Developments in Human Pluripotent Stem Cell Genome Engineering and *in situ* Sequencing Technologies

A dissertation presented

by

Joyce Lichi Yang

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biological and Biomedical Sciences

Harvard University

Cambridge, Massachusetts

April 2015

Dissertation Advisor: George M. Church, Ph.D. Author: Joyce Lichi Yang

Developments in Human Pluripotent Stem Cell Genome Engineering and *in situ* Sequencing Technologies

Abstract

Technology is a key driving force in the advancement of scientific discoveries. While DNA sequencing uncovered the blueprint of life encoded in the human genome, functional roles of sequence variants remain largely unknown. This thesis focuses on developing enabling technologies with broad applications for the study of genetic variations and gene regulation.

Recent advances in CRISPR/Cas9-based genome engineering technology have revolutionized biomedical research. The facilitated genome editing system employs a programmable RNA that guides the Cas9 nuclease to its target DNA. Furthermore, gene targeting in human induced pluripotent stem cells (hiPSCs) offers the unprecedented potential for dissecting gene function and correcting disease mutations to fulfill the vision of personalized regenerative medicine. Despite phenomenal progress, the efficiency of targeted modifications remained low in hiPSCs. In part I of this thesis, we developed an efficient genome editing platform by reversibly integrating doxycycline-inducible Cas9 into the genome (iPS-Cas9). We characterized and optimized critical parameters for efficient targeting, generated precise mutations for disease modeling, and demonstrated the potential of multiplexed and continuous editing. Additionally, we initiated efforts to improve homology directed repair (HDR) frequency relative to nonhomologous end joining (NHEJ)

via coupling strategies. This versatile platform enables rapid generation of mutant hiPSCs for the study of genome function and provides a test bed for further engineering of Cas9-based tools.

While DNA stores the genetic code to life, gene regulation inferred from RNA expression defines cell identity and function. Transcriptome analysis is essential for understanding developmental regulations of complex organisms by deducing gene function from expression pattern and detecting altered gene expression in disease. Traditional gene expression assays are limited by the lack of specificity, spatial context, single-cell resolution, or scalability. In part II, we explored two strategies – padlock probe (PLP) and fluorescence *in situ* sequencing (FISSEQ) – to develop highly multiplexed *in situ* RNA sequencing with single cell resolution. We concluded that PLP-based method is suitable for targeted analysis of few transcripts, while FISSEQ represents a transcriptome-wide method for *in situ* RNA profiling. The technologies presented will greatly accelerate the understanding of gene regulation in complex biological samples with broad applications in biology and medicine.

Contents

List of figures

List of tables

Acknowledgments

"Where do you come from? And where are you going?"

–Through the Looking Glass, Lewis Carroll

My journey into science is much like Alice's adventures in Wonderland – stumbling upon a fascinating world unbeknownst to me through serendipity, which to this day makes me wonder whether it has all been but a dream.

Over the past several years, I have met a number of amazing individuals no less extraordinary than the fantastical characters Alice encountered, reassuring my adventures in science will not vanish at the sound of a finger snap. Before I begin describing the work of this thesis, I would like to first reflect on how I came to be and acknowledge all who has helped me along this journey, making my pursuit of Ph.D. a reality.

My thesis work would have been impossible without my advisor, George M. Church. I thank George for taking me under his wing and supporting me with rich resources, countless troubleshooting discussions, tremendous freedom and encouragements whether in the face of success or failure. George is known for his words, "Impossible, that word is banned from our lab" and "Negative data is still data!" I thank him for always encouraging me to dream big and seek nothing but the truth. With George's exceptional passion and vision in science, I have obtained a crispr (misspelling intended) understanding of the most intriguing questions in the field. Above all, even as a man of brilliance and an overall epitome of success, George's generosity and kindness transcends all else. It has truly been an

honor to work with George during my graduate career, and he will forever be a source of inspiration and role model as a visionary scientist as well as a kind human being.

I would not have been able to sail through the waves and undercurrents of graduate research if it were not for the help of many incredible scientists in our lab. I thank Billy Li for mentoring me during my rotation, guiding me with patience and instilling in me a sense of courage to navigate in a big lab, which at the time was around 30 people. Now, I have a whole new definition of big. I thank Francois Vigneault for helping me setup the in situ sequencing project, ensuring me it was more than a crazy idea on paper. I thank Jay Lee for helpful discussions regarding the in situ sequencing project and am glad to have joined efforts with him to develop FISSEQ in its early stage. Having a comrade certainly made the seemingly endless days of troubleshooting and lamenting over poor signal in the scope room more tolerable, and allowed us to encourage each other to keep believing. I thank Luhan Yang for helping me setup the genome editing project and so much more. Our chats ranging from science to career to life goals, our silly jokes, the home-cooked meals and the dorm when I miss the last bus have all made lab life that much more fun and memorable. I cannot imagine going through the Church lab without Luhan, my classmate, labmate, and dear Pang friend. I thank Marc Guell for his patience and help with computational analysis and Raj Chari for being a great sounding board for different ideas. I thank Ben Stranges for help with crystal structure visualization and Reza Kalhor for discussions on FISSEQ specificity. I am grateful to have John Aach's input starting from the PQE proposal to the final thesis; his insightful comments have often pointed me towards the right directions. I thank Po-Yi Huang for being a great friend in life and companion in lab. In particular, I am thankful for Po-Yi, Nikolai Eroshenko, Michael Napolitano, and Bobby Dhadwar's comments during my thesis practice talk. I thank Prashant Mali, Vatsan Raman, Sri Kosuri, Adrian Briggs, Eswar

Iyer, Alex Chavez, Xavier Rios, Mike Chou, Susan Byrne, Jonathan Scheiman, and Seth Shipman for helpful discussions and advice throughout my time in the Church lab. I thank Dima Ter-Ovanesyan for being an awesome tissue culture buddy with his rather satirical sense of humor and enlightening conversations. I am grateful to be surrounded in my bay by some of the most interesting characters in the lab, including Raj Chari, Dan Mandell, and Michael Napolitano – the successor of Uri Laserson. I am also thankful to be in the company of few but amazingly energetic female scientists including Margo Monroe, Sandrine Boissel, Su Vora, Stephanie Yaung, Nili Ostrov, Jeantine Lunshof, Madeleine Ball, Julie Norville, and our former lab manager Sara Vassallo. I have also greatly enjoyed being a Teaching Fellow for iGEM with Vatsan, Dan Goodman, Jamie Rogers, Noah Taylor, and Jun Li. As a lab that seems to be growing exponentially in size, it would be impractical to acknowledge each person but instead, I would like to thank everyone for kindly offering advice, constructive criticism, encouragements or even a simple smile over the years, and for making Church lab the unique place it is.

I would not have been able to move forward with the in situ sequencing project if it were not for the kind help from the Mats Nilsson lab. I thank Mats Nilsson for inviting me to his lab to learn the technology, Rongqin Ke and Marco Mignardi for guiding me with experiments and together with Rachel Nong and Spyros Darmanis for being great hosts that made the dark and bitterly cold Swedish winter warm at heart, and to this day remain one of the most memorable moments of my graduate life.

I would not have been able to drive my thesis to completion without the guidance and feedback from my DAC committee. I thank Susan Dymecki, Jon Seidman, and John Rinn for being supportive of the turns in my thesis directions and keeping me on track for graduation. I greatly appreciate their feedback and guidance over the years. I am also

thankful to have during the final stretch, a wonderful thesis committee that includes Jon Seidman, David Frank, Ralph Scully, and Mark Bathe. Their insightful questions and comments have been illuminating and mark a memorable milestone in my graduate career.

I am thankful for the BBS and Leder Human Biology and Translational Medicine programs for providing wonderful mentors, amazing resources and courses. I thank David Cardozo for his unrelenting ability to bring a smile to my face with his cheerful ways, David Frank for being the gentlest mentor guiding me through the frantic first year in BBS, Fred Winston for the best genetics course with his deadpan humor and dedication to mentorship, Davie van Vactor for creating a warm and welcoming atmosphere in BBS, Michael Goldberg for project discussions and encouragements along the way, Gerald Greenhouse for the simply breathtaking human anatomy course that offers graduate students an opportunity to explore the wonders of the human body, and Kate Hodgins for all the candid advice and warm support in my early years. I would also like to thank Connie Cepko and Thomas Michel for creating the Leder program and organizing wonderful courses and events to enrich our graduate curriculum with a clinical angle.

I am very fortunate to have access to the amazing resources available at Harvard. Even as a graduate student in the sciences, I have been exposed to an eye-opening array of courses, seminars, and events overflowing with opportunities to learn from the best in the field, be it in science, engineering, art, or business. I would like to especially thank Daniel Henderson, Omar Thomas, and Nick Grondin for the highly energetic Jazz harmony course that wakes me up with such eagerness, and David Malan with his formidable force of CS50 staff for the fiercely beautiful introduction to computer science. I only wished I had more capacity to soak in everything here, but will be forever grateful for having once immersed in the sea of educational resources at Harvard.

My mentors at Berkeley were instrumental in helping me pave the way for graduate studies at Harvard. I thank Robert Tjian for taking me into his lab as an undergraduate with no research experience and mentoring me with advice, encouragements, and his luminary vision. I thank Wei-Li Liu for being my first official mentor in scientific experiments, grounding in me the fundamental principles of being an exceptional scientist with great attention to detail. Her patience, kindness, and rigorous training have together built a strong foundation for me. Besides being my scientific mentor, Wei-Li together with Rob Coleman also took great care of me as their own throughout my undergraduate and graduate careers. Their home in NYC has been a home away from home during my graduate life, and I cannot thank them enough for all the advice, encouragements, and support in science and in life alike throughout the past many years.

I would not have made it through Harvard while preserving my joyful spirits, or sanity even, if it were not for the many wonderful friends I have met here. I thank all my friends who I have laughed, cried, or simply had a good time with. In particular, I thank Melissa Lin, Xuyu Cai, Luhan Yang, Chewie Lin, and Alejandro de Los Angeles for being awesome BBS buddies. I thank Pei-Chen Tsung and Efan Chu for all the fun "Peabody Girls" times and Hsiao-Han Chang, Louis Liu, Tsung-Han Lin, Cheng-Sheng Lee, and Chi-Ming Chang for wonderful times exploring Boston together as the G1-forever club. I also thank many good friends that I have met through HTSA, including Ching-Fu Lin, Hai-Yin Wu, Judy Hsu, Yun-Ru Chen, and the list continues in my heart.

I am also very fortunate to have met many great friends from MIT. I thank Ya-Hui Chang, Hung-Wen Chen, and Syuan-Ming Guo profusely for the countless gatherings, trips, and candid conversations that have made my graduate life so colorful and enjoyable. Memories made in 7H, be it funny, warm, painful, or outright ridiculous, will live on in my

heart. I am also grateful to have roomed with my best friend from childhood, Wan-Yu Huang, who came at the perfect timing. Even though our times together in Boston often felt too short to be true, I have greatly enjoyed every moment of her company and still marvel at the workings of fate. I thank Jen Lee for sharing a love of jazz and philosophies in life with me, and being an awesome travel buddy. I am grateful to have met Chia-Ching Chou and Shu-Wei Chang and enjoyed board game nights in their warm home. I especially want to thank Chia-Ching for being a part of the "thesis support group", teaching me how to use word like never before and cheering each other on to reach light at the end of the tunnel. I owe many thanks to Owen Chen for always being there for me, whether for tutoring, moving, or just bubble tea-ing. I thank Ellen Guo for her sharp and witty comments that always make me laugh. I am grateful to have met Yu-Chih Ko and Emmy Lin, a beautiful couple with a lovely family that I admire dearly. I also thank Hsiang-Chieh Lee, Chien-Jen Lai, Sidney Tsai, and many other members of MIT ROCSA for all the great times, especially in the SP karaoke suite. To maintain my love of music, I have sung with MIT CCCS and Musingers, as well as jammed with Aditya Pathak, Jeff Liu and Chia-Hui Lin. I thank all the musicians I have performed with for the immense fun of creating music and for keeping my dream alive.

I may very well have never stepped into science if it were not for Sam, that special friend of mine since the days at Berkeley. It was him that encouraged me to look beyond the business major and taught me the earliest lessons in chemistry and physics. I still cannot help but smile at the fond memory, although admittedly a bit nerve-wracking at the time, of my first assignment to memorize the periodic table. A rigorous teacher he was. After sparking my curiosity in science, he soothed the rocky start by supporting me tirelessly, more than sufficient to earn him a 24/7 GSI award if one existed, and providing much-needed

encouragements through times of struggle. I thank him for always believing in me and giving me strength to take a leap of faith. His strong will, passion for science, and uncompromising work ethic to the point of perfectionism have all been a source of inspiration and have driven me to exceed my own limits. Of course, precious memories of our travels and wanderings, particularly the road trips with in-car karaoke, will forever live in my heart. From Berkeley to Boston, we have grown together in the most formative years of our lives, and I am very thankful to have your cherished company and support along this journey.

Finally, I would never have made it so far if it were not for my loving family. I thank my parents for bringing me into this world and to the land of opportunity, for always providing me with the best they can afford, and for being supportive in all that I choose to pursue. I thank my brother for being the best partner in crime since childhood and for supporting each other through thick and thin. I realize how fortunate I am to have a loving and caring family, and will be forever grateful for their unconditional love that shaped me into who I am today.

Now looking back in time, I see how all the dots are connected – where I came from, how I came to be – and I give a huge heartfelt THANK YOU to everyone for all the kindness and support you have shown me over the years. I may not know where the path will lead me next, just as I never would have imagined being here 10 years ago, but as Steve Jobs once advised, I will follow my heart and trust that the dots will eventually connect in the future. After the historic winter of 2015, comes an especially beautiful spring. As the Charles flows, the flowers blossom, and Boston gives way to spring, I wave goodbye to Wonderland with all my might while hiding a hint of nostalgia, confident that it has been more than a curious dream, then embark on the next chapter in life to compose the adventures of a whole new journey.

-Joyce

Part I:

Human Pluripotent Stem Cell Genome Engineering Technology

CHAPTER 1 Introduction

Science and technology go hand in hand in the pursuit of knowledge and human welfare. While scientific breakthroughs lead to technological advancements, technologies often drive further scientific discoveries. In the field of biology, the discovery of the DNA double helix structure in 1953 marked a historic milestone that has empowered decades of ensuing research in analyzing, synthesizing, and manipulating DNA up to the present day[1–3]. In the 1970s, the development of recombinant DNA technology further enabled scientists to isolate genes for laboratory study and applications in biotechnology and medicine. With the invention of genome sequencing technologies, the Human Genome Project aimed at mapping the entire human genome came into reality, setting the stage for identifying genetic roots to variations in developmental traits and diseases[4]. As anticipated, the human genome is highly complex and variable among individuals as over 1.4 million single nucleotide polymorphisms (SNPs) have been identified, contributing to personal traits and human evolution. Given the wealth of genetic information available, the next step lies in elucidating gene functions and pinpointing causal genetic variants. Recent advances in genome engineering technologies, particularly the repurposing of bacterial CRISPR-Cas9 (clustered

regularly interspaced short palindromic repeats - CRISPR-associated protein 9 nuclease) system for genome manipulation, have set forth a new revolution in biomedical research. It is now possible to introduce targeted changes into the genome of living cells and organisms in their endogenous context, facilitated by the ease of use and efficiency of the CRISPR-Cas9 system. The ability to make site-specific modifications to the genome holds tremendous promise in scientific research, biotechnology and medicine through the dissection of gene functions, engineering of useful biological systems for industrial production, and correction of genetic diseases for gene therapy applications.

1.1 **Brief Overview of Genome Engineering**

Early approaches of manipulating the eukaryotic genome made use of homologous recombination (HR) to introduce exogenous repair templates that consist of sequence homology to the target site along with the desired modifications into the genome. Although HR-mediated gene targeting enabled the construction of knockin and knockout animal models, the overall low efficiency of recombination events (1 in 10^6 – 10^9 cells) prevented wide adoption of the technology[5].

A series of studies led to the discovery that the creation of a DNA double-stranded break (DSB) at the target genomic locus could greatly enhance genome editing efficiency[6–9]. DSBs represent one of the most critical forms of DNA damage, thus cells have developed efficient DNA repair mechanisms to maintain genome integrity and cell survival. Classically, two major DSB repair pathways have been defined: the error-prone nonhomologous end-joining (NHEJ) and faithful homology-directed repair (HDR). As NHEJ directly ligates the broken DNA ends, repair through this pathway can induce insertion/deletion mutations

(indels) that disrupt gene function. On the other hand, HDR uses an undamaged homologous sequence to serve as donor template for repair. Therefore, HDR-mediated repair can be manipulated to induce desired modifications through the incorporation of exogenously supplied homology repair donor templates into the target locus.

1.1.1 Programmable nucleases for genome editing

To take advantage of enhanced genome editing with the introduction of site-specific DSBs, four major classes of programmable nucleases have been developed including meganucleases[10], zinc finger nucleases (ZFN)[11–13], transcription activator-like effectors nucleases (TALEN)[14–16], and RNA-guided nucleases represented by the type II bacterial adaptive immune CRISPR-Cas9 system[17–19]. Meganuclease, ZFN, and TALEN bind to specific DNA sequences through protein-DNA interactions. Meganucleases are engineered from naturally occurring restriction enzymes with extended DNA recognition sequences (14-40 basepairs), whereas ZFNs and TALENs are artificial fusion proteins consisting of a customizable DNA binding domain fused to a nonspecific nuclease domain from the restriction enzyme FokI, with each ZF and TALE module recognizing 3 and 1 nucleotides (nt) of DNA, respectively. Although these platforms have made important advances in the genome editing field, each faces its own set of limitations. Meganucleases have been challenging to engineer because the DNA recognition and cleavage functions are entwined in a single domain. On the contrary, ZFNs and TALENs are composed of distinct DNA binding and FokI cleavage domains, thus allowing scientists to engineer the proteins with customized DNA-binding specificities. However, crosstalk between individual ZF domains in an array leads to context-dependent binding preferences, making the robust construction of ZFNs limited to few laboratories with the relevant expertise[20]. Likewise, while TALE

repeat domains have less context-dependent effects, the assembly of TALE repeats can still be challenging and the highly repetitive nature of TALEN-coding sequences may also hinder their delivery with viral vectors such as lentiviruses[21]. Given the challenges in engineering programmable DNA-binding proteins, the recent development of RNA-guided nucleases based on a bacterial CRISPR-associated protein 9 nuclease from *Streptococcus pyogenes* (Cas9) represent a significant breakthrough in genome engineering technologies. Instead of targeting DNA through protein-DNA interactions, Cas9 recognizes its targets through simple Watson-Crick base pairing guided by an engineered RNA. The simplicity and efficiency of the Cas9 system have led to its broad adoption in applications ranging from basic research to biotechnology and medicine.

1.1.2 CRISPR-Cas9 biology: from immune defense to genome targeting

While the CRISPR-Cas9 revolution dawned on the genome editing field in 2013, the discovery of these elements unfolded in 1987 when a group of Japanese scientists observed a set of 29 nt repeats while studying the *iap* enzyme in *E. coli*[22]. With decades of basic research to understand the biological function and mechanism of the mysterious repetitive elements known as CRISPR (clustered regularly interspaced short palindromic repeats), it finally became clear that CRISPR systems function as an adaptive immune defense system that safeguards organisms from invading foreign nucleic acids, such as viruses and plasmids[23–26].

The genomic CRISPR loci consist of an array of identical repeats interspersed with invader DNA-targeting spacers and an operon of *cas* genes[27,28]. CRISPR/Cas-mediated immunity occurs in three stages. First in the adaptive phase, bacteria and archaea react to invading foreign DNA by integrating short fragments of foreign nucleic acid (protospacers)

into the host chromosome at the proximal end of the CRISPR array. In the expression phase, CRISPR loci are transcribed into precursor CRISPR RNA (pre-crRNA) and further processed into a library of short CRISPR RNAs (crRNAs) that can target complementary sequences from invading viral or plasmid DNA[29–34]. Finally, in the interference phase, crRNAs together with trans-activating crRNA (tracrRNA) and Cas proteins, form ribonucleoprotein complexes that detect and destroy foreign sequences [35–40].

A key insight for repurposing the bacterial defense mechanism for genome engineering came in 2012, demonstrating the type II CRISPR system from *Streptococcus pyrogenes* can be engineered to induce Cas9-mediated double stranded breaks in a sequence-specific manner in vitro by providing a synthetic single guide RNA (sgRNA) composed of a crRNA-tracrRNA fusion[41]. Before long, the CRISPR-Cas9 system has been successfully engineered to function in human cells with the use of human codon-optimized Cas9 and customizable 20-nt sgRNAs[17–19]. In Cas9-mediated genome editing, the sgRNA first identifies its 20-bp target followed by a NGG PAM (protospacer-adjacent motif) sequence, subsequently Cas9 nuclease cleaves the target sequence and creates a DSB for DNA repair through the HDR or NHEJ pathway[41]. This RNA-guided two component system greatly facilitates genome engineering as the variable DNA recognition component is dictated by RNA sequences that are easy to design, synthesize, and deliver, holding tremendous promise in a broad range of applications.

The powerful Cas9 genome editing technology has been utilized in a wealth of applications ranging from generation of cellular and animal models or genome-wide perturbation experiments to study gene function and causal genetic variants, as well as correction of genetic mutations in inherited disorders. Furthermore, a catalytically deactivated version of Cas9 (dCas9) can be used as a modular RNA-guided DNA binding

protein, and repurposed for transcriptional regulation and live cell imaging when fused with an appropriate effector protein[42–47].

1.2 Promise of genome engineering in pluripotent stem cells

Cas9-mediated gene targeting has been widely used as a research tool, but perhaps one of the most tantalizing future directions is the development of Cas9-based therapy for treatment of genetic disorders. Building precise and tractable disease models is the first step in understanding the molecular mechanisms of disease pathogenesis, hence driving the development of novel therapies. As stem cells have the ability to self-renew and differentiate, they represent an ideal system for creating disease models and downstream applications in drug development or transplantation therapies. The first successful isolation of human embryonic stem (ES) cells from blastocysts by Thomson and colleagues in 1998 stimulated great excitement for the prospect of understanding human development and disease mechanisms, as well as developing therapeutic applications. However, the derivation of ES cells from human embryos has provoked controversy in the United States over ethical concerns, resulting in restricted government funding and limitations on cell lines approved for research use[48]. In 2006, Takahashi and Yamanaka demonstrated in a landmark study that induced pluripotent stem (iPS) cells can be derived directly from mouse somatic cells through the ectopic co-expression of four reprogramming transcription factors[49]. With the promise of an alternative source for stem cell research, a flurry of follow-up studies translated the results into human fibroblasts and a number of other cell types[50–52]. Besides circumvention of ethical issues, a key advantage of human iPS cells is the potential to

generate patient-autologous iPS cell derivatives free from immune rejection for transplantation therapy.

The use of genome engineering methods can further enhance the potential of human iPS cells, as the synergy of the technologies offers a promising step toward precise disease modeling and personalized cell therapy. In disease modeling, gene targeting can create isogenic human iPS cell lines that differ only at specific loci of interest, thus enabling the dissection of gene or mutation function free from confounding effects of the genetic background. From the therapeutic angle, precise gene editing can be utilized to correct disease-causing mutations in patient iPS cells and differentiated into the appropriate cell type for autologous cell transplantation therapy, a concept heralded as the future of regenerative medicine. Given the potential of multiplexed gene targeting by Cas9, the study and treatment of complex diseases is also within reach.

Despite having made great strides in genome engineering and iPS biology, several challenges lie ahead. First, the efficiency of delivery and targeting in human iPS cells needs to be improved to facilitate simple and efficient disease modeling in the iPS system. In addition, to realize its therapeutic value for a broad range of genetic disorders, the choice of DSB repair pathway will have to shift towards HDR. Finally, before Cas9-genome edited iPS cells can reach translational clinic, it will be of outmost importance to thoroughly characterize the safety and long-term implications. To harness the potential of genome engineering in human iPS cells, we proceeded to develop an efficient platform for gene targeting in human iPS cells and devised strategies to enhance the HDR efficiency.

1.3 **Thesis Outline – Part I**

In part I of the thesis, we describe the development of an efficient genome engineering platform in human iPS cells. CHAPTER 2 focuses on the investigation of gene targeting rate in iPS cells, development of an efficient targeting system, characterization and optimization of critical parameters for efficient gene editing. We then apply the knowledge gained to experimental design in CHAPTER 3, targeting disease mutations identified in genetic disorders, investigating the potential of multiplexed and continuous editing, devising strategies to further improve the HDR to NHEJ ratio, and exploring the underlying cause of variable HDR:NHEJ ratio observed at different genomic target sites.

The experiments presented in part I of this thesis demonstrate a simple, efficient, and multiplexable genome editing technology in human iPS cells. Efforts made in promoting HDR also provide insights for future designs. Overall, the versatile platform provides a valuable resource for rapid generation of mutant human iPS cells to dissect gene functions and pinpoint causal disease variants in an isogenic background that can also be scaled for high-throughput analysis.

CHAPTER 2 Method Development for Efficient Human Stem Cell Genome Editing via CRISPR/Cas9 System

2.1 Introduction

Targeted human genome editing enables functional studies of genetic variation in human biology and disease, holding immense potential in therapeutic applications. With the advent of human induced pluripotent stem cells (hiPSCs), it is now possible to reprogram fibroblasts to a human embryonic stem cell (hESC)-like state with maintained pluripotency, self-renewal and differentiation capacity, paving way for the future of developmental biology studies, drug testing, and regenerative medicine[49–54]. Genome engineering technologies can further enhance the value of hiPSCs through the creation of disease models or correction of genetic mutations in its original context, allowing the dissection of gene function and disease mechanism in an isogenic background. To harness the full potential of hiPSCs, methods for simple, efficient, and precise genetic manipulations are needed.

Recently, the type II bacterial CRISPR (clustered regularly interspaced short palindromic repeats)/Cas (CRISPR-associated) system has been developed as a versatile genome editing technology in eukaryotic cells and whole organisms [17–19,41,55–59]. The type II CRISPR system from *Streptococcus pyrogenes* has been engineered to induce Cas9-mediated double stranded breaks (DSBs) in a sequence-specific manner in vitro, guided by a synthetic guide RNA composed of crRNA fused to tracrRNA [41]. Moreover, the system has been successfully adapted to function in human cells with the use of a human codon-optimized Cas9 and programmable 20-nt single guide RNA (sgRNA) [17–19,60]. Once delivered into the cells, the sgRNA identifies its genomic target followed by a 5'-NGG-3' PAM (protospacer-adjacent motif) sequence, Cas9 nuclease then cleaves the target sequence creating a DSB [41]. The resulting DSB will either be repaired by the error-prone NHEJ (non-homologous end joining) pathway resulting in non-specific mutations disrupting gene function, or the HDR (homology directed repair) pathway generating specific modifications dictated by an exogenous repair donor template [61–63]. This two-component system greatly facilitates genome engineering through the creation of targeted DSBs guided by RNA sequences that are easy to design, synthesize, and deliver.

While the Cas9 system has demonstrated high gene targeting efficiency in multiple model systems, the overall efficiency in human iPS cells remained low[17]. To achieve simple, efficient, and easily multiplexable genome editing in hiPSCs, we have developed a genome engineering platform with an inducible Cas9-integrated iPS cell line. We created the doxycycline-inducible iPS Cas9 cell line using a piggyBac transposon system, for the prospect of scarless genome editing. We proceeded to develop a complementary lipid-based transfection approach for genome editing due to its low cytotoxicity compared to electroporation and potential for multiplexed gene editing. Finally, we characterized key

parameters guiding the design for highly efficient single-stranded oligodeoxynucleotide (ssODN)-mediated genome targeting. We present this versatile platform as a valuable resource for rapid generation of mutant human iPS cells for the dissection of gene functions and identification of causal disease variants in an isogenic background with the potential to be scaled for high-throughput analysis.

2.2 **Results and Discussion**

2.2.1 Efficiency of Cas9-mediated gene editing in human iPS cells

The type II bacterial CRISPR-Cas9 system has been successfully adapted for sequence-specific genome editing of mammalian genomes[17,64]. Given the promising demonstration of targeting efficiency and ease of design[17,18], we sought to further expand the toolbox to enable efficient genome editing in human iPS cells. As an initial study of gene targeting efficiency in PGP1 human iPS cells, we delivered via nucleofection the human codon-optimized Cas9 protein with a C-terminal SV40 nuclear localization signal cloned into a mammalian expression system[17], along with a single-guide RNA (sgRNA) expressed from the human U6 polymerase III promoter and a corresponding HDR donor oligo to each genomic target of interest. For gene editing experiments introducing precise SNPs or small alterations to the genome, we chose to use single-stranded oligodeoxynucleotide (ssODN) as the HDR donor template because of its small size, ease and flexibility in design, and demonstrated efficiency[61].

As a starting point, we designed sgRNAs to target two loci in the adenosine deaminase (ADA) gene – ADA exon 7 (ADA7) and ADA exon 10 (ADA10) following the

form of 5'-GN(19)NGG-3'. We then designed corresponding 90-nt HDR donor oligos for each site, introducing a GGG to AGG transition mutation in exon 7 and a frameshift mutation (Del(GAAGA)) in exon 10, both of which have been identified as molecular defects in an adenosine deaminase deficiency-related severe combined immunodeficiency (ADA-SCID) patient iPS cell line[65] (Figure 2-1).

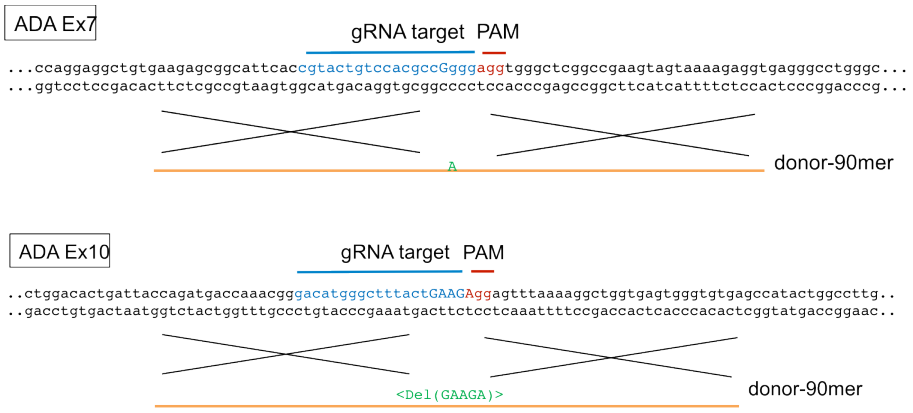

Figure 2-1: Design of ADA7 and ADA10 sgRNA and HDR donor oligos.

The HDR donor oligos have been designed in such a way that the mutations also disrupts the sgRNA or PAM sequence, so that once the donor oligo has been successfully incorporated, the site should no longer be targeted for further editing. To assess gene editing efficiencies in human iPS cells, we co-nucleofected the Cas9 plasmid, sgRNA construct, and HDR donor oligo into PGP1 iPS cells. Three days post-nucleofection, we harvested the cells and analyzed the gene editing rates using the MiSeq Personal Sequencer. Deep sequencing of the ADA7 and ADA10 loci showed indel formation near the predicted cutting site (Figure 2-2) with 5.98% NHEJ at the ADA7 locus and 2.93% NHEJ at the ADA10 locus, consistent with earlier results targeting human iPS cells[17]. The HDR efficiency on the other hand was virtually undetectable, with 0.61% at the ADA7 locus and 0.38% at the ADA10 locus. This suggested that Cas9 was functional and facilitated indel formation at the target site in

combination with sgRNA, but was not an efficient system for introducing specific mutations via HDR in its current state. The variable and generally low gene editing efficiencies in human pluripotent stem cells have also been observed in other studies[17,66], highlighting the need for a more efficient method of Cas9-mediated gene targeting in hiPSCs.

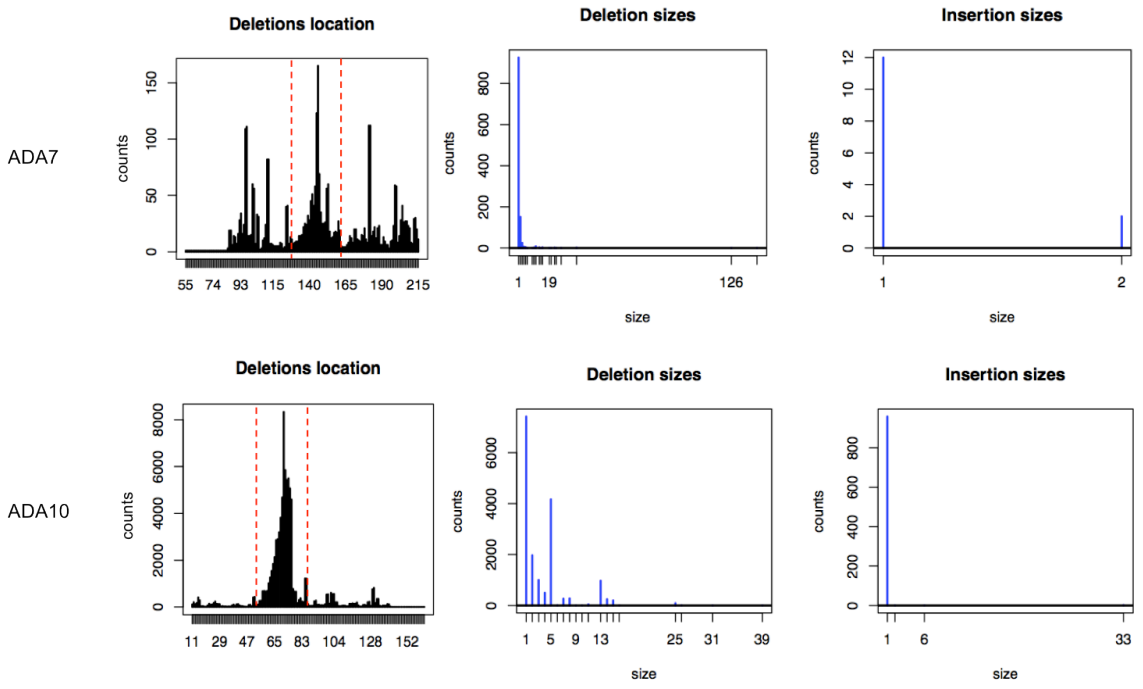

Figure 2-2: NHEJ profile of ADA7 and ADA10 target site, showing indel formation near the expected cleavage site (marked by the red dashed lines) and the sizes of indels.

While there have been improvements in several gene delivery methods, transfecting DNA into human stem cells remained an often capricious process[67]. To understand the low gene targeting rate in the initial test, we next examined the nucleofection efficiency of Cas9 protein to determine whether it posed as a rate limiting step. We delivered a Cas9-GFP plasmid and a control pmax GFP plasmid into PGP1 iPS cells through nucleofection. Two days after nucleofection, we used fluorescence-activated cell sorting (FACS) to identify the percentage of cells with GFP-positive signal, indicating successful delivery of the construct.

FACS analysis revealed that delivering and expressing pmaxGFP positive control plasmid was 4X more efficient than Cas9-GFP plasmid, suggesting that the delivery of the relatively large Cas9 proteins may be a barrier to efficient genome editing (Figure 2-3).

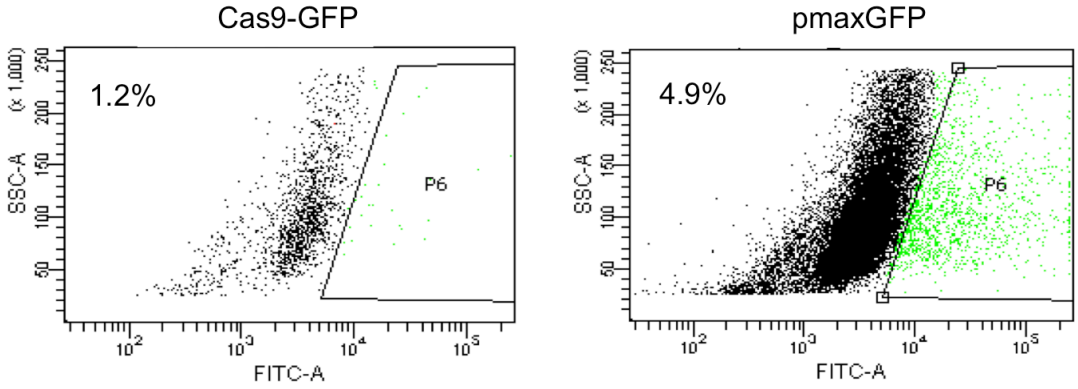

Figure 2-3: Nucleofection efficiency of Cas9-GFP and pmaxGFP (positive control) plasmid.

In addition to the ADA loci, we targeted Shwachman-Bodian-Diamond syndrome (SBDS), a congenital disorder that affects the bone marrow, pancreas, and skeletal system, to test the possibility of isolating clonal populations of edited iPS cells from Cas9-mediated gene targeting. We designed sgRNAs and 90-nt HDR donor oligos to introduce two point mutations found in a SBDS patient cell line, namely a T to C mutation at the intron 2 splice donor site (IV2 + 2 T > C) and a G to A mutation just before exon 3 (IV3 – 1G > A)[65]. To introduce the desired mutations, we co-nucleofected the Cas9 plasmid, sgRNA construct, and HDR donor oligo into PGP1 iPS cells. One week after nucleofection, the transfected cells were single-cell sorted by FACS into 96-well plates and allowed to grow into monoclonal colonies for one week. Finally, single iPS colonies were harvested and screened by Sanger sequencing to identify clones with the desired mutations. Since human iPS cells have poor viability as single cells, we have optimized the single cell FACS sorting and culture conditions, as well as a rapid genotyping system to enable large-scale genotyping of edited

iPS clones without selection[68]. Results from Sanger sequencing identified one clone out of 43 screened with biallelic modifications (2.3%) and one with monoallelic modification (2.3%) at the SBDS – IV2 loci (Figure 2-4).

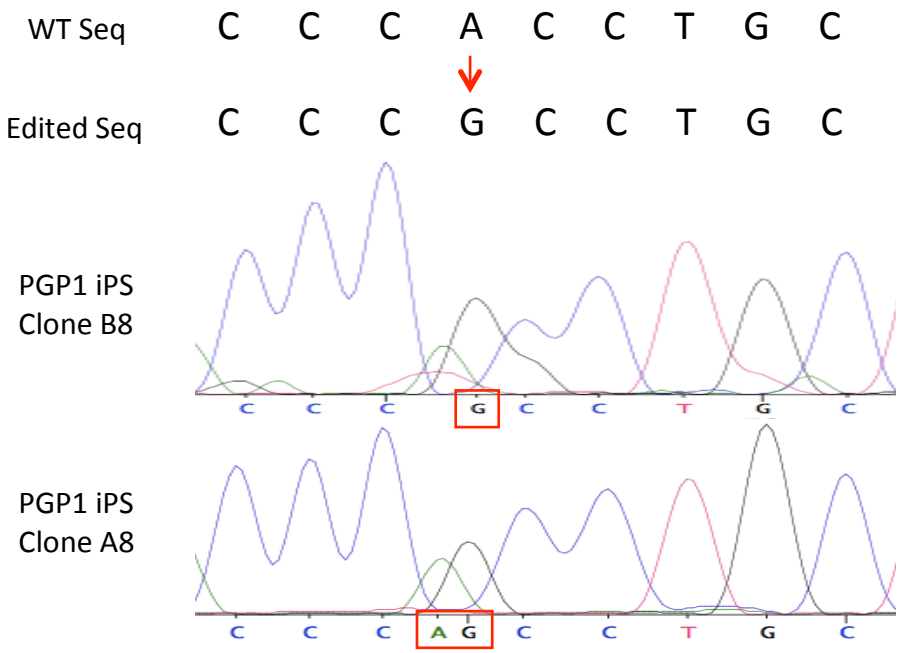

Figure 2-4: Sanger sequencing identified 2 SBDS – IV2 clones with correctly edited sequences.

While successfully edited clonal iPS cells have been isolated for the SBDS – IV2 target site, Sanger sequencing of 79 SBDS – IV3 targeted colonies, 78 ADA7 targeted colonies, and 12 ADA10 targeted colonies failed to identify any clones with the desired modification. Therefore, in our initial test of Cas9-mediated genome editing in human iPS cells, we observed that the Cas9 system was functional and introduced indels at the target site, but was generally low in editing efficiency consistent with literature[17], in part due to the inefficient delivery of Cas9 plasmid into the cells. While the HDR targeting efficiency was limited, we were able to isolate human iPS clones of correctly edited cells for the SBDS – IV2 target without the use of drug selection. However, the low targeting efficiency required a

laborious and time-consuming screening process to identify the rare edited clones. Hence, this prompted us to develop a more efficient system for Cas9-mediated gene targeting in human iPS cells.

2.2.2 Development of iPS-Cas9 platform for efficient genome editing in human iPS cells

Given the low efficiency of Cas9-mediated genome editing in human iPS cells, we sought an alternative strategy to enhance the gene targeting potential. Based on the observation that delivery of Cas9 plasmid by nucleofection was inefficient, we reasoned that integrating Cas9 into the genome with the PiggyBac (PB) transposon system may lift this barrier altogether. The PB transposon is a mobile genetic element that can be efficiently moved between vectors and chromosomes through a cut and paste mechanism mediated by the PB transposase enzyme. During transposition, the PB transposase recognizes inverted terminal repeat sequences (ITRs) flanking the ends of the transposon vector, and catalyzes the excision and insertion event, moving the contents between ITRs on the PB transposon into TTAA chromosomal sites. By integrating Cas9 into the genome, we obviate the need to deliver the large Cas9 plasmid into iPS cells for each experiment, hence avoiding the inefficient delivery problem. Additionally, an elegant feature of the PB transposon system is that the integrated cassette can be seamlessly removed by the PB transposase. Therefore, once the editing experiments have been completed and successfully modified cells isolated, theoretically the integrated Cas9 cassette can be removed from the target genome without leaving a trace. A schematic of the genome editing roadmap using the genomically integrated Cas9 iPS system is presented in Figure 2-5. As a powerful tool for potential applications in gene therapy and regenerative medicine, the PB transposon system has already been adapted

for use in a broad range of cell types, including the process of reprogramming human embryonic fibroblasts into human induced pluripotent stem cells[69,70].

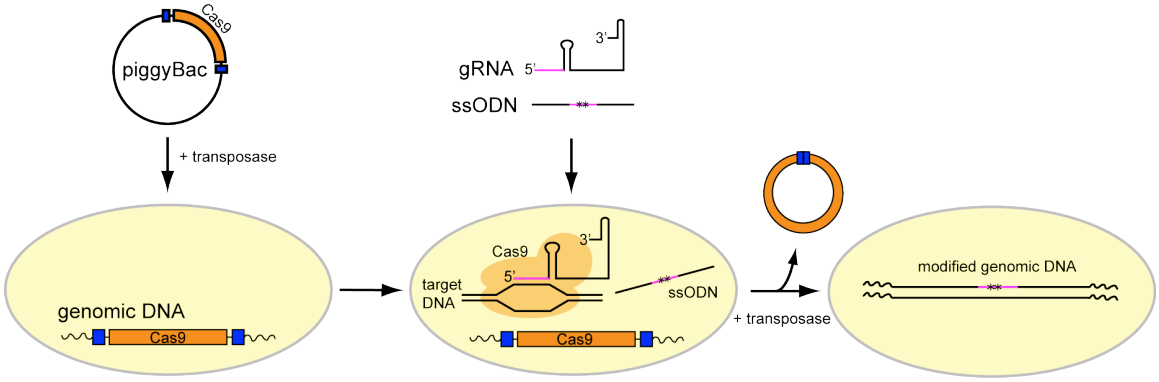

Figure 2-5: Overview of genome editing roadmap using the genomically integrated Cas9 iPS cell line.

Driven by the potential of enhancing genome editing efficiencies, we set out to build an iPS cell line with genomically integrated Cas9 (iPS-Cas9) with the prospect of ultimately enabling efficient scarless genome editing in human iPS cells. We constructed the PGP1 iPS-Cas9 cell line by inserting a reverse tetracycline-controlled transactivator (rtTA) and a human codon-optimized Cas9 under the control of a tet response element (TRE) into PGP1 iPS cells via the piggyBac transposon. The system allows Cas9 expression to be tightly controlled and activated only by the addition of doxycycline (a tetracycline derivatives) into the culture media. The transfected cells were selected with puromycin and Cas9 integration was confirmed by qPCR and Sanger sequencing.

Having established the iPS-Cas9 cell line, we next sought to develop a simple and compatible method to introduce sgRNAs and HDR donor oligos into these cells for genome editing experiments. Several methods exist for delivering nucleic acids into cultured mammalian cells, including electroporation, nucleofection, transfection via cationic lipid vehicles, and lentiviral vectors, while sgRNAs can be delivered in the form of RNA or

plasmid DNA. Given the potential of *in vivo* RNA delivery for gene therapy applications and risk of insertional mutagenesis of DNA-based approaches, we chose to deliver sgRNAs in the *in vitro* transcribed RNA form to provide a simple, non-mutagenic, and highly flexible system for testing various designs. Moreover, we broadened the utility of our system by coupling the dox-inducible Cas9 cell line with a lipid-based transfection method, with the reasoning that this would allow for repeated and/or multiplexed transfections, potentially leading to higher genome editing efficiencies. Therefore, we proceeded with developing a lipid-based transfection method directly delivering *in vitro* transcribed RNAs and ssODN donor oligos into PGP1 iPS-Cas9 cells.

We first tested the feasibility of RNA transfections in the PGP1 iPS-Cas9 cell line by transfecting EGFP mRNA at various concentrations to gauge the efficiency and begin optimizing transfection conditions. We transfected modified EGFP mRNA ranging from 2.4 pmol to 36 pmol using Lipofectamine-based transfection reagents and measured the percentage of GFP positive cells by FACS one day after transfection. GFP expression was detected across the spectrum of mRNA concentrations tested, with increasing concentrations resulting in higher percentages of GFP-positive cells as expected (Figure 2-6).

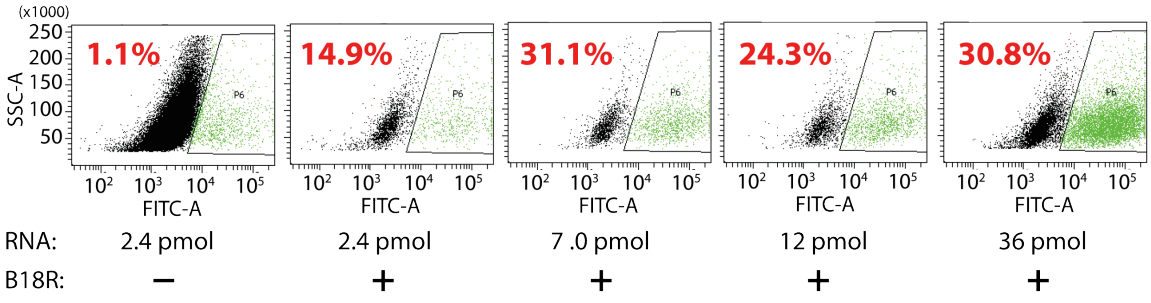

Figure 2-6: Initial assessment of RNA transfection efficiency in PGP1 iPS-Cas9 cells.

In addition, we tested the role of B18R recombinant protein in RNA transfection of human iPS cells. B18R protein is a Vaccinia virus-encoded protein that acts as a decoy receptor for Type I Interferons and has been found to increase cell viability during RNA transfection protocols for reprogramming somatic cells to pluripotency[71]. It was interesting to note, the addition of recombinant B18R protein greatly enhanced the percentage of GFP-positive cells, suggesting that media supplementation with B18R was also crucial for enhancing cell viability during RNA transfection of human iPS cells.

2.2.3 Impact of RNA modifications on gene targeting efficiency

Earlier experiments have shown that introducing exogenous single-stranded RNA (ssRNA) into mammalian cells may trigger innate immune reactions through the interferon and NF-κB-dependent pathway[72–76]. It is also known that eukaryotic mRNA is modified extensively *in vivo*, where the modified nucleobases have been observed to reduce response from the ssRNA sensor RIG-I and endosomal ssRNA sensors TLR7 and TLR8[77–79]. In order to reduce the innate antivral defense response against transfected RNA, several strategies have been employed to modify RNAs. Based on an earlier study reprogramming human cells to pluripotency with synthetic modified mRNA, the RNA modifications included incorporating a 5' guanine cap to stabilize RNA, synthesizing RNA with modified ribonucleoside bases substituting 5-methylcytidine (5mC) for cytidine and pseudoruidine (psi) for uridine to improve cell viability, and treating synthesized RNA with phosphatase to remove 5' triphosphates to prevent signaling by RIG-I[71]. As the combination of the RNA modifications approach outlined above has shown to evade the innate antivral surveillance system, we incorporated these RNA modifications into our initial sgRNA design and synthesis.

To test the functionality of the RNA transfection system on PGP1 iPS-Cas9 cells, we first targeted the AAVS1 locus located in the PPP1R12C gene on chromosome 19. The AAVS1 locus served as an ideal proof of concept target site because it has been shown to be a transcriptionally active safe harbor with no known adverse phenotypic effects from disruption of the site. To begin, we designed an sgRNA targeting the AAVS1 locus following the form of 5'-GN(19)NGG-3' along with a corresponding 70-nt HDR donor oligo with a CC to GG 2 base pair (bp) mismatch in the middle (Figure 2-7).

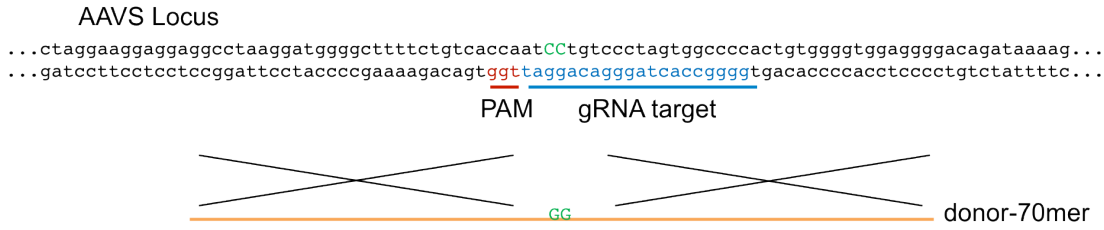

Figure 2-7: Schematic of AAVS1 sgRNA target and HDR donor oligo design.

The HDR donor oligos have been designed in such a way that the mutations also disrupts the sgRNA sequence, so once the donor oligo has been successfully incorporated, the site should not be subjected to further editing. We synthesized the sgRNAs by *in vitro* transcription following the modifications outlined above.

To assess the genome editing efficiency of human iPS-Cas9 cells using the RNA transfection approach, we co-transfected *in vitro* transcribed sgRNA at various concentrations and HDR donor oligo targeting the AAVS1 site into dox-induced PGP1 iPS-Cas9 cells in B18R supplemented media. Three days post-transfection, we harvested the cells and analyzed the gene editing rates using the MiSeq Personal Sequencer. Deep sequencing results showed no clear indel formation near the predicted cutting site and no detectable HDR above background levels (Figure 2-8).

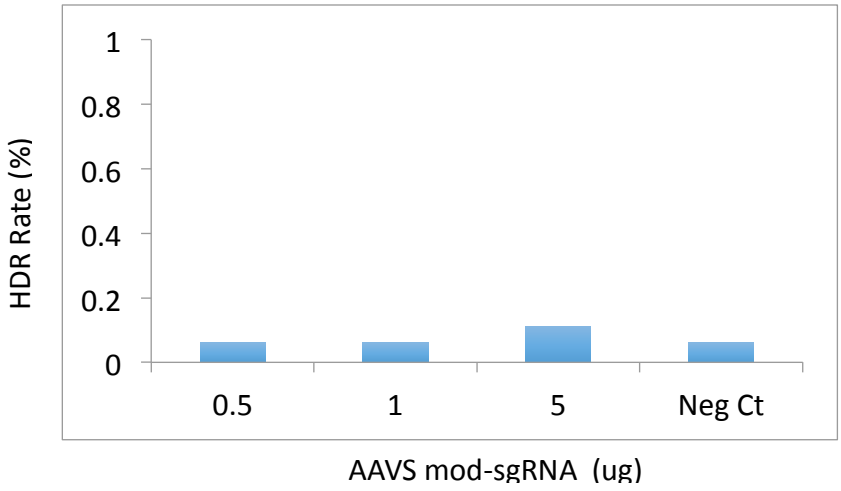

Figure 2-8: HDR efficiency of modified sgRNA transfection in PGP1 iPS-Cas9 cells

In the same experiment, using nucleofection to deliver AAVS1 sgRNA under the U6 promoter in the DNA form along with the HDR donor oligo into dox-induced PGP1 iPS-Cas9 cells with B18R media supplementation showed 2.35% HDR rate and indel formation around the predicted targeting site (Figure 2-9).

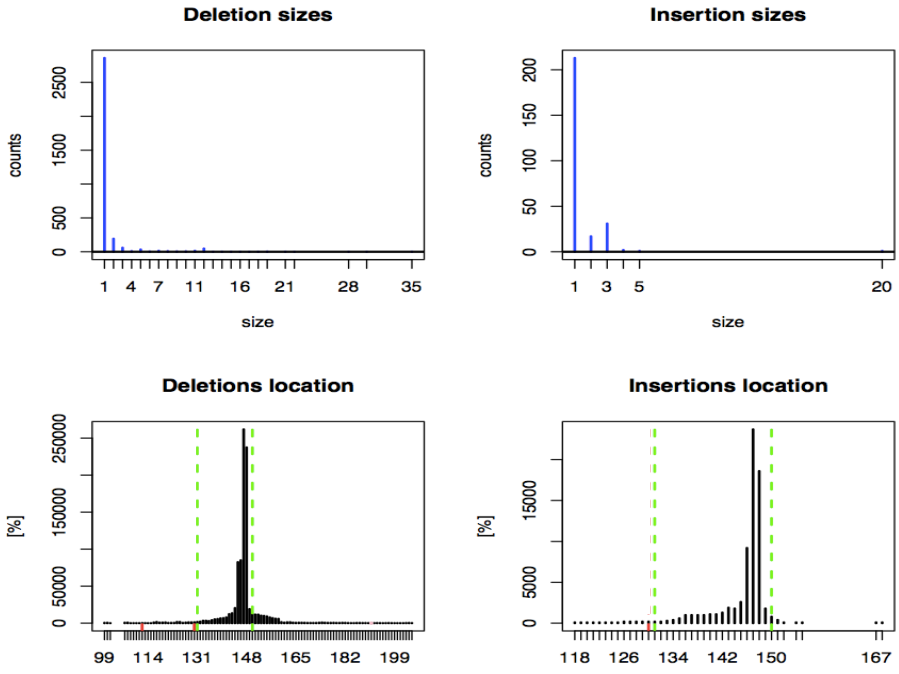

Figure 2-9: NHEJ profile of AAVS1 sgRNA targeting via nucleofection.

Together, the results demonstrated that the genomically-integrated Cas9 was functional and facilitated indel formation via the nucleofection method, while genome editing with the RNA transfection method was not optimal in the current configuration.

To diagnose potential problems with the RNA transfection method, we focused on the sgRNA as it was the main component that differed in the nucleofection versus transfection methods. First, we verified the sequence of the PCR template used for *in vitro* transcription of the AAVS1 sgRNA by Sanger sequencing. Next, we checked the size of the *in vitro* transcribed modified sgRNA along with sgRNAs carrying varying degrees of modifications on a 10% denaturing polyacrylamide gel electrophoresis (PAGE) gel against a low range ssRNA ladder to see whether the modifications affect the sgRNA size. All four versions of the AAVS1 sgRNA produced a major product band around 100 bp, at the expected sgRNA size (Figure 2-10).

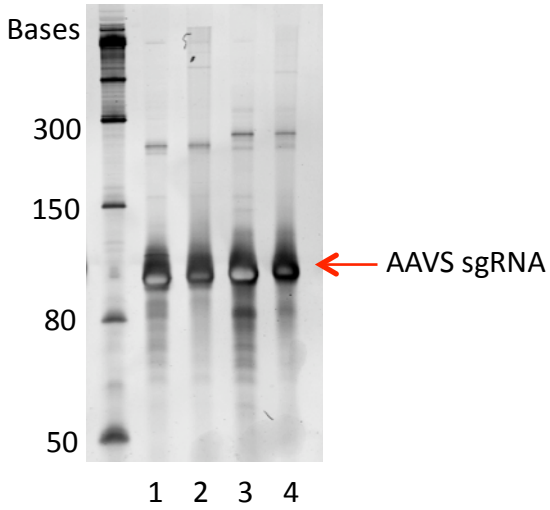

Figure 2-10: Assessment of AAVS1 sgRNA size on a 10% denaturing PAGE gel. 1) Native sgRNA. 2) Native sgRNA with 5' guanine cap. 3) Modified sgRNA. 4) Modified sgRNA with 5' guanine cap.

Having verified the sgRNA sequence and size, we next inquired whether the RNA modifications play a role in inhibiting Cas9-mediated gene editing. For instance, although 5' guanine capping is known to prevent degradation by exonucleases hence promoting RNA half-life, it also serves as a nuclear export signal to shuttle mature messenger RNA (mRNA) to the cytoplasm for translation[80,81]. Moreover, despite the crystal structure of Cas9 in complex with sgRNA was not yet solved at the time of these experiments, it was known that sgRNA interacts with Cas9 to guide it to the target DNA. Therefore, incorporating modified ribonucleoside bases into sgRNA posed the risk of disrupting its interactions with the Cas9 protein, potentially resulting in low cutting efficiency.

To evaluate the effect of RNA modifications on Cas9-mediated genome editing efficiency, we generated four versions of sgRNA – native (Table 2-1), native with 5' cap (Table 2-2), modified (Table 2-3), and modified with 5' cap (Table 2-4)– targeting the AAVS1 locus for comparison. We co-transfected each form of the AAVS1 sgRNA along with the HDR donor oligo into dox-induced PGP1 iPS-Cas9 cells in B18R supplemented media. Three days post-transfection, we harvested the cells and analyzed the gene editing efficiency by next-generation sequencing. Deep sequencing results revealed efficient indel formation and editing rates at the predicted AAVS1 target site for sgRNA in the native form, with 30.4% HDR and 40.7% NHEJ. Furthermore, we observed reduced efficiency with the 5' capped native sgRNA at 17.2% HDR and 21.6% NHEJ, and considerably lower gene targeting rates with modified sgRNA at 1% HDR and 0.99% NHEJ, and 5' capped modified sgRNA at 1.33% HDR and 1.04% NHEJ, consistent with earlier results (Figure 2-11).

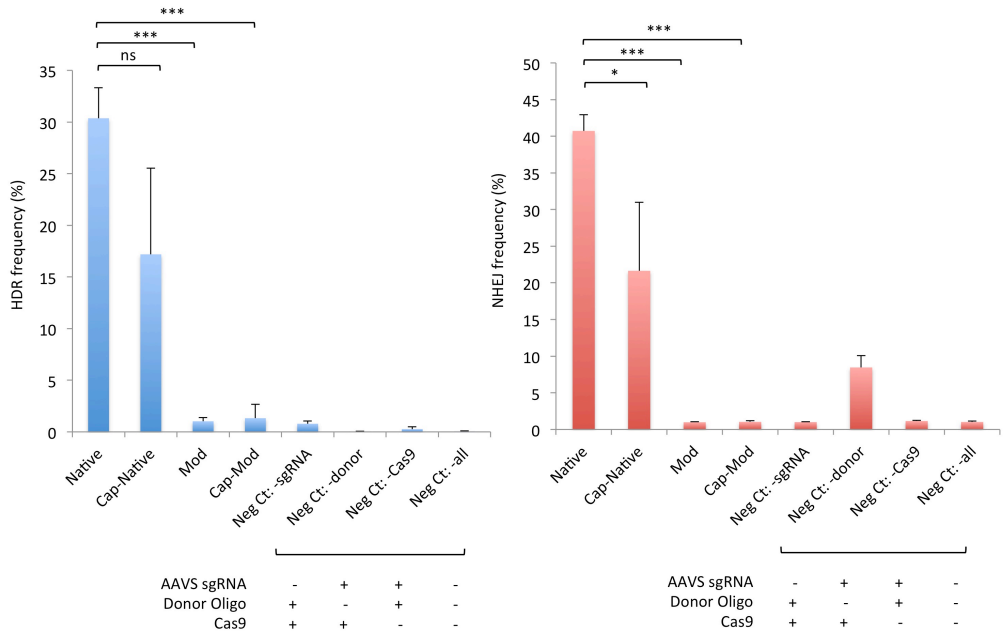

Figure 2-11: Impact of various sgRNA modifications on gene editing efficiency. Significance calculated by two-tailed *t*-test: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, ns, not significant.

Although the significant increase in gene editing efficiency by simply converting from modified to native sgRNA came much to our surprise, it was also theoretically reasonable. The modifications incorporated in the earlier study were used to synthesize mRNAs which function in the cytoplasm and inherently harbor the modifications *in vivo*, whereas we were attempting to generate sgRNAs for activity in the nucleus which functions in complex with Cas9 proteins[71]. The reduced efficiency of capped and modified sgRNA implied that the addition of a 5' guanine cap may preclude the sgRNA from entering the nucleus – its site of action, while the incorporation of modified ribonucleoside bases potentially disrupted the interaction between sgRNA and Cas9 protein, consequently contributing to the substantially reduced gene targeting efficiency. Having observed the dramatic impact of RNA modifications on genome editing rates, we proceeded with the native form of sgRNA to ensure efficient gene targeting in further experiments.

2.2.4 Characterization and optimization of Cas9-mediated genome editing in iPS-Cas9 system

Having established the PGP1 iPS-Cas9 cell line and an efficient RNA transfection methodology for genome editing, we next sought to evaluate the efficiency of our system in comparison to the established method of nucleofecting PGP1 iPS cells. We co-transfected AAVS1 sgRNA in the native RNA form along with the HDR donor oligo into dox-induced PGP1 iPS-Cas9 cells. Side by side, we co-nucleofected AAVS1 sgRNA expressed from the human U6 polymerase III promoter together with the HDR donor oligo and Cas9 plasmid into PGP1 iPS cells. The same components except for the Cas9 plasmid were also nucleofected into dox-induced PGP1 iPS-Cas9 cells. All conditions included B18R media supplementation. Three days post-transfection/nucleofection, we harvested the cells and analyzed the gene editing efficiency by next-generation sequencing. Deep sequencing results revealed RNA transfection of the iPS-Cas9 cell line as the superior gene editing method (HDR = 30.37%, NHEJ = 40.73%), with a marked 152-fold improvement in HDR rates and 10-fold increase in NHEJ rates compared to the nucleofection of iPS cells approach (HDR = 0.2%, NHEJ = 3.94%) (Figure 2-12).

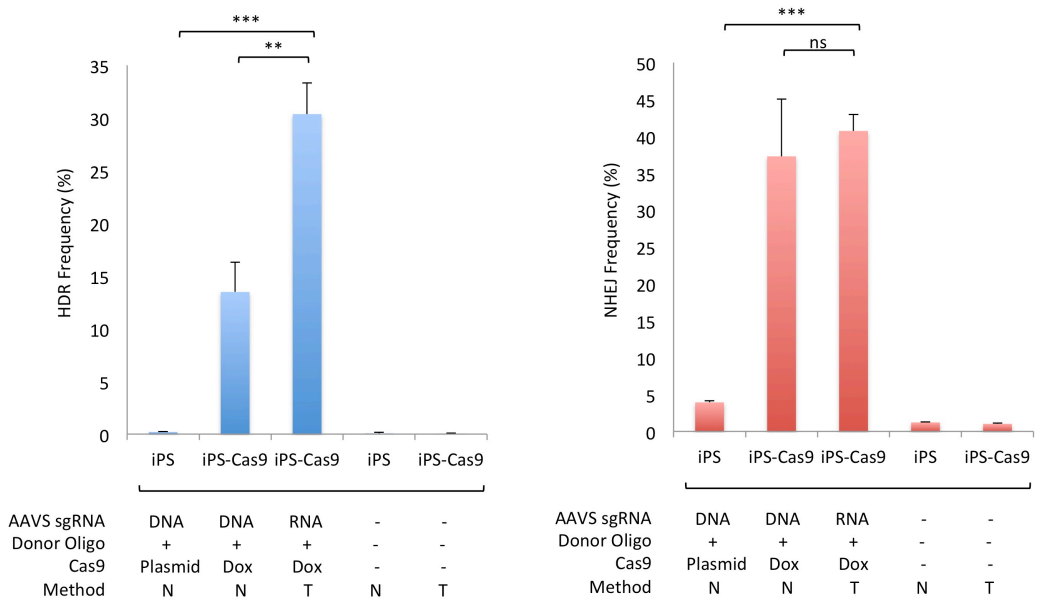

Figure 2-12: Comparison of gene editing efficiencies using various methods. AAVS sgRNA was either expressed from the human U6 polymerase III promoter (DNA) or delivered as *in vitro* transcribed RNA (RNA). Cas9 expression was induced either from a plasmid or with doxycycline (Dox). Under methods, N denotes nucleofection and T denotes lipid-based transfection. Significance calculated by two-tailed *t*-test: ** $P < 0.01$, *** $P < 0.001$, ns, not significant.

Moreover, within the iPS-Cas9 system, direct transfection of sgRNA in the native RNA form proved more efficient than nucleofection of sgRNA expressed from the human U6 polymerase III promoter (HDR = 13.49%, NHEJ = 37.29%), potentially owing to the observations that the RNA form of sgRNA was smaller and easier to deliver and that nucleofection appear to induce more cellular toxicity than lipid-based transfections (Figure 2-12).

To dive further into the differences observed between genome editing efficiencies of iPS and iPS-Cas9 cell lines, we examined gene targeting rates at the AAVS1 locus using the same RNA transfection method, with the only difference in Cas9 delivery. For the iPS cell line, Cas9 was co-transfected as mRNA along with the sgRNA and donor oligo, while Cas9

was induced by doxycycline in iPS-Cas9 cells. All the reagents were transfected using lipofectamine-based reagents. This procedure eliminated the difference in delivery method from the equation, and allowed us to probe the inherent differences in gene editing efficiency between the two systems. Having calibrated the two cell lines to differ only in the method of Cas9 delivery, we again observed enhanced gene editing efficiency in the iPS-Cas9 cell line (HDR = 18.89%, NHEJ = 20.84%) with 10-fold improvement in HDR rates and 8.5-fold improvement in NHEJ rates compared to iPS cells (HDR = 1.83%, NHEJ = 2.45%) (Figure 2-13).

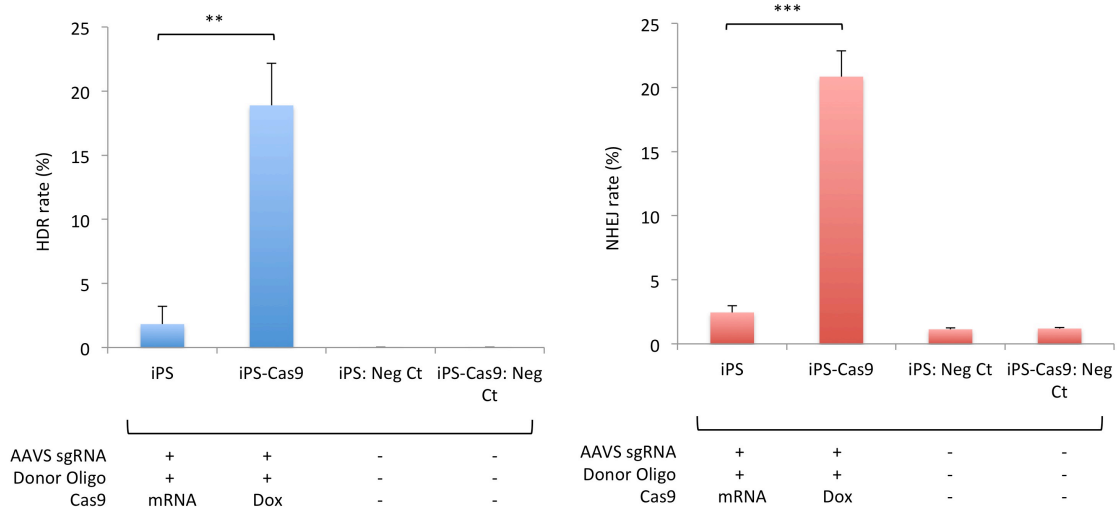

Figure 2-13: Comparison of gene editing efficiency in iPS and iPS-Cas9 cells via RNA transfection. Significance calculated by two-tailed *t*-test: ** $P < 0.01$, *** $P < 0.001$.

Taken together, these results highlighted the utility of PGP1 iPS-Cas9 cell line as an enhanced system for conducting genome editing experiments and the remarkable improvement in editing rates achievable when performed in combination with the direct native RNA transfection method. This provides a simple, efficient, and highly flexible platform for genome engineering applications in human induced pluripotent stem cells.

After demonstrating the functionality of the PGP1 iPS-Cas9 cell line, we next sought to optimize the RNA transfection parameters to maximize the gene editing efficiency. Using the RNA transfection method described earlier, we targeted the AAVS1 locus by co-transfecting sgRNA and HDR donor oligo into dox-induced iPS-Cas9 cells, with variations in donor oligo concentration, cell density, Cas9 induction level, donor length, and donor orientation. We harvested the cells three days post transfection and determined the gene editing efficiency by next-generation sequencing. From deep sequencing analysis, we identified the optimal gene editing efficiency was reached when using 3.8X molar ratio of donor oligo to sgRNA (Figure 2-14).

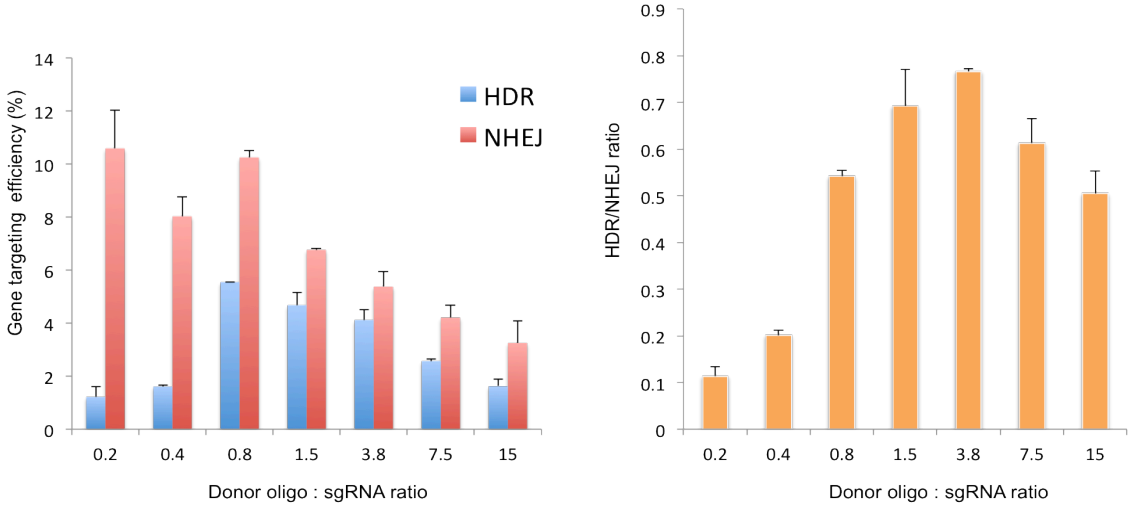

Figure 2-14: Optimization of HDR donor oligo to sgRNA ratio.

From the analysis of cell density and dox induction curve, both parameters appeared to play minor roles in gene editing rates, where lower seeding density showed slightly better HDR/NHEJ ratio and 0.5 ug/ml doxycycline exhibited the best ratio (Figure 2-15).

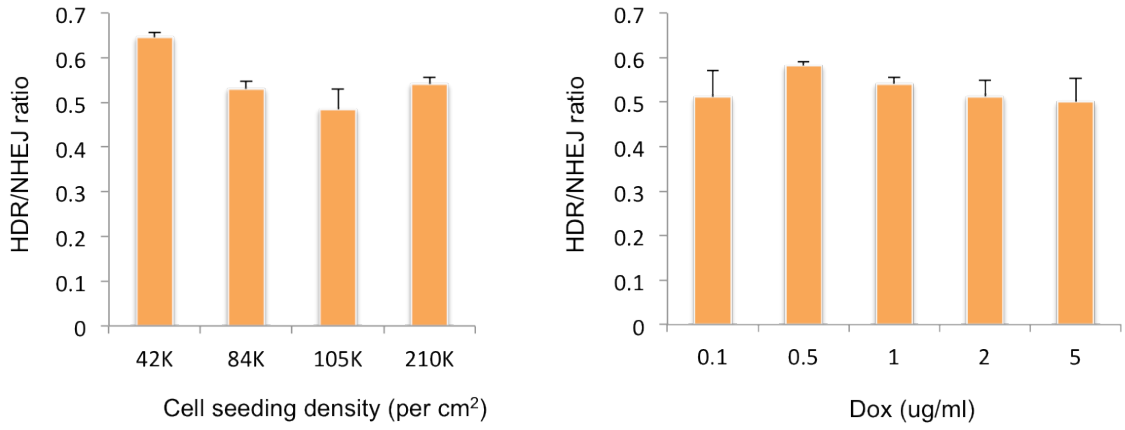

Figure 2-15: Optimization of cell density and dox induction.

Having optimized the fundamental parameters of RNA transfection method, we next performed a systematic study of ssODN designs to determine the guiding principles for designing high performing HDR donor oligo. As incorporation of phosphorothioate bonds at the terminal bases have been known to protect oligos from exonuclease degradation inside the cell, we tested whether the addition of terminal phosphorothioated bonds promotes gene editing efficiency by stabilizing the donor oligo. We designed a set of 70-mer oligos targeting the AAVS1 locus carrying a 2 baspair (bp) mismatch in the middle, with 0, 1(*), or 2(**) phosphorothioated bonds between bases at the 5', 3', or both 5' and 3' termini. From the deep sequencing result, we observed negligible differences in gene targeting rates amongst the various donor oligo modifications, suggesting the unmodified HDR donor oligo was not subject to extensive exonuclease degradation within the cell (Figure 2-16).

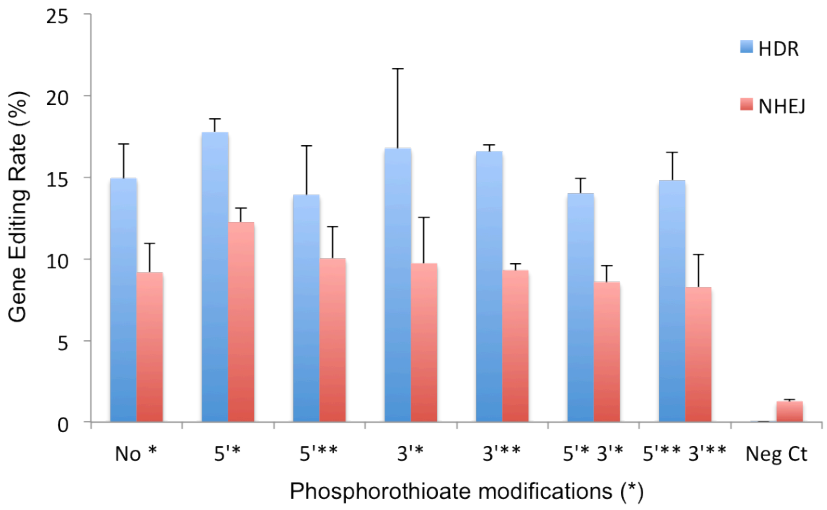

Figure 2-16: Impact of terminal phosphorothioate bonds on gene editing efficiency.

To evaluate the efficiency of introducing varying numbers of mismatches into the genome through ssODN donor oligos, we designed a set of 70-mer oligos targeting the AAVS1 locus, with 1, 2, 3, 5, 10, 20, 30 mismatches stemming from the middle of the oligo (referred to as mismatch number oligos). Deep sequencing analysis revealed the efficiency of generating mismatch modifications is dependent on the amount of homologous sequence between the donor oligo and its genomic target. As the number of mismatches increases to over 5 bases, the modification efficiency drops greatly, suggesting that ssODNs are most suitable for introducing small and precise changes to the genome (Figure 2-17a).

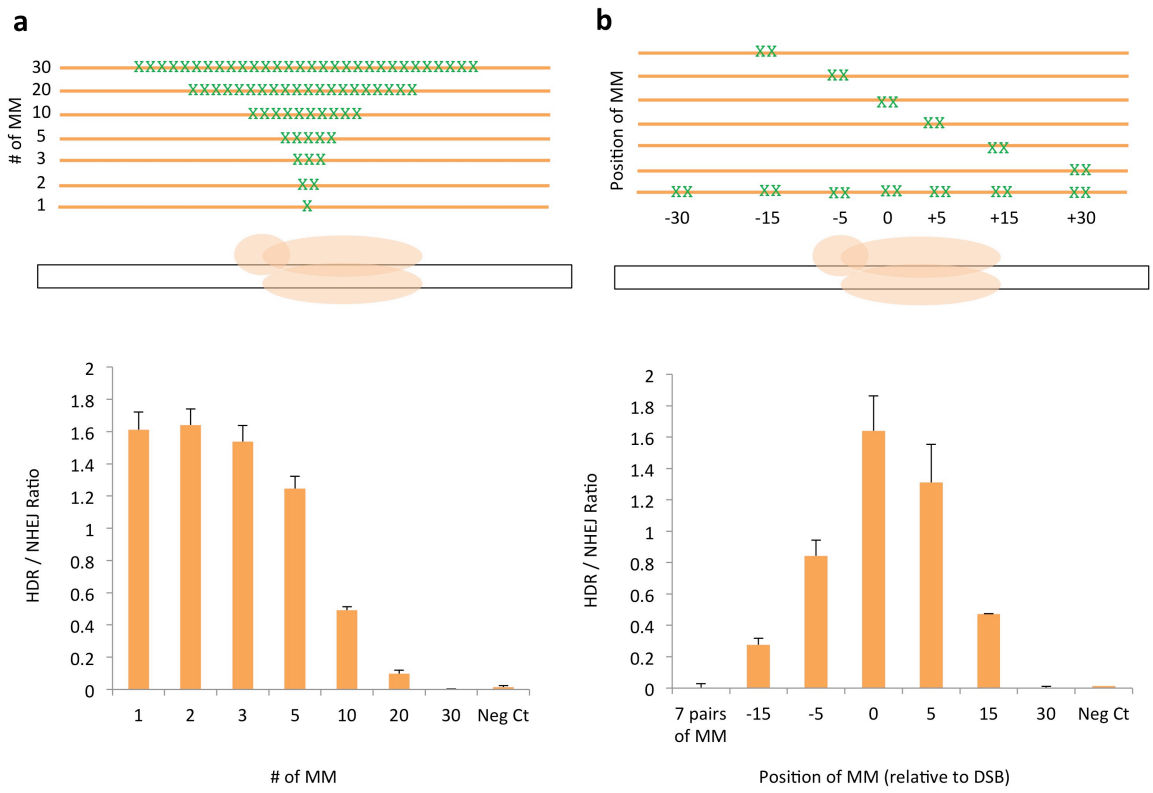

Figure 2-17: Characterization of gene editing efficiency as a function of (a) number of mismatches (MM) and (b) position of MM relative to the DSB.

Next, we examined the positional effect of mismatches on gene editing rates. To investigate how mismatch incorporation efficiency varies with the position of the mismatch relative to the double stranded break (DSB), we designed a set of seven 70-mer ssODN donor oligos targeting the AAVS1 locus, with a 2 bp mismatch placed at the -15, -5, 0, +5, +15, +30 positions relative to the DSB and one donor oligo with 7 pairs of 2 bp mismatches positioned at -30, -15, -5, 0, +5, +15, +30 positions relative to the DSB (referred to as mismatch position oligos). From the analysis of deep sequencing data, we discovered a strong positional effect where the highest gene editing efficiency occurred when the mismatch was placed right on the DSB and the efficiency tapered off as the distance between the mismatch and DSB increased to over 5 bp (Figure 2-17b). This result suggests

that there exists a small and precise window near the DSB where the intended modifications carried by the HDR donor oligo can be efficiently incorporated, likely dictated by the extent of DSB repair processing.

Furthermore, we performed a donor competition assay to assess the efficiency of various donor oligo designs at inducing modifications when competing in a pool of donors. To this end, we mixed at equal molar ratio the set of seven mismatch number oligos, the set of seven mismatch position oligos, and the entire set of 13 mismatch number and mismatch position oligos, and co-transfected each oligo mix with AAVS1 sgRNA into dox-induced iPS-Cas9 cells. In the mismatch number oligo pool, the mutation incorporation rate were comparable within 5 mismatches, but was greatly reduced with over 10 mismatches, similar to earlier observations. The donor competition of mismatch position oligos closely recapitulated prior results where mutation incorporation was most efficient at the DSB and dwindled as the distance between the DSB and mutation widened past 5 bp (Figure 2-18).

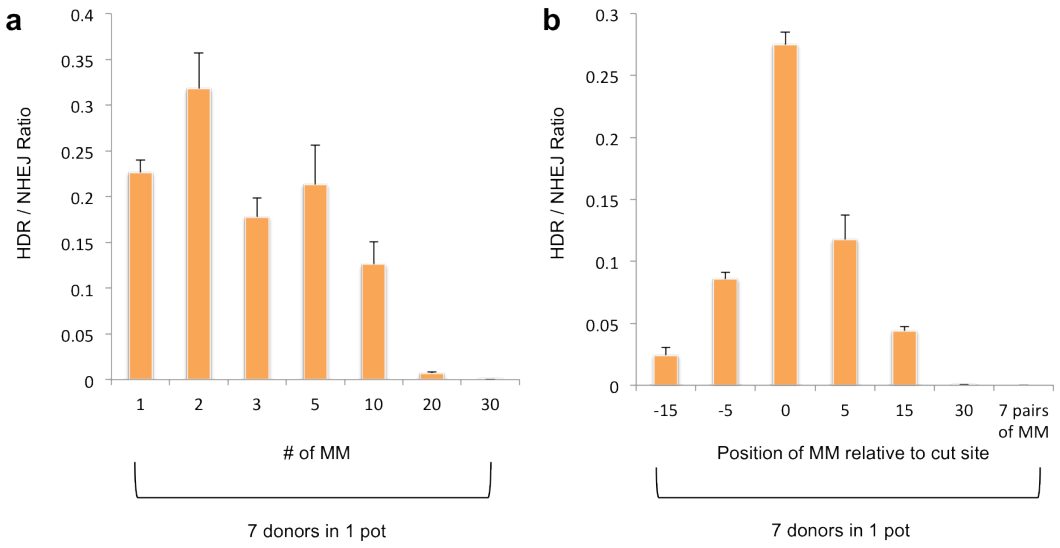

Figure 2-18: Donor competition assay with a) 7 mismatch number donor oligos or b) 7 mismatch position donor oligos in one pot.

Interestingly, the donor competition assay with the pool of 13 total mismatch number oligos and mismatch position oligos revealed a similar pattern of editing efficiency within each sub pool, whereas the mismatch number oligos with under 10 mismatches tend to have outcompeted the mismatch position oligos with mismatches placed greater than 5 bp away from the DSB (Figure 2-19). The donor competition assay results implied that the position of mutations relative to the DSB is a critical design parameter for efficient ssODN-mediated gene editing, whereas the number of mismatches introduced is better tolerated by the repair system. We also noted the total editing efficiency (HDR+NHEJ) for multiple donor assays were comparable to the total with single donors. This is expected as the total amount of sgRNA and donor oligos added were kept the same, the level of total editing should also be similar.

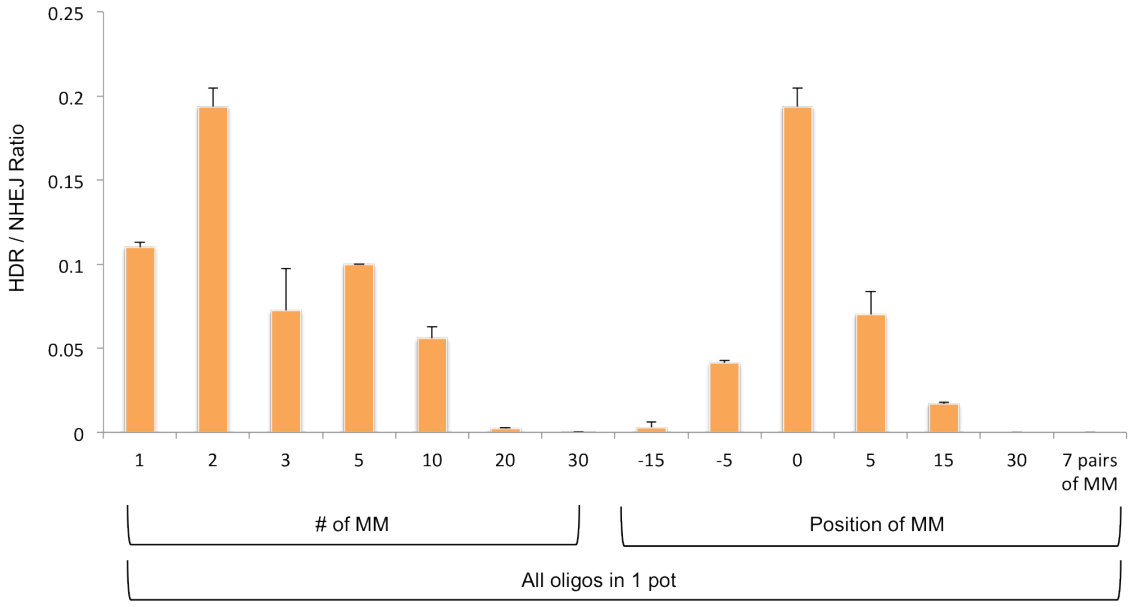

Figure 2-19: Donor competition assay of all 13 mismatch number and position donor oligos in one reaction.

In addition, analysis of ssODN donor length and orientation indicated that the 70-mer donor was more efficient than the 90-mer donor, and that donors in the complementary

orientation to sgRNA (D_c) achieved higher gene editing rates than donors in the non-complementary orientation (D_{nc}) (Figure 2-20), in agreement with earlier findings[82].

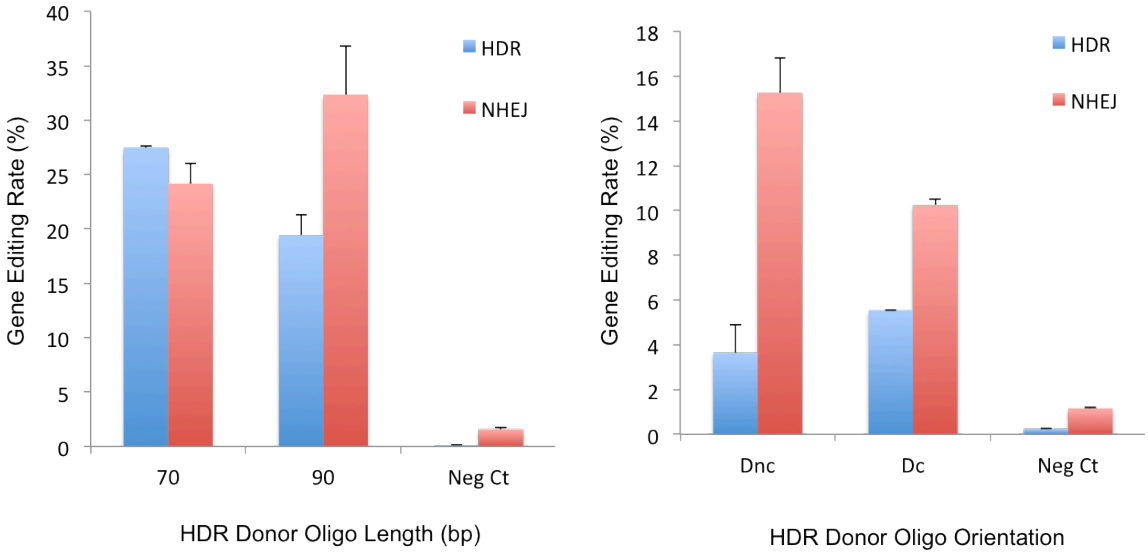

Figure 2-20: Effect of donor oligo length and orientation on gene editing efficiency. (D_c symbolizes donor is complementary to sgRNA. D_{nc} denotes donor is non-complementary to sgRNA.

We reasoned that extending the homology arm of the donor oligo beyond a certain point may cease to promote homologous recombination, due to the increased likelihood of secondary structure formation that potentially precludes bases on the oligo to hybridize to its chromosomal target. Moreover, while the kinetics and interactions between the Cas9 components and HDR repair machinery are not yet well-understood, we hypothesize that the D_c donor oligo may achieve better gene editing efficiency due to the fact that it can interact with the relatively free chromosomal target DNA strand whereas the sgRNA may still be bound to the other target strand, preventing the D_{nc} donor from accessing it after cleavage.

2.3 **Experimental Methods**

All oligonucleotide sequences used were synthesized by Integrated DNA Technologies (IDT).

2.3.1.1 Cell Line Construction

PGP1 human induced pluripotent stem (iPS) cell line was obtained from the Personal Genome Project (http://www.personalgenomes.org)[83]. The PGP1 iPS cell line with Cas9 integrated in the genome (iPS-Cas9) was constructed by delivering into PGP1 iPS cells a piggyBac transposon that expresses the reverse tetracycline-controlled transactivator (rtTA) and a human codon-optimized Cas9 under the control of a tet response element (TRE). Transfected cells were selected with puromycin at a concentration of 2 ug/ml starting 3 days post transfection. Selected cells were allowed to grow into a colony before verifying Cas9 genomic integration by qPCR and Sanger sequencing.

2.3.1.2 Cell Culture

PGP1 iPS cell line and PGP1 iPS-Cas9 cell line were maintained in feeder-free conditions on Matrigel (BD Biosciences)-coated plates in mTeSR-1 basal medium (Stemcell Technologies). Cultures were passaged every 5–7 days with TrypLE Express Enzyme (Life Technologies) and maintained in 10 uM Y-27632, a Rho-associated kinase (ROCK) inhibitor (Millipore), for one day after passaging to enhance cell survival[84]. All cells were maintained in a humidified incubator at 37°C with 5% CO2.

2.3.1.3 Design of sgRNA

sgRNA target sites with minimal potential off-targets were identified using sequence analysis tools such as ZiFit (http://zifit.partners.org/ZiFiT/)[85,86] and CRISPR Design Tool (http://crispr.mit.edu). After selecting specific sgRNA target sites, we designed sgRNA templates consisting of a U6 or T7 promoter, followed by the selected guide sequence (in the form of 5'-N19-NGG-3') and the 3' ss scaffold sequence (77 nucleotides in length). See Table 2-5 for a complete list of sgRNA sequences.

2.3.1.4 Design of HDR Donor Oligo

Single-stranded DNA oligonucleotides (ssODNs) have been used to introduce small modifications within a defined locus with reasonable efficiency. To achieve optimal HDR efficiencies, ssODNs have been designed with at least 34 bp flanking sequences on each arm that are homologous to the target region with the intended mutations positioned near the cutting site. The mutations were also designed to destroy the PAM and/or sgRNA targeting site, so that once the cells have been edited by HDR with the desired change, it would no longer be targeted by the CRISPR/Cas9 machinery. See Table 2-6 for a complete list of HDR donor oligo sequences.

2.3.1.5 Design of MiSeq Primers

Optimal primer sets that amplify 200 to 300 bp around the CRISPR targeting site were identified using primer design tools such as Primer3 (http://biotools.umassmed.edu/bioapps/primer3_www.cgi). Illumina forward and reverse adapter sequences were appended to the 5' end of the target site's forward primer and reverse primer sequences. See Table 2-7 for a complete list of MiSeq primer sequences.

2.3.1.6 sgRNA synthesis

sgRNA templates with U6 promoter were synthesized by IDT as gBlock gene fragments and PCR amplified for use. For transfection experiments directly introducing sgRNA in the RNA form, sgRNA templates were PCR amplified with T7 forward primer and sgRNA scaffold reverse primer to construct templates for in vitro transcription (IVT). PCR reactions were performed with HIFI Hotstart (KAPA Biosystems) and purified with QIAquick PCR Purification Kit (Qiagen) following the manufacturer's instructions.

sgRNAs were transcribed *in vitro* with T7 polymerase (MEGAscript T7 Kit, Ambion), using 1000 ng of purified PCR product as template for each 40 ul reaction. For the various modified versions of sgRNAs, the custom ribonucleoside mix is made according to Table 2-1 for native RNA,

Table 2-2 for capped native RNA, Table 2-3 for modified RNA, and Table 2-4 for capped modified RNA. GTP, ATP, CTP, and UTP were provided with MEGAscript T7 Kit, while 3'-O-Me-m7G Cap structure analog was purchased from NEB (New England Biolabs) and 5-me-CTP and pseudo-UTP from TriLink (TriLink Biotechnologies).

Table 2-1. Custom native RNA mixture for IVT.

#1 Native RNA Mix	**[Stock] (mM)**	**[Final] (mM)**	**Vol/reaction(ul)**
GTP	**75**	**7.5**	**4**
ATP	**75**	**7.5**	**4**
CTP	**75**	**7.5**	**4**
UTP	**75**	**7.5**	**4**

Table 2-2. Custom capped native RNA mixture for IVT.

#2 Capped Native RNA Mix	[Stock] (mM)	[Final] (mM)	Vol/reaction(ul)
3'-O-Me-m7G Cap structure analog	**60**	**6**	**4**
GTP	**75**	**1.5**	**0.8**
ATP	**75**	**7.5**	**4**
CTP	**75**	**7.5**	**4**
UTP	**75**	**7.5**	**4**

Table 2-3. Custom modified RNA mixture for IVT.

#3 Modified RNA Mix	[Stock] (mM)	[Final] (mM)	Vol/reaction(ul)
GTP	**75**	**7.5**	**4**
ATP	**75**	**7.5**	**4**
5-Me-CTP	**100**	**7.5**	**3**
Pseudo-UTP	**100**	**7.5**	**3**

Table 2-4. Custom capped modified RNA mixture for IVT.

#4 Capped/Modified RNA Mix	[Stock] (mM)	[Final] (mM)	Vol/reaction(ul)
3'-O-Me-m7G Cap structure analog	**60**	**6**	**4**
GTP	**75**	**1.5**	**0.8**
ATP	**75**	**7.5**	**4**
5-Me-CTP	**100**	**7.5**	**3**
Pseudo-UTP	**100**	**7.5**	**3**

Reactions were incubated overnight at 37°C to increase yield. The IVT reactions were then DNase treated and purified with MEGAclear (Ambion) following manufacturer's instructions. Next, sgRNA was treated with Antarctic Phosphatase (New England Biolabs)

for 30 - 60 min at 37°C to remove residual 5'-triphosphates and purified again with MEGAclear (Ambion). Finally, the concentrations of sgRNA products from IVT reactions were quantified by Nanodrop (Thermo Scientific) and adjusted to 100 ng/ul working concentration by adding elution buffer. To check the integrity of the *in vitro* transcribed sgRNA, small aliquots were run on a 10% denaturing PAGE gel (Life Technologies) with a low range ssRNA ladder (NEB). After verifying sgRNA size and quality, the sgRNAs was stored in -80°C until use.

2.3.1.7 Nucleofection Experiments

For nucleofection of PGP1 iPS or iPS-Cas9 cell line, 1–2 × 10^5 cells were seeded per 48-well plate one day before transfection. Nucleofections were performed with the P3 Primary Cell Nucleofector Kit (Lonza) following manufacturer's procedures. For each 48-well reaction, 1ul of 1 ug/ul sgRNA gene block expressed from the U6 promoter, 1 ug donor oligo and 1 ul of 1 ug/ul Cas9 plasmid were mixed with 16.4 ul P3 and 3.6 ul supplement reagent from the Nucleofector kit. Single cell suspensions were generated by dissociating with TrypLE Express (Invitrogen) and resuspending in mTeSR1 medium. Count cells and dilute accordingly to have at least 1 x 10^6 cells/nucleofection reaction to ensure survival post nucleofection. Cell pellets were collected by centrifuging at 200 x g for 5 minutes at room temperature and aspirating the supernatant. Then, cell pellets were resuspended in the P3 reagent mastermix and transferred to Nucleocuvette strip. Cells were nucleofected using the program CB150. Immediately after nucleofection, cells were recovered by adding pre-warmed mTeSR1 medium containing 2 ul/ml ROCK inhibitor and plated onto matrigel-coated plates to ensure maximal survival. Plates were centrifuged at 70 x g for 3 minutes at room temperature and placed into 37°C incubator. Twenty four hours

post nucleofection, ROCK inhibitor supplemented media was replaced with plain mTeSR1 medium. Three days post nucleofection, cells were harvested for MiSeq library preparation.

2.3.1.8 RNA Transfection

For RNA transfection of PGP1 iPS-Cas9 cell line, 1–2 × 10^5 cells were seeded per 48-well plate one day before transfection. Two hours prior to transfections, change to fresh media containing 0.5 – 1 ug/ml Doxycycline (Sigma-Aldrich) to induce Cas9 expression. RNA transfections were performed with RNAiMAX (Invitrogen) cationic lipid delivery vehicles following manufacturer's procedures, while specific parameters may be modified according to our optimization experiments. In general for each 48-well reaction, ~30 pmol sgRNA and ~80 pmol donor oligo were first diluted in 25ul Opti-MEM basal media (Invitrogen) while 6ul of RNAiMAX reagent was diluted in 25ul Opti-MEM basal media. Then, these components were mixed and incubated for 15 minutes at room temperature before adding to cells. Transfected cells were maintained in 37°C incubator and changed with mTeSR-1 Basal Medium (STEMCELL technologies) daily until harvest 3 days post-transfection. In certain experiments, the interferon inhibitor B18R (eBioscience) was added at 200 ng/ml as a media supplement.

For RNA transfection of PGP1 iPS cell line, the basic procedure outlined above was employed with some key differences. Instead of adding Dox to the media prior to transfection, iPS cells were maintained in plain mTeSR-1 Basal Medium. To introduce Cas9 into iPS cells, 0.5 – 1 ug Cas9 mRNA or plasmid was diluted along with sgRNA and donor oligo in 25ul Opti-MEM basal media. Then, the components were pooled with RNAiMAX reagent, incubated for 15 minutes at room temperature before dispensing onto cells. iPS cells were maintained at 37°C and changed with mTeSR-1 Basal Medium daily until harvest 3

days post-transfection. In certain experiments, the interferon inhibitor B18R (eBioscience) was added at 200 ng/ml as a media supplement.

2.3.1.9 RNA Transfection with Donor Coupling

General RNA transfection procedures as outlined above were performed, with a slight modification to pre-anneal sgRNA-docks with HDR donor oligos before mixing with the transfection reagents. In brief, for each 48-well reaction, ~30 pmol sgRNA and ~80 pmol HDR donor oligo were first diluted in 25ul Opti-MEM basal media. The sgRNA and donor oligo mixture was then heated to 70°C for 30 sec and cooled gradually at -0.1°C/sec to 25°C to allow hybridization. The pre-annealed sgRNA and HDR donor oligo mixture was then mixed with the RNAiMAX reagent and incubated for 15 minutes at room temperature before dispending onto cells. Transfected cells were maintained in 37°C incubator and changed with mTeSR-1 Basal Medium (STEMCELL technologies) daily until harvest 3 days post-transfection.

2.3.1.10 MiSeq Library Preparation

Cells were harvested 3 days post-transfection with TrypLE express after a quick wash with PBS (Life Technologies). Cell pellets were collected after spinning down at top speed in a tabletop centrifuge. Genomic DNA was extracted using prepGEM tissue kit (ZyGEM). For each sample, 0.1 ul of prepGEM tissue protease enzyme and 1 ul of prepGEM gold buffer were mixed with 8.9 ul of 2–5 × 10^5 cells and incubated at 75°C for 15 minutes followed by 95°C for 5 minutes. To amplify the genomic region flanking the CRISPR target site and attach Illumina sequence adapters and unique barcodes for each sample, two rounds of PCR was performed.

In the first-round PCR amplifying the genomic region of interest, 2.5 ul of the cell lysis reaction was added to 22.5 ul of PCR mastermix containing 12.5 ul of 2X KAPA Hifi Hotstart Readymix and 100 uM of corresponding PCR amplification primer pairs. Reactions were incubated at 95°C for 5 min followed by 13-15 cycles of 98°C for 20 sec, 60°C for 20 sec and 72°C for 20 sec, and a final extension step of 72°C for 4 min.

A second-round PCR was performed to add Illumina sequence adaptors and unique index for sample barcoding. For this, 5 ul of PCR round-1 products were added to 20 ul of PCR mastermix containing 12.5 ul of 2X KAPA HIFI Hotstart Readymix and 100 uM ScriptSeq Index PCR primers pairs (Illumina). Reactions were incubated at 95°C for 5 min followed by 25 cycles of 98°C for 20 sec, 60°C for 20 sec and 72°C for 20 sec, and a final extension step of 72°C for 4 min. PCR products were run on a 2% agarose gel to verify the correct amplicon length and purified with QIAquick PCR purification kit (QIAGEN). Next, the concentrations of PCR products were quantified using Nanodrop (Thermo Scientific) or Qubit (Life Technologies), and pooled in equimolar ratio to ensure equal sequencing coverage. The mixed barcoded library was then sequenced with an Illumina MiSeq Personal Sequencer (Life Technologies).

2.3.1.11 Sequencing Data Analysis

MiSeq reads were analyzed with custom scripts from the genome editing assessment system as previously described[87]. The analysis method has also been established as a web platform for easy access (http://crispr-ga.net)[88].

The method presented in the "Sequencing Data Analysis" subsection has been published and adapted from the following paper to fit the format of this dissertation:

- Yang L, Guell M, Byrne S, Yang, JL, De Los Angelos, A, Mali, P, Aach, J, Kim-Kiselak, C, Briggs, AW, Rios, X. Huang, P, Daley, G, and Church G. Optimization of scarless human stem cell genome editing. *Nucleic Acids Res*. 2013;41:9049-9061.

We wrote a pipeline to analyze the genome engineering data. This pipeline is integrated in one single Unix module, which uses different tools such as R, BLAT and FASTX Toolkit.

Barcode splitting: Groups of samples were pooled together and sequenced using MiSeq 150 bp paired end (PE150) (Illumina Next Gen Sequencing) and later separated based on DNA barcodes using FASTX Toolkit.

Quality filtering: We trimmed nucleotides with lower sequence quality (pDed score <20). After trimming, reads shorter than 80 nt were discarded.

Mapping: We used BLAT to map the paired reads independently to the reference genome and we generated .psl files as output.

Indel calling: We defined indels as the full-length reads containing two blocks of matches in the alignment. Only reads following this pattern in both paired end reads were considered. As a quality control, we required the indel reads to possess minimal 70 nt matching with the reference genome and both blocks to be at least 20 nt long. Size and position of indels were calculated by the positions of each block to the reference genome. Non-homologous end joining (NHEJ) has been estimated as the percentage of reads containing indels (see Equation 2.1). The majority of NHEJ event have been detected at the targeting site vicinity.

Homology-directed recombination (HDR) efficiency: Pattern matching (grep) within a 12 bp window centering over DSB was used to count specific signatures corresponding to reads containing the reference sequence, modifications of the reference sequence (2 bp intended mismatches) and reads containing only 1 bp mutation within the 2 bp intended mismatches (see Equation 2.1).

$$\text{HDR efficiency} = \left(100\ \text{X}\ \frac{\text{B}}{\text{A} + \text{B} + \text{C} + \text{D}} \right)\%$$

$$\text{NHEJ efficiency} = \left(100\ \text{X}\ \frac{\text{D}}{\text{A} + \text{B} + \text{C} + \text{D}} \right)\%$$

A=reads identical to the reference: XXXXXABXXXXX

B =reads containing 2 bp mismatch programed by ssODN: XXXXXabXXXXX

C = reads containing only 1 bp mutation in the target site: such as XXXXXaBXXXXX or XXXXXAbXXXXX

D = reads containing indels as described above

Equation 2.1. Estimation of HDR and NHEJ efficiency.

Table 2-5. sgRNA design sequences.

ADA_ex7_U6	TGTACAAAAAAGCAGGCTTTAAAGGAACCAATTCAGTCGACTGGATCCGGTACCAA GGTCGGGCAGGAAGAGGGCCTATTTCCCATGATTCCTTCATATTTGCATATACGAT ACAAGGCTGTTAGAGAGATAATTAGAATTAATTTGACTGTAAACACAAAGATATTA GTACAAAATACGTGACGTAGAAAGTAATAATTTCTTGGGTAGTTTGCAGTTTTAAA ATTATGTTTTAAAATGGACTATCATATGCTTACCGTAACTTGAAAGTATTTCGATT TCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGcgtactgtccacgccgg ggGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGA AAAAGTGGCACCGAGTCGGTGCTTTTTTTCTAGACCCAGCTTTCTTGTACAAAGTT GGCATTA
ADA_ex10_U6	TGTACAAAAAAGCAGGCTTTAAAGGAACCAATTCAGTCGACTGGATCCGGTACCAA GGTCGGGCAGGAAGAGGGCCTATTTCCCATGATTCCTTCATATTTGCATATACGAT ACAAGGCTGTTAGAGAGATAATTAGAATTAATTTGACTGTAAACACAAAGATATTA GTACAAAATACGTGACGTAGAAAGTAATAATTTCTTGGGTAGTTTGCAGTTTTAAA ATTATGTTTTAAAATGGACTATCATATGCTTACCGTAACTTGAAAGTATTTCGATT TCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGgacatgggctttactga agGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGA AAAAGTGGCACCGAGTCGGTGCTTTTTTTCTAGACCCAGCTTTCTTGTACAAAGTT GGCATTA
SBDS_IV2_U6	TGTACAAAAAAGCAGGCTTTAAAGGAACCAATTCAGTCGACTGGATCCGGTACCAA GGTCGGGCAGGAAGAGGGCCTATTTCCCATGATTCCTTCATATTTGCATATACGAT ACAAGGCTGTTAGAGAGATAATTAGAATTAATTTGACTGTAAACACAAAGATATTA GTACAAAATACGTGACGTAGAAAGTAATAATTTCTTGGGTAGTTTGCAGTTTTAAA ATTATGTTTTAAAATGGACTATCATATGCTTACCGTAACTTGAAAGTATTTCGATT TCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGctgaaatctgtaagcag gtGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGA AAAAGTGGCACCGAGTCGGTGCTTTTTTTCTAGACCCAGCTTTCTTGTACAAAGTT GGCATTA

Table 2-5 (Continued).

SBDS_IV 3-1_U6	TGTACAAAAAAGCAGGCTTTAAAGGAACCAATTCAGTCGACTGGATCCGGTACCAA GGTCGGGCAGGAAGAGGGCCTATTTCCCATGATTCCTTCATATTTGCATATACGAT ACAAGGCTGTTAGAGAGATAATTAGAATTAATTTGACTGTAAACACAAAGATATTA GTACAAAATACGTGACGTAGAAAGTAATAATTTCTTGGGTAGTTTGCAGTTTTAAA ATTATGTTTTAAAATGGACTATCATATGCTTACCGTAACTTGAAAGTATTTCGATT TCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGttttttagattttgact aaGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGA AAAAGTGGCACCGAGTCGGTGCTTTTTTTCTAGACCCAGCTTTCTTGTACAAAGTT GGCATTA
SBDS_IV 3-2_U6	TGTACAAAAAAGCAGGCTTTAAAGGAACCAATTCAGTCGACTGGATCCGGTACCAA GGTCGGGCAGGAAGAGGGCCTATTTCCCATGATTCCTTCATATTTGCATATACGAT ACAAGGCTGTTAGAGAGATAATTAGAATTAATTTGACTGTAAACACAAAGATATTA GTACAAAATACGTGACGTAGAAAGTAATAATTTCTTGGGTAGTTTGCAGTTTTAAA ATTATGTTTTAAAATGGACTATCATATGCTTACCGTAACTTGAAAGTATTTCGATT TCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGagtgatttcttaaatgt gtGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGA AAAAGTGGCACCGAGTCGGTGCTTTTTTTCTAGACCCAGCTTTCTTGTACAAAGTT GGCATTA
AAVS_U6	TGTACAAAAAAGCAGGCTTTAAAGGAACCAATTCAGTCGACTGGATCCGGTACCAA GGTCGGGCAGGAAGAGGGCCTATTTCCCATGATTCCTTCATATTTGCATATACGAT ACAAGGCTGTTAGAGAGATAATTAGAATTAATTTGACTGTAAACACAAAGATATTA GTACAAAATACGTGACGTAGAAAGTAATAATTTCTTGGGTAGTTTGCAGTTTTAAA ATTATGTTTTAAAATGGACTATCATATGCTTACCGTAACTTGAAAGTATTTCGATT TCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGatcctgtccctagtggc cccGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTG AAAAAGTGGCACCGAGTCGGTGCTTTTTTTCTAGACCCAGCTTTCTTGTACAAAGT TGGCATTA
ADA_ex7 _T7	TAATACGACTCACTATAGGcgtactgtccacgccggggGTTTTAGAGCTAGAAATA GCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGT GCT
ADA_ex1 0_T7	TAATACGACTCACTATAGGgacatgggctttactgaagGTTTTAGAGCTAGAAATA GCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGT GCT
AAVS_T7	TAATACGACTCACTATAGGGGGGCCACTAGGGACAGGATGTTTTAGAGCTAGAAAT AGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGG TGCT

Table 2-6. ssODN HDR donor template sequences.

ADA-ex7	CCAGGAGGCTGTGAAGAGCGGCATTCACCGTACTGTCCACGCCAGGGAGGTGG GCTCGGCCGAAGTAGTAAAAGAGGTGAGGGCCTGGGC
ADA-ex10	CTGGACACTGATTACCAGATGACCAAACGGGACATGGGCTTTACT<DEL(GAA GA)>GGAGTTTAAAAGGCTGGTGAGTGGGTGTGAGCCATACTGGCCTTG
SBDS-IV2	CAGTGCGTTTGGAACAGATGACCAAACTGAAATCTGTAAGCAGGCGGGTAACA GCTGCAGCATAGCTAACCCTAATAACCATTTATAACG
SBDS-IV3	GATAGAGAAAGATAGTGATTTCTTAAATGTGTTGGCATTTTTTTAAATTTTGA CTAAAGGAGAAGTTCAAGTATCAGATAAAGAAAGACA
AAVS_70	ggaggcctaaggatggggcttttctgtcaccaatGGtgtccctagtggcccca ctgtggggtggagggga
AAVS_90	ctaggaaggaggaggcctaaggatggggcttttctgtcaccaatGGtgtccct agtggccccactgtggggtggaggggacagataaaag
AAVS mut oligo 70_1mm	ggaggcctaaggatggggcttttctgtcaccaatcGtgtccctagtggcccca ctgtggggtggagggga
AAVS mut oligo 70_3mm	ggaggcctaaggatggggcttttctgtcaccaaCGGtgtccctagtggcccca ctgtggggtggagggga
AAVS mut oligo 70_5mm	ggaggcctaaggatggggcttttctgtcaccaTCGGCgtccctagtggcccca ctgtggggtggagggga
AAVS mut oligo 70_10mm	ggaggcctaaggatggggcttttctgtcacGTTCGGCACGcctagtggcccca ctgtggggtggagggga
AAVS mut oligo 70_20mm	ggaggcctaaggatggggcttttctACGTGGTTCGGCACGGGCTAtggcccca ctgtggggtggagggga
AAVS mut oligo 70_30mm	ggaggcctaaggatggggctCCCGCACGTGGTTCGGCACGGGCTACAAGGcca ctgtggggtggagggga
AAVS mut oligo 70_-15	ggaggcctaaggatggggcCCttctgtcaccaatcctgtccctagtggcccca ctgtggggtggagggga
AAVS mut oligo 70_-5	ggaggcctaaggatggggcttttctgtcaGGaatcctgtccctagtggcccca ctgtggggtggagggga

Table 2-6 (Continued).

AAVS mut oligo 70_+5	ggaggcctaaggatggggcttttctgtcaccaatcctgtGGctagtggcccca ctgtggggtggagggga
AAVS mut oligo 70_+15	ggaggcctaaggatggggcttttctgtcaccaatcctgtccctagtggcGGca ctgtggggtggagggga
AAVS mut oligo 70_+30	ggaggcctaaggatggggcttttctgtcaccaatcctgtccctagtggcccca ctgtggggtggTAggga
AAVS mut oligo 70_7pair MM	ggagAGctaaggatggggcCCttctgtcaGGaatGGtgtGGctagtggcGGca ctgtggggtggTAggga
AAVS mut oligo 70_5*	g*gaggcctaaggatggggcttttctgtcaccaatGGtgtccctagtggcccc actgtggggtggagggga
AAVS mut oligo 70_5**	g*g*aggcctaaggatggggcttttctgtcaccaatGGtgtccctagtggccc cactgtggggtggagggga
AAVS mut oligo 70_3*	ggaggcctaaggatggggcttttctgtcaccaatGGtgtccctagtggcccca ctgtggggtggagggg*a
AAVS mut oligo 70_3**	ggaggcctaaggatggggcttttctgtcaccaatGGtgtccctagtggcccca ctgtggggtggaggg*g*a
AAVS mut oligo 70_5*3*	g*gaggcctaaggatggggcttttctgtcaccaatGGtgtccctagtggcccc actgtggggtggagggg*a
AAVS mut oligo 70_5**3**	g*g*aggcctaaggatggggcttttctgtcaccaatGGtgtccctagtggccc cactgtggggtggaggg*g*a

Table 2-7. MiSeq PCR Primers.

ADA_ex7_F-primer	ACACTCTTTCCCTACACGACGCTCTTCCGATCTgtatgggaggaggcagtgag
ADA_ex7_R-primer	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTaggcagcatgactaggatgg
ADA_ex10_F-primer	ACACTCTTTCCCTACACGACGCTCTTCCGATCTtgacccgctcatcttcaagt
ADA_ex10_R-primer	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTgcagactcactccctctctc
SBDS_IV2_F-primer	ACACTCTTTCCCTACACGACGCTCTTCCGATCTaggaagatctcatcagtgcgt
SBDS_IV2_R-primer	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTtgatttcaggaggttttggca
SBDS_IV3_F-primer	ACACTCTTTCCCTACACGACGCTCTTCCGATCTctgctccagttgtgtgtgtc
SBDS_IV3_R-primer	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTgggcaaagctcaaaccattac
AAVS-F	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGGTTAATGTGGCTCTGGTT
AAVS-R	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTACAGGAGGTGGGGGTTAGAC
Universal-PCR	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
ScripSeq Index Primers	CAAGCAGAAGACGGCATACGAGATN1N2N3N4N5N6GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

2.4 **Acknowledgments**

I thank Luhan Yang for sharing expertise to help with initial setup of this project. I thank Marc Guell for assistance with the MiSeq data analysis. I thank Xavier Rios for sharing the piggyBac plasmid. I thank Alejandro de Los Angeles for discussion regarding the disease targets. I thank Sam Chunte Peng for assistance with schematic illustration. I thank Prashant Mali and Susan Byrne for helpful discussions. I thank John Aach for reading and providing insightful comments on this chapter. I thank George M. Church for advice and support throughout the project.

CHAPTER 3

Applications of Facilitated Human Stem Cell Genome Editing and Strategies for Enhancement of HDR Efficiency

3.1 Introduction

The two major DSB DNA repair pathways – HDR and NHEJ, compete in the repair of DSBs generated by the Cas9 system. While NHEJ occurs throughout the cell cycle[89], HDR only takes place during S and G2 phase[17,55,90]. Thereby, Cas9-mediated gene knockout through NHEJ has been reported to work efficiently, reaching 20-60% in mouse embryonic stem cells and zygotes[55,90]. Since NHEJ is error-prone and introduces unpredictable indels, it is mainly used for inducing small mutations for gene disruption. To achieve precise genome editing, the DSB must be repaired efficiently through the HDR pathway with the incorporation of the desired change specified in the exogenous donor template. However,

the efficiency of HDR for precise genome editing remains low[17,55,90]. In order to isolate correctly edited cells, a long and labor-intensive screening process is often required. The low HDR efficiency also poses a significant challenge for generating sufficient numbers of genome edited animals, and ultimately using the technology in clinical applications.

To demonstrate the utility of the PGP1 iPS-Cas9 system, we first applied the method to disease gene targeting and showed enhanced efficiency compared to earlier methods. We then proceeded to investigate the prospect of multiplexed and continuous genome editing using the RNA transfection system. Finally, we explored various strategies based on the concept of coupling sgRNA with donor template to enhance HDR over NHEJ ratio, and examined the site to site variability in gene editing efficiency. Overall, we demonstrate the versatility of the iPS-Cas9 system that can be easily adapted for disease gene targeting, as well as continuous and multiplexed editing, and provide insights to guide future methods for further promoting HDR efficiency.

3.2 **Results and Discussion**

3.2.1 Application of method to disease gene targeting

Recall that we initiated this work with the original goal of achieving efficient genome editing in human iPS cells without the need for drug selection, targeting genes with potential applications in disease modeling and gene therapy. Having developed the PGP1 iPS-Cas9 cell line with improved gene editing efficiency as well as the optimized RNA transfection method, we went on to demonstrate efficient gene editing of disease genes.

First, we revisited the two loci – ADA exon 7 (ADA7) and ADA exon 10 (ADA10) – implicated in adenosine deaminase deficiency, which have caused severe combined immunodeficiency (ADA-SCID) in a patient iPS cell line[65]. In an effort to induce the disease genotype in PGP1 iPS-Cas9 cell line, we designed 90-nt HDR donor oligos for each site, introducing a GGG to AGG transition mutation in ADA7 and a frameshift mutation (Del(GAAGA)) in ADA10, as described earlier (Figure 2-1). To assess gene editing efficiencies of endogenous disease gene targets, we co-trasnfected the ADA7 or ADA10 sgRNAs along with its corresponding HDR donor oligo into dox-induced PGP1 iPS-Cas9 cells with B18R media supplementation. Three days post-transfection, we harvested the cells and analyzed the gene targeting efficiency with MiSeq Personal Sequencer. Deep sequencing of the ADA7 and ADA10 loci displayed indel formation near the predicted cleavage site with 6.91% NHEJ at the ADA7 locus and 19.56% NHEJ at the ADA 10 locus (Figure 3-1). The introduction of disease mutations was also detectable with 4.07% HDR at the ADA7 locus and 5.76% HDR at the ADA10 locus, a 6-fold and 15-fold improvement, respectively, compared to nucleofection of PGP1 iPS cells.

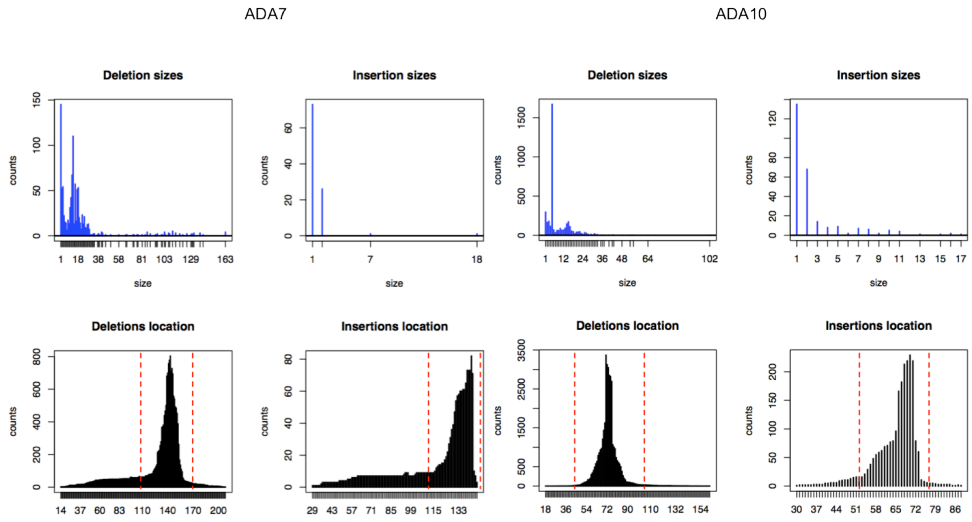

Figure 3-1: NHEJ profile of ADA7 and ADA10 gene editing via RNA transfection.

Concurrently, we re-examined the impact of RNA modifications at the two target sites, and observed the same pattern of decreased efficiency as sgRNA was capped and modified (Figure 3-2), confirming our previous finding that sgRNA in the native RNA form maximized genome editing efficiency.

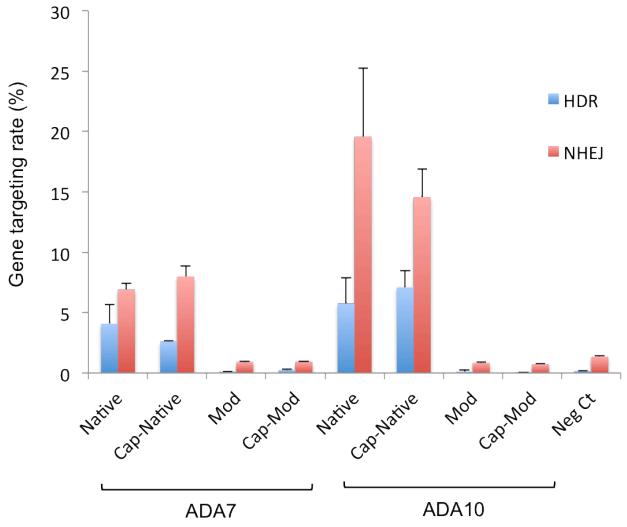

Figure 3-2: Gene editing efficiency at ADA7 and ADA10 target sites with various sgRNA modifications.

In addition, we targeted the *TAZ* gene encoding the protein tafazzin, which is mutated in patients with Barth syndrome. Barth syndrome is an X-linked condition characterized by dilated cardiomyopathy, skeletal myopathy, neutropenia, and short stature. Mutations in the *TAZ* gene prevent tafazzin from carrying out its normal function – acylation of cardiolipin, the main phospholipid of the mitochondrial inner membrane responsible for maintaining mitochondrial shape and energy production, hence leading to cardiac and skeletal mitochondrial myopathy[91,92]. As the mechanism of disease phenotype onset was not well understood, it would be useful to model the disease in a human iPS cell line with the specified *TAZ* mutation to establish a causative role of the mutation in an otherwise isogenic background. Toward this end, we tested the utility of the PGP1 iPS-Cas9

transfection system for facilitating gene editing at the *TAZ* locus. We designed 3 HDR donor oligos targeting the *TAZ* gene, one introducing a frameshift mutation (517del(G)) known to cause Barth syndrome in a patient cell line[93], and two donor oligos each with 1 or 2 mismatch mutations (G to C) as a reference (Figure 3-3).

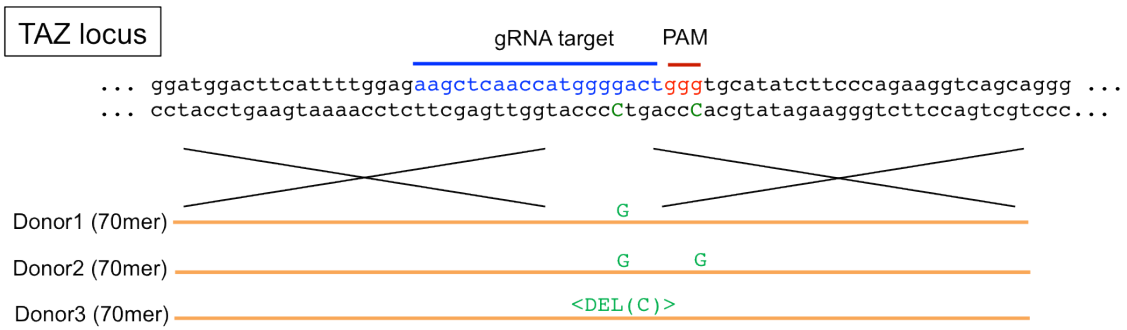

Figure 3-3: Schematic of HDR donor designs targeting the *TAZ* locus. Donors depicted in the bottom strand orientation.

To evaluate the efficiency of introducing disease mutations into PGP1 iPS-Cas9 cells, we co-transfected the sgRNA targeting *TAZ* gene, along with HDR donor oligos into dox-induced cells. Three days post-transfection, we harvested the cells and determined the gene editing rates by next-generation sequencing. Sequencing results showed comparable gene editing efficiencies across the three donor oligo designs, with NHEJ ranging from 10.88% to 13.35%, and HDR ranging from 1.31% to 2.76% (Figure 3-4). Together, these results demonstrate the efficiency and applicability of the streamlined iPS-Cas9 transfection platform, where starting from target design, one can expect to obtain edited cells in less than two weeks, greatly facilitating genome editing studies in the traditionally challenging human iPS cell line.

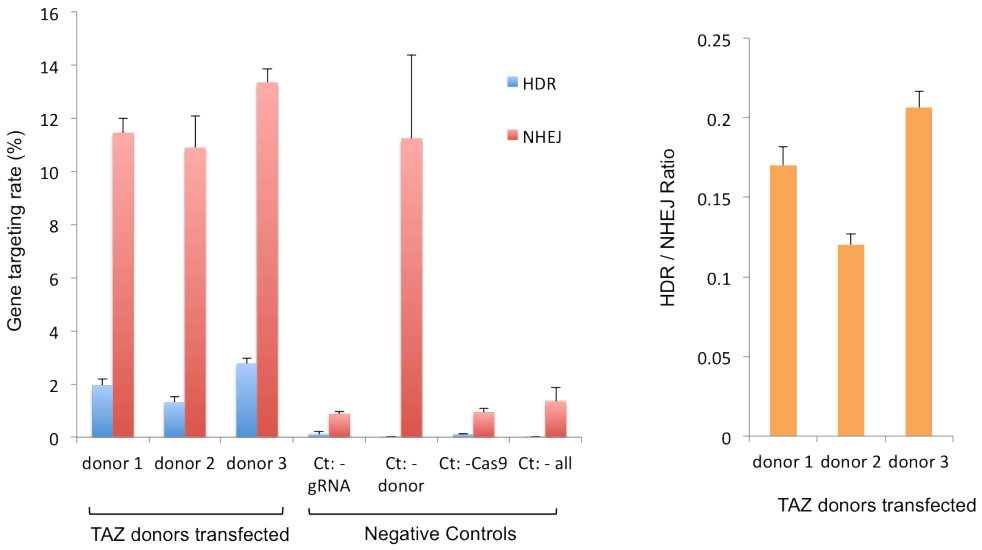

Figure 3-4: Efficiency of introducing disease genotype or other point mutations into *TAZ* locus.

3.2.2 Potential of multiplexed and continuous gene targeting

A key advantage of the CRISPR-Cas system over zinc finger nucleases (ZFNs) or Transcription activator-like effector nucleases (TALENs) lies in its potential for multiplexed genome editing since target recognition is determined by an easily programmable sgRNA. The ability to perform multiplexed genome targeting will enable efficient generation of cell lines or model organisms carrying multiple mutations, eliminating the laborious process of sequential targeting and selection, thus greatly facilitating the genetic dissection of development, multigenic, or complex diseases.

To evaluate the iPS-Cas9 system for potential of multiplexed gene editing, we first targeted the ADA7 and ADA10 loci simultaneously as an initial demonstration of efficiency. We co-trasnfected sgRNAs and HDR donor oligos targeting the ADA7 and ADA10 loci as described in Chapter 2.2.1, either individually or together, into dox-induced PGP1 iPS-Cas9 cells. Three days after transfection, we harvested the cells and analyzed the gene editing

efficiency by next-generation sequencing. As with single targeting, simultaneous targeting of ADA7 and ADA10 loci produced indels around each predicted cutting site with 4.91% NHEJ observed at ADA7 and 15.27% at ADA10. Likewise, HDR rate was detected at 1.22% for ADA7 and 3.47% for ADA10 (Figure 3-5).

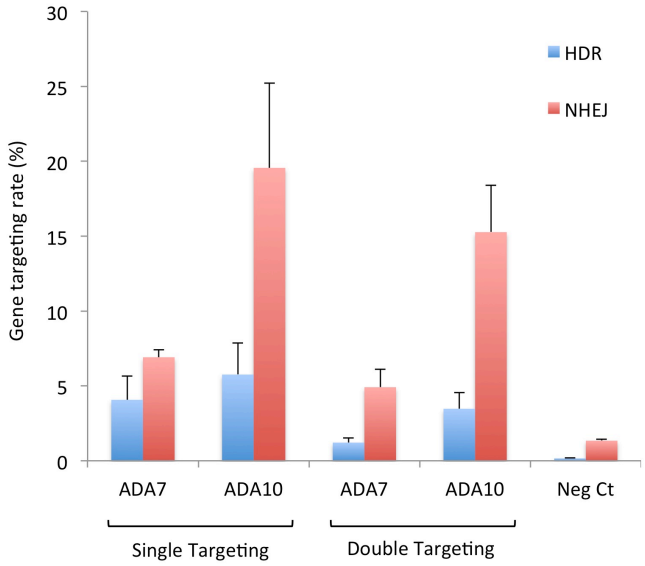

Figure 3-5: Multiplex capacity of Cas9-mediated gene targeting.

We showed that multiplex gene editing was feasible with our system, although with slightly reduced efficiencies compared to single targeting events. Since the maximal total editing efficiency (Total Edits = HDR + NHEJ) remained similar, we reasoned the decreased targeting rate observed in each double targeting sites could be attributed to potentially less efficient delivery of all the necessary components into the cells and saturation of cutting or repair enzymes. As such, we believe efficient multiplexed targeting can be achieved with further optimizations of the experimental conditions.

Having demonstrated the feasibility of multiplexed genome editing of endogenous genes implicated in ADA disorder, we next pursued the possibility of continuous editing to further enhance the overall gene targeting efficiency. Initially, we developed the RNA transfection method with the iPS-Cas9 system precisely for the prospect of repeated transfections. Given the ease of controlling Cas9 expression simply by the addition of doxycycline, we reasoned the RNA transfection method would allow repeated delivery of the gene editing components into the cells, increasing the chances of successful editing events. To test the notion of continuous editing, we co-transfected the sgRNA, HDR donor oligo targeting the ADA7 or ADA10 loci into dox-induced PGP1 iPS-Cas9 cells with B18R media supplementation, repeating the same transfection dose daily up to seven days. Three days after the final dose of transfection, we harvested the cells and analyzed the gene editing rates using next-generation sequencing. Deep sequencing of ADA10 loci revealed a 2.6-fold increase in the absolute value of HDR rates and 2.7-fold increase in NHEJ rates comparing Day 7 versus Day 1, while the overall ratio of HDR to NHEJ remained relatively flat (Figure 3-6).

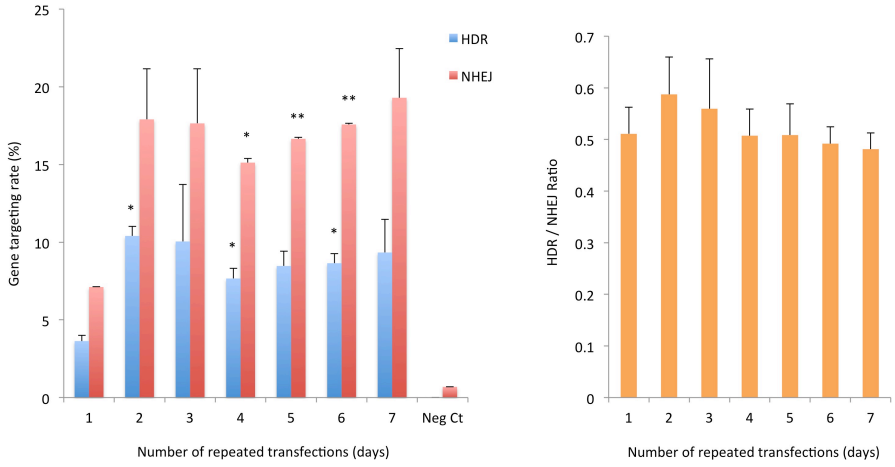

Figure 3-6: Continuous editing of ADA10 loci. Significance calculated by two-tailed t-test in comparison to Day 1 values: * $P < 0.05$, ** $P < 0.01$.

Intriguingly, sequencing analysis of the ADA7 loci also showed enhanced gene targeting rates from repeated transfections, but the effect was more pronounced on HDR events with a 9.7-fold improvement as opposed to 2-fold increase in NHEJ when comparing Day 1 with Day 7 results. The enhancement of HDR rates tipped the balance towards HDR with a favorable 4.9-fold increase in HDR to NHEJ ratio simply by continuous editing (Figure 3-7).

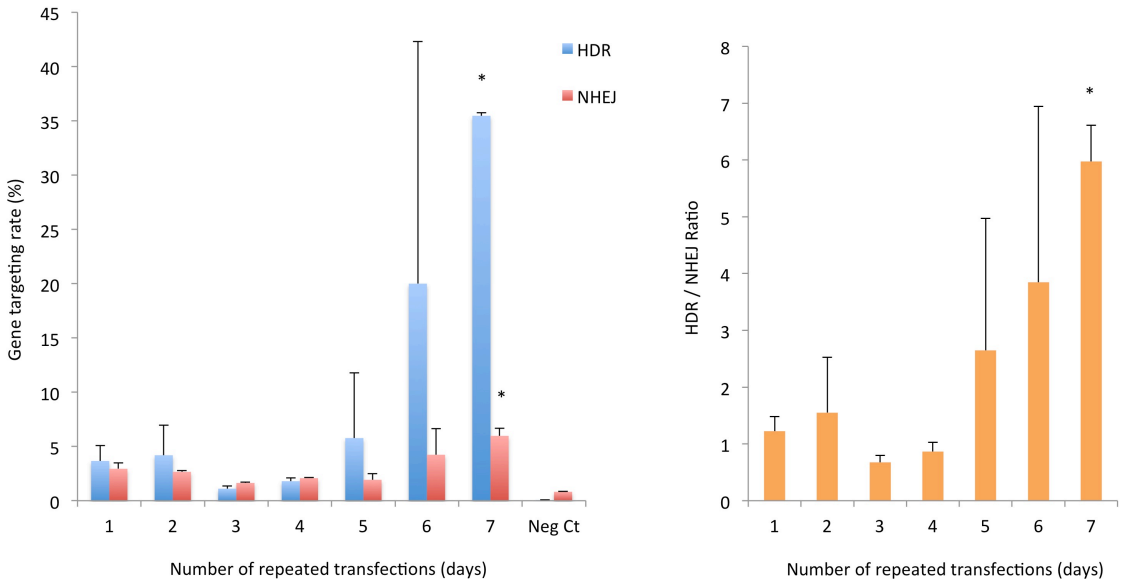

Figure 3-7: Continuous editing of ADA7 loci. Significance calculated by two-tailed *t*-test in comparison to Day 1 values: * $P < 0.05$.

Possible reasons for this observation may be that repeated transfections allow higher possibility of delivering all gene editing components into the same cell as NHEJ only requires the presence of sgRNA but HDR requires both the sgRNA and donor oligo to function. Additionally, since the HDR donor oligo was designed to disrupt the sgRNA recognition sequence, once the target site is repaired by the HDR pathway, it should be insusceptible to further targeting. On the other hand, repair through the NHEJ pathway often comprise of small indels. It is conceivable that as long as the sgRNA recognition site

and PAM sequence remains intact after NHEJ repair, the target site may be subject to further cutting, and hence another chance for repair through the HDR pathway. Also of note, the HDR to NHEJ ratio at ADA7 appears to be inherently better than the ADA10 loci perhaps due to differences in sequence compositions or epigenetic states near the target sites, therefore repeated targeting may further enhance the disparity observed in the HDR to NHEJ ratio.

3.2.3 Strategies for enhancing HDR-mediated repair rate

Strategies for shifting the balance away from NHEJ-mediated indel mutations toward precise modifications through the HDR pathway remain a major challenge for many genome editing applications. Although Cas9-mediated gene knockout through indel mutations introduced by NHEJ has demonstrated high efficiency in many systems, accurate modification of the genome through the HDR pathway continues to show limited efficiency[17,55,90]. Due to the low efficiency, a long and labor-intensive screening process is often required to isolate correctly edited cells (see Chapter 2.2.1). It will be particularly crucial to shift the balance away from competing NHEJ pathway for therapeutic applications dependent on successful HDR.

One of the perceived risk of manipulating the HDR:NHEJ ratio by inhibition of NHEJ is that it may not be well tolerated by most cells, as NHEJ plays a critical role in routine DNA repair. Therefore, we began by devising strategies to promote HDR rather than inhibit NHEJ. Given that precise repair through HDR requires a coordinated two-step process involving cleavage by Cas9 then repair by HDR machinery in the presence of a

donor template, we reasoned that the likelihood of proceeding with HDR may be promoted by coupling the donor oligo to the cutting site.

First, we explored the possibility of using single-stranded RNA as template for DSB repair or DNA as the single-guide sequence for Cas9-mediated cleavage. If either strategy succeeds, it would facilitate the design of a single construct, fusing the guide sequence with the single-stranded donor oligo in order to couple the cleavage and repair events. The use of RNA as DSB repair template was especially promising, as it has been demonstrated to work in yeast, albeit with lower efficiency than DNA templates[94]. To evaluate the feasibility of using RNA as repair template and DNA as single-guide sequence for Cas9, we designed the same constructs targeting AAVS1 locus except with RNA as the donor oligo and DNA as the single-guide sequence. We compared the two variations (sgRNA with ssRNA donor template, and sgDNA with ssDNA donor template) against the original setup (sgRNA with ssDNA donor template) to gauge the gene editing efficiency of each approach. Three days after co-transfecting the appropriate sgRNA-donor pairs into dox-induced cells, we harvested the cells and analyzed the gene editing rates with next-generation sequencing. Deep sequencing results revealed efficient gene editing rates with the original setup as expected (HDR = 5.54%, NHEJ = 10.25%) (Figure 3-8).

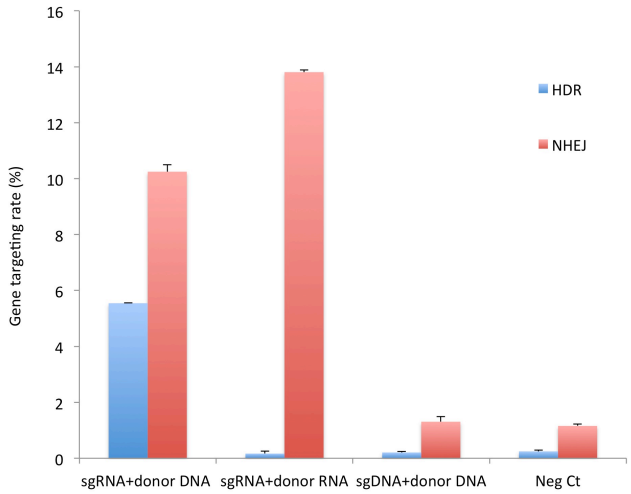

Figure 3-8: Gene editing efficiency of alternative pairings of guide sequence and HDR donor oligos.

While the NHEJ rate for the sgRNA/RNA donor pair was comparable to the original (NHEJ = 13.81%), the HDR rate was virtually undetectable (HDR = 0.17%), suggesting that while the sgRNA can complex with Cas9 and mediate cleavage reactions, the ssRNA donor oligo failed to serve as an efficient donor for DNA repair. Meanwhile, we observed minimal NHEJ and HDR using the sgDNA/DNA donor pair, implying that Cas9 could not function with single-guide sequences in the DNA form, resulting in deficient cutting and essentially undetectable repair through either pathway. These observations informed us that linking the guide sequence and donor oligo directly as one molecule with the same nucleotide composition was not a viable path forward.

Next, we examined the possibility of enhancing HDR efficiency with the coupling strategy by directly fusing the sgRNA and ssODN donor template together into a RNA/DNA chimera. We designed an sgRNA targeting the AAVS1 locus fused with a corresponding ssODN serving as repair template as a single molecule (Figure 3-9a). To test

the functionality of the sgRNA-ssODN fusion, we transfected the chimera into dox-induced PGP1 iPS-Cas9 cells and compared its gene editing efficiency with the individual sgRNA and ssODN co-transfections by MiSeq analysis. Deep sequencing results showed the functionality of sgRNA-ssODN chimera (HDR = 2.76%, NHEJ = 4.6%), although with approximately 3-fold reduction in efficiency compared to separate sgRNA and ssODN system (HDR = 8.7%, NHEJ = 13.8%) (Figure 3-9, b and c).

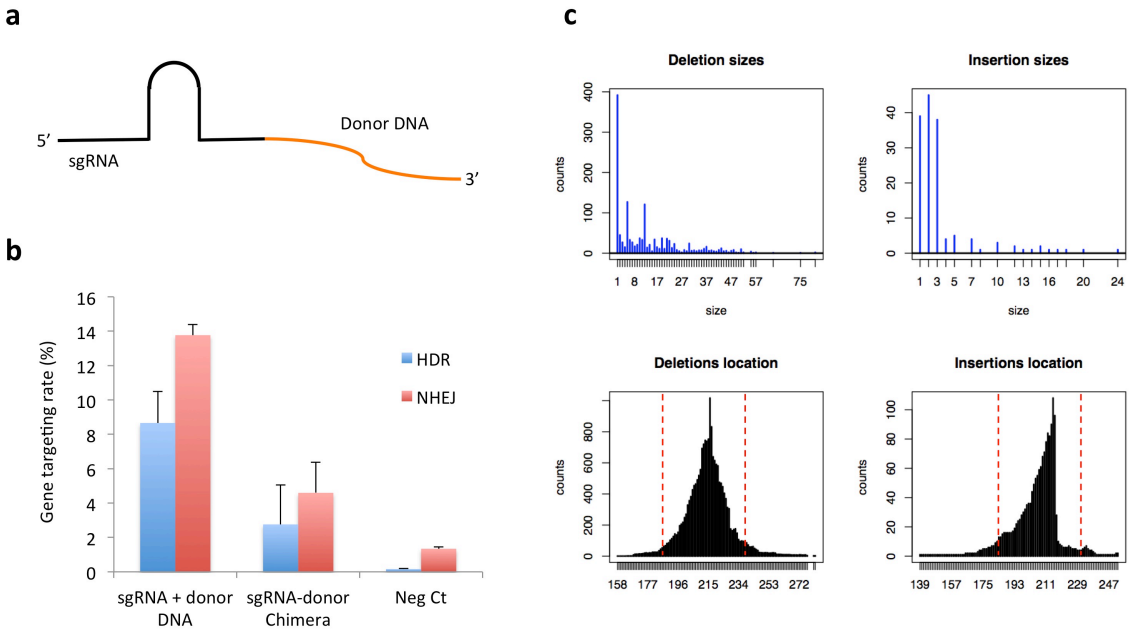

Figure 3-9: Potential of gene editing using single chimera sgRNA-donor molecule. (**a**) Design of sgRNA-donor DNA fusion (sgRNA in black, ssODN donor template in orange). (**b**) Comparison of gene targeting efficiency between separate sgRNA and donor DNA constructs, and single sgRNA-donor chimera design. (**c**) NHEJ profile of sgRNA-donor chimera shows indel formations around the target site.

A few factors may be at play. For one, the current design of sgRNA-ssODN chimera may not yet be optimal for facilitating coordinated action of Cas9 and HDR machinery. With the determination of Cas9 crystal structure in complex with sgRNA and target DNA, it was observed that the 3' end of the sgRNA protrudes through the back of the Cas9 complex,

positioning it a fair distance away from the active sites of RuvC and HNH cleavage domains (Figure 3-10)[95]. It is conceivable that adding a linker between the sgRNA and ssODN will further enhance its ability to interact with the DSB by extending its length and flexibility. Thereby, optimizing the sgRNA-ssODN chimera design with various linker lengths may further improve its efficiency. Another possibility is that Cas9-mediated cleavage and HDR repair occur sequentially through time and space. As such, the Cas9 protein needs to first release the target DNA, likely bringing the sgRNA-ssODN chimera along, before the HDR machinery can access the DSB for repair. In this scenario, having a covalently linked sgRNA-ssODN chimera would not provide much benefit, as the ssODN may be secluded within the Cas9 complex that has fallen off the target DNA.

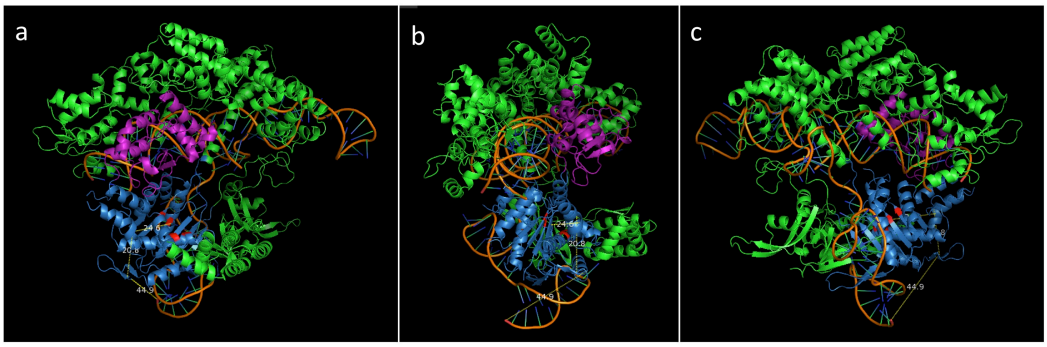

Figure 3-10: Crystal structure of Cas9 nuclease in complex with sgRNA and target DNA as solved by Nishimasu et al[95]. sgRNA and target DNA are depicted in orange. HNH domain is illustrated in magenta and RuvC domain in blue. (a) Front view. (b) Side view. (c) Back view. Structure downloaded from Protein Data Bank (PDBID 4oo8).

Illuminated by the observations with chimeras and the Cas9 crystal structure, we next sought to couple the ssODN repair template to sgRNA through non-covalent interactions. We designed various sgRNAs with an elongated 3' tail, serving as a docking site for ssODN hybridization through Watson-Crick base-pairing interactions. By anchoring

ssODN onto sgRNA through non-covalent interactions, we reasoned that sgRNA and ssODN can be pre-annealed and both be present at the target DNA site. After the sgRNA mediates Cas9 cleavage, the ssODN may be recognized as the repair template by the incoming repair machinery and subsequently released from the sgRNA dock to mediate HDR repair. To test the docking strategy, we designed a number of sgRNA-docks targeting the AAVS1 and TAZ loci, varying the linker length between the sgRNA and docking site, the length of the docking site, as well as the orientation of the free ssODN donor arm (Figure 3-11).

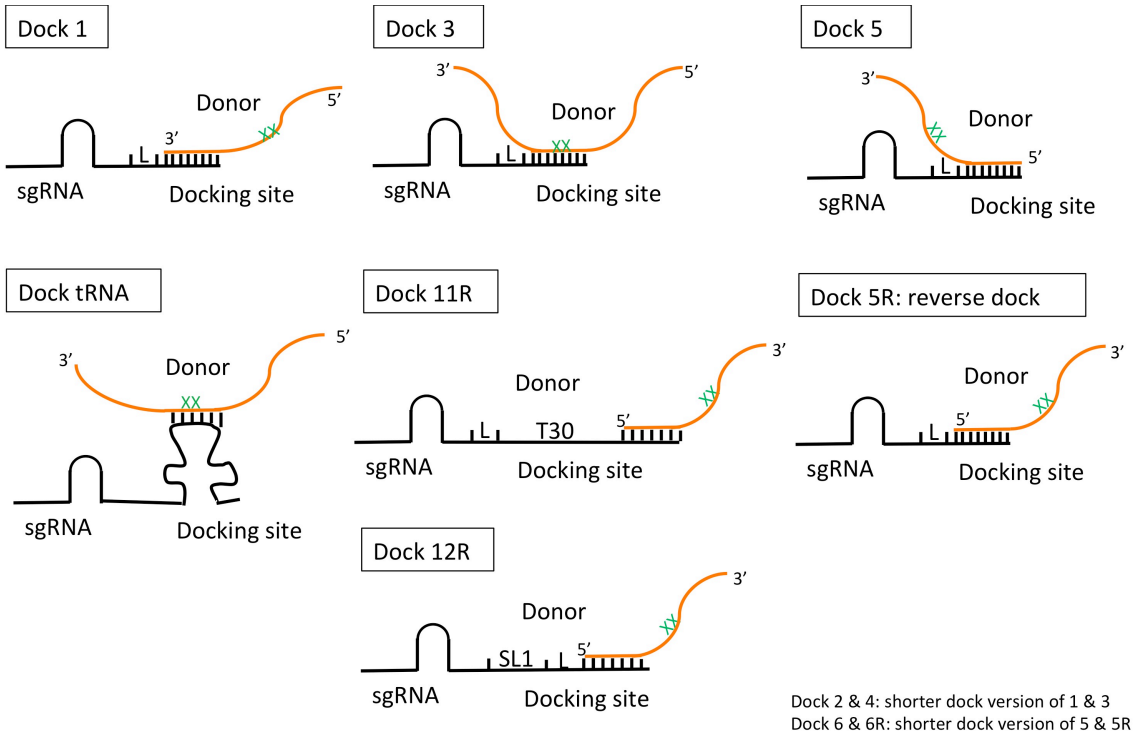

Figure 3-11: Schematic of sgRNA-dock designs.

To evaluate the efficiency of the docking strategy, we co-transfected the pre-hybridized sgRNA-dock and ssODN donor oligo into dox-induced iPS-Cas9 cells and analyzed the gene editing efficiency by MiSeq three days post-transfection. For both target

sites, the original sgRNA with free ssODN donor oligos still maintain the highest absolute percentage of gene editing. The reduced gene editing rates of dock designs may be attributed to the extended 3' tail as it can potentially hinder the interaction with Cas9. However, we noticed most of the docking strategy designs still induced reasonable HDR and NHEJ rates. In particular, dock 3 design achieved slightly better HDR:NHEJ ratio compared to the original system although the increase was not statistically significant (Figure 3-12). Several factors may come into play in the docking strategy. First, *in vitro* experiments have shown that Cas9- RNA remains bound to cleaved DNA[96], implying it will only be displaced by the incoming repair machinery. Since the mechanism for handing off the DSB from the Cas9 complex to the repair machinery is still unclear, it may be beneficial to protect the docked oligos with phosphorothioate bonds to prevent it from being degraded or resected by the repair machinery. In addition, the preliminary version of the sgRNA docks provide landing site for a single docking oligo, thus localizing only one copy of the donor near the cut. Since increasing intracellular donor concentration will enhance the HDR frequency up to a saturation point, the docking strategy may be improved by designing sgRNA docks with a longer 3' tail that allows multiple docking donor oligos to anneal and localize to the cleavage site. Taken together, the result implies that the docking strategy is functional in mediating gene editing, and may be engineered to further enhance the HDR:NHEJ ratio.

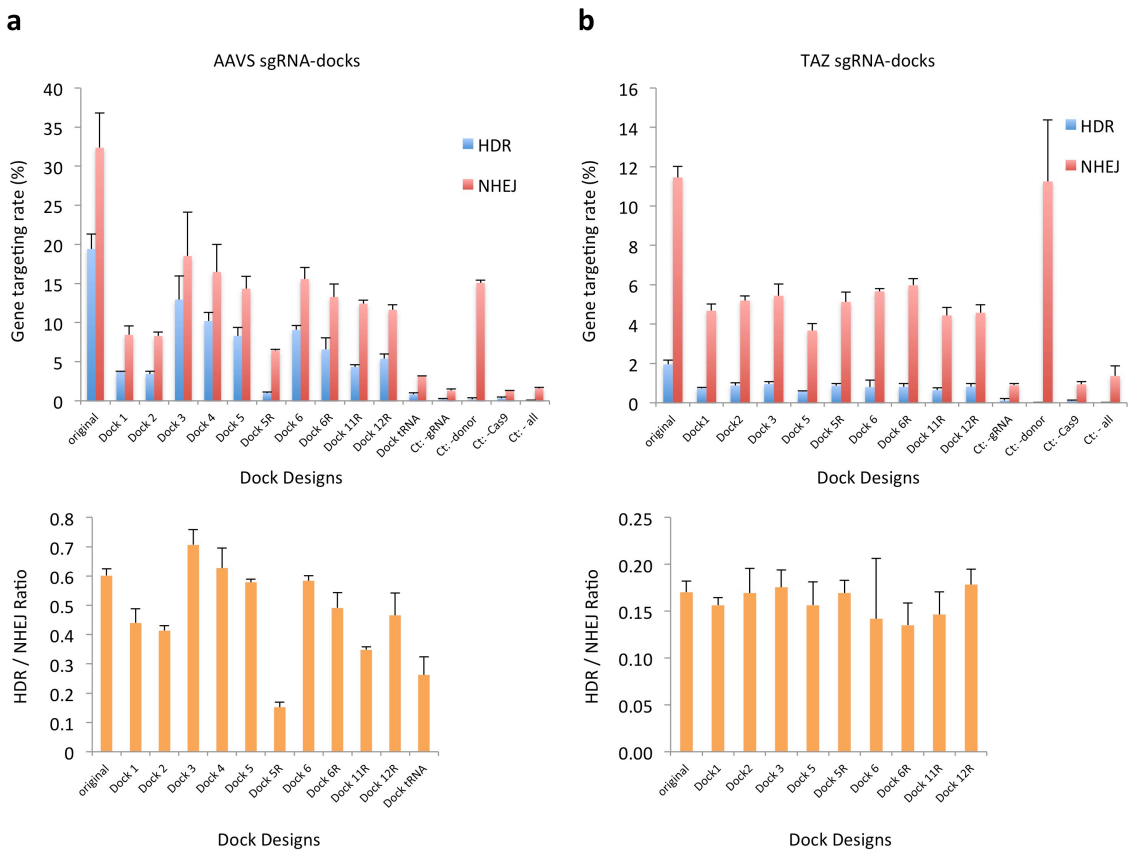

Figure 3-12: Evaluation of docking strategy for promoting HDR over NHEJ rate. (**a**) Gene targeting efficiency and HDR/NHEJ ratio for AAVS sgRNA-dock designs. (**b**) Gene targeting efficiency and HDR/NHEJ ratio for TAZ sgRNA-dock constructs.

Another approach reported to enhance genome editing specificity was through the use of double Cas9 D10A mutant nickases (Cas9n), a catalytically inactive version of Cas9 that binds but does not cleave target DNA, and a pair of offset sgRNAs. Studies of Cas9 specificity have shown that while the 20nt sgRNA contributes to target DNA recognition, mismatches can be tolerated depending on its position, quantity, and base identify[66,97]. Paired nickases have been developed to improve specificity by requiring the nicking of both target DNA strands by a pair of Cas9 nickases to induce site-specific DSB, whereas single nicks are mainly repaired by the high-fidelity base excision repair pathway (BER)[98]. As the double nickase method has been demonstrated to enhance genome editing specificity, we proceeded

to construct a PGP1 iPS-Cas9n cell line as a resource for nickase-based experiments. As before, we constructed the PGP1 iPS-Cas9n cell line by inserting a reverse tetracycline-controlled transactivator (rtTA) and Cas9n under the control of a tet response element (TRE) into PGP1 iPS cells via the piggyBac transposon. The system allows Cas9n expression to be tightly controlled and activated only by the addition of doxycycline into the culture media. The transfected cells were selected with puromycin and Cas9n integration was confirmed by qPCR. We envision the iPS-Cas9n cell line to be a useful research tool for further developments aimed at improving genome editing specificity.

3.2.4 Investigation of site-to-site variability in genome editing efficiency

Given that we have observed variability in HR:NHEJ ratio at different target loci, we next explored the impact of chromatin state on genome editing efficiency. We first analyzed the DNaseI hypersensitivity (HS) signal from an ENCODE iPS cell line (NIHi7)[99] for three sgRNA targets of interest – AAVS, ADA7, ADA10 – given that they have been used extensively in our experimental system. Interestingly, we observed a positive correlation (R^2 = 0.83, P = 0.27) between DNaseI HS signal and HR:NHEJ ratio, implying that more open chromatin is more prone to repair through the HDR pathway (Figure 3-13). Although the initial dataset was too small to be of significance, it revealed a potential role of chromatin state in regulating gene editing efficiency.

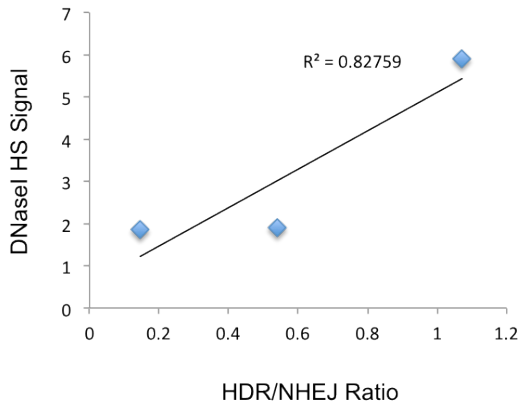

Figure 3-13: Correlation of DNaseI hypersensitivity (HS) signal with the HDR/NHEJ ratio at AAVS1, ADA7, and ADA10 genomic loci. (R^2 = 0.83, P = 0.27).

To further explore the effect of chromatin state on gene editing rate, we designed 16 sgRNAs and corresponding HDR donor oligos tiling from ADA exon 7 to exon 10 to better understand the landscape (Figure 3-14).

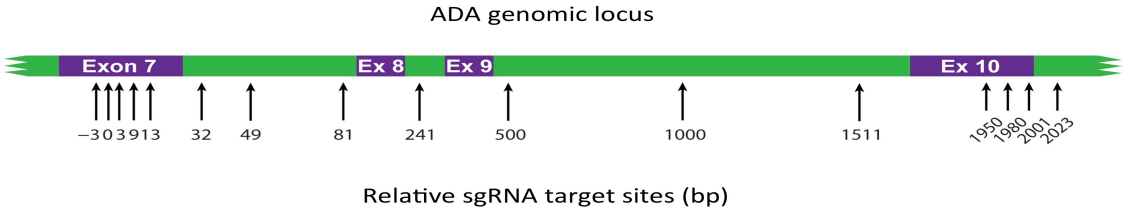

Figure 3-14: Design of tiling sgRNAs spanning ADA exon 7 to exon 10. sgRNA target sites indicated by arrow, numbers represent relative position in the genome. Note: scale for target sites <100bp has been enlarged for clarity.

We co-transfected each pair of sgRNA and HDR donor oligos into dox-induced PGP1 iPS-Cas9 cells and harvested the cells three days post-transfection for MiSeq analysis. Then, we assessed the gene editing rates from deep sequencing data and compared it against the DNaseI HS signal from the NIHi7 cell line. Intriguingly, the HDR/NHEJ ratio correlated poorly with DNase HS signal (R^2 = 0.15, P =0.14) as opposed to our earlier findings (Figure 3-15).

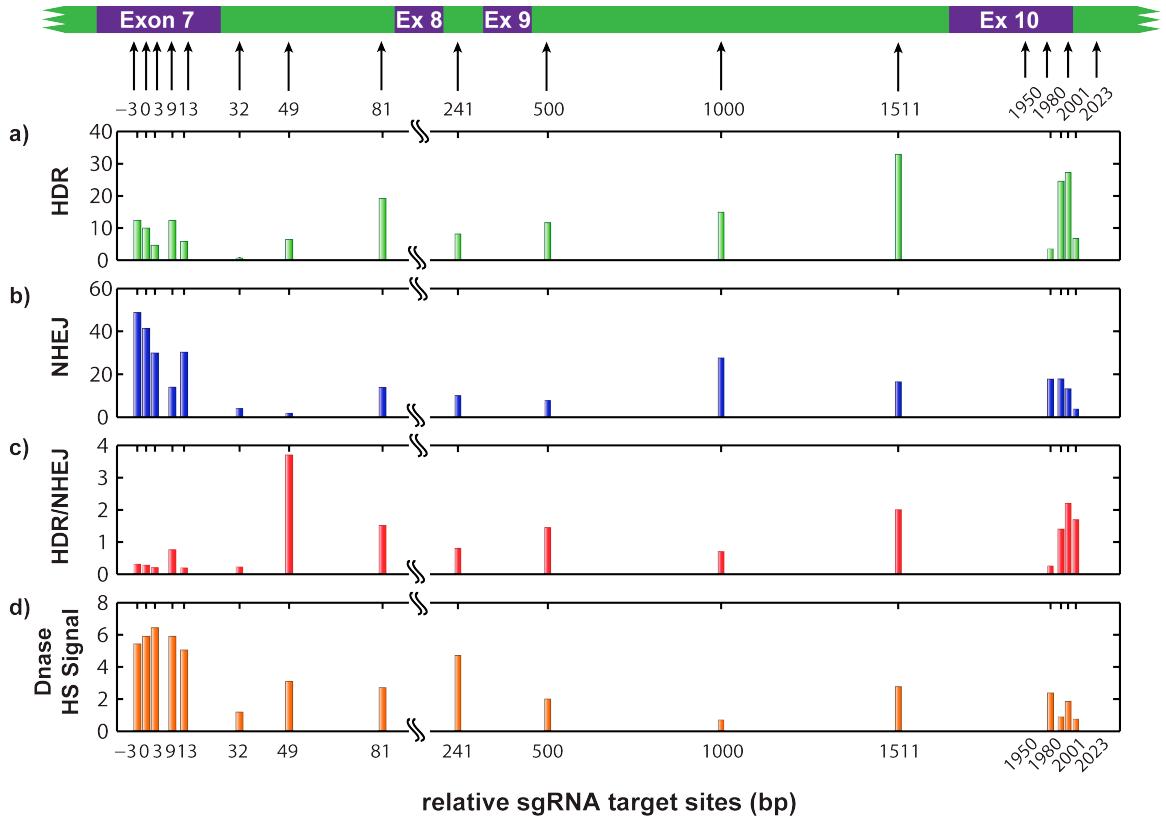

Figure 3-15: Comparison of gene targeting efficiency relative to DNase hypersensitivity (HS) landscape across ADA exon 7 and exon 10. The panels show (**a**) HDR frequency (**b**) NHEJ frequency (**c**) HDR/NHEJ ratio and (**d**) DNase HS signal across the ADA sgRNA target sites. Note: scale for target sites <100bp has been enlarged for clarity.

We reasoned the discrepancy may be due to the proximity of the sites targeted in the ADA landscape investigation, thus making it more difficult to pinpoint defined DNaseI HS signal across various sites. Additionally, it is possible that the chromatin status of the iPS cells continues to shift while in culture, potentially contributing to the variable HR:NHEJ ratio we have observed. Also of note, the NIHi7 iPS cell line used for DNaseI HS signal analysis is different from the PGP1 iPS-Cas9 cell line used for gene editing and only served as an initial reference point for comparison. Since cell lines derived from different sources have distinct genetic and epigenetic features, it would be ideal to compare the DNaseI HS signal and gene editing efficiency in the same cell line.

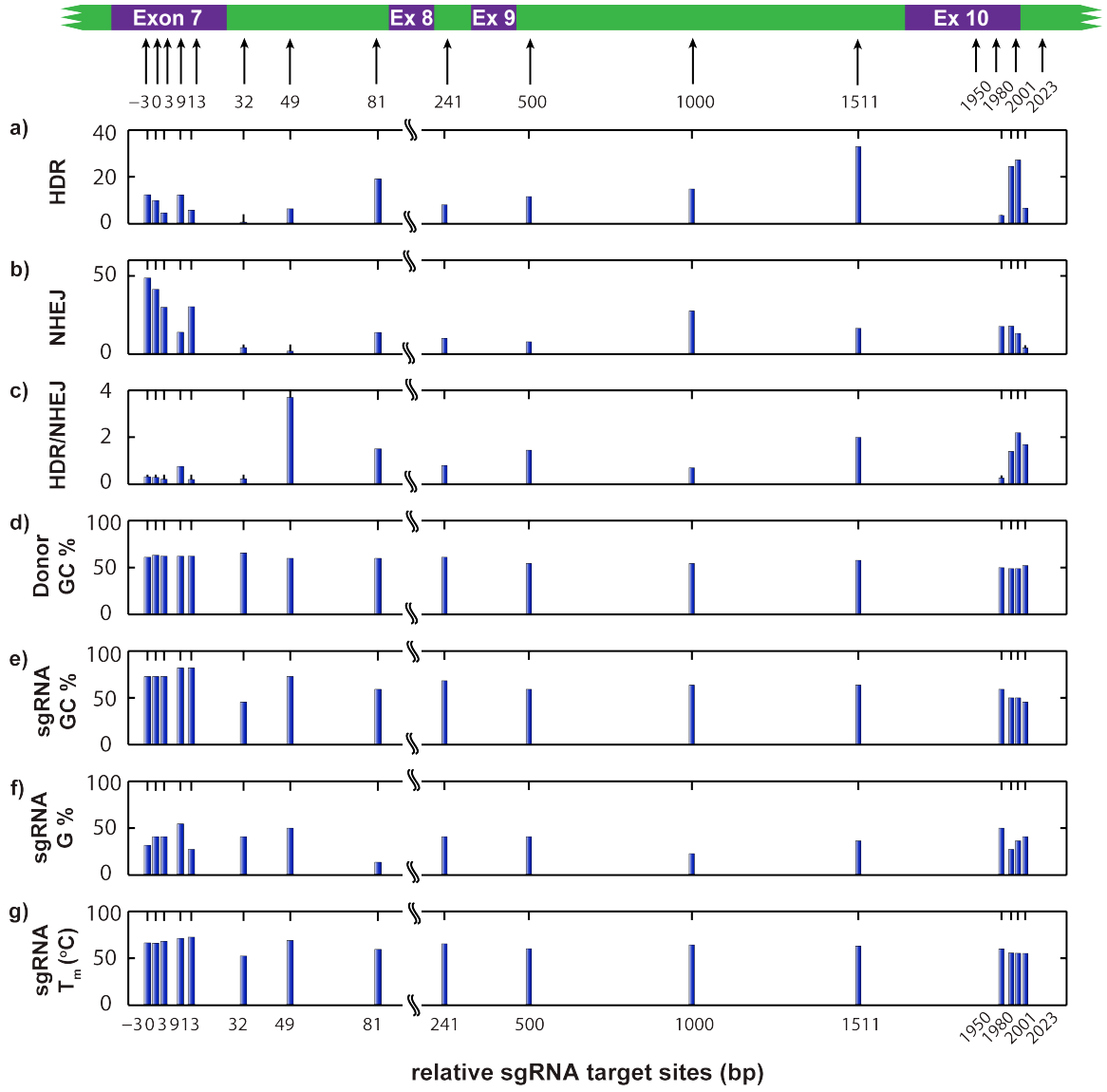

Figure 3-16: Comparison of gene targeting efficiency relative to donor DNA and sgRNA parameters. The panels show (**a**) HDR frequency (**b**) NHEJ frequency (**c**) HDR/NHEJ ratio (**d**) donor GC content (**e**) sgRNA GC content (**f**) sgRNA G content and (**g**) sgRNA melting temperature across the ADA sgRNA target sites. Note: scale for target sites <100bp has been enlarged for clarity.

In addition, we investigated various donor DNA and sgRNA parameters including GC content, G percentage, and melting temperature, but found no significant correlations to HDR and NHEJ frequency (Figure 3-16). Given the considerable range of possible variables, future studies investigating a wider spectrum of genomic targets and their corresponding

DNaseI HS signals across different time points will help elucidate the role of chromatin state in impacting gene editing ratios. It will also be interesting to examine other factors that may come into play, including positional effects, target sequence contexts, sgRNA and donor parameters, and transcriptional level of target genes. A deeper understanding of the critical parameters governing the HDR to NHEJ ratio will enable better strategies to engineer Cas9-based tools to further shift the DSB repair pathway choice towards HDR.

3.3 **Experimental Methods**

All oligonucleotide sequences used were synthesized by Integrated DNA Technologies (IDT).

3.3.1.1 Cell Line and Cell Culture

PGP1 iPS cell line and PGP1 iPS-Cas9 cell line were obtained and maintained with the same methods described in 2.3 Experimental Methods.

3.3.1.2 Design of sgRNA, HDR Donor Oligo, MiSeq Primers

sgRNA, HDR donor oligo, and MiSeq primers for each new target was designed and synthesized using the same principles outlined in described in 2.3 Experimental Methods. See Table 3-1, Table 3-2, Table 3-3, and Table 3-4 for a complete list of sequences.

3.3.1.3 RNA Transfection

For one-day single-target RNA transfections of PGP1 iPS-Cas9 cells, the procedures were similar to those outlined in Chapter 2.3. Briefly, $1–2 \times 10^5$ cells were seeded per 48-well plate one day before transfection and changed to media containing 0.5 – 1 ug/ml Doxycycline (Sigma-Aldrich) two hours prior to transfections to induce Cas9 expression.

RNA transfections were performed with RNAiMAX (Invitrogen) cationic lipid delivery vehicles following manufacturer's procedures, while specific parameters may be modified according to our optimization experiments. In general for each 48-well reaction, ~30 pmol sgRNA and ~70 pmol donor oligo were first diluted in 25ul Opti-MEM basal media (Invitrogen) while 6ul of RNAiMAX reagent was diluted in 25ul Opti-MEM basal media. Then, these components were mixed and incubated for 15 minutes at room temperature before adding to cells. Transfected cells were maintained in 37°C incubator and changed with mTeSR-1 Basal Medium (STEMCELL technologies) daily until harvest 3 days post-transfection. In certain experiments, the interferon inhibitor B18R (eBioscience) was added at 200 ng/ml as a media supplement.

For one-day double targeting of ADA7 and ADA10 experiments, the same procedures described above were employed with the difference of each sample receiving twice the amount of sgRNA, HDR donor oligio, and Dox.

For one-day multiple-donor competition experiments, the method outlined for one-day single-target conditions was performed with the difference that each sample received ~30 pmol sgRNA and ~80 pmol of pooled HDR donor oligos at equimolar ratio.

For continuous editing experiments, the same procedure for single-target RNA transfections was carried out, but repeated daily for 7 – 14 days. Cells were harvested 3 days post transfection as before.

3.3.1.4 MiSeq Library Preparation

MiSeq library was prepared using the same method as described in 2.3 Experimental Methods.

3.3.1.5 Sequencing Data Analysis

MiSeq reads were analyzed using the same platform as described in 2.3 Experimental Methods.

3.3.1.6 Isolation of Edited iPS Cell Clones

Human iPS cells grown on feeder-free conditions were treated with mTeSr-1 media supplemented with SMC4 (5 uM thiazovivin, 1 uM CHIR99021, 0.4 uM PD0325901, 2 uM SB431542)[100] for at least 2 h before fluorescence-activated cell sorting (FACS) sorting. Single-cell suspensions were generated by adding Accutase (Millipore) and resuspending in mTeSr-1 media supplemented with SMC4 and 0.5 ul of the viability dye ToPro-3 (Invitrogen). Next, live human iPS cells were single-cell sorted using a BD FACSAria II SORP UV (BD Biosciences) with 100 μm nozzle under sterile conditions to minimize stress on the hiPSCs. The cells were sorted into 96-well plates coated with irradiated CF-1 mouse embryonic fibroblasts (Global Stem) and hES cell medium[101] supplemented with 100 ng/ml recombinant human basic Fibroblast Growth Factor (Millipore), SMC4, and 5 μg/ml fibronectin (Sigma). After sorting, plates were centrifuged at 70*g* for 3 min and incubated in 37°C incubator. Colony formation was observed 4 – 7 days post sorting, and culture media was changed to cell medium with SMC4. Eight days after sorting, media was replaced with hES cell medium without supplements.

A few thousand cells were harvested 1 – 2 weeks after FACS. To extract genomic DNA, 0.1 µl of prepGEM tissue protease enzyme (ZyGEM) and 1 µl of prepGEM gold buffer (ZyGEM) were added to 8.9 µl of cells in the medium. The reactions were then added to 25 µl of PCR mix containing 12.5 ul of 2X KAPA Hifi Hotstart Readymix and 100 uM of forward and reverse primers. Reactions were incubated at 95°C for 5 min followed by 30 cycles of 98°C for 20 sec, 60°C for 20 sec, and 72°C for 20 sec. Products were Sanger sequenced, and sequences were analyzed with Lasergene (DNASTAR).

3.3.1.7 Analysis of DNaseI Hypersensitivity (HS) Sites

The DNaseI HS Overlap signal of the NIHi7 iPS cell line was downloaded from UCSC ENCODE dataset (https://genome.ucsc.edu/ENCODE/). The DNaseI HS signals for the regions of interest were obtained by running the bigWigAverageOverBed command on the base overlap signals. The resulting data for each target region were compiled for further analysis with MiSeq data.

Table 3-1. sgRNA design sequences.

ADA_ex7_T7	TAATACGACTCACTATAGGcgtactgtccacgccgggGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCT
ADA_ex10_T7	TAATACGACTCACTATAGGgacatgggctttactgaagGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCT
AAVS_T7	TAATACGACTCACTATAGGGGGGCCACTAGGGACAGGATGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCT
TAZ_ T7	TAATACGACTCACTATAGGGaagctcaaccatggggactGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCT

Table 3-1 (Continued).

AAVS-sgRNA-donor-chimera	rGrGrGrCrCrArCrUrArGrGrGrArCrArGrGrArUrGrUrUrUrUrArGrArG rCrUrArGrArArArUrArGrCrArArGrUrUrArArArArUrArArGrGrCrUrA rGrUrCrCrGrUrUrArUrCrArArCrUrUrGrArArArArArGrUrGrGrCrArC rCrGrArGrUrCrGrGrUrGrCrUrGGGAGGCCTAAGGATGGGGCTTTTCTGTCAC CAATggTGTCCCTAGTGGCCCCACTGTGGGGTGGAGGGGA Red: RNA bases. Black: DNA bases.
Dock_U6_gblock	TGTACAAAAAAGCAGGCTTTAAAGGAACCAATTCAGTCGACTGGATCCGGTACCAA GGTCGGGCAGGAAGAGGGCCTATTTCCCATGATTCCTTCATATTTGCATATACGAT ACAAGGCTGTTAGAGAGATAATTAGAATTAATTTGACTGTAAACACAAAGATATTA GTACAAAATACGTGACGTAGAAAGTAATAATTTCTTGGGTAGTTTGCAGTTTTAAA ATTATGTTTTAAAATGGACTATCATATGCTTACCGTAACTTGAAAGTATTTCGATT TCTTGGCTTTATATATCTTGTGGAAAGGACGAAACACCGGGGGCCACTAGGGACAG GATGTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTG AAAAAGTGGCACCGAGTCGGTGCTTTTTTTATCcttttatctgtcccctccacccc
ADA7 gRNA-2 (-3)	TAATACGACTCACTATAGGGcaccgtactgtccacgccgGTTTTAGAGCTAGAAAT AGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGG TGCT
ADA7 gRNA-3 (+3)	TAATACGACTCACTATAGGGactgtccacgccggggaggGTTTTAGAGCTAGAAAT AGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGG TGCT
ADA7 gRNA-4 bottom (+13)	TAATACGACTCACTATAGGGtcggccgagcccacctcccGTTTTAGAGCTAGAAAT AGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGG TGCT
ADA7 gRNA-5 (+9)	TAATACGACTCACTATAGGGcacgccggggaggtgggctGTTTTAGAGCTAGAAAT AGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGG TGCT
ADA7 gRNA-6 (+32)	TAATACGACTCACTATAGGGcgaagtagtaaaagaggtgGTTTTAGAGCTAGAAAT AGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGG TGCT
ADA gRNA-7 (+49)	TAATACGACTCACTATAGGGtgagggcctgggctggccaGTTTTAGAGCTAGAAAT AGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGG TGCT
ADA gRNA-8 (+81)	TAATACGACTCACTATAGGGcactgcctcctcccatactGTTTTAGAGCTAGAAAT AGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGG TGCT
ADA gRNA-9 (+241)	TAATACGACTCACTATAGGGagtggggaggaaccatcccGTTTTAGAGCTAGAAAT AGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGG TGCT

Table 3-1 (Continued).

ADA gRNA-10 (+500)	TAATACGACTCACTATAGGGgttccaggaaggccaaagaGTTTTAGAGCTAGAAAT AGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGG TGCT
ADA gRNA-11 (+1000)	TAATACGACTCACTATAGGGagccgccttccccaagacaGTTTTAGAGCTAGAAAT AGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGG TGCT
ADA gRNA-12 (+1511)	TAATACGACTCACTATAGGGtggagtagccagctcccagGTTTTAGAGCTAGAAAT AGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGG TGCT
ADA10 gRNA-13 Bottom (+1950)	TAATACGACTCACTATAGGGgggtggacttgaagatgagGTTTTAGAGCTAGAAAT AGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGG TGCT
ADA10 gRNA-14 (+1980)	TAATACGACTCACTATAGGGgattaccagatgaccaaacGTTTTAGAGCTAGAAAT AGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGG TGCT
ADA10 gRNA-15 (+2023)	TAATACGACTCACTATAGGGagtttaaaaggctggtgagGTTTTAGAGCTAGAAAT AGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGG TGCT

Table 3-2. PCR Primers for constructing dock sgRNAs.

AAVS_dock_F	TAATACGACTCACTATAGGGGCCACTAGGGACAGGATGTTTTAGAGCT AGAAATAGCA
Dock1_PCR-R (3'_Long_90donor)	ggggtggaggggacagataaaagGATAAAAAAAGCACCGACTCGGTGCC ACTTTTTCAAGTTG
Dock2_PCR-R (3'_Short_90donor)	ggaggggacagataaaagGATAAAAAAAGCACCGACTCGGTGCCACTTT TTCAAGTTG
Dock3_PCR-R (mid_Long_90donor)	gtcaccaatGGtgtccctagGATAAAAAAAGCACCGACTCGGTGCCACT TTTTCAAGTTG
Dock4_PCR-R (mid_Short_90donor)	tGGtgtccctagGATAAAAAAAGCACCGACTCGGTGCCACTTTTTCAAG TTG

Table 3-2 (Continued).

Dock5_PCR-R (5'_Long_90 donor)	ctaggaaggaggaggcctaagGATAAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTG
Dock5R_PCR-R (5'_Long_90 donor)	gaatccggaggaggaaggatcGATAAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTG
Dock6_PCR-R (5'_Short_90donor)	gaggaggcctaagGATAAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTG
Dock6R_PCR-R (5'_Short_90donor)	GAATCCGGAGGAGGATAAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTG
Dock7_PCR-R (3'_Long_180donor)	cggccctgggaatataaggtggGATAAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTG
Dock8_PCR-R (3'_Short_180donor)	ctgggaatataaggtggGATAAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTG
Dock9R_PCR-R (5'_Long_180donor)	gggtggaggacaatccgtctaGATAAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTG
Dock10R_PCR-R (5'_Short_180donor)	gggtggaggacaatcGATAAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTG
Dock11R_PCR-R (5'_polyT_90donor)	gaatccggaggaggaaggatcAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAGATAAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTG
Dock12R_PCR-R (5'_stem loop_90donor)	gaatccggaggaggaaggatcGATAACGGACTAGCCTTAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTG
TAZ_T7_PCR-F	TAATACGACTCACTATAGGGaagctcaaccatggggactGTTTTAGAGCTAGAAATAGCA

Table 3-2 (Continued).

Taz Dock1-PCR R (3' Long)	CTCCAAAATGAAGTCCATCCGATAAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTG
Taz Dock2-PCR R (3' Short)	tgaagtccatccGATAAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTG
Taz Dock 3- PCR R (mid Long)	acccagtGcccatggttgGATAAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTG
Taz Dock 4- PCR R (mid Short)	tGcccatggttgGATAAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTG
Taz Dock 5- PCR R (5' Long)	ccctgctgaccttctgggaaGATAAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTG
Taz Dock 5R- PCR R (5' Long)	aagggtcttccagtcgtcccGATAAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTG
Taz Dock 6- PCR R (5' Short)	ccctgctgacctGATAAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTG
Taz Dock 6R- PCR R (5' Short)	TCCAGTCGTCCCGATAAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTG
Taz Dock 11R- PCR R (5' polyT)	aagggtcttccagtcgtcccAAAAAAAAAAAAAAAAAAAAAAAAAAAAAGATAAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTG
Taz Dock 12R- PCR R (5' Stem Loop)	aagggtcttccagtcgtcccGATAACGGACTAGCCTTAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTG

Table 3-3. ssODN HDR donor template sequences.

ADA-ex7	CCAGGAGGCTGTGAAGAGCGGCATTCACCGTACTGTCCACGCCAGGGAGGTGGG CTCGGCCGAAGTAGTAAAAGAGGTGAGGGCCTGGGC
ADA-ex10	CTGGACACTGATTACCAGATGACCAAACGGGACATGGGCTTTACT<DEL(GAAG A)>GGAGTTTAAAAGGCTGGTGAGTGGGTGTGAGCCATACTGGCCTTG
SBDS-IV2	CAGTGCGTTTGGAACAGATGACCAAACTGAAATCTGTAAGCAGGCGGGTAACAG CTGCAGCATAGCTAACCCTAATAACCATTTATAACG
SBDS-IV3	GATAGAGAAAGATAGTGATTTCTTAAATGTGTTGGCATTTTTTTAAATTTTGAC TAAAGGAGAAGTTCAAGTATCAGATAAAGAAAGACA
AAVS_70	ggaggcctaaggatggggcttttctgtcaccaatGGtgtccctagtggccccac tgtggggtggagggga
AAVS_90	ctaggaaggaggaggcctaaggatggggcttttctgtcaccaatGGtgtcccta gtggccccactgtggggtggaggggacagataaaag
TAZ mut oligo 1	ccctgctgaccttctgggaagatatgcacccagtGcccatggttgagcttctcc aaaatgaagtccatcc
TAZ mut oligo 2	ccctgctgaccttctgggaagatatgcaGccagtGcccatggttgagcttctcc aaaatgaagtccatcc
TAZ mut oligo 3	ccctgctgaccttctgggaagatatgcacccagt<DEL (G)> cccatggttgagcttctccaaaatgaagtccatcc
ADA7 Mut Oligo-2 (-3)	ccttccaggaggctgtgaagagcggcattcaccgtactgtccacAccggggagg tgggctcggccgaagtagtaaaagaggtgagggcct
ADA7 Mut Oligo-3 (+3)	aggaggctgtgaagagcggcattcaccgtactgtccacgccgggAaggtgggct cggccgaagtagtaaaagaggtgagggcctgggctg
ADA7 Mut Oligo-4 bottom (+13)	cagcccaggccctcacctcttttactacttcggccgagcccacctTcccggcgt ggacagtacggtgaatgccgctcttcacagcctcct
ADA7 Mut Oligo-5 (+9)	ctgtgaagagcggcattcaccgtactgtccacgccggggaggtgAgctcggccg aagtagtaaaagaggtgagggcctgggctggccatg
ADA7 Mut Oligo-6 (+32)	actgtccacgccggggaggtgggctcggccgaagtagtaaaagaAgtgagggcc tgggctggccatggggtccctcctcactgcctcctc

Table 3-3 (Continued).

ADA Mut Oligo-7 (+49)	ggtgggctcggccgaagtagtaaaagaggtgagggcctgggctgAccatggggt ccctcctcactgcctcctcccatacttggctctatt
ADA Mut Oligo-8 (+81)	gggcctgggctggccatggggtccctcctcactgcctcctcccaCacttggctc tattctgcttctctacaggctgtggacatactcaag
ADA Mut Oligo-9 (+241)	acatgcacttcgaggtaagcgggccagggagtggggaggaaccaCccccggctg tcccaacttcctgtatagagaggcagaaagcagggc
ADA Mut Oligo-10 (+500)	gctctgttcccctgggcctgttcaattttgttccaggaaggccaTagagggaag aaactttagggattgggcatcagcccatgccgcgtc
ADA Mut Oligo-11 (+1000)	cacaggagcagtatcaggccttaggaaaaagccgccttccccaaAacaaggaca gcaagaactcagggtgaccatggtcaggccagcact
ADA Mut Oligo-12 (+1511)	tcgtgccaagaacagcttccatggtatgttggagtagccagctcGcagtgggac tgaggaacaagcagggtagggtgcagaggggaaggc
ADA10 Mut Oligo-13 Bottom (+1933)	ccgtttggtcatctggtaatcagtgtccagggtggacttgaagaCgagcgggtc atctgtgttgagcgagtagttagcctggtcattttt
ADA10 Mut Oligo-14 (+1980)	cgctcatcttcaagtccaccctggacactgattaccagatgaccTaacgggaca tgggctttactgaagaggagtttaaaaggctggtga
ADA10 Mut Oligo-15 (+2022)	caaacgggacatgggctttactgaagaggagtttaaaaggctggCgagtgggtg tgagccatactggccttgactcgggtttgggagtat
ADA10 Mut Oligo-16	tggacactgattaccagatgaccaaacgggacatgggctttactAaagaggagt ttaaaaggctggtgagtgggtgtgagccatactggc

Table 3-4. MiSeq PCR primer sequences.

ADA_ex7_F-primer	ACACTCTTTCCCTACACGACGCTCTTCCGATCTgtatgggaggaggcagtgag
ADA_ex7_R-primer	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTaggcagcatgactaggatgg
ADA_ex10_F-primer	ACACTCTTTCCCTACACGACGCTCTTCCGATCTtgacccgctcatcttcaagt
ADA_ex10_R-primer	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTgcagactcactccctctctc
SBDS_IV2_F-primer	ACACTCTTTCCCTACACGACGCTCTTCCGATCTaggaagatctcatcagtgcgt
SBDS_IV2_R-primer	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTtgatttcaggaggtttggca
SBDS_IV3_F-primer	ACACTCTTTCCCTACACGACGCTCTTCCGATCTctgctccagttgtgtgtgtc
SBDS_IV3_R-primer	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTgggcaaagctcaaaccattac
AAVS-F	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCGGTTAATGTGGCTCTGGTT
AAVS-R	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTACAGGAGGTGGGGGTTAGAC
ADA Miseq Primer F-6	ACACTCTTTCCCTACACGACGCTCTTCCGATCT ttccaggaggctgtgaagag
ADA Miseq Primer R-6	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTggtgtggtagccgtgtcc
ADA Miseq Primer F-9	ACACTCTTTCCCTACACGACGCTCTTCCGATCTgctaccacaccctggaagac
ADA Miseq Primer R-9	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTgccacccctcgagttcct
ADA Miseq Primer F-10	ACACTCTTTCCCTACACGACGCTCTTCCGATCTggtccagctacctcactggt
ADA Miseq Primer R-10	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTcccagggtgtcgaagagat

Table 3-4 (Continued).

ADA Miseq Primer F-11	ACACTCTTTCCCTACACGACGCTCTTCCGATCTcccatcctggagtctaacca
ADA Miseq Primer R-11	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTcagaacatcagagccgaagc
ADA Miseq Primer F-12	ACACTCTTTCCCTACACGACGCTCTTCCGATCTtggtcagagctaggaaagatcc
ADA Miseq Primer R-12	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTgagccaagaaagcaacatcc
ADA_ex10_F Primer-2	ACACTCTTTCCCTACACGACGCTCTTCCGATCTgctgattctctcctcctccc
ADA_ex10_R Primer-2	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTaccatactcccaaaccccgag
TAZ Miseq F	ACACTCTTTCCCTACACGACGCTCTTCCGATCTccccgagaatggttactgat
TAZ Miseq R	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTccatcccgctcatactgg

3.4 **Acknowledgments**

I thank Luhan Yang for helpful discussions throughout the project. I thank Marc Guell for assistance with the MiSeq data analysis. I thank Raj Chari for helping with DNase Hypersensitivity Signal analysis. I thank Ben Stranges for assisting with Cas9 crystal structure visualization. I thank Sam Chunte Peng for assistance with graphic illustration. I thank Margo Monroe and Dima Ter-Ovanesyan for stimulating discussions. I thank John Aach for reading and providing insightful comments on this chapter. I thank George M. Church for advice and support throughout the project.

CHAPTER 4 Conclusion and Outlook

In summary, we describe the development of a versatile and efficient genome engineering platform in human iPS cells in part I of the thesis. The system was built by the integration of Cas9 into the PGP1 iPS cell genome via a piggyBac transposon system. We have further optimized and characterized the system to achieve simple, efficient, continuous and multiplexable gene targeting in human iPS cells. In addition, we have explored various coupling strategies to enhance the HDR over NHEJ rate, paving way for future method developments to further shift the balance toward precise HDR repair. We anticipate the versatile platform to be a valuable resource for rapid generation of mutant human iPS cells for the study of gene functions and causal disease variants in an isogenic background that is scalable for high-throughput analysis. The high efficiency of the system allows rapid creation of disease models with greatly reduced screening complexity and without the need for drug selection, facilitating the functional study of genetic variations in development and disease. The ease of multiplexing will also advance the study of complex multigenic diseases and epistatic relationship of multiple genes. A similar approach may also be applied for constructing alternative Cas9-effector variants to repurpose the platform for investigation of

transcriptional regulation, epigenetic control, and live cell DNA imaging. The knowledge gained from genetic studies in the facilitated iPS-Cas9 genome editing system will ultimately provide valuable insights for future applications in clinical medicine, as genome engineering of patient iPS cells to correct genetic mutations via the HDR pathway may open the door to the future of hiPSC-based autologous transplantation therapy.

The pace of developments in Cas9 based genome engineering technology over the past couple of years has been phenomenal, however, several challenges remain to be addressed before clinical applications can be realized. The specificity of Cas9 targeting will need to be significantly improved before reaching the clinic. Toward this end, development of unbiased methods to globally assess off-target activity such as GUIDE-seq[102] will paint a more complete picture of the off-target effects of different genome editing platforms and help identify the safest route forward. Having a better understanding of Cas9 activity in the context of chromatin accessibility will also guide computational design of sgRNA targets with increased on-target efficiency and minimized off-target effects. Furthermore, methods for efficient delivery of Cas9 and sgRNA to specific cell types and tissues or developmental stages will be especially important for human gene therapy. Deeper structural and biochemical understanding of the CRISPR-Cas9 system may enable further engineering to identify Cas9 variants with increased specificity and broader targeting range that are also smaller and easier to deliver. Finally, strategies to shift the DSB repair pathway balance to enhance HDR will be critical for achieving therapeutic efficacy in a broad range of genetic disorders.

Given the unprecedented progress and power of Cas9-mediated genome editing technology, the scientific community recently raised the issue of ethical concerns regarding genome editing of the human germ line. Although the Cas9 system has been demonstrated

to facilitate genomic modifications in differentiated somatic cells, pluripotent stem cells, and animal eggs or embryos, a comprehensive analysis of safety is lacking as it may be impossible to fully comprehend the spectrum of potential problems arising from genetic modifications in an embryo until years after birth. Our limited knowledge of human evolutionary biology and disease, as well as gene-environment interactions makes consequences of heritable germline modifications hard to predict. Even assuming the safety and efficacy of genome editing technologies, the ethical issue of whether and when the use of this technology to modify the human germ line would be ethically justifiable remains a philosophical dilemma for humanity. As George Q. Daley, a professor of biological chemistry and molecular pharmacology at Harvard Medical School, sums up the issue in a recent New York Times article, "It raises the most fundamental of issues about how we are going to view our humanity in the future and whether we are going to take the dramatic step of modifying our own germline and in a sense take control of our genetic destiny, which raises enormous peril for humanity"[103]. The recent call by leading scientists within the genome editing community for a voluntary moratorium on human germline modification experiments and for the creation of open dialogue involving experts in science, bioethics, law, governmental agencies, as well as the public will initiate the process for educating, debating, and together formulating a responsible path forward for genome editing applications, taking all the risks and benefits into consideration.

The discovery and development of the powerful CRISPR-Cas9 genome engineering technology from the bacterial immune defense system underscores the importance of basic science research, attesting to the notion that science and technology together drives the advancement of human knowledge. The simplicity, efficiency, and versatility of the Cas9 system have revolutionized basic and biomedical research. Great strides made by the

scientific community and those to come will undoubtedly lead to future avenues for innovation in basic science, biotechnology, and clinical medicine.

Part II:

in situ Sequencing Technology

CHAPTER 5 Introduction

Each cell in a multicellular organism holds the same genetic code, yet the intricate regulatory mechanisms encoded by the genome leads to substantial spatial heterogeneity and complexity in different tissues, largely as a result of the differentially regulated intermediate gene expression pathway. As the first step in the gene expression pathway, transcriptional activity is tightly regulated in order to execute distinct developmental programs by expressing or repressing target genes whose expression is important to cellular fate. Therefore, the ability to detect and quantify mRNA levels is of significant importance in order to dissect biological and developmental questions of interest.

Traditionally, most gene expression analysis methods such as RT-PCR, microarray, and nanostring make ensemble measurements of averaged expression levels. Most of these methods require disaggregating and lysing cell populations, while certain methods also involve synthesis and amplification of cDNA which are then subject to quantitative analysis[104–106]. One limitation of these methods is the requirement for large numbers of cells to isolate sufficient material to perform an experiment. Yet perhaps the most critical shortcoming of these techniques is the lack of single-cell resolution as it is becoming

increasingly apparent that gene expression in individual cells may deviate significantly from the average behavior of cell populations. This is of particular importance when investigating a heterogeneous mixture of cells whose expression profile may differ largely due to regulatory programs in distinct developmental stages. The developing embryo, solid tumors, and brain tissue all represent examples of inherently heterogeneous biological samples. Furthermore, single cell analysis allows noise from stochastic biological fluctuations to be observed[107]. As a result, several new techniques have been developed such as single-cell RNA-Seq and microfluidic-based single-cell gene expression analysis[108,109]. Although achieving single-cell resolution, these methods are still limited by the need for cell lysis, thus leading to the partial degradation of RNA as well as the loss of spatial information. As steps along the gene expression pathway take place in distinct cellular compartments, the ability to maintain structural information is critical in obtaining a sharper picture of cellular processes. Therefore, to study the localization and fluctuation of expressed genes, single cell methods with single molecule sensitivity while preserving spatial orientation are essential.

5.1 **Brief overview of *in situ* RNA profiling technologies**

To satisfy the need for single-cell mRNA analysis with spatial information, methods including MS2 mRNA detection and in vivo hybridization of target mRNAs with molecular beacons have been developed with real-time imaging capability[110]. However, the need to create transgenes with long untranslated regions may affect normal mRNA dynamics and be unsuitable for multiplexing. A traditional technique that enables *in situ* single-cell mRNA analysis is fluorescent *in situ* hybridization (FISH) developed by Robert Singer and colleagues[111]. The method involves hybridizing five oligonucleotide probes, each conjugated

with five fluorophore moieties, to the mRNA target to create high-intensity signals that can be imaged and quantified. Although FISH showed single molecule sensitivity, it was also estimated that over 30% of the transcripts hybridized to zero or only one of the probes. Since a single probe binding event cannot be definitively discriminated between targeted binding and merely nonspecific binding, the specificity of the assay is limited.

More recently, several novel methods have been developed to address the issue of *in situ* mRNA detection from different standpoints. The first of which modified the Singer protocol by probing target mRNAs with more (> 30) and shorter oligonucleotides (20 bases), each hybridizing to a different region of the target mRNA[112]. The rationale is to label each oligonucleotide with a single fluorophore at its 3' end. Upon hybridization, many fluors are brought near the target and generate a diffraction-limited spot detectable by wide-field fluorescence microscope. This clever approach achieves *in situ* single mRNA detection with single-cell resolution.

Another interesting strategy employs orthogonal amplification with hybridization chain reactions (HCR) that function independently in the same sample to achieve detection of five mRNA targets simultaneously[113]. Once a specific RNA initiator is detected, the metastable fluorescent RNA hairpins self-assemble into polymers while recreating the initiator sequence for further signal amplification. This technique has been applied to *in situ* imaging of target mRNAs in fixed whole-mount and sectioned zebrafish embryos.

Lastly, the padlock probe (PLP) technology originally developed for detecting DNA molecules has been shown to be applicable for detecting and genotyping individual mRNA molecules *in situ*[114]. The strategy involves first converting the mRNA into localized complementary DNA (cDNA) molecules that are then hybridized by PLP and undergoes

target-primed rolling circle amplification (RCA), the signal of which are ultimately detected by fluorescently labeled detection probes.

Although several of the abovementioned technologies enable *in situ* detection of mRNA molecules in single cells, they are not highly scalable beyond detecting three to five transcripts, as the number of spectrally distinguishable fluorophores available becomes the limiting factor. Ideally, to fully appreciate the complexity of different transcriptional states in heterogeneous cell mixtures or even structured tissues, we would need to develop a novel method for highly multiplexed *in situ* sequencing of individual mRNA targets in single cells.

5.2 **Thesis Outline – Part II**

In CHAPTER 6, we describe the development of *in situ* sequencing method building upon the padlock probe (PLP) technology in collaboration with the Mats Nilsson group from Uppsala University in Sweden and investigate the specificity and scalability of the platform. We conclude that while the PLP method demonstrates high specificity, it is limited in terms of scalability, thus most suitable for the interrogation of few transcripts requiring high specificity. In CHAPTER 7, we describe the development of fluorescent *in situ* sequencing (FISSEQ) from our laboratory, an alternative and potentially complementary method with high multiplex capacity, providing a transcriptome-wide sampling method for RNA expression *in situ*. Overall, we demonstrate a highly multiplexed *in situ* sequencing technology that will enable systems level understanding of transcriptional profiles in various stages of biological development and disease.

CHAPTER 6 Development of *in situ* RNA sequencing with padlock probes (PLP)

6.1 Introduction

Gene expression level is tightly regulated by complex mechanisms and executes distinct developmental programs in a cell-type specific manner. Several methods have been developed over the past few decades to address the important challenge of detecting and quantifying mRNA levels for the study of gene expression and its biological consequences. However, most current methods suffer from various limitations such as the averaging of ensemble measurements in population-based assays when single cell resolution is needed, loss of spatial information, insufficient specificity, and low multiplex capacity. Therefore, we aim to develop a novel technology for highly multiplexed *in situ* sequencing of individual mRNA targets in single cells in order to obtain a better grasp of the complex transcriptional states in heterogeneous cell populations.

Considering the recent developments in *in situ* RNA profiling methods described in CHAPTER 5, the PLP technology is most readily adaptable to our ambitions. It is

conceivable to insert an anchor primer sequence into the backbone of the PLP followed by unique sequencing barcodes in the variable region of the PLP to detect the presence of target cDNAs with great specificity via sequencing by ligation. This strategy is also highly multiplexible as the number of transcripts sequenced equals 4^n where "n" is the number of bases sequenced. Therefore, we initially set out to develop an *in situ* RNA sequencing technology building upon the PLP approach. In this chapter, we first established a robust procedure for *in situ* mRNA detection for single targets and then extended the method for *in situ* sequencing. In addition, we assessed the specificity and scalability of the method and noted the high specificity of the system but relative difficulty to scale up. Therefore, we concluded that this method is best suited for applications interrogating few transcripts that require high specificity. The PLP method development described in this chapter has been performed in collaboration with the Mats Nilsson group from Uppsala University.

6.2 **Results and Discussion**

6.2.1 Establishment of *in situ* single mRNA detection

To develop a novel technology for *in situ* sequencing, we first set out to achieve *in situ* mRNA detection with currently available methods as a foundation to build upon. With the increasing importance of examining single cell gene expression *in situ*, various methods have been developed and modified. Two of the protocols that we have tested include the *in situ* reverse transcription (RT)-PCR method[115] and the *in situ* single mRNA detection method using PLPs[114]. The procedure for *in situ* RT-PCR involves fixing cells onto a slide, treating with proteinase K to partially permeabilize the cell membrane and with DNAse to eliminate

endogenous DNA, performing a one step RT-PCR amplification, and lastly detecting the signal by *in situ* hybridization with a fluorescently labeled probe. To test the efficacy of this protocol, we first attempted to detect *B-actin* and *GAPDH* transcripts in Hela cells with the rationale that housekeeping genes should be robustly detectable. After preliminary optimizations of specific steps in the protocol however, we observed only a low level of diffused signals (Figure 6-1).

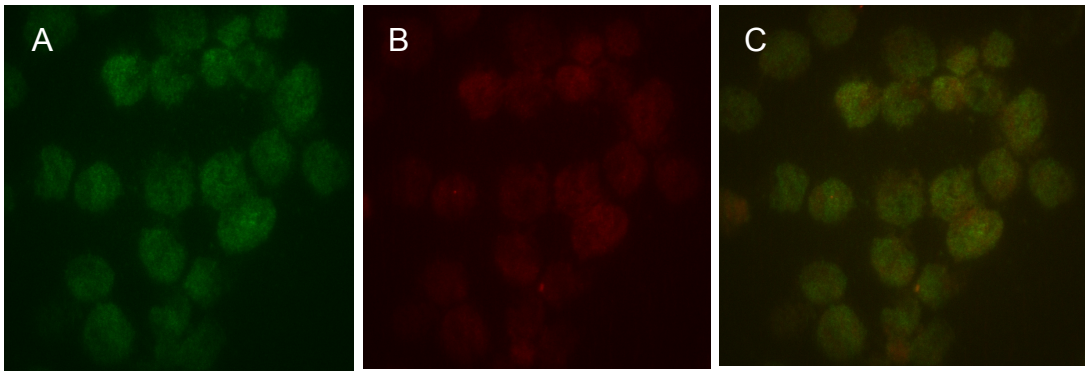

Figure 6-1: Preliminary results from *in situ* RT-PCR detection of *B-actin* transcript. RT-PCR for *B-actin* was performed and cells were probed both with (A) *B-actin* probes (Cy5, Green) for signal detection, and (B) *GAPDH* probes (Cy3, Red) as control for nonspecific probe binding. (C) shows the overlay of Cy3 and Cy5 channels.

Subsequently, we investigated the *in situ* PLP system in Hela cells probing for *B-actin* and *c-Myc* transcripts individually. The PLP protocol required first fixing cells onto the glass slide and partially permeabilizing the membrane, then converting the mRNA into localized cDNA molecules by reverse transcriptase. After padlock probe hybridization and target-primed rolling circle amplification (RCA), the signal was detected with fluorescently labeled detection probes. The preliminary images generated with the PLP method appeared more promising as we observed the corresponding Cy3 signal from *B-actin* detection and Cy5 signal from *c-Myc* detection (Figure 6-2). However, the fluorescent signal was diffuse, unlike the single molecule resolution that has been previously demonstrated. We reasoned that the

initial fixation and permeabilization steps may be critical parameters to optimize since RNA can be easily degraded or leaked out of the cells once the cell membranes are permeabilized.

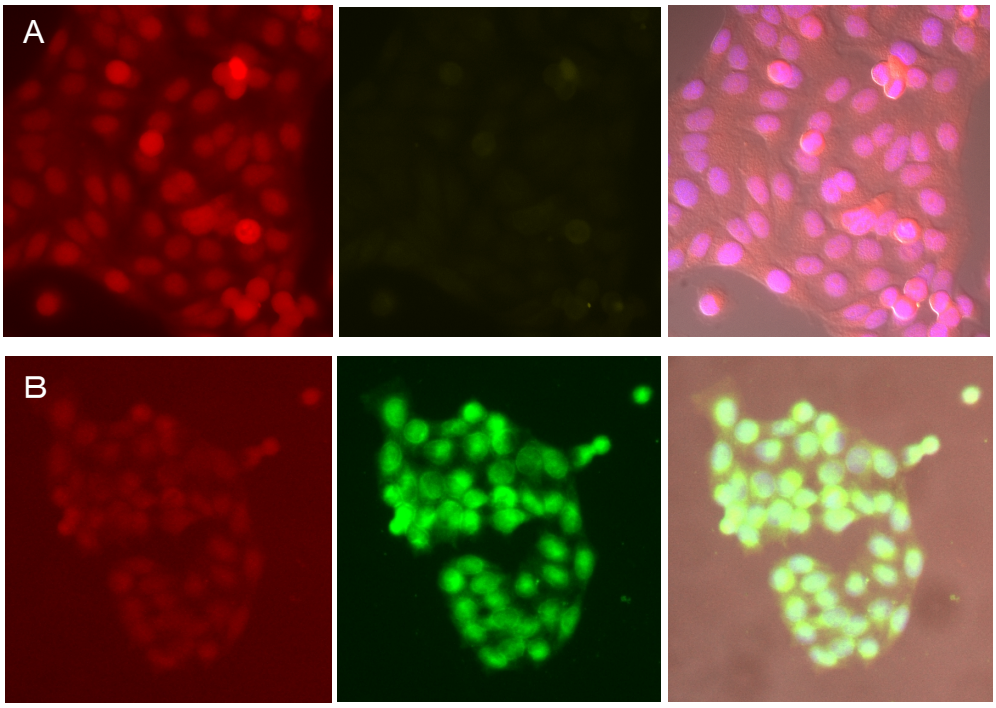

Figure 6-2: Preliminary investigation of *in situ* mRNA detection using PLP method. (A) *B-actin* mRNA (Cy3, Red). (B) *c-Myc* mRNA (Cy5, Green).

Given that several lab members have expressed similar problems with successfully reproducing the PLP method, we decided to reach out to the Mats Nilsson group, the developers of the PLP mRNA detection technology, for their expertise. After contacting the Nilsson group to express our vision for this project, we were fortunate to have established a collaboration with the ultimate goal of achieving highly multiplexed *in situ* sequencing of mRNA transcripts by combining their method for *in situ* detection and quantification of mRNA transcripts with our multiplex and automation powers on the Polonator. With their guidance, we have been able to generate beautiful single-molecule resolution images of individual *B-actin* transcripts detected *in situ* in human BJ fibroblast cells (Figure 6-3). The

PLP approach proved to be superior to the RT-PCR protocol for our purposes as the former yields much higher specificity attributed to PLP technology with localized signals resulting from target-primed RCA, not to mention the overall steps are simpler and less time-consuming to conduct.

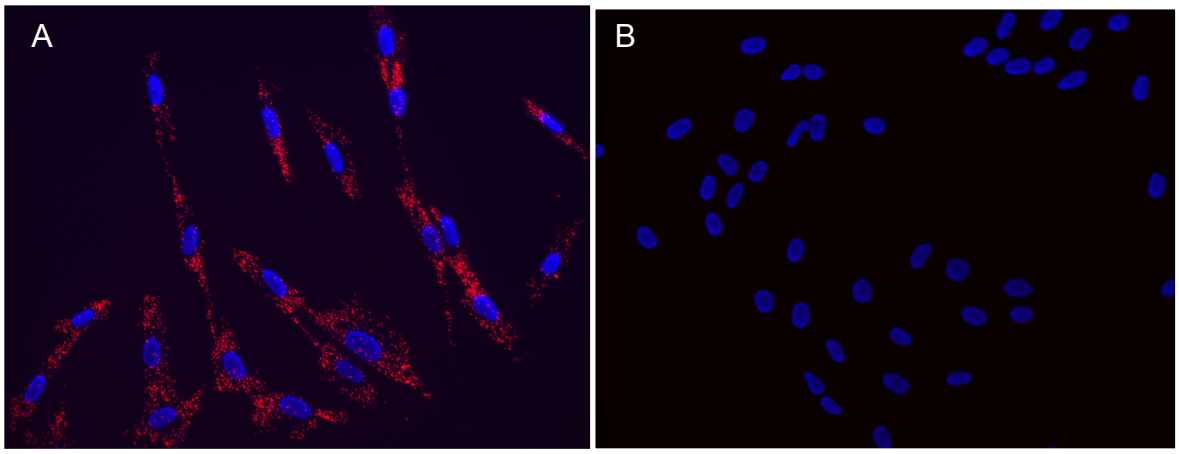

Figure 6-3: *in situ* detection of individual *B-actin* transcripts with PLP method. (A) *B-actin* (Cy3, red). Nuclei (Dapi, blue). (B) Negative control without RT.

To further validate the selectivity of the PLP, we reproduced the detection of a single nucleotide difference in the *B-actin* transcripts of co-cultured human fibroblast and mouse embryonic fibroblast cells (Figure 6-4). These results confirmed that the PLP method is sensitive for detecting single mRNA molecules, highly specific, and relatively simple to setup.

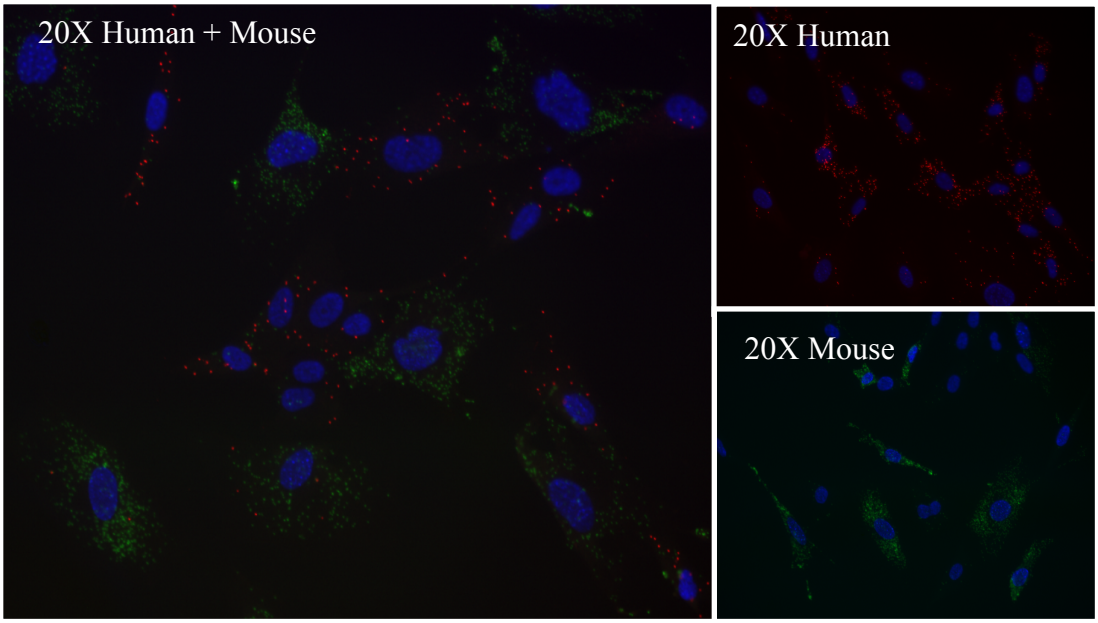

Figure 6-4: Demonstration of PLP specificity as it distinguishes a single nucleotide difference in *B-actin* transcripts in co-cultured human fibroblast and mouse embryonic fibroblast cells. For all three images, PLPs targeting both the human and mouse transcripts were added. Labels indicate the specific detection probes hybridized before imaging.

To assess the applicability of the PLP method, we performed *in situ* detection of *B-actin* in the PGP1 iPS cell line and observed robust rolony formation, indicating the ease of adapting this method for use in different systems (Figure 6-5).

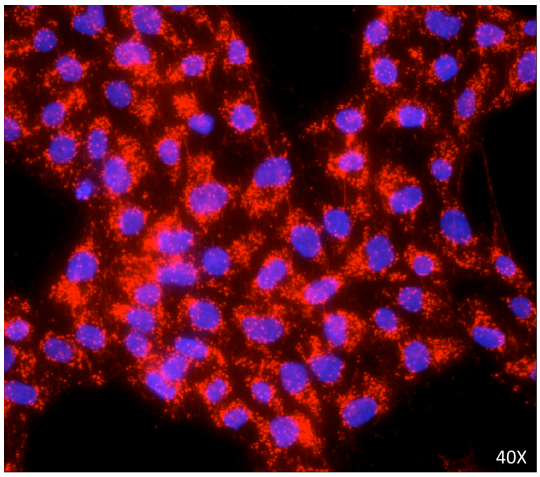

Figure 6-5: *in situ* detection of single mRNA molecule in PGP1 iPS cell line. (*B-actin*: red, Nuclei: blue).

In addition, we investigated the compatibility of the PLP detection method with a micropattern array chip from Cytoo. The 2 cm by 2cm, 170μm thick chips are organized into a grid of 144 micropattern arrays that allow cells to morph into the customized patterns specified on the chip, standardizing cell position, shape, and polarity. To assess the possibility of studying how cell shape affects gene expression, we next sought to adapt the PLP method for cells grown on the micropatterned chips. Human BJ fibroblasts were cultured on the standard micropattern chip at a density to allow one cell per pattern. After the cells settle into the shapes, we proceed with the PLP method as described earlier. Imaging analysis revealed that the PLP method is readily adaptable for cells grown on micropatterned chips, while further optimizations may increase the number of cells attached to the grids and the number of rolonies detected (Figure 6-6). The combination of *in situ* mRNA detection on micropatterned chips offers the potential to study the interplay between internal cell organization and gene expression patterns, particularly valuable for cells displaying high polarity such as neurons and epithelial cells.

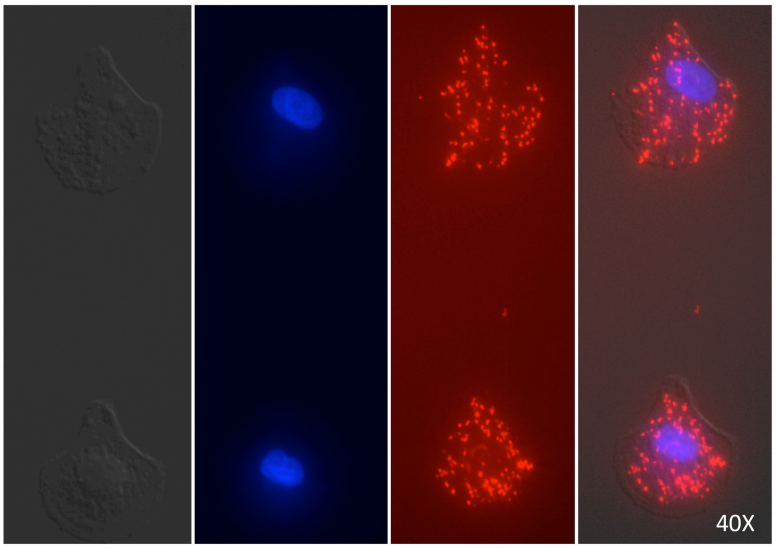

Figure 6-6: *in situ* detection of single mRNA molecules in human BJ fibroblast cell line on Cytoo micropattern arrays. (*B-actin*: red, Nuclei: blue).

6.2.2 Establishment of *in situ* sequencing of single mRNA molecules

After achieving *in situ* detection of individual mRNA molecules by probing, the next logical step was to develop *in situ* sequencing techniques for one transcript by manual sequencing. Preliminary testing indicated that sequencing by ligation method was more feasible compared to other sequencing chemistries for this system. Thus for the initial development, we investigated the potential of *in situ* sequencing by ligation, targeting the *B-actin* transcript in human BJ fibroblasts (Figure 6-7).

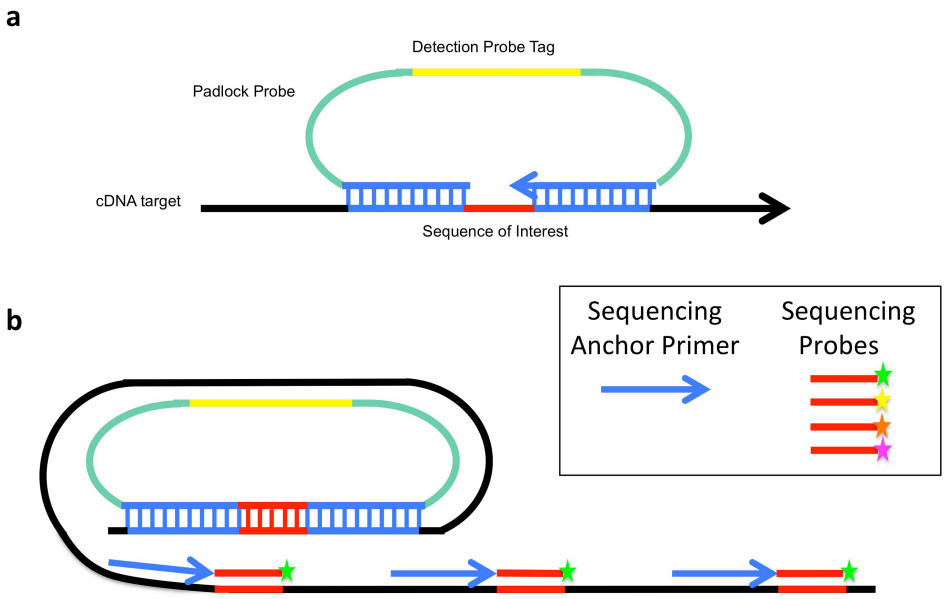

Figure 6-7: Schematic of padlock probe design for manual sequencing by ligation. (**a**) PLP design in the context of the cDNA target. Blue: 5' and 3' arms of PLP complementary to cDNA target. Green: Common PLP backbone. Yellow: Tag sequence for detection probe hybridization. Red: Sequence of interest to be captured by PLP. (**b**) Sequencing schematic: Sequencing anchor primer is first hybridized to the RCA product (black). Then, sequencing by ligation cycles are performed with all combinations of sequencing probes.

In addition to the modified probe design, several adjustments to the detection by probing protocol have been made to accommodate the later sequencing steps (Figure 6-8).

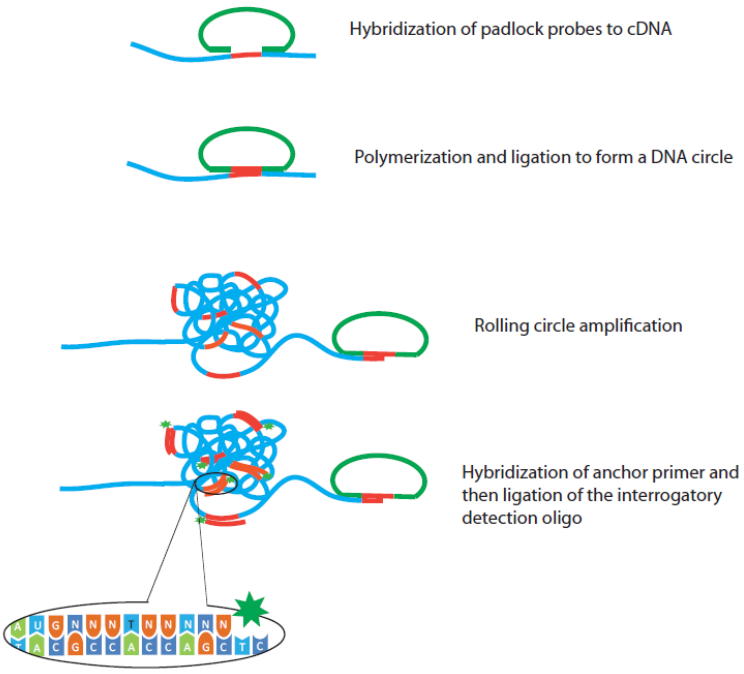

Figure 6-8: Overview of *in situ* sequencing by ligation procedure. (Figure courtesy of the Mats Nilsson group).

As before, the cells were fixed, permeabilized with ethanol washes, and reverse transcribed with a specific RT primer. In the second enzymatic reaction, stoffel fragment was added to fill in the 4-nucleotide gap between the two arms of the PLP, along with ampligase and Rnase H in the original reaction. Then, RCA and detection probe hybridization was performed as usual to check the outcome of the molecular reactions. If RCA signals can be successfully detected by probing, then we proceed to manually sequence the transcript by ligation. For the sequencing reactions, we first stripped off the probes with 65% formamide washes. Next, we hybridized the anchor primers, which in this case recognize the region corresponding to the 3' arm of the PLP. Subsequently, we added T4 ligase along with the 4 sequencing probes for interrogating the first base, each tagged with a different fluorophore, to allow the correct probe to be ligated to the anchor primer. After visualizing the results of 1st base sequencing under the microscope, the sample was incubated

with UNG enzyme to digest the U bases incorporated into the anchor primer to yield smaller fragments that can be easily washed off by formamide. From this point on, subsequent bases of interest can be sequenced by repeating the procedures outlined above. As demonstrated in Figure 6-9, the initial attempt at *in situ* sequencing of *B-actin* transcripts in human fibroblasts appeared promising as the fluorescence signal from each sequencing cycle corresponded to the target sequence TGCA. In summary, it was feasible to manually sequence at least four bases *in situ* by combining the previously developed PLP technology with sequencing by ligation method.

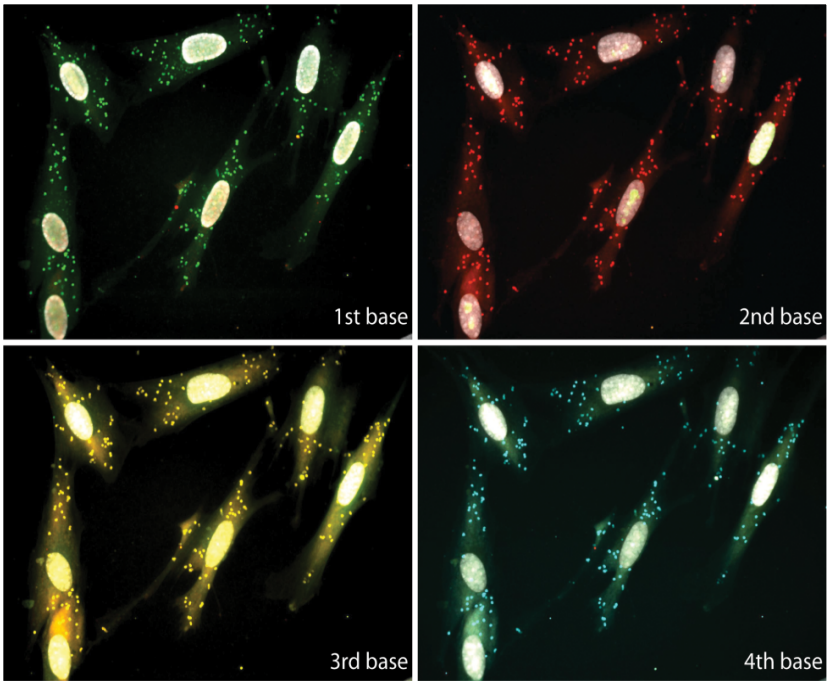

Figure 6-9: Preliminary data for *in situ* manual sequencing of four bases of the *B-actin* transcript in human BJ fibroblasts. The target sequence is TGCA. First base: T (FITC, green), second base: G (Cy3, red), third base: C (TexasRed, yellow), fourth base: A (Cy5, aqua). (Figure courtesy of the Mats Nilsson group).

6.2.3 Assessment of multiplex potential of PLP-based *in situ* sequencing technology

After achieving *in situ* sequencing for one transcript, we next sought to take a step further towards developing *in situ* sequencing technology for transcriptomic assays in single cells by highly multiplexing the method. With the recent remarkable progress in next generation DNA sequencing technologies, it is now imaginable to apply certain concepts like miniaturizing, localizing, and parallelizing to *in situ* transcript detection and bring *in situ* transcriptome sequencing into reality. For initial proof of concept experiments, we took a targeted sequencing approach that aims to detect and quantify a specific subset of transcripts in individual cells. Specifically, we first designed and tested PLPs for ~10 transcripts to investigate the multiplex capacity and efficiency of the method.

To begin, we designed PLPs and RT primers targeting a set of 10 transcripts with allele-specific expression (ASE) from a well-studied network in which the expression levels are known[116]. As illustrated in Figure 6-10, the PLP consisted of a 5' arm that hybridizes to the cDNA followed by an anchor primer sequence, a unique sequencing tag of 4 bases, a linker sequence to increase the backbone length and improve hybridization efficiency, and a 3' arm that binds to the cDNA on the other end and specifies the allele under question. In general, besides the 5' and 3' cDNA hybridization sequences and the sequencing tag used to identify each transcript, the backbone of all PLPs are designed to be identical to minimize variations.

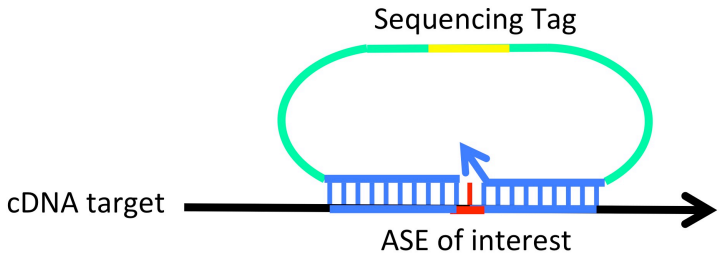

Figure 6-10: PLP design for multiplexed *in situ* sequencing of ASE transcripts.

For the design of RT primers, we followed the basic principles laid out in the original PLP method[114]. For instance, the RT primers are found to be most efficient when they target a region as close to the PLP hybridization site as possible, incorporate at least 5 locked nucleic acid (LNA) modified bases to increase hybridization efficiency, and exclude LNA bases from the stem of a hairpin as well as in the PLP hybridization site.

In terms of the sequencing detection probe design, we employed 9-mers each specifying the base in question followed by degenerate bases and tagged with a different fluorophore as shown in Figure 6-11. To sequence the first base, a pool of 4 different probes (ANNNNNNNN, TNNNNNNNN, CNNNNNNNN, GNNNNNNNN) were added and the complementary probe would be ligated onto the anchor primer and give off a fluorescent signal detected by an epifluorescent microscope. For sequencing of subsequent bases, the corresponding sequencing probes would be added to the reaction. A method to reduce the complexity of the pool will be to fix the last 5 bases of the sequencing oligo since the sequences can be purposely made identical in the PLP design. This will be sufficient for our initial test, as unique 4 base barcodes will theoretically enable the detection of 256 (4^4) transcripts. The advantage of this approach is that while reducing pool complexity, we

increase the concentration of the perfectly matching sequencing probe and thus may potentially increase the specificity and reduce the background of the sequencing reactions.

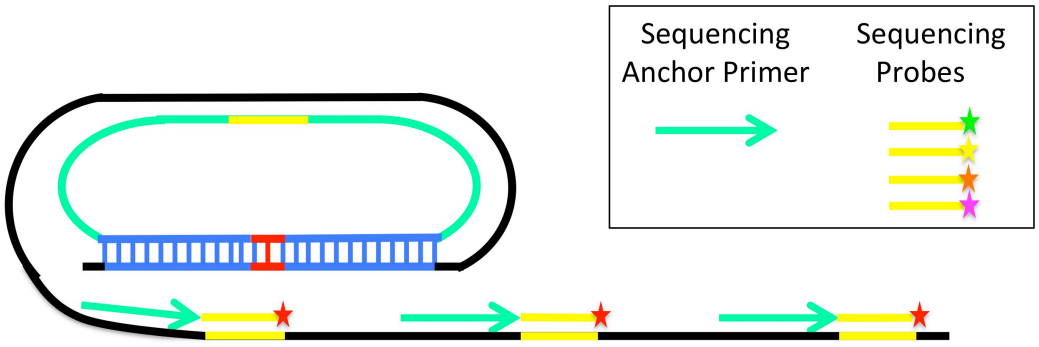

Figure 6-11: Overview of sequencing by ligation using PLP targeting ASE transcripts.

Given that the designs are based on theoretical assumptions and prior empirical data, each PLP and RT primer would have to be tested to ensure their functionality and efficiency. We began by testing the efficiency of MRFAP1 and TBL1X designs in the PGP1 fibroblast and iPS cell lines. We observed that while the initial PLP design for each target yielded detectable rolonies, the number of transcripts detected was very low in both the fibroblast and iPS cell lines (Figure 6-12).

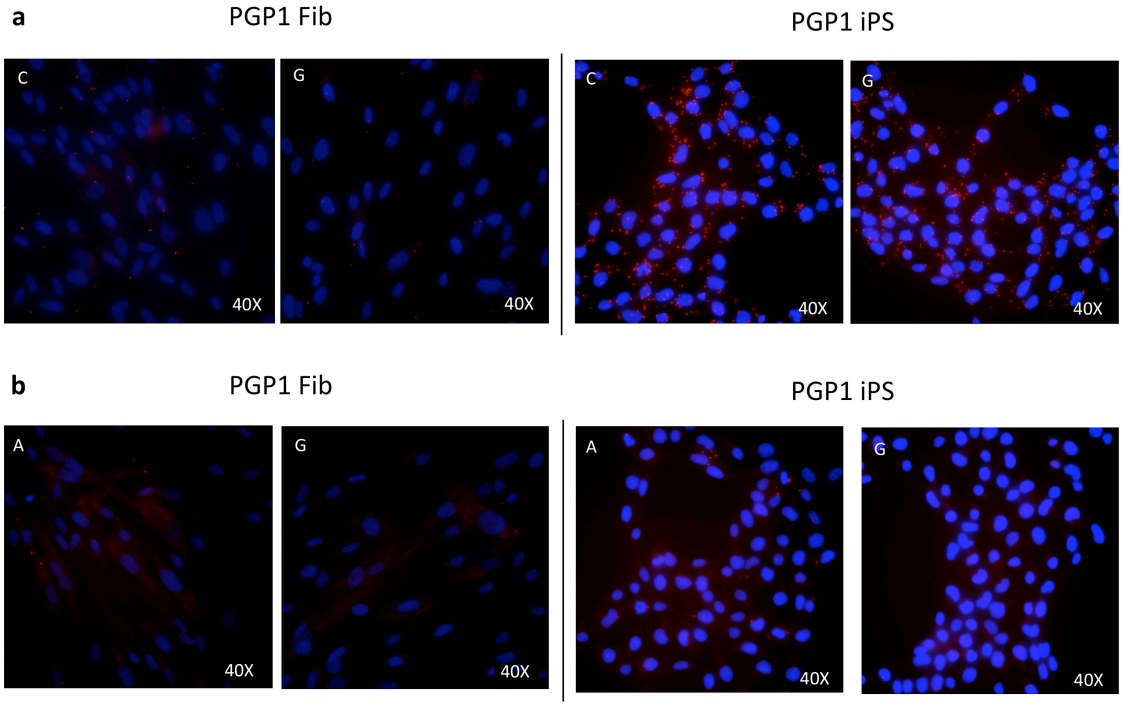

Figure 6-12: Validation of probe designs by *in situ* detection of ASE transcripts targeting the (**a**) MRFAP1 transcript and (**b**) TBL1X transcript. Letter in the upper left hand corner denotes the specific allele being probed in the image.

For the MRFAP1 transcript, both alleles showed relatively equal representation while the TBL1X transcript revealed allele specific expression favoring the A allele. The biased expression in TBL1X gene was expected, as the gene is located on the X chromosome of a male donor cell line, only the maternal copy would be expressed (Figure 6-13).

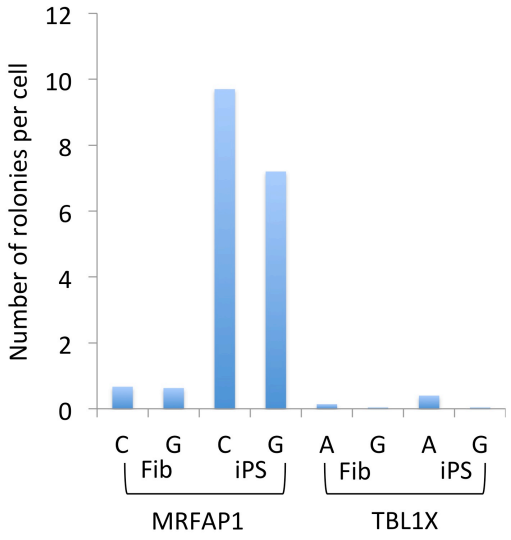

Figure 6-13: Quantification of ASE transcripts detected by PLP in PGP1 fibroblast and iPS cell lines.

Upon closer inspection of the gene expression data for the ASE transcripts we have initially selected, we noted that most of the targets exhibit low or moderate gene expression levels compared to the reference B-actin. Based on the B-actin reference and assuming similar PLP detection efficiency for all the transcripts, the number of expected rolonies per cell would be 71 for MRFAP1 and 22 for TBL1X (Table 6-1). Given variations in sequence contexts and PLP efficiencies, the actual number of detectable rolonies would likely be even smaller, thus explaining the low number of rolonies detected in our preliminary studies.

Table 6-1. Gene expression profile of the selected ASE targets and the expected PLP detection efficiency.

Gene	Illumina Gene Expression	Expected # of rolonies/ fibroblast
MRFAP1	5698	71
TBL1X	1783	22
HNRNPA0	4020	50
VANGL1	119	1.5
ARL15	202	2.5
PLCB3	92	1
RCL1	379	4.7
FBXO42	142	1.8
REEP5	4958	62
AKAP12	172	2
GBAS	1188	14.8
B-actin (ref)	48,000	600

Given this result and our experience with optimizing the PLP method for multiplexed transcript detection, we noticed the following limitations. First, while the sensitivity for detecting abundant transcripts like *B-actin* is more than sufficient, it becomes very difficult to probe low or even moderately expressed genes. In addition, although PLP are highly specific, they exhibit inherent bias and variability, thus needs to be optimized and validated individually. This may be a deterrent for the high multiplex capacity we would like to achieve. With increased number of targets, we anticipate the detection efficiency to be further reduced. Also, while the ASE detection relies on simple PLP ligation, the earlier *de novo* sequencing of four bases in *B-actin* required an additional polymerase fill in step to complete the gapped PLP for ligation. Although it has been demonstrated to work on *B-actin*, the additional molecular reaction greatly reduces the detection efficiency and would

likely require highly abundant transcripts and optimized PLP designs to function efficiently. Due to the inefficient polymerase gap-fill in step, the read length of the *de novo* sequencing by ligation would also be limited. Lastly, the use of LNA RT primers would also be prohibitively expensive for highly multiplexed applications. Given the technical limitations, we conclude that the PLP technology is best suited for detecting fewer transcripts with high specificity, but may not be the ideal platform for developing highly multiplexed *in situ* RNA sequencing. Although we did not proceed further to develop the PLP technology for highly multiplexed *in situ* RNA sequencing, the experience gained through working with *in situ* PLP detection still proved to be informative and contributed to design principles in the alternative FISSEQ strategy that will be presented in the following chapter.

6.3 **Experimental Methods**

6.3.1.1 Cell culture and sample preparation

The human BJ fibroblast cell line was cultured in DMEM (Gibco) supplemented with 10% FBS (Sigma) and penicillin-stretomycin (Life Technologies). Cells were incubated at 37°C, 5% CO_2.

To prepare cells for *in situ* detection, confluent cells were dissociated with TrypLE Express Enzyme (Life Technologies) and resuspended in culturing medium. Three milliliters of cells were then seeded onto five Superfrost Plus slides (Thermo) in a 150 150 mm x 25 mm petri dish (Corning) with 25 ml final culture medium. Cells were incubated and allowed to attach for 12-24 hours. Slides with co-cultured cell lines were prepared similarly except with a mixture of different cell lines. Next, cells were washed in PBS two times and fixed in 3% (w/v) paraformaldehyde (Sigma) in DEPC-treated PBS for 30 min at RT. After fixation,

slides were washed twice in DEPC-treated PBS and dehydrated in a series of 70%, 85%, and 100% ethanol washes for 5 min each.

6.3.1.2 Sequence Designs

See Table 6-2 for a list of RT primer, PLP, detection probe, anchor primer, and sequencing oligo designs. All oligonucleotide sequences used were synthesized by Integrated DNA Technologies (IDT). All LNA sequences were synthesized by Exiqon.

6.3.1.3 Reverse transcription

All subsequent molecular reactions were performed in Secure-Seal hybridization chambers (Invitrogen). First, cells were washed with DEPC-PBS-Tween (DEPC-PBS-T) followed by incubation in 0.1 M HCl in DEPC water for 5 min and two washes with DEPC-PBS-T. Next, reverse transcription reaction containing 1 uM LNA-modified cDNA primer, 20 U/ul of reverse transcriptase (Fermentas), 500 uM dNTPs (Fermentas), 0.2 ug/ul BSA (NEB) and 1 U/ul RiboLock RNase Inhibitor (Fermentas) were mixed in the reverse transcription buffer and added to the reaction chamber. The reverse transcription reaction was incubated at 37°C for 1 h. Slides were washed twice with PBS-T followed by a postfixation step in 3% (w/v) paraformaldehyde in DEPC-PBS for 10 min at RT and two more washes in DEPC-PBS-T.

6.3.1.4 PLP hybridization and ligation

All padlock probes were 5' phosphorylated to allow ligation. After reverse transcription, the next reaction simultaneously degrades RNA, hybridizes the PLP, ligates the barcode PLP or fills in the gap for gapped PLP, and forms a complete DNA circle.

For the barcoded PLP, ligation reaction containing a mix of 100 nM of padlock probe, 0.5 U/ul Ampligase, 0.4 U/ul RNaseH (Fermentas), 1 U/ul RiboLock Rnase Inhibitor, 50 mM KCl and 20% formamide was mixed in 1x Ampligase buffer, and added to the reaction chamber. The slides were incubated at 37°C for 30 min and 45°C for 45 min, followed by two washes in 1x DEPC-PBS-T.

For the gapped PLP, the reaction containing 0.2 U/ul Stoffel fragment (Applied Biosystems), 0.5 U/ul Ampligase, 0.4 U/ul RNase H (Fermentas), 100 nM PLP, 50 uM dNTPs, 1 U/ul RiboLock RNase Inhibitor, 50 mM KCl and 20% formamide were mixed in 1× Ampligase buffer (20 mM Tris-HCl, pH 8.3, 25 mM KCl, 10 mM $MgCl_2$, 0.5 mM NAD and 0.01% Triton X-100). The slides were incubated at 37°C for 30 min and 45°C for 45 min, followed by two washes in 1x DEPC-PBS-T.

6.3.1.5 Rolling Circle Amplification (RCA) and signal detection

For RCA, a reaction containing 1 U/ul phi29 polymerase (Fermentas), 0.25 mM dNTPs, 0.2 ug/ul BSA, 5% glycerol in DEPC water were mixed in 1x phi29 polymerase buffer and added to the reaction chamber. The slides were incubated for 2 h at 37°C. Following the incubation, the slides were washed three times in DEPC-PBS-T. For signal detection, 100nM of detection probe in 2x SSC and 20% formamide was added to the reaction chamber and incubated at 37°C for 30 min to allow hybridization. Excess detection probes were washed away with three DEPC-PBS-T washes. Next, the secure seal chambers were removed and slides were dehydrated through an ethanol series. The slides were then prepared in Vectashield mouting medium (Vector) containing 100ng/ml DAPI (4', 6'-diamidino-2-phenylindole) for nuclei counterstaining and fluorescence signal were analyzed using an epifluorescence microscope (Leica).

6.3.1.6 Sequencing by ligation

Prior to the sequencing reaction, detection probes from signal detection were first stripped off. The slides were washed in an ethanol series to remove the mounting medium and allowed to dry at RT. For detection probes without uracils, the samples were washed with DEPC-PBS-T, incubated in 65% formamide three times for 30 s, then washed twice with DEPC-PBS-T. For detection probes that contained uracils, samples were first treated with UNG treating buffer (0.02 U/ul UNG (Fermentas), 0.2ug/ul BSA, 1x phi29 polymerase buffer (Fermentas)) for 10 min, then washed twice with DEPC-PBS-T before formamide incubation.

Next, a mix of 500 nM anchor primers in 2x SSC and 20% formamide were added to the sample and incubated at RT for 30 min followed by two washes in DEPC-PBS-T. A ligation mix containing the 100 nM of each interrogation probe, 0.1 U/ul T4 ligase (Fermentas), 1mM ATP (Fermentas) were mixed in 1x T4 ligase buffer (Fermentas). The ligation mixture was added to the samples and incubated for 30 min at RT. Unligated probes were washed away with three 1 min incubations in DEPC-PBS-T. The slides were mounted in Vectashield mounting medium with 100 ng/ml DAPI. After imaging, the sequencing cycle was repeated by stripping the probes, hybridizing the next set of anchor primers, ligating corresponding interrogation probes, and imaging.

6.3.1.7 Image acquisition and analysis

Fluorescence images were acquired using an epifluorescence microscope (Leica). Subsequent image analysis was performed with CellProfiler cell image analysis software (v.2.0) available at http://www.cellprofiler.org/examples.shtml.

Table 6-2. Sequence designs of RT primers, PLP, detection oligos, anchor primers, and sequencing oligos.

Seq Type	Target	Sequence
LNA RT Primer	Actin	5' - ATCATCCATGGTGAGCTGGCGGCGG - 3' Positions 2,4,6,8,10,12,14 = LNA bases
PLP	Actin	AGCCTCGCCTTTGCCTTCCTTTTACGACCTCAATGCTGCTGCTGTACTAC TCTTCGCCCCGCGAGCACAG
Gapped PLP	Actin_4 nt	5'- AGGCCGGCTTCGCGGGCGACGGCGACTATGATTACTGACTGCGTCTATTT AGTGGAGCCCTATCTTCTTTCAACGGCTCCGGCATG - 3'
Detection Oligo	Actin_P LP	5'-TGCGTCTATTTAGTGGAGCC-3'
Detection Oligo	Actin_G apPLP	CCTCAATGCTGCTGCTGTACTAC
LNA RT Primer	pLNA_MR FAP1	TCCAGTTTAGGGTCCAATGCAGACA
PLP	Pd_MRFP 1_C	AGTATGGAACTGTCTGCACCTCAATGCTGCTGCTGTACTACGATAAGTCG GAAGTACTACTCTCTGGGCTTGGGGATGAC
PLP	Pd_MRFP 1_G	AGTATGGAACTGTCTGCACCTCAATGCTGCTGCTGTACTACCATAAGTCG GAAGTACTACTCTCTGGGCTTGGGGATGAG
LNA RT Primer	pLNA_HN RNPA0	TTGAAGTAAATTGTCATAGAAATGA
PLP	Pd_HNRN PA0_A	TTGTTTATCATCTTGACATGCCTCAATGCTGCTGCTGTACTACGGTAAGT CGGAAGTACTACTCTCTATCTCTTGTACTAAGCGAA
PLP	Pd_HNRN PA0_G	TTGTTTATCATCTTGACATGCCTCAATGCTGCTGCTGTACTACCGTAAGT CGGAAGTACTACTCTCTATCTCTTGTACTAAGCGAG
LNA RT Primer	pLNA_TB L1X	CAAACTGGTCTTTGAACCTCCTTTG
PLP	Pd_TBL1 X_A	ATTGCTCTCACAAAGGAGCCTCAATGCTGCTGCTGTACTACGACAAGTCG GAAGTACTACTCTCTCTAACAATTTGGACACTACA
PLP	Pd_TBL1 X_G	ATTGCTCTCACAAAGGAGCCTCAATGCTGCTGCTGTACTACCACAAGTCG GAAGTACTACTCTCTCTAACAATTTGGACACTACG

Table 6-2 (Continued).

LNA RT Primer	pLNA_VANGL1	AATCTTTTAAATGTCATTTACATTG
PLP	Pd_VANGL1_T	TGCAGTGGGAACAATGTCCTCAATGCTGCTGCTGTACTACGAAAAGTCGGAAGTACTACTCTCTTTTATAGCTTGAGTTACTTT
PLP	Pd_VANGL1_G	TGCAGTGGGAACAATGTCCTCAATGCTGCTGCTGTACTACCAAAAGTCGGAAGTACTACTCTCTTTTATAGCTTGAGTTACTTG
LNA RT Primer	pLNA_ARL15	CAGATAATCCATGTACAGAAACGCC
PLP	Pd_ARL15_A	ATAACTGAGGCGTTTCTGCCTCAATGCTGCTGCTGTACTACGTTAAGTCGGAAGTACTACTCTCTGGATGTCTGATCTCCGA
PLP	Pd_ARL15_G	ATAACTGAGGCGTTTCTGCCTCAATGCTGCTGCTGTACTACCTTAAGTCGGAAGTACTACTCTCTGGATGTCTGATCTCCGG
LNA RT Primer	pLNA_PLCB3	GTTCTCGAAGGAGAGGATGACGGGG
PLP	Pd_PLCB3_G	CCCTACCCCGTCATCCCTCAATGCTGCTGCTGTACTACGGAAAGTCGGAAGTACTACTCTCTTGCCTTCAAGACCTCG
PLP	Pd_PLCB3_A	CCCTACCCCGTCATCCCTCAATGCTGCTGCTGTACTACCGAAAGTCGGAAGTACTACTCTCTTGCCTTCAAGACCTCA
LNA RT Primer	pLNA_RCL1	GTGTCCGTGGCAGCTTCTGCTTTGT
PLP	Pd_RCL1_A	TGCCTACAGACAAAGCAGCCTCAATGCTGCTGCTGTACTACGTAAAGTCGGAAGTACTACTCTCTCACAAGATAAGGCCCCAA
PLP	Pd_RCL1_G	TGCCTACAGACAAAGCAGCCTCAATGCTGCTGCTGTACTACCTAAAGTCGGAAGTACTACTCTCTCACAAGATAAGGCCCCAG
LNA RT Primer	pLNA_FBXO42	AGCACAGTTTCTTCCTGGTCCACAG
PLP	Pd_FBXO42_A	GACAGTGAAGATGACAGTCCTCAATGCTGCTGCTGTACTACGCAAAGTCGGAAGTACTACTCTCTCATGGCCAGCTCCTCA
PLP	Pd_FBXO42_G	GACAGTGAAGATGACAGTCCTCAATGCTGCTGCTGTACTACCCAAAGTCGGAAGTACTACTCTCTCATGGCCAGCTCCTCG
LNA RT Primer	pLNA_REEP5	TAAAGCTATCCTGGTATTCATATGC

Table 6-2 (Continued).

PLP	Pd_REEP5_G	TAGTATATGGCATATGAATACCCTCAATGCTGCTGCTGTACTACGCTAAGTCGGAAGTACTACTCTCTGGCCTGGTTGTTTCCG
PLP	Pd_REEP5_C	TAGTATATGGCATATGAATACCCTCAATGCTGCTGCTGTACTACCCTAAGTCGGAAGTACTACTCTCTGGCCTGGTTGTTTCCC
LNA RT Primer	pLNA_AKAP12	***A***C***A***C***G***T***T***T***C***T***T***T***G***AGCTTCACCAGC
PLP	Pd_AKAP12_T	GAACCTGCCAAGGAGCCCTCAATGCTGCTGCTGTACTACGAGAAGTCGGAAGTACTACTCTCTCAGAACCTCAGGAAGCT
PLP	Pd_AKAP12_C	GAACCTGCCAAGGAGCCCTCAATGCTGCTGCTGTACTACCAGAAGTCGGAAGTACTACTCTCTCAGAACCTCAGGAAGCC
LNA RT Primer	pLNA_GBAS	***A***T***T***C***A***A***C***G***T***G***G***A***C***CCCTTCTGGAGG
PLP	Pd_GBAS_G	AGCCACTTCTCCCCACCCTCAATGCTGCTGCTGTACTACGTGAAGTCGGAAGTACTACTCTCTCAGTATACCTTATAACACTG
PLP	Pd_GBAS_T	AGCCACTTCTCCCCACCCTCAATGCTGCTGCTGTACTACCTGAAGTCGGAAGTACTACTCTCTCAGTATACCTTATAACACTT
Anchor Primer	AP_ASE	CCUCAAUGCUGCUGCUGUACUAC
Seq Oligo	Cy50A	ANNAAGTCG
Seq Oligo	Cy30G	GNNAAGTCG
Seq Oligo	TR0C	CNNAAGTCG
Seq Oligo	FITC0T	TNNAAGTCG
Seq Oligo	Cy51A	NANAAGTCG
Seq Oligo	Cy31G	NGNAAGTCG
Seq Oligo	TR1C	NCNAAGTCG
Seq Oligo	FITC1T	NTNAAGTCG
Seq Oligo	Cy52A	NNAAAGTCG
Seq Oligo	Cy32G	NNGAAGTCG
Seq Oligo	TR2C	NNCAAGTCG
Seq Oligo	FITC2T	NNTAAGTCG

6.4 **Acknowledgments**

I thank Francois Vigneault for guidance on the initial setup of this project. The Nilsson group has published the PLP detection method[114] and developed a working protocol for *in situ* sequencing by ligation. I thank the Mats Nilsson group, particularly Rongqin Ke and Marco Mignardi, for hosting me and sharing their expertise on the PLP detection and sequencing method, which was instrumental for us to successfully setup the system in our lab. I initiated the efforts to develop *in situ* sequencing, setup a working PLP system within our lab and repeated the Nilsson group's findings on mRNA detection and sequencing. I also designed the ASE experiments with the Nilsson group and validated the initial targets. I thank Frederick Vigneault for advice on the use of Polonator flow cells for sequencing automation. I thank Jay Lee for providing the ASE target sites. I thank John Aach for helpful discussions. I thank George M. Church for guidance and support throughout the project.

CHAPTER 7 Development of fluorescent *in situ* sequencing (FISSEQ)

The work presented in this chapter has been published in the following paper[117]:

- From [Lee JH*, Daugharthy ER*, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SSF, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, Church GM. Highly multiplexed subcellular RNA sequencing in situ. *Science*. 2014;343(6177):1360-1363.] Reprinted with permission from AAAS.

Given the observation that the PLP approach was limited in scalability, we explored alternative strategies to achieve transcriptome-wide RNA profiling and developed the method termed fluorescent *in situ* sequencing (FISSEQ). The work of several lab members over many years was integral to the initial demonstration of FISSEQ. The author contributions will be listed in the acknowledgment section at the end of the chapter. Here we present the progress in FISSEQ development as published in the paper cited above with text and figures modified to fit the format of this dissertation.

7.1 Introduction

Understanding the spatial organization of gene expression with single-nucleotide resolution requires localizing the sequences of expressed RNA transcripts within a cell in situ. Here, we describe fluorescent in situ RNA sequencing (FISSEQ), in which stably cross-linked complementary DNA (cDNA) amplicons are sequenced within a biological sample. Using 30-base reads from 8102 genes in situ, we examined RNA expression and localization in human primary fibroblasts with a simulated wound-healing assay. FISSEQ is compatible with tissue sections and whole-mount embryos and reduces the limitations of optical resolution and noisy signals on single-molecule detection. Our platform enables massively parallel detection of genetic elements, including gene transcripts and molecular barcodes, and can be used to investigate cellular phenotype, gene regulation, and environment in situ.

7.2 Results and Discussion

The spatial organization of gene expression can be observed within a single cell, tissue, and organism, but the existing RNA localization methods are limited to a handful of genes per specimen, making it costly and laborious to localize RNA transcriptome-wide [118–120]. We originally proposed fluorescent in situ sequencing (FISSEQ) in 2003 and subsequently developed methods to sequence DNA amplicons on a solid substrate for genome and transcriptome sequencing [121–124]; however, sequencing the cellular RNA in situ for gene expression profiling requires a spatially structured sequencing library and an imaging method capable of resolving the amplicons.

We report here the next generation of FISSEQ. To generate cDNA amplicons within the cell (Figure 7-1), RNA was reverse-transcribed in fixed cells with tagged random hexamers (Figure 7-2A).

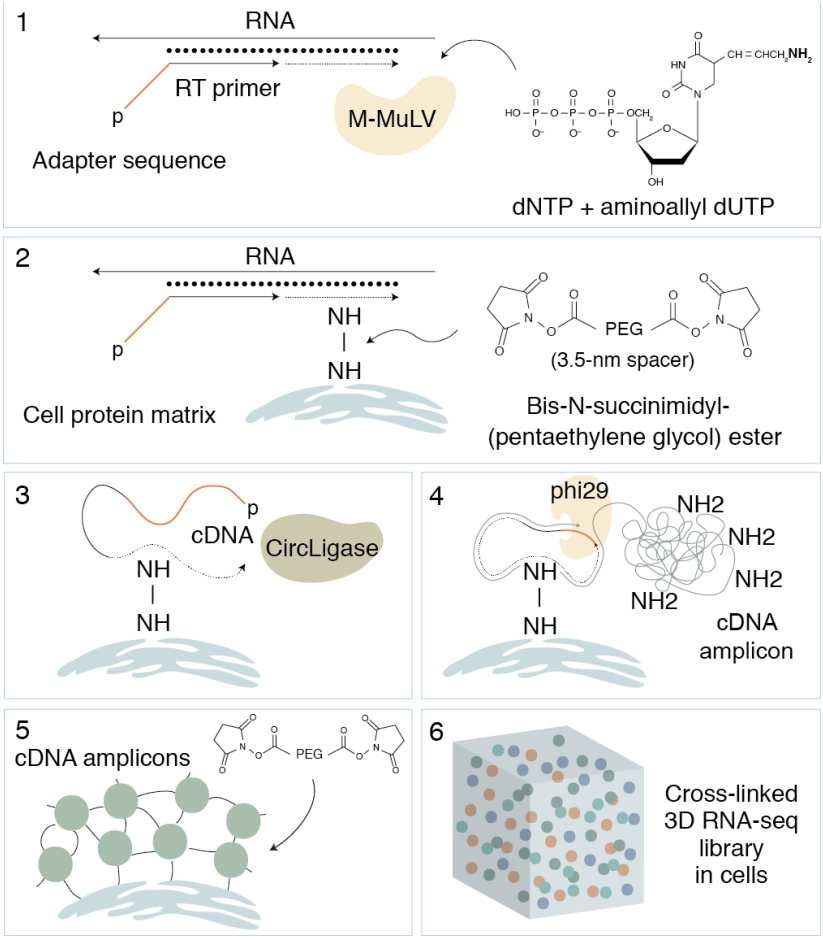

Figure 7-1: Fluorescent *in situ* sequencing (FISSEQ) library construction. (**1**) A tagged random hexamer primer is used to prime M-MuLV reverse transcriptase to generate aminoallyl dUTP- modified cDNA fragments in fixed cells or tissues. (**2**) BS(PEG)9 permanently cross-links the modified cDNA and the cellular protein matrix. (**3**) After cDNA circularization, (**4**) Phi29 DNA polymerase generates cDNA amplicons (**5**) cross-linked to form (**6**) the 3D *in situ* RNA sequencing library within the cell.

From [Lee JH*, Daugharthy ER*, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SSF, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, Church GM. Highly multiplexed subcellular RNA sequencing in situ. *Science*. 2014;343(6177):1360-1363.] Reprinted with permission from AAAS.

We incorporated aminoallyl deoxyuridine 5′-triphosphate (dUTP) during reverse transcription (RT) (Figure 7-2B) and refixed the cells using BS(PEG)9, an amine-reactive linker with a 4-nm spacer. The cDNA fragments were then circularized before rolling circle amplification (RCA) (Figure 7-2C), and BS (PEG)9 was used to cross-link the RCA amplicons containing aminoallyl dUTP (Figure 7-2, D and E).

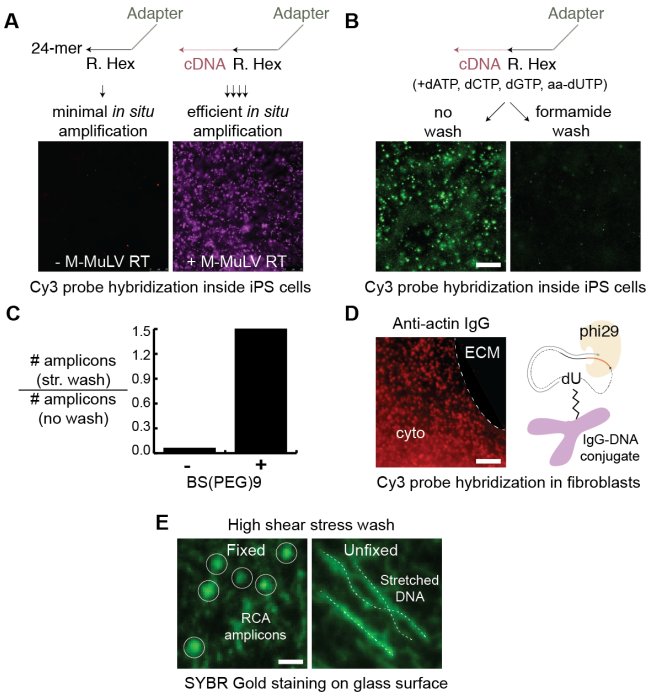

Figure 7-2: Spatial stabilization of the RNA-seq library ***in situ***. (**A**) A short RT primer (24-mer) cannot circularize and amplify ***in situ***, unlike the cDNA (>40-bases). (**B**) Aminoallyl dUTP can be incorporated during RT, providing the reactive group for BS(PEG)9. Without cross-linking, formamide removes much of the cDNA in iPS cells, reducing the number of amplicons (bar: 10 um). (**C**) With cross-linking, a formamide wash does not reduce the number of amplicons in iPS cells. (**D**) A circular template covalently linked to the protein matrix amplifies efficiently ***in situ***. Here an anti-actin antibody is conjugated to the synthetic DNA, purified using ion exchange chromatography, and used to label primary fibroblasts prior to RCA (bar: 10 um). (**E**) RCA amplicons on an aminosilane-treated glass surface are washed aggressively, stretching >95% of unfixed amplicons vs. ~25% of fixed amplicons (bar: 3 um).

From [Lee JH*, Daugharthy ER*, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SSF, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, Church GM. Highly multiplexed subcellular RNA sequencing in situ. *Science*. 2014;343(6177):1360-1363.] Reprinted with permission from AAAS.

We found that random hexamer-primed RT was inefficient (Figure 7-3A), but cDNA circularization was complete within hours (Figure 7-3, B to D).

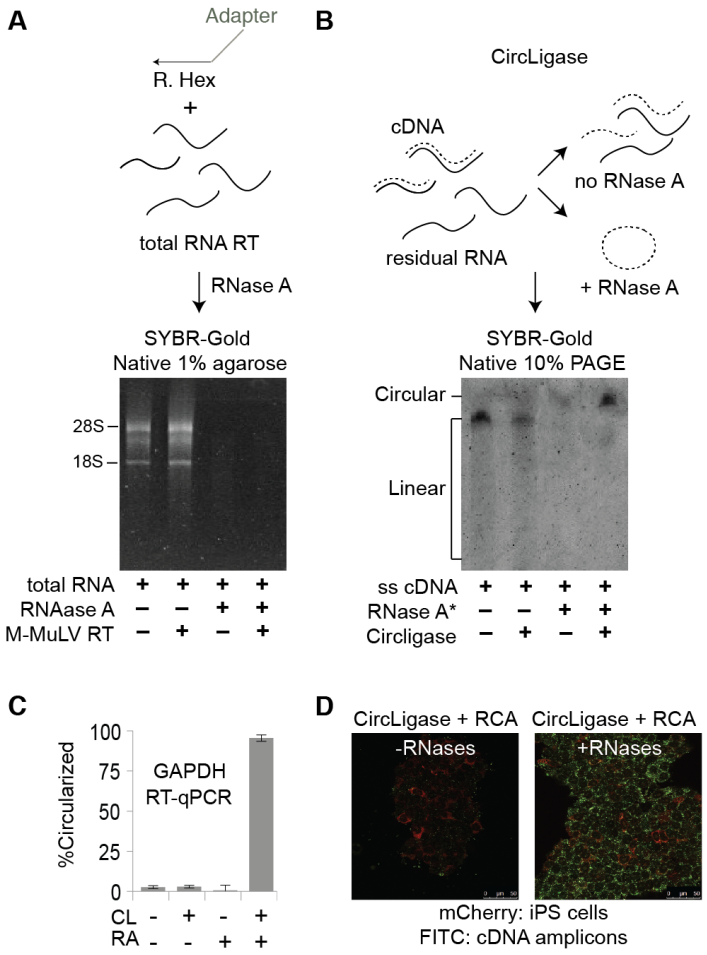

Figure 7-3: Improving the amplicon density *in situ*. (**A**) The cDNA yield is typically low when random hexamers are used to prime reverse transcription, even in solution. (**B**) The resulting single-stranded cDNA fragments do not circularize efficiently (upward mobility shift) unless the residual RNA is degraded using RNase A (* denotes exonuclease-contamination). (**C**) Using real-time qPCR, the amount of circular GAPDH cDNA was estimated in the random hexamer-primed cDNA library with and without 1 hour CircLigase (CL) or RNase A (RA) after Exonuclease I treatment. (**D**) Without RNases, the density of cDNA amplicons *in situ* is variable and typically low in iPS cells.

From [Lee JH*, Daugharthy ER*, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SSF, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, Church GM. Highly multiplexed subcellular RNA sequencing in situ. *Science*. 2014;343(6177):1360-1363.] Reprinted with permission from AAAS.

The result was single-stranded DNA nanoballs 200 to 400 nm in diameter (Figure 7-4A), consisting of numerous tandem repeats of the cDNA sequence. BS(PEG)9 reduced nonspecific probe binding (Figure 7-4B), and amplicons were highly fluorescent after probe hybridization (Figure 7-4C). As a result, the amplicons could be rehybridized many times, with minimal changes in their signal-to-noise ratio or position (Figure 7-4, D and E).

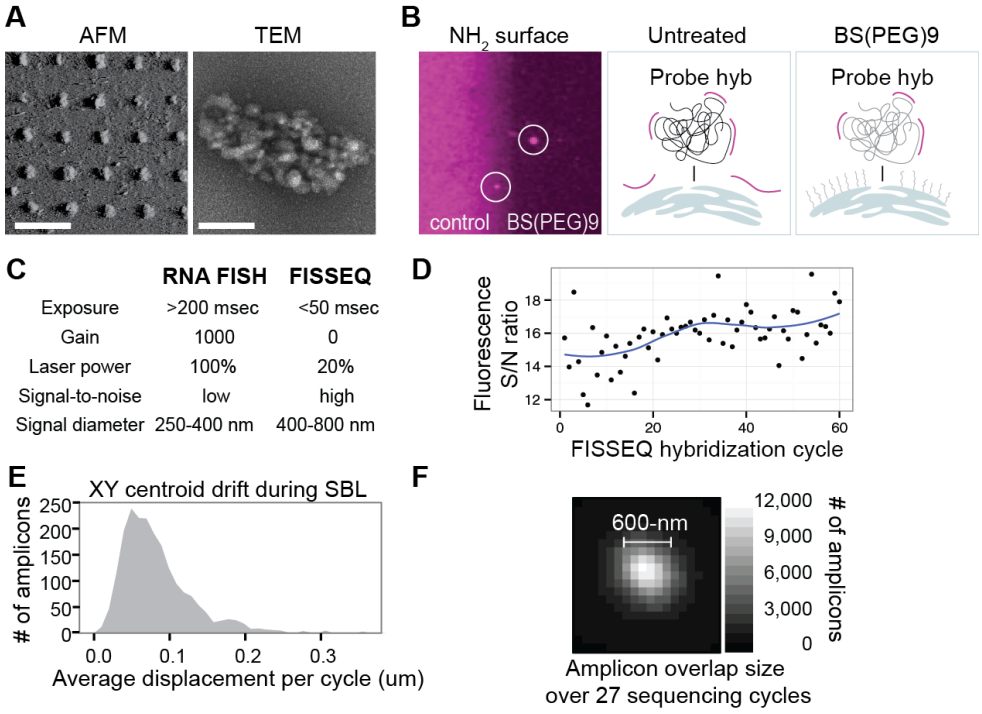

Figure 7-4: Characterization of the cDNA amplicons. (**A**) A standard RCA amplicon generated in solution is a negatively charged DNA molecule containing a large number of the template sequence (AFM bar: 1 um, TEM bar: 100 nm). (**B**) BS(PEG)9 renders the surrounding area less prone to non-specific probe binding. (**C**) Compared to single molecule RNA FISH (20 probes to human GAPDH), FISSEQ amplicons retain their signal-to-ratio even after a prolonged exposure to an excitation laser. (**D**) The amplicons can be re-hybridized 60 or more cycles without a reduction in the signal-to-noise ratio. (**E**) The average centroid drift during sequencing in iPS cells. (**F**) An average amplicon in fibroblasts (overlay of 14,960 centroid-aligned amplicons with 27 nt reads). On average, signals from 27 sequencing cycles overlap by ~36 pixels (600 nm diameter) using a 20x objective (N.A. 0.75).

From [Lee JH*, Daugharthy ER*, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SSF, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, Church GM. Highly multiplexed subcellular RNA sequencing in situ. *Science*. 2014;343(6177):1360-1363.] Reprinted with permission from AAAS.

Using SOLiD sequencing by ligation (Figure 7-5), the signal overlap over 27 consecutive sequencing reactions was ~ 600 nm in diameter (Figure 7-4F).

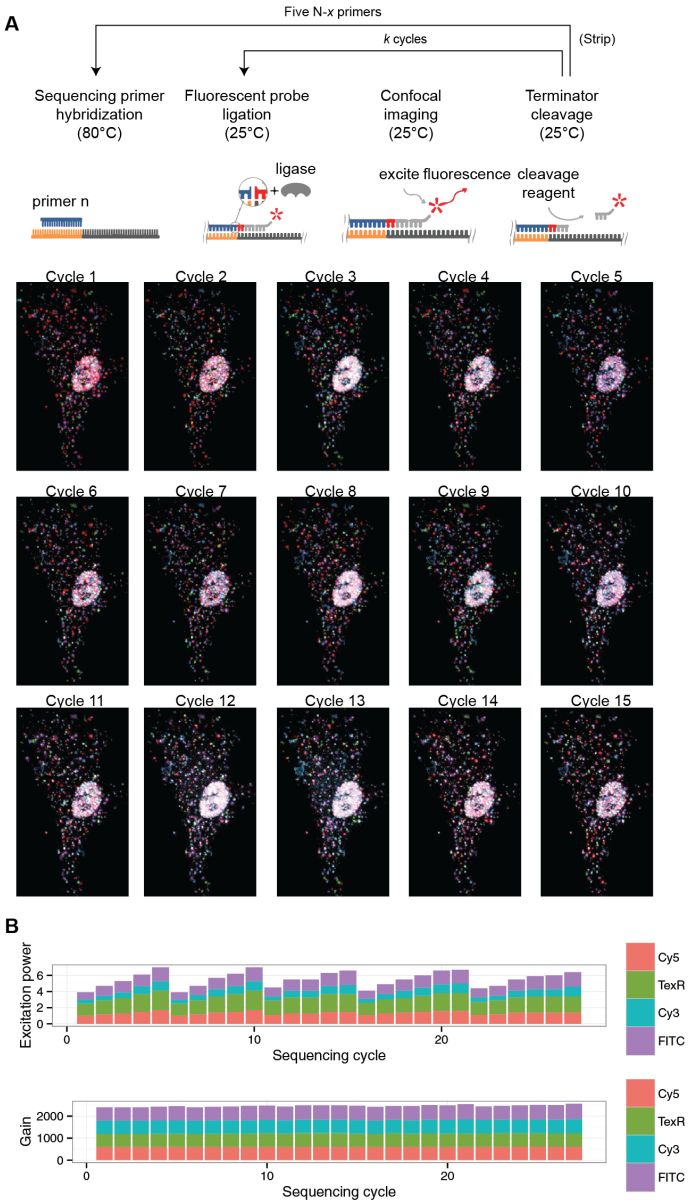

Figure 7-5: Sequencing reaction cycles and imaging. (**A**) Modified SOLiD sequencing-by-ligation for FISSEQ in a primary fibroblast, and deconvolved images for the first 15 cycles. (**B**) Laser excitation power and gain settings over the thirty imaging cycles.

From [Lee JH*, Daugharthy ER*, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SSF, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, Church GM. Highly multiplexed subcellular RNA sequencing in situ. *Science*. 2014;343(6177):1360-1363.] Reprinted with permission from AAAS.

In induced pluripotent stem (iPS) cells, the amplicons counterstained subcellular structures, such as the plasma membrane, the nuclear membrane, the nucleolus, and the chromatin (Figure 7-6A, Figure 7-7). We were able to generate RNA sequencing libraries in different cell types, tissue sections, and whole-mount embryos for three-dimensional (3D) visualization that spanned multiple resolution scales (Figure 7-6, B and C).

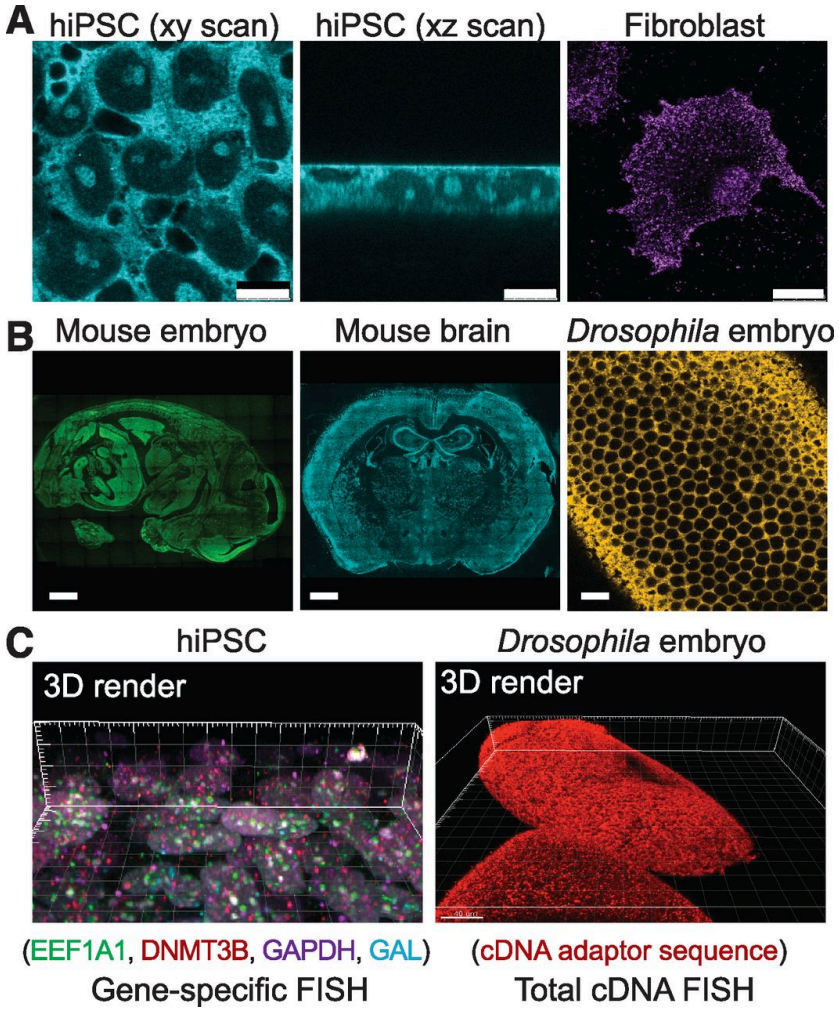

Figure 7-6: Construction of 3D RNA-seq libraries in situ. After RT using random hexamers with an adapter sequence in fixed cells, the cDNA is amplified and cross-linked in situ. (A) A fluorescent probe is hybridized to the adapter sequence and imaged by confocal microscopy in human iPS cells (hiPSC) (scale bar: 10 mm) and fibroblasts (scale bar: 25 mm). (B) FISSEQ can localize the total RNA transcriptome in mouse embryo and adult brain sections (scale bar: 1 mm) and whole-mount Drosophila embryos (scale bar: 5 mm), although we have not sequenced these samples. (C) 3D rendering of gene-specific or adapter-specific probes hybridized to cDNA amplicons. FISH, fluorescence in situ hybridization.

Figure 7-6 (Continued): From [Lee JH*, Daugharthy ER*, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SSF, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, Church GM. Highly multiplexed subcellular RNA sequencing in situ. *Science*. 2014;343(6177):1360-1363.] Reprinted with permission from AAAS.

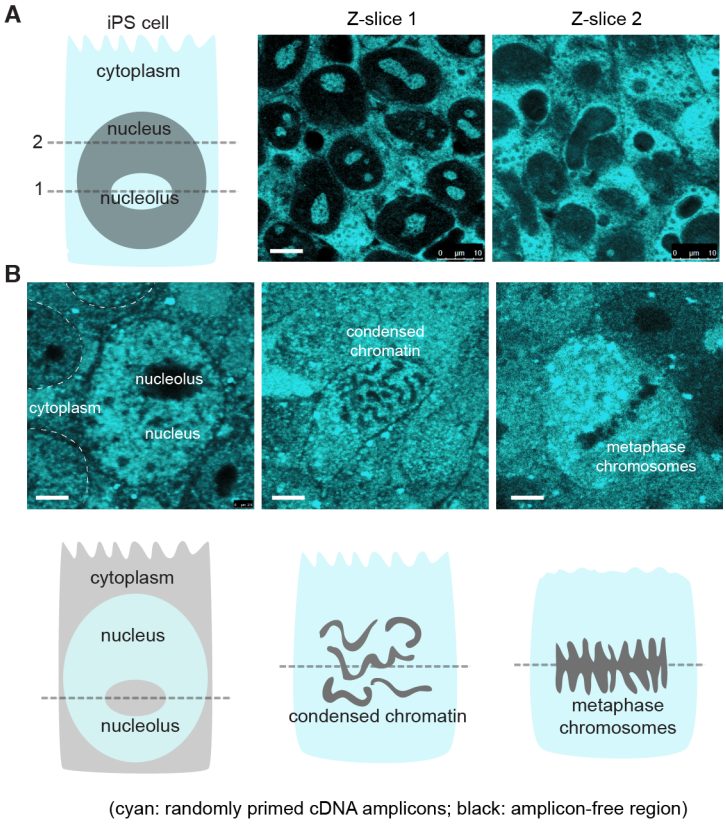

Figure 7-7: The high amplicon density enables visualization of the RNA-rich subcellular compartments in iPS cells. (**A**) In some iPS cell populations, cDNA amplicons are enriched in the cytoplasm and the nucleolus, but not the nucleus. (**B**) In other iPS cell populations, the cDNA amplicons are enriched in the nucleus, counter-staining the nucleolus or condensed chromatin bodies. Amplicon-free speckles are frequently observed in the nucleus and the cytoplasm, potentially indicating regions without RNA or not accessible to FISSEQ.

From [Lee JH*, Daugharthy ER*, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SSF, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, Church GM. Highly multiplexed subcellular RNA sequencing in situ. *Science*. 2014;343(6177):1360-1363.] Reprinted with permission from AAAS.

High numerical aperture and magnification are essential for imaging RNA molecules in single cells [111,112,125], but many gene expression patterns are most efficiently detected in a

low-magnification and wide-field mode, where it typically becomes difficult to distinguish single molecules because of the optical diffraction limit and low sensitivity [126]. To obtain a spot density that is high enough to yield statistically significant RNA localization, and yet sufficiently low for discerning individual molecules, we developed partition sequencing, in which pre-extended sequencing primers are used to reduce the number of molecular sequencing reactions through random mismatches at the ligation site (Figure 7-8A). Progressively longer sequencing primers result in exponential reduction of the observed density, and the sequencing primer can be changed during imaging to detect amplicon pools of different density.

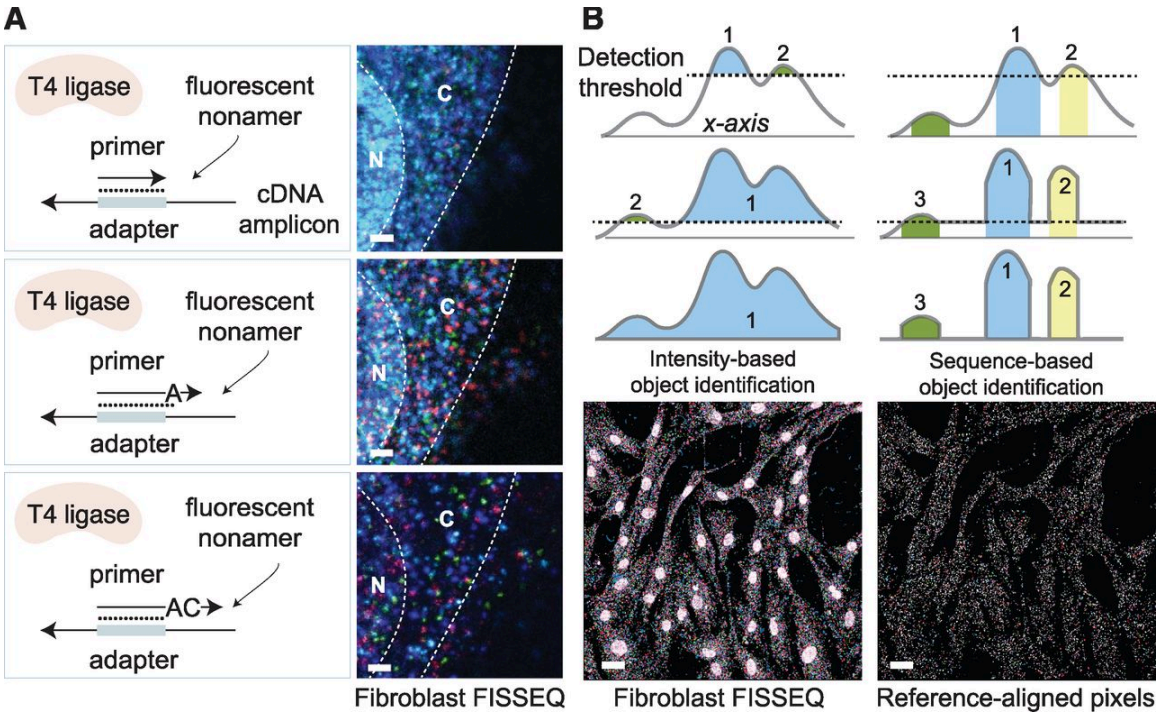

Figure 7-8: Overcoming resolution limitations and enhancing the signal-to-noise ratio. (A) Ligation of fluorescent oligonucleotides occurs when the sequencing primer ends are perfectly complementary to the template. Extending sequencing primers by one or more bases, one can randomly sample amplicons at 1/4th, 1/16th, and 1/256th of the original density in fibroblasts (scale bar: 5 mm). N, nucleus; C, cytoplasm. (B) Rather than using an arbitrary intensity threshold, color sequences at each pixel are used to identify objects. For sequences of L bases, the error rate is approximately $n/4^L$ per pixel, where n is the size of

Figure 7-8 (Continued): the reference. By removing unaligned pixels, the nuclear background noise is reduced in fibroblasts (scale bar: 20 mm).

From [Lee JH*, Daugharthy ER*, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SSF, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, Church GM. Highly multiplexed subcellular RNA sequencing in situ. *Science*. 2014;343(6177):1360-1363.] Reprinted with permission from AAAS.

Fluorescence microscopy can be accompanied by tissue-specific artifacts and autofluorescence, which impede accurate identification of objects. If objects are nucleic acids, however, discrete sequences, rather than the analog signal intensity, can be used to analyze the image. For FISSEQ, putative nucleic acid sequences are determined for all pixels. The sequencing reads are then compared with reference sequences, and a null value is assigned to unaligned pixels. With a suitably long read length (L), a large number of unique sequences (n) can be used to identify transcripts or any other objects with a false-positive rate of approximately $n/4^L$ per pixel. Because the intensity threshold is not used, even faint objects are registered on the basis of their sequence, whereas background noise, autofluorescence, and debris are eliminated (Figure 7-8B).

We applied these concepts to sequence the transcription start site of inducible mCherry mRNA in situ, analogous to 5′ rapid amplification of cDNA ends–polymerase chain reaction (RACE-PCR) [127]. After RT and molecular amplification of the 5′ end followed by fluorescent probe hybridization (Figure 7-9A), we quantified the concentration- and time-dependent mCherry gene expression in situ (Figure 7-9B). Using sequencing-by-ligation, we then determined the identity of 15 contiguous bases from each amplicon in situ, corresponding to the transcription start site (Figure 7-9C). When the sequencing reads were mapped to the vector sequence, 7472 (98.7%) amplicons aligned to the positive strand of

mCherry, and 3967 (52.4%) amplicons mapped within two bases of the predicted transcription start site (Figure 7-9D).

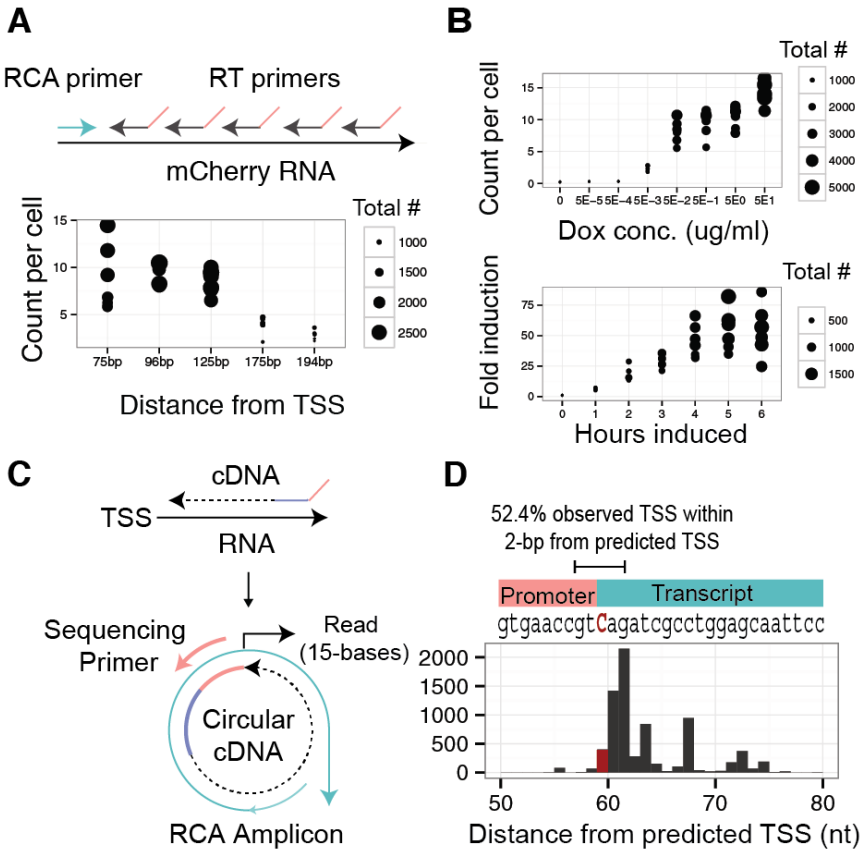

Figure 7-9: Single gene capture and sequencing *in situ* in HeLa cells. (**A**) The doxycycline-inducible mCherry transcripts are reverse transcribed using a series of RT primers along the 5' region in order to assess the effect of cDNA length on the sensitivity. (**B**) Upon doxycycline treatment, the mCherry cDNA amplicons are detected in a concentration- and time-dependent manner (each point is a replicate experiment of ~300 cells). (**C**) The 3' end of the cDNA corresponds to the transcription start site (TSS), which abuts the 5' end of the adapter sequence when circularized. (**D**) When the amplicons are sequenced *in situ*, 3,967 out of 7,492 (52.4%) amplicons map to the four base window spanning the predicted TSS.

From [Lee JH*, Daugharthy ER*, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SSF, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, Church GM. Highly multiplexed subcellular RNA sequencing in situ. *Science*. 2014;343(6177):1360-1363.] Reprinted with permission from AAAS.

We then sequenced the transcriptome in human primary fibroblasts in situ (Figure 7-10A) and generated sequencing reads of 27 bases with a median per-base error rate of 0.64% (Figure 7-11).

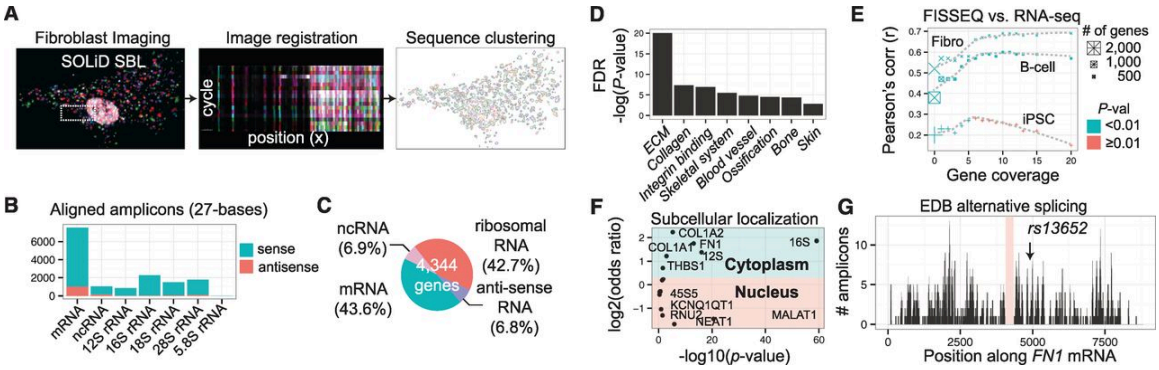

Figure 7-10: Whole-transcriptome in situ RNA-seq in primary fibroblasts. (A) From deconvolved confocal images, 27-base reads are aligned to the reference, and alignments are spatially clustered into objects. (B) Of the amplicons, 90.6% align to the annotated (+) strand. (C) mRNA and non- coding RNA make up 43.6% and 6.9% of the amplicons, respectively. (D) GO term clustering for the top 90 ranked genes. (E) FISSEQ of 2710 genes from fibroblasts compared with RNA-seq for fibroblast, B cell, and iPS cells. Pearson's correlation is plotted as a function of the gene expression level. (F) Subcellular localization enrichment compared to the whole transcriptome distribution. (G) Of the amplicons, 481 map to the FN1 mRNA, showing an alternatively spliced transcript variant and a single-nucleotide polymorphism (arrow).

From [Lee JH*, Daugharthy ER*, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SSF, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, Church GM. Highly multiplexed subcellular RNA sequencing in situ. *Science*. 2014;343(6177):1360-1363.] Reprinted with permission from AAAS.

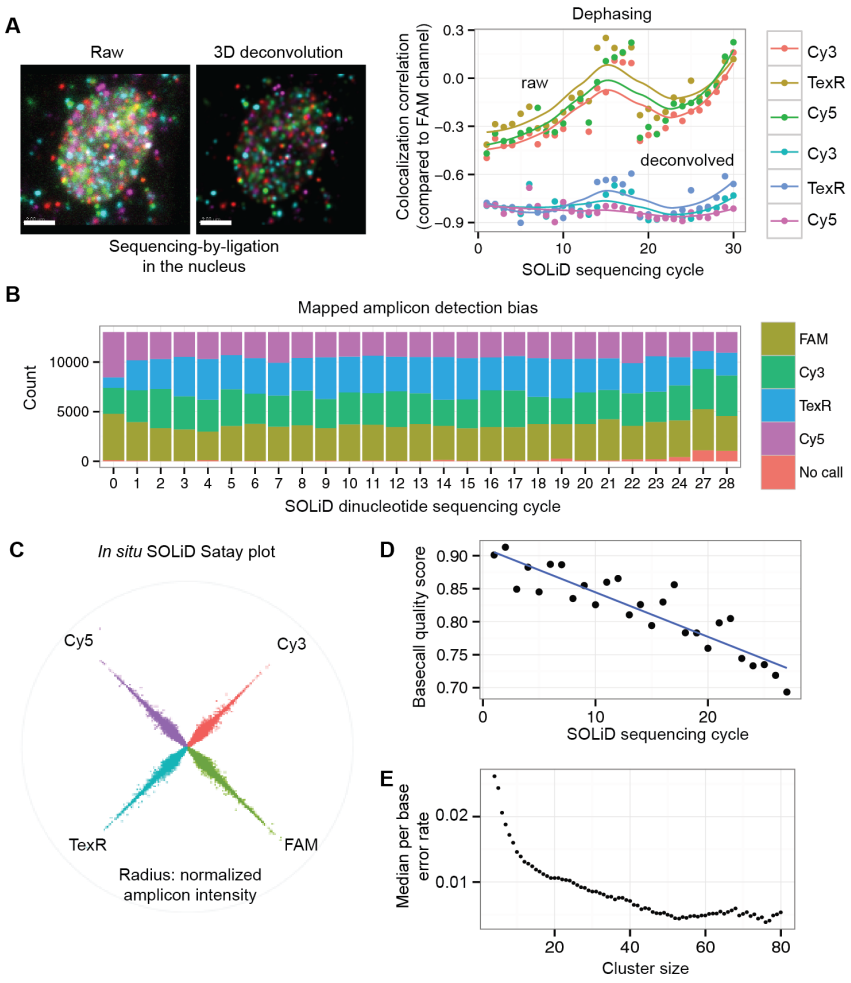

Figure 7-11: Imaging and base calling statistics. (**A**) 3D deconvolution enhances the signal-to-noise ratio and decreases dephasing of the base-specific fluorescence over multiple sequencing cycles in a fibroblast (bar: 5 um). The high background represents the nucleus. (**B**) The number of amplicons associated with specific fluorescence is plotted as a function of sequencing cycles. (**C**) A Satay plot of amplicon fluorescence from all 27 sequencing cycles. The radius is the intensity of the signal, and the angle is the separation between the signal and the axis of the called base. (**D**) The basecall quality over 27 sequencing cycles. (**E**) The median per-base error rate from the whole transcriptome FISSEQ as a function of cluster size.

From [Lee JH*, Daugharthy ER*, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SSF, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, Church GM. Highly multiplexed subcellular RNA sequencing in situ. *Science*. 2014;343(6177):1360-1363.] Reprinted with permission from AAAS.

Using an automated analysis pipeline (Figure 7-12), we identified 14,960 amplicons with size >5 pixels, representing 4171 genes, of which 13,558 (90.6%) amplicons mapped to the correct annotated strand (Figure 7-10B, Figure 7-13, and Table 7-1).

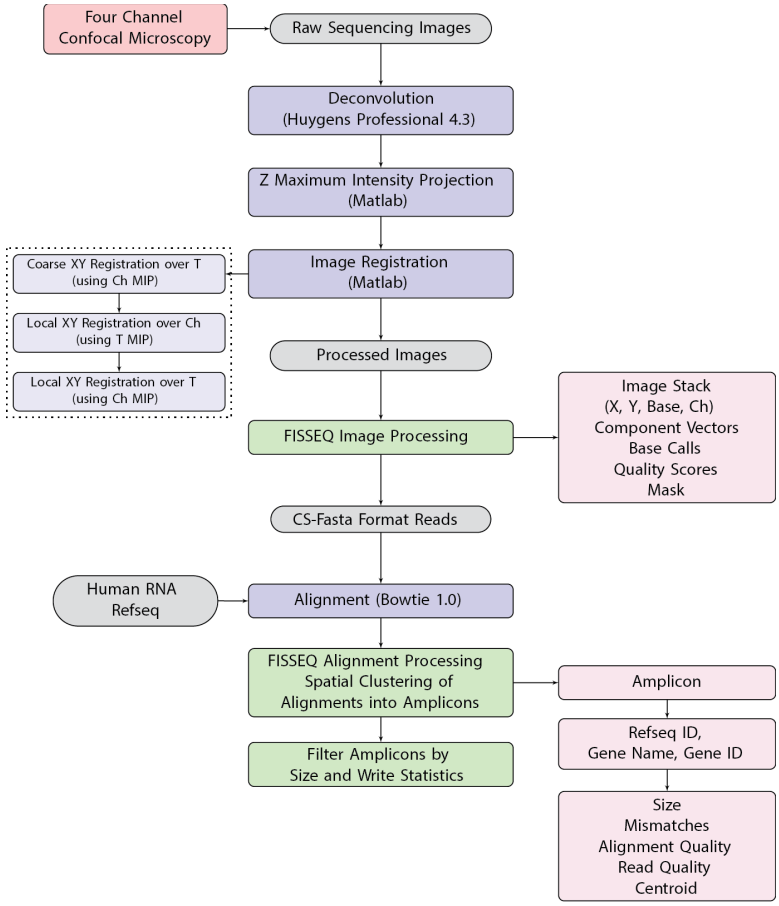

Figure 7-12: FISSEQ image and data analysis pipeline. Confocal image stacks deconvoluted to reduce dephasing of base calls. Depending on the sample thickness and the amplicon density, the number of z-slices are reduced to minimize the processing time. Images were corrected for chromatic shifts and registered to one another using a block-based algorithm prior to applying sequenced-based quality filters. The remaining pixels are spatially clustered to identify specific sequence-associated amplicons of size >5 pixels.

From [Lee JH*, Daugharthy ER*, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SSF, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, Church GM. Highly multiplexed subcellular RNA sequencing in situ. *Science*. 2014;343(6177):1360-1363.] Reprinted with permission from AAAS.

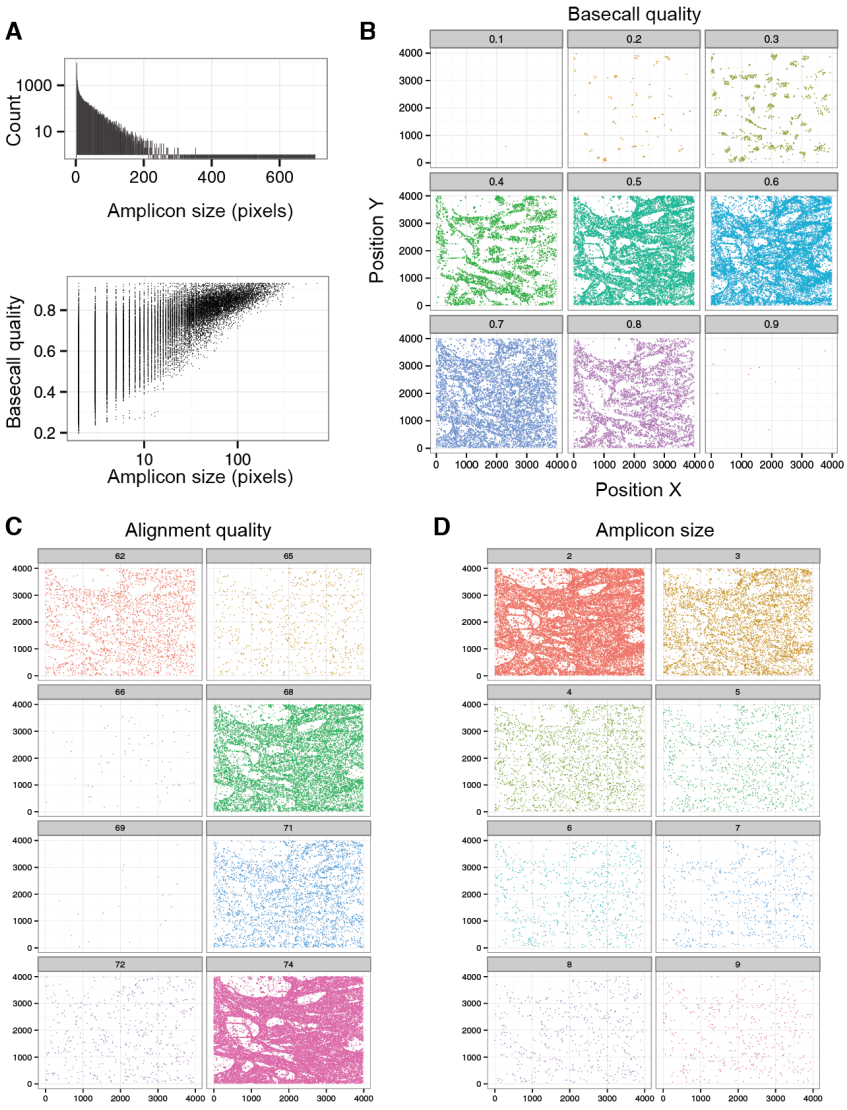

Figure 7-13: Basecall and alignment quality in primary fibroblasts. (**A**) The distribution of the base quality as a function of the amplicon pixel cluster size. (**B**) The centroid of aligned amplicons is plotted as a function of the basecall quality. (**C**) The alignment quality and (**D**) the amplicon size (or the number of pixels per amplicon) are relatively uniform throughout the whole image and within the cell.

From [Lee JH*, Daugharthy ER*, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SSF, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, Church GM. Highly multiplexed subcellular RNA sequencing in situ. *Science*. 2014;343(6177):1360-1363.] Reprinted with permission from AAAS.

Table 7-1. FISSEQ summary statistics from human primary fibroblasts in FBS media.

From [Lee JH*, Daugharthy ER*, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SSF, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, Church GM. Highly multiplexed subcellular RNA sequencing in situ. *Science*. 2014;343(6177):1360-1363.] Reprinted with permission from AAAS.

Total bases from aligned amplicons (>5 pixels)	**420,714 nt**
Amplicons with >5 pixels	**15,582 (100%)**
Belongs to an annotated RNA class	**15,126 (97.1%)**
Maps to a single gene	**14,960 (96.0%)**
Unique gene names	**4,171**
Mean per-base error %	**1.74%**
Median per-base error %	**0.64%**

We found that mRNA (43.6%) was relatively abundant even though random hexamers were used for RT (Figure 7-10C). Ninety genes with the highest expression counts included fibroblast markers [128], such as fibronectin (FN1); collagens (COL1A1, COL1A2, COL3A1); matrix metallopeptidases and inhibitors (MMP14, MMP2, TIMP1); osteonectin (SPARC); stanniocalcin (STC1); and the bone morphogenesis–associated transforming growth factor (TGF)–induced protein (TGFBI), representing extracellular matrix, bone development, and skin development [Benjamini-Hochberg false discovery rate (FDR) $<10^{-19}$, 10^{-5}, and 10^{-3}, respectively] (Figure 7-10D) [129]. We made Illumina sequencing libraries to compare FISSEQ to RNA-seq. Pearson's r correlation coefficient between RNA-seq and FISSEQ ranged from 0.52 to 0.69 ($P < 10^{-16}$), excluding one outlier (FN1). For 854 genes with more than one observation, Pearson's r was 0.57 ($P < 10^{-16}$), 0.47 ($P < 10^{-16}$), and 0.23 ($P < 10^{-3}$) between

FISSEQ and RNA-seq from fibroblasts, lymphocytes, and iPS cells, respectively (Figure 7-10E). When FISSEQ was compared with gene expression arrays, Pearson's r was as high as 0.73 ($P < 10^{-16}$) among moderately expressed genes, whereas genes with low or high expression levels correlated poorly ($r < 0.4$) (Figure 7-14).

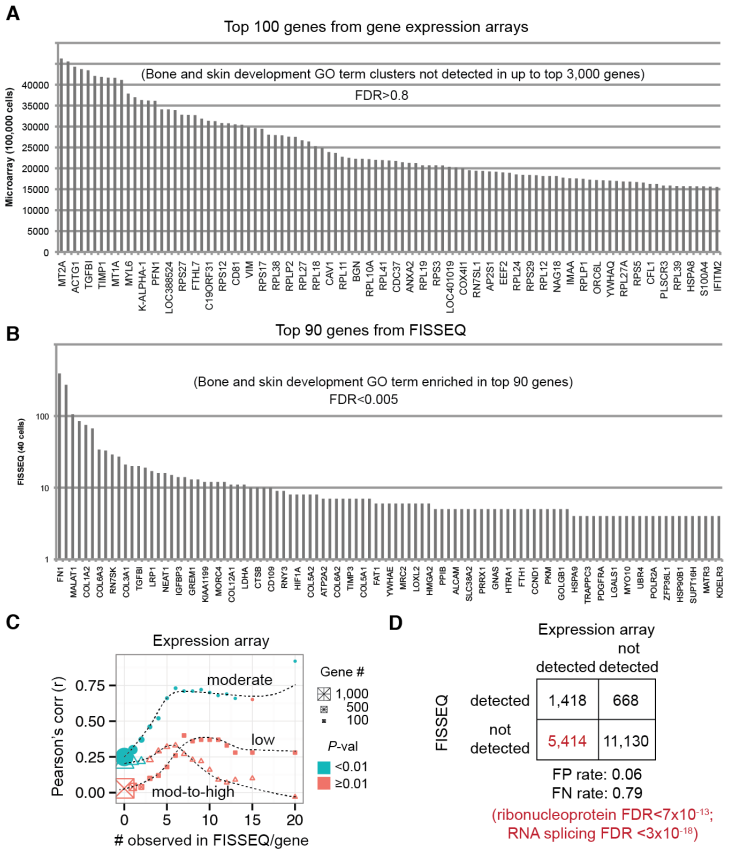

Figure 7-14: Comparison of the gene expression data from expression arrays and FISSEQ. (**A**) The top 100 ranked genes from microarrays are shown. (**B**) The top 90 ranked genes from FISSEQ are enriched in genes known to be markers of the fibroblast activity. (**C**) The correlation between the array and FISSEQ dataset among genes with different expression array levels. Pearson's correlation is plotted as a function of the expression array level. The number of genes used for correlation is indicated on the right. (**D**) Compared to the expression array, FISSEQ has a false positive rate of 6% and detects fewer genes related to RNA processing and splicing (FDR<10^{-13}).

From [Lee JH*, Daugharthy ER*, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SSF, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, Church GM. Highly multiplexed subcellular RNA sequencing in situ. *Science*. 2014;343(6177):1360-1363.] Reprinted with permission from AAAS.

Highly abundant genes in RNA-seq and gene expression arrays were involved in translation and splicing (Figure 7-14 and Figure 7-15), whereas such genes were underrepresented in FISSEQ. We examined 12,427 (83.1%) and 2533 (16.9%) amplicons in the cytoplasm and nuclei, respectively, and found that nuclear RNA was 2.1 [95% confidence interval (CI) 1.9 to 2.3] times more likely to be non-coding ($P < 10^{-16}$), and antisense mRNA was 1.8 [95% CI 1.7 to 2.0] times more likely to be nuclear ($P < 10^{-16}$). We confirmed nuclear enrichment of MALAT1 and NEAT1 by comparing their relative distribution against all RNAs (Figure 7-10F) or mitochondrial 16S ribosomal RNA (rRNA) (Table 7-2), whereas mRNA, such as COL1A1, COL1A2, and THBS1, localized to the cytoplasm (Table 7-3). We also examined splicing junctions of FN1, given its high read coverage (481 reads over 8.9 kilobases).

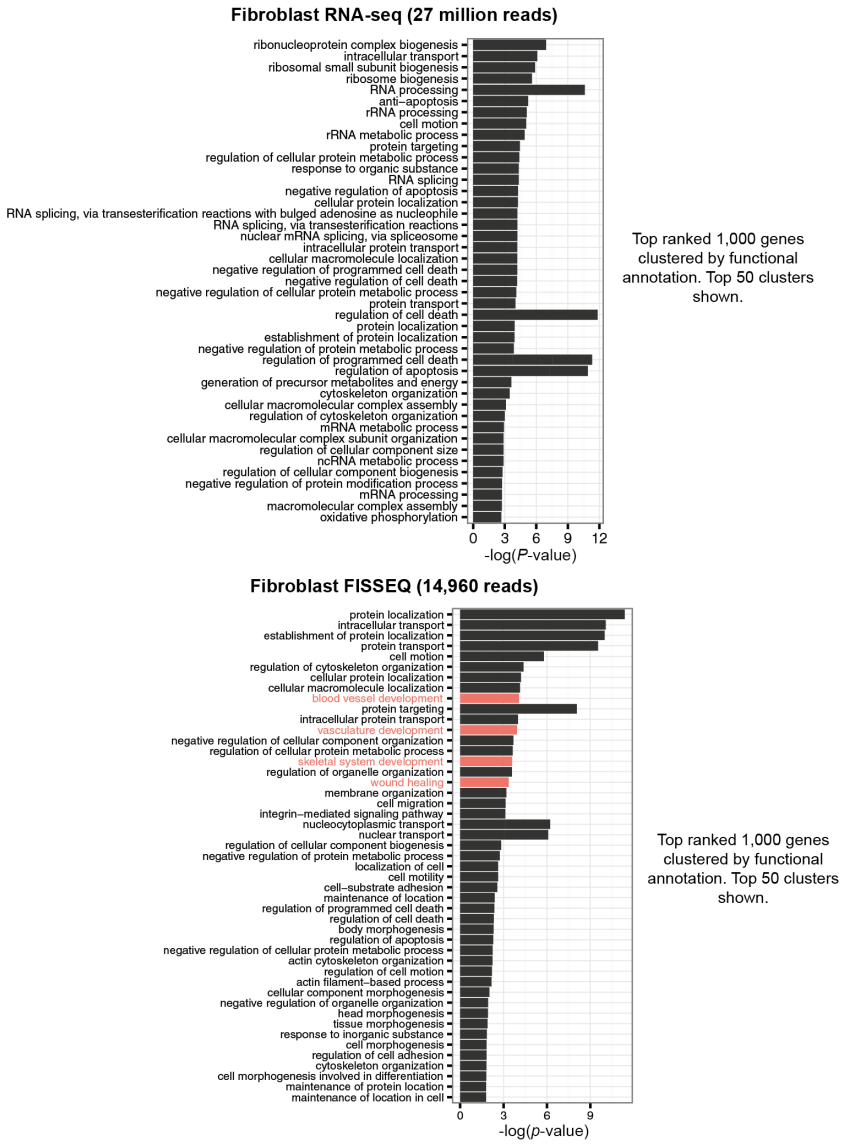

Figure 7-15: Comparison of the functional term enrichment between RNA-seq and FISSEQ. Despite a much deeper coverage and read depth, the top 1,000 ranked genes from RNA-seq do not form cell type-specific functional clusters. The top 90 ranked genes from FISSEQ are related to the bone and skin development (Figure 7-10), while the top 1,000 ranked genes from FISSEQ are associated with wound healing and skeletal development.

From [Lee JH*, Daugharthy ER*, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SSF, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, Church GM. Highly multiplexed subcellular RNA sequencing in situ. *Science*. 2014;343(6177):1360-1363.] Reprinted with permission from AAAS.

Table 7-2. **The RNA localization likelihood compared to 16S mitochondrial rRNA.** Amplicons, 12,427 (83.1%) cytoplasmic and 2,533 (16.9%) nuclear, are compared against 16S rRNA localization. 164 genes with more than five observations are chosen for Fisher's exact test. All genes were more likely to be found in the nucleus compared to 16S mitochondrial rRNA, given their nuclear origin (odds ratio < 1). But non-coding RNA (*MALAT1*, *NEAT*, *KCNQ1OT1*), small nuclear RNA (*RN7SK*, *RNU2-1*), and pre-ribosomal RNA (*RNA45S5*) were notably more enriched in the nucleus.

From [Lee JH*, Daugharthy ER*, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SSF, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, Church GM. Highly multiplexed subcellular RNA sequencing in situ. *Science*. 2014;343(6177):1360-1363.] Reprinted with permission from AAAS.

Gene	Cytoplasmic	Nuclear	*P*-value	OR	95% CImin	95% CImax
MALAT1	295	164	1.2E-64	0.10	0.08	0.12
RIBO28S	1774	319	5.7E-29	0.31	0.25	0.37
RIBO18S	1493	258	9.0E-25	0.32	0.26	0.39
NEAT1	51	33	3.1E-19	0.09	0.05	0.12
RNA45S5	193	48	2.1E-13	0.23	0.15	0.30
RN7SK	75	19	1.0E-06	0.22	0.13	0.31
RNU2-1	18	9	8.3E-06	0.11	0.05	0.18
KCNQ1OT1	19	8	6.8E-05	0.13	0.05	0.21
GLS	7	5	2.7E-04	0.08	0.02	0.14
MYO10	5	4	8.5E-04	0.07	0.01	0.13
THBS1	186	23	1.7E-03	0.45	0.28	0.62
FNDC1	3	3	2.7E-03	0.06	0.01	0.10
MAP3K4	3	3	2.7E-03	0.06	0.01	0.10
VIM	22	6	3.4E-03	0.21	0.08	0.33
SIPA1L1	4	3	4.6E-03	0.07	0.01	0.14
VPS13C	4	3	4.6E-03	0.07	0.01	0.14
HNRNPA2B1	5	3	7.1E-03	0.09	0.02	0.17
SNORD50A	5	3	7.1E-03	0.09	0.02	0.17
DST	12	4	9.1E-03	0.17	0.05	0.29
CHD1	7	3	1.4E-02	0.13	0.03	0.23
IGFBP4	7	3	1.4E-02	0.13	0.03	0.23

Table 7-3. **The RNA localization likelihood compared to MALAT1, a non-coding RNA known to localize to the nuclear speckles.** Amplicons, 12,427 (83.1%) cytoplasmic and 2,533 (16.9%) nuclear, are compared against MALAT1 localization. Also, 164 genes with more than five observations are chosen for Fisher's exact test. Most genes were more likely to be found in the cytoplasm compared to MALAT1 (odds ratio > 1).

From [Lee JH*, Daugharthy ER*, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SSF, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, Church GM. Highly multiplexed subcellular RNA sequencing in situ. *Science*. 2014;343(6177):1360-1363.] Reprinted with permission from AAAS.

Gene	Cytoplasmic	Nuclear	*P*-value	OR	95% Cimin	95% Cimax
RIBO16S	2271	127	1.2E-64	9.93	7.58	12.28
RIBO12S	844	66	3.5E-38	7.10	5.14	9.06
FN1	481	29	1.9E-33	9.20	5.99	12.41
RIBO18S	1493	258	6.3E-22	3.22	2.53	3.90
RIBO28S	1774	319	2.1E-21	3.09	2.45	3.73
COL1A2	138	6	3.5E-16	12.75	5.53	19.97
COL1A1	160	14	1.4E-13	6.34	3.52	9.15
THBS1	186	23	5.4E-12	4.49	2.77	6.21
COL6A3	50	4	5.6E-06	6.93	2.48	11.38
COL6A1	29	0	6.4E-06	Inf	4.02	Inf
RNA45S5	193	48	1.4E-05	2.23	1.53	2.94
TGFBI	42	3	2.0E-05	7.76	2.42	13.11
SPARC	30	1	4.6E-05	16.63	2.71	30.54
ITGB1	38	3	9.2E-05	7.02	2.18	11.87
COL12A1	26	1	2.3E-04	14.41	2.32	26.50
COL3A1	28	2	5.4E-04	7.76	1.92	13.61
RNY3	18	0	5.9E-04	Inf	2.40	Inf
FBN1	35	4	7.0E-04	4.85	1.69	8.01
EEF1A1	30	3	1.0E-03	5.55	1.68	9.41

FN1 has three variable domains referred to as EDA, EDB, and IIICS, which are alternatively spliced [130]. We did not observe development-associated EDB, but observed adult tissue-associated EDA and IIICS (Figure 7-10G).

We also sequenced primary fibroblasts in situ after simulating a response to injury, obtaining 156,762 reads (>5 pixels), representing 8102 annotated genes (Figure 7-16A and Figure 7-17, A to D). Pearson's r was 0.99 and 0.91 between different wound sites and growth conditions, respectively (Figure 7-16B and Figure 7-17, E and F).

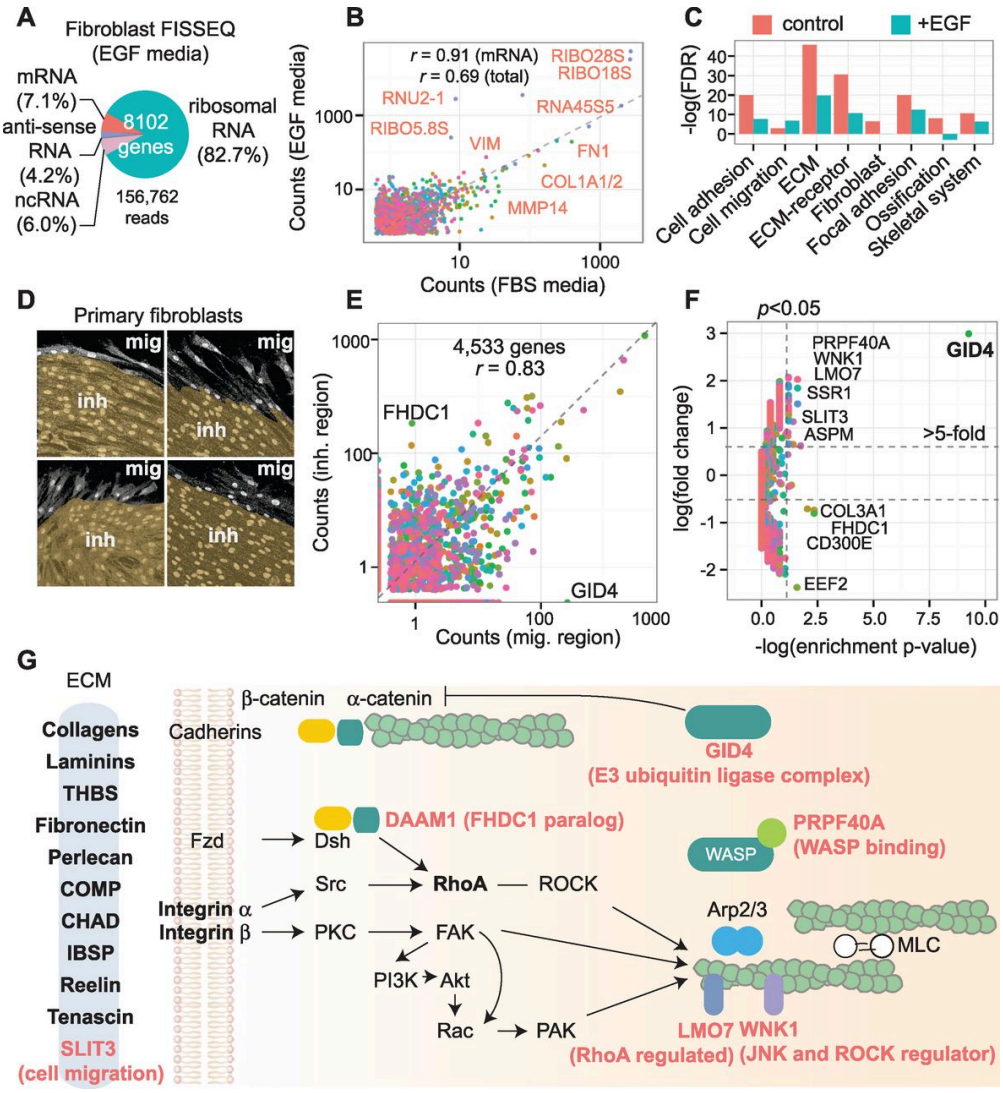

Figure 7-16: Functional analysis of fibroblasts during simulated wound healing. (A) In EGF medium, rRNA makes up 82.7% of the amplicons. (B) EGF medium 147,610 reads compared with 13,045 reads from FBS medium (different colors denote genes). (C) The top 100 ranked genes from FBS versus EGF FISSEQ clustered for functional annotation. (D) An in vitro wound-healing assay allows cells to migrate (mig) into the wound gap. inh, contact-inhibited cells. The image segments are based on the cell morphology. (E) Comparison of 4533 genes from migrating and contact-inhibited cells. (F) Twelve genes are differentially expressed (Fisher's exact test $P < 0.05$ and >fivefold; 180 genes). (See Table 7-4.) (G) The top 100 genes in fibroblasts are enriched for terms associated with ECM-receptor interaction and focal adhesion kinase complex (bold letters). During cell migration, genes involved in ECM-receptor-cytoskeleton signaling and remodeling are differentially expressed (red letters). THBS, thrombospodin; COMP, cartilage oligomeric matrix protein; CHAD, chondroadherin; IBSP, integrin-binding sialoprotein; PKC, protein kinase C; FAK, focal adhesion kinase; PI3K, phos- phatidylinositol 3-kinase; MLC, myosin light chain; PAK, p21-activated protein kinase; WASP, Wiskott- Aldrich syndrome protein.

Figure 7-16 (Continued): From [Lee JH*, Daugharthy ER*, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SSF, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, Church GM. Highly multiplexed subcellular RNA sequencing in situ. *Science*. 2014;343(6177):1360-1363.] Reprinted with permission from AAAS.

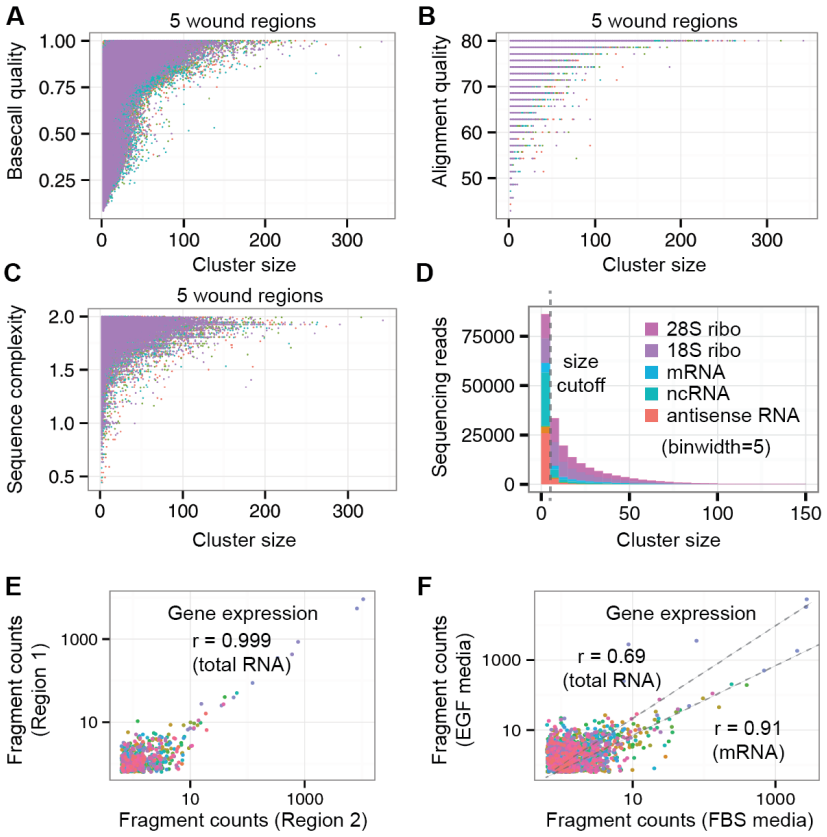

Figure 7-17: Wound healing FISSEQ across five different regions. (**A**) Basecall quality as a function of the amplicon size (in pixels). (**B**) Alignment quality as a function of the amplicon size. (**C**) Sequence complexity as a function of the amplicon size. (**D**) Gene categories as a function of the amplicon size. (**E**) Gene expression comparison between wound regions (n=658 genes). Pearson's r is >0.999 for all genes, 0.91 for genes with 10-100 counts, and 0.41 for genes with 0-10 counts. (**F**) Gene expression comparison between slow and fast growing fibroblasts in separate experiments (n=2,309 all genes; n=1,621 mRNA).

From [Lee JH*, Daugharthy ER*, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SSF, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, Church GM. Highly multiplexed subcellular RNA sequencing in situ. *Science*. 2014;343(6177):1360-1363.] Reprinted with permission from AAAS.

In medium with epidermal growth factor (EGF), 82.7% of the amplicons were rRNA compared to 42.7% in fetal bovine serum (FBS) medium. When the 100 highest ranked genes were clustered, cells in FBS medium were enriched for fibroblast-associated GO terms, whereas rapidly dividing cells in EGF medium were less fibroblast-like (Figure 7-16C) with alternative splicing of FN1 (Figure 7-18).

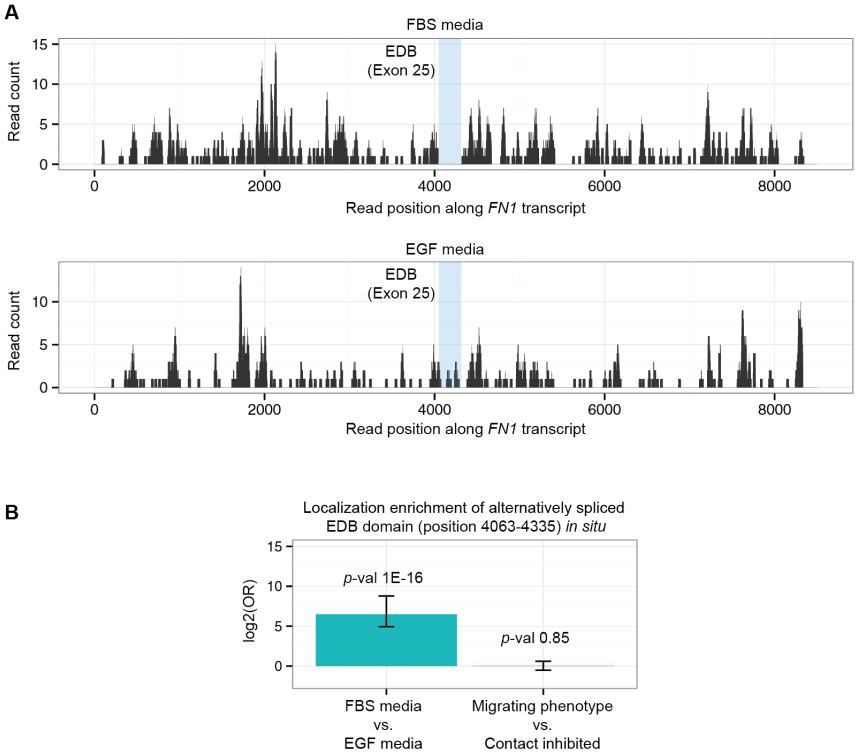

Figure 7-18: Analysis of alternative splicing of *FN1 in situ*. (**A**) Approximately 500 *FN1* reads from the cells grown in FBS media and EGF media are compared. Reads aligned to Exon 25 (EDB) are enriched in fibroblasts from EGF media. (**B**) Retention of Exon 25 (EDB) is 90-fold higher in EGF media, compared to slower growing fibroblasts in FBS media; however, no statistical enrichment of EDB retention is seen between migrating and contact inhibited fibroblasts. Error bars indicate a 95% confidence interval of the estimated odds ratio (Fisher's exact test).

From [Lee JH*, Daugharthy ER*, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SSF, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, Church GM. Highly multiplexed subcellular RNA sequencing in situ. *Science*. 2014;343(6177):1360-1363.] Reprinted with permission from AAAS.

In regions containing migrating cells versus contact-inhibited cells, 12 genes showed differences in relative gene expression (Fisher's exact test $P < 0.05$ and >fivefold change) (Figure 7-16, D to F, and Table 7-4), eight of which were associated with the extracellular matrix (ECM)–receptor–cytoskeleton interaction, including GID4, FHDC1, PRPF40A, LMO7, and WNK1 (Figure 7-16G and Table 7-4).

Table 7-4. **The likelihood table of differentially expressed genes (180 genes with >5 observations) reveals biological pathway enrichment in migrating vs. stationary fibroblasts.** For the purpose of generating fold-change plots (Figure 7-16F), a small positive value (0.01) was added to normalized mRNA counts. Fisher's exact test used 28S rRNA counts in migrating cells vs. contact inhibited cells for comparison.

From [Lee JH*, Daugharthy ER*, Scheiman J, Kalhor R, Yang JL, Ferrante TC, Terry R, Jeanty SSF, Li C, Amamoto R, Peters DT, Turczyk BM, Marblestone AH, Inverso SA, Bernard A, Mali P, Rios X, Aach J, Church GM. Highly multiplexed subcellular RNA sequencing in situ. *Science*. 2014;343(6177):1360-1363.] Reprinted with permission from AAAS.

Gene	Migrating	Stationary	*P*-value	Pubmed ID
GID4	35	0	5.41E-10	PMID17467196, PMID18215130
CD300E	1	15	4.48E-03	PMC3009906
FHDC1	1	15	4.48E-03	PMID11779461
COL3A1	7	26	8.92E-03	
SSR1	12	2	1.79E-02	PMID22314232
LMO7	6	0	2.44E-02	PMC3147800
PRPF40A	6	0	2.44E-02	PMC3201212
WNK1	6	0	2.44E-02	PMID10660600
EEF2	0	8	2.58E-02	
ASPM	7	1	3.89E-02	PMID15972725
SLIT3	7	1	3.89E-02	PMID9813312
COL1A2	47	28	4.57E-02	

In summary, we present a platform for transcriptome-wide RNA sequencing in situ and demonstrate imaging and analytic approaches across multiple specimen types and spatial scales. FISSEQ correlates well with RNA-seq, except for genes involved in RNA and protein processing, possibly because some cellular structures or classes of RNA are less accessible to FISSEQ. It is notable that FISSEQ generates far fewer reads than RNA-seq but predominantly detects genes characterizing cell type and function. If this finding can be

generalized, FISSEQ may be used to identify cell types based on gene expression profiles in situ. Using partition sequencing to control the signal density, it may even be possible to combine transcriptome profiling and in situ mutation detection in a high-throughput manner [116,131,132]. Using RNA barcodes from expression vectors, one can label up to 4^N (N = barcode length) cells uniquely, much more than is possible using a combination of fluorescent proteins [133]. Similar to next-generation sequencing, we expect advances in read length, sequencing depth and coverage, and library preparation (i.e., fragmentation, rRNA depletion, targeted sequencing). Such advances may lead to improved stratification of diseased tissues in clinical medicine. Although more work remains, our present demonstration is an important first step toward a new era in biology and medicine.

7.3 **Experimental Methods**

7.3.1.1 Cell Lines

An openly consented and IRB-approved Personal Genome Project (PGP) iPS cell line can be obtained from Coriell (GM23338). The donor primary fibroblast is GM23248. The doxycycline-inducible mCherry HeLa cell line is constructed by combining the TetON-3G system (Clontech) and the Piggybac transposon (Systems Bioscience) into a single doxycycline- inducible transposable vector.

7.3.1.2 Cell Culture

iPS cells are grown in mTeSR1 media (Stem Cell Technologies) on ES cell- qualified Matrigel-coated plates (EMD Millipore). Primary fibroblasts are grown in 10% fetal bovine serum D-MEM or 15% newborn calf serum D-MEM/F12 with 10 ng/ml human EGF, 1×

L-glutamine, pen/strep, and nonessential amino acid (Life Technologies). HeLa cells are grown in 10% FBS DMEM with GlutaMAX, sodium pyruvate, pen/strep, and 10 ug/mL puromycin (Life Technologies). All cells are grown on glass bottom Petri dishes (Mattek dish part No. P35GC-1.5-14-C).

7.3.1.3 Wound closure assay

Human primary fibroblasts are grown to 100 percent confluency in EGF-supplemented media on a glass bottom Petri dish. Using the corner of a glass cover slip, multiple scratches (~200-um wide) are made on the cell monolayer. Eight to twelve hours after the first scratch, the cells are fixed for analysis.

7.3.1.4 Amplicon stretching on glass assay

RCA amplicons are generated *in vitro* using a synthetic DNA template under FISSEQ RCA conditions. Aminosilane-treated glass is prepared by incubating cleaned #1.5 glass coverslips for 10 minutes in 1% (3-Aminopropyl)triethoxysilane (Sigma) in acetone. RCA amplicons are bound to the glass in PBS and fixed under FISSEQ fixation conditions. Amplicons were counted using ImageJ.

7.3.1.5 Cell fixation and permeabilization

Cells are fixed with 4% formaldehyde (Sigma) for 15 minutes. iPS cells are permeabilized using 0.25% Triton-X100 (Calbiochem) for 15 minutes, and primary fibroblasts are permeabilized using 70% ethanol for 5 minutes. Samples are then incubated with 0.1 N hydrochloric acid for 2 minutes. For tissue sections, 0.01% of pepsin (Roche) in 0.1 N hydrochloric acid is used for permeabilization, followed by three PBS washes to neutralize pepsin.

7.3.1.6 In situ reverse transcription and amplification

A 200 uL mixture containing 4,000 U M-MuLV reverse transcriptase (Enzymatics), 250 uM dNTP (Enzymatics), 40 uM aminoallyl dUTP (Anaspec), 50 U RNase inhibitor (Enzymatics), and 100 pmol tagged random hexamers (/5Phos/TCTCGGGAACGCTGAAGANNNNNN), prepared on ice is added to cells at 25°C for 10 minutes. The concentration of aminoallyl dUTP can vary depending the cell type and the application. Generally, a high incorporation rate of aminoallyl dUTP results in better cross- linking and reduced cDNA diffusion but a lower amplicon density. The sample is then incubated overnight in a humidified 37°C chamber. The sample is washed using 1x PBS and cross-linked using BS(PEG)9 (Thermo Scientific), diluted to 50 mM in PBS, for 1 hour at 25°C. 1 M Tris (G Biosciences) is added to quench the reaction for 30 minutes at 25°C. A mixture of DNase-free RNases (Roche Diagnostics) and RNase H (Enzymatics) is added to degrade residual RNA for 1 hour at 37°C. A 100 uL circularization reaction mixture (1x reaction buffer, 2.5 mM $MnCl_2$, 1 M Betaine and 5 uL CircLigase II from Illumina/Epicentre) is then added to the sample well and incubated at 60°C for 2 hours. After circularization, the sample is washed using H_20 and incubated with a 200 uL mixture containing 0.1 uM RCA primer (TCTTCAGCGTTCCCGA*G*A from IDT) in 2x SSC and 30% formamide for 15 minutes at 60°C. The sample is washed using 2x SSC, and a 200 uL amplification mixture containing 500 U Phi29 DNA polymerase (Enzymatics), 250 uM dNTP, 40 uM, and aminoallyl dUTP is added. The sample is incubated in a dry 30°C chamber overnight and cross-linked using BS(PEG)9 diluted to 50 mM in PBS for 1 hour at 25°C. After a rinse with PBS, 1 M Tris is added to quench the reaction for 30 minutes. At this point, the sample can be stored in nuclease-free 1x PBS at 4°C.

7.3.1.7 SOLiD dinucleotide sequencing

We designed five sequencing primers specific to our universal adaptor (N, N-1, N-2, N-3, N-4, where N-*x* is recessed at the 5' end by *x*-nt). These sequencing primers are annealed to the sample sequentially, and *k* ligation reactions are performed for each primer (*k*+1 ligation reactions for primers N-2, N-3, and N-4). Each sequencing primer is annealed to the sample at 2.5 uM in 200 uL 80°C 5X SASC (0.75 M sodium acetate and 75 mM tri-sodium citrate, pH 7.5), incubating for 10 minutes at 25°C. The sample is washed twice for one minute each with 1 mL 1x Instrument Buffer (SOLiD Instrument Buffer Kit, Applied Biosystems Cat# 4389784). 200 uL sequencing mix is freshly prepared on ice using 165 uL nuclease-free H_20, 20 uL T4 DNA ligase buffer, 10 uL T4 DNA Ligase (Enzymatics), and 5 uL SOLiD sequencing oligos (the dark purple tubes from the SOLiD ToP Sequencing Kit Fragment Library F3 Tag MM50 Cat# 4449388). After aspirating the Instrument Buffer, the sequencing mix is added to the sample and incubated at 25oC for 45 minutes. The sequencing mix is aspirated, and the sample is washed with 1x Instrument Buffer (four 1 mL washes for 5 minutes each). Imaging is done in 1 mL 1x Instrument Buffer. After aspirating the Instrument Buffer, the fluorophore is cleaved to allow for subsequent ligation. The sample is incubated twice for 5 minutes each in 200 uL 1x Cleave Solution 1 (SOLiD ToP Instrument Buffer Kit Component 4406489), followed by two incubations for five minutes each in 200 uL 1X Cleave Mix 2.1 (SOLiD ToP Instrument Buffer Kit Component 4445677, prepared fresh with 106.7 uL Cleave Solution 2.1 Part 1 and 293.3 uL Cleave Solution Part 2). After the second incubation with Cleave Mix 2.1, the sample is washed three times for 5 minutes each with 1x Instrument Buffer. After repeating the cyclic ligation process *k* (or *k*+1) times, the ligated strands are stripped by four 5 minute washes in 80oC strip buffer (80%

formamide, 0.01% Triton-X100). Another sequencing primer is annealed, and the cyclic ligation process is repeated.

7.3.1.8 Partition Sequencing.

Fundamentally, any set of orthogonal primers can selectively sequence the amplicons for improved spatial resolution, provided that RT primers contain complementary adapter sequences. Our method uses sequencing primers that extend into the cDNA sequence by several nucleotides. The orthogonality defined by single nucleotide differences is based on the high specificity of T4 DNA ligase near the ligation junction (<6 bases away). With the template 3'-Adapter-NNN...-5', the sequencing primer can be one of 5'-Adapter (reverse complement)-A/G/C/T-3'. N is degenerate, so ligation occurs on $1/4^{th}$ of the amplicons that start with a defined base (T/C/G/A). The sequencing primer can also be one of 5'-Adapter (reverse complement)-$(A/G/C/T)_2$ -3' or 5'-Adapter (reverse complement)-$(A/G/C/T)_3$ -3'for detecting $1/16^{th}$ or $1/256^{th}$ of the amplicons. Sequencing reads from multiple primers can be combined, increasing spatial resolution, sequencing time, and cost. Partition sequencing can be useful for short barcode sequencing and ***in situ*** quantitation by effective serial dilution. Because SOLiD sequencing uses recessed primers itself, partition sequencing using SOLiD requires synthesis of specific bridge oligonucleotides. Instead, we use sequencing-by-ligation detailed in http://openwetware.org/wiki/Church_Lab [121].

7.3.1.9 Image acquisition.

The sample is firmly clamped to the stage of a confocal microscope. Imaging is done on a Zeiss Axio Observer with LSM 710 scanning laser confocal system, using the following excitation and emission profile. FITC: 25 mW Argon laser (458/488/514 nm) and 490-560

nm emission filter. Cy3: 20 mW DPSS laser (561 nm) and 563-593 nm emission filter. Texas Red: 2 mW HeNe laser (594 nm) and 597-647 nm emission filter. Cy5: 5 mW HeNe laser (633 nm) and 637-758 nm emission filter. Typically, we choose pixel resolution and optical slice thickness close to the Nyquist sampling rate (http://www.svi.nl/NyquistCalculator).

7.3.1.10 Image processing

The raw images are first deconvolved using Huygens Professional 4.3 running on a Windows 7 Intel Xeon workstation (typically less than five iterations) to increase the signal to noise ratio and reduce dephasing over sequencing cycles. For relatively flat cells such as fibroblasts, maximum intensity projection (MIP) is created from multiple optical slices in order to compensate for z-drift over time and to reduce the computational time required for image alignment. The separate images from different sequencing cycles form a time series (up to 30 time points) containing four colors. Using custom Matlab and Python scripts, individual images are corrected for chromatic shifts and registered locally (using a brick-based algorithm with 100 bricks per image). Software can be downloaded from www.arep.med.harvard.edu.

7.3.1.11 Base calling and spatial clustering

In order to identify the amplicon sequences, as well as their alignment to the reference sequence library, amplicons are treated as a set of spatially connected pixels with the similar color transitions over the sequencing run. Pixels at the amplicon border can have missing base calls due to reduced quality or misalignment, but short-read aligners can tolerate small mismatches. To cluster pixels with a shared alignment to the reference into objects, we first import the Tiff images and use the vector of fluorescent intensities to

calculate a component vector for each pixel: $C_x = I_x / (I_1^2 + I_2^2 + I_3^2 + I_4^2)^{1/2}$, where I_x indicates the intensity for channel x, and C_x represents the component value for channel x. The base call is determined by the largest component value at each pixel. Pixels without a largest component value are masked as not having a base call. The quality of the base call is determined as the geometric distance between the unit component vector and the unit vector in the direction of the base call. Another measure of quality, used for Satay plots, is the angle between the unit component vector and the unit vector in the direction of the base call, calculated as arccos(1-θ), where $\theta = C_x$ for the component value corresponding to the base call. Reads are generated for all pixels with less than 3-6 missing bases per read. Reads are then written to the csfasta format (0 = FAM, 1 = CY3, 2 = TXR, 3 = CY5, '.' = no base call) for alignment using the SOLiD color space-compatible short-read aligner Bowtie v1. The alignment output file from Bowtie is loaded back into the custom Python program, where alignments are re-assigned to the corresponding pixels using the read ID. For each alignment class (defined as having a particular reference sequence, strand, and position of alignment), all pixels with mapped reads are spatially clustered with a user-specified kernel. (We defined connectedness as being separated by no more than one pixel.) Clusters are redefined as amplicons occupying a set of pixels and having various statistics such as centroid, consensus mismatches (>50% of the pixels in the amplicon share a particular Bowtie mismatch), mean per-base quality, etc. Amplicons can have additional alignment classes if more than 1/2 of the pixels are shared between an existing amplicon and a new cluster generated from a different alignment class. This allows an amplicon to have multiple alignments, such as when a read aligns identically to several variants of a transcript. For each amplicon, a best alignment class is defined as the set of alignments with the highest strata of

mean alignment quality over all bases and pixels. The best alignment class is exclusively used for filtering and in downstream analysis. Amplicons are then filtered by size and summary statistics are written to file. Per-base sequencing error rate is calculated as the total number of consensus mismatches at each base divided by the total number of consensus reads (e.g. Base 1 Error Rate = (# Amplicons w/Consensus Mismatch at Base 1) / (Total # Amplicons). The mean and median per-base error statistics are generated using these values.

7.3.1.12 Bowtie read alignment.

Reads are aligned separately to several references to allow granularity in the Bowtie settings: Human RefSeq RNA (containing all RefSeq NM and NR class annotations), human ribosomal RNA reference (including mitochondrial rRNAs not found in the human RefSeq reference), and human tRNA reference. For the first experiment, the 27 sequenced bases correspond to nucleotides 1-25, plus nucleotides 28 and 29, after the sequencing primer. Since base calls are missing for nucleotides 26 and 27, we allow up to 3 mismatches of any base call quality value in the seed region (Bowtie flag: -n 3 –l 15 –e 240). In both cases, all alignments in the best strata (Bowtie flags: --best --strata) are reported for reads with less than 20 alignments (Bowtie flag: -m 20). All statistics are calculated using a read length value reflecting the number of bases actually sequenced. In plots and figures, any statistics for the missing bases are excluded but can be understood to be zero, since no sequencing data are acquired for these bases.

7.3.1.13 Expression array and RNA-Seq

The total RNA is isolated from ~500,000 PGP1 primary fibroblasts, immortalized B-cells, and iPS cells (RNeasy, Qiagen) for BeadChip HuRef-8 v3 expression arrays (Illumina) and RNA-seq. For RNA-seq, the cDNA library is generated using random hexamers and

poly dT reverse transcription primers for linear displacement amplification (Ovation RNA-seq System, NuGEN). After size selection, the sequencing library is prepared using SPRIworks Fragment Library System (Beckman Coulter) and sequenced on HiSeq2000 (Illumina) for 75-base paired-end reads (Partners HealthCare Center for Personalized Genetic Medicine, Harvard). We use fastq-mcf, BWA, and eXpress 1.4.0 for read processing, alignment (Human RNA RefSeq), and gene quantification. The total number of paired-end reads mapped to the transcriptome from primary fibroblasts, B-cells, and iPS cells are 27.2 million, 14.1 million, and 12.4 million, respectively.

7.3.1.14 Subcellular transcript localization & differential expression

Since the nuclei are brighter in our images than the rest of the sample, we use the MATLAB image processing toolbox and manual annotation to generate nuclear masks, and determine whether the centroid of each amplicon is located inside or outside of the nuclear mask. For each gene, the cytoplasmic to nuclear ratio is compared to that of other genes with a known subcellular localization profile using Fisher's exact test in R to generate *p*-values, odds ratios, and 95% confidence intervals. In most cases, we are comparing <200 genes, so our *p*-values are not multiple hypotheses corrected. For differential expression we used genes with >5 total observations to calculate Fisher's exact test *p*-values, odds ratios, and 95% confidence intervals.

7.3.1.15 Data visualization & plots

We use Bitplane Imaris 7.6 for visualizing 3D images during the sequencing run and for creating movies. Gene expression data visualization is done using ggplot2 in R. In cases where multiple genes have the same expression counts, we add random noise ('jitter plot') to avoid over-plotting and use multiple colors to denote different genes.

7.3.1.16 Statistics

For gene ontology analysis, a default setting in DAVID 6.7 is used. For comparing FISSEQ to RNA-seq, Pearson's correlation is generated between the datasets after applying various minimum and maximum count thresholds to the FISSEQ dataset. Increasing the minimum count threshold increases the correlation between FISSEQ and RNA-seq. Reducing the maximum count threshold initially improves the correlation by excluding *FN1* (>500 counts; most genes are under 200 counts). A further reduction of the maximum count threshold leads to a lower correlation. Thus, we treat *FN1* as an outlier for expression level analysis.

7.4 **Acknowledgments**

The original idea for FISSEQ has been conceived by George M. Church over a decade ago. Je Hyuk (Jay) Lee re-initiated efforts for in situ sequencing with the FISSEQ approach and developed the initial protocol. Given my experience with *in situ* PLP method, I joined efforts with Jay to further develop, troubleshoot, and fine-tune the FISSEQ procedure. Our early efforts led to the development of a robust method for *in situ* RNA-seq library construction and image visualization under fluorescence microscopy in various cell types, serving as the foundation for subsequent developments. After generating detectable signal from RNA, the development of sequencing capabilities was spearheaded by Jay and Evan R. Daugharthy. Evan also led the efforts to develop an image analysis pipeline for large-scale quantitative analysis of FISSEQ images. Reza Kalhor performed the mCherry quantification experiment to demonstrate specificity of the system. Jonathan Scheiman

prepared and validated the FISSEQ libraries while Jay and Evan performed sequencing and imaging analysis in the final stages of method development.

Thomas C. Ferrante assisted in developing high-resolution imaging for FISSEQ. Rich Terry, Samuel A. Inverso, Chao Li, Derek T. Peters, and Brian M. Turczyk assisted with automation and Derek assisted with partition sequencing. Ryoji Amamoto assisted with additional *in situ* RNA-seq library construction optimization. Adam H. Marblestone generated the AFM and TEM images. Prashant Mali and Xavier Rios generated the fluorescently tagged cell lines, and Sauveur S.F. Jeanty maintained the human iPS cells. Amy Bernard generated the mouse brain section. John Aach advised on the study design. Jay, Evan, and George wrote the paper, while Jonathan, Reza, and I helped to edit. Jay and Evan contributed equally to this manuscript. From RNA to library construction, to signal detection, to sequencing, and to image analysis, the collective effort of several researchers over many years finally brought the vision of FISSEQ into reality.

Data can be downloaded from http://arep.med.harvard.edu/FISSEQ_Science_2014/ and Gene Expression Omnibus (gene expression arrays: GSM313643, GSM313646, and GSM313657; RNA-seq: GSE54733). We thank S. Kosuri, K. Zhang, and M. Nilsson for discussions; A. DePace for Drosophila embryos; and I. Bachelet for antibody conjugation. Funded by NIH Centers of Excellence in Genomic Sciences grant P50 HG005550. J.H.L. and co-workers were funded by the National Heart, Lung, and Blood Institute, NIH, grant RC2HL102815; the Allen Institute for Brain Science, and the National Institute of Mental Health, NIH, grant MH098977. E.R.D. was funded by NIH grant GM080177 and NSF Graduate Research Fellowship Program grant DGE1144152. A.H.M. was funded by the Hertz Foundation. Potential conflicts of interests for G.M.C. are listed on http://arep.med.harvard.edu/gmc/tech.html.

J.H.L., E.R.D., R.T., and G.M.C. are authors on a patent application from the Wyss Institute that covers the method of generating three-dimensional nucleic acid–containing matrix.

CHAPTER 8 Conclusion and Outlook

In summary, we present a platform for *in situ* transcriptome-wide RNA profiling to quantitatively unravel gene regulation of single cells in heterogeneous biological samples within the spatial context. While DNA stores a constant set of genetic instructions, the dynamic regulation of gene expression, inferred from RNA and protein levels in response to developmental cues and external stimuli, leads to differential cellular fate. As biological samples are inherently heterogeneous, understanding gene expression programs with single-cell resolution while preserving the spatial context is of great importance for the dissection of cellular and tissue function during development in disease.

A few challenges remain to be improved with the FISSEQ technology. Currently, the protocol requires manual sequencing over the course of 2-3 weeks which can be difficult to coordinate and labor-intensive. Efforts are being made to automate the sequencing process to streamline the ease of performing FISSEQ. In addition, the current protocol lacks rRNA depletion, detecting 40-80% rRNA within primary fibroblasts. The ability to deplete rRNA from the transcriptome will greatly enrich the number of mRNA reads per cell. Finally, we have a limited understanding of the inherent bias in FISSEQ. We have observed that

FISSEQ enriches for biologically active genes, yet the parameters governing such enrichment is unclear. We speculate that active RNA molecules may be more accessible, thus enriched in detection by FISSEQ.

Transcriptome-wide RNA profiling *in situ* reveal cellular and tissue heterogeneity at the molecular level. Having explored two approaches to achieve highly multiplexed RNA profiling *in situ*, we observed that the PLP method offers higher specificity[114,131], but is difficult for scaling up given the use of expensive LNA primers and the need for calibrating individual PLPs. On the other hand, FISSEQ provides a transcriptome-wide approach for visualizing RNA *in situ*. In certain applications, the two methods may work synergistically where FISSEQ discovers enriched genes and pathways *de novo*, and PLP further validates and investigates the underlying interplay between biomarkers of a particular cellular phenotype or gene regulation pathway. Given sufficient sensitivity and specificity, it is conceivable to identify different cell types based on their gene expression profile, enabling basic research in complex tissues such as the differentiation process of the developing embryo as well as clinical applications with the detection of cancer cells within heterogeneous solid tumors as an example. We are hopeful that with further technical improvements, *in situ* RNA sequencing technologies will paint a more complete picture of the transcriptional landscape within the spatial context of complex biological samples and open the doors for the study of transcriptional regulation in biology and medicine.

CHAPTER 9 Reference and Appendix

References

1. Watson JD, Crick FHC. Molecular structure of nucleic acids. *Nature*. 1953;171(4356):737-738. doi:10.1097/BLO.0b013e3181468780.

2. Wilkins MHF, Stokes a R, Wilson HR. Molecular structure of deoxypentose nucleic acids. *Nature*. 1953;171(4356):738-740. doi:10.1038/171738a0.

3. Franklin RE, Gosling RG. The structure of sodium thymonucleate fibres. I. The influence of water content. *Acta Crystallogr*. 1953;6(8):673-677. doi:10.1107/S0365110X53001939.

4. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001;409(6822):860-921. doi:10.1038/35057062.

5. Capecchi MR. Altering the genome by homologous recombination. *Science*. 1989;244(4910):1288-1292. doi:10.1126/science.2660260.

6. Rudin N, Sugarman E, Haber JE. Genetic and physical analysis of double-strand break repair and recombination in Saccharomyces cerevisiae. *Genetics*. 1989;122(3):519-534.

7. Plessis A, Perrin A, Haber JE, Dujon B. Site-specific recombination determined by I-SceI, a mitochondrial group I intron-encoded endonuclease expressed in the yeast nucleus. *Genetics*. 1992;130(3):451-460.

8. Rouet P, Smih F, Jasin M. Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. *Mol Cell Biol*. 1994;14(12):8096-8106. doi:10.1128/MCB.14.12.8096.

9. Choulika A, Perrin A, Dujon B, Nicolas JF. Induction of homologous recombination in mammalian chromosomes by using the I-SceI system of Saccharomyces cerevisiae. *Mol Cell Biol*. 1995;15(4):1968-1973.

10. Smith J, Grizot S, Arnould S, et al. A combinatorial approach to create artificial homing endonucleases cleaving chosen sequences. *Nucleic Acids Res*. 2006;34(22). doi:10.1093/nar/gkl720.

11. Bibikova M, Carroll D, Segal DJ, et al. Stimulation of homologous recombination through targeted cleavage by chimeric nucleases. *Mol Cell Biol.* 2001;21(1):289-297. doi:10.1128/MCB.21.1.289-297.2001.

12. Urnov FD, Miller JC, Lee Y-L, et al. Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature.* 2005;435(7042):646-651. doi:10.1038/nature03556.

13. Miller JC, Holmes MC, Wang J, et al. An improved zinc-finger nuclease architecture for highly specific genome editing. *Nat Biotechnol.* 2007;25(7):778-785. doi:10.1038/nbt1319.

14. Boch J, Scholze H, Schornack S, et al. Breaking the code of DNA binding specificity of TAL-type III effectors. *Science.* 2009;326(5959):1509-1512. doi:10.1126/science.1178811.

15. Moscou MJ, Bogdanove AJ. A simple cipher governs DNA recognition by TAL effectors. *Science.* 2009;326(5959):1501. doi:10.1126/science.1178817.

16. Christian M, Cermak T, Doyle EL, et al. Targeting DNA double-strand breaks with TAL effector nucleases. *Genetics.* 2010;186(2):756-761. doi:10.1534/genetics.110.120717.

17. Mali P, Yang L, Esvelt KM, et al. RNA-guided human genome engineering via Cas9. *Science.* 2013;339(6121):823-826. doi:10.1126/science.1232033.

18. Cong L, Ran FA, Cox D, et al. Multiplex genome engineering using CRISPR/Cas systems. *Science.* 2013;339(6121):819-823. doi:10.1126/science.1231143.

19. Jinek M, East A, Cheng A, Lin S, Ma E, Doudna J. RNA-programmed genome editing in human cells. *Elife.* 2013;2:e00471. doi:10.7554/eLife.00471.

20. Maeder ML, Thibodeau-Beganny S, Osiak A, et al. Rapid "Open-Source" Engineering of Customized Zinc-Finger Nucleases for Highly Efficient Gene Modification. *Mol Cell.* 2008;31(2):294-301. doi:10.1016/j.molcel.2008.06.016.

21. Holkers M, Maggio I, Liu J, et al. Differential integrity of TALE nuclease genes following adenoviral and lentiviral vector gene transfer into human cells. *Nucleic Acids Res.* 2013;41(5). doi:10.1093/nar/gks1446.

22. Ishino Y, Shinagawa H. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. *J ….* 1987. http://jb.asm.org/content/169/12/5429.short. Accessed October 29, 2013.

23. Barrangou R, Fremaux C, Deveau H, et al. CRISPR provides acquired resistance against viruses in prokaryotes. *Science*. 2007;315(5819):1709-1712. doi:10.1126/science.1138140.

24. Horvath P, Barrangou R. CRISPR/Cas, the immune system of bacteria and archaea. *Science*. 2010;327(5962):167-170. doi:10.1126/science.1179555.

25. Fineran PC, Charpentier E. Memory of viral infections by CRISPR-Cas adaptive immune systems: Acquisition of new information. *Virology*. 2012;434(2):202-209. doi:10.1016/j.virol.2012.10.003.

26. Wiedenheft B, Sternberg SH, Doudna J a. RNA-guided genetic silencing systems in bacteria and archaea. *Nature*. 2012;482(7385):331-338. doi:10.1038/nature10886.

27. Bhaya D, Davison M, Barrangou R. CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. *Annu Rev Genet*. 2011;45:273-297. doi:10.1146/annurev-genet-110410-132430.

28. Terns MP, Terns RM. CRISPR-based adaptive immune systems. *Curr Opin Microbiol*. 2011;14(3):321-327. doi:10.1016/j.mib.2011.03.005.

29. Carte J, Wang R, Li H, Terns RM, Terns MP. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes Dev*. 2008;22(24):3489-3496. doi:10.1101/gad.1742908.

30. Deltcheva E, Chylinski K, Sharma CM, et al. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature*. 2011;471(7340):602-607. doi:10.1038/nature09886.

31. Gesner EM, Schellenberg MJ, Garside EL, George MM, Macmillan AM. Recognition and maturation of effector RNAs in a CRISPR interference pathway. *Nat Struct Mol Biol*. 2011;18(6):688-692. doi:10.1038/nsmb.2042.

32. Haurwitz RE, Jinek M, Wiedenheft B, Zhou K, Doudna J a. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science*. 2010;329(5997):1355-1358. doi:10.1126/science.1192272.

33. Sashital DG, Jinek M, Doudna J a. An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3. *Nat Struct Mol Biol*. 2011;18(6):680-687. doi:10.1038/nsmb.2043.

34. Wang R, Preamplume G, Terns MP, Terns RM, Li H. Interaction of the Cas6 riboendonuclease with CRISPR RNAs: recognition and cleavage. *Structure*. 2011;19(2):257-264. doi:10.1016/j.str.2010.11.014.

35. Brouns SJJ, Jore MM, Lundgren M, et al. Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science*. 2008;321(5891):960-964. doi:10.1126/science.1159689.

36. Hale C, Kleppe K, Terns RM, Terns MP. Prokaryotic silencing (psi)RNAs in Pyrococcus furiosus. *RNA*. 2008;14(12):2572-2579. doi:10.1261/rna.1246808.

37. Jore MM, Lundgren M, van Duijn E, et al. Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nat Struct Mol Biol*. 2011;18(5):529-536. doi:10.1038/nsmb.2019.

38. Wiedenheft B, Lander GC, Zhou K, et al. Structures of the RNA-guided surveillance complex from a bacterial immune system. *Nature*. 2011;477(7365):486-489. doi:10.1038/nature10402.

39. Lintner N, Kerou M, Brumfield S. Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (. *J Biol* …. 2011. http://www.jbc.org/content/286/24/21643.short. Accessed October 29, 2013.

40. Wiedenheft B, Duijn E Van, Bultema JB, et al. RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proc Natl Acad Sci*. 2011;108(36):15010-15010. doi:10.1073/pnas.1111854108.

41. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna J a, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*. 2012;337(6096):816-821. doi:10.1126/science.1225829.

42. Gilbert LA, Larson MH, Morsut L, et al. XCRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell*. 2013;154(2). doi:10.1016/j.cell.2013.06.044.

43. Konermann S, Brigham MD, Trevino AE, et al. Optical control of mammalian endogenous transcription and epigenetic states. *Nature*. 2013;500(7463):472-476. doi:10.1038/nature12466.

44. Maeder ML, Linder SJ, Cascio VM, Fu Y, Ho QH, Joung JK. CRISPR RNA-guided activation of endogenous human genes. *Nat Methods*. 2013;10(10):977-979. doi:10.1038/nmeth.2598.

45. Mali P, Aach J, Stranges PB, et al. CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat Biotechnol*. 2013;31(9):833-838. doi:10.1038/nbt.2675.

46. Perez-Pinera P, Kocak DD, Vockley CM, et al. RNA-guided gene activation by CRISPR-Cas9-based transcription factors. *Nat Methods*. 2013;10(10):973-976. doi:10.1038/nmeth.2600.

47. Chen B, Gilbert LA, Cimini BA, et al. Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell*. 2013;155(7):1479-1491. doi:10.1016/j.cell.2013.12.001.

48. Lo B, Parham L. Ethical issues in stem cell research. *Endocr Rev*. 2009;30(3):204-213. doi:10.1210/er.2008-0031.

49. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*. 2006;126(4):663-676. doi:10.1016/j.cell.2006.07.024.

50. Takahashi K, Tanabe K, Ohnuki M, et al. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*. 2007;131(5):861-872. doi:10.1016/j.cell.2007.11.019.

51. Yu J, Vodyanik M a, Smuga-Otto K, et al. Induced pluripotent stem cell lines derived from human somatic cells. *Science*. 2007;318(5858):1917-1920. doi:10.1126/science.1151526.

52. Park I-H, Zhao R, West J a, et al. Reprogramming of human somatic cells to pluripotency with defined factors. *Nature*. 2008;451(7175):141-146. doi:10.1038/nature06534.

53. Maherali N, Sridharan R, Xie W, et al. Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. *Cell Stem Cell*. 2007;1(1):55-70. doi:10.1016/j.stem.2007.05.014.

54. Wernig M, Meissner A, Foreman R, et al. In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature*. 2007;448(7151):318-324. doi:10.1038/nature05944.

55. Wang H, Yang H, Shivalila CS, et al. One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell*. 2013;153(4):910-918. doi:10.1016/j.cell.2013.04.025.

56. DiCarlo JE, Norville JE, Mali P, Rios X, Aach J, Church GM. Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. *Nucleic Acids Res*. 2013;41(7):4336-4343. doi:10.1093/nar/gkt135.

57. Gratz SJ, Cummings AM, Nguyen JN, et al. Genome engineering of Drosophila with the CRISPR RNA-guided Cas9 nuclease. *Genetics*. 2013;194(4):1029-1035. doi:10.1534/genetics.113.152710.

58. Hwang WY, Fu Y, Reyon D, et al. Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nat Biotechnol*. 2013;31(3):227-229. doi:10.1038/nbt.2501.

59. Friedland AE, Tzur YB, Esvelt KM, Colaiácovo MP, Church GM, Calarco J a. Heritable genome editing in C. elegans via a CRISPR-Cas9 system. *Nat Methods*. 2013;10(8):741-743. doi:10.1038/nmeth.2532.

60. Cho SW, Kim S, Kim JM, Kim J-S. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nat Biotechnol.* 2013;31(3):230-232. doi:10.1038/nbt.2507.

61. Chen F, Pruett-Miller SM, Huang Y, et al. High-frequency genome editing using ssDNA oligonucleotides with zinc-finger nucleases. *Nat Methods.* 2011;8(9):753-755. doi:10.1038/nmeth.1653.

62. Saleh-Gohari N, Helleday T. Conservative homologous recombination preferentially repairs DNA double-strand breaks in the S phase of the cell cycle in human cells. *Nucleic Acids Res.* 2004;32(12):3683-3688. doi:10.1093/nar/gkh703.

63. Urnov FD, Rebar EJ, Holmes MC, Zhang HS, Gregory PD. Genome editing with engineered zinc finger nucleases. *Nat Rev Genet.* 2010;11(9):636-646. doi:10.1038/nrg2842.

64. Cong L, Ran FA, Cox D, et al. Multiplex genome engineering using CRISPR/Cas systems. *Science.* 2013;339:819-823. doi:10.1126/science.1231143.

65. Park I-H, Arora N, Huo H, et al. Disease-specific induced pluripotent stem cells. *Cell.* 2008;134:877-886. doi:10.1016/j.cell.2008.07.041.

66. Hsu PD, Scott D a, Weinstein J a, et al. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol.* 2013;31(9):827-832. doi:10.1038/nbt.2647.

67. Yates F, Daley GQ. Progress and prospects: gene transfer into embryonic stem cells. *Gene Ther.* 2006;13:1431-1439. doi:10.1038/sj.gt.3302854.

68. Yang L, Yang JL, Byrne S, Pan J, Church GM. CRISPR/Cas9-Directed Genome Editing of Cultured Cells. *Curr Protoc Mol Biol.* 2014;107:31.1.1-31.1.17. doi:10.1002/0471142727.mb3101s107.

69. Woltjen K, Michael IP, Mohseni P, et al. piggyBac transposition reprograms fibroblasts to induced pluripotent stem cells. *Nature.* 2009;458:766-770. doi:10.1038/nature07863.

70. Kaji K, Norrby K, Paca A, Mileikovsky M, Mohseni P, Woltjen K. Virus-free induction of pluripotency and subsequent excision of reprogramming factors. *Nature.* 2009;458:771-775. doi:10.1038/nature07864.

71. Warren L, Manos PD, Ahfeldt T, et al. Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mRNA. *Cell Stem Cell.* 2010;7:618-630. doi:10.1016/j.stem.2010.08.012.

72. Diebold SS, Kaisho T, Hemmi H, Akira S, Reis e Sousa C. Innate antiviral responses by means of TLR7-mediated recognition of single-stranded RNA. *Science.* 2004;303:1529-1531. doi:10.1126/science.1093616.

73. Hornung V, Ellegast J, Kim S, et al. 5'-Triphosphate RNA is the ligand for RIG-I. *Science*. 2006;314:994-997. doi:10.1126/science.1132505.

74. Pichlmair A, Schulz O, Tan CP, et al. RIG-I-mediated antiviral responses to single-stranded RNA bearing 5'-phosphates. *Science*. 2006;314:997-1001. doi:10.1126/science.1132998.

75. Kawai T, Akira S. Antiviral signaling through pattern recognition receptors. *J Biochem*. 2007;141:137-145. doi:10.1093/jb/mvm032.

76. Uematsu S, Akira S. Toll-like receptors and Type I interferons. *J Biol Chem*. 2007;282:15319-15323. doi:10.1074/jbc.R700009200.

77. Karikó K, Buckstein M, Ni H, Weissman D. Suppression of RNA recognition by Toll-like receptors: The impact of nucleoside modification and the evolutionary origin of RNA. *Immunity*. 2005;23:165-175. doi:10.1016/j.immuni.2005.06.008.

78. Karikó K, Muramatsu H, Welsh FA, et al. Incorporation of pseudouridine into mRNA yields superior nonimmunogenic vector with increased translational capacity and biological stability. *Mol Ther*. 2008;16:1833-1840. doi:10.1038/mt.2008.200.

79. Uzri D, Gehrke L. Nucleotide sequences and modifications that determine RIG-I/RNA binding and signaling activities. *J Virol*. 2009;83:4174-4184. doi:10.1128/JVI.02449-08.

80. Visa N, Izaurralde E, Ferreira J, Daneholt B, Mattaj IW. A nuclear cap-binding complex binds Balbiani ring pre-mRNA cotranscriptionally and accompanies the ribonucleoprotein particle during nuclear export. *J Cell Biol*. 1996;133:5-14. doi:10.1083/jcb.133.1.5.

81. Lewis JD, Izaurralde E. The role of the cap structure in RNA processing and nuclear export. *Eur J Biochem*. 1997;247:461-469. doi:10.1111/j.1432-1033.1997.00461.x.

82. Ran FA, Hsu PD, Wright J, Agarwala V, Scott D a, Zhang F. Genome engineering using the CRISPR-Cas9 system. *Nat Protoc*. 2013;8:2281-2308. doi:10.1038/nprot.2013.143.

83. Ball MP, Thakuria J V., Zaranek AW, et al. Inaugural Article: A public resource facilitating clinical use of genomes. *Proc Natl Acad Sci*. 2012;109:11920-11927. doi:10.1073/pnas.1201904109.

84. Watanabe K, Ueno M, Kamiya D, et al. A ROCK inhibitor permits survival of dissociated human embryonic stem cells. *Nat Biotechnol*. 2007;25:681-686. doi:10.1038/nbt1310.

85. Sander JD, Zaback P, Joung JK, Voytas DF, Dobbs D. Zinc Finger Targeter (ZiFiT): An engineered zinc finger/target site design tool. *Nucleic Acids Res.* 2007;35. doi:10.1093/nar/gkm349.

86. Sander JD, Maeder ML, Reyon D, Voytas DF, Joung JK, Dobbs D. ZiFiT (Zinc Finger Targeter): An updated zinc finger engineering tool. *Nucleic Acids Res.* 2010;38. doi:10.1093/nar/gkq319.

87. Yang L, Guell M, Byrne S, et al. Optimization of scarless human stem cell genome editing. *Nucleic Acids Res.* 2013;41:9049-9061. doi:10.1093/nar/gkt555.

88. Güell M, Yang L, Church GM. Genome editing assessment using CRISPR Genome Analyzer (CRISPR-GA). *Bioinformatics.* 2014:1-3. doi:10.1093/bioinformatics/btu427.

89. Panier S, Boulton SJ. Double-strand break repair: 53BP1 comes into focus. *Nat Rev Mol Cell Biol.* 2014;15(1):7-18. doi:10.1038/nrm3719.

90. Yang H, Wang H, Shivalila CS, Cheng AW, Shi L, Jaenisch R. Resource One-Step Generation of Mice Carrying Reporter and Conditional Alleles by CRISPR / Cas-Mediated Genome Engineering. *Cell.* 2013;154(6):1370-1379. doi:10.1016/j.cell.2013.08.022.

91. Bione S, D'Adamo P, Maestrini E, Gedeon AK, Bolhuis PA, Toniolo D. A novel X-linked gene, G4.5. is responsible for Barth syndrome. *Nat Genet.* 1996;12(4):385-389. doi:10.1038/ng0496-385.

92. Houtkooper RH, Turkenburg M, Poll-The BT, et al. The enigmatic role of tafazzin in cardiolipin metabolism. *Biochim Biophys Acta - Biomembr.* 2009;1788(10):2003-2014. doi:10.1016/j.bbamem.2009.07.009.

93. Wang G, McCain ML, Yang L, et al. Modeling the mitochondrial cardiomyopathy of Barth syndrome with induced pluripotent stem cell and heart-on-chip technologies. *Nat Med.* 2014;20(6):616-623. doi:10.1038/nm.3545.

94. Storici F, Bebenek K, Kunkel TA, Gordenin DA, Resnick MA. RNA-templated DNA repair. *Nature.* 2007;447(7142):338-341. doi:10.1038/nature06114.

95. Nishimasu H, Ran FA, Hsu PD, et al. Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell.* 2014;156(5):935-949. doi:10.1016/j.cell.2014.02.001.

96. Sternberg SH, Redding S, Jinek M, Greene EC, Doudna J a. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature.* 2014;507(7490):62-67. doi:10.1038/nature13011.

97. Fu Y, Foden J a, Khayter C, et al. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat Biotechnol.* 2013;31(9):822-826. doi:10.1038/nbt.2623.

98. Dianov GL, Hübscher U. Mammalian base excision repair: The forgotten archangel. *Nucleic Acids Res*. 2013;41(6):3483-3490. doi:10.1093/nar/gkt076.

99. Thurman RE, Rynes E, Humbert R, et al. The accessible chromatin landscape of the human genome. *Nature*. 2012;489(7414):75-82. doi:10.1038/nature11232.

100. Valamehr B, Abujarour R, Robinson M, et al. A novel platform to enable the high-throughput derivation and characterization of feeder-free human iPSCs. *Sci Rep*. 2012;2. doi:10.1038/srep00213.

101. Park I-H, Lerou PH, Zhao R, Huo H, Daley GQ. Generation of human-induced pluripotent stem cells. *Nat Protoc*. 2008;3:1180-1186. doi:10.1038/nprot.2008.92.

102. Tsai SQ, Zheng Z, Nguyen NT, et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat Biotechnol*. 2014;33(2):187-198. doi:10.1038/nbt.3117.

103. Wade N. Scientists Seek Ban on Method of Editing the Human Genome. *New York Times*. 2015:1-4. http://www.nytimes.com/2015/03/20/science/biologists-call-for-halt-to-gene-editing-technique-in-humans.html.

104. Eberwine J, Kacharmina JE, Andrews C, et al. mRna expression analysis of tissue sections and single cells. *J Neurosci*. 2001;21(21):8310-8314. doi:21/21/8310 [pii].

105. Kurimoto K, Yabuta Y, Ohinata Y, Saitou M. Global single-cell cDNA amplification to provide a template for representative high-density oligonucleotide microarray analysis. *Nat Protoc*. 2007;2(3):739-752. doi:10.1038/nprot.2007.79.

106. Geiss GK, Bumgarner RE, Birditt B, et al. Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nat Biotechnol*. 2008;26(3):317-325. doi:10.1038/nbt1385.

107. Raj A, van Oudenaarden A. Single-molecule approaches to stochastic gene expression. *Annu Rev Biophys*. 2009;38:255-270. doi:10.1146/annurev.biophys.37.032807.125928.

108. Tang F, Barbacioru C, Wang Y, et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods*. 2009;6(5):377-382. doi:10.1038/nmeth.1315.

109. Toriello NM, Douglas ES, Thaitrong N, et al. Integrated microfluidic bioprocessor for single-cell gene expression analysis. *Proc Natl Acad Sci U S A*. 2008;105(51):20173-20178. doi:10.1073/pnas.0806355106.

110. Vargas DY, Raj A, Marras SAE, Kramer FR, Tyagi S. Mechanism of mRNA transport in the nucleus. *Proc Natl Acad Sci U S A*. 2005;102(47):17008-17013. doi:10.1073/pnas.0505580102.

111. Femino AM, Fay FS, Fogarty K, Singer RH. Visualization of single RNA transcripts in situ. *Science*. 1998;280(5363):585-590. doi:10.1126/science.280.5363.585.

112. Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Methods*. 2008;5(10):877-879. doi:10.1038/nmeth.1253.

113. Choi HMT, Chang JY, Trinh LA, Padilla JE, Fraser SE, Pierce NA. Programmable in situ amplification for multiplexed imaging of mRNA expression. *Nat Biotechnol*. 2010;28(11):1208-1212. doi:10.1038/nbt.1692.

114. Larsson C, Grundberg I, Söderberg O, Nilsson M. In situ detection and genotyping of individual mRNA molecules. *Nat Methods*. 2010;7(5):395-397. doi:10.1038/nmeth.1448.

115. Bagasra O. Protocols for the in situ PCR-amplification and detection of mRNA and DNA sequences. *Nat Protoc*. 2007;2(11):2782-2795. doi:10.1038/nprot.2007.395.

116. Lee J-H, Park I-H, Gao Y, et al. A robust approach to identifying tissue-specific gene expression regulatory variants using personalized human induced pluripotent stem cells. *PLoS Genet*. 2009;5(11):e1000718. doi:10.1371/journal.pgen.1000718.

117. Lee JH, Daugharthy ER, Scheiman J, et al. Highly multiplexed subcellular RNA sequencing in situ. *Science*. 2014;343(6177):1360-1363. doi:10.1126/science.1250212.

118. Diez-Roux G, Banfi S, Sultan M, et al. A high-resolution anatomical atlas of the transcriptome in the mouse embryo. *PLoS Biol*. 2011;9(1). doi:10.1371/journal.pbio.1000582.

119. Zeng H, Shen EH, Hohmann JG, et al. Large-scale cellular-resolution gene profiling in human neocortex reveals species-specific molecular signatures. *Cell*. 2012;149(2):483-496. doi:10.1016/j.cell.2012.02.052.

120. Lécuyer E, Yoshida H, Parthasarathy N, et al. Global Analysis of mRNA Localization Reveals a Prominent Role in Organizing Cellular Architecture and Function. *Cell*. 2007;131(1):174-187. doi:10.1016/j.cell.2007.08.003.

121. Shendure J, Porreca GJ, Reppas NB, et al. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*. 2005;309(5741):1728-1732. doi:10.1126/science.1117389.

122. Kim JB, Porreca GJ, Song L, et al. Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science*. 2007;316(5830):1481-1484. doi:10.1126/science.1137325.

123. Mitra RD, Butty VL, Shendure J, Williams BR, Housman DE, Church GM. Digital genotyping and haplotyping with polymerase colonies. *Proc Natl Acad Sci U S A*. 2003;100(10):5926-5931. doi:10.1073/pnas.0936399100.

124. Drmanac R, Sparks AB, Callow MJ, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science*. 2010;327(5961):78-81. doi:10.1126/science.1181498.

125. Levsky JM, Shenoy SM, Pezo RC, Singer RH. Single-cell gene expression profiling. *Science*. 2002;297(5582):836-840. doi:10.1126/science.1072241.

126. Itzkovitz S, van Oudenaarden A. Validating transcripts with probes and imaging technology. *Nat Methods*. 2011;8(4 Suppl):S12-S19. doi:10.1038/nmeth.1573.

127. Frohman MA, Dush MK, Martin GR. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proc Natl Acad Sci U S A*. 1988;85(23):8998-9002. doi:10.1073/pnas.85.23.8998.

128. McAnulty RJ. Fibroblasts and myofibroblasts: Their source, function and role in disease. *Int J Biochem Cell Biol*. 2007;39(4):666-671. doi:10.1016/j.biocel.2006.11.005.

129. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*. 2009;4(1):44-57. doi:10.1038/nprot.2008.211.

130. Kornblihtt AR, Pesce CG, Alonso CR, et al. The fibronectin gene as a model for splicing and transcription studies. *FASEB J*. 1996;10(2):248-257.

131. Ke R, Mignardi M, Pacureanu A, et al. In situ sequencing for RNA analysis in preserved tissue and cells. *Nat Methods*. 2013;10(9):857-860. doi:10.1038/nmeth.2563.

132. Zhang K, Li JB, Gao Y, et al. Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nat Methods*. 2009;6(8):613-618. doi:10.1038/nmeth.1357.

133. Livet J, Weissman TA, Kang H, et al. Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature*. 2007;450(7166):56-62. doi:10.1038/nature06293.

134. Lee JH, Daugharthy ER, Scheiman J, et al. Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat Protoc*. 2015. doi:10.1038/nprot.2014.191.

Appendix 9.A: CRISPR/Cas9-Directed Genome Editing of Cultured Cells

The work presented in this chapter has been published in the following paper[68]:

- Yang L, Yang JL, Byrne S, Pan J, Church GM. CRISPR/Cas9-Directed Genome Editing of Cultured Cells. *Curr Protoc Mol Biol.* 2014;107:31.1.1-31.1.17. doi:10.1002/0471142727.mb3101s107.

CRISPR/Cas9-Directed Genome Editing of Cultured Cells

Luhan Yang,[1,3] Joyce L. Yang,[1,2,3] Susan Byrne,[1,3] Joshua Pan,[2] and George M. Church[1]

[1]Department of Genetics, Harvard Medical School, Boston, Massachusetts
[2]Biological and Biomedical Sciences Program, Harvard Medical School, Boston, Massachusetts
[3]These authors contributed equally to this work.

ABSTRACT

Human genome engineering has been transformed by the introduction of the CRISPR (clustered regularly interspaced short palindromic repeats)/Cas (CRISPR-associated) system found in most bacteria and archaea. Type II CRISPR/Cas systems have been engineered to induce RNA-guided genome editing in human cells, where small RNAs function together with Cas9 nucleases for sequence-specific cleavage of target sequences. Here we describe the protocol for Cas9-mediated human genome engineering, including construct building and transfection methods necessary for delivering Cas9 and guide RNA (gRNA) into human-induced pluripotent stem cells (hiPSCs) and HEK293 cells. Following genome editing, we also describe methods to assess genome editing efficiency using next-generation sequencing and isolate monoclonal hiPSCs with the desired modifications for downstream applications. *Curr. Protoc. Mol. Biol.* 107:31.1.1-31.1.17.

Keywords: genome engineering • CRISPR • human stem cells

INTRODUCTION

Targeted human genome editing enables functional studies of genetic variation in biology and disease, and holds tremendous potential for clinical applications. To facilitate genome engineering, technologies such as Zinc-Finger Nucleases (ZFNs) and Transcription Activator-Like Effector Nucleases (TALENs) have been developed to enable targeted and programmable modification of endogenous genomic sequences (Miller et al., 2007; Hockemeyer et al., 2011). However, the need to design new complex nucleases for each target site limits the utility of these methods, particularly in multiplexed gene targeting applications.

Recently, the type II bacterial CRISPR (clustered regularly interspaced short palindromic repeats)/Cas (CRISPR-associated) system has been developed as an efficient and versatile technology for genome editing in eukaryotic cells and whole organisms (Jinek et al., 2012, 2013; Cong et al., 2013; DiCarlo et al., 2013; Friedland et al., 2013; Gratz et al., 2013; Hwang et al., 2013; Mali et al., 2013a; Wang et al., 2013). The CRISPR/Cas system was first identified in bacteria and archaea as an RNA-mediated adaptive defense system that safeguards organisms from invading viruses and plasmids (Ishino et al., 1987; Horvath and Barrangou, 2010; Wiedenheft et al., 2012). The hallmark of the CRISPR/Cas system consists of CRISPR arrays composed of spacers interspersed with direct repeats and *cas* genes present in the operons (Bhaya et al., 2011; Terns and Terns, 2011). In CRISPR/Cas-mediated immunity, bacteria and archaea react to viral or plasmid attack in the adaptive phase by first integrating short fragments of foreign nucleic acid (protospacers) into the host chromosome at the proximal end of the CRISPR array. In the expression phase, CRISPR loci are transcribed into precursor CRISPR RNA (pre-crRNA) and further

Current Protocols in Molecular Biology 31.1.1-31.1.17, July 2014
Published online July 2014 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/0471142727.mb3101s107

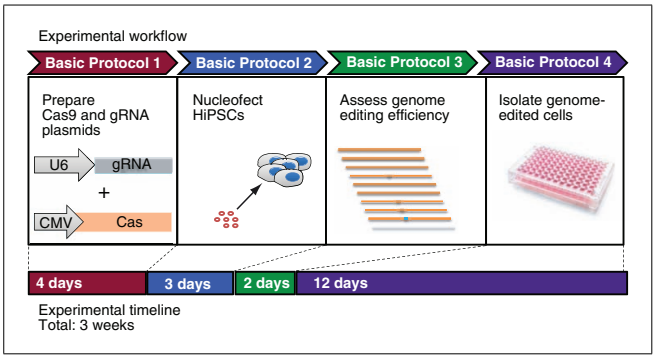

Figure 31.1.1 Workflow and timeline for hiPSC genome engineering. In the complete workflow, the Cas9 and gRNA plasmid constructs are built and subsequently transfected into cells. Genome editing efficiency is assessed using deep sequencing. Finally, monoclonal hiPSC colonies with desired genotype can be isolated using cell sorting. The entire workflow takes 3 weeks to perform.

processed into a library of short CRISPR RNAs (crRNAs) that can recognize and pair with complementary sequences from invading viral or plasmid targets (Carte et al., 2008; Haurwitz et al., 2010; Deltcheva et al., 2011; Gesner et al., 2011; Sashital et al., 2011; Wang et al., 2011). In the final interference phase, crRNAs are packaged with trans-activating crRNA (tracrRNA) and Cas proteins to form ribonucleoprotein complexes that together detect and destroy foreign sequences (Brouns et al., 2008; Hale et al., 2008; Jore et al., 2011; Lintner et al., 2011; Wiedenheft et al., 2011a,b).

It has been recently demonstrated that the type II CRISPR system from *Streptococcus pyrogenes* can be engineered to induce Cas9-mediated double-stranded breaks (DSBs) in a sequence-specific manner in vitro by providing a synthetic guide RNA (gRNA) composed of crRNA fused to tracrRNA (Jinek et al., 2012). Moreover, the system has been successfully adapted to function in human cells with the use of human codon-optimized Cas9 and customizable 20-nt gRNAs (Cho et al., 2013; Cong et al., 2013; Jinek et al., 2013; Mali et al., 2013a). Once the gRNA identifies its 20-bp target followed by a PAM (protospacer-adjacent motif) sequence–NGG, Cas9 nuclease then cleaves the target sequence, creating a DSB (Jinek et al., 2012). The resulting DSB will either generate nonspecific mutations knocking out a gene through the error-prone NHEJ (non-homologous end joining) pathway, or produce specific modifications dictated by an exogenous repair template through the HDR (homology-directed repair) pathway (Saleh-Gohari and Helleday, 2004; Urnov et al., 2010; Chen et al., 2011). This system greatly enhances the ease of genome engineering through the creation of desired DSBs targeted by RNA sequences that are easy to design, synthesize, and deliver, holding great promise for multiplexed genome editing.

With the advent of human induced pluripotent stem cells (hiPSCs) that can be re-programmed from fibroblasts to a human embryonic stem cell (hESC)–like state with maintained pluripotency, self-renewal, and differentiation capacity, a better understanding of human biology and potential clinical applications is now possible (Takahashi and Yamanaka, 2006; Maherali et al., 2007; Takahashi et al., 2007; Wernig et al., 2007; Yu et al., 2007; Park et al., 2008). hiPSC technology presents a promising tool for supplying various cell types for transplantation therapy, regenerative medicine, drug testing, and developmental biology experiments. The potential of hiPSCs can be further enhanced

by genome engineering, which may be used to study human gene function, track cells or endogenous proteins with a knock-in reporter, and correct genetic defects for gene therapy.

To harness the full potential of hiPSC technology, this unit provides a streamlined method for conducting genome editing in hiPSCs (Fig. 31.1.1). Basic Protocol 1 describes the construction of Cas9 and gRNA plasmids, including the purification of the Cas9 plasmid from stab cultures obtained from Addgene, bioinformatic analysis to determine an appropriate target sequence, and construction of gRNA plasmid from IDT gBlocks. Basic Protocol 2 describes the transfection of hiPSCs, while the Alternate Protocol outlines the same process for HEK293 cells. Basic Protocol 3 describes the assessment of genome editing efficiency in successfully transfected cells. Finally, Basic Protocol 4 describes a method to isolate monoclonal hiPSC colonies with desired genotype.

PREPARATION OF Cas9 AND gRNA PLASMIDS

BASIC PROTOCOL 1

Plasmids containing Cas9 and the guide RNA are necessary for Cas9-mediated genome editing. This basic protocol outlines the steps necessary to prepare both plasmids for transfection.

Materials

Cas9 plasmid (Addgene, plasmid ID 41815) as bacterial stab in agar
LB agar plate containing 100 μg/ml ampicillin (*UNIT 1.1*)
LB liquid medium containing 100 μg/ml ampicillin (*UNIT 1.1*)
HiSpeed Plasmid Maxi Kit (Qiagen)
PCR-grade sterile deionized water
PCR-Blunt II-Topo kit (Invitrogen, cat. no. K2800-20) including One Shot Top10 Chemically Competent *E. coli* cells (other competent cells for cloning may also be used)
Sterilized glass beads (EMD Millipore, cat. no. 71013-3)
LB agar plate containing 50 μg/ml kanamycin (*UNIT 1.1*)
M13 Forward (5′-GTTTTCCCAGTCACGACG-3′) and M13 Reverse (5′-AACAGCTATGACCATG-3′) universal sequencing primers
LB liquid medium containing 50 μg/ml kanamycin (*UNIT 1.1*)
Qiagen plasmid Mini Kit (Qiagen)

Sterile pipet tips or toothpicks for picking colonies from agar plates
37°C incubator-shaker
Nanodrop microspectrophotometer (*http://www.nanodrop.com*)
Sequence analysis software (e.g., NCBI BLAST, UCSC Genome Browser BLAT, LaserGene)
DNA synthesis facility
42°C incubator for heat-shocking cells
10-ml bacterial culture tubes
Access to Sanger sequencing facility

Additional reagents and equipment for DNA synthesis (*UNIT 2.11*) and Sanger sequencing (*UNIT 7.1*)

Prepare Cas9 plasmid

1. Obtain plasmid from Addgene.
2. Use a sterile pipet tip or toothpick to scrape the bacterial stock from the Addgene bacterial stab, and streak it onto an LB agar plate containing 100 μg/ml ampicillin. Incubate plate at 37°C for 10 hr or overnight.

3. Once colonies are formed, pick a single colony from the plate to inoculate 200 ml of LB liquid medium containing 100 μg/ml ampicillin. Grow overnight at 37°C with shaking at 200 rpm.

4. Isolate plasmid DNA using a plasmid Maxiprep kit. Use Nanodrop microspectrophotometer to measure DNA concentration. Resuspend DNA at ~1 μg/μl in water. Use this product for transfection.

Identify appropriate gRNA targeting sequence

5. Using sequence analysis software, identify all 22-bp regions within 50 bp of the intended genomic target in the form of 5′-N19-NGG-3′.

   These 22-bp regions may be located on either strand and should ideally overlap the target sequence.

   The selected target sequence must follow the standard sequence structure of 5′-G-N19-NGG-3′. The 5′ G is necessary for the U6 promoter used on the gRNA plasmid, while the 3' NGG is the protospacer adjacent motif (PAM) that is necessary for Cas9 recognition. It must also be unique to the genomic target site and have minimal alternate targets on the genome.

6. For each candidate sequence, query for alternate binding sites in the reference genome. Because of the higher tolerance of mismatches in the first 7 bp of the target sequence, search the reference genome for the last 13 bp of the target sequence with the NGG protospacer adjacent motif (S_{13}NGG). Use NCBI BLASTN or other online software to choose the one with minimal off-target sites at region of interest. Finalize the design of the customized gRNA expression fragment (455 bp) by including the selected target sequence (N19) in the gRNA expression fragment below.

   This final sequence will contain everything necessary for gRNA expression, including the U6 promoter, customized target sequence, gRNA scaffold, and termination signal, as annotated in Figure 31.1.2.

Create gRNA plasmid construct from IDT gBlock

7. Synthesize the final gRNA expression fragment (455 bp) as a standard gBlock without any 5′ modifications from gene synthesis companies.

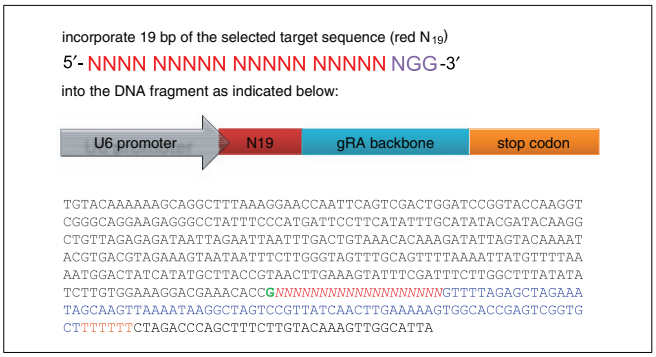

Figure 31.1.2 Overview of customized gRNA expression fragment. The chosen selected target sequence is inserted in the red region of the construct above. Of note, G in green indicates the start of the U6-driven transcript. For the color version of this figure, go to *http://www.currentprotocols.com/protocol/mb3101*.

8. Resuspend the gBlock (delivered at 200 ng) in 20 μl of water for a final concentration of 10 ng/μl.
9. Pipet 1 μl gBlock, 1 μl pCRII-Blunt-TOPO vector, and 4 μl salt solution (from PCR-Blunt II-Topo kit) in a 1.5-ml microcentrifuge tube, mixing gently. Incubate at room temperature for at least 5 min.
10. To transform 5 μl of product into Top10 Chemically Competent *E. coli* cells, thaw one aliquot of Top10 cells in ice for 10 min, add 5 μl of the TOPO cloning reaction from the previous step, and incubate on ice for 30 min. Heat-shock the cells at 42°C, then return to ice for 2 min. Add 250 μl of room temperature SOC medium (from PCR-Blunt II-Topo kit) and incubate at 37°C with shaking for 1 hr.
11. Spread 100 μl of the transformation mixture using sterilized glass beads onto a prewarmed LB agar plate containing 50 μg/ml kanamycin by gently swirling the plate or, alternatively, using an inoculating loop for spreading. Incubate overnight at 37°C.

    Expect 10 to 100 colonies, with the majority containing the desired insert.

12. After incubation, pick ~5 colonies for Sanger sequencing (*UNIT 7.7*) using the M13 Forward and M13 Reverse universal sequencing primers.
13. After identifying the colonies with the correct sequence, grow a maxiprep culture of the correct transformant by inoculating 200 ml of LB medium containing 50 μg/ml kanamycin with 100 μl of the original culture (step 11). Grow overnight at 37°C with shaking at 200 rpm.
14. Isolate plasmid DNA using a plasmid maxiprep kit. Resuspend plasmid DNA at ~1 μg/ml in water. Use this product for transfection (see Basic Protocol 2 and Alternate Protocol).

TRANSFECTION OF hiPSCs

BASIC PROTOCOL 2

Genome editing in hiPSCs holds great potential for gene therapy as well as the functional study of genetic variation when hiPSCs differentiate into relevant cell types. While the Cas9 and gRNA plasmids are being prepared, initiate hiPSC culture to prepare for transfection. The proper procedure for genome editing on tissue-cultured hiPSCs is described in this protocol.

NOTE: All culture incubations should be performed in a humidified 37°C, 5% CO_2 incubator unless otherwise specified.

NOTE: All reagents and equipment coming into contact with live cells must be sterile, and aseptic technique should be used accordingly.

Materials

PGP1 hiPSC cells adapted for growth on Matrigel (see personal genome project Web site: *http://www.personalgenomes.org/*)
Matrigel (hESC-qualified; BD Sciences, cat. no. 354277)
DMEM/F12 medium (Invitrogen)
mTeSR1 medium (StemCell Technologies, cat. no. 05850)
InSolution Rho kinase (ROCK) inhibitor (Calbiochem, cat. no. Y-27632)
P3 Primary Cell 4D-Nucleofector X kit containing P3 and Supplement 1 solutions in addition to 16-well Nucleocuvette Strips (Lonza, cat. no. V4XP-4032)
Cas9 plasmid DNA (see Basic Protocol 1)
gRNAexpression vector (see Basic Protocol 1)

Phosphate-buffered saline (PBS; Life Technologies, cat. no. 20012-050)
TrypLE Express (Invitrogen, cat. no. 12604-013)

6- and 48-well tissue culture–treated plates
15- and 50-ml conical centrifuge tubes (e.g., BD Falcon)
Countess automated cell counter (Invitrogen)
Tabletop centrifuge and plate adapter
Amaxa 4D-Nucleofector System (Lonza, cat. no. CD-MN025)

Additional reagents and equipment for culture of hiPSC in mTeSR medium (see Technical Manual Version 3.0.0 from Stem Cell Technologies; *http://www.stemcell.com/~/media/Technical%20Resources/B/C/A/2/B/29106MAN_3_0_0.pdf*)

Prepare for transfection

1. Culture PGP1 hiPSCs using standard protocol for hiPSC in mTeSR1 medium (in 6-well Matrigel-coated plates, until the cells are 40% confluent.

   To coat plates with Matrigel, do the following:
   a. Thaw a vial of 300 µl Matrigel on ice.
   b. Transfer 24 ml cold DMEM/F12 into a 50-ml conical polypropylene tube.
   c. Transfer 300 µl Matrigel into the tube. Invert to mix.
   d. Add 1 ml of this mixture per well of a 6-well plate, then leave the plate at room temperature for 1 hr.
   e. Aspirate Matrigel and replace with 2 ml cells/medium.

2. At a time point 2 hr before electroporation, replace the medium of the hiPSCs with 2 ml prewarmed mTeSR1 medium containing 2 µl/ml ROCK inhibitor.

3. At a time point 1 hr before electroporation, prepare destination wells for transfected cells:
   a. Thaw a vial of 300 µl Matrigel on ice
   b. Transfer 24 ml cold DMEM/F12 into a 50-ml conical polypropylene tube.
   c. Transfer 300 µl Matrigel into the tube. Invert to mix.
   d. Add 500 µl of this mixture per well of 48-well plate (one well will be needed per transfection), then leave the plate at room temperature for 1 hr.
   e. Aspirate Matrigel and replace with prewarmed 500 µl mTeSR1 medium with 2 µl/ml ROCK inhibitor.

   The small surface area of the wells of 48-well plates promotes high cell density and healthy growth after transfection.

4. Prepare a transfection master mix (scale appropriately):
   16.4 µl P3 and 3.6 µl Supplement 1 from Nucleofactor X kit
   1 µl 1 µg/µl Cas9 plasmid
   1 µl 1 µg/µl gRNA plasmid
   22 µl per reaction, total.

Transfect hiPSCs

5. Aspirate the ROCK inhibitor–containing medium from the wells containing hiPSCs and wash each well with 2 ml room temperature PBS.

6. Aspirate PBS, add 1 ml TrypLE Express, and incubate the plate at 37°C for 5 min.

7. Resuspend cells with 3 ml mTeSR1 medium and gently pipet up and down several times to generate a single-cell suspension. Transfer disassociated cells into a 15-ml centrifuge tube containing 10 ml mTeSR1 medium.

8. Count cells with cell counter and calculate total volume required for 1×10^6 cells/transfection, scaling as needed.

   Given the toxicity of transfection, the minimum number of cells per transfection required to isolate transfectants is 200,000. However, higher cell counts decrease the efficiency of transfection by increasing the number of targets. A titration of cell counts ranging from 200,000 to 1×10^6 may help find the optimal balance.

9. Place desired quantity of cells (in this case 1×10^6) in 15-ml centrifuge tube, centrifuge at $200 \times g$ for 5 min at room temperature, and aspirate supernatant.
10. Resuspend each unit of 1×10^6 cells in 22 µl of the transfection master mix prepared in step 4.
11. Quickly transfer cells into the central chamber of one well of a Nucleocuvette strip. Place the strip into 4-D Nucleofector device.
12. Nucleofect cells using program CB150.
13. Quickly add 80 µl of prewarmed mTESR1 medium containing 2 µl/ml ROCK inhibitor to each well of electroporated cells. Pipet up and down once or twice to mix.
14. Transfer cells from the strip to wells of the Matrigel-coated plate containing mTeSR1 medium with 2 µl/ml ROCK inhibitor prepared in step 3.
15. Centrifuge the plate 3 min at $70 \times g$, room temperature. Place cells into 37°C incubator.
16. After 24 hr, change to fresh mTeSR1 medium without ROCK inhibitor.
17. Harvest cells 3 days after electroporation. Follow protocol for assessing targeting efficiency in Basic Protocol 3.

TRANSFECTION OF HUMAN HEK293 CELLS

ALTERNATE PROTOCOL

Genome editing in HEK293 cells is efficient and convenient, thus serving as an ideal system to test and optimize reagent before moving to hiPSCs. Here, we describe the transfection procedure on HEK293 cells with the Cas9/gRNA plasmids.

NOTE: All culture incubations should be performed in a humidified 37°C, 5% CO_2 incubator unless otherwise specified.

NOTE: All reagents and equipment coming into contact with live cells must be sterile, and aseptic technique should be used accordingly.

Additional Materials *(also see Basic Protocol 2)*

HEK 293 cells (Invitrogen)
Complete DMEM medium (see recipe)
Lipofectamine 20000 (Invitrogen, cat. no. 11668027)
Opti-MEM medium (Invitrogen, cat. no. 31985062)
Cas9 plasmid DNA (see Basic Protocol 1)
gRNAexpression vector (see Basic Protocol 1)
12-well tissue culture treated plates

Plate 293 cells for transfection

1. Culture HEK 293 cells in complete DMEM medium in 6-well plates until the cells are ~70% confluent.
2. Aspirate medium and wash cells with 2 ml room temperature PBS.

3. Aspirate PBS, add 1 ml TrypLE Express, and incubate at 37°C for 2 min.
4. Resuspend cells with 5 ml prewarmed complete DMEM medium
5. Count cells using an automated cell counter and calculate volume required for 200,000 cells per transfection.
6. Place desired volume of cells into 15-ml centrifuge tube. Centrifuge 5 min at 200 × *g*, room temperature, and aspirate supernatant.
7. Resuspend cell pellet in 1 ml complete DMEM medium.
8. Plate cells in a 12-well tissue culture plate and return to incubator.

Transfect 293 cells

9. After a day of incubation, replace medium on cells with 1 ml fresh prewarmed complete DMEM medium. Return to incubator and allow to incubate while preparing DNA mix.
10. Add 5 μl Lipofectamine 2000 to 50 μl Opti-MEM in a 1.5-ml microcentrifuge tube. Invert several times to mix. Incubate the mixture at room temperature for 5 min.
11. Add 1 μg Cas9 plasmid and 1 μg gRNA to 50 μlOpti-MEM in a 1.5-ml microcentrifuge tube.
12. Add diluted DNA from step 11 to diluted Lipofectamine mixture from step 10, flicking the tube several times to mix.
13. Incubate the mixture 15 min at room temperature.
14. Add 100 μl of the mixture dropwise to the cells.
15. Replace medium after 24 hr with fresh prewarmed complete DMEM medium.

    High concentrations of Lipofectamine can be toxic. Monitor cell conditions. If high cell toxicity is observed, change to fresh DMEM medium after 8 hr.

16. Harvest cells 3 days after transfection.

BASIC PROTOCOL 3

GENOTYPING TRANSFECTED CELLS USING NEXT-GENERATION SEQUENCING

After transfection, the targeting efficiency needs to be assessed to determine whether isolation of genome-targeted cells from a heterogeneous population is feasible. Normally, targeting efficiency on the order of 1% is expected for hiPSCs using Cas9-gRNA system without selection. This basic protocol describes the assessment of the targeting efficiency using next-generation sequencing techniques that can yield high read depths on the targeted site from a population of nucleofected cells.

Materials

Illumina forward sequence (ACACTCTTTCCCTACACGACGCTCTTCCGATCT)
Illumina reverse sequence
(GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT)
Transfected hiPSCs (Basic Protocol 1 or Alternate Protocol) growing in culture
Phosphate-buffered saline (PBS; Life Technologies, cat. no. 20012-050)
mTeSR1 medium (StemCell Technologies, cat. no. 05850)
prepGEM gold buffer (ZyGEM)
prepGEM tissue protease enzyme (ZyGEM)
KAPA Hifi Hotstart Readymix (KAPA Biosystems)
Illumina amplification primers (see step 3)

Illumina index primers (ScriptSeq Index PCR Primers)
Illumina PCR primer (AATGATACGGCGACCACCGAGATCTACACTCTTTCC-CTACACGACGCTCTTCCGATCT)
2-log DNA ladder (New England Biolabs)
QIAquick PCR purification kit (Qiagen)

Computer running Primer3 software (*http://primer3.sourceforge.net/*) for primer identification
15-ml conical tubes (BD Falcon)
Tabletop centrifuge
Thermal cycler
Access to MiSeq sequencer

Additional reagents and equipment for agarose gel electrophoresis (*UNIT 20.5A*) and measuring DNA concentration (*APPENDIX 3D*)

Design Illumina amplification primers for the targeting region

1. Select a ~500-bp region around the targeting site.
2. Use Primer3 to identify optimal targeting primer sets that amplify 200 to 300 bp around the targeting site.
3. Finalize the design of and order the customized Illumina amplification primers:
   a. Append the Illumina forward sequence (ACACTCTTTCCCTACAC-GACGCTCTTCCGATCT) to the 5′ end of the forward primer from step 2.
   b. Append the Illumina reverse sequence (GTGACTGGAGTTCAGACGTGT-GCTCTTCCGATCT) to the 5′ end of the reverse primer from step 2.

   The Illumina amplification scheme is summarized in Figure 31.1.3.

Harvest cells and create sequencing library

4. Aspirate the mTeSR1 medium from the cultured, transfected hiPSCs and wash the cells gently with PBS.
5. Aspirate PBS, add 1 ml TrypLE Express, and incubate the plate at 37°C for 5 min.
6. Transfer disassociated cells into a 15-ml conical tube containing 10 ml mTeSR1 medium and centrifuge 5 min at 200 × *g*, room temperature.
7. Aspirate supernatant and resuspend the cell pellet with the residual medium left in the conical tube.
8. Prepare a 10-μl cell lysis reaction with the following reagents in a PCR strip:
   8.9 μl cell pellet suspension
   1 μl prepGEM gold buffer (ZyGEM)
   0.1 μl of prepGEM tissue protease enzyme (ZyGEM)
9. Incubate the reaction in a thermal cycler:
   75°C for 15 min
   95°C for 5 min.
10. Prepare a 20-μl PCR reaction to obtain the amplicon of the targeting region.
    1 μl of the reaction from step 9
    10 μl 2 × KAPA Hifi Hotstart Readymix
    0.2 μl 100 mM each Illumina amplification primer (see step 3)
    Water to 20 μl.
11. Perform PCR with the following parameters:

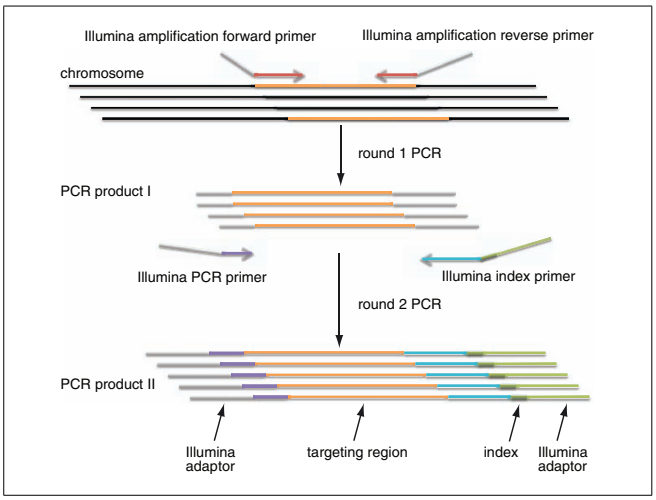

Figure 31.1.3 Schematic of Illumina sequencing library preparation. The first round of PCR amplifies the targeting region with the universal forward and reverse sequences necessary to anneal to the Illumina proprietary primers. The second round of PCR adds an index primer necessary for deconvoluting separate sequencing pools, as well as the adaptor necessary for attachment to the sequencing flow cell.

1 cycle:	5 min	95°C	(initial denaturation)
15 to 25 cycles:	20 sec	98°C	(denaturation)
	20 sec	65°C	(annealing)
	20 sec	72°C	(extension).

If you do not get clear product bands, try more cycles to amplify desired product.

12. Prepare the second round of PCR reaction to add the Illumina sequence adaptor.
    5 µl of the reaction from step 11
    10 µl KAPA Hifi Hotstart Readymix
    1 µl Illumina index primer
    0.1 µl of 100mM Illumina PCR primer
    Water to 20 µl.

    There are 48 orthogonal Illumina index primers from the ScriptSeq Index PCR Primers kit. Choose independent index primers for different reactions.

13. Perform the second round of PCR with the following parameters:

1 cycle:	5 min	95°C	(initial denaturation)
15 to 25 cycles:	20 sec	98°C	(denaturation)
	20 sec	65°C	(annealing)
	20 sec	72°C	(extension)
1 cycle:	4 min	72°C	(final extension).

If you do not get clear product bands, try more cycles to amplify desired product.

14. Run PCR product on a 2% agarose gel (*UNIT 20.5A*) against a 2-log DNA ladder and verify the correct amplicon length.

The Illumina sequencing adapter adds 160 bp to the genomic amplicon.

15. PCR purify the product with QIAquick PCR purification kit. Measure concentration of each sample (*APPENDIX 3D*) and pool each sample at the same concentration to ensure equal sequencing coverage. Submit for sequencing with MiSeq Personal Sequencer.

16. After the sequencing results arrive, analyze the results using the bioinformatics platform of choice.

    The authors recommend the CRISPR genome analyzer: http://54.80.152.219/.

    The incorporation frequency is defined by the percentage of sequences that have mutated away from the wild-type sequence from the original sample. An incorporation frequency of 1% or more is ideal for the downstream protocols.

SINGLE-CELL ISOLATION OF GENOME-TARGETED MONOCLONAL hiPSCs

BASIC PROTOCOL 4

After a successful round of Cas9-mediated genome engineering, the next step is to isolate monoclonal hiPSC colonies with the desired genotype. This can be accomplished by fluorescence-activated cell sorting (FACS) and genotyping of single cell–derived colonies. Once genome-edited monoclonal hiPSC colonies have been isolated, they can be utilized for downstream applications—e.g., differentiating into relevant tissue types to interrogate functionality in biology and disease.

Materials

0.1% (w/v) gelatin (StemCell Technologies, cat. no. 07903)
Irradiated CF-1 mouse embryonic fibroblasts (Michalska, 2007)
hES cell medium (see recipe)
Recombinant fibroblast growth factor (Millipore)
SMC4 (BD Biosciences; for 1× SMC4, supplement 500 ml of medium with one vial of SMC4 purchased from BD)
Fibronectin (StemCell Technologies)
Heterogenous pool of edited hiPSC (Basic Protocol 2)
mTeSR1 medium (StemCell Technologies, cat. no. 05850) supplemented with SMC4 (BD Biosciences) at final concentration of 1× (one vial per 500 ml medium)
mTeSR1 medium (StemCell Technologies, cat. no. 05850), unsupplemented
Phosphate-buffered saline (PBS; Invitrogen, cat. no. 20012-050)
Accutase (Millipore)
ToPro-3 viability dye (Invitrogen)
Matrigel (hESC-qualified; BD Sciences, cat. no. 354277)

96-well plates
BD FACSAria II SORP UV (BD Biosciences) with 100-mm nozzle
Centrifuge for 96-well plates
Access to Sanger sequencing facility

Additional reagents and equipment for obtaining amplicons of the targeting region (see Basic Protocol 3, steps 4 to 11)

1. One day before the experiment, prepare 96-well plates with CF-1 mouse embryonic fibroblast (MEF) as follows:
   a. Coat the plate by incubating 15 min with 50 μl/well of 0.1% (w/v) gelatin at room temperature, and wash with PBS.

b. Thaw and plate MEF in the gelatin-coated 96-well at a concentration of 1×10^6 cells/well in complete DMEM medium.
c. Incubate the MEF plate in the 37°C incubator overnight.
d. Following the overnight incubation, change the medium to hES cell medium supplemented with 100 ng/ml fibroblast growth factor, 1× SMC4 (one vial per 500 ml medium), and 5 mg/ml fibronectin.

2. Replace the medium on the hiPSCs in a 48-well plate from Basic Protocol 2 with mTeSR1 medium supplemented with 1× SMC4 (one vial per 500 ml medium) for at least 2 hr before FACS analysis is to be performed.
3. Aspirate the medium from the cultured hiPSCs, then wash the cells gently with PBS.
4. Aspirate PBS, add 200 μl/well Accutase (or enough to cover the well), and incubate at 37° for 5 to 10 min.
5. Generate the single-cell suspension by adding 1 ml mTeSR1 (unsupplemented) to each well and pipetting up and down gently several times.
6. Place cell suspension in a 15-ml conical tube, then centrifuge 5 min at 200 × *g*, room temperature. Aspirate supernatant.
7. Resuspend the cells with 1 ml mTeSR1 and add 0.5 μl of the viability dye ToPro-3.
8. Using a BD FACSAria II SORP UV with 100-mm nozzle under sterile conditions, sort single cells into individual wells of the 96-well plates prepared in step 1.

   The 100-mm nozzle is critical for the FACS experiment, to minimize the stress on hiPSCs.

9. After collection, centrifuge plates 3 min at 70 × *g*, room temperature. and place the plate into the tissue culture incubator
10. Four days after sorting, colony formation should be apparent; at this point replace the culture medium with hES cell medium supplemented with 1× SMC4.
11. Eight days after sorting, replace medium with hES medium (unsupplemented).

    SMC4 is beneficial for cell viability post sorting. However, long-duration exposure of cells to SMC4 may lead to cell differentiation. We recommend removing SMC4 from the culture medium once the colony formation is stable.

12. Passage the monoclonal hiPSC cells into Matrigel-coated 96-well plate and save half of the cells for genotyping.

    To coat wells with Matrigel, do the following:
    a. Thaw a vial of 300 μl Matrigel on ice.
    b. Transfer 24 ml cold DMEM/F12 into a 50-ml conical polypropylene tube.
    c. Transfer 300 μl Matrigel into the tube. Invert to mix.
    d. Add 100 μl of this mixture per well of a 96-well plate, then leave the plate at room temperature for 1 hr.
    e. Aspirate Matrigel and replace with 200 μl cells/medium.
13. Perform steps 4 to 11 in Basic Protocol 3 to obtain amplicons of the targeting region.

    The amplicon produced in the first round of Illumina PCR is sufficient for Sanger sequencing. Either the forward or the reverse primer can be used as the sequencing primer.

14. Perform Sanger sequencing to check the genotype of the targeting region.
15. Choose a colony containing the correct mutation for downstream differentiation or processing.

Table 31.1.1 Troubleshooting Common Problems with the CRISPR/Cas9 System

Problem	Possible cause	Solution
Low hiPSC viability after electroporation	DNA plasmid purity is low	Use maxiprep kit to generate high-quality DNA
	Too much DNA	Reduce the amount of DNA
	Cell density is too high before electroporation	Transfect cells under exponential growth phase
	Cell number is not sufficient	Use a minimum of 300,000 cells per transfection
	Delay of cell recovery after electroporation	Speed the recovery after electroporation by preparing all the necessary plates and pipets in advance and recovering the cells as soon as the electroporation is complete
	Cell is sensitive to trypsin treatment	Use nonenzymatic method to generate single-cell suspension, such as EDTA treatment
	gRNA off-target effect is prevalent and toxic to the cell	Try alternative gRNA targeting site
Low genome targeting efficiency	DNA transfection efficiency is low, or cell viability is low after transfection	Use Cas9-GFP construct followed by FACS to enrich transfected cell
	The targeting site is not accessible/targetable	There is no systematic knowledge yet regarding the impact of the targeting sequence on the targeting efficiency. We recommend that users design/generate/test multiple gRNAs near the region of interest.
Unable to obtain amplicon of the targeting region	Primer design is not optimal	Use Primer3 or other primer design software to optimize the design of primer on the targeting region
	Insufficient cell number	Start with at least >1000 cells
	Lysis reaction is not sufficient	Elongate the prepGEM digestion time
	Too much lysis reaction in the PCR reaction	Use no more than 1/10 volume of lysis reaction in the final PCR reaction

REAGENTS AND SOLUTIONS

Use deionized, distilled water in all recipes and protocol steps. For common stock solutions, see ***APPENDIX 2****; for suppliers, see* ***APPENDIX 4****.*

Complete medium for HEK 293 cells

High-glucose DMEM medium (Invitrogen) supplemented with:
10% fetal bovine serum (FBS)

continued

Table 31.1.2 Time Considerations for CRISPR/Cas9 Protocols

Procedure	Substep	Hands-on time	Total experiment time	Stopping point
Basic Protocol 1—Preparation of hCas9 and gRNA plasmids		5 hr	~ 4 days	Yes
	Prepare hCas9 plasmid	2 hr	2 hr	Yes
	Identify appropriate gRNA targeting sequence	1 hr	~3 days	Yes
	Plasmid construction of gRNA construct from IDT gBlock	2 hr	1 day	Yes
Basic Protocol 2—Transfection of Human iPS cells		3 hr	~3 days (recovery after transfection)	No
Basic Protocol 3—Genotyping transfected cells using next generation sequencing		4 hr	~2 days	Yes
	Harvest cells and create sequencing library	3 hr	3 hr + 1 day (sequencing)	Yes
Basic Protocol 4—Isolate genome-edited hiPSCs			~12 days	
	FACS sorting	3 hr	1 day (prepare MEF plate) +1 day (FACS) + 8 days (colony growth)	Yes
	Harvest cells and create sequencing amplicon	3 hr	3 hr + 1 day (Sanger sequencing)	Yes

1× nonessential amino acids (NEAA)
1× penicillin/streptomycin solution (pen/strep)
Store up to 3 to 6 months at 4ºC

hES cell medium

DMEM/F12 medium (e.g., Invitrogen) containing:
20% (v/v) knockout serum replacement (KOSR)
5 to 10 ng/ml bFGF
1 mM L-glutamine
100 μM nonessential amino acids
100 μM 2-mercaptoethanol
1× penicillin/streptomycin solution (pen/strep)
Store up to 3 to 6 months at 4ºC

COMMENTARY

Background Information

Cas9 is a tool for easily editing the genome of human cells. Compared with other genome editing methods, such as ZFNs and TALENs, the RNA-guided Cas9 system has certain advantages.

First, the simplicity of its design and construction make the tool more accessible. Second, the mere requirement of small RNAs for each new target allows for multiplexible genome targeting. Third, independent studies indicate that the Cas9 system is more efficient than other tools targeting the same region (Hwang et al., 2013; Mali et al., 2013a). However, the specificity of Cas9-mediated genome targeting is still under investigation. Judicious selection of the targeting site is necessary to minimize off-target effects.

Critical Parameters

As with any transfection, the quality of the DNA plasmid is paramount. We recommend plasmid maxiprep kits providing transfection-level DNA, especially for hiPSCs. The amount of DNA used in the transfection is also an important parameter, since higher transfection efficiency and dosage usually yield higher genome-targeting efficiency. When a new cell line is used for the first time, the amount of DNA needed for optimal transfection efficiency should be determined by titration. Finally, the concentration of DNA used in transfection is another important parameter, since the DNA volume used for hiPSC electroporation should be less than 1/10 of the total reaction volume to achieve effective transfection without incurring significant cell death. If the concentration is too low to satisfy this requirement, use a Speedvac evaporator to evaporate some of the water in the DNA solution, thereby increasing concentration.

The hiPSC density before transfection is important, as we have observed that genome editing on cells at exponential growth phase yields higher efficiency. We recommend that users conduct transfection on cells that have reached 30% to 40% confluence.

Finally, when assessing the efficiency of genome editing, the number of cells used in genotyping is critical for successful genotyping following this protocol. We tested the sensitivity of genotyping and found that a minimum of four cells is required to enable the amplification reaction. However, empirically, robust target region amplification occurs with >1000 cells.

Troubleshooting

Table 31.1.1 describes some problems commonly encountered with the protocols described in this unit, along with accompanying solutions.

Anticipated Results

We can achieve ~2% genome targeting efficiency in hiPSCs and ~30% genome targeting efficiency in HEK293 cells using the methods described above. The efficiency in hiPSCs varies with the targeting sites and locations. We detected 0.2% to 15% targeting efficiency in hiPSCs and 1% to >50% targeting efficiency in HEK 293 cells without any transfection enrichment and selection. We recommend that a transfection-enrichment strategy be used to maximize the efficiency.

Double-nickases represent an alternative approach for genome editing with mitigated off-targeted effects. It has been shown that the efficiency achieved by double-nickases is comparable to that of nuclease in HEK293 (Mali et al., 2013b; Ran et al., 2013). However, it is still under investigation whether the double-nickase strategy would work in hiPSCs.

Time Considerations

See Table 31.1.2 for a description of the time required for the protocols described in this unit.

Literature Cited

Bhaya, D., Davison, M., and Barrangou, R. 2011. CRISPR-Cas systems in bacteria and archaea: Versatile small RNAs for adaptive defense and regulation. *Annu. Rev. Genetics* 45:273-297.

Brouns, S.J.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J.H., Snijders, A.P.L., Dickman, M.J., Makarova, K.S., Koonin, E.V., and van der Oost, J. 2008. Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* 321:960-964.

Carte, J., Wang, R., Li, H., Terns, R.M., and Terns, M. P. 2008. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes Dev.* 22:3489-3496.

Chen, F., Pruett-Miller, S.M., Huang, Y., Gjoka, M., Duda, K., Taunton, J., Collingwood, T.N., Frodin, M., and Davis, G.D. 2011. High-frequency genome editing using ssDNA oligonucleotides with zinc-finger nucleases. *Nat. Methods* 8:753-755.

Cho, S. W., Kim, S., Kim, J. M., and Kim, J.-S. 2013. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nat. Biotechnol.* 31:230-232.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., and Zhang, F. 2013. Multiplex genome engineering using CRISPR/Cas systems. *Science* 339:819-823.

Deltcheva, E., Chylinski, K., Sharma, C.M., Gonzales, K., Chao, Y., Pirzada, Z.A., Eckert, M.R., Vogel, J., and Charpentier, E. 2011. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* 471:602-607.

DiCarlo, J.E., Norville, J.E., Mali, P., Rios, X., Aach, J., and Church, G.M. 2013. Genome engineering in *Saccharomyces cerevisiae* using CRISPR-Cas systems. *Nucleic Acids Res.* 41:4336-4343.

Friedland, A.E., Tzur, Y.B., Esvelt, K.M., Colaiácovo, M.P., Church, G.M., and Calarco, J.A. 2013. Heritable genome editing in *C. elegans* via a CRISPR-Cas9 system. *Nat. Methods* 10:741-743.

Gesner, E.M., Schellenberg, M.J., Garside, E.L., George, M.M., and Macmillan, A.M. 2011. Recognition and maturation of effector RNAs in a CRISPR interference pathway. *Nat. Struct. Mol. Biol.* 18:688-692.

Gratz, S.J., Cummings, A.M., Nguyen, J.N., Hamm, D.C., Donohue, L.K., Harrison, M.M., Wildonger, J., and O'Connor-Giles, K.M. 2013. Genome engineering of Drosophila with the CRISPR RNA-guided Cas9 nuclease. *Genetics* 194:1029-1035.

Hale, C., Kleppe, K., Terns, R.M., and Terns, M.P. 2008. Prokaryotic silencing (psi)RNAs in *Pyrococcus furiosus*. *RNA* 14:2572-2579.

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

Hockemeyer, D., Wang, H., Kiani, S., Lai, C.S., Gao, Q., Cassady, J.P., Cost, G.J., Zhang, L., Santiago, Y., Miller, J.C., Zeitler, B., Cherone, J.M., Meng, X., Hinkley, S.J., Rebar, E.J., Gregory, P.D., Urnov, F.D., and Jaenisch, R. 2011. Genetic engineering of human pluripotent cells using TALE nucleases. *Nat. Biotechnol.* 29:731-734.

Horvath, P. and Barrangou, R. 2010. CRISPR/Cas, the immune system of bacteria and archaea. *Science* 327:167-170.

Hwang, W.Y., Fu, Y., Reyon, D., Maeder, M.L., Tsai, S.Q., Sander, J.D., Peterson, R.T., Yeh, J.R., and Joung, J.K. 2013. Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nat. Biotechnol.* 31:227-229.

Ishino, Y., Shinagawa, H., Makino, K., Amemura, M., and Nakata, A. 1987. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. *J. Bacteriol.* 169:5429-5433.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337:816-821.

Jinek, M., East, A., Cheng, A., Lin, S., Ma, E., and Doudna, J. 2013. RNA-programmed genome editing in human cells. *eLife* 2:e00471.

Jore, M.M., Lundgren, M., van Duijn, E., Bultema, J.B., Westra, E.R., Waghmare, S.P., Wiedenheft, B., Pul, U., Wurm, R., Wagner, R., Beijer, M.R., Barendregt, A., Zhou, K., Snijders, A.P., Dickman, M.J., Doudna, J.A., Boekema, E.J., Heck, A.J., van der Oost, J., and Brouns, S.J. 2011. Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nat. Struct. Mol. Biol.* 18:529-536.

Lintner, N., Kerou, M., Brumfield, S., Graham, S., Liu, H., Naismith, J.H., Sdano, M., Peng, N., She, Q., Copié, V., Young, M.J., White, M.F., and Lawrence, C.M. 2011. Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE). *J. Biol. Chem.* 286:21643-21656.

Maherali, N., Sridharan, R., Xie, W., Utikal, J., Eminli, S., Arnold, K., Stadtfeld, M., Yachechko, R., Tchieu, J., Jaenisch, R., Plath, K., and Hochedlinger, K. 2007. Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. *Cell Stem Cell* 1:55-70.

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., and Church, G.M. 2013a. RNA-guided human genome engineering via Cas9. *Science* 339:823-826.

Mali, P., Aach, J., Stranges, P.B., Esvelt, K.M., Moosburner, M., Kosuri, S., Yang, L., and Church, G.M. 2013b. CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat. Biotechnol.* 31:833-838.

Michalska, A.E. 2007. Isolation and propagation of mouse embryonic fibroblasts and preparation of mouse embryonic feeder layer cells. *Curr. Protoc. Stem Cell Biol.* 3:1C.3.1-1C.3.17.

Miller, J.C., Holmes, M.C., Wang, J., Guschin, D.Y., Lee, Y.-L., Rupniewski, I., Beausejour, C.M., Waite, A.J., Wang, N.S., Kim, K.A., Gregory, P.D., Pabo, C.O., and Rebar, E.J. 2007. An improved zinc-finger nuclease architecture for highly specific genome editing. *Nat. Biotechnol.* 25:778-785.

Park, I.-H., Zhao, R., West, J.A, Yabuuchi, A., Huo, H., Ince, T.A, Lerou, P.H., Lensch, M.W., and Daley, G.Q. 2008. Reprogramming of human somatic cells to pluripotency with defined factors. *Nature* 451:141-146.

Ran, F.A., Hsu, P.D., Lin, C.Y., Gootenberg,, J.S., Konermann, S., Trevino, A.E., Scott, D.A., Inoue, A., Matoba, S., Zhang, Y., and Zhang, F. 2013. Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell* 154:1380-1389.

Saleh-Gohari, N. and Helleday, T. 2004. Conservative homologous recombination preferentially repairs DNA double-strand breaks in the S phase of the cell cycle in human cells. *Nucleic Acids Res.* 32:3683-3688.

Sashital, D.G., Jinek, M., and Doudna, J.A. 2011. An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3. *Nat. Struct. Mol. Biol.* 18:680-687.

Takahashi, K. and Yamanaka, S. 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 126:663-676.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. 2007. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* 131:861-872.

Terns, M.P. and Terns, R.M. 2011. CRISPR-based adaptive immune systems. *Curr. Opin. Microbiol.* 14:321-327.

Urnov, F.D., Rebar, E.J., Holmes, M.C., Zhang, H.S., and Gregory, P.D. 2010. Genome editing with engineered zinc finger nucleases. *Nat. Rev. Genetics* 11:636-646.

Wang, H., Yang, H., Shivalila, C.S., Dawlaty, M.M., Cheng, A.W., Zhang, F., and Jaenisch, R. 2013.

One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell* 153:910-918.

Wang, R., Preamplume, G., Terns, M.P., Terns, R.M., and Li, H. 2011. Interaction of the Cas6 riboendonuclease with CRISPR RNAs: Recognition and cleavage. *Structure* 19:257-264.

Wernig, M., Meissner, A., Foreman, R., Brambrink, T., Ku, M., Hochedlinger, K., Bernstein, B.E., and Jaenisch, R. 2007. In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature* 448:318-324.

Wiedenheft, B., van Duijn, E., Bultema, J.B., Waghmare, S., Zhou, K., Barendregt, A., Westphal, W., Heck, A.J., Boekema, E.J., Dickman, M.J., and Doudna, J.A. 2011a. RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proc. Natl. Acad. Sci. U.S.A.* 108:10092-10097.

Wiedenheft, B., Lander, G.C., Zhou, K., Jore, M.M., Brouns, S.J., van der Oost, J., Doudna, J.A., and Nogales, E. 2011b. Structures of the RNA-guided surveillance complex from a bacterial immune system. *Nature* 477:486-489.

Wiedenheft, B., Sternberg, S.H., and Doudna, J.A. 2012. RNA-guided genetic silencing systems in bacteria and archaea. *Nature* 482:331-338.

Yu, J., Vodyanik, M.A, Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J.L., Tian, S., Nie, J., Jonsdottir, G.A., Ruotti, V., Stewart, R., Slukvin, I.I., and Thomson, J.A. 2007. Induced pluripotent stem cell lines derived from human somatic cells. *Science* 318:1917-1920.

Appendix 9.B: Optimization of Scarless Human Stem Cell Genome Editing

The work presented in this chapter has been published in the following paper[87]:

- Yang L, Guell M, Byrne S, Yang JL, De Los Angeleos A, Mali P, Aach J, Kim-Kiselak C, Briggs AW, Rios X, Huang PY, Daley G, Church GM. Optimization of scarless human stem cell genome editing. *Nucleic Acids Res.* 2013;41:9049-9061. doi:10.1093/nar/gkt555.

Nucleic Acids Research Advance Access published July 31, 2013

Nucleic Acids Research, 2013, 1–13
doi:10.1093/nar/gkt555

Optimization of scarless human stem cell genome editing

Luhan Yang[1,2], Marc Guell[1], Susan Byrne[1], Joyce L. Yang[1,2], Alejandro De Los Angeles[3], Prashant Mali[1], John Aach[1], Caroline Kim-Kiselak[2], Adrian W Briggs[1], Xavier Rios[1], Po-Yi Huang[1,4], George Daley[3] and George Church[1,5,*]

[1]Department of Genetics, Harvard Medical School, Boston, 02115 MA, USA, [2]Biological and Biomedical Sciences Program, Harvard Medical School, Boston, 02115 MA, USA, [3]Children's Hospital, Boston, 02115 MA, USA, [4]Chemistry and Chemical Biology program, Harvard, 02138 Cambridge, MA, USA and [5]Wyss Institute for Biologically Inspired Engineering, Harvard University, Cambridge, 02138 MA, USA

Received March 25, 2013; Revised May 17, 2013; Accepted May 28, 2013

ABSTRACT

Efficient strategies for precise genome editing in human-induced pluripotent cells (hiPSCs) will enable sophisticated genome engineering for research and clinical purposes. The development of programmable sequence-specific nucleases such as Transcription Activator-Like Effectors Nucleases (TALENs) and Cas9-gRNA allows genetic modifications to be made more efficiently at targeted sites of interest. However, many opportunities remain to optimize these tools and to enlarge their spheres of application. We present several improvements: First, we developed functional re-coded TALEs (reTALEs), which not only enable simple one-pot TALE synthesis but also allow TALE-based applications to be performed using lentiviral vectors. We then compared genome-editing efficiencies in hiPSCs mediated by 15 pairs of reTALENs and Cas9-gRNA targeting *CCR5* and optimized ssODN design in conjunction with both methods for introducing specific mutations. We found Cas9-gRNA achieved 7–8× higher non-homologous end joining efficiencies (3%) than reTALENs (0.4%) and moderately superior homology-directed repair efficiencies (1.0 versus 0.6%) when combined with ssODN donors in hiPSCs. Using the optimal design, we demonstrated a streamlined process to generated seamlessly genome corrected hiPSCs within 3 weeks.

INTRODUCTION

Precise genome editing in human-induced pluripotent cells (hiPSCs) will enable functional studies of human genetic variation and enhance the potential use of hiPSCs for regenerative medicine. Currently, genome editing via sequence-specific nucleases represents the most efficient way to precisely edit human cell genomes (1 3). A nuclease-mediated double-stranded DNA (dsDNA) break in the genome can be repaired by two main mechanisms (4): non-homologous end joining (NHEJ), which frequently results in the introduction of non-specific insertions and deletions (indels), or homology-directed repair (HDR), which incorporates a homologous strand as a repair template. When a sequence-specific nuclease is delivered along with a homologous donor DNA construct containing the desired mutations, gene targeting efficiencies are increased by 1000-fold compared with just the donor construct alone (5). Thus, the development of programmable nucleases has greatly facilitated the practice of targeted genome engineering.

Despite large advances in gene editing tools, many challenges and questions remain regarding the use of custom-engineered nucleases in hiPSC engineering. First, despite their design simplicity, Transcription Activator-Like Effectors Nucleases (TALENs) target particular DNA sequences with tandem copies of Repeat Variable Diresidue (RVD) domains (6). Although the modular nature of RVDs simplifies TALEN design, their repetitive sequences complicate methods for synthesizing their DNA constructs (7 10) and also impair their use with lentiviral gene delivery vehicles, most likely by causing sequence instabilities (11).

Next, we sought to improve the ease and sensitivity of current detection methods for assessing genome editing. In

*To whom correspondence should be addressed. Tel: +1 617 432 3675; Fax: +1 617 432 6513; Email: gchurch@genetics.med.harvard.edu

current practice, NHEJ and HDR are frequently evaluated using separate assays. Mismatch-sensitive endonuclease assays (12) are often used for assessing NHEJ, but the quantitative accuracy of this method is variable, and the sensitivity is limited to NHEJ frequencies greater than ~3% (12). Meanwhile, HDR is frequently assessed by cloning and sequencing, a completely different and often cumbersome procedure. Sensitivity is still an issue because, although high editing frequencies on the order of 50% are frequently reported for some cell types, such as U2OS and K562 (10,13), frequencies are generally lower in hiPSCs (14). Recently, high editing frequencies have been reported in hiPSC and hESC using TALENs (15) and even higher frequencies with the CRISPR Cas9-gRNA system (16 19). However, editing rates at different sites appear to vary widely (17), and editing is sometimes not detectable at all at some sites (20). Moreover, although the recent successes in editing hiPSC genomes with TALENs and Cas9 are striking, genome editing using these tools has not yet been systematically explored and compared. To come to a fuller understanding of these issues and optimize inefficiencies will require simple and efficient collection and analysis of NHEJ and HDR rates at large numbers of sites using tools that accurately capture low as well as high rates. To this end, we developed a robust and user-friendly package using next generation sequencing to screen HR and NHEJ events in hiPSCs together.

As a demonstration of how our improved synthesis method for TALEs, and our genome editing assessment tool, can expedite data gathering, analysis and optimization, we used these tools to compare reTALEN and Cas9 efficiencies in hiPSCs at 15 sites near the *CCR5* locus. As with TALEN and Cas9 editing of hiPSCs, generally, use of ssODNs as DNA donors has been reported (21,22), but the optimal design and scope of ssODNs for this purpose have not been systematically explored. We then used our tools to optimize the design of ssODNs used as donors for scarless genome engineering.

Another area for improvement in editing procedures for hiPSC relates to the clonal isolation of the hiPSCs themselves, an operation that is difficult in part because hiPSC are difficult to grow out from isolated single cells because in the absence of appropriate cell-to-cell contacts with other hiPSCs or feeder cells. However, procedures that improve clonal hiPSC isolation have recently been reported (23), and we adapted these to integrate with the other procedures we report here. Taken all together, we demonstrate that it is possible to obtain clonal, precisely genome-edited hiPSCs within 3 weeks, including within this the amount of time required to synthesize optimal reagents and perform rapid prospective screening of target events.

MATERIALS AND METHODS

gRNA assembly

We incorporated 19 bp of the selected target sequence (i.e. 5′-N_{19} of 5′-N_{19}-NGG-3′) into two complementary 100 mer oligonucleotides (TTCTTGGCTTTATATATCTTG TGGAAAGGACGAAACACCGN19GTTTTAGAGCT AGAAATAGCAAGTTAAAATAAGGCTAGTCC). Each 100 mer oligonucleotide was suspended at 100 mM in water, mixed with equal volume and annealed in thermocycle machine (95°C, 5 min; Ramp to 4°C, 0.1°C/s). To prepare the destination vector, we linearized the gRNA cloning vector (Addgene plasmid ID 41824, Supplementary Sequence S3) using AflII and purified the vector through purification. We carried out the (10 µl) gRNA assembly reaction with 10 ng annealed 100 bp fragment, 100 ng destination backbone, 1× Gibson assembly reaction mix (New England Biolabs) at 50°C for 30 min, and reaction can be processed directly for bacterial transformation to colonize individual assemblies.

re-TALEs design and assembly

re-TALEs were optimized at different levels to facilitate assembly and improve expression. re-TALE DNA sequences were first co-optimized for a human codon-usage and low mRNA folding energy at the 5′ end (GeneGA, Bioconductor). The obtained sequence was evolved through several cycles to eliminate repeats (direct or inverted) longer than 11 bp (Supplementary Figure S8). In each cycle, synonymous sequences for each repeat are evaluated. Those with the largest hamming distance to the evolving DNA are selected. The sequence of one of re-TALE possessing 16.5 monomers is listed in Supplementary Sequence S1.

re-TALE dimer blocks encoding two RVDs (Supplementary Figure S2A) were generated by two rounds of PCR under standard Kapa HIFI (KPAP) PCR conditions, in which the first round of PCR introduced the RVD coding sequence and the second round of PCR generated the entire dimer blocks with 36 bp overlaps with the adjacent blocks. PCR products were purified using QIAquick 96 PCR Purification Kit (QIAGEN), and the concentrations were measured by Nano-drop. The primer and template sequences are listed in Supplementary Tables S1 and S2.

re-TALENs and re-TALE-TF destination vectors were constructed by modifying the TALE-TF and TALEN cloning backbones (24). We re-coded the 0.5 RVD regions on the vectors and also incorporated SapI cutting site at the designated re-TALE cloning site. The sequences of re-TALENs and re-TALE-TF backbones are listed in Supplementary Sequence S2. Plasmids can be pre-treated with SapI (New England Biolabs) with manufacturer recommended conditions and purified with QIAquick PCR purification kit (QIAGEN).

We carried out the (10 µl) one-pot TALE Single-incubation Assembly (TASA) assembly reaction with 200 ng of each block, 500 ng of destination backbone, 1× TASA enzyme mixture [2U SapI, 100 U Ampligase (Epicentre), 10 mU T5 exonuclease (Epicentre), 2.5U Phusion DNA polymerase (New England Biolabs)] and 1× isothermal assembly reaction buffer as described before (25) [5% PEG-8000, 100 mM Tris HCl (pH 7.5), 10 mM MgCl2, 10 mM DTT, 0.2 mM each of the four dNTPs and 1 mM NAD]. Incubations were performed at 37°C for 5 min and 50°C for 30 min. TASA assembly

reaction can be processed directly for bacterial transformation to colonize individual assemblies. The efficiency of obtaining full-length construct is ~20% with this approach. Alternatively, >90% efficiency can be achieved by three-steps assembly. First, 10 µl of re-TALE assembly reactions were performed with 200 ng of each block, 1× re-TALE enzyme mixture (100 U Ampligase, 12.5 mU T5 exonuclease, 2.5 U Phusion DNA polymerase) and 1× isothermal assembly buffer at 50°C for 30 min, followed by standardized Kapa HIFI PCR reaction, agarose gel electrophoresis and QIAquick Gel extraction (Qiagen) to enrich the full-length re-TALEs. In all, 200 ng of re-TALE amplicons can then be mixed with 500 ng of Sap1-pre-treated destination backbone, 1× re-TALE assembly mixture and 1× isothermal assembly reaction buffer and incubated at 50°C for 30 min. The re-TALE final assembly reaction can be processed directly for bacterial transformation to colonize individual assemblies. Additional notes of the assembly methods can be found in Supplementary Note S1.

Cell line and cell culture

PGP1 iPS cells were maintained on Matrigel (BD Biosciences)-coated plates in mTeSR1 (Stemcell Technologies). Cultures were passaged every 5 7 days with TrypLE Express (Invitrogen). The 293 T and 293FT cells were grown and maintained in Dulbecco's modified Eagle's medium (DMEM, Invitrogen) high glucose supplemented with 10% fetal bovine serum (Invitrogen), penicillin/streptomycin (pen/strep, Invitrogen) and non-essential amino acids (Invitrogen). K562 cells were grown and maintained in RPMI (Invitrogen) supplemented with 10% fetal bovine serum (Invitrogen 15%) and penicillin/streptomycin (pen/strep, Invitrogen). All cells were maintained at 37°C and 5% CO2 in a humidified incubator.

We established a stable 293T cell line for detecting HDR efficiency as described before (26). Specifically, the reporter cell lines bear genomically integrated GFP-coding sequences disrupted by the insertion of a stop codon and a 68 bp genomic fragment derived from the AAVS1 locus.

Test of reTALENs activity

We seeded 293 T reporter cells at densities of 2×105 cells per well in 24-well plate and transfected them with 1 µg of each re-TALENs plasmid and 2 µg DNA donor plasmid using Lipofectamine 2000 following the manufacturer's protocols. Cells were harvested using TrypLE Express (Invitrogen) ~18 h after transfection and resuspended in 200 µl of media for flow cytometry analysis using an LSRFortessa cell analyzer (BD Biosciences). The flow cytometry data were analyzed using FlowJo (FlowJo). At least 25 000 events were analyzed for each transfection sample. For endogenous AAVS1 locus targeting experiment in 293 T, the transfection procedures were identical as described earlier in the text, and we conducted puromycin selection with drug concentration at 3 µg/ml 1 week after transfection.

Functional lentivirus generation assessment

The lentiviral vectors were created by standard PCR and cloning techniques. The lentiviral plasmids were transfected by Lipofectamine 2000 with Lentiviral Packaging Mix (Invitrogen) into cultured 293FT cells (Invitrogen) to produce lentivirus. Supernatant was collected 48 and 72 h post-transfection, sterile filtered and 100 µl of filtered supernatant was added to 5×10^5 fresh 293 T cells with polybrene. Lentivirus titration was calculated based on the following formula: virus titration = (percentage of GFP+ 293 T cell × initial cell numbers under transduction)/(the volume of original virus collecting supernatant used in the transduction experiment). To test the functionality of lentivirus, 3 days after transduction, we transfected lentivirus transduced 293 T cells with 30 ng of plasmids carrying mCherry reporter and 500 ng of pUC19 plasmids using Lipofectamine 2000 (Invitrogen). Cell images were analyzed using Axio Observer Z.1 (Zeiss) 18 h after transfection and harvested using TrypLE Express (Invitrogen) and resuspended in 200 µl of media for flow cytometry analysis using LSRFortessa cell analyzer (BD Biosciences). The flow cytometry data were analyzed using BD FACSDiva (BD Biosciences).

Test of re-TALENs and Cas9-gRNA genome editing efficiency

PGP1 iPSCs were cultured in Rho kinase (ROCK) inhibitor Y-27632 (Calbiochem) 2 h before nucleofection. Transfections were done using P3 Primary Cell 4D-Nucleofector X Kit (Lonza). Specifically, cells were harvested using TrypLE Express (Invitrogen), and 2×10^6 cells were resuspended in 20 µl of nucleofection mixture containing 16.4 µl of P3 Nucleofector solution, 3.6 µl of supplement, 1 µg of each re-TALENs plasmid or 1 µg of Cas9 and 1 µg of gRNA construct, 2 µl of 100 µM ssODN. Subsequently, we transferred the mixtures to 20 µl of Nucleocuvette strips and conducted nucleofection using CB150 program. Cells were plated on Matrigel-coated plates in mTeSR1 medium supplemented with ROCK inhibitor for the first 24 h. For endogenous AAVS1 locus-targeting experiment with dsDNA donor, we used the identical procedure except we used 2 µg of dsDNA donor, and we supplement the mTeSR1 media with puromycin at the concentration of 0.5 µg/ml 1 week after transfection.

The information of reTALENs, gRNA and ssODNs used in this study are listed in Supplementary Tables S3 and S6.

Amplicon library preparation of the targeting regions

Cells were harvested 6 days after nucleofection and 0.1 µl of prepGEM tissue protease enzyme (ZyGEM) and 1 µl of prepGEM gold buffer (ZyGEM) were added to 8.9 µl of the 2 5×10^5 cells in the medium. In all, 1 µl of the reactions were then added to 9 µl of PCR mix containing 5 µl 2 × KAPA Hifi Hotstart Readymix (KAPA Biosystems) and 100 nM corresponding amplification primer pairs. Reactions were incubated at 95°C for 5 min followed by 15 cycles of 98°C, 20 s; 65°C, 20 s and 72°C, 20 s. To add

the Illumina sequence adaptor, 5 µl of reaction products were then added to 20 µl of PCR mix containing 12.5 µl of 2 × KAPA HIFI Hotstart Readymix (KAPA Biosystems) and 200 nM primers carrying Illumina sequence adaptors. Reactions were incubated at 95°C for 5 min followed by 25 cycles of 98°C, 20 s; 65°C, 20 s and 72°C, 20 s. PCR products were purified by QIAquick PCR purification kit, mixed at roughly the same concentration and sequenced with MiSeq Personal Sequencer. All the PCR primers can be found in the Supplementary Table S5.

Genome editing assessment system

We wrote a pipeline to analyze the genome engineering data. This pipeline is integrated in one single Unix module, which uses different tools such as R, BLAT and FASTX Toolkit.

Barcode splitting: Groups of samples were pooled together and sequenced using MiSeq 150 bp paired end (PE150) (Illumina Next Gen Sequencing) and later separated based on DNA barcodes using FASTX Toolkit.

Quality filtering: We trimmed nucleotides with lower sequence quality (phred score <20). After trimming, reads shorter than 80 nt were discarded.

Mapping: We used BLAT to map the paired reads independently to the reference genome and we generated .psl files as output.

Indel calling: We defined indels as the full-length reads containing two blocks of matches in the alignment. Only reads following this pattern in both paired end reads were considered. As a quality control, we required the indel reads to possess minimal 70 nt matching with the reference genome and both blocks to be at least 20 nt long. Size and position of indels were calculated by the positions of each block to the reference genome. Non-homologous end joining (NHEJ) has been estimated as the percentage of reads containing indels [see Equation (1)]. The majority of NHEJ event have been detected at the targeting site vicinity.

Homology-directed recombination (HDR) efficiency: Pattern matching (grep) within a 12 bp window centering over DSB was used to count specific signatures corresponding to reads containing the reference sequence, modifications of the reference sequence (2 bp intended mismatches) and reads containing only 1 bp mutation within the 2 bp intended mismatches [see Equation (1)].

Equation 1. Estimation of NHEJ and HDR

A = reads identical to the reference: XXXXXABXXXX

B = reads containing 2 bp mismatch programed by ssODN: XXXXXabXXXXX

C = reads containing only 1 bp mutation in the target site: such as XXXXXaBXXXXX or XXXXXAbXXXXX

D = reads containing indels as described above

$$\text{NHEJ efficiency} = \left(100 \times \frac{D}{A+B+C+D}\right)\%$$

$$\text{HDR efficiency} = \left(100 \times \frac{B}{A+B+C+D}\right)\%$$

The statistic analysis of the GEAS can be found in Supplementary Note S2.

Genotype screening of colonized hiPSCs

Human iPS cells on feeder-free cultures were pre-treated with mTesr-1 media supplemented with SMC4 (5 uM thiazovivin, 1 uM CHIR99021, 0.4 uM PD0325901, 2 uM SB431542) (23) for at least 2 h before fluorescence-activated cell sorting (FACS) sorting. Cultures were dissociated using Accutase (Millipore) and resuspended in mTesr-1 media supplemented with SMC4 and the viability dye ToPro-3 (Invitrogen) at concentration of 1 2 × 107 /ml. Live hiPS cells were single-cell sorted using a BD FACSAria II SORP UV (BD Biosciences) with 100 µm nozzle under sterile conditions into 96-well plates coated with irradiated CF-1 mouse embryonic fibroblasts (Global Stem). Each well contained hES cell medium (27) with 100 ng/ml recombinant human basic Fibroblast Growth Factor (Millipore) supplemented with SMC4 and 5 µg/ml fibronectin (Sigma). After sorting, plates were centrifuged at 70*g* for 3 min. Colony formation was seen 4 days post sorting, and the culture media was replaced with hES cell medium with SMC4. SMC4 can be removed from hES cell medium 8 days after sorting.

A few thousand cells were harvested 8 days after FACS and 0.1 µl of prepGEM tissue protease enzyme (ZyGEM) and 1 µl of prepGEM gold buffer (ZyGEM) were added to 8.9 µl of cells in the medium. The reactions were then added to 40 µl of PCR mix containing 35.5 ml of platinum 1.1× Supermix (Invitrogen), 250 nM of each dNTP and 400 nM primers. Reactions were incubated at 95°C for 3 min followed by 30 cycles of 95°C, 20 s; 65°C, 30 s and 72°C, 20 s. Products were Sanger sequenced using either one of the PCR primers (Supplementary Table S5), and sequences were analyzed using DNASTAR (DNASTAR).

Immunostaining and teratoma assays of hiPSCs

Cells were incubated in the KnockOut DMEM/F-12 medium at 37°C for 60 min using the following antibody: Anti-SSEA-4 PE (Millipore) (1: 500 diluted); Tra-1-60 (BD Pharmingen) (1:100 diluted). After the incubation, cells were washed three times with KnockOut DMEM/F-12 and imaged on the Axio Observer Z.1 (ZIESS).

To conduct teratoma formation analysis, we harvested human iPSCs using collagenase type IV (Invitrogen) and resuspended the cells into 200 µl of Matrigel and injected intramuscularly into the hind limbs of Rag2gamma knockout mice. Teratomas were isolated and fixed in formalin between 4 and 8 weeks after the injection. The teratomas were subsequently analyzed by hematoxylin and eosin staining.

RESULTS

ReTALENs target genomic loci effectively in human somatic and stem cells

TALEs have proven to be a powerful and easy-to-design tool for targeted genome manipulation in multiple cell

the Illumina sequence adaptor, 5 µl of reaction products were then added to 20 µl of PCR mix containing 12.5 µl of 2 × KAPA HIFI Hotstart Readymix (KAPA Biosystems) and 200 nM primers carrying Illumina sequence adaptors. Reactions were incubated at 95°C for 5 min followed by 25 cycles of 98°C, 20 s; 65°C, 20 s and 72°C, 20 s. PCR products were purified by QIAquick PCR purification kit, mixed at roughly the same concentration and sequenced with MiSeq Personal Sequencer. All the PCR primers can be found in the Supplementary Table S5.

Genome editing assessment system

We wrote a pipeline to analyze the genome engineering data. This pipeline is integrated in one single Unix module, which uses different tools such as R, BLAT and FASTX Toolkit.

Barcode splitting: Groups of samples were pooled together and sequenced using MiSeq 150 bp paired end (PE150) (Illumina Next Gen Sequencing) and later separated based on DNA barcodes using FASTX Toolkit.

Quality filtering: We trimmed nucleotides with lower sequence quality (phred score <20). After trimming, reads shorter than 80 nt were discarded.

Mapping: We used BLAT to map the paired reads independently to the reference genome and we generated .psl files as output.

Indel calling: We defined indels as the full-length reads containing two blocks of matches in the alignment. Only reads following this pattern in both paired end reads were considered. As a quality control, we required the indel reads to possess minimal 70 nt matching with the reference genome and both blocks to be at least 20 nt long. Size and position of indels were calculated by the positions of each block to the reference genome. Non-homologous end joining (NHEJ) has been estimated as the percentage of reads containing indels [see Equation (1)]. The majority of NHEJ event have been detected at the targeting site vicinity.

Homology-directed recombination (HDR) efficiency: Pattern matching (grep) within a 12 bp window centering over DSB was used to count specific signatures corresponding to reads containing the reference sequence, modifications of the reference sequence (2 bp intended mismatches) and reads containing only 1 bp mutation within the 2 bp intended mismatches [see Equation (1)].

Equation 1. Estimation of NHEJ and HDR

A = reads identical to the reference: XXXXXABXXXX

B = reads containing 2 bp mismatch programed by ssODN: XXXXXabXXXXX

C = reads containing only 1 bp mutation in the target site: such as XXXXXaBXXXXX or XXXXXAbXXXXX

D = reads containing indels as described above

$$\text{NHEJ efficiency} = \left(100 \times \frac{D}{A+B+C+D}\right)\%$$

$$\text{HDR efficiency} = \left(100 \times \frac{B}{A+B+C+D}\right)\%$$

The statistic analysis of the GEAS can be found in Supplementary Note S2.

Genotype screening of colonized hiPSCs

Human iPS cells on feeder-free cultures were pre-treated with mTesr-1 media supplemented with SMC4 (5 uM thiazovivin, 1 uM CHIR99021, 0.4 uM PD0325901, 2 uM SB431542) (23) for at least 2 h before fluorescence-activated cell sorting (FACS) sorting. Cultures were dissociated using Accutase (Millipore) and resuspended in mTesr-1 media supplemented with SMC4 and the viability dye ToPro-3 (Invitrogen) at concentration of 1 2 × 107 /ml. Live hiPS cells were single-cell sorted using a BD FACSAria II SORP UV (BD Biosciences) with 100 µm nozzle under sterile conditions into 96-well plates coated with irradiated CF-1 mouse embryonic fibroblasts (Global Stem). Each well contained hES cell medium (27) with 100 ng/ml recombinant human basic Fibroblast Growth Factor (Millipore) supplemented with SMC4 and 5 µg/ml fibronectin (Sigma). After sorting, plates were centrifuged at 70*g* for 3 min. Colony formation was seen 4 days post sorting, and the culture media was replaced with hES cell medium with SMC4. SMC4 can be removed from hES cell medium 8 days after sorting.

A few thousand cells were harvested 8 days after FACS and 0.1 µl of prepGEM tissue protease enzyme (ZyGEM) and 1 µl of prepGEM gold buffer (ZyGEM) were added to 8.9 µl of cells in the medium. The reactions were then added to 40 µl of PCR mix containing 35.5 ml of platinum 1.1× Supermix (Invitrogen), 250 nM of each dNTP and 400 nM primers. Reactions were incubated at 95°C for 3 min followed by 30 cycles of 95°C, 20 s; 65°C, 30 s and 72°C, 20 s. Products were Sanger sequenced using either one of the PCR primers (Supplementary Table S5), and sequences were analyzed using DNASTAR (DNASTAR).

Immunostaining and teratoma assays of hiPSCs

Cells were incubated in the KnockOut DMEM/F-12 medium at 37°C for 60 min using the following antibody: Anti-SSEA-4 PE (Millipore) (1: 500 diluted); Tra-1-60 (BD Pharmingen) (1:100 diluted). After the incubation, cells were washed three times with KnockOut DMEM/F-12 and imaged on the Axio Observer Z.1 (ZIESS).

To conduct teratoma formation analysis, we harvested human iPSCs using collagenase type IV (Invitrogen) and resuspended the cells into 200 µl of Matrigel and injected intramuscularly into the hind limbs of Rag2gamma knockout mice. Teratomas were isolated and fixed in formalin between 4 and 8 weeks after the injection. The teratomas were subsequently analyzed by hematoxylin and eosin staining.

RESULTS

ReTALENs target genomic loci effectively in human somatic and stem cells

TALEs have proven to be a powerful and easy-to-design tool for targeted genome manipulation in multiple cell

lines and organisms (2,13 15, 28 30). Several strategies have been developed to assemble the repetitive TALE RVD array sequences (7 10). However, once assembled, the TALE sequence repeats remain unstable, which limits the wide utility of this tool, especially for viral gene delivery vehicles (11,31). We thus thought that complete elimination of repeats would not only enable faster and simple synthesis of extended TALE RVD arrays but also address this important post-synthesis problem.

To eliminate repeats, we computationally evolved the nucleotides sequence of TALE RVD arrays to minimize the number of sequence repeats while maintaining the amino acid composition. Re-coded TALE (Re-TALEs) encoding 16 tandem RVD DNA recognition monomers, plus the final half RVD repeat, are devoid of any 12 bp repeats (Supplementary Figure S1a). Notably, this level of recoding is sufficient to allow PCR amplification of any specific monomer or sub-section from a full-length re-TALE construct (Supplementary Figure S1b). The improved design of re-TALEs makes it possible to order them directly from gene synthesis companies using standard DNA synthesis technology (32), without incurring the additional costs or procedures associated with repeat-heavy sequences. Furthermore, the recoded sequence design also enabled us to efficiently assemble re-TALE constructs using a modified isothermal assembly reaction ('Materials and Methods' section, Supplementary Note S1, Supplementary Figure S2).

We next sought to test the function of reTALEN in comparison with the corresponding non-recoded TALEN in human cells. To this end, we used a HEK 293 cell line containing a GFP reporter cassette carrying a frame-shifting insertion as previously described (33) (Figure 1a). Delivery of TALENs or reTALENs targeting the insertion sequence, together with a promoter-less GFP donor construct, leads to DSB-induced HDR repair of the GFP cassette so that GFP repair efficiency can be used to evaluate the nuclease cutting efficiency (34). We found that reTALENs induced GFP repair in 1.4% of the transfected cells, similar to that achieved by TALENs (1.2%) (Figure 1b). We further tested the activity of reTALENs at the AAVS1 locus in PGP1 hiPSCs (Figure 1c) and successfully recovered cell clones containing specific insertions (Figure 1d and e), confirming that reTALENs are active in both somatic and pluripotent human cells.

We then confirmed that the elimination of repeats would enable us to generate functional lentivirus with a re-TALE cargo. Specifically, we packaged lentiviral particles encoding re-TALE-2 A-GFP and obtained lentiviral particles with tittering of 1.3×10^6 We then tested the activity of the re-TALE-TF encoded by viral particles by transfecting a mCherry reporter into a pool of lenti-reTALE-2 A-GFP-infected 293 T cells. The 293 T cells transduced by lenti-re-TALE-TF showed 36× reporter expression activation compared with the reporter only negative (Supplementary Figure S3a c). We further checked the sequence integrity of the re-TALE-TF in the lentiviral infected cells and detected full-length reTALEs in all 10 of the clones tested (Supplementary Figure S3d).

Comparison of ReTALEs and Cas9-gRNA efficiency in hiPSCs with GEAS

To compare the editing efficiencies of re-TALENs versus Cas9-gRNA in hiPSCs, we developed a next-generation sequencing platform to precisely pinpoint and quantify both NHEJ and HDR gene-editing events, which we refer to as Genome Editing Assessment System (GEAS). First, we designed and constructed a re-TALEN pair and a Cas9-gRNA, both targeting the upstream region of CCR5 (re-TALEN, Cas9-gRNA pair #3 in Supplementary Table S3), along with a 90 nt ssODN donor identical to the target site except for a 2 bp mismatch (Figure 2a). We then transfected the nuclease constructs and donor ssODN into hiPSCs. To precisely quantitate the gene-editing efficiency, we conducted paired-end deep sequencing on the target genomic region 3 days after transfection. HDR efficiency was measured by the percentage of reads containing the precise 2 bp mismatch. NHEJ efficiency was measured by the percentage of reads carrying indels.

Delivery of the ssODN alone into hiPSCs resulted in minimal HDR and NHEJ rates, whereas delivery of the re-TALENs and the ssODN led to efficiencies of 1.7% HDR and 1.2% NHEJ (Figure 2b). The introduction of the Cas9-gRNA with the ssODN led to 1.2% HDR and 3.4% NHEJ efficiencies. Notably, the rate of genomic deletions and insertions peaked in the middle of the spacer region between the two reTALENs binding site, but peaked 3 4 bp upstream of the protospacer associated motif (PAM) sequence of Cas9-gRNA-targeting site (Figure 2b) as would be expected from the fact that DSBs take place in these regions. We observed a median genomic deletion size of 6 bp and insertion size of 3 bp generated by the re-TALENs and a median deletion size of 7 bp and insertion of 1 bp by the Cas9-gRNA (Figure 2b), consistent with DNA lesion patterns usually generated by NHEJ (4). Several analyses of our next-generation sequencing platform revealed that GEAS can detect HDR detection rates as low as 0.007%, which is both highly reproducible (coefficient of variation between replicates = ± 15% × measured efficiency) and 400× more sensitive than most commonly used mismatch sensitive endonuclease assays (Supplementary Figure S4).

After confirming the reliability of GEAS, we next sought to test the scalability of our tools by building and assessing re-TALEN pairs and Cas9-gRNAs targeted to 15 sites at the CCR5 genomic locus (Figure 2c, Supplementary Table S3). Anticipating that editing efficiency might depend on chromatin state, these sites were selected to represent a wide range of DNaseI sensitivities (35). The nuclease constructs were transfected with the corresponding ssODNs donors (Supplementary Table S3) into PGP1 hiPSCs. Six days after transfection, we profiled the genome-editing efficiencies at these sites (Supplementary Table S4). For 13 of 15 re-TALEN pairs with ssODN donors, we detected NHEJ and HDR at levels above our statistical detection thresholds, with an average NHEJ efficiency of 0.4% and an average HDR efficiency of 0.6% (Figure 2c). In addition, a statistically significant positive correlation ($r^2 = 0.81$) was found

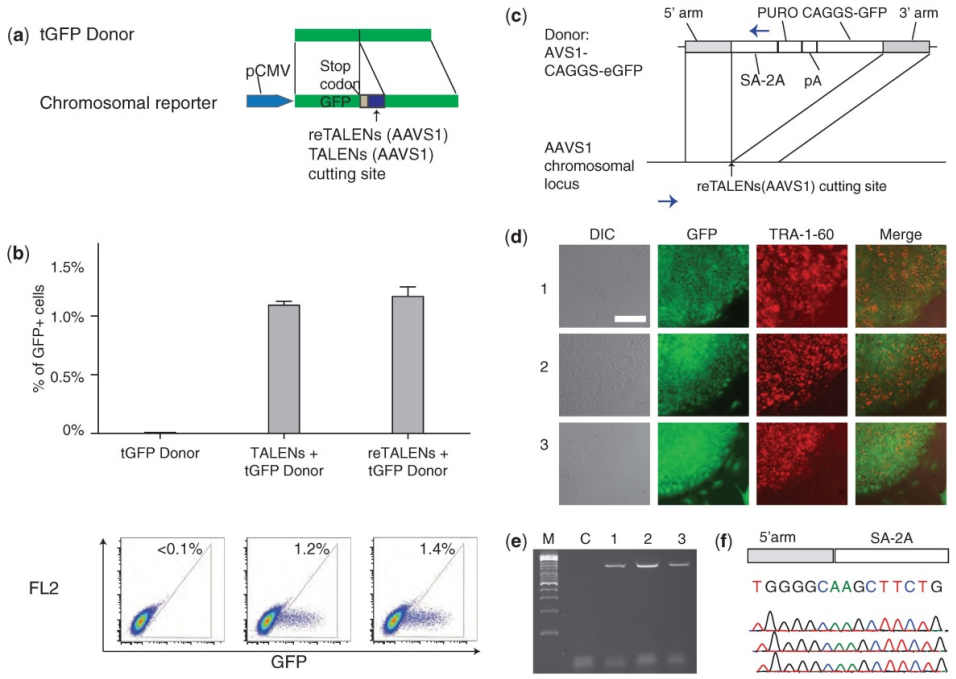

Figure 1. Functional tests of re-TALENs in human somatic and stem cells. (**a**) Schematic representation of experimental design for testing genome targeting efficiency. A genomically integrated GFP-coding sequence is disrupted by the insertion of a stop codon and a 68 bp genomic fragment derived from the AAVS1 locus (bottom). Restoration of the GFP sequence by nuclease-mediated homologous recombination with tGFP donor (top) results in GFP+ cells that can be quantitated by FACS. Re-TALENs and TALENs target identical sequences within AAVS1 fragments. (**b**) Bar graph depicting GFP+ cell percentage introduced by tGFP donor alone, TALENs with tGFP donor and re-TALENs with tGFP donor at the target locus, as measured by FACS ($N = 3$, error bar = SD). Representative FACS plots are shown later in the text. (**c**) Schematic overview depicting the targeting strategy for the native AAVS1 locus. The donor plasmid, containing splicing acceptor (SA)- 2 A (self-cleaving peptides), puromycin resistant gene (PURO) and GFP were described before (14). The locations of PCR primers used to detect successful editing events are depicted as blue arrows. (**d**) Successfully targeted clones of PGP1 hiPSCs were selected with puromycin (0.5 μg/ml) for 2 weeks. Microscopy images of three representative GFP+ clones are shown. Cells were also stained for the pluripotency markers TRA-1-60. Scale bar: 200 μm. (**e**) PCR assays performed on these the monoclonal GFP+ hiPSC clones demonstrated successful insertions of the donor cassettes at the AAVS1 site (lanes 1–3), whereas plain hiPSCs show no evidence of successful insertion (lane C). (**f**) Sanger sequencing of the PCR amplicon from the three targeted hiPSC colonies confirmed that the expected DNA bases at the genome-insertion boundary is present.

between HR and NHEJ efficiency at the same targeting loci ($P < 1 \times 10^{-4}$) (Supplementary Figure S5a), suggesting that DSB generation, the common upstream step of both HDR and NHEJ, is a rate-limiting step for reTALEN-mediated genome editing.

In contrast, all 15 Cas9-gRNA pairs showed significant levels of NHEJ and HR, with an average NHEJ efficiency of 3% and an average HDR efficiency of 1.0% (Figure 2c). In addition, a positive correlation was also detected between the NHEJ and HDR efficiency introduced by Cas9-gRNA (Supplementary Figure S5b) ($r^2 = 0.52$, $P = 0.003$), consistent with what we had observed with our reTALENs. The NHEJ efficiency achieved by Cas9-gRNA was significantly higher than that achieved by reTALENs (*t*-test, paired-end, $P = 0.02$). Interestingly, we observed a moderate but statistically significant correlation between NHEJ efficiency and the melting temperature of the gRNA targeting sequence (Supplementary Figure S5c) ($r^2 = 0.28$, $P = 0.04$), suggesting that the strength of base pairing between the gRNA and its genomic target could explain as much as 28% of the variation in the efficiency of Cas9-gRNA-mediated DSB generation. Even though Cas9-gRNA produced NHEJ levels at an average of seven times higher than the corresponding reTALEN, Cas9-gRNA only achieved HDR levels (average = 1.0%) similar to that of the corresponding reTALENs (average = 0.6%), suggesting either that the ssODN concentration at the DSB is the limiting factor for HDR or that the genomic break structure created by the Cas9-gRNA is not

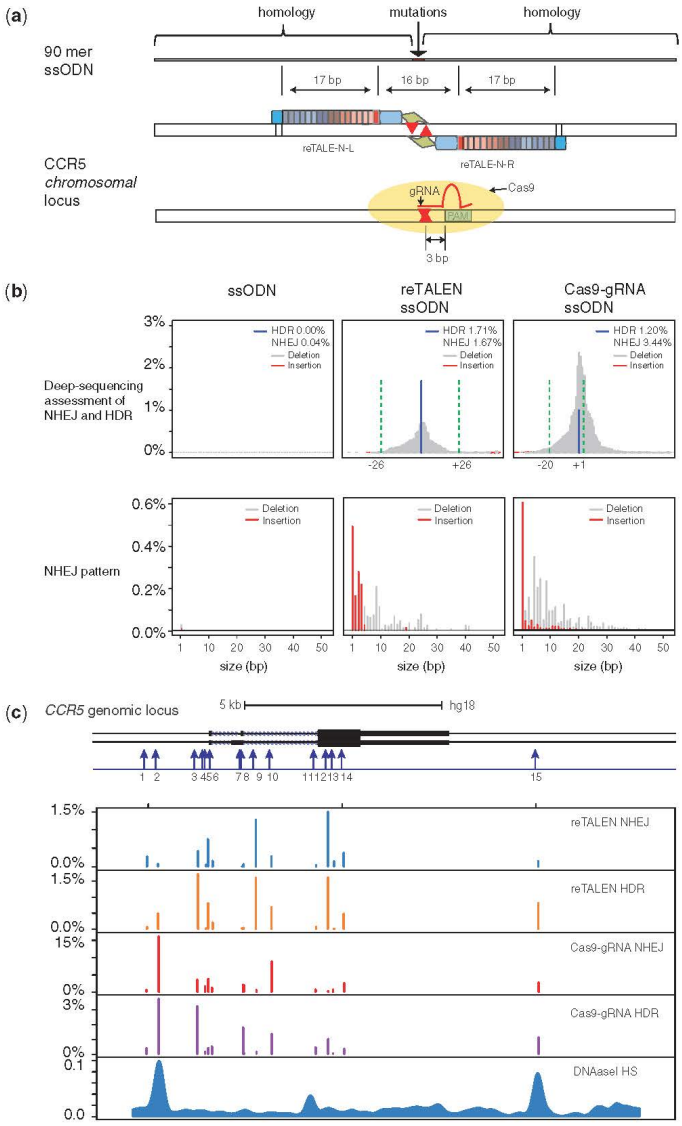

Figure 2. Comparison of reTALENs and Cas9-gRNAs genome targeting efficiency on *CCR5* in iPSCs. (**a**) Schematic representation of genome engineering experimental design. At the re-TALEN pair or Cas9-gRNA targeting site, a 90 mer ssODN carrying a 2 bp mismatch against the genomic DNA was delivered along with the reTALEN or Cas9-gRNA constructs into PGP1 hiPSCs. The cutting sites of the nucleases are depicted as red arrows in the figure. (**b**) Deep-sequencing analysis of HDR and NHEJ efficiencies for re-TALEN pairs (CCR5 #3) and ssODN, or the Cas9-gRNA and ssODN. Alterations in the genome of hiPSCs were analyzed from high-throughput sequence data by GEAS. Top: HDR was quantified from the fraction of reads that contained a 2 bp point mutation built into the center of the ssODN (blue), and NHEJ activity was quantified from the fraction of deletions (gray)/Insertions (red) at each specific position in the genome. For the reTALEN and ssODN graphs, we plot green dashed lines to mark

(continued)

favorable for effective HDR (see 'Discussion' section). Of note, within our data, we did not observe any correlation between DNaseI HS and the genome targeting efficiencies achieved by either method (Supplementary Figure S6).

Optimization of ssODN donor design for HDR

Although ssODNs have been found to be effective as donor DNA in genome editing [see earlier in the text, (21,22)], many questions remain regarding how to optimize their design. Having compared the efficiencies of reTALEN and Cas9-gRNA nucleases, we next developed strategies for the design of highly performing ssODNs in hiPSCs.

We first designed a set of ssODNs donors of different lengths (50 170 nt), all carrying the same 2 bp mismatch in the middle of the spacer region of the CCR5 re-TALEN pair #3 target sites. HDR efficiency was observed to vary with ssODN length, and an optimal HDR efficiency of ~1.8% was observed with a 90 nt ssODN, whereas longer ssODNs decreased HDR efficiency (Figure 3a). As longer homology regions improve HDR rates when dsDNA donors are used with nucleases (36), possible reasons for this result may be that ssODNs are used in an alternative genome repair process; longer ssODNs are less available to the genome repair apparatus or that longer ssODNs incur negative effects that offset any improvements gained by longer homology, compared with dsDNA donors (37). Yet, if either of the first two reasons were the case, then NHEJ rates should either be unaffected or would increase with longer ssODNs because NHEJ repair does not involve the ssODN donor. However, NHEJ rates were observed to decline along with HDR (Figure 3a), suggesting that the longer ssODNs present offsetting effects. Possible hypotheses would be that longer ssODNs are toxic to the cell (38) or that transfection of longer ssODNs saturates the DNA processing machinery, thereby causing decreased molar DNA uptake and reducing the capacity of the cells to take up or express re-TALEN plasmids.

Next, we examined how rate of incorporation of a mismatch carried by the ssODN donor varies with its distance to the DSB. To this end, we designed a series of 90 nt ssODNs all possessing the same 2 bp mismatch (A) in the center of the spacer region of re-TALEN pair #3. Each ssODN also contained a second 2 bp mismatch (B) at varying distances from the center (Figure 3b). An ssODN possessing only the center 2 bp mismatch was used as a control. Each of these ssODNs was introduced individually with re-TALEN pair #3, and the outcomes were analyzed with GEAS. We found that overall HDR as measured by the rate at which the A mismatch was incorporated (A only or A + B) decreased as the B mismatches became farther from the center (Figure 3b, Supplementary Figure S7a). The higher overall HDR rate observed when B is only 10 bp away from A may reflect a lesser need for annealing of the ssODN against genomic DNA immediately proximal to the dsDNA break.

For each distance of B from A, a fraction of HDR events only incorporated the A mismatch, whereas another fraction incorporated both A and B mismatches [Figure 3b (A only and A + B)], These two outcomes may be due to gene conversion tracts (39) along the length of the ssDNA oligo, whereby incoporation of A + B mismatches resulted from long conversion tracts that extended beyond the B mismatch, and incorporation of the A-only mismatch resulted from shorter tracts that did not reach B. Under this interpretation, we estimated a distribution of gene conversion lengths in both directions along the ssODN (Supplementary Figure S7b). The estimated distribution implies that gene conversion tracts progressively become less frequent as their lengths increase, a result similar to gene conversion tract distributions seen with dsDNA donors (39), but on a highly compressed distance scale of tens of bases for the ssDNA donor versus hundreds of bases for dsDNA donors. Consistent with this result, an experiment with a ssODN containing three pairs of 2 bp mismatches spaced at intervals of 10 nt on either side of the central 2 bp mismatch 'A's gave rise to a pattern in which A alone was incorporated 86% of the time, with multiple B mismatches incorporated at other times (Supplementary Figure S7c). Although the numbers of B only incorporation events were too low to estimate a distribution of tract lengths <10 bp, it is clear that the short tract region within 10 bp of the nuclease site predominates (Supplementary Figure S7b). Finally, in all of our experiments with single B mismatches, we see a small fraction of B-only incorporation events (0.04 0.12%) that is roughly constant across all B distances from A. The nature of these events is unclear.

Furthermore, we tested how far the ssODN donor can be placed from the re-TALEN-induced dsDNA break and

Figure 2. Continued
the outer boundary of the re-TALEN pair's binding sites, which are at positions −26 bp and +26 bp relative to the center of the two re-TALEN-binding sites. For Cas9-gRNA and ssODN graphs, the green dashed lines mark the outer boundary of the gRNA targeting site, which are at positions −20 and −1 bp relative to the Protospacer Associated Motif sequence. Bottom: Deletion/Insertion size distribution in hiPSCs analyzed from the entire NHEJ population with treatments indicated earlier in the text. (c) The genome-editing efficiency of re-TALENs and Cas9-gRNAs targeting *CCR5* in PGP1 hiPSCs. Top: schematic representation of the targeted genome-editing sites in *CCR5*. The 15 targeting sites are illustrated by blue arrows later in the text. For each site, cells were co-transfected with a pair of re-TALENs and their corresponding ssODN donor carrying 2 bp mismatches against the genomic DNA. Genome-editing efficiencies were assayed 6 days after transfection. Similarly, we transfected 15 Cas9-gRNAs with their corresponding ssODNs individually into PGP1-hiPSCs to target the same 15 sites and analyzed the efficiency 6 days after transfection. Bottom: the genome-editing efficiency of re-TALENs and Cas9-gRNAs targeting CCR5 in PGP1 hiPSCs. Panels 1 and 2 indicate NHEJ and HDR efficiencies mediated by reTALENs. Panels 3 and 4 indicate NHEJ and HDR efficiencies mediated by Cas9-gRNAs. NHEJ rates were calculated by the frequency of genomic alleles carrying deletions or insertions at the targeting region; HDR rates were calculated by the frequency of genomic alleles carrying 2 bp mismatches. Panel 5, the DNaseI HS profile of a hiPSC cell line from ENCODE database (Duke DNase HS, iPS NIHi7 DS). Of note, the scales of different panels are different.

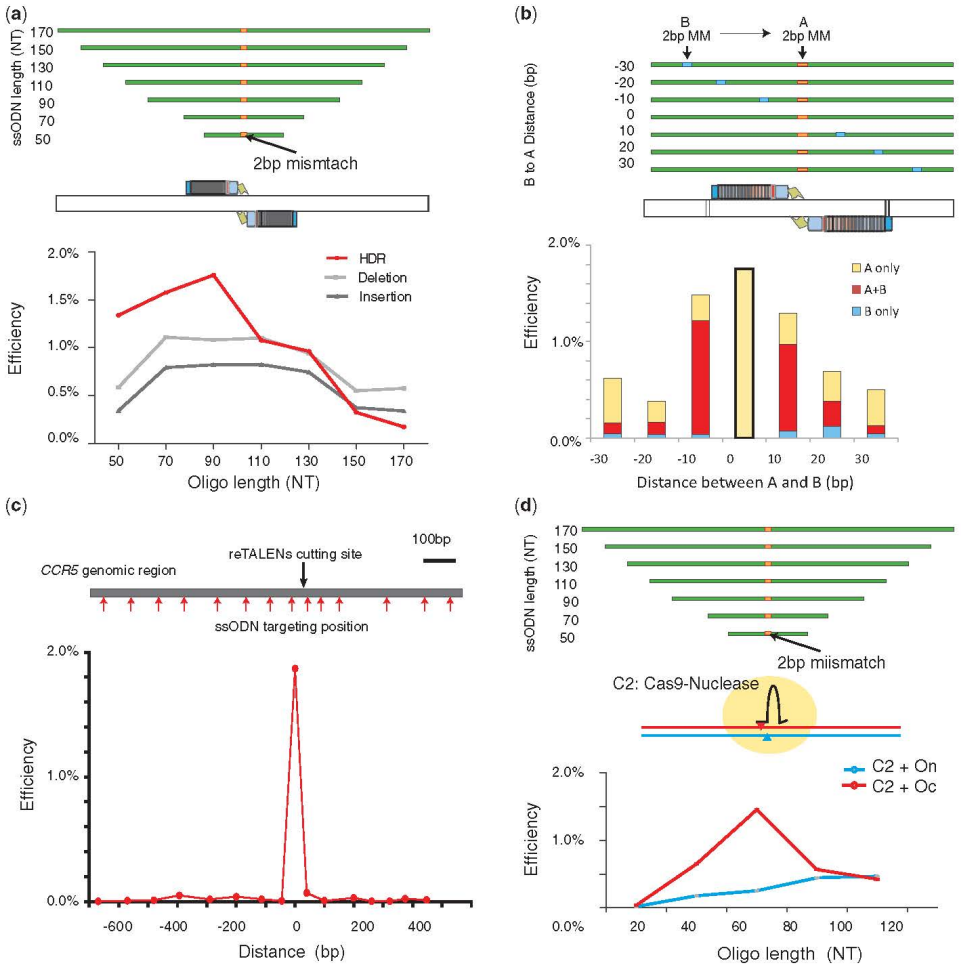

Figure 3. Study of functional parameters governing ssODN-mediated HDR with re-TALENs or Cas9-gRNAs in PGP1 hiPSCs. (**a**) PGP1 hiPSCs were co-transfected with re-TALENs pair (#3) and ssODNs of different lengths (50, 70, 90, 110, 130, 150 and 170 nt). All ssODNs possessed an identical 2 bp mismatch against the genomic DNA in the middle of their sequence. A 90 mer ssODN achieved optimal HDR in the targeted genome. The assessment of HDR, NHEJ-incurred deletion and insertion efficiency is described in the 'Materials and Methods' section. (**b**) 90 mer ssODNs corresponding to re-TALEN pair #3 each containing a 2 bp mismatch (A) in the center and an additional 2 bp mismatch (B) at different positions offset from A (where offsets varied from −30 to 30 bp) were used to test the effects of deviations from homology along the ssODN. Genome-editing efficiency of each ssODN was assessed in PGP1 hiPSCs. The bottom bar graph shows the incorporation frequency of A only, B only and A + B in the targeted genome. HDR rates decrease as the distance of homology deviations from the center increase (see text and Supplementary Figure S7a and b). (**c**) ssODNs targeted to sites with varying distances (−620~480 bp) away from the target site of re-TALEN pair #3 were tested to assess the maximum distance within which we can place ssODNs to introduce mutations. All ssODNs carried a 2 bp mismatch in the middle of their sequences. We observed minimal HDR efficiency (≤0.06%) when the ssODN mismatch was positioned 40 bp away from the middle of re-TALEN pair's binding site. (**d**) PGP1 hiPSCs were co-transfected with Cas9-gRNA (AAVS1) and ssODNs of different orientation (O_c: complement to gRNA; O_n: non-complement to gRNA) and different lengths (30, 50, 70, 90 and 110 nt). All ssODNs possessed an identical 2 bp mismatch against the genomic DNA in the middle of their sequence. A 70 mer O_c achieved optimal HDR in the targeted genome.

still observe incorporation. A set of 90 nt ssODNs with central 2 bp mismatches targeting a range of larger distances (−600 to +400 bp) away from the re-TALEN-induced dsDNA break site were tested. When the ssODNs matched ≥40 bp away, we observed >30× lower HDR efficiencies compared with the control ssODN positioned centrally over the cut region (Figure 3c). The low level of incorporation that was observed may be due to processes unrelated to the dsDNA cut, as seen in experiments in which genomes are altered by a ssDNA donor alone (38). Meanwhile, the low level of HDR present when the ssODN is ~40 bp away may be due to a combination of weakened homology on the mismatch-containing side of the dsDNA cut along with insufficient ssODN oligo length on the other side of the dsDNA break.

We similarly tested the ssODNs DNA donor design for Cas9-gRNA-mediated targeting. First, we constructed Cas9-gRNA (C_2) targeting the AAVS1 locus and designed ssODN donors of variable orientations (O_c: complementary to the gRNA and O_n: non-complementary to the gRNA) and lengths (30, 50, 70, 90 and 110 nt). We found O_c achieved better efficiency than O_n, with a 70 mer O_c achieving an optimal HDR rate of 1.5%. (Figure 3d) The same ssODN strand bias was detected using a Cas9-derived nickase (C_c: Cas9_D10A), despite the fact that the HDR efficiencies mediated by C_c with ssODN were significantly less than C_2 (*t*-test, paired-end, $P = 0.02$) (Supplementary Figure S8). Future investigation will further elucidate the factors that may contribute to this bias, including sequence bias, direction of transcription and replication.

hiPSC clonal isolation of corrected cells

GEAS revealed that re-TALEN pair #3 achieved precise genome editing with an efficiency of ~1% in hiPSCs, a level at which correctly edited cells can usually be isolated by screening clones. HiPSCs have poor viability as single cells, but recent advances in culture conditions have facilitated outgrowth of hiPSCs from single cells (23). We optimized these protocols along with a single-cell FACS sorting procedure to establish a robust platform for single hiPSCs sorting and maintenance, where hiPSC clones can be recovered with survival rates of >25% (see 'Materials and Methods' section). We combined this method with a rapid and efficient genotyping system where we can conduct chromosomal DNA extraction and targeted genome amplification in 1-h single tube reactions, enabling large-scale genotyping of edited hiPSCs. Together, these methods comprise a pipeline for robustly obtaining genome-edited hiPSCs without selection.

To demonstrate this system (Figure 4a), we first transfected PGP1 hiPSCs with a pair of re-TALENs and an ssODN targeting CCR5 at site #3 (Supplementary Table S3), and we performed GEAS with a portion of the transfected cells, finding an HDR frequency of 1.7% (Figure 4b). This information, along with the 25% recovery of sorted single-cell clones, allowed us to estimate that we could obtain at least one correctly edited clone from five 96-well plates with Poisson probability 98% (assuming $\mu = 0.017 \times 0.25 \times 96 \times 5 \times 2$). Six days after transfection, hiPSCs were FACS sorted and 8 days after sorting, 100 hiPSC clones were screened. Sanger sequencing revealed that 2 of 100 of these unselected hiPSC colonies contained a heterozygous genotype possessing the 2 bp mutation introduced by the ssODN donor (Figure 4c). The targeting efficiency of 1% ($1\% = 2/2 \times 100$, 2 mono-allelic corrected clones out of 100 cell screened) was consistent with the next-generation sequencing analysis (1.7%) (Figure 4b). The pluripotency of the resulting hiPSCs was confirmed with immunostaining for SSEA4 and TRA-1-60 (Figure 4d). The successfully targeted hiPSCs clones were able to generate mature teratomas with features of all three germ layers (Figure 4e).

DISCUSSION

Here, we developed and demonstrated several improvements to the design and assessment of genome-editing reagents and demonstrated a streamlined method for efficient human stem cell editing. We first developed reTALENs, which simplify TALEN construction and enables the generation of functional lenti-viruses, which are important tools for delivering the reagents into many cell types and animals (33).

We then built a highly sensitive GEAS assay system to easily and precisely pinpoint and quantify HDR and NHEJ events in hiPSCs. In comparison with other methods of assessing design parameters for genome-editing, our genome-editing assessment tool provides simultaneous information on rates of HDR, NHEJ and other mutagenic processes through a single experimental and statistical analysis method versus performing different experiments and applying separate statistical methods for each individually. In the course of this study, we routinely pooled ~50 barcoded samples together and used the Illumina MiSeq system to obtain the sequence data, which was analyzed with our genome-editing assessment software. Currently, MiSeq can deliver ~20 Million paired-end 150 bp reads within 27 h so that up to 200 sample-barcoded targeting regions can be covered with ~100 K reads each at a cost of approximately $5 per sample. If desired, sample throughput can be traded off for higher sensitivity by allotting more reads per sample and processing fewer samples. Software and documentation for our genome-editing assessment system is available to provide researchers with the means to improve and standardize their genome-editing methods and extend them to additional cell lines and types.

Using our developed reTALENs, Cas9-gRNAs and GEAS method, we compared HDR and NHEJ efficiencies across 15 pairs of reTALENs and Cas9-gRNA (Supplementary Table S3 and S4) on the CCR5 locus. We found 13/15 of reTALEN pairs and all 15 Cas9-gRNAs exhibited detectable activities in hiPSCs, suggesting that both nuclease platforms serve as robust tools for genome editing. We confirmed the activity of the two failed reTALEN pairs in K562 cells and found 4 and 3% cutting efficiency, respectively, suggesting some

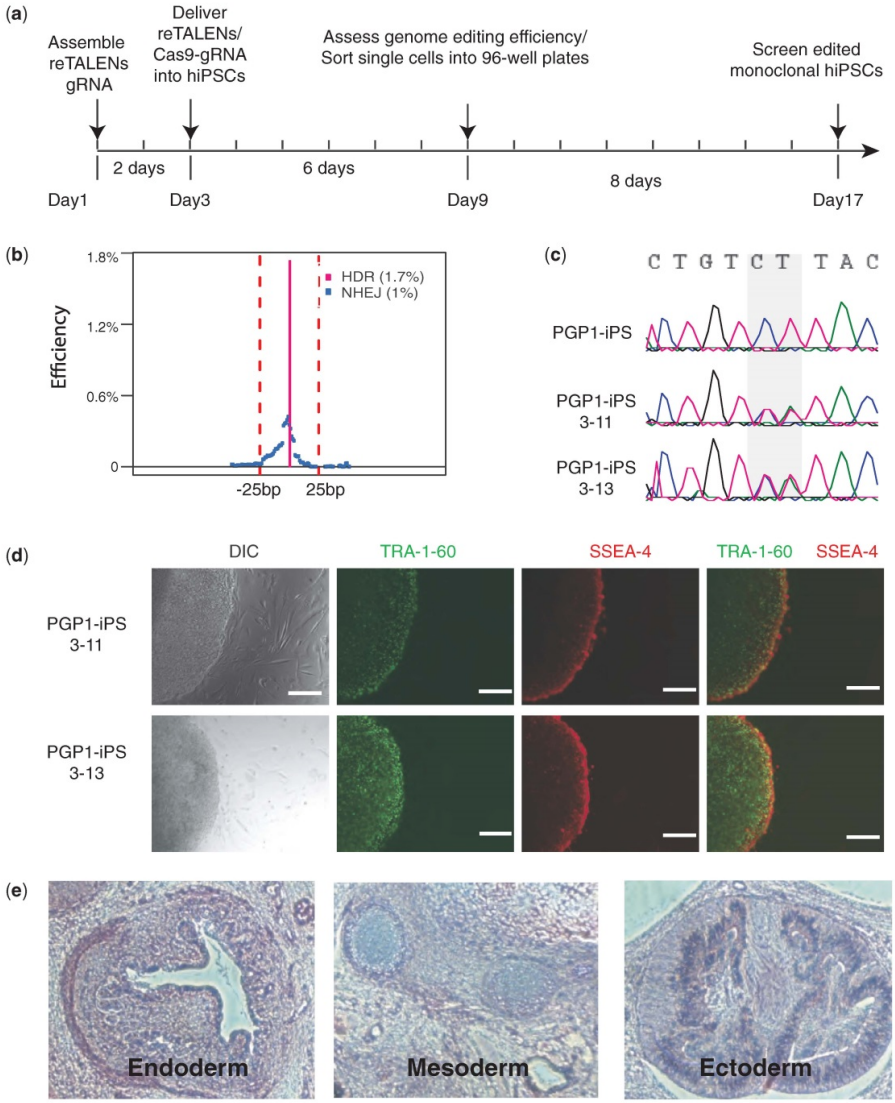

Figure 4. Using re-TALENs and ssODNs to obtain monoclonal genome-edited hiPSC without selection. (**a**) Timeline of the experiment. (**b**) Genome engineering efficiency of re-TALENs pair and ssODN (#3) assessed by the NGS platform described in Figure 2b. (**c**) Sanger sequencing results of monoclonal hiPSC colonies after genome editing. Of note, the 2 bp heterogeneous genotype (CT/CT→TA/CT) was successfully introduced into the genome of PGP1-iPS-3-11, PGP1-iPS-3-13 colonies. (**d**) Immunofluorescence staining of targeted PGP1-iPS-3-11. Cells were stained for the pluripotency markers Tra-1-60 and SSEA4. (**e**) Hematoxylin and eosin staining of teratoma sections generated from monoclonal PGP1-iPS-3-11 cells.

pertinent factors in hiPSCs, such as heterochromatin of methylation at the targeting regions make them resistant to reTALEN activity. In addition, we found that Cas9-gRNA induced on average 7 8× greater NHEJ rates than reTALEN, similar to recent reports (15). The effective concentration of Cas9-gRNA complexes or the intrinsic enzyme kinetics may contribute to this difference. Surprisingly, we did not see an equivalent increase of HDR with Cas9-gRNA and ssODN. Although ssODN concentration may reach saturating levels during construct delivery, ssODN availability at the DSB might be the limiting factor for HDR. Future studies using Cas9-gRNA nickases to generate defined DSB resections more favorable for HDR (36) can be conducted to test this hypothesis and further increase HDR efficiencies. Although we have compared the genome-targeting efficiencies achieved by reTALENs and Cas9-gRNA, a critical issue will also be to determine the generation of off-target mutations. It will be imperative to address the specificity of both targeting tools to improve the potential of hiPSCs genome engineering.

Finally, we demonstrated a streamlined pipeline for obtaining scarlessly edited human stem cells using our reagents. The pipeline comprises of the following: (i) reTALEN or Cas9-gRNA synthesis; (ii) prospective screening of reagents using GEAS; and (iii) high-throughput isolation of hiPSC clones. We note that with 1% HDR efficiency, it is feasible to generate isogeneic hiPSCs with mono-allelic mutations, which will facilitate hiPSC-based modeling of dominant alleles, allele-specific expression or X-linked mutations. However, targeting efficiencies must be improved to generate of homozygous mutations in hiPSCs. Other strategies such as transfection enrichment (15,17), or transient hypothermia (40), can be used together with our tools to achieve this goal. Last, we emphasize the versatility of our tools in that re-TALEs/Cas-gRNA can be engineered and used for other genomic-targeting technologies such as customized transcriptional factors and epigenetic modifiers, whereas GEAS can be applied to other gene-editing techniques, such as ZFNs, targeted nickases and meganucleases. We envision that our pipeline of efficiently generating scarlessly engineered human stem cells will allow the research community to resolve the causal underpinnings of numerous important biological problems, as well as to precisely engineer hiPSCs and other cell lines for autologous cell therapy.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank all the Church laboratory members for suggestion and support; and Daniel Gibson (J. Craig Venter Institute) for providing advice on assembly reactions.

FUNDING

National Human Genome Research Institute (NHGRI) Center for Excellence in Genomics Science [P50 HG005550, G.M.C.]; funded by Human Frontiers Science Program long-term fellowship (to M.G.). Funding for open access charge: NHGRI Center for Excellence in Genomics Science [P50 HG005550, G.M.C.].

Conflict of interest statement. G.M.C., L.Y, M.G. and J.Y. are inventors on a patent application describing the reTALE concept and assembly method.

REFERENCES

1. Carroll,D. (2011) Genome engineering with zinc-finger nucleases. *Genetics*, **188**, 773–782.
2. Wood,A.J., Lo,T.W., Zeitler,B., Pickle,C.S., Ralston,E.J., Lee,A.H., Amora,R., Miller,J.C., Leung,E., Meng,X. *et al.* (2011) Targeted genome editing across species using ZFNs and TALENs. *Science*, **333**, 307.
3. Perez-Pinera,P., Ousterout,D.G. and Gersbach,C.A. (2012) Advances in targeted genome editing. *Curr. Opin. Chem. Biol.*, **16**, 268–277.
4. Symington,L.S. and Gautier,J. (2011) Double-strand break end resection and repair pathway choice. *Annu. Rev. Genet.*, **45**, 247–271.
5. Urnov,F.D., Miller,J.C., Lee,Y.-L., Beausejour,C.M., Rock,J.M., Augustus,S., Jamieson,A.C., Porteus,M.H., Gregory,P.D. and Holmes,M.C. (2005) Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature*, **435**, 646–651.
6. Boch,J., Scholze,H., Schornack,S., Landgraf,A., Hahn,S., Kay,S., Lahaye,T., Nickstadt,A. and Bonas,U. (2009) Breaking the code of DNA binding specificity of TAL-type III effectors. *Science*, **326**, 1509–1512.
7. Cermak,T., Doyle,E.L., Christian,M., Wang,L., Zhang,Y., Schmidt,C., Baller,J.A., Somia,N.V., Bogdanove,A.J. and Voytas,D.F. (2011) Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res.*, **39**, e82.
8. Briggs,A.W., Rios,X., Chari,R., Yang,L., Zhang,F., Mali,P. and Church,G.M. (2012) Iterative capped assembly: rapid and scalable synthesis of repeat-module DNA such as TAL effectors from individual monomers. *Nucleic Acids Res.*, **40**, e117.
9. Zhang,F., Cong,L., Lodato,S., Kosuri,S., Church,G.M. and Arlotta,P. (2011) LETTErs Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. *Nat. Biotechnol.*, **29**, 149–154.
10. Reyon,D., Tsai,S.Q., Khayter,C., Foden,J.A., Sander,J.D. and Joung,J.K. (2012) FLASH assembly of TALENs for high-throughput genome editing. *Nat. Biotechnol.*, **30**, 460–465.
11. Holkers,M., Maggio,I., Liu,J., Janssen,J.M., Miselli,F., Mussolino,C., Recchia,A., Cathomen,T. and Gonçalves,M.A. (2012) Differential integrity of TALE nuclease genes following adenoviral and lentiviral vector gene transfer into human cells. *Nucleic Acids Res.*, **41**, e63.
12. Qiu,P., Shandilya,H., D'Alessio,J.M., O'Connor,K., Durocher,J. and Gerard,G.F. (2004) Mutation detection using Surveyor nuclease. *Biotechniques*, **36**, 702–707.
13. Miller,J.C., Tan,S., Qiao,G., Barlow,K.A., Wang,J., Xia,D.F., Meng,X., Paschon,D.E., Leung,E., Hinkley,S.J. *et al.* (2011) A TALE nuclease architecture for efficient genome editing. *Nat. Biotechnol.*, **29**, 143–148.
14. Hockemeyer,D., Wang,H., Kiani,S., Lai,C.S., Gao,Q., Cassady,J.P., Cost,G.J., Zhang,L., Santiago,Y., Miller,J.C. *et al.* (2011) Genetic engineering of human pluripotent cells using TALE nucleases. *Nat. Biotechnol.*, **29**, 731–734.
15. Ding,Q., Lee,Y., Schaefer,E.A.K., Peters,D.T., Veres,A., Kim,K., Kuperwasser,N., Motola,D.L., Meissner,T.B., Hendriks,W.T. *et al.* (2013) Resource A TALEN genome-editing system for

generating human stem cell-based disease models. *Cell Stem Cell*, **12**, 238–251.
16. Mali,P., Yang,L., Esvelt,K.M., Aach,J., Guell,M., DiCarlo,J.E., Norville,J.E. and Church,G.M. (2013) RNA-guided human genome engineering via Cas9. *Science*, **339**, 823–826.
17. Ding,Q., Regan,S.N., Xia,Y., Oostrom,L.A., Cowan,C.A. and Musunuru,K. (2013) Enhanced efficiency of human pluripotent stem cell genome editing through replacing TALENs with CRISPRs. *Cell Stem Cell*, **12**, 393–394.
18. Cong,L., Ran,F.A., Cox,D., Lin,S., Barretto,R., Habib,N., Hsu,P.D., Wu,X., Jiang,W., Marraffini,L.A. *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science*, **339**, 819–823.
19. Cho,S.W., Kim,S., Kim,J.M. and Kim,J.S. (2013) Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nat. Biotechnol.*, **31**, 230–232.
20. Hwang,W.Y., Fu,Y., Reyon,D., Maeder,M.L., Tsai,S.Q., Sander,J.D., Peterson,R.T., Yeh,J.-R.J. and Joung,J.K. (2013) Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nat. Biotechnol.*, **31**, 227–229.
21. Chen,F., Pruett-Miller,S.M., Huang,Y., Gjoka,M., Duda,K., Taunton,J., Collingwood,T.N., Frodin,M. and Davis,G.D. (2011) High-frequency genome editing using ssDNA oligonucleotides with zinc-finger nucleases. *Nat. Methods*, **8**, 753–755.
22. Soldner,F., Laganière,J., Cheng,A.W., Hockemeyer,D., Gao,Q., Alagappan,R., Khurana,V., Golbe,L.I., Myers,R.H., Lindquist,S. *et al.* (2011) Generation of isogenic pluripotent stem cells differing exclusively at two early onset Parkinson point mutations. *Cell*, **146**, 318–331.
23. Valamehr,B., Abujarour,R., Robinson,M., Le,T., Robbins,D., Shoemaker,D. and Flynn,P. (2012) A novel platform to enable the high-throughput derivation and characterization of feeder-free human iPSCs. *Sci. Rep.*, **2**, 213.
24. Sanjana,N.E., Cong,L., Zhou,Y., Cunniff,M.M., Feng,G. and Zhang,F. (2012) A transcription activator-like effector toolbox for genome engineering. *Nat. Protoc.*, **7**, 171–192.
25. Gibson,D.G., Young,L., Chuang,R., Venter,J.C., Iii,C.A.H., Smith,H.O. and America,N. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*, **6**, 12–16.
26. Zou,J., Maeder,M.L., Mali,P., Pruett-Miller,S.M., Thibodeau-Beganny,S., Chou,B.K., Chen,G., Ye,Z., Park,I.H., Daley,G.Q. *et al.* (2009) Gene targeting of a disease-related gene in human induced pluripotent stem and embryonic stem cells. *Cell Stem Cell*, **5**, 97–110.
27. Park,I.H., Lerou,P.H., Zhao,R., Huo,H. and Daley,G.Q. (2008) Generation of human-induced pluripotent stem cells. *Nat. Protoc.*, **3**, 1180–1186.
28. Hockemeyer,D., Wang,H., Kiani,S., Lai,C.S., Gao,Q., Cassady,J.P., Cost,G.J., Zhang,L., Santiago,Y., Miller,J.C. *et al.* (2011) Genetic engineering of human pluripotent cells using TALE nucleases. *Nature biotechnology*, **29**, 731–734.
29. Mussolino,C., Morbitzer,R., Lütge,F., Dannemann,N., Lahaye,T. and Cathomen,T. (2011) A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity. *Nucleic Acids Res.*, **39**, 9283–9293.
30. Bedell,V.M., Wang,Y., Campbell,J.M., Poshusta,T.L., Starker,C.G., Krug Ii,R.G., Tan,W., Penheiter,S.G., Ma,A.C., Leung,A.Y.H. *et al.* (2012) *In vivo* genome editing using a high-efficiency TALEN system. *Nature*, **490**, 114–118.
31. Pathak,V.K. and Temin,H.M. (1990) Broad spectrum of in vivo forward mutations, hypermutations, and mutational hotspots in a retroviral shuttle vector after a single replication cycle: substitutions, frameshifts, and hypermutations. *Proc. Natl Acad. USA*, **87**, 6019–6023.
32. Tian,J., Ma,K. and Saaem,I. (2009) Advancing high-throughput gene synthesis technology. *Mol. Biosyst.*, **5**, 714–722.
33. Zou,J., Mali,P., Huang,X., Dowey,S.N. and Cheng,L. (2011) Site-specific gene correction of a point mutation in human iPS cells derived from an adult patient with sickle cell disease. *Blood*, **118**, 4599–4608.
34. Mali,P., Yang,L., Esvelt,K.M., Aach,J., Guell,M., Dicarlo,J.E., Norville,J.E. and Church,G.M. (2013) RNA-guided human genome engineering via Cas9. *Science (New York, N.Y.)*, **339**, 823–826.
35. Boyle,A.P., Davis,S., Shulha,H.P., Meltzer,P., Margulies,E.H., Weng,Z., Furey,T.S. and Crawford,G.E. (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322.
36. Orlando,S.J., Santiago,Y., DeKelver,R.C., Freyvert,Y., Boydston,E.A., Moehle,E.A., Choi,V.M., Gopalan,S.M., Lou,J.F., Li,J. *et al.* (2010) Zinc-finger nuclease-driven targeted integration into mammalian genomes using donors with limited chromosomal homology. *Nucleic Acids Res.*, **38**, e152.
37. Wang,Z., Zhou,Z.J., Liu,D.P. and Huang,J.D. (2008) Double-stranded break can be repaired by single-stranded oligonucleotides via the ATM/ATR pathway in mammalian cells. *Oligonucleotides*, **18**, 21–32.
38. Rios,X., Briggs,A.W., Christodoulou,D., Gorham,J.M., Seidman,J.G. and Church,G.M. (2012) Stable gene targeting in human cells using single-strand oligonucleotides with modified bases. *PloS One*, **7**, e36697.
39. Elliott,B., Richardson,C., Winderbaum,J., Jac,A., Jasin,M. and Nickoloff,J.A.C.A. (1998) Gene conversion tracts from double-strand break repair in mammalian cells. *Mol. Cell. Biol.*, **18**, 93–101.
40. Doyon,Y., Choi,V.M., Xia,D.F., Vo,T.D., Gregory,P.D. and Holmes,M.C. (2010) Transient cold shock enhances zinc-finger nuclease-mediated gene disruption. *Nat. Methods*, **7**, 459–460.

Appendix 9.C: Fluorescent *in situ* sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues

The work presented in this chapter has been published in the following paper[134]:

- Lee JH, Daugharthy ER, et al. Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat Protoc.* 2015. doi:10.1038/nprot.2014.191.

PROTOCOL

Fluorescent *in situ* sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues

Je Hyuk Lee[1,6], Evan R Daugharthy[1–3,6], Jonathan Scheiman[1,2], Reza Kalhor[2], Thomas C Ferrante[1], Richard Terry[1], Brian M Turczyk[1], Joyce L Yang[2], Ho Suk Lee[4], John Aach[2], Kun Zhang[5] & George M Church[1,2]

[1]Wyss Institute, Harvard Medical School, Boston, Massachusetts, USA. [2]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [3]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, USA. [4]Department of Electrical and Computer Engineering, University of California San Diego, California, USA. [5]Department of Bioengineering, University of California San Diego, La Jolla, California, USA. [6]These authors contributed equally to this work. Correspondence should be addressed to J.H.L. (jehyuklee@mac.com) or G.M.C. (gchurch@genetics.med.harvard.edu).

Published online 12 February 2015; doi:10.1038/nprot.2014.191

RNA-sequencing (RNA-seq) measures the quantitative change in gene expression over the whole transcriptome, but it lacks spatial context. In contrast, *in situ* hybridization provides the location of gene expression, but only for a small number of genes. Here we detail a protocol for genome-wide profiling of gene expression *in situ* in fixed cells and tissues, in which RNA is converted into cross-linked cDNA amplicons and sequenced manually on a confocal microscope. Unlike traditional RNA-seq, our method enriches for context-specific transcripts over housekeeping and/or structural RNA, and it preserves the tissue architecture for RNA localization studies. Our protocol is written for researchers experienced in cell microscopy with minimal computing skills. Library construction and sequencing can be completed within 14 d, with image analysis requiring an additional 2 d.

INTRODUCTION

Background

Cell type and function in tissues can be inferred from RNA or protein markers[1,2], but this approach to functional classification requires well-characterized biomarkers. Ideally, it would be preferable to define cell or tissue types using high-throughput molecular profiling *in situ* with high-resolution imaging. Indeed, several studies have surveyed global gene expression *in situ*, in which hundreds of organ tissue slices from multiple animals were individually interrogated using gene-specific probes[3–6]; however, such approaches represent a massive experimental undertaking, and they produce only an average view of tissue-specific gene expression.

In theory, multiplexed *in situ* RNA detection demands fewer samples, but so far this approach is limited by the number of spectrally distinct fluorophores and the optical diffraction limit of microscopy[7–11]. Alternatively, padlock probes[12–16] can capture specific RNA sequences from dozens of genes in parallel for targeted sequencing *in situ*[12]; however, padlock probes can have a substantial amount of probe-specific bias[17], and the approach cannot easily be scaled to the transcriptome. Given these challenges, *in situ* RNA profiling is typically restricted to a small number of well-annotated genes, and they can miss differences arising from unexpected signaling pathways or noncoding RNAs. In contrast, we wanted to develop an unbiased and transcriptome-wide sampling method for quantitative visualization of RNA *in situ*, preferably using direct molecular sequencing[18,19] for the detection of tissue-specific gene expression, RNA splicing and post-transcriptional modifications while preserving their spatial context; we call our method fluorescence *in situ* sequencing of RNA (FISSEQ).

Overview of the FISSEQ procedure

FISSEQ begins with fixing cells on a glass slide and performing reverse transcription (RT) *in situ*. After RT, the residual RNA is degraded to prevent it from competitively inhibiting CircLigase, and cDNA fragments are circularized at 60 °C. To prevent cDNA fragments from diffusing away, primary amines are incorporated into cDNA fragments during RT via aminoallyl-dUTP, and the primary amines are then cross-linked using BS(PEG)9. Each cDNA circle is linearly amplified using rolling-circle amplification (RCA) into a single molecule containing multiple copies of the original cDNA sequence, and the amine-modified RCA amplicons are cross-linked to create a highly porous and 3D nucleic acid matrix inside the cell (**Fig. 1a**).

In SOLiD sequencing by ligation, crucial enzymatic steps can be performed directly on a standard microscope at room temperature (25 °C). First, a sequencing primer is hybridized to multiple copies of the adapter sequence in RCA amplicons, followed by ligation of dinucleotide-specific fluorescent oligonucleotides. After imaging, the fluorophores are cleaved from the ligation complex, and ligation of fluorescent oligonucleotides is repeated six more times to interrogate dinucleotide pairs at every fifth position (**Fig. 1b**). To fill in the gaps between dinucleotide pairs, the whole ligation complex is stripped off, and four additional sequencing primers with a single base offset are used to repeat dinucleotide interrogation starting from positions *N*–1, *N*–2, *N*–3 and *N*–4, generating up to 35 raw 3D image stacks representing dinucleotide compositions at all base positions over time.

The raw images are enhanced using standard 3D deconvolution techniques to reduce the background noise, and our freely available MATLAB script performs image alignment to produce TIFF images that are then used for base calling using a separate python script (http://arep.med.harvard.edu/FISSEQ_Nature_Protocols_2014/). The base calls from individual pixels are then aligned to the reference transcriptome using Bowtie, and neighboring pixels with highly similar sequences are grouped into a single object generating a consensus sequence. The final data set includes the number of individual pixels per object, gene ID, consensus sequence, *x* and *y* centroid positions, number of mismatches, base call quality and alignment quality.

One of the key considerations early in the development of FISSEQ was imaging. Biological patterns, including RNA localization, occur

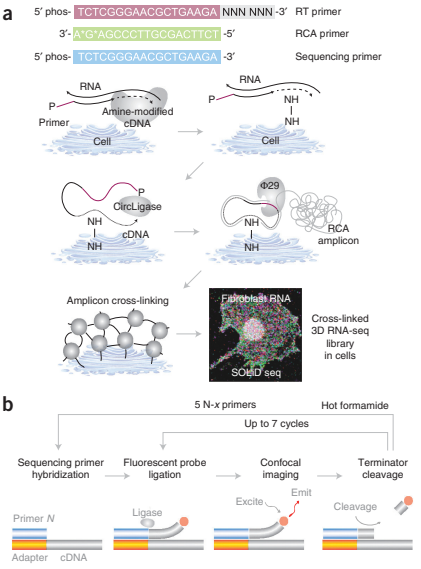

Figure 1 | Schematic overview of FISSEQ library construction and sequencing. (**a**) Fixed cells or tissues are permeabilized and reverse-transcribed *in situ* in the presence of aminoallyl-dUTP and adapter sequence-tagged random hexamers. The cDNA fragments are fixed to the cellular protein matrix using a nonreversible amine cross-linker and circularized after degrading the RNA. The circular templates are amplified using RCA primers complementary to the adapter sequence in the presence of aminoallyl-dUTP and stably cross-linked. The nucleic acid amplicons in cells are then ready for sequencing and imaging (fibroblast shown). (**b**) Each amplicon contains numerous tandem copies of the cDNA template and adapter sequence. A sequencing primer hybridizes to the adapter sequences in individual amplicons, and fluorescent eight-base probes interrogate the adjacent dinucleotide pair. After imaging, the three bases attached to a fluorophore are cleaved, generating a phosphorylated 5′ end at the ligation complex suitable for additional ligation cycles interrogating every fifth dinucleotide pairs. The whole process is repeated using four other sequencing primers with an offset to interrogate intervening base positions.

in a scale-dependent manner, in which some patterns are visible at one scale but disappear at another. Therefore, we developed our sequencing method specifically for confocal microscopy using a wide range of objectives, magnification, numerical apertures (NAs), scanning speed and depth. In addition, autofluorescence, cell debris and background noise are common in cell imaging, unlike in standard next-generation sequencing. Therefore, we developed an approach to classify individual pixels on the basis of their specific color transitions to detect true signals even in the noisy and/or low-intensity environment. Finally, we also developed a way to control the imaging density of single molecules, which enables the sequencing of a large number of molecules in single cells regardless of the microscopy resolution.

Comparisons with single-cell RNA-seq

More than one million mRNA reads per cell can be obtained from a single-cell RNA-seq experiment[20], but typically <100,000 reads per cell are from unique cDNA fragments, and PCR amplification accounts for the remainder[20–22]. In one study, the detection sensitivity of single-cell RNA-seq was estimated as ~10% or ~3% compared with single-molecule fluorescence *in situ* hybridization (FISH) or spiked-in controls, respectively[20]. This means that only ~300 genes are expected to have a coefficient of variation of <23% based on Poisson distribution; however, such genes are generally uninformative, and they include many housekeeping genes such as ribosomal subunit proteins (**Fig. 2a**), requiring that reads from multiple cells are combined to detect biologically meaningful gene expression differences between groups of single cells.

In FISSEQ, only ~200 mRNA reads per cell are obtained without rRNA depletion[23] (versus ~40,000 in single-cell RNA-seq); however, functionally important transcripts are enriched in FISSEQ by more than tenfold compared with single-cell RNA-seq (**Fig. 2b**). When examining a single spatial region of ~40 cells (~8,000 mRNA reads), the top-ranked genes lie substantially above the detection threshold, and they form highly reproducible cell type–specific annotation clusters[23]. Because of the relative absence of housekeeping genes, the high correlation (Pearson's r >0.9) between biological replicates in FISSEQ is driven by cell type– and/or function-specific genes rather than housekeeping genes.

To attain truly single-cell gene expression profiling that is biologically meaningful, FISSEQ may require a read depth per cell that is ~40 times deeper (~8,000 amplicons per cell).

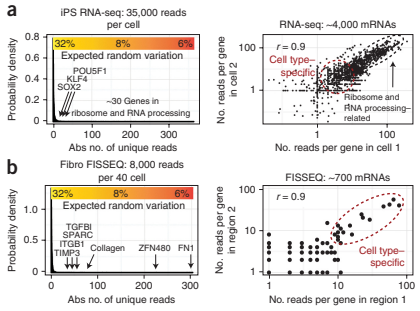

Figure 2 | Comparing single-cell RNA-seq with FISSEQ. (**a**) A typical single-cell RNA-seq[20] can generate more than one million reads per cell, but <10% represent unique reads from cDNAs, and they are composed largely of structural and/or housekeeping genes (i.e., ribosome-related). Many genes of interest are found near the detection limit with a large coefficient of variation, and the high correlation reported for single-cell RNA-seq is typically due to housekeeping genes. (**b**) The current version of FISSEQ combines mRNA reads from ~40 cells to obtain a comparable result, but the high correlation between biological replicates in FISSEQ results from mostly cell type–specific expression markers.

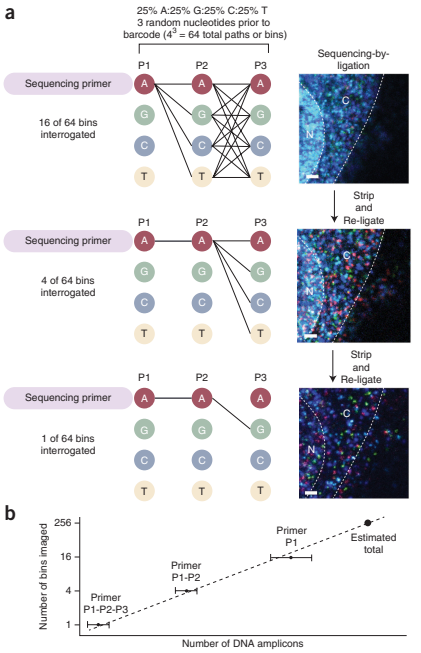

Figure 3 | Counting resolution-limited amplicons using partition sequencing. (**a**) The cDNA or padlock probe template can include three random nucleotides in equal proportions. By controlling the length of the complementary portion of the sequencing primer to the random bases, one can ligate fluorescent probes to different amplicon pools of varying sizes (fibroblasts; scale bars, 1 µm). This scheme works for single-base sequencing-by-ligation, and the SOLiD sequencing chemistry requires additional modifications to the bridge oligonucleotide. C, cytoplasm; N, nucleus. (**b**) Serial ligation reactions using the sequencing primers with 0–3 complementary bases to the random partitioning bases are analogous to doing a serial dilution experiment. The average count from each primer category can be used to extrapolate and estimate the actual amplicon count, regardless of the limitations in optical microscopy.

As the rRNAs comprise >80% of the reads in FISSEQ[23], it may be possible to increase the read depth by about fivefold by simply depleting rRNA *in situ*[24]. We expect another fivefold increase in the amplicon density by optimizing our reaction conditions, and a read depth of ~5,000 non-rRNA reads per cell may soon be possible. As individual amplicons of any density can be discriminated using partition sequencing[23] (**Fig. 3**), the actual size of each amplicon now becomes a limiting factor in the number of reads generated per cell.

Single-cell RNA-seq and FISSEQ are fundamentally limited by the efficiency of mRNA to cDNA conversion. In single-cell RNA-seq, this is estimated to be ~10% compared with single-molecule FISH[20], with a detection threshold of ~5–10 mRNA molecules per cell[21]. This means that most low-abundance genes are not detected in single-cell RNA-seq for a given cell. For FISSEQ, this value is harder to determine because not all genes are enriched in the same manner, but we estimate the current detection threshold at ~200–400 mRNA molecules per cell. After rRNA depletion and other improvements, the detection threshold may improve to ~10–20 mRNA molecules per cell; however, a large fraction of low-abundance genes will still remain undetected.

Comparisons with other approaches

Compared with microdissection[25,26] or photo-activated mRNA capture–based[27] single-cell RNA-seq[21,28–31], FISSEQ scales to large tissues more efficiently[32], and it can compare multiple RNA localization patterns in a nondestructive manner[23]. In addition, other methods require RNA isolation and PCR that can introduce a substantial amount of technical variability[20–22], assuming a Poisson distribution model of transcript abundance. In contrast, all samples can be processed together in a single well from cell culture to sequencing in FISSEQ.

Single-molecule FISH remains the gold standard for high-sensitivity detection of RNA in single cells[7–9,33–37]; however, spectral discrimination of hybridized probes can be difficult to multiplex, and it requires high-resolution microscopy. Recently, highly scalable FISH was demonstrated in single cells, in which sequential hybridization is used to barcode a color sequence for each transcript[10]. In theory, only seven hybridization cycles are required to interrogate 4^7 or >16,000 genes using four colors; however, this approach is limited by the sheer number of probes needed, and the optical diffraction limit prevents accurate quantification of highly abundant or aggregated transcripts.

The sensitivity of padlock probes is two orders of magnitude higher than FISSEQ for a given gene[12,13], but the use of locked nucleic acid makes this approach prohibitively expensive for multiplexing, and individual probes must be calibrated for measuring the relative RNA abundance. For certain applications, it may be possible to combine FISSEQ and padlock probes to interrogate a large number of loci *in situ*. In a recent study, sequencing was limited to short barcodes from dozens of gene-specific padlock probes[12], but now hundreds of thousands of padlock probes[17,38–41] can be discriminated using a 20-base barcode. In the same study, the microscopy resolution limited the number of targeted genes[12], but our partition sequencing[23] bypasses such limitations for highly multiplexed amplicon discrimination *in situ*.

Limitations

On a practical level, equipping a microscope for four-color imaging can currently cost up to $20,000 for a new filter set and a laser. Most users will need to reserve the microscope for 2–3 weeks so that sequencing can proceed uninterrupted. We have used laser-scanning confocal, wide-field epifluorescence and spinning-disk confocal microscopes and obtained comparable sequencing data that differ mainly in the read density. With the laser-scanning confocal microscope, imaging can take over 30 min per stack, but wide-field or spinning-disk confocal microscopes can image the same volume in 1–2 min. Reagent exchanges are done manually in the current protocol, but FISSEQ samples can remain on the microscope and be sequenced over 2–3 weeks.

On a technical level, a major limitation of our current protocol is the lack of rRNA depletion. Initially, we used rRNA as an internal control for library construction, sequencing and bioinformatics; however, this reduced the number of mRNA reads per cell. In primary fibroblasts, the rRNA reads comprised 40–80% of the total (ref. 23); therefore, if one were to deplete the rRNA[24], it might be possible to increase the number of mRNA reads per cell by about fivefold.

Another limitation is the lack of information on biases in our method. FISSEQ enriches for biologically active genes, enabling discrimination of cell type–specific processes with a small number of reads[23]; however, it is not clear how such enrichment occurs. We hypothesize that active RNA molecules are more accessible to FISSEQ, whereas RNA molecules involved in ribosome biogenesis, RNA splicing or heat-shock responses are trapped in ribonucleoproteins, spliceosomes or stress granules. It is now important to investigate and understand the molecular basis of such enrichment across multiple cell types and conditions and to correlate the result with the observed cellular phenotype.

Applications

The current FISSEQ protocol is suitable for most cultured cells and tissue sections, including formalin-fixed and paraffin-embedded (FFPE) tissue sections. Whole-mount *Drosophila* embryos, induced pluripotent stem cell (iPSC)-derived embryo bodies (EBs) and organoids are also compatible (**Table 1**). In FISSEQ, each sequencing read has a spatial coordinate, and the reads are binned according to the cellular morphology, subcellular location, protein localization or GFP fluorescence. A statistical test is then applied to identify enriched genes and pathways *de novo* and to discover possible biomarkers of the cellular phenotype[23]. This approach may be combined with padlock probes to detect evolving mutations and RNA biomarkers in cancers[12,13] or to compare gene expression in asymmetric cells or tissues.

FISSEQ may also sequence molecular barcodes in individual cells and transcripts, where expression or reporter (i.e., cDNA, promoter-GFP) libraries are examined in a pool of single cells for massively parallel functional assays and cell-lineage tracing. In essence, a practically unlimited number of DNA-associated cellular features may now be imaged, enumerated and analyzed across multiple spatial scales using the DNA sequence as a temporal barcode.

Experimental design

General considerations. This protocol details the method described in our original report[23], in which endogenous RNAs in cultured fibroblasts were sequenced on a confocal microscope. The availability of a microscope and computational resources will guide the general experimental approach (**Table 2**). We provide basic computational tools along with a sample data set, but a background in python, MATLAB, ImageJ and/or R is helpful for analyzing a large number of images. If such expertise is not available, we recommend focusing on a few regions of interest with well-demarcated features for comparing gene expression using our custom scripts[23]. After outlining the experiment, one should download our sample image, software and data set and become familiar with image and data analysis (http://arep.med.harvard.edu/FISSEQ_Nature_Protocols_2014/). One should then finalize the experimental design and define the imaging parameters (i.e., area, thickness, resolution and magnification).

Cell and tissue fixation. We have been able to fix and generate *in situ* sequencing libraries in a wide number of biological specimens (**Table 1**). The only case in which we failed was a hard piece of bone marrow embedded in Matrigel, which detached from the glass surface after several wash steps. Fixation artifacts can include changes in subcellular RNA localization, cell swelling, incomplete permeabilization and RNA leakage. Certain primary cell types are also sensitive to cold[42], whereas transformed cell lines or stem cells appear to be less sensitive (**Supplementary Fig. 1**). If you are using FISSEQ to study subcellular localization, we recommend fixing cells by adding warm formalin directly into the growth medium to a final concentration of 10% (vol/vol).

Cell and tissue sample mounting. For high-resolution imaging, we recommend poly lysine– or Matrigel-coated glass-bottom

TABLE 1 | Specimens tested for FISSEQ library construction.

Types	Fixation	Mounting substrate	Permeabilization	Notes
HeLa, 293A, COS1, U2OS, iPSC, primary fibroblasts and bipolar neurons	10% (vol/vol) formalin or 4% (vol/vol) PFA	Poly lysine–coated coverslip (Matrigel for iPSCs)	70% (vol/vol) ethanol or 0.25% (vol/vol) Triton X-100 (0.1 N HCl optional)	Changes in temperature can cause altered mRNA localization
Mouse embryo FFPE section (20 µm)	Already fixed	Superfrost Plus glass slide	0.1% (wt/vol) pepsin in 0.1 N HCl	Use silicone isolators (Grace Bio-Labs)
Mouse brain fresh-frozen section (20 µm)	10% (vol/vol) formalin	Poly lysine–coated coverslip	0.1% (wt/vol) pepsin in 0.1 N HCl	Use silicone isolators (Grace Bio-Labs)
iPS-derived 3D organoids	10% (vol/vol) formalin	Poly lysine–coated coverslip (embed in Matrigel and fix with 4% PFA)	0.25% (vol/vol) Triton X-100 and 0.1 N HCl	10% (vol/vol) formalin is less effective for fixing Matrigel
Dechorionated whole-mount *Drosophila* embryos	10% (vol/vol) formalin	Poly lysine–coated coverslip (embed in Matrigel and fix with 4% PFA)	100% (vol/vol) methanol then PBS with 0.2% (vol/vol) Triton X-100 and 0.2% (vol/vol) Tween-20	10% (vol/vol) formalin is less effective for fixing Matrigel

TABLE 2 | Comparison of the microscopy platforms tested for FISSEQ.

	Model	Pros	Cons	Uses
Wide-field epifluorescence	Nikon TE-2000	Fast imaging Simple setup	Poor axial resolution Low signal-to-noise ratio Lower read depth	Thin cells and tissue sections Whole-cell barcode labeling
Scanning confocal	Zeiss LSM 710 confocal Leica TCS SP5 Confocal	Good axial resolution Scanning zoom Flexible pixel density	Slow imaging	High-resolution FISSEQ of a single region
Spinning-disk confocal	Yokogawa CSU-W1	Fast imaging Good axial resolution	Fixed pixel density	All purpose

dishes, but 96-well plastic-bottom plates can be used for simple protocol optimization. Tissue sections can be mounted using a standard mounting procedure, and we advise inexperienced users to consult those who have experience in the art of tissue mounting. For nonadherent cell types and whole-mount specimens, we recommend fixing samples embedded in Matrigel using 4% (vol/vol) paraformaldehyde (PFA) on a glass-bottom dish.

RT *in situ*. The length of RT primers should be <25 bases to prevent self-circularization. We perform RT overnight for most samples, but 1 h is often sufficient for cell monolayers. A negative control without RT should be included to rule out self-circularization of the primer. A positive control primer with the adapter sequence plus a synthetic sequence (~30 additional bases) can be used to check RCA and imaging parameters. Other than the 5′ region of highly abundant mCherry transcripts[23], we have not had consistent results with targeted RT; we typically see very few amplicons regardless of the primer design. In contrast, random hexamers (24 bases) and poly-dT primers (33 bases) work well across all conditions. Some of the possible reasons for failure may include poor target accessibility and competitive inhibition of CircLigase by nonspecifically bound sequence-specific RT primers that are capable of self-circularization. Possible solutions include the use of locked nucleic acid (LNA)-based RT primers for high-temperature hybridization[13], ligation of the adapter sequence after RT and tiling multiple RT probes across a gene target. We have yet to try these alternatives.

Generation of amplicon matrix. Aminoallyl-dUTP is a dTTP analog commonly used in fluorescence labeling of cDNA[43], which we use for cross-linking nucleic acids; however, the efficiency of RT and RCA is inversely correlated with the concentration of aminoallyl-dUTP[23]. The cross-linker, bis(succinimidyl)-nona-(ethylene glycol) or BS(PEG)9, is functionalized with *N*-hydroxysuccinimide (NHS) ester groups at both ends[44], and it forms a stable covalent bond with primary amine groups provided by aminoallyl-dUTP at pH 7–9. The cross-linking density can be enhanced by increasing the concentration of aminoallyl-dUTP or BS(PEG)9, or by increasing the pH. Cross-linking after RT is optional, but cross-linking of RCA amplicons is essential for high-quality sequencing reads.

Sequencing. We use sequencing-by-ligation[18,19,45] (SOLiD[46,47]) because it works well at room temperature, and so a heated stage is not required. SOLiD uses a dinucleotide detection scheme in which a base position is interrogated twice per sequencing run[46,47], and this can reduce the base calling error rate; however, converting the color sequence to the base sequence is not straightforward because of its propensity to propagate errors, and sequence analysis must remain in the color space (**Box 1** and **Fig. 4**). In comparison, sequencing-by-synthesis (Illumina) works at 65 °C for primer extension and cleavage, and it uses proprietary fluorophores that require a heated flow-cell and a custom imaging setup. As sequencing-by-synthesis can generally yield a much longer read length, we are currently investigating its compatibility with FISSEQ.

Box 1 | SOLiD sequencing chemistry

The SOLiD sequencing chemistry consists of multiple reaction cycles in which a sequencing primer is extended using fluorescent eight-base probes via sequential DNA ligation. The fluorescent amplicons are then imaged, and the last three bases and the fluorophore are cleaved, followed by the ligation of another eight-base probe. These steps are repeated using four additional sliding primers to record the dinucleotide color values from starting positions 1-6-11-16-20-26-31 (primer *N*), 0-5-10-15-20-25-30 (primer *N*–1), 4-9-14-19-24-29-34 (primer *N*–2), 3-8-13-18-23-28-33 (primer *N*–3) and 2-7-12-16-22-27-32 (primer *N*–4; **Fig. 4a**). Most bases are represented by two sequential colors, and although each color represents up to four possible dinucleotide combinations the exact nucleotide sequence can be determined if the identity of any one base is known (i.e., the base identity in the sequencing primer). For example, AAGCAGTCA is equivalent to BORGOGOG (B: blue, O: orange, R: red and G: green; **Fig. 4b**); however, the conversion table alone cannot assign the base identity from color codes. However, if one base is known (i.e., first base is A in BORGOGOG), assigning the base identity is relatively straightforward (**Fig. 4c**). One disadvantage is that any missing or wrong base calls can affect the whole read, and it makes sequence-to-sequence comparisons impossible. Therefore, the SOLiD sequencing reads and the reference database must remain in the color space for sequence alignment, and the user should keep this in mind when designing a custom sequence analysis pipeline.

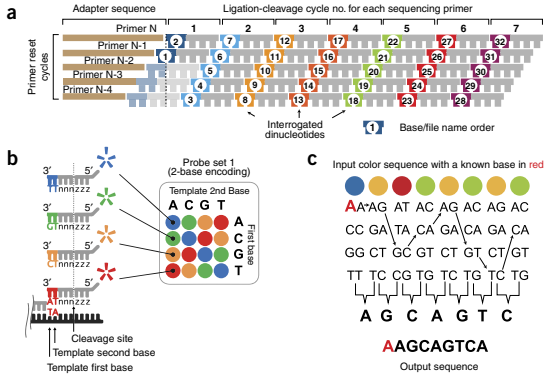

Figure 4 | Schematic overview of the SOLiD color-coding and decoding scheme. (**a**) The base position within the template sequence is enclosed by white circles and should be used for naming the image files, and the actual sequencing cycle numbers are noted on both sides. Each ligation extension is shown in different colors, and cycles 15, 22 and 29 are shown in gray, as no images are acquired for these cycles. The red box at cycle 8 denotes a known base identity. (**b**) SOLiD dinucleotide coding scheme. (**c**) SOLiD color space decoding scheme. As long as any one of the base identities are known (here in red), the color space sequence can be converted to the nucleotide sequence. Image reproduced from Life Technologies (ref. 47). © 2014 Thermo Fisher Scientific, Inc. Used under permission.

Partition sequencing. T4 DNA ligase has a single-base specificity at the ligation junction[18], and sequencing primers differing by one base can recognize different sets of amplicons[23]. By dividing imaging over multiple separate runs, spatially overlapping amplicons can be enumerated using multiple sequencing primers even on a low-resolution microscope; however, this requires full automation for the increased number of sequencing runs per sample. Without automation, partition sequencing is better suited for quantifying short barcode sequences rather than full RNA sequences *in situ*[12] (**Fig. 3**).

Imaging. Epifluorescence microscopy can generate a reasonable number of alignable reads from relatively thin specimens (<5 µm), such as HeLa cells[23], but thicker samples require confocal microscopy to obtain high-density reads. Spinning-disk confocal microscopy is markedly faster than laser-scanning confocal microscopy, and it has a good balance of imaging speed and axial resolution. An automated stage capable of finding a *z*-stack across multiple *x*-*y* tiles is highly desirable (**Table 2**).

In FISSEQ, individual amplicons can be detected using objectives with a NA of 0.4 or greater. The magnification required is determined by the biological question and the amplicon density[48]. Typically, we use a 20× NA 0.75 objective to examine tissue sections and cultured cell monolayers, whereas 40× NA 0.8 and 63× NA 1.2 water-immersion objectives are used for high-resolution imaging of single cells. We have observed noticeable chromatic aberration in our experiments, depending on the objectives used. The degree of chromatic aberration should be measured using image calibration beads (i.e., FocalCheck fluorescence microscope test slide) before sequencing, and they should be calibrated by the microscope vendor if necessary.

For each imaging setup, the user should determine the ideal Nyquist rate. This value can be calculated using http://www.svi.nl/NyquistCalculator. The *x*-*y* pixel and *z*-step sizes should not be >1.7 times the Nyquist value for image deconvolution. Four-color imaging should proceed from the longest to the shortest wavelength (i.e., Cy5, Texas Red, Cy3 and FAM), and an intensity histogram should be used to adjust the laser power to prevent saturated pixels. The intensity histogram should be consistent across fluorescence channels and sequencing cycles. To use our software, the image file name must be standardized: <Position>_<Primer #>_ <Ligation #>_<Date_Time>.extension (e.g., 06_N1_2_2013_10_25_11_57_18.czi).

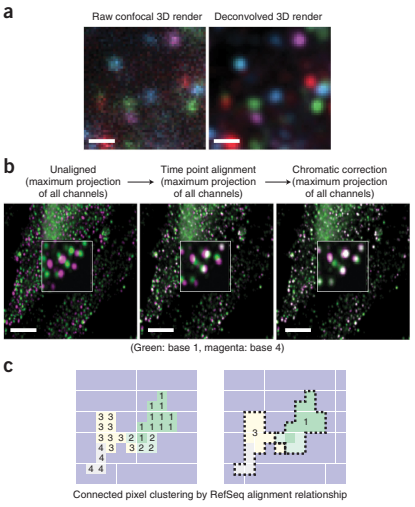

Figure 5 | Example of image analysis, registration and sequence clustering. (**a**) A four-color confocal image stack is deconvolved using a CLME algorithm with ten iterations and a signal-to-noise ratio of two (scale bars, 1 µm). (**b**) Sequencing images for base 1 and base 4 (left) are aligned using a composite channel across various time points (middle) and then using a composite time projection across various channels (right; scale bars, 5 µm). (**c**) Individual nonzero pixels are aligned to the reference sequence database (i.e., human RefSeq). Highly related sequences connected to the neighboring pixels are then grouped into a single cluster.

Image analysis tools. In practice, the extent of image processing and analysis is dictated by the available imaging tools and computing resources[49]. We use Bitplane Imaris for data visualization and movie creation and Scientific Volume Imaging (SVI) Huygens for 3D deconvolution. Although they are easy to use, scalable and relatively fast, their cost may be out of reach for small laboratories; however, free and/or open-source alternatives are also available[49–51].

Image deconvolution. We use 3D deconvolution[52] to reduce the out-of-focus background and to improve the quality of base calls (**Fig. 5a**). High-quality 3D deconvolution requires sampling near the Nyquist rate, but this increases the image acquisition and deconvolution time, as well as the file size. We generally recommend using high-quality confocal imaging and minimal 3D deconvolution for FISSEQ. The use of 3D deconvolution to compensate for low-quality imaging will not necessarily improve the quality or the number of sequencing reads. We provide a sample data set containing raw and deconvolved image stacks from a successful 30-base sequencing experiment for practice (http://arep.med.harvard.edu/FISSEQ_Nature_Protocols_2014/).

Image registration, base calling and sequence alignment. As long as the input image files are correctly named, our software will generate the maximum intensity projection, register the images and correct for chromatic shifts[23] (**Fig. 5b**). The resulting images are used for base calling and sequence alignment to human RefSeq (**Fig. 5c**), but our software does not generate *z*-coordinates for sequencing reads, as it uses maximum intensity projection for base calling. We provide a sample data output and screen logs for troubleshooting our bioinformatics software (http://arep.med.harvard.edu/FISSEQ_Nature_Protocols_2014/).

Data analysis. Our software generates a tab delimited text file that contains 10,000–50,000 aligned reads per field of view. We recommend RStudio with the latest version of R installed for plotting reads by RNA classes, position, cluster size, quality, gene name, strand and so on. We provide a sample R session file that is used for FISSEQ data analysis as an introduction to statistical computing and for assessing the quality of FISSEQ data set (http://arep.med.harvard.edu/FISSEQ_Nature_Protocols_2014/).

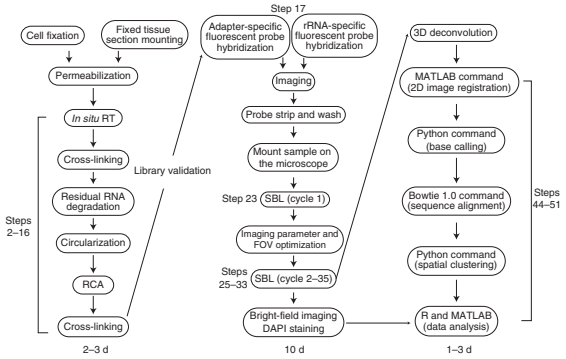

Figure 6 | Schematic overview of FISSEQ experimental and analysis steps. FOV, field of view; SBL, SOLiD sequencing-by-ligation.

Level of expertise required for the protocol. FISSEQ is at the intersection of cell imaging and functional genomics, and it has generated much interest from cell biologists who are not familiar with RNA-seq. Our protocol is aimed at such researchers, who are familiar with cell image analysis but have few computing skills (**Fig. 6**). FISSEQ library construction can be performed by anyone with basic molecular biology skills, but image acquisition is best done with help from an imaging core specialist for the initial setup. Once the equipment, software, imaging and deconvolution parameters are finalized, a capable technician, graduate student or post-doc can perform manual sequencing on a microscope with some training and practice. Image and sequence analysis using our software can be performed by anyone familiar with the Unix environment, but statistical data analysis requires either a graduate student or post-doc familiar with statistical tools and concepts.

Considerations about the laboratory facilities. All steps in FISSEQ library construction can be carried out in a standard laboratory setting. A vacuum line facilitates solution aspiration and reagent exchanges, and we do not find RNA degradation or PCR contamination to be a notable problem in our method. We advise having a dedicated microscope with proper excitation and emission filters on a vibration isolation table in a low-traffic area.

MATERIALS

REAGENTS

- Starting material of interest. The PROCEDURE is written for cultured cells in glass-bottom dishes or for tissue sections on glass-bottom dishes or coverslips. However, it can be adapted for use on a range of starting materials (**Table 1**)
- Acetone (for detaching a Petri dish glued to the microscope stage)
  ! CAUTION Acetone is highly flammable. Work in a well-ventilated area.
- Aminoallyl-dUTP, 4 mM (AnaSpec, cat. no. 83203)
- Betaine, 5 M (included in CircLigase II kit, Epicentre, cat. no. CL9025K)
- BS(PEG)9, 100 mg (Thermo Scientific, cat. no. 21582)
  ▲ CRITICAL BS(PEG)9 loses its effectiveness 1 month after reconstitution in DMSO. Prepare a fresh batch every month, especially if it has been frozen and thawed repeatedly.
- CircLigase II kit (Epicentre, cat. no. CL9025K)
- Cleave solution 1 (Applied Biosystems, cat. no. 4406489)
- Cleave solution 2.1 kit (Applied Biosystems, cat. no. 4445677)
  ! CAUTION Contains toxic organoamine. Wear gloves and work in a well-ventilated area.

- Cyanoacrylate adhesive, optical grade (VWR International, cat. no. 19806-00-1)
- Diethylpyrocarbonate (DEPC)-treated water (Santa Cruz Biotechnology, cat. no. sc-204391)
- DMSO (Sigma-Aldrich, cat. no. D8418)
- dNTP, 25 mM (Enzymatics, cat. no. N2050L)
- Ethanol, 70% (vol/vol) (in DEPC-treated water)
- Formalin, 10% (vol/vol) (Electron Microscopy Sciences, cat. no. 15740)
  ! CAUTION Wear gloves and work in a well-ventilated area. Dispose of waste per institutional guideline.
- Formamide (Sigma-Aldrich, cat. no. 221198)
- HCl, 0.1 N (in DEPC-treated water)
- Immersol W 2010 (ne = 1.33) for water-immersion lens (Carl Zeiss Microscopy, cat. no. 444969-0000-000)
- Instrument buffer, 10× (Applied Biosystems, cat. no. 4389784)
- Moloney-murine leukemia virus (M-MuLV) reverse transcriptase (Enzymatics, cat. no. P7040L)
- Mineral oil (Sigma-Aldrich, M5904)
- $MnCl_2$ (included in CircLigase II kit, Epicentre, cat. no. CL9025K)
- Nuclease-free water, not DEPC-treated (Life Technologies, cat. no. AM9932)
- Pepsin, 1 g (dissolve in 10 ml of H_2O and store it at −20 °C; Affymetrix, cat. no. 20010)
- ϕ29 DNA polymerase (Enzymatics, cat. no. P7020-HC-L)
- PBS (Life Technologies, cat. no. 10010023)
- RNase, DNase-free (Roche Applied Science, cat. no. 11579681001)
- RNase H (Enzymatics, cat. no. Y9220F)
- RNase inhibitor (Enzymatics, cat. no. P9240L)
- RNaseZap (Life Technologies, cat. no. P9780)
- Silicone isolator (Grace Bio-Labs, cat. no. 664304)
- Sodium acetate, 3 M (pH 7.5)
- SOLiD ToP sequencing kit fragment library F3 tag MM50 (Applied Biosystems, cat. no. 4449388)
- Saline sodium citrate (SSC), 20× (Roche Applied Science, cat. no. 11666681001)
- Streptavidin–Alexa Fluor 647 (Life Technologies, cat. no. S32357)
- T4 DNA ligase (Enzymatics, cat. no. L6030-LC-L)
- Tris solution, 1 M (G-Biosciences, cat. no. R002)
- Trisodium citrate dihydrate (Sigma-Aldrich, cat. no. C8532)
- Triton X-100, 10% (vol/vol) solution (Sigma-Aldrich, cat. no. 93443)

RT, RCA and sequencing primers (all are in 5′–3′ orientation)_

- Random hexamer RT primer, 100 µM in nuclease-free H_2O (/5phos/TCTCGGGAACGCTGAAGANNNNNN; hand-mixed, Integrated Data Technologies (IDT))
- RCA primer, 100 µM (TCTTCAGCGTTCCCGA*G*A; * is phosphorothioate)
- Sequencing primer *N*: /5phos/TCTCGGGAACGCTGAAGA (HPLC purified)
- Sequencing primer *N*–1: /5phos/CTCGGGAACGCTGAAGA (HPLC purified)
- Sequencing primer *N*–2: /5phos/TCGGGAACGCTGAAGA (HPLC purified)
- Sequencing primer *N*–3: /5phos/CGGGAACGCTGAAGA (HPLC purified)
- Sequencing primer *N*–4: /5phos/GGGAACGCTGAAGA (HPLC purified)

Control primers (5′–3′)_

- Adapter-specific probes, 100 µM (/56-FAM/TCTCGGGAACGCTGAAGA)
- Adapter-specific probes, 100 µM (/5TYE563/TCTCGGGAACGCTGAAGA)
- Adapter-specific probes, 100 µM (/5TEX615/TCTCGGGAACGCTGAAGA)
- Adapter-specific probes, 100 µM (/5TYE665TCTCGGGAACGCTGAAGA)
- 18S rRNA detection primer1: /5biotin/ GCTACTGGCAGGATCAACCAGGTA
- 18S rRNA detection primer2: /5biotin/ TACGCTATTGGAGCTGGAATTACC
- 18S rRNA detection primer3: /5biotin/ GTTGAGTCAAATTAAGCCGCAGGC
- 18S rRNA detection primer4: /5biotin/ TTGCAATCCCCGATCCCCATCACG
- 28S rRNA detection primer1: /5biotin/ CCACGTCTGATCTGAGGTCGCG
- 28S rRNA detection primer2: /5biotin/ CACGCCCTCTTGAACTCTCTCTTC
- 28S rRNA detection primer3: /5biotin/ CTCCACCAGAGTTTCCTCTGGCT
- 28S rRNA detection primer4: /5biotin/ TGAGTTGTTACACACTCCTTAGCG
- 28S rRNA detection primer5: /5biotin/ CGACCCAGCCCTTAGAGCCAATC
- 28S rRNA detection primer6: /5biotin/ GACAGTGGGAATCTCGTTCATCCA
- 28S rRNA detection primer7: /5biotin/ GCACATACACCAAATGTCTGAACC

EQUIPMENT

- 4 °C and −20 °C storage units
- Centrifuge for 1.5- and 2-ml tubes
- Dry-block heater for microtubes at 80 °C
- Falcon conical centrifuge tubes (15 and 50 ml; Fisher Scientific, cat. nos. 14-959-49B and 14-432-22)
- Flexible plastic i.v. catheter for reagent aspiration (Terumo, cat. no. SR*FF2419)
  ! CAUTION The catheter comes with a plastic outer sheath and a sharp needle in the middle. The needle must be carefully removed and discarded into a sharps container.
- FocalCheck fluorescence microscope test slide (Life Technologies, cat. no. F36909)
- Glass bottom MatTek dish (Poly lysine-treated: cat. no. P35GC-1.5-14-C, Poly lysine-treated 96-well plate: cat. no. P96GC-1.5-5-F)
- Glass Pasteur pipettes (autoclaved)
- Incubators at 30 °C, 37 °C (humidified) and 60 °C
- Inverted confocal microscope, PC and image acquisition software (see Equipment Setup)
- Microscope stage insert, metal (for securely gluing the specimen holder)
- Nonsterile syringes, 10 ml (BD Biosciences, cat. no. 301029)
- RNase-free microtubes (Eppendorf, cat. no. 0030 121.589)
- Sealable plastic container or Ziploc bags (for CircLigase reaction at 60 °C)
- Vacuum flask, trap and tubing

PC and software requirements

- Access to a high-performance computing cluster (remote host)
- Bowtie 1.0 or earlier (http://bowtie-bio.sourceforge.net) on the remote host
  ▲ **CRITICAL** Bowtie 2.0 or higher does not work with SOLiD sequencing.
- Fiji/ImageJ (http://fiji.sc/Fiji) on a PC
- MATLAB (http://www.mathworks.com) on the remote host
- Python 2.7 (https://www.enthought.com/products/canopy/) on the remote host ▲ **CRITICAL** Other versions of python lack the required modules for running our script.
- R (http://www.r-project.org) and RStudio (http://www.rstudio.com) on a PC
- Windows PC or Mac with 16 GB RAM minimum
- Optional: SVI Huygens 3D deconvolution software (commercial), Bitplane Imaris 3D rendering software (commercial)

REAGENT SETUP

Triton X-100, 0.25% (vol/vol) Dilute 0.25 ml of 10% (vol/vol) Triton X-100 in DEPC-treated H_2O to a total volume of 10 ml. Store it at room temperature for up to 6 months.

SSC, 2× Dilute 20× SSC in H_2O and adjust the final volume to 50 ml. Store it at room temperature for up to 6 months.

SSC, 1× Dilute 20× SSC in H_2O and adjust the final volume to 50 ml. Store it at room temperature for up to 6 months.

SASC, 5× Make 0.75 M sodium acetate, 75 mM tri-sodium citrate, and then adjust the pH to 7.5 using acetic acid in H_2O to a final volume of 50 ml. Store it at room temperature for up to 6 months.

RCA primer hybridization buffer Dilute 20× SSC, 2× SASC and 30% (vol/vol) formamide in H_2O. Store the buffer at room temperature for up to 6 months.

Strip buffer Strip buffer is 80% (vol/vol) formamide in H_2O and 0.01% (vol/vol) Triton X-100 in a final volume of 50 ml. Store the buffer at room temperature for up to 6 months.

Cleave solution 2.1, reconstituted Mix 1 ml of cleave solution 2.1 Part 1 with 2.75 ml of cleave solution 2.1 Part 2. Store it at 4 °C in the dark for up to 24 h.

EQUIPMENT SETUP

Microscope setup Configure a four-channel microscope with appropriate excitation light sources and emission filters: FITC-488 excitation, 490–560-nm emission; Cy3-561-nm excitation, 563–593-nm emission;

Texas Red-594-nm excitation, 597–647-nm emission; and Cy5–633-nm excitation, 637–758-nm emission. Suggested microscope objectives are plan-Apochromat dry 20× NA 0.75, dry 40× NA 0.8 and water-immersion 63× NA 1.3.

Software installation Verify that Bio-Formats (http://loci.wisc.edu/software/bio-formats) plug-ins are available for Fiji/ImageJ. Download a free academic version of Canopy Python 2.7 in the home directory on the remote host, and follow the installation instructions (http://docs.enthought.com/canopy/quick-start/install_linux.html). Canopy Python 2.7 is easy to install, and it has all the required packages for our FISSEQ software. Install the latest version of the ggplot2 and data.table packages in RStudio.

PROCEDURE

FISSEQ library construction in cultured cells or tissue sections ● TIMING 2–3 d

▲ **CRITICAL** All reagents and washes are at room temperature unless indicated otherwise.

1| To construct libraries for FISSEQ, follow option A for cultured adherent cells or follow option B for tissue sections.

(A) Cultured adherent cells on a glass-bottom dish ● TIMING 30 min

(i) Fix the cells using 2 ml of 10% formalin in PBS for 15 min at 25 °C.

(ii) Wash the cells with 2 ml of PBS three times.

(iii) Add 2 ml of 0.25% (vol/vol) Triton X-100 in DEPC-PBS for 10 min, or 70% (vol/vol) ethanol for 2 min. Triton X-100 tends to maintain the subcellular structures better than 70% (vol/vol) ethanol.

(iv) Wash the cells with 2 ml of PBS three times.

(v) (Optional) Some cell types may require acid treatment for improved permeabilization: add 0.1 N HCl in DEPC-treated H_2O for 10 min, followed by three PBS washes (**Supplementary Fig. 2**).

? TROUBLESHOOTING

(B) Tissue sections on a glass-bottom dish ● TIMING 1 h

(i) Mount 10–20-µm-thick formalin-fixed tissue sections onto an RNase-free glass coverslip using a standard mounting procedure.

(ii) Remove the glass coverslip attached to a MatTek glass-bottom dish by gently pressing around the coverslip with a razor blade.

(iii) Attach the glass coverslip with a mounted tissue section to the MatTek dish using double-sided adhesive tape.

(iv) Wash the tissue section twice using DEPC-treated H_2O for 5 min each.

(v) Add 0.25% (vol/vol) Triton X-100 in DEPC-treated H_2O for 15 min, and aspirate.

(vi) Wash the sample with DEPC-treated H_2O twice.

(vii) Add 200 µl of 0.1% (wt/vol) pepsin in 0.1 N HCl for up to 10–30 min. Most tissue sections are permeabilized after 10–15 min. We recommend optimizing the permeabilization conditions for each tissue type.

(viii) Wash the tissue sections with 2 ml of PBS three times to inactivate pepsin.

? TROUBLESHOOTING

2| Prepare an RT mixture on ice, as indicated below, with and without reverse transcriptase.

▲ **CRITICAL STEP** Chilling the assembled mix to 4 °C before RT improves the efficiency of primer annealing.

Component	Amount (µl)	Final
DEPC-H_2O	159	
M-MuLV RT buffer, 10×	20	1×
dNTP, 25 mM	2	250 µM
Aminoallyl-dUTP, 4 mM	2	40 µM
RT primer, 100 µM (/5Phos/TCTCGGGAACGCTGAAGANNNNNN)	5	2.5 µM
RNase inhibitor (40 U µl[1])	2	0.4 U μl^{-1}
M-MuLV reverse transcriptase (100 U μl^{-1})	10	5 U μl^{-1}
Total	200	

3| Incubate the specimen with the reaction mixture for 10 min at 4 °C, and then transfer it to 37 °C overnight. Typically, 1–2 h is sufficient, but more time may be required for thicker samples. Aspirate and wash the specimen with PBS once.

4| To cross-link cDNA molecules containing aminoallyl-dUTP, add 20 µl of reconstituted BS(PEG)9 in 980 µl of PBS to the sample for 1 h at room temperature.

5| Aspirate and wash the sample with PBS and quench it with 1 M Tris (pH 8.0) for 30 min.
■ **PAUSE POINT** The sample can be stored in PBS for up to 1 week at 4 °C.

6| Aspirate and add 10 µl of DNase-free RNase and 5 µl of RNase H in 1× RNase H buffer for 1 h at 37 °C (**Supplementary Fig. 3**).
▲ **CRITICAL STEP** Skipping this step results in few amplicons.

7| Rinse the sample with 2 ml of nuclease-free H_2O twice to remove traces of phosphate.

8| Prepare a CircLigase reaction mixture on ice, as tabulated below, and add it to the glass-bottom dish containing the sample.

Component	Amount (µl)	Final
Nuclease-free H_2O	128	
CircLigase buffer, 10×	20	1×
$MnCl_2$, 50 mM	10	2.5 mM
Betaine, 5 M	40	0.5 M
CircLigase II (100 U µl^{-1})	2	1 U µl^{-1}
Total	200	

9| Place the glass-bottom dish in a tightly sealed plastic container or in a Ziploc bag with moist wipes, and incubate it at 60 °C for 1 h. If a longer reaction time is desired, 1 ml of mineral oil can be layered on top of the sample.

10| Aspirate the reaction mixture, and wash with PBS. Mineral oil can be removed using PBS with 0.1% (vol/vol) Triton X-100.
■ **PAUSE POINT** The sample can be stored in PBS at 4 °C indefinitely.

11| Add 200 µl of RCA primer hybridization buffer containing 500 nM RCA primer to the glass-bottom dish and incubate at 60 °C for 1 h.

12| Aspirate and wash the sample with RCA hybridization buffer at 60 °C for 10 min.

13| Aspirate and wash the sample with 2× SSC, 1× SSC and PBS once each.

14| Prepare an RCA reaction mixture on ice, as tabulated below. Add this mixture to the sample and incubate it overnight at 30 °C. Additional dNTP (up to 10 µl) and ϕ29 DNA polymerase (up to 10 µl) can enhance the fluorescence signal from DNA amplicons.
▲ **CRITICAL STEP** Aminoallyl-dUTP is required for cross-linking and should not be omitted.

Component	Amount (µl)	Final
Nuclease-free H_2O	174	
ϕ29 buffer, 10×	20	1×
dNTP, 25 mM	2	250 µM
Aminoallyl-dUTP, 4 mM	2	40 µM
ϕ29 DNA polymerase (100 U µl^{-1})	2	1 U µl^{-1}
Total	200	

15| To cross-link cDNA molecules containing aminoallyl-dUTP, wash them gently with PBS, add 20 µl of reconstituted BS(PEG)9 in 980 µl of PBS to the sample and incubate the mixture for 1 h at room temperature.
▲ **CRITICAL STEP** BS(PEG)9 expires after 2–3 weeks with multiple freeze-thaw cycles, and using expired BS(PEG)9 can lead to unstable amplicons and poor sequencing results.

16| Wash the sample with PBS, aspirate and add 1 M Tris, pH 8.0, for 30 min.
■ **PAUSE POINT** Store the sample in PBS at 4 °C for up to 4 weeks.

17| Aspirate and add 2.5 µM control probe in 200 µl of 5× SASC, preheated to 80 °C, to the sample and incubate the mixture for 10 min at room temperature. Use the adapter- or rRNA–specific probes as positive controls to image all amplicons or rRNA amplicons, respectively. RT-negative controls should not produce any amplicons.

18| Wash the sample two times for 1 min each with 1 ml of 1× instrument buffer. If you are using adapter sequence–specific probe, proceed directly to Step 19 for imaging. If you are using the biotinylated rRNA probes, incubate in 2 µg ml^{-1} streptavidin–Alexa Fluor in PBS for 5 min, followed by three 2-ml PBS washes before continuing with Step 19.

19| Image on a microscope and inspect the amplicon density and distribution. Amplicons should be distributed uniformly throughout the sample across the glass-bottom dish. Obtain an axial view, and check to see whether the amplicon density is similar between regions near the glass and cell surface.
▲ **CRITICAL STEP** The sample can be imaged while immersed in 1× instrument buffer. If an alternative immersion liquid is used, do not add Tris-EDTA or other chelating agents.
? TROUBLESHOOTING

20| Aspirate and incubate the sample twice for 5 min each in 1 ml of strip buffer at room temperature, preheated to 80 °C.

21| Wash the sample twice for 5 min each with 1 ml of 1× instrument buffer at room temperature.
■ **PAUSE POINT** We have kept samples in 1× instrument buffer at 4 °C for up to several months without suffering a substantial loss in the fluorescence signal.

SOLiD sequencing-by-ligation ● TIMING 10 d for 30 cycles
22| Clamp the sample firmly to the microscope stage, and use cyanoacrylate adhesive to secure any potential sources of movement, such as adjustable stage inserts. Cyanoacrylate adhesive can be applied directly to metal components, and it can be removed with acetone after sequencing.
▲ **CRITICAL STEP** Use only optical-grade cyanoacrylate adhesive, as standard cyanoacrylate adhesives degas and ruin nearby objectives.

23| Add 2.5 µM sequencing primer N in 200 µl of 5× SASC, preheated to 80 °C, to the sample and incubate the mixture for 10 min at room temperature. Aspiration can be performed using a vacuum aspirator or a flexible plastic catheter attached to a syringe.

24| Wash the sample two times for 1 min each with 1 ml of 1× instrument buffer at room temperature.

25| Sequence the sample by adding a freshly prepared T4 DNA ligation mixture and incubating it for 45 min at room temperature.

Component	Amount (µl)	Final
Nuclease-free H_2O	165	
T4 DNA ligase buffer, 10×	20	1×
T4 DNA ligase, 120 U $µl^{-1}$	10	6 U $µl^{-1}$
SOLiD sequencing oligos		
(dark purple tube from the SOLiD ToP sequencing kit)	5	
Total	200	

26| Wash the sample four times for 5 min each with 1 ml of 1× instrument buffer at room temperature.

27| Acquire images.
▲ **CRITICAL STEP** The first ligation cycle for recessed primers *N*–2, *N*–3 and *N*–4 produces a fluorescence signal in just one channel. These images should not be included in the final data set.
? TROUBLESHOOTING

28| Aspirate and cleave the fluorophore by incubating the sample two times for 5 min each in cleave solution 1, and then two times for 5 min each in reconstituted cleave solution 2.1.
? TROUBLESHOOTING

29| Aspirate and wash the sample three times for 5 min each with 1 ml of 1× instrument buffer.
■ **PAUSE POINT** The sample is stable for 2–3 d in 1× instrument buffer at room temperature.

30| Repeat Steps 25–29 up to a total of seven cycles.

31| Incubate the sample four times for 5 min each in 1 ml of strip buffer, preheated to 80 °C.

32| Wash the sample two times for 1 min each with 1 ml of 1× instrument buffer.
■ **PAUSE POINT** The sample is stable for 2–3 d in 1× instrument buffer at room temperature.

33| Repeat Steps 23–32 using different sequencing primers (*N*–1, *N*–2, *N*–3 and *N*–4).
? TROUBLESHOOTING

Image pre-processing ● TIMING 6–12 h
34| If necessary, use ImageJ to crop image stacks for faster 3D deconvolution.

35| Determine the optimal 3D deconvolution parameters using a smaller cropped test image from the experiment. In Huygens Professional, we typically use a Nyquist sampling rate of 1.7, CMLE mode, 5–10 iterations and a signal-to-noise ratio of 2–5.

36| Deconvolve all sequencing images, and save images as .ics/.ids files with the following names in a folder named 'decon_images' (**Supplementary Fig. 4**). Filename: <Position>_<Primer #>_<Ligation #>_<Date__Time>.<ext>; Position: dinucleotide position as two-digit integers 01 to 30; Primer number: N followed by one-digit integers N0 to N4; Cycle number: ligation cycle per primer from 1 to 7; Date/time: An alphanumeric string using underscores; and File extension: .ics and .ids.
? TROUBLESHOOTING

Image analysis ● TIMING 6–12 h
▲ **CRITICAL** Some users of our method may have little or no background in bioinformatics. Here we introduce common computational environments and tools, but novice users should obtain additional help from experienced users, network administrators and online resources (i.e., http://www.ee.surrey.ac.uk/Teaching/Unix/).

37| Download fisseq.zip (http://arep.med.harvard.edu/FISSEQ_Nature_Protocols_2014/) and copy it to a remote host using a command-line terminal on a PC.

```
local:~$ scp fisseq.zip <user@remote_host_name:~/>
```

▲ **CRITICAL STEP** One must have an account to a designated remote host. Ask the network administrator at your institution.

38| Download and unzip decon_images.zip (http://arep.med.harvard.edu/FISSEQ_Nature_Protocols_2014/). Copy the decon_images folder (Step 36) to a scratch space on the remote host.

```
local:~$ scp -r ~/decon_images/ <user@remote_host_name:scratch_space/>
```

▲ **CRITICAL STEP** Analysis of multiple high-resolution image stacks requires a large amount of disk space. Contact your network administrator for the location of a temporary scratch space.

39| Log on to the remote host and submit a job request to work on a high-memory queue interactively. We recommend at least 100 GB (memory below is in MB).

```
local:~$ ssh <user@remote_host_name>
remote:~$ bsub -R "rusage[mem=100000]" -q <queue_name> -Is bash
```

▲ **CRITICAL STEP** Running CPU or memory-intensive tasks incorrectly can bring down the remote host. Make sure that you are working on a designated node. Contact your network administrator for more information before proceeding.

40| Unzip fisseq.zip and change the working directory to fisseq.

```
remote:~$ unzip fisseq.zip
remote:~$ cd fisseq
```

▲ **CRITICAL STEP** Working from folders other than ~/fisseq results in missing file errors when entering our commands as written below.

41| Download and decompress the RefSeq-to-Gene ID conversion table.

```
remote:~/fisseq$ wget ftp://ftp.ncbi.nih.gov/gene/DATA/gene2refseq.gz
remote:~/fisseq$ gzip -d gene2refseq.gz
```

42| Download the organism-specific RefSeq RNA FASTA file and unzip the file.

```
remote:~/fisseq$ wget
ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/mRNA_Prot/human.rna.fna.gz
remote:~/fisseq$ gzip -d human.rna.fna.gz
## Use the following address for mouse or rat:
## ftp.ncbi.nlm.nih.gov/refseq/M_musculus/mRNA_Prot/mouse.rna.fna.gz
## ftp.ncbi.nlm.nih.gov/refseq/R_norvegicus/mRNA_Prot/rat.rna.fna.gz
```

43| Build the reference index of [ref_name] in color space. Here [ref_name] is refseq_human. This process can take several hours.

```
remote:~/fisseq$ bowtie-build -C -f human.rna.fna refseq_human
```

? TROUBLESHOOTING

44| Start MATLAB and add a search path:

```
remote:~/fisseq$ matlab
>> addpath('~/fisseq', '~/fisseq/bfmatlab')
```

45| Define the input and output directories and run image registration (**Supplementary Fig. 5**).

```
>> input_dir='<scratch_space>/decon_images/'
>> output_dir='registered_images/'
>> register_FISSEQ_images(input_dir,output_dir,10,0.1,1)
>> quit()
```

Set the number of blocks per axis for local registration (default = 10); set the fraction overlap between neighboring blocks (default = 0.1); and adjust the alignment precision, where 10 will register images to 1/10 of a pixel (default = 1).
? TROUBLESHOOTING

46| Copy files in ~/fisseq/registered_images/ to a PC. Use ImageJ to open TIFF files (File > Import > Bio-Formats) as a time series, and check alignment in channel 4 by scrolling through the timeline (**Supplementary Videos 1–4**). Maximum-projected TIFF files (channel 4 is a composite of channels 0–3); Routput.mat file: block-wise registration offsets between bases; Rchadj.mat file: block-wise chromatics shifts as a matrix; and Rtadj.mat file: registration offsets over time for the whole image (not block-wise).
? TROUBLESHOOTING

47| Start python, and write base calls to read_data_*.csfasta. The maximum number of missing base calls allowed per read is 6 by default. * denotes an automatically generated time stamp.

```
remote:~/fisseq$ python

>>> import FISSEQ

>>> FISSEQ.ImageData('registered_images','.',6)

>>> quit()
```

? TROUBLESHOOTING

48| Align reads to refseq_human (Step 43) using Bowtie 1.0 or earlier, and write mapped reads to bowtie_output.txt. The exact name of read_data_*.csfasta can be determined by listing files in the directory (ls -l).

```
remote:~/fisseq$ bowtie -C -n 3 -l 15 -e 240 -a -p 12 -m 20 --chunkmbs 200 -f --best
--strata --refidx refseq_human read_data_*.csfasta bowtie_output.txt
```

? TROUBLESHOOTING

49| Spatially cluster the Bowtie reads (Step 48), annotate clusters using gene2refseq (Step 41) and write to results.tsv. The default kernel size of 3 performs a 3 × 3 dilation before clustering.

```
remote:~/fisseq$ python

>>> import FISSEQ

>>> G = FISSEQ.ImageData('registered_images',None,6)

>>> FISSEQ.AlignmentData('bowtie_output.txt',3,G,'results.tsv',

   'human.rna.fna','gene2refseq','9606')

>>> quit()

## Use the following command for mouse or rat:

>>> FISSEQ.AlignmentData('bowtie_output.txt',3,G,'results.tsv',

   'mouse.rna.fna','gene2refseq','10090')

>>> FISSEQ.AlignmentData('bowtie_output.txt',3,G,'results.tsv',

   'rat.rna.fna','gene2refseq','10116')
```

Data analysis ● TIMING 1 d
▲ CRITICAL Data analysis can be done on any software package, but R is convenient for interactive analysis and high-quality graphs[23]. Novice users may find RStudio more intuitive than the command-line interface. We provide a sample R session containing a sample data set and a list of commands (http://arep.med.harvard.edu/FISSEQ_Nature_Protocols_2014/).

50| Open the FISSEQ RStudio project file (*Menu* → *File* → *Open project*…).

51| Find the HISTORY tab on the upper right console window, and double-click on individual commands in order to re-execute the previous R session (**Supplementary Fig. 6**) and learn how to: import and filter data using a specific criterion (i.e., cluster size); plot a distribution of reads by a specific criterion (i.e., RNA classes and strands); convert a table of reads into a table of gene expression level; correlate gene expression from different images; and find statistically enriched genes in different regions.
? TROUBLESHOOTING

? TROUBLESHOOTING
Troubleshooting advice can be found in **Table 3**.

TABLE 3 | Troubleshooting table.

Step	Problem	Possible reason	Solution
1	Cells wash away during PBS washes or fixation	Cell attachment dependent on Ca^{++}	Use PBS with Ca^{++}, or add formalin directly to the growth medium
	Membrane blebs	Diluting formalin using 1× PBS makes it hypotonic	Use 10× PBS and water to dilute the formalin
	Tissue sections falling apart or off slide	Pepsin overdigestion and/or small contact area between sample and glass	Shorter pepsin digestion; embed in Matrigel and re-fix
19	Few amplicons limited to the cell surface	Poor cell permeabilization	Use 0.1 N HCl after Triton or ethanol-based cell permeabilization
	Amplicons in no-RT control	RT primer is too long	Use shorter RT primers
	High background	Excess RCA primer	More stringent washes at Step 13
	Dim amplicons	Low template copy number per amplicon	More dNTP and ϕ29 enzyme at Step 14
19,27	Dim, fuzzy or stretched amplicons	Poor cross-linking	Fresh BS(PEG)9 at Step 15
28	White precipitate buildup	Silver reacts with chloride	Eliminate chloride-containing buffers
33	Progressive loss of signal	Photodamage to amplicons	Low laser exposure
36	Deconvolution takes too much time	Large images	Crop unused areas Smaller images Fewer iterations
43	Bowtie command is not found	Bowtie v1.0 environment is not set up	Check Bowtie version (which bowtie); ask administrator for assistance
45	MATLAB out of memory error	Low RAM Low heap space for Java virtual memory (VM)	Allocate >100 GB RAM Increase Java VM in java.opts
	Input or output folders are not found	Incorrect slash use with folder name	No slash before and one slash after
46	ImageJ does not open TIFF files correctly	Image dimensions are not correctly read	Check 'Group files…', 'Swap dim…' and 'Concatenate…' when importing
47	Cannot find input images	Undefined path	Registered image directory must be in ~/fisseq
	Extension error messages	Missing package	Use Canopy Python 2.7

(continued)

TABLE 3 | Troubleshooting table (continued).

Step	Problem	Possible reason	Solution
48	Bowtie is not found	Bowtie 1.0 is not loaded	Check available versions and load Bowtie 1.0 or earlier versions
	Extra parameter(s) error	Typo, or option flags are not in the correct order	Copy and paste the command from Step 48
51	Unexpectedly high number of anti-sense mRNA reads	Noisy image, many missing or incorrect base calls at the 3′ end	Obtain better images, use deconvolution to reduce noise, sequence longer reads, trim reads or increase the cluster size threshold

● **TIMING**

Steps 1–21, FISSEQ library construction: 2–3 d
Steps 22–33, sequencing and imaging: 10 d
Steps 34–36, image pre-processing: 6–12 h
Steps 37–49, image analysis: 6–12 h
Steps 50 and 51, data analysis: 1 d

ANTICIPATED RESULTS

The size of subcellular cDNA amplicons is slightly larger than the diffraction limit after 3D deconvolution. At 20× NA 0.75, the diameter of cDNA amplicons is ~400–800 nm after image deconvolution. A typical amplicon contains hundreds of fluorescent probe–binding sites, and this results in images that are 20–50 times brighter and that have a markedly improved signal-to-noise ratio than single-molecule FISH. A good FISSEQ library should yield many intensely bright amplicons that are distinct from cell debris and spurious amplification products. If long exposure time and high gain have to be used to visualize objects, it is likely that they represent contamination, reaction precipitates or cell debris.

When fluorescent probes are stripped, nearly all of the fluorescence is completely removed, except possibly in the nucleus. Stripping is a good way to distinguish a DNA amplicon from fluorescent debris, and we recommend alternately hybridizing the sample with FAM, Cy3 or Cy5 probes while the sample is still on the microscope. If the fluorescent object is a DNA amplicon, it should fluoresce in distinct colors sequentially with little or no cross-talk. The amplicon density varies depending on the cell size, but we typically see several hundreds of amplicons per cell in cultured cell lines (i.e., iPSCs, fibroblasts, HeLa cells and bipolar neurons). We have detected up to 4,000 amplicons using synthetic DNA per cell in fibroblasts, suggesting that the RT efficiency may be a limiting factor.

The signal-to-noise ratio from SOLiD sequencing-by-ligation is high, especially for early ligation cycles. The quality drops after the fourth re-ligation cycle for each primer, and the image quality degrades significantly after 25 total cycles. Much of the image degradation results from the laser-induced damage during imaging. Typically, unimaged regions remain pristine even after 30 cycles of sequencing, and it may be possible to obtain a longer read length with appropriate free-radical scavengers in the imaging buffer, but we have not attempted this yet.

Depending on the camera sensor size, density and bit depth, one image stack containing multiple optical planes across four channels can be 800 MB–2 GB per field of view. Our image registration software then creates a separate folder containing TIFF images (five channels per base) of 20–50 M in size. Once our software processes and analyzes the images, it generates a tab-delimited file containing the gene ID, name, cluster size, strand, class, base quality, alignment quality, color space sequence and *x*-*y* position. We recommend performing a quick data check by selecting a gene cluster size of >5 to compare the number of sense and anti-sense reads, and we also recommend comparing the number of reads from different RNA classes. Typically, >90% of all reads should map to the positive sense strand. The rRNA read should comprise 50–80% of the total number of reads. We typically get 15,000–40,000 reads per image containing 30–50 cells. Regional or subcellular localization is measured in statistically significant enrichment scores, rather than absolute counts, owing to a small number of reads distributed over a large area. We recommend making B&W image masks on the basis of the cell morphology, DAPI stains, immunohistochemistry and other types of spatial masks, and measuring the relative enrichment of individual genes using Fisher's exact test or other similar tests[23]. With a high read density, it may be possible to use unsupervised local clustering of reads for regional identification of biological processes[2].

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS This study was funded by US National Institutes of Health (NIH) Centers of Excellence in Genomic Sciences (CEGS) grant no. P50 HG005550. J.H.L. and co-workers were funded by National Heart, Blood and Lung Institute (NHBLI) grant no. RC2HL102815, by the Allen Institute for Brain Science and by National Institute of Mental Health (NIMH) grant no. MH098977. E.R.D. was funded by NIH grant no. GM080177 and by National Science Foundation (NSF) Graduate Research Fellowship grant no. DGE1144152.

AUTHOR CONTRIBUTIONS J.H.L. and E.R.D. conceived FISSEQ library construction, sequencing, image analysis and bioinformatics. J.S., R.K., J.L.Y., B.M.T., H.S.L. and J.A. provided key feedbacks during the FISSEQ method development. R.T. and T.C.F. assisted with automated microscopy and image analysis. K.Z. and G.M.C. oversaw the project. J.H.L. wrote the paper, and E.R.D. wrote the FISSEQ software.

COMPETING FINANCIAL INTERESTS The authors declare competing financial interests: details are available in the online version of the paper.

1. Rifai, N., Gillette, M.A. & Carr, S.A. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat. Biotechnol.* **24**, 971–983 (2006).
2. Jaitin, D.A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
3. Battich, N., Stoeger, T. & Pelkmans, L. Image-based transcriptomics in thousands of single human cells at single-molecule resolution. *Nat. Methods* **10**, 1127–1133 (2013).
4. Lein, E.S. *et al.* Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168–176 (2007).
5. Zeng, H. *et al.* Large-scale cellular-resolution gene profiling in human neocortex reveals species-specific molecular signatures. *Cell* **149**, 483–496 (2012).
6. Diez-Roux, G. *et al.* A high-resolution anatomical atlas of the transcriptome in the mouse embryo. *PLoS Biol.* **9**, e1000582 (2011).
7. Femino, A.M., Fay, F.S., Fogarty, K. & Singer, R.H. Visualization of single RNA transcripts *in situ*. *Science* **280**, 585–590 (1998).
8. Raj, A., van den Bogaard, P., Rifkin, S.A., van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* **5**, 877–879 (2008).
9. Levsky, J.M., Shenoy, S.M., Pezo, R.C. & Singer, R.H. Single-cell gene expression profiling. *Science* **297**, 836–840 (2002).
10. Lubeck, E., Coskun, A.F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell *in situ* RNA profiling by sequential hybridization. *Nat. Methods* **11**, 360–361 (2014).
11. Choi, H.M. *et al.* Programmable *in situ* amplification for multiplexed imaging of mRNA expression. *Nat. Biotechnol.* **28**, 1208–1212 (2011).
12. Ke, R. *et al. In situ* sequencing for RNA analysis in preserved tissue and cells. *Nat. Methods* **10**, 857–860 (2013).
13. Larsson, C., Grundberg, I., Söderberg, O. & Nilsson, M. *In situ* detection and genotyping of individual mRNA molecules. *Nat. Methods* **7**, 395–397 (2010).
14. Larsson, C. *et al. In situ* genotyping individual DNA molecules by target-primed rolling-circle amplification of padlock probes. *Nat. Methods* **1**, 227–232 (2004).
15. Lagunavicius, A. *et al.* Novel application of Phi29 DNA polymerase: RNA detection and analysis *in vitro* and *in situ* by target RNA-primed RCA. *RNA* **15**, 765–771 (2009).
16. Merkiene, E., Gaidamaviciute, E., Riauba, L., Janulaitis, A. & Lagunavicius, A. Direct detection of RNA *in vitro* and *in situ* by target-primed RCA: the impact of *E. coli* RNase III on the detection efficiency of RNA sequences distanced far from the 3′-end. *RNA* **16**, 1508–1515 (2010).
17. Lee, J.H. *et al.* A robust approach to identifying tissue-specific gene expression regulatory variants using personalized human induced pluripotent stem cells. *PLoS Genet.* **5**, e1000718 (2009).
18. Shendure, J. *et al.* Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732 (2005).
19. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).
20. Grun, D., Kester, L. & van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nat. Methods* **11**, 637–640 (2014).
21. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* **14**, 618–630 (2013).
22. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11**, 163–166 (2014).
23. Lee, J.H. *et al.* Highly multiplexed subcellular RNA sequencing *in situ*. *Science* **343**, 1360–1363 (2014).
24. Adiconis, X. *et al.* Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nat. Methods* **10**, 623–629 (2013).
25. Yachida, S. *et al.* Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467**, 1114–1117 (2010).
26. Frumkin, D. *et al.* Amplification of multiple genomic loci from single cells isolated by laser micro-dissection of tissues. *BMC Biotechnol.* **8**, 17 (2008).
27. Lovatt, D. *et al.* Transcriptome *in vivo* analysis (TIVA) of spatially defined single cells in live tissue. *Nat. Methods* **11**, 190–196 (2014).
28. Schmid, M.W. *et al.* A powerful method for transcriptional profiling of specific cell types in eukaryotes: laser-assisted microdissection and RNA sequencing. *PLoS ONE* **7**, e29685 (2012).
29. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* **21**, 1160–1167 (2011).
30. Ramskold, D. *et al.* Full-length mRNA-seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).
31. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-seq: single-cell RNA-seq by multiplexed linear amplification. *Cell Rep.* **2**, 666–673 (2012).
32. Avital, G., Hashimshony, T. & Yanai, I. Seeing is believing: new methods for *in situ* single-cell transcriptomics. *Genome Biol.* **15**, 110 (2014).
33. Buxbaum, A.R., Wu, B. & Singer, R.H. Single β-actin mRNA detection in neurons reveals a mechanism for regulating its translatability. *Science* **343**, 419–422 (2014).
34. Hanna, J. *et al.* Direct cell reprogramming is a stochastic process amenable to acceleration. *Nature* **462**, 595–601 (2009).
35. Buganim, Y. *et al.* Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell* **150**, 1209–1222 (2012).
36. Itzkovitz, S., Blat, I.C., Jacks, T., Clevers, H. & van Oudenaarden, A. Optimality in the development of intestinal crypts. *Cell* **148**, 608–619 (2012).
37. Hansen, C.H. & van Oudenaarden, A. Allele-specific detection of single mRNA molecules *in situ*. *Nat. Methods* **10**, 869–871 (2013).
38. Porreca, G.J. *et al.* Multiplex amplification of large sets of human exons. *Nat. Methods* **4**, 931–936 (2007).
39. Kosuri, S. *et al.* Composability of regulatory sequences controlling transcription and translation in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **110**, 14024–14029 (2013).
40. Li, J.B. *et al.* Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing. *Science* **324**, 1210–1213 (2009).
41. Zhang, K. *et al.* Digital RNA allelotyping reveals tissue-specific and allele-specific gene expression in human. *Nat. Methods* **6**, 613–618 (2009).
42. Michael, W.M., Choi, M. & Dreyfuss, G. A nuclear export signal in hnRNP A1: a signal-mediated, temperature-dependent nuclear protein export pathway. *Cell* **83**, 415–422 (1995).
43. Kaposi-Novak, P., Lee, J.S., Mikaelyan, A., Patel, V. & Thorgeirsson, S.S. Oligonucleotide microarray analysis of aminoallyl-labeled cDNA targets from linear RNA amplification. *Biotechniques* **37**, 580, 582–586, 588 (2004).
44. Nanda, J.S. & Lorsch, J.R. Labeling a protein with fluorophores using NHS ester derivitization. *Methods Enzymol.* **536**, 87–94 (2014).
45. Mardis, E.R. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* **9**, 387–402 (2008).
46. Massingham, T. & Goldman, N. Error-correcting properties of the SOLiD Exact Call Chemistry. *BMC Bioinformatics* **13**, 145 (2012).
47. Applied Biosystems. *SOLiD System Accuracy with the Exact Call Chemistry Module* (https://www3.appliedbiosystems.com/cms/groups/global_marketing_group/documents/generaldocuments/cms_091372.pdf).
48. Itzkovitz, S. & van Oudenaarden, A. Validating transcripts with probes and imaging technology. *Nat. Methods* **8**, S12–S19 (2011).
49. Eliceiri, K.W. *et al.* Biological imaging software tools. *Nat. Methods* **9**, 697–710 (2012).
50. Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
51. Kankaanpaa, P. *et al.* BioImageXD: an open, general-purpose and high-throughput image-processing platform. *Nat. Methods* **9**, 683–689 (2012).
52. Pawley, J.B. *Handbook of Biological Confocal Microscopy* 3rd edn. (Springer, 2006).