Whole organism lineage tracing by combinatorial and cumulative genome editing

Aaron McKenna

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

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Jay Shendure, Chair

Brian Reid

Marshall Horwitz

Program Authorized to Offer Degree:
Department of Genome Sciences

University of Washington

Abstract

Whole organism lineage tracing by combinatorial and cumulative genome editing

Aaron McKenna

Chair of the Supervisory Committee:
Professor Jay Shendure
Genome Sciences

Each of us begins life as a single fertilized cell, or zygote. The act of fertilization brings about an endeavored series of cell divisions, ultimately resulting in the forty trillion cells in each adult human. How are these cell divisions coordinated to produce the sophisticated organs and tissues of the body? This is a long-standing and open question in biology. Many fate-mapping and lineage tracing techniques have been developed to sample from the underlying developmental process, but we are still far from a holistic lineage map of vertebrates or mammals. Such a map would be transformative, placing physiology, cell biology, genomics, and anatomy onto a unified scaffold of development.

Despite the enormous strides made over the last one hundred years, our understanding of lineage remains enormously fragmented for most organisms. Only one animal has had its lineage fully described, the thousand-cell worm *Caenorhabditis elegans*. There have been many challenges in expanding to larger, more complex animals: the orders of magnitude increase in cell numbers, the stochastic nature of development, and the non-transparency of the target animals. Many technological advances have been applied to further refine existing lineage maps, including microscopy, vital dyes, and genetic markers, but none of these technologies scale to the number of cells in even the simplest vertebrates. My goal in this dissertation is to harness technologies in the nascent field of genome engineering to encode information into individual cells, with the ultimate goal of recording whole organism

lineage maps.

In my first chapter, I describe methods and tools to harness the power of the CRISPR genome editing technology. I detail an approach for rapidly discovering target sites within arbitrary genome segments, aggregating potential off-target sequences, and ranking the results using a wide array of community developed scoring schemes. I then apply this method to characterize the enhancer region of the human gene *MYC*, as well as to aid in the design of a deletion scan for the regulatory region of the *HPRT1* gene. The underlying tool for finding and characterizing CRISPR sites, FlashFry, has also been released to the community.

I then leverage these computational tools in conjunction with published work in the field to generate a compact lineage recording technology. I combine multiple CRISPR targets into a compact barcode, and read this barcode out of individual cells in both RNA and DNA. This system is then shown to accumulate a diversity of alleles, a prerequisite for lineage tracing. Using a synthetic lineage created in cell culture we test the reconstructive power of our approach, and then apply this technology to recover the lineage of both embryonic and adult *Danio rerio* (zebrafish). Leveraging the large embryo size, I directly inject our marking reagents at the single cell stage, and observe diverse alleles through progressive stages of development. I then recover editing patterns consistent with marking of cells as early as the two cell stage. Using techniques borrowed from phylogenetics, I then reconstruct coarse lineage trees from many embryonic stages, recovering trees with upwards of 8 layers of cell division. In the adult zebrafish I successfully recover germ layer and organ relationships, and observe a lineage bottleneck for all the organ systems sampled.

TABLE OF CONTENTS

LIST OF FIGURES

GLOSSARY

ALLELE: A unique GESTALT barcode sequence, observed in one or more cells

BARCODE: An array of CRISPR/Cas9 target sites for genome editing to mark lineages

CAS9: A type II CRISPR protein first isolated from *Streptococcus pyogenes*

CRISPR: Clustered regularly interspaced short palindromic repeats. A bacterial defense system, used in genome engineering to create programmed double-stranded breaks

GUIDE: A short RNA composed of a scaffold sequence necessary for CRISPR protein binding and protospacer sequence that matches the intended target sequence

GESTALT: *Genome editing of synthetic target arrays for lineage tracing* (GESTALT), our lineage tracing technology based on CRISPR/Cas9

HPRT1: Hypoxanthine(-guanine) phosphoribosyltransferase, the X-linked housekeeping gene underlying Lesch-Nyhan syndrome, a recessive Mendelian disorder

PAM: The protospacer adjacent motif (PAM) is a short sequence motif recognized by the CRISPR nuclease. This must be present in the DNA sequence for binding

PROTOSPACER: A short RNA sequence (20 bases in Cas9) loaded in the CRISPR protein that confers the specificity for the target site

PROSPECTIVE LINEAGE TRACING: Lineage tracing technologies that mark individual cells or cell populations ahead of time, and recover those marks later in development

RETROPROSPECTIVE LINEAGE TRACING: Lineage tracing technologies that rely on naturally occurring marks. The marks appear stochastically throughout development, and are only sampled from at tissue collection

TARGET: A sequence within the genome of interest that matches the protospacer of the guide, and is flanked by the required PAM sequence of the CRISPR protein

UMI: Unique molecular identifier. UMIs are added in the first rounds of PCR, tagging individual molecules before standard PCR. Each UMI barcode should then correspond to a single fragment of input DNA

ACKNOWLEDGMENTS

I am deeply indebted to so many people who have offered their emotional, financial, and scientific support during my time here at the University of Washington.

I would like to first thank my supervisor and mentor Jay Shendure for his tremendous support throughout my PhD. He had the generosity to take in a computational trainee with no lab experience, the wisdom to guide me into a deeply meaningful area of work, and the insight and motivation to drive that work to completion. He is a tireless supporter of his trainees, something I will be forever grateful for.

I would also like to thank the members of my PhD committee who have provided critical support throughout my dissertation. I am grateful to Marshall Horwitz who brought me into the field of lineage tracing, to which he had contributed deeply, and provided countless insights when we needed them the most. Brian Reid has indulged so many of my wandering conversations on cancer genetics, tumor heterogeneity, and evolution, and has been a wonderful sounding board throughout all of this. Bob Waterston and Stan Fields always had their door open to me when I needed either project or career advice.

The Shendure lab has been a wonderful place to learn and grow over the last four years. I am eternally grateful to Riza Daza, who spent months training this naive software engineer in countless experimental techniques. I also owe a huge debt to Ruolan Qiu and Beth Martin, who provided key experimental assistance, and have answered an endless number of questions without asking anything in return. Choli Lee has put with my failed sequencing runs, bizarre requests, and demanding schedules without a complaint. Thanks Choli. Thanks to Andrew Adey, Akash Kumar, Steve Salipante, Molly Gasperini, Matthew Snyder, Lea Starita, Seungsoo Kim, Vijay Ramani, and Ron Hause for making the the lab an exciting

and fun place to spend way too much of my time. I also want to thank Ben Vernot for his friendship, and for being the one person in the building who was willing tell me when I screwed up. Lastly and most importantly I want to thank Greg Findlay for his huge contribution to this project at a critical juncture, and his endless stream of sound advice. Without him I don't think I'd be writing this dissertation. Thanks for everything Greg.

I'd also like to thank my family for their endless support through this rather intense period in my life; your unwavering support and optimism kept me afloat. Also thanks to Spec, our crazy dog. Our morning walks through Ravenna park have centered me on so many tough days. You'll get those squirrels yet.

Lastly I'd like to thanks my wife Carrie, who put up with endless late nights, weekends in the lab, and all the while cheered me on through countless failures. None of this would have happened without you. I love you dearly.

DEDICATION

This work is dedicated to my parents Linda and Peter McKenna.

Mom you have been my most tireless supporter, and your endless optimism has lifted me in so many dark moments.

Despite his medical training my dad was a scientist at heart, and would have so deeply enjoyed this. I cherish the memory of touring undergraduate schools, and without fail having to track him down, wandering somewhere in a science department. We all miss you Dad.

Chapter 1

INTRODUCTION

Figure 1.1 has been adapted from: McKenna A, Findlay GM, Gagnon JA, Horwitz MS, Schier AF, and Shendure J. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* 29 July 2016

Every cell in your body is the result of a cell division, and has a lineage that traces back through successive mitoses to the fertilized embryo. Throughout each of these cell divisions a concerted program of development is underway, creating the complex patterns and hierarchies of organs and tissues. Our goal in lineage tracing is to describe this process at the single-cell level, creating a branching, directed graph of individual cell divisions called a lineage tree, or a fate map. Such a tree provides a scaffold on which biologists can arrange the rich descriptions of cellular state in both normal development and pathological disease states. Unfortunately we currently lack such a tree in vertebrate or mammalian model organisms.

In this chapter I review the field of lineage tracing. I describe the major milestones over the last 150 years, and the biological insight gained by using lineage tracing in various model organisms. I give the most focus to the technologies themselves, especially the prospective and retrospective techniques developed in the last two decades. Due to space limitations some methodologies such as cell engraftment are omitted, and left as explorations for the curious reader. Lastly I detail some of the information storage underpinnings of recent technological advances and their potential application to whole-organism fate maps.

1.1 Lineage tracing by direct observation

Advances in biology are often dictated by the advances in technology. In the nineteenth century the preeminent tool used to study biology was the light microscope. The experiments of Pasteur had effectively disproved the theory of spontaneous generation and replaced it with the theory of biogenesis. "omnis cellula e cellula" – all cells from cells (Amos, 2000). Leveraging this theory, biologists began to explore cell biology and development in a new light. Nageli, Hofmeister, and Strasburger observed mitosis using a light microscope, followed by Boveri's observation of meiosis in the nematode *Ascaris megalocephala* (Amos, 2000). But how could these simple cell divisions be combined and orchestrated to create the diversity of the natural world? To understand this process, biologists would have to track individual cells from the genesis of life through adulthood.

Biologists began this effort by observing life at the point of inception, the fertilized ovum. Pioneers like Charles O. Whitman studied the early cell cleavages of the leech embryo using light microscopes (Moore, 1993). Here Whitman was able to track the first two cell divisions of the leech, resulting in four equally sized cells, which he labeled *a*, *b*, *c*, and *x*. These four cells split into pairs of small and large cells, with the smaller set becoming the basis for the ectoderm of the leach. Follow-on work by Edwin Conklin in the Ascidian egg went further into development, tracing cell lineages for eight cycles of division (representing 218 cells), and into one of the most critical stages of development, gastrulation. Importantly, his observations of early cell division patterns mirrored what Whitman had seen in the leach (Moore, 1999).

Direct observation also played a central role in the most famous lineage tracing experiment of the last century, the whole organism fate map of the flatworm *Caenorhabditis elegans*. Technological improvements in the form of Nomarski DIC microscopy enabled Sulston and colleagues to track individual worm cells *in situ* (Sulston, 2003). Using this technology they watched individual cells progress through successive cell division, tracing each cell lineage to its terminal differentiation or death. When they integrated these observations over many

C. elegans, a clear pattern emerged. The lineage tree of these eutelic worms was invariant: the progeny of every division was reproducible across all members of the species. This fully resolved lineage tree was transformative, as it provides a developmental scaffold for any *Caenorhabditis elegans* work.

The works of Sulston, Whitman, and Conklin demonstrate an importance lesson in lineage tracing: to be successful, your technology must scale to the biological question at hand. Their efforts were possible given the available technologies and the choice of organism: they all were transparent, they had a traceable number of cells to study, and had highly or fully reproducible lineages. In Conklin's own works: "The cleave of the egg is beautifully regular and can be observed so readily in life that it is not surprising that ascidians were among the first animals to which the 'cell-lineage' method was applied" (Conklin, 1905). Direct observation is still in active use in such animals, and has been repurposed in newer approaches such as lineage tracing using *in situ* sequencing (Frieda et al., 2017). Unfortunately the vast majority of development is inaccessible to visual inspection, and new technologies had to be developed to peer into the complex structure of vertebrate and mammalian cell lineages.

1.2 Prospective lineage tracing

The fact that direct observation would be impossible for so many organisms of interest was appreciated even at the turn of the 20th century. In conjunction with visual inspection, new systems were devised to tag individual cells in a way that these introduced marks were retained by all cellular progeny (fig. 1.1). Initial approaches by Vogt and others used "vital dyes", embedded into agar chips, which were taken up by cells of interest. Cells that carry the mark later in development can be inferred to be descendants of the original tagged cell population. The challenge is the marking agent: these marks eventually dilute out, limiting the developmental windows that could be studied. Additionally, these marks were specific only to a large region of cells, and did not mark individual lineages. Many other chemical marking strategies have been used to overcome some of these limitations, from horseradish peroxidase to BrdU nucleotide incorporation (Kretzschmar and Watt, 2012).

One recent innovation is to tag individual cells using lentiviral viral particles. When cells are transduced, the RNA viral genome is reverse transcribed and a payload containing a unique barcode is inserted randomly into the genome of the cell. These tagged subpopulations can then be monitored over time to characterize the dynamics of cell population outgrowth, competition, and cellular senescence. This system has been used to track cancer cell populations in culture, as well as to track mouse neuronal development(Walsh and Cepko, 1992; Porter et al., 2014). One issue is that many cell types are refractory to viral integration, including many postmitotic cell types, leaving blind spots in the resulting lineage map.

These cell-marking approaches have many drawbacks, especially when our goal is whole organism lineage tracing. First, these techniques require non-destructive access to the cell of interest during development. This eliminates studying cells that arise internally or in a stochastic fashion. These approaches are prospective, in that they require *a priori* knowledge of the cell to mark. Lastly, such approaches cannot delineate the relationship between cells that emerged from a single tagged clone. Once a cell is tagged, any progeny will contain the same identical mark or barcode. To resolve the whole lineage of a complex animal, an enormous number of timepoints and cells would need to be marked, making these techniques impracticable for anything but highly focused efforts.

1.3 Genetic Marking

Site-specific recombinases have been used in a variety of ways to encode information into the genome for lineage tracing. These recombinase enzymes recognize and bind to a pair of DNA sequence motifs, and when two sites are bound and colocalized, these proteins can either invert or excise the intervening segment of DNA depending on the relative orientation of the motifs (Grindley et al., 2006). Commonly recombinases include the Flp/FRT, Cre/loxP, and RCI systems. Their use in lineage tracing can be categorized into three main approaches, with increasing information storage and complexity: single marker systems, combinatorial color systems, and DNA shuffling systems.

Single marker systems are the simplest recombinase-based approach. Here a transcriptional stop signal is placed between a reporter gene and its transcriptional promoter. The stop signal is flanked on each side by recognition sites, and on expression of the recombinase, the stop signal is excised and the reporter gene is expressed. This reporter is commonly a fluorescent marker like GFP or mCherry, but can include other reporters like luciferase (Kretzschmar and Watt, 2012). Animal lines can then be generated with the marker gene integrated into the genome, and crossed with lines that express the recombinase either globally or in a tissue specific fashion. This simplifies the experimental setup, as new lines only have to be generated for the each activator of interest. The field has used this to great effect, enabling lineage tracing in a wide variety of tissues of disease models(Nagy, 2000). The challenge, much like in vital dyes, is knowing which tissues to mark and which expression system will work to activate lineage recording at the correct timepoint.

These single marker systems have recently been expanded to include multicolor systems, such as Brainbow and Confetti (Livet et al., 2007; Snippert et al., 2010). Here the integrated cassette contains multiple fluorescent reporters. The activation of the recombinase randomly inverts one color into the path of transcription. These systems then use the mixing of colors from many stochastic inversion cassettes to label cells with one of a hundred mixed colors. Although the information content of such systems is a huge step over single marker approaches, the complexity is still relatively low in comparison to the cell count in even the simplest animals. These approaches are also constrained to the spectrum of colors that can be captured using light microscopy.

Given the limited combinatorics of the color space mixing approaches above, other groups began to experiment with readouts of recombinase-shuffled DNA segments. Peikon et al.'s approach uses the RCI recombinase, which inverts DNA between two marker sequences without excision to shuffle a large DNA region (Peikon et al., 2014); we tried a nearly identical approach, but found both cloning and reliable RCI expression challenging (fig. 1.2). As a demonstration of it's capabilities, the authors used a cassette with 11 flippable regions. After recombination they recovered 1786 sequences, of which 96% were unique. The information

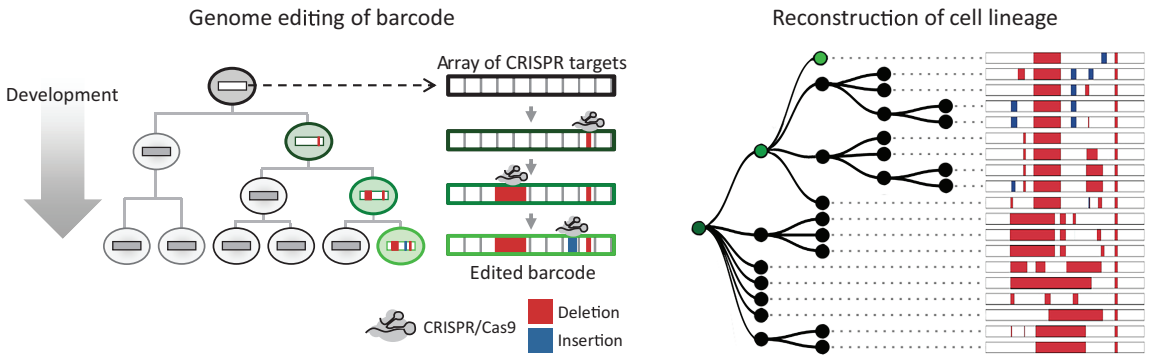


Figure 1.1: Reconstructing lineage with engineered marks. (Left) A barcode of CRISPR/Cas9 target sites is progressively edited over many cell divisions. (Right) Edited barcode sequences are related to one another on the basis of shared mutations in order to reconstruct lineage trees.

capacity of such a system is a huge leap over previous recombinase systems (the theoretical diversity of their 11 region cassette is 176 million combinations), although they observe a much smaller set in their limited sample. There are two major challenges. The first challenge, much like the color based systems, is at the readout stage. Their 11 region cassette was over 1500 basepairs in length, forcing the authors to use PacBio's long-read technology to sequence each cassette haplotype. Future sequencing technology improvements will render this point moot, but there is a second more fundamental issue. Without excision, the RCI flanked regions are free to recombine through any possible set of states, including overlapping and reversible inversion paths. This bidirectional shuffling will seriously confound any lineage reconstruction as any mark is reversible. This system has great utility in generating diverse DNA segments in cell populations, which could be useful for other genome engineering projects or in combination with other lineage tracing systems.

1.4 CRISPR/Cas9 approaches

Our ability to freely modify DNA has been greatly expanded by the discovery and adaptation of the CRISPR prokaryotic immune system. A full treatment of the CRISPR system is out of the scope of this introduction, but readers are pointed to many great reviews on the subject (Sander and Joung, 2014; Doudna and Charpentier, 2014). Briefly, the prokaryotic immune proteins of the CRISPR family are nucleases that can be directed to make double stranded breaks at a programmed DNA sequence. Unlike the recombinases above, the recognized sequence is not solely determined by the protein structure. Instead it's determined in two parts. First the CRISPR protein scans the genome of interest for a very short motif required by the protein called the protospacer adjacent motif (PAM). For instance, the protein SpCas9 from *Streptococcus pyogenes* requires a 3' sequence motif of NGG. When a match is discovered in the genome, the protein checks its RNA protospacer or 'guide' for complementarily to the flanking bases. This protospacer is what makes the CRISPR/Cas9 systems programmable. Any sequence of sufficient length can be used to guide CRISPR/Cas9 to it's target, given the scaffolding sequence is appended.

Once the CRISPR/Cas9 complex has matched its programmed protospacer to the target sequence, it creates double-stranded breaks with high specificity and efficiency (Tsai et al., 2014). These double-stranded breaks are often resolved into small insertions or deletions (indels) after repair by the nonhomologous end-joining pathway (van Overbeek et al., 2016). These marks or 'edits' can serve as markers of lineage, in a similar fashion as the inverted or excised DNA in the recombinase approaches. Unlike the recombinase approaches, the inherent diversity of editing outcomes from a stochastic indel process has the potential to encode much more information at a single locus. In our work we see many diverse deletions and insertions at individual target sites, with thousands of editing combinations over a target array. These outcomes are stochastic, as we see little to no overlap between animals. These indels are also irreversible: once the target sequence is changed, Cas9 will no longer bind, preventing additional editing.

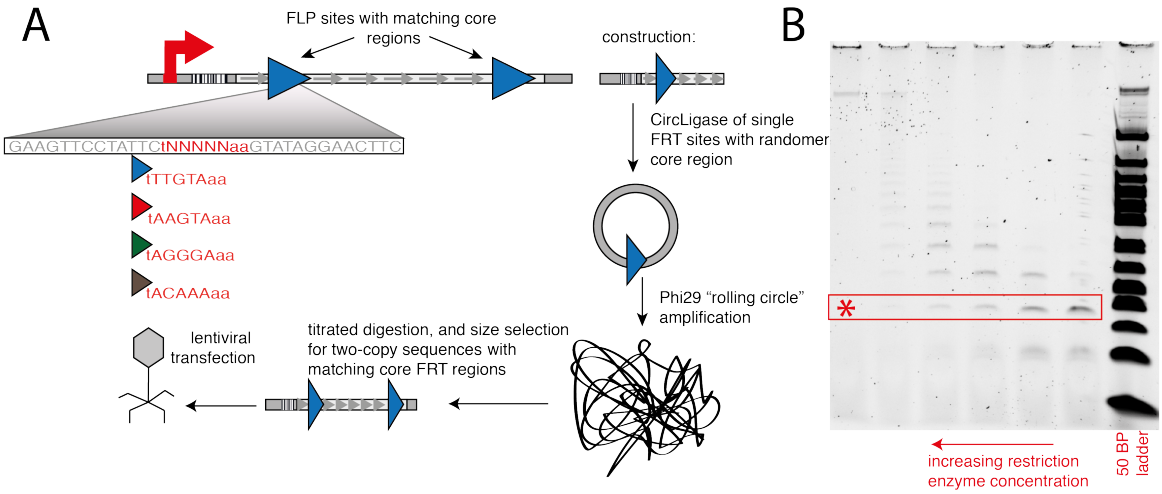


Figure 1.2: Generating diverse recombinase substrates for lineage tracing. **(A)** Recombinase proteins recognize a core motif, the orientation of which determines if the intervening DNA will be excised or inverted upon protein binding. We demonstrated a Phi29 polymerase 'rolling-circle' approach to generate degenerate libraries of motifs for RCI, FLP, or other common recombinases. **(B)** Size selection for DNA fragments with increasing numbers of motifs.

The field has used this exciting new technology to generate a number of innovative lineage tracing technologies. A general schematic of each technology is shown in figure 1.1. All of these CRISPR lineage tracing technologies rely on the CRISPR/Cas9 protein (specifically SpCas9, from *Streptococcus pyogenes*). These technologies can be subdivided into two approaches: those that use self-targeting guides, and those that separate expression of the guide from the target sequence.

The self-targeting technologies, mSCRIBE and Homing guides, both reengineer the RNA backbone sequence of the guide (Perli et al., 2016; Kalhor et al., 2016). This canonical crRNA backbone contains a conserved stem loop, and these reengineering efforts change the protospacer proximal bases to match the PAM sequence motif NGG (Perli et al., 2016; Kalhor et al., 2016). This enables the resulting Cas9/guide complex to self-recognize and generate double-stranded breaks in the guide's own DNA. Any indels created are then included in any subsequent transcripts. This process can be repeated; the DNA sequence can be edited again, creating an indel aggregation loop, which is terminated in one of two ways. First, a deletion can extend into the PAM sequence, ablating Cas9 binding. Second, deletions can shorten the distance between the promoter and the guide sequence, creating guides that are too short to activate editing, usually less than 17 bases (Zheng et al., 2017).

Given these limitations, how much information recording can we expect? Kalhor et al. characterized the information capacity of their captured homing guides, estimating that their system contained five bits of information per guide (Kalhor et al., 2016). Using this figure, they go on to claim six such homing guides could label all the neurons in an adult mouse. There is an important caveat here: in dynamic tagging approaches, lineage tracing is not just labeling. In both papers, deletions were the predominant outcome of double stranded break repair. These self-targeting systems then will lose relationship information over time, which makes relating individual barcodes challenging. These approaches would greatly benefit from methods that directed double-strand break repair to produce insertions in greater proportions.

The other class of CRISPR lineage tracing techniques use guides that are expressed

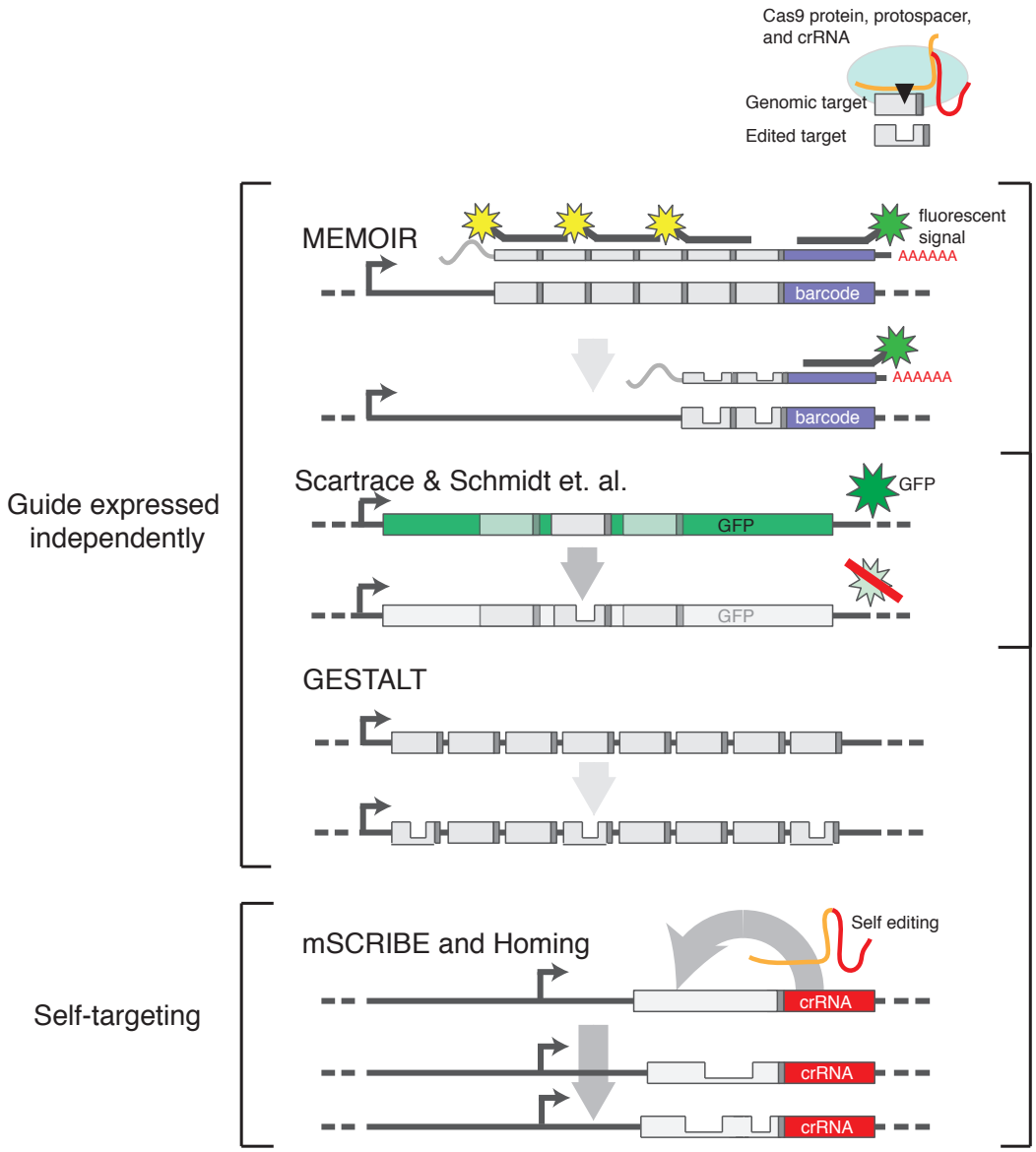


Figure 1.3: Catalog of CRISPR/Cas9 lineage tracing methods. Schematics for GESTALT, Scartrace, Schmidt et al., MEMOIR, mSCRIBE, and Homing guide lineage tracing technologies. CRISPR/Cas9 is used to introduce insertions or deletions at engineered target sites, and these genome perturbations are measured using DNA sequencing, *in situ* hybridization, or microscopy (McKenna et al., 2016; Frieda et al., 2017; Junker et al., 2017; Schmidt et al., 2017; Perli et al., 2016; Kalhor et al., 2016).

independently from their targeted loci. These approaches include ScarTrace, Schmidt et al., GESTALT, and MEMOIR (McKenna et al., 2016; Frieda et al., 2017; Junker et al., 2017; Schmidt et al., 2017). The first two, ScarTrace and Schmidt et al., use editing of a single target sequence within the coding sequence of GFP. Any out-of-frame indels acquired in double-stranded break repair will disrupt fluorescence (fig. 1.3). This approach links phenotype and genotype, allowing readouts of editing status using both visual inspection and sequencing technologies. Such gene-body targeting approaches are limited by the number of fluorescent proteins that can be engineered into an organism. If the fluorescent readout is ignored, then this technique is only limited by the total number of suitable Cas9 targets within each protein, although this wasn't explored by either group.

GESTALT and MEMOIR engineer target arrays into the genome of interest, and independently express a guide to edit that array. GESTALT, covered in great detail later, recovers the resulting indels using second-generation sequencing. To label individual cells, unique molecular identifiers (UMIs) are attached to the DNA target array in the first PCR cycle, followed by conventional PCR. This simple approach allowed us to recover the GESTALT barcodes of hundreds of thousands of cells. This approach is not without flaws. In some cases deletions can extend into PCR primer sites, obscuring cell lineages, and for recovered cells the complex editing patterns can make sequence alignment and indel recovery challenging. Lastly, GESTALT cannot assign spatial information to the barcode, an important feature of MEMOIR.

MEMOIR also uses an engineered locus, where a series of target sequences are arrayed in front of a static barcode (Frieda et al., 2017). This array or 'scratchpad' is interrogated using *in situ* single-cell analysis by sequential smFISH (SeqFISH) (Shah et al., 2017). When the target array is intact, a two-color signal is produced by hybridizing fluorescent oligos to first the scratchpad and then the barcode region. After CRISPR/Cas9 destroys the scratchpad region, probes will only hybridize to the barcode region. They were able to use MEMOIR to track the lineage of mouse ES cells through the first 3-4 divisions in culture, faithfully reproducing the lineage recovered from direct visual observation. A big advantage of MEMOIR

is that hybridization can be performed on intact tissues, which preserves important spatial information, on of the main challenges for GESTALT. Unfortunately in MEMOIR the binary scratchpads are uni-directional, and all will eventually fail to produce fluorescent signal from the scratchpad. Once a significant number are lost, no signal will be shared between cells and the relevant lineage information is lost.

An interesting aspect of the MEMOIR paper is the use of their scratchpad approach to record cellular signaling events. Here the authors tied expression of a separate guide and scratchpad sequence to the *Wnt* pathway, using an enhancer and promoter sequence activated by binding of the *Wnt3a* ligand. Such an approach could be expanded to record the dynamics of a much large set of key factors during development, many of which are transient in nature.

In aggregate there are many advantages of the CRISPR based lineage tracing systems compared to the previous generation of prospective lineage tracing technologies. First, all of the CRISPR systems record into a compact locus or set of loci across the genome. This simplifies the multiplexed recovery of these marks using sequencing technologies. These loci can also be expressed as RNA transcripts and captured using RNA-Seq, an advantage when working with emerging single-cell sequencing systems.

Many of the challenges in using these CRISPR based lineage approaches are shared with the other prospective approaches. First, system components must be integrated into the genome of interest. This can mean significant time and effort to breed and verify such systems in animal models, like the mouse. This also precludes the use of these systems in humans, a major limitation with emerging efforts such as the Human Cell Atlas. There is also the chance that such systems could be cytotoxic; we have observed large deletions in some of our targets that have not been well characterized. The stochastic nature of editing makes direct interrogation of the target site with *in situ* sequencing probes challenging. The authors of MEMOIR converted the CRISPR editing events into binary readouts, but more ingenuity is needed to capture the full spectrum of NHEJ repair outcomes at a single locus. In summary such CRISPR editing systems represent a clear step forward for the field of

lineage tracing, and the potential to combine these advancements with retrospective lineage tracing approaches could mean orders of magnitude increases in our cellular reconstruction power.

1.5 Endogenous, retrospective lineage tracing approaches

In comparison to the prospective techniques above, retrospective lineage tracing relies on an underlying mutational process to introduce divergent marks into an organism's genome. These mutations aren't programmed externally, but accumulate stochastically as cell divide and various lesions fix into the genome. All progeny of this mutationally-marked cell will inherit these lesions, and can also acquire additional mutations during subsequent cell divisions. Much of the groundwork for these retrospective lineage tracing approaches come from the field of cancer research, where the increased mutational frequency enabled early study.

Stanley Gartler and David Linder's 1965 seminal Science paper was an early demonstration of the power of such an approach (Linder and Gartler, 1965). Here the authors extended their previous use of allelic differences in the glucose-6-phosphate dehydrogenase protein (G6PD) to study the lineage of Leiomyomas. G6PD is encoded on the X chromosome, and in heterozygous women only one copy will be expressed after X-inactivation. In sampled patches of Leiomyoma tumor they consistently observed a single active allele of G6PD, in contrast to the heterogenous patterns they saw in healthy uterine tissue. The implications were striking: all tumor cells descended from a single progenitor cell. Moreover, this elegant experiment demonstrated the power of genetic marks as carriers of lineage information.

Again technological constraints limited progress in the field. Large advances in retrospective lineage tracing did not come about until the emergence of sequencing technologies that could sample genetic variation at the nucleotide level. One of the first to realize this potential was Salipante and Horwitz (Salipante and Horwitz, 2006). The authors sampled from polyguanine DNA repeats, which are both highly prevalent in the genome and have an estimated mutational rate as high as 10^4 per cell per generation (Salipante et al., 2010). By sampling from diverse anatomical regions of an adult mouse, they were able to use these

polyguanine repeats to differentiate between two hepatocyte lineages, one of which shared ancestry with the remaining organs. Much like recombination approaches, these microsatellites are free to both gain and lose repeats over time, which could confound reconstruction. The sensitivity of the ABI sequencing platform also limited their resolution, but this study, and others like it (Frumkin et al., 2005) demonstrated the enormous potential of retrospective lineage tracing. These microsatellite approaches were later utilized to characterize crypt dynamics in the mouse (Reizel et al., 2011). Many groups have since expanded on these ideas, using both diverse mutational patterns and nascent sequencing technologies to broaden the applications of retrospective lineage tracing. For instance, Carlson et al. used random single primer PCR to reconstruct synthetic lineages of mouse fibroblasts (Carlson et al., 2011). In this work they demonstrate that single nucleotide variants (SNVs) could be used to reconstruct fate, and that these marks were rich enough to reconstruct the lineage of their synthetic cell-culture system.

Recent efforts have expanded retrospective lineage tracing into human somatic tissue, inferring lineage in both terminally differentiated cells as well as early embryonic stages. Lodato et al. used whole-genome sequencing from 36 individual neurons from post-mortem brain tissue, discovering over 1500 SNVs capable of delineating lineage, albeit with a high false positive rate (Lodato et al., 2015). They then used these lineage marks to perform targeting sequencing of another 136 neurons, discovering at least five neuronal clades that predate gastrulation. Related approaches have discovered the similar early lineage bifurcations in human blood (Ju et al., 2017). These somatic lineage tracing efforts have borrowed heavily from the techniques and lessons-learned in the study of tumor heterogeneity: many of the low frequency variant calling tools and modeling approaches were first pioneered in the study of blood cancers and tumor modeling (Carter et al., 2012; Ding et al., 2012; Cibulskis et al., 2013).

The power of retrospective lineage tracing comes from the enormous information capacity of the genome. Each human cell contains three billion bases: this has the capacity to encode a practically infinite amount of information. Adding on non-canonical bases, mi-

crosatellite expansions, base modifications, and other modifying marks further enriches this already expansive informational landscape. Additionally many of these three billion bases are unconstrained by obvious functional designs, and are open to perturbation. There is also the potential to enrich endogenous mutational rates with directed somatic mutators such as cytidine deaminases (Hess et al., 2016). The challenge of retrospective lineage tracing is then a more practical one: how do we recover all this scattered information for millions or even billions of cells?

As single-cell sequencing technologies advance and the per-cell costs decline, it will become feasible to sequence the cellular contents of whole organs, or even whole organisms. But for each additional cell sequenced, there will be a tremendous overlap, as new cells recover increasingly redundant information. Additionally, recovering this scattered lineage signal from such a sparse DNA landscape poses challenges for spatially resolved technologies such as *in situ* sequencing. A grander vision is to incorporate many of the prospective and retrospective lineage technologies mentioned above into an integrated approach, capturing a holistic view of development. Such efforts will maximize the information recovered from each experiment, a critical consideration given the stochastic nature of both development and lineage tracing technologies.

1.6 The organization of this dissertation

In this dissertation, I aim to improve the resolution, accuracy, and capabilities of prospective lineage tracing techniques using the CRISPR/Cas9 genome editing system. I explain our approach for lineage tracing, GESTALT, and demonstrate its synthetic and biological applications. Further, I look to future improvements of our technology, as well as its application to the broader field.

In my first chapter, I develop tooling to aid in the adoption of genome editing technologies, specifically software for CRISPR target selection, processing, and visualization. Here these improvements are harnessed for use in a scan of regulatory DNA regions in the HPRT1 gene. The underlying tool for finding and characterizing CRISPR sites, FlashFry, has also been

released to the community. This work forms much of the foundation for the GESTALT technique in the later chapter.

In the second chapter of the dissertation, we generate an initial proof of concept for our lineage tracing approach, GESTALT, leveraging our computational tooling in conjunction with published experimental work. We demonstrate that a tiled series of targets in a compact locus can be edited *in vivo*, and that such edits accumulate in a stochastic and progressive fashion. We then generate a synthetic lineage system in cell culture to demonstrate GESTALT's lineage recording capacity and to develop tools for lineage reconstruction. We then integrate our GESTALT system into the model organism *Danio rerio* (zebrafish). By sampling diverse tissues from an adult zebrafish, we show that our recorded lineage marks are consistent with existing knowledge of germ layer division and organ specification. Finally, we find that barcode diversity is bottlenecked in the adult tissues in comparison to the sampled embryos, and that many organ systems are composed of very few progenitors.

In the final chapter, I present my vision for the future of whole organism lineage tracing, and discuss both future opportunities and challenges. I close with thoughts about the integration of diverse technologies to comprehensively characterize both normal development and disease states at the resolution of single cells.

Chapter 2

FLASHFRY: A TOOL FOR RAPID, GENOME-WIDE CHARACTERIZATION OF CRISPR TARGET SITES

Figure 2.6 and its accompanying text are adapted from the follow paper: Molly Gasperini*, Gregory M. Findlay*, Aaron McKenna, Jennifer H. Milbank, Choli Lee, Melissa D. Zhang, Darren A. Cusanovich, and Jay Shendure. Multiplex deletion scanning of 206 kilobases encompassing HPRT1 for functionally critical distal regulatory elements. *BiorXiv*, 8 December 2016.

2.1 Summary

FlashFry is a fast and flexible command-line tool for characterizing large numbers of CRISPR target sequences. Many similar tools exist, but there is a clear need for a simple, lightweight framework that can quickly discover and score thousands of candidate guides from any arbitrary DNA sequence. With FlashFry, users can specify an unconstrained number of mismatches to putative off-targets, richly annotate discovered sites, and tag guides with common on and off-target scoring metrics. FlashFry also runs at speeds comparable to genome-wide aligners for large target sets. Output is provided as an easy-to-manipulate text file.

2.2 Availability and implementation

FlashFry is written in Scala and bundled as a stand-alone Jar file, easily run on any system with an installed Java virtual machine (JVM). The tool is freely licensed under version 3 of the GPL, and code and documentation are available on GitHub:
http://aaronmck.github.io/FlashFry/

2.3 Introduction

The CRISPR prