTTACCACGTCCAAGTCAGAGGGGTAA ACTTTCCCTCCAATGGTCCCGTGATGCAGAAGAAGACCAAGGGTTGGGAGCCTAATACAGAGATGATGTATCCAGC AGATGGTGGTCTGAGGGGATACACTCATATGGCACTGAAAGTTGATGGTGGTGGCCATCTGTCTTGCTCTTTCGTAA CAACTTACAGGTCAAAAAAGACCGTCGGGAACATCAAGATGCCCGGTATCCATGCCGTTGATCACCGCCTGGAAAG GTTAGAGGAAAGTGACAATGAAATGTTCGTAGTACAACGCGAACACGCAGTTGCCAAGTTCGCCGGGCTTGGTGGT GGGATGGACGAGCTGTACAAGTAAGAATTCTGCAGATATCCATCACACTGGCGGCCGCTCGAGCATGCATCTAGAG GGCCCTATTCTATAGTGTCACCTAAATGCTAGAGCTCGCTGATCAGCCTCGACTGTGCCTTCTAGTTGCCAGCCATC TGTTGTTTGCCCCTCCCCCGTGCCTTCCTTGACCCTGGAAGGTGCCACTCCCACTGTCCTTTCCTAATAAAATGAGG AAATTGCATCGCATTGTCTGAGTAGGTGTCATTCTATTCTGGGGGGTGGGGTGGGGCAGGACAGCAAGGGGGAGG ATTGGGAAGACAATAGCAGGCATGCTGGGGATGCGGTGGGCTCTATGGCATGGCACTAGGACTAAAGGTTGGCCA AAGTACAAGATATTTGTCTTATCTGATGACAACTCTGTGTCCTGGACTCTCTTCCAGAATAAGACCTTTCCTGCAGCA CTGCTTGAACTCCTCTTAGCAAGAGGGAAACATGTGAAATGCTACCAAAATAGAATAGAAGTAAATTCTTATTATATT CCTTTGTTCACTCATATCCTGAAGTGCATCAAATCAGGTTTTCTCACCTGTATAATGCTGTATTTTACTTGAGTTGGAA TAATTTTGCTTAGAAATAAATAAGTAAAACAGCACCTG

Donor name	Donor sequence (GSH-targeting homology arms in green)
GSH2 CMV-mRuby	CATTACATCCAAGTTTAGACTCATTGAGCTCTAAATATTTGGGAAAACATATTTAAAGAAATTATATAGGTTTGATCCAA AATCTCTTTGGCACAACTTGAAATATGGGTAATCGTCATGTGAAATTTGTGAATAGGAGAACCCACTGTAGGATACTTA ACATAAATCAGCCACATAATTTCTATCACTGATATCCAGGGAATTTCAATGACAAATCTAGTGATAAAAATTGATAAAAC ATTTTTGATAGTTTTGATACAAGTGAAAGTCATGGGATATCAGACTTAAAAGAAACCTCAGGACATTGATTATTGACTA GTTATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCCGCGTTACATAACTTACGGTAAA TGGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCAATAATGACGTATGTTCCCATAGTAACGCCAA TAGGGACTTTCCATTGACGTCAATGGGTGGACTATTTACGGTAAACTGCCCACTTGGCAGTACATCAAGTGTATCATA TGCCAAGTACGCCCCCTATTGACGTCAATGACGGTAAATGGCCCGCCTGGCATTATGCCCAGTACATGACCTTATGG GACTTTCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGGTGATGCGGTTTTGGCAGTACATCAATG GGCGTGGATAGCGGTTTGACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGTCAATGGGAGTTTGTTTTGGCAC CAAAATCAACGGGACTTTCCAAAATGTCGTAACAACTCCGCCCCATTGACGCAAATGGGCGGTAGGCGTGTACGGTG GGAGGTCTATATAAGCAGAGCTCTCTGGCTAACTAGAGAACCCACTGCTTACTGGCTTATCGAAATTAATACGACTCA CTATAGGGAGACCCAAGCTTGCGGCCGCCACCATGGTGCGGGGTTCTCATCATCATCATCATCATGGTATGGCTAGC ATGACTGGTGGACAGCAAATGGGTCGGGATCTGTACGACGATGACGATAAGGATCCGATGGTGTCTAAGGGCGAAG AGCTGATCAAGGAAAATATGCGTATGAAGGTGGTCATGGAAGGTTCGGTCAACGGCCACCAATTCAAATGCACAGGT GAAGGAGAAGGCAATCCGTACATGGGAACTCAAACCATGAGGATCAAAGTCATCGAGGGAGGACCCCTGCCATTTG CCTTTGACATTCTTGCCACGTCGTTCATGTATGGCAGCCGTACTTTTATCAAGTACCCGAAAGGCATTCCTGATTTCTT TAAACAGTCCTTTCCTGAGGGTTTTACTTGGGAAAGAGTTACGAGATACGAAGATGGTGGAGTCGTCACCGTCATGC AGGACACCAGCCTTGAGGATGGCTGTCTCGTTTACCACGTCCAAGTCAGAGGGGTAAACTTTCCCTCCAATGGTCCC GTGATGCAGAAGAAGACCAAGGGTTGGGAGCCTAATACAGAGATGATGTATCCAGCAGATGGTGGTCTGAGGGGAT ACACTCATATGGCACTGAAAGTTGATGGTGGTGGCCATCTGTCTTGCTCTTTCGTAACAACTTACAGGTCAAAAAAGA CCGTCGGGAACATCAAGATGCCCGGTATCCATGCCGTTGATCACCGCCTGGAAAGGTTAGAGGAAAGTGACAATGA AATGTTCGTAGTACAACGCGAACACGCAGTTGCCAAGTTCGCCGGGCTTGGTGGTGGGATGGACGAGCTGTACAAG TAAGAATTCTGCAGATATCCATCACACTGGCGGCCGCTCGAGCATGCATCTAGAGGGCCCTATTCTATAGTGTCACC TAAATGCTAGAGCTCGCTGATCAGCCTCGACTGTGCCTTCTAGTTGCCAGCCATCTGTTGTTTGCCCCTCCCCCGTG CCTTCCTTGACCCTGGAAGGTGCCACTCCCACTGTCCTTTCCTAATAAAATGAGGAAATTGCATCGCATTGTCTGAGT AGGTGTCATTCTATTCTGGGGGGTGGGGTGGGGCAGGACAGCAAGGGGGAGGATTGGGAAGACAATAGCAGGCAT GCTGGGGATGCGGTGGGCTCTATGGTGCTATCAAGTCTGATGTCAGTAATTTTTGGAGGAGACTGAAGTGCAGTGAG ACTATCCAAAGTCAGACATGGGGAAAAGCAGAGTCATCCCTCCTAGGCTGCCAAAATCCTCCCCATCCAAGCTCATC CTTGAAGCCCTCACTTAAGACAAAGTTCCTCCCATCCCTTCTGCCTGCTCTGGCATGGTCTGAACCATTTGCCTATTA ATTGCCCTGCCTGGTTTCATTTGTTCTTTTTGCTGTATTTAAACTGTGGGAATTCTATTGTTAACCTTTTTCTTGCTCAA CTGAACTGTGACA

Donor name	Donor sequence (GSH-targeting homology arms in green)
GSH1 CMV-LAMB3-T2A-GFP	CTGCATTTAAGTAGGATTCAATAATTTTAAAGTGCAGGGACAAAATTTCCTCATATGGCTCACTAGCTACATTGCAAATTT CTTGAAATCAGAACACAGAAGTGCAGTCCTGTGCTCGCAATGCAGACTTGCAGGGTGTAGAGGCATAAATGGCTCCAG AGCCAGGGACATGGGTCCAGAGGGGGGTAGTCTCCAGAAGACTCCTTTCGGGCCTATTACCATGCCTCAGAGGTCCA AGTGGGGCATGGTGAATATATTATCCTTTATATTATATTTCTTATATGTCTACAACTGCCACTTGACATTGATTATTGACTA GTTATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCCGCGTTACATAACTTACGGTAAAT GGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCAATAATGACGTATGTTCCCATAGTAACGCCAATA GGGACTTTCCATTGACGTCAATGGGTGGACTATTTACGGTAAACTGCCCACTTGGCAGTACATCAAGTGTATCATATGC CAAGTACGCCCCCTATTGACGTCAATGACGGTAAATGGCCCGCCTGGCATTATGCCCAGTACATGACCTTATGGGACT TTCCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGGTGATGCGGTTTTGGCAGTACATCAATGGGCGT GGATAGCGGTTTGACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGTCAATGGGAGTTTGTTTTGGCACCAAAATC AACGGGACTTTCCAAAATGTCGTAACAACTCCGCCCCATTGACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGGTCT ATATAAGCAGAGCTCTCTGGCTAACTAGAGAACCCACTGCTTACTGGCTTATCGAAATTAATACGACTCACTATAGGGA GACCCAAGCTTGCGGCCGCCACCATGAGACCATTCTTCCTCTTGTGTTTTGCCCTGCCTGGCCTCCTGCATGCCCAAC AAGCCTGCTCCCGTGGGGCCTGCTATCCACCTGTTGGGGACCTGCTTGTTGGGAGGACCCGGTTTCTCCGAGCTTCAT CTACCTGTGGACTGACCAAGCCTGAGACCTACTGCACCCAGTATGGCGAGTGGCAGATGAAATGCTGCAAGTGTGACT CCAGGCAGCCTCACAACTACTACAGTCACCGAGTAGAGAATGTGGCTTCATCCTCCGGCCCCATGCGCTGGTGGCAGT CCCAGAATGATGTGAACCCTGTCTCTCTGCAGCTGGACCTGGACAGGAGATTCCAGCTTCAAGAAGTCATGATGGAGT TCCAGGGGCCCATGCCTGCCGGCATGCTGATTGAGCGCTCCTCAGACTTCGGTAAGACCTGGCGAGTGTACCAGTAC CTGGCTGCCGACTGCACCTCCACCTTCCCTCGGGTCCGCCAGGGTCGGCCTCAGAGCTGGCAGGATGTTCGGTGCCA GTCCCTGCCTCAGAGGCCTAATGCACGCCTAAATGGGGGGAAGGTCCAACTTAACCTTATGGATTTAGTGTCTGGGAT TCCAGCAACTCAAAGTCAAAAAATTCAAGAGGTGGGGGAGATCACAAACTTGAGAGTCAATTTCACCAGGCTGGCCCC TGTGCCCCAAAGGGGCTACCACCCTCCCAGCGCCTACTATGCTGTGTCCCAGCTCCGTCTGCAGGGGAGCTGCTTCT GTCACGGCCATGCTGATCGCTGCGCACCCAAGCCTGGGGCCTCTGCAGGCCCCTCCACCGCTGTGCAGGTCCACGAT GTCTGTGTCTGCCAGCACAACACTGCCGGCCCAAATTGTGAGCGCTGTGCACCCTTCTACAACAACCGGCCCTGGAGA CCGGCGGAGGGCCAGGACGCCCATGAATGCCAAAGGTGCGACTGCAATGGGCACTCAGAGACATGTCACTTTGACCC CGCTGTGTTTGCCGCCAGCCAGGGGGCATATGGAGGTGTGTGTGACAATTGCCGGGACCACACCGAAGGCAAGAACT GTGAGCGGTGTCAGCTGCACTATTTCCGGAACCGGCGCCCGGGAGCTTCCATTCAGGAGACCTGCATCTCCTGCGAG TGTGATCCGGATGGGGCAGTGCCAGGGGCTCCCTGTGACCCAGTGACCGGGCAGTGTGTGTGCAAGGAGCATGTGC AGGGAGAGCGCTGTGACCTATGCAAGCCGGGCTTCACTGGACTCACCTACGCCAACCCGCAGGGCTGCCACCGCTGT GACTGCAACATCCTGGGGTCCCGGAGGGACATGCCGTGTGACGAGGAGAGTGGGCGCTGCCTTTGTCTGCCCAACGT GGTGGGTCCCAAATGTGACCAGTGTGCTCCCTACCACTGGAAGCTGGCCAGTGGCCAGGGCTGTGAACCGTGTGCCT GCGACCCGCACAACTCCCTCAGCCCACAGTGCAACCAGTTCACAGGGCAGTGCCCCTGTCGGGAAGGCTTTGGTGGC CTGATGTGCAGCGCTGCAGCCATCCGCCAGTGTCCAGACCGGACCTATGGAGACGTGGCCACAGGATGCCGAGCCTG TGACTGTGATTTCCGGGGAACAGAGGGCCCGGGCTGCGACAAGGCATCAGGCCGCTGCCTCTGCCGCCCTGGCTTG ACCGGGCCCCGCTGTGACCAGTGCCAGCGAGGCTACTGCAATCGCTACCCGGTGTGCGTGGCCTGCCACCCTTGCTT CCAGACCTATGATGCGGACCTCCGGGAGCAGGCCCTGCGCTTTGGTAGACTCCGCAATGCCACCGCCAGCCTGTGGT CAGGGCCTGGGCTGGAGGACCGTGGCCTGGCCTCCCGGATCCTAGATGCAAAGAGTAAGATTGAGCAGATCCGAGCA GTTCTCAGCAGCCCCGCAGTCACAGAGCAGGAGGTGGCTCAGGTGGCCAGTGCCATCCTCTCCCTCAGGCGAACTCT CCAGGGCCTGCAGCTGGATCTGCCCCTGGAGGAGGAGACGTTGTCCCTTCCGAGAGACCTGGAGAGTCTTGACAGAA GCTTCAATGGTCTCCTTACTATGTATCAGAGGAAGAGGGAGCAGTTTGAAAAAATAAGCAGTGCTGATCCTTCAGGAGC CTTCCGGATGCTGAGCACAGCCTACGAGCAGTCAGCCCAGGCTGCTCAGCAGGTCTCCGACAGCTCGCGCCTTTTGG

	ACCAGCTCAGGGACAGCCGGAGAGAGGCAGAGAGGCTGGTGCGGCAGGCGGGAGGAGGAGGAGGCACCGGCAGCC CCAAGCTTGTGGCCCTGAGGCTGGAGATGTCTTCGTTGCCTGACCTGACACCCACCTTCAACAAGCTCTGTGGCAACT CCAGGCAGATGGCTTGCACCCCAATATCATGCCCTGGTGAGCTATGTCCCCAAGACAATGGCACAGCCTGTGGCTCCC GCTGCAGGGGTGTCCTTCCCAGGGCCGGTGGGGCCTTCTTGATGGCGGGGCAGGTGGCTGAGCAGCTGCGGGGCTT CAATGCCCAGCTCCAGCGGACCAGGCAGATGATTAGGGCAGCCGAGGAATCTGCCTCACAGATTCAATCCAGTGCCC AGCGCTTGGAGACCCAGGTGAGCGCCAGCCGCTCCCAGATGGAGGAAGATGTCAGACGCACACGGCTCCTAATCCAG CAGGTCCGGGACTTCCTAACAGACCCCGACACTGATGCAGCCACTATCCAGGAGGTCAGCGAGGCCGTGCTGGCCCT GTGGCTGCCCACAGACTCAGCTACTGTTCTGCAGAAGATGAATGAGATCCAGGCCATTGCAGCCAGGCTCCCCAACGT GGACTTGGTGCTGTCCCAGACCAAGCAGGACATTGCGCGTGCCCGCCGGTTGCAGGCTGAGGCTGAGGAAGCCAGG AGCCGAGCCCATGCAGTGGAGGGCCAGGTGGAAGATGTGGTTGGGAACCTGCGGCAGGGGACAGTGGCACTGCAGG AAGCTCAGGACACCATGCAAGGCACCAGCCGCTCCCTTCGGCTTATCCAGGACAGGGTTGCTGAGGTTCAGCAGGTA CTGCGGCCAGCAGAAAAGCTGGTGACAAGCATGACCAAGCAGCTGGGTGACTTCTGGACACGGATGGAGGAGCTCCG CCACCAAGCCCGGCAGCAGGGGGCAGAGGCAGTCCAGGCCCAGCAGCTTGCGGAAGGTGCCAGCGAGCAGGCATT GAGTGCCCAAGAGGGATTTGAGAGAATAAAACAAAAGTATGCTGAGTTGAAGGACCGGTTGGGTCAGAGTTCCATGCT GGGTGAGCAGGGTGCCCGGATCCAGAGTGTGAAGACAGAGGCAGAGGAGCTGTTTGGGGAGACCATGGAGATGATG GACAGGATGAAAGACATGGAGTTGGAGCTGCTGCGGGGCAGCCAGGCCATCATGCTGCGCTCGGCGGACCTGACAG GACTGGAGAAGCGTGTGGAGCAGATCCGTGACCACATCAATGGGCGCGTGCTCTACTATGCCACCTGCAAGGAGGGC AGAGGAAGTCTTCTAACATGCGGTGACGTGGAGGAGAATCCCGGCCCTAGCATGGTGAGCAAGGGCGAGGAGCTGTT CACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGC GAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCAC CCTCGTGACCACCCTGACCTACGGCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAA GTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGCCG AGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATC CTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCATGGCCGACAAGCAGAAGAACGGCATCAAG GTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCCCAT CGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAGCACCCAGTCCGCCCTGAGCAAAGACCCCAACGAG AAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGTA AAGCGGACTAGTCTAGCAATCAACCTCTGGATTACAAAATTTGTGAAAGATTGACTGGTATTCTTAACTATGTTGCTCCT TTTACGCTATGTGGATACGCTGCTTTAATGCCTTTGTATCATGCTATTGCTTCCCGTATGGCTTTCATTTTCTCCTCCTTG TATAAATCCTGGTTAGTTCTTGCCACGGCGGAACTCATCGCCGCCTGCCTTGCCCGCTGCTGGACAGGGGCTCGGCTG TTGGGCACTGACAATTCCGTGGTGTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCATTCTAGCTTTATTT GTGAAATTTGTGATGCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAAACAAGTTAACAACAACAATTGCATTCATT TTATGTTTCAGGTTCAGGGGGAGATGTGGGAGGTTTTTTAAAGCCATGGCACTAGGACTAAAGGTTGGCCAAAGTACAA GATATTTGTCTTATCTGATGACAACTCTGTGTCCTGGACTCTCTTCCAGAATAAGACCTTTCCTGCAGCACTGCTTGAAC TCCTCTTAGCAAGAGGGAAACATGTGAAATGCTACCAAAATAGAATAGAAGTAAATTCTTATTATATTCCTTTGTTCACTC ATATCCTGAAGTGCATCAAATCAGGTTTTCTCACCTGTATAATGCTGTATTTTACTTGAGTTGGAATAATTTTGCTTAGAA ATAAATAAGTAAAACAGCACCTG

Donor name	Donor sequence (GSH-targeting homology arms in green)
GSH2 CMV-LAMB3 -T2A-GFP	CATTACATCCAAGTTTAGACTCATTGAGCTCTAAATATTTGGGAAAACATATTTAAAGAAATTATATAGGTTTGATCCAAAA TCTCTTTGGCACAACTTGAAATATGGGTAATCGTCATGTGAAATTTGTGAATAGGAGAACCCACTGTAGGATACTTAACAT AAATCAGCCACATAATTTCTATCACTGATATCCAGGGAATTTCAATGACAAATCTAGTGATAAAAATTGATAAAACATTTTT GATAGTTTTGATACAAGTGAAAGTCATGGGATATCAGACTTAAAAGAAACCTCAGGACATTGATTATTGACTAGTTATTAA TAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCCGCGTTACATAACTTACGGTAAATGGCCCGCC TGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCAATAATGACGTATGTTCCCATAGTAACGCCAATAGGGACTTTC CATTGACGTCAATGGGTGGACTATTTACGGTAAACTGCCCACTTGGCAGTACATCAAGTGTATCATATGCCAAGTACGCC CCCTATTGACGTCAATGACGGTAAATGGCCCGCCTGGCATTATGCCCAGTACATGACCTTATGGGACTTTCCTACTTGGC AGTACATCTACGTATTAGTCATCGCTATTACCATGGTGATGCGGTTTTGGCAGTACATCAATGGGCGTGGATAGCGGTTT GACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGTCAATGGGAGTTTGTTTTGGCACCAAAATCAACGGGACTTTCC AAAATGTCGTAACAACTCCGCCCCATTGACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGGTCTATATAAGCAGAGCT CTCTGGCTAACTAGAGAACCCACTGCTTACTGGCTTATCGAAATTAATACGACTCACTATAGGGAGACCCAAGCTTGCGG CCGCCACCATGAGACCATTCTTCCTCTTGTGTTTTGCCCTGCCTGGCCTCCTGCATGCCCAACAAGCCTGCTCCCGTGG GGCCTGCTATCCACCTGTTGGGGACCTGCTTGTTGGGAGGACCCGGTTTCTCCGAGCTTCATCTACCTGTGGACTGACC AAGCCTGAGACCTACTGCACCCAGTATGGCGAGTGGCAGATGAAATGCTGCAAGTGTGACTCCAGGCAGCCTCACAAC TACTACAGTCACCGAGTAGAGAATGTGGCTTCATCCTCCGGCCCCATGCGCTGGTGGCAGTCCCAGAATGATGTGAACC CTGTCTCTCTGCAGCTGGACCTGGACAGGAGATTCCAGCTTCAAGAAGTCATGATGGAGTTCCAGGGGCCCATGCCTG CCGGCATGCTGATTGAGCGCTCCTCAGACTTCGGTAAGACCTGGCGAGTGTACCAGTACCTGGCTGCCGACTGCACCT CCACCTTCCCTCGGGTCCGCCAGGGTCGGCCTCAGAGCTGGCAGGATGTTCGGTGCCAGTCCCTGCCTCAGAGGCCTA ATGCACGCCTAAATGGGGGGAAGGTCCAACTTAACCTTATGGATTTAGTGTCTGGGATTCCAGCAACTCAAAGTCAAAAA ATTCAAGAGGTGGGGGAGATCACAAACTTGAGAGTCAATTTCACCAGGCTGGCCCCTGTGCCCCAAAGGGGCTACCAC CCTCCCAGCGCCTACTATGCTGTGTCCCAGCTCCGTCTGCAGGGGAGCTGCTTCTGTCACGGCCATGCTGATCGCTGC GCACCCAAGCCTGGGGCCTCTGCAGGCCCCTCCACCGCTGTGCAGGTCCACGATGTCTGTGTCTGCCAGCACAACACT GCCGGCCCAAATTGTGAGCGCTGTGCACCCTTCTACAACAACCGGCCCTGGAGACCGGCGGAGGGCCAGGACGCCCA TGAATGCCAAAGGTGCGACTGCAATGGGCACTCAGAGACATGTCACTTTGACCCCGCTGTGTTTGCCGCCAGCCAGGG GGCATATGGAGGTGTGTGTGACAATTGCCGGGACCACACCGAAGGCAAGAACTGTGAGCGGTGTCAGCTGCACTATTT CCGGAACCGGCGCCCGGGAGCTTCCATTCAGGAGACCTGCATCTCCTGCGAGTGTGATCCGGATGGGGCAGTGCCAG GGGCTCCCTGTGACCCAGTGACCGGGCAGTGTGTGTGCAAGGAGCATGTGCAGGGAGAGCGCTGTGACCTATGCAAG CCGGGCTTCACTGGACTCACCTACGCCAACCCGCAGGGCTGCCACCGCTGTGACTGCAACATCCTGGGGTCCCGGAG GGACATGCCGTGTGACGAGGAGAGTGGGCGCTGCCTTTGTCTGCCCAACGTGGTGGGTCCCAAATGTGACCAGTGTGC TCCCTACCACTGGAAGCTGGCCAGTGGCCAGGGCTGTGAACCGTGTGCCTGCGACCCGCACAACTCCCTCAGCCCACA GTGCAACCAGTTCACAGGGCAGTGCCCCTGTCGGGAAGGCTTTGGTGGCCTGATGTGCAGCGCTGCAGCCATCCGCCA GTGTCCAGACCGGACCTATGGAGACGTGGCCACAGGATGCCGAGCCTGTGACTGTGATTTCCGGGGAACAGAGGGCC CGGGCTGCGACAAGGCATCAGGCCGCTGCCTCTGCCGCCCTGGCTTGACCGGGCCCCGCTGTGACCAGTGCCAGCGA GGCTACTGCAATCGCTACCCGGTGTGCGTGGCCTGCCACCCTTGCTTCCAGACCTATGATGCGGACCTCCGGGAGCAG GCCCTGCGCTTTGGTAGACTCCGCAATGCCACCGCCAGCCTGTGGTCAGGGCCTGGGCTGGAGGACCGTGGCCTGGC CTCCCGGATCCTAGATGCAAAGAGTAAGATTGAGCAGATCCGAGCAGTTCTCAGCAGCCCCGCAGTCACAGAGCAGGA GGTGGCTCAGGTGGCCAGTGCCATCCTCTCCCTCAGGCGAACTCTCCAGGGCCTGCAGCTGGATCTGCCCCTGGAGGA GGAGACGTTGTCCCTTCCGAGAGACCTGGAGAGTCTTGACAGAAGCTTCAATGGTCTCCTTACTATGTATCAGAGGAAG AGGGAGCAGTTTGAAAAAATAAGCAGTGCTGATCCTTCAGGAGCCTTCCGGATGCTGAGCACAGCCTACGAGCAGTCAG

CCCAGGCTGCTCAGCAGGTCTCCGACAGCTCGCGCCTTTTGGACCAGCTCAGGGACAGCCGGAGAGAGGCAGAGAGG
CTGGTGCGGCAGGCGGGAGGAGGAGGAGGCACCGGCAGCCCCAAGCTTGTGGCCCTGAGGCTGGAGATGTCTTCGTT
GCCTGACCTGACACCCACCTTCAACAAGCTCTGTGGCAACTCCAGGCAGATGGCTTGCACCCCAATATCATGCCCTGGT
GAGCTATGTCCCCAAGACAATGGCACAGCCTGTGGCTCCCGCTGCAGGGGTGTCCTTCCCAGGGCCGGTGGGGCCTT
CTTGATGGCGGGGCAGGTGGCTGAGCAGCTGCGGGGCTTCAATGCCCAGCTCCAGCGGACCAGGCAGATGATTAGGG
CAGCCGAGGAATCTGCCTCACAGATTCAATCCAGTGCCCAGCGCTTGGAGACCCAGGTGAGCGCCAGCCGCTCCCAGA
TGGAGGAAGATGTCAGACGCACACGGCTCCTAATCCAGCAGGTCCGGGACTTCCTAACAGACCCCGACACTGATGCAG
CCACTATCCAGGAGGTCAGCGAGGCCGTGCTGGCCCTGTGGCTGCCCACAGACTCAGCTACTGTTCTGCAGAAGATGA
ATGAGATCCAGGCCATTGCAGCCAGGCTCCCCAACGTGGACTTGGTGCTGTCCCAGACCAAGCAGGACATTGCGCGTG
CCCGCCGGTTGCAGGCTGAGGCTGAGGAAGCCAGGAGCCGAGCCCATGCAGTGGAGGGCCAGGTGGAAGATGTGGTT
GGGAACCTGCGGCAGGGGACAGTGGCACTGCAGGAAGCTCAGGACACCATGCAAGGCACCAGCCGCTCCCTTCGGCT
TATCCAGGACAGGGTTGCTGAGGTTCAGCAGGTACTGCGGCCAGCAGAAAAGCTGGTGACAAGCATGACCAAGCAGCT
GGGTGACTTCTGGACACGGATGGAGGAGCTCCGCCACCAAGCCCGGCAGCAGGGGGCAGAGGCAGTCCAGGCCCAG
CAGCTTGCGGAAGGTGCCAGCGAGCAGGCATTGAGTGCCCAAGAGGGATTTGAGAGAATAAAACAAAAGTATGCTGAG
TTGAAGGACCGGTTGGGTCAGAGTTCCATGCTGGGTGAGCAGGGTGCCCGGATCCAGAGTGTGAAGACAGAGGCAGA
GGAGCTGTTTGGGGAGACCATGGAGATGATGGACAGGATGAAAGACATGGAGTTGGAGCTGCTGCGGGGCAGCCAGG
CCATCATGCTGCGCTCGGCGGACCTGACAGGACTGGAGAAGCGTGTGGAGCAGATCCGTGACCACATCAATGGGCGC
GTGCTCTACTATGCCACCTGCAAGGAGGGCAGAGGAAGTCTTCTAACATGCGGTGACGTGGAGGAGAATCCCGGCCCT
AGCATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGG
CCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCAC
CGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGACCTACGGCGTGCAGTGCTTCAGCCGCTACCCCG
ACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGA
CGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCA
TCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCATGGC
CGACAAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTGCAGCTCGCCGA
CCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAGCACCCAGTCCGC
CCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGG
CATGGACGAGCTGTACAAGTAAAGCGGACTAGTCTAGCAATCAACCTCTGGATTACAAAATTTGTGAAAGATTGACTGGT
ATTCTTAACTATGTTGCTCCTTTTACGCTATGTGGATACGCTGCTTTAATGCCTTTGTATCATGCTATTGCTTCCCGTATGG
CTTTCATTTTCTCCTCCTTGTATAAATCCTGGTTAGTTCTTGCCACGGCGGAACTCATCGCCGCCTGCCTTGCCCGCTGC
TGGACAGGGGCTCGGCTGTTGGGCACTGACAATTCCGTGGTGTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGT
AACCATTCTAGCTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCATTATAAGCTGCAATAAACAAGTTAACA
ACAACAATTGCATTCATTTTATGTTTCAGGTTCAGGGGGAGATGTGGGAGGTTTTTTAAAGCTGCTATCAAGTCTGATGTC
AGTAATTTTTGGAGGAGACTGAAGTGCAGTGAGACTATCCAAAGTCAGACATGGGGAAAAGCAGAGTCATCCCTCCTAG
GCTGCCAAAATCCTCCCCATCCAAGCTCATCCTTGAAGCCCTCACTTAAGACAAAGTTCCTCCCATCCCTTCTGCCTGCT
CTGGCATGGTCTGAACCATTTGCCTATTAATTGCCCTGCCTGGTTTCATTTGTTCTTTTTGCTGTATTTAAACTGTGGGAA
TTCTATTGTTAACCTTTTTCTTGCTCAACTGAACTGTGACA

6. References

1. Blaese, R. M. *et al.* T Lymphocyte-Directed Gene Therapy for ADA$^-$ SCID: Initial Trial Results After 4 Years. *Sci. New Ser.* **270**, 475–480 (1995).

2. Raper, S. E. *et al.* Fatal systemic inflammatory response syndrome in a ornithine transcarbamylase deficient patient following adenoviral gene transfer. *Mol. Genet. Metab.* **80**, 148–158 (2003).

3. Carmen, I. H. A Death in the Laboratory: The Politics of the Gelsinger Aftermath. *Mol. Ther.* **3**, 425–428 (2001).

4. Russell, S. *et al.* Efficacy and safety of voretigene neparvovec (AAV2-hRPE65v2) in patients with RPE65 -mediated inherited retinal dystrophy: a randomised, controlled, open-label, phase 3 trial. *The Lancet* **390**, 849–860 (2017).

5. Feins, S., Kong, W., Williams, E. F., Milone, M. C. & Fraietta, J. A. An introduction to chimeric antigen receptor (CAR) T-cell immunotherapy for human cancer. *Am. J. Hematol.* **94**, S3–S9 (2019).

6. Wagner, J., Wickman, E., DeRenzo, C. & Gottschalk, S. CAR T Cell Therapy for Solid Tumors: Bright Future or Dark Reality? *Mol. Ther.* **28**, 2320–2339 (2020).

7. Eyquem, J. *et al.* Targeting a CAR to the TRAC locus with CRISPR/Cas9 enhances tumour rejection. *Nature* **543**, 113–117 (2017).

8. Lee, C. S. *et al.* Adenovirus-mediated gene delivery: Potential applications for gene and cell-based therapies in the new era of personalized medicine. *Genes Dis.* **4**, 43–63 (2017).

9. Atasheva, S. & Shayakhmetov, D. M. Adenovirus sensing by the immune system. *Curr. Opin. Virol.* **21**, 109–113 (2016).

10. Cots, D., Bosch, A. & Chillón, M. Helper Dependent Adenovirus Vectors: Progress and Future Prospects. *Curr. Gene Ther.* **13**, 370–381 (2013).

11. Kremer, E. J. Pros and Cons of Adenovirus-Based SARS-CoV-2 Vaccines. *Mol. Ther.* **28**, 2303–2304 (2020).

12. Musayev, F. N., Zarate-Perez, F., Bishop, C., Burgner, J. W. & Escalante, C. R. Structural

Insights into the Assembly of the Adeno-associated Virus Type 2 Rep68 Protein on the Integration Site AAVS1. *J. Biol. Chem.* **290**, 27487–27499 (2015).

13. Wang, D., Tai, P. W. L. & Gao, G. Adeno-associated virus vector as a platform for gene therapy delivery. *Nat. Rev. Drug Discov.* **18**, 358–378 (2019).

14. Keeler, A. M. & Flotte, T. R. Recombinant Adeno-Associated Virus Gene Therapy in Light of Luxturna (and Zolgensma and Glybera): Where Are We, and How Did We Get Here? *Annu. Rev. Virol.* **6**, 601–621 (2019).

15. Hocquemiller, M., Giersch, L., Audrain, M., Parker, S. & Cartier, N. Adeno-Associated Virus-Based Gene Therapy for CNS Diseases. *Hum. Gene Ther.* **27**, 478–496 (2016).

16. Moss, K. H., Popova, P., Hadrup, S. R., Astakhova, K. & Taskova, M. Lipid Nanoparticles for Delivery of Therapeutic RNA Oligonucleotides. *Mol Pharm.* 13 (2019).

17. Floch, V. *et al.* Cation Substitution in Cationic Phosphonolipids: A New Concept To Improve Transfection Activity and Decrease Cellular Toxicity. *J. Med. Chem.* **43**, 4617–4628 (2000).

18. Semple, S. C. *et al.* E¤cient encapsulation of antisense oligonucleotides in lipid vesicles using ionizable aminolipids: formation of novel small multilamellar vesicle structures. *Biochim. Biophys. Acta* 15.

19. Pardi, N., Hogan, M. J., Porter, F. W. & Weissman, D. mRNA vaccines — a new era in vaccinology. *Nat. Rev. Drug Discov.* **17**, 261–279 (2018).

20. Jackson, L. A. *et al.* An mRNA Vaccine against SARS-CoV-2 — Preliminary Report. *N. Engl. J. Med.* **383**, 1920–1931 (2020).

21. Milone, M. C. & O'Doherty, U. Clinical use of lentiviral vectors. *Leukemia* **32**, 1529–1541 (2018).

22. Maetzig, T., Galla, M., Baum, C. & Schambach, A. Gammaretroviral Vectors: Biology, Technology and Application. *Viruses* **3**, 677–713 (2011).

23. Lewinski, M. K. *et al.* Retroviral DNA Integration: Viral and Cellular Determinants of Target-Site Selection. *PLoS Pathog.* **2**, e60 (2006).

24. Lewis, P. F. & Emerman, M. Passage through mitosis is required for oncoretroviruses but

not for the human immunodeficiency virus. *J. Virol.* **68**, 510–516 (1994).

25. Vannucci, L., Lai, M., Chiuppesi, F., Ceccherini-Nelli, L. & Pistello, M. Viral vectors: a look back and ahead on gene transfer technology. *Viral Vectors* 22.

26. Yacoub, N. al, Romanowska, M., Haritonova, N. & Foerster, J. Optimized production and concentration of lentiviral vectors containing large inserts. *J. Gene Med.* **9**, 579–584 (2007).

27. Howe, S. J. *et al.* Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients. *J. Clin. Invest.* **118**, 3143–3150 (2008).

28. Hacein-Bey-Abina, S. *et al.* Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. *J. Clin. Invest.* **118**, 3132–3142 (2008).

29. Ayoub, E. *et al.* EVI1 overexpression reprograms hematopoiesis via upregulation of Spi1 transcription. *Nat. Commun.* **9**, 4239 (2018).

30. Stein, S. *et al.* Genomic instability and myelodysplasia with monosomy 7 consequent to EVI1 activation after gene therapy for chronic granulomatous disease. *Nat. Med.* **16**, 198–204 (2010).

31. Epinat, J.-C. A novel engineered meganuclease induces homologous recombination in yeast and mammalian cells. *Nucleic Acids Res.* **31**, 2952–2962 (2003).

32. Carroll, D. Genome Engineering With Zinc-Finger Nucleases. *Genetics* **188**, 773–782 (2011).

33. Urnov, F. D., Rebar, E. J., Holmes, M. C., Zhang, H. S. & Gregory, P. D. Genome editing with engineered zinc finger nucleases. *Nat. Rev. Genet.* **11**, 636–646 (2010).

34. Gaj, T., Gersbach, C. A. & Barbas, C. F. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol.* **31**, 397–405 (2013).

35. Li, T. *et al.* TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. *Nucleic Acids Res.* **39**, 359–372 (2011).

36. Moscou, M. J. & Bogdanove, A. J. A Simple Cipher Governs DNA Recognition by TAL Effectors. *Science* **326**, 1501–1501 (2009).

37. Horvath, P. & Barrangou, R. CRISPR/Cas, the Immune System of Bacteria and Archaea.

Science **327**, 167–170 (2010).

38. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).

39. Jinek, M. *et al.* A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **337**, 816–821 (2012).

40. Mali, P. *et al.* RNA-Guided Human Genome Engineering via Cas9. *Science* **339**, 823-826 (2013).

41. Pardo, B., Gómez-Gonzµlez, B. & Aguilera, A. DNA double-strand break repair: how to fix a broken relationship. **66**, 18 (2009).

42. Zhang, X.-H. Off-target Effects in CRISPR/Cas9-mediated Genome Engineering. *Mol. Ther.* **4** (2015).

43. Kelton, W. J., Pesch, T., Matile, S. & Reddy, S. T. Surveying the Delivery Methods of CRISPR/Cas9 for ex vivo Mammalian Cell Engineering. *Chimia* **70**, 439-442 (2016).

44. Sadelain, M., Papapetrou, E. P. & Bushman, F. D. Safe harbours for the integration of new DNA in the human genome. *Nat. Rev. Cancer* **12**, 51–58 (2012).

45. Shin, S. *et al.* Comprehensive Analysis of Genomic Safe Harbors as Target Sites for Stable Expression of the Heterologous Gene in HEK293 Cells. *ACS Synth. Biol.* **9**, 1263–1269 (2020).

46. Papapetrou, E. P. & Schambach, A. Gene Insertion Into Genomic Safe Harbors for Human Gene Therapy. *Mol. Ther.* **24**, 678–684 (2016).

47. Pellenz, S. *et al.* New Human Chromosomal Sites with “Safe Harbor” Potential for Targeted Transgene Insertion. *Hum. Gene Ther.* **30**, 814–828 (2019).

48. Gaidukov, L. *et al.* A multi-landing pad DNA integration platform for mammalian cell engineering. *Nucleic Acids Res.* **46**, 4072–4086 (2018).

49. Papapetrou, E. P. *et al.* Genomic safe harbors permit high β-globin transgene expression in thalassemia induced pluripotent stem cells. *Nat. Biotechnol.* **29**, 73–78 (2011).

50. Kashima, Y. Single-cell sequencing techniques from individual to multiomics analyses. *Mol. Med.* 9 (2020).

51. Nakai, H. *et al.* Extrachromosomal Recombinant Adeno-Associated Virus Vector Genomes Are Primarily Responsible for Stable Liver Transduction In Vivo. *J VIROL* **75**, 8 (2001).

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

53. Roybal, K. T. & Lim, W. A. Synthetic Immunology: Hacking Immune Cells to Expand Their Therapeutic Capabilities. 27 (2017).

54. Abraham, R. T. & Weiss, A. Jurkat T cells and development of the T-cell receptor signalling paradigm. *Nat. Rev. Immunol.* **4**, 301–308 (2004).

55. Morgan, R. A., Gray, D., Lomova, A. & Kohn, D. B. Hematopoietic Stem Cell Gene Therapy: Progress and Lessons Learned. *Cell Stem Cell* **21**, 574–590 (2017).

56. Bardhan, A. *et al.* Epidermolysis bullosa. *Nat. Rev. Dis. Primer* **6**, 78 (2020).

57. Gorell, E., Nguyen, N., Lane, A. & Siprashvili, Z. Gene Therapy for Skin Diseases. *Cold Spring Harb Perspect Med* **4** (2014).

58. Werner, S. & Grose, R. Regulation of Wound Healing by Growth Factors and Cytokines. *Physiol Rev* **83**, 36 (2003).

59. Chin, C. L. *et al.* A human expression system based on HEK293 for the stable production of recombinant erythropoietin. *Sci. Rep.* **9**, 16768 (2019).

60. Bestor, T. H. Gene silencing as a threat to the success of gene therapy. *J. Clin. Invest.* **105**, 409–411 (2000).

61. Ellis, D. J. Silencing and Variegation of Gammaretrovirus and Lentivirus Vectors. *Human Gene Therapy* **16**, 1241-1246 (2005).

62. Lee, J. S., Kildegaard, H. F., Lewis, N. E. & Lee, G. M. Mitigating Clonal Variation in Recombinant Mammalian Cell Lines. *Trends Biotechnol.* **37**, 931–942 (2019).

63. Chen, W. *et al.* AAVS1 site-specific integration of the CAR gene into human primary T cells using a linear closed-ended AAV-based DNA vector. *J. Gene Med.* **22**, (2020).

64. Richardson, N. H. *et al.* Tuning the performance of CAR T cell immunotherapies. *BMC Biotechnol.* **19**, 84 (2019).

65. Droz-Georget Lathion, S. *et al.* A single epidermal stem cell strategy for safe *ex vivo* gene therapy. *EMBO Mol. Med.* **7**, 380–393 (2015).

66. Hirsch, T. *et al.* Regeneration of the entire human epidermis using transgenic stem cells. *Nature* **551**, 327–332 (2017).

67. Maeder, M. L. & Gersbach, C. A. Genome-editing Technologies for Gene and Cell Therapy. *Mol. Ther.* **24**, 430–446 (2016).

68. Barzel, A. *et al.* Promoterless gene targeting without nucleases ameliorates haemophilia B in mice. *Nature* **517**, 360–364 (2015).

69. Oceguera-Yanez, F. *et al.* Engineering the AAVS1 locus for consistent and scalable transgene expression in human iPSCs and their differentiated derivatives. *Methods* **101**, 43–55 (2016).

70. Hong, S. G. *et al.* Rhesus iPSC Safe Harbor Gene-Editing Platform for Stable Expression of Transgenes in Differentiated Cells of All Germ Layers. *Mol. Ther.* **25**, 44–53 (2017).

71. Ordovás, L. *et al.* Efficient Recombinase-Mediated Cassette Exchange in hPSCs to Study the Hepatocyte Lineage Reveals AAVS1 Locus-Mediated Transgene Inhibition. *Stem Cell Rep.* **5**, 918–931 (2015).

72. Jiao, X. *et al.* Recent Advances Targeting CCR5 for Cancer and Its Role in Immuno-Oncology. *Cancer Res.* **79**, 4801–4807 (2019).

73. Silva, E. & Stumpf, M. P. H. HIV and the CCR5-Δ32 resistance allele. *FEMS Microbiol. Lett.* **241**, 1–12 (2004).

74. Lombardo, A. *et al.* Site-specific integration and tailoring of cassette design for sustainable gene transfer. *Nat. Methods* **8**, 861–869 (2011).

75. Sather, B. D. *et al.* Efficient modification of CCR5 in primary human hematopoietic cells using a megaTAL nuclease and AAV donor template. *Sci Transl Med* **7** (2015).

76. Joy, M. T. *et al.* CCR5 Is a Therapeutic Target for Recovery after Stroke and Traumatic Brain Injury. *Cell* **176**, 1143-1157.e13 (2019).

77. Irion, S. *et al.* Identification and targeting of the ROSA26 locus in human embryonic

stem cells. *Nat. Biotechnol.* **25**, 1477–1482 (2007).

78. Friedrich, G. & Soriano, P. Promoter traps in embryonic stem cells: a genetic screen to identify, and mutate developmental genes m mice. *Genes Dev* **5**, 1513-1523 (1991).

79. Zambrowicz, B. P. *et al.* Disruption of overlapping transcripts in the ROSA geo 26 gene trap strain leads to widespread expression of -galactosidase in mouse embryos and hematopoietic cells. *Proc. Natl. Acad. Sci.* **94**, 3789–3794 (1997).

80. Filipowicz, W., Bhattacharyya, S. N. & Sonenberg, N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat. Rev. Genet.* **9**, 102–114 (2008).

81. Schoenfelder, S. & Fraser, P. Long-range enhancer–promoter contacts in gene expression control. *Nat. Rev. Genet.* **20**, 437–455 (2019).

82. Vangala, P. *et al.* High-Resolution Mapping of Multiway Enhancer-Promoter Interactions Regulating Pathogen Detection. *Mol. Cell* **80**, 359-373.e8 (2020).

83. Chen, C.-K. *et al.* Xist recruits the X chromosome to the nuclear lamina to enable chromosome-wide silencing. *Science* **354**, 468–472 (2016).

84. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).

85. Schimmel, P. The emerging complexity of the tRNA world: mammalian tRNAs beyond protein synthesis. *Nat. Rev. Mol. Cell Biol.* **19**, 45–58 (2018).

86. Villasante, A., Abad, J. P. & Mendez-Lago, M. Centromeres were derived from telomeres during the evolution of the eukaryotic chromosome. *Proc. Natl. Acad. Sci.* **104**, 10542–10547 (2007).

87. Roybal, K. T. *et al.* Engineering T Cells with Customized Therapeutic Response Programs Using Synthetic Notch Receptors. *Cell* **167**, 419-432.e16 (2016).

88. Vazquez-Lombardi, R. *et al.* CRISPR-targeted display of functional T cell receptors enables engineering of enhanced specificity and prediction of cross-reactivity. http://biorxiv.org/lookup/doi/10.1101/2020.06.23.166363 (2020).

89. Sakuma, T., Nakade, S., Sakane, Y., Suzuki, K.-I. T. & Yamamoto, T. MMEJ-assisted

gene knock-in using TALENs and CRISPR-Cas9 with the PITCh systems. *Nat. Protoc.* **11**, 118–133 (2016).

90. Nakade, S. *et al.* Microhomology-mediated end-joining-dependent integration of donor DNA in cells and animals using TALENs and CRISPR/Cas9. *Nat. Commun.* **5**, (2014).

91. Sfeir, A. & Symington, L. S. Microhomology-Mediated End Joining: A Back-up Survival Mechanism or Dedicated Pathway? *Trends Biochem. Sci.* **40**, 701–714 (2015).

92. Gutierrez-Triana, J. A. *et al.* Efficient single-copy HDR by 5' modified long dsDNA donors. *eLife* **7**, e39468 (2018).

93. Robbins, P. B. *et al.* In vivo restoration of laminin 5 3 expression and function in junctional epidermolysis bullosa. *Proc. Natl. Acad. Sci.* **98**, 5193–5198 (2001).

94. Nielsen, A. A. & Voigt, C. A. Multi-input CRISPR/Cas genetic circuits that interface host regulatory networks. *Mol. Syst. Biol.* **10**, 763 (2014).

95. Schwarz, K. A. & Leonard, J. N. Engineering cell-based therapies to interface robustly with host physiology. *Adv. Drug Deliv. Rev.* **105**, 55–65 (2016).

96. Baeuerle, P. A. *et al.* Synthetic TRuC receptors engaging the complete T cell receptor for potent anti-tumor response. *Nat. Commun.* **10**, 2087 (2019).

97. Yeku, O. O. Armored CAR T cells enhance antitumor efficacy and overcome the tumor microenvironment. *Sci. Rep.* **7** (2017).

98. Fromm, B. *et al.* MirGeneDB 2.0: the metazoan microRNA complement. *Nucleic Acids Res.* **48**, D132–D141 (2020).

99. Gao, T. & Qian, J. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res.* **48**, D58-D64 (2019)

100. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

101. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).

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

103. Liao, Y., Smyth, G. K. & Shi, W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* **41**, e108–e108 (2013).

104. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930 (2014).

105. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

106. Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology* **11** (2010).

107. Yermanos, A. *et al.* Platypus: an open-access software for integrating lymphocyte single-cell immune repertoires with transcriptomes. *NAR Genomics and Bioinformatics* **3** (2021).

108. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

109. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).

110. Stoeger, T., Gerlach, M., Morimoto, R. I. & Nunes Amaral, L. A. Large-scale investigation of the reasons why potentially important genes are ignored. *PLOS Biol.* **16**, e2006643 (2018).

111. Gebert, L. F. R. & MacRae, I. J. Regulation of microRNA function in animals. *Nat. Rev. Mol. Cell Biol.* **20**, 21–37 (2019).

112. Moreira de Mello, J. C., Fernandes, G. R., Vibranovski, M. D. & Pereira, L. V. Early X chromosome inactivation during human preimplantation development revealed by single-cell RNA-sequencing. *Sci. Rep.* **7**, 10794 (2017).

113. Shay, J. W. & Wright, W. E. Telomeres and telomerase: three decades of progress. *Nat. Rev. Genet.* **20**, 299–309 (2019).

114. Rajagopal, N. *et al.* High-throughput mapping of regulatory DNA. *Nat. Biotechnol.* **34**, 167–174 (2016).

115. Kappel, S., Matthess, Y., Kaufmann, M. & Strebhardt, K. Silencing of mammalian genes by tetracycline-inducible shRNA expression. *Nat. Protoc.* **2**, 3257–3269 (2007).

116. Henriksen, J. R. *et al.* Comparison of RNAi efficiency mediated by tetracycline-responsive H1 and U6 promoter variants in mammalian cell lines. *Nucleic Acids Res.* **35**, e67–e67 (2007).

117. Chang, H. H. Y., Pannunzio, N. R., Adachi, N. & Lieber, M. R. Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nat. Rev. Mol. Cell. Biol.* **18**, 495-506 (2017).

118. He, X. *et al.* Knock-in of large reporter genes in human cells via CRISPR/Cas9-induced homology-dependent and independent DNA repair. *Nucleic Acids Res.* **44**, e85–e85 (2016).

119. Yu, W. *et al.* Repair of G1 induced DNA double-strand breaks in S-G2/M by alternative NHEJ. *Nat. Commun.* **11**, 5239 (2020).

120. Suzuki, K. *et al.* In vivo genome editing via CRISPR/Cas9 mediated homology-independent targeted integration. *Nature* **540**, 144–149 (2016).

121. Chylinski, K. CRISPR-Switch regulates sgRNA activity by Cre recombination for sequential editing of two loci. *Nat Comms* **10** (2019).

122. Lázaro, I., Cossu, G. & Kostarelos, K. Transient transcription factor OSKM expression is key towards clinical translation of in vivo cell reprogramming. *EMBO Mol. Med.* **9**, 733–736 (2017).

123. Miranda, M., Morici, J. F., Zanoni, M. B. & Bekinschtein, P. Brain-Derived Neurotrophic Factor: A Key Molecule for Memory in the Healthy and the Pathological Brain. *Front. Cell. Neurosci.* **13**, 363 (2019).

124. Greb, J. E. *et al.* Psoriasis. *Nat. Rev. Dis. Primer* **2**, 16082 (2016).

125. Schukur, L., Geering, B., Charpin-El Hamri, G. & Fussenegger, M. Implantable synthetic cytokine converter cells with AND-gate logic treat experimental psoriasis. *Sci. Transl. Med.* **7**, 318ra201-318ra201 (2015).

126. Iriguchi, S. *et al.* A clinically applicable and scalable method to regenerate T-cells from iPSCs for off-the-shelf T-cell immunotherapy. *Nat. Commun.* **12**, 430 (2021).

127. Kim, S. CTCF as a multifunctional protein in genome regulation and gene expression. *Exp. Mol. Med.* **47** (2015).

128. Belton, J.-M. *et al.* Hi–C: A comprehensive technique to capture the conformation of genomes. *Methods* **58** (2012).

129. Quinodoz, S. A. *et al.* Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell* **174**, 744-757.e24 (2018).

130. Guijas, C., Montenegro-Burke, J. R., Warth, B., Spilker, M. E. & Siuzdak, G. Metabolomics activity screening for identifying metabolites that modulate phenotype. *Nat. Biotechnol.* **36**, 316–320 (2018).

ERIK AZNAURYAN

Rosentalstrasse 17, Basel, Switzerland, 4058
+41 79 867 8060 • erik.aznauryan@bsse.ethz.ch

EDUCATION

ETH Zürich, Switzerland **2016-2021**
PhD, Systems Biology

Columbia University, NY, USA **2013-2014**
MA, Biotechnology

Newcastle University, UK **2010-2013**
BSc, Biomedical sciences

EXPERIENCE

ETH Zürich, Switzerland **2016-2021**
PhD candidate; Advisor: Dr Sai Reddy
Discovery and validation of novel human genomic safe harbor sites for gene and cell therapies

- Computationally predicted novel genomic safe harbor sites satisfying existing and newly introduced safety criteria.
- Experimentally validated predicted sites for durable expression of reporter and therapeutic genes of interest in various cellular contexts.
- Verified safety of newly identified safe harbor sites using bulk and single-cell transcriptomics.

Harvard University, MA, USA **2019-2020**
Visiting graduate student; Advisor: Dr George Church
Use of genomic safe harbor sites for gene therapy of skin disorders and anti-aging skin therapy

- Developed ex-vivo gene therapy approach for inherited skin disorders using CRISPR/Cas9 technology and novel human genomic safe harbor sites.
- Phenotypically screened age-associated transcription factors to identify potential targets for skin anti-aging therapy.

Flagship Pioneering, MA, USA **2020**
Flagship Pioneering Fellow
Exploration and creation of scientifically novel platform technologies

- Worked in a team of fellows to ideate on new venture hypotheses revolving around medical, industrial and agricultural biotechnology.
- Presented our team's ideas to the leadership of the company for evaluation and potential company creation.

California Institute of Technology, CA, USA **2015-2016**
Research assistant; Advisor: Dr Mitchell Guttman
Role of Xist lncRNA and its associated proteins in X-chromosome inactivation

- Used genome engineering tools to create novel mammalian ES cell lines for elucidation of the X-chromosome inactivation mechanism by Xist lncRNA and its interacting proteins.
- Implemented genome engineering techniques to develop single-molecule lncRNA imaging methods.

Atlas Biomed, Russian Federation **2016**
Summer intern; Advisor: Dr Vladislav Mileyko
Genetic testing for inherited disorders

- Developed protocols for population-wide breast cancer screening using clinical and genetic diagnostics.

Columbia University, NY, USA **2014**
Master's research project; Advisor: Dr Virginia Cornish
Overproduction of terpenes in yeast *S. cerevisiae*

- Modified metabolic pathways in yeast *S. cerevisiae* by expressing synthetic enzymatic complexes in an effort to increase the production levels of terpenes.
- Developed fermentation and mass-spectrometry protocols for measuring overproduction of terpenes in yeast.

Newcastle University, UK **2013**
Undergraduate research project; Advisor: Dr Robert Taylor
Effects of novel mutations in *LRPPRC* gene in patients with Leigh Syndrome

- Characterized the structural composition of complex IV of electron transport chain in patients with novel mutations in *LRPPRC* gene.
- Studied consequences of novel *LRPPRC* mutations on functional parameters of mitochondria and cellular respiration.

Association of Pediatric Ophthalmology Clinics, Russian Federation **2011**
Summer intern; Advisor: Dr Victoria Balasanyan
Biomechanical formula for precise strabismus surgery dosage

- Participated in derivation of a biomechanical formula for precise and efficient dosage of surgical correction of strabismus.

PAPERS & PATENTS

Aznauryan, E., Yermanos, A., Kinzina, E., Kapetanovic, E., Milanova, D., Church, G., Reddy, S. (2021) Discovery and validation of novel human genomic safe harbor sites for gene and cell therapies. *Cell Genomics. In review.*

Aznauryan, E., Milanova, D., Reddy, S., Church, G. (2021) Use of novel genomic safe harbor sites for skin gene therapy. *Patent application filed.*

Vazques-Lombardi, R., Jung, J., Bieberich, F., Kapetanovic, E., **Aznauryan, E.**, Weber, C., Reddy, S. (2020) Synthetic T cell receptors engineered for enhanced activity and specificity to tumor antigen. *Cell. In review.*

Quinodoz, S., Ollikainen, N., Tabak, B., Palia, A., Schmidt, J., Detmar E., Lai, M., Shishkin, A., Bhat, P., Takei, Y., Trinh, V., **Aznauryan, E.**, Russell, P., Cheng, C., Jovanovic, M., Chow, A., Cai, L., McDonel, P., Garber, M., Guttman, M. (2018) Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell* 174, 744–757.

Chen, C., Blanco, M., Jackson, C., **Aznauryan, E.**, Ollikainen, N., Surka, C., Chow, A., Cerase, A., McDonel, P., Guttman, M. (2016) Xist recruits the X-chromosome to the nuclear lamina to enable chromosome-wide silencing. *Science* 354, 468-72.

Oláhová, M., Hardy, S., Hall, J., Yarham, J., Haack, T., Wilson, W., Alston, C., He, L., **Aznauryan, E.**, Brown, R., Brown, G., Morris, A., Mundy, H., Broomfield, A., Barbosa, J., Simpson, M., Deshpande, C., Moeslinge, D., Koch, J., Stettner, G., Bonnen, P., Prokisch, H., Lightowlers, R., McFarland, R., Chrzanowska-Lightowlers, Z., Taylor, R. (2015) LRPPRC mutations cause early-onset multisystem mitochondrial disease and COX deficiency outside of the French Canadian population. *Brain* 138, 3503-19.

LEADERSHIP & VOLUNTEER EXPERIENCE

Foundation for Armenian Science and Technology **2017-present**
Founder of the Next Generation council

TUMO, Armenia **2016-present**
Biotechnology workshop leader

Birthright Armenia, Armenia **2016**
Teaching volunteer

Columbia University, NY, USA **2014**
Teaching assistant

Royal Victoria Infirmary, UK **2010-2012**
Hospital volunteer

SCHOLARSHIPS & AWARDS

Synthego **2019**
Genome Engineering Innovation Grant

Armenian General Benevolent Union **2013**
US Graduate Scholarship

Newcastle University **2010**
International Undergraduate Merit Scholarship

REFERENCES

Sai Reddy, Ph.D.

Associate Professor of Systems and Synthetic Immunology

Department of Biosystems Science and Engineering

ETH Zürich

sai.reddy@bsse.ethz.ch

George Church, Ph.D.

Robert Winthrop Professor of Genetics
Harvard Medical School

Harvard University

gchurch@genetics.med.harvard.edu

Randall Platt, Ph.D.

Associate Professor of Biological Engineering

Department of Biosystems Science and Engineering

ETH Zürich

rplatt@ethz.ch

Roger Hajjar, M.D.

Head of the Research and Development

Ring Therapeutics

Arthur & Janet C. Ross Professor of Medicine

Mount Sinai School of Medicine
rhajjar@flagshippioneering.com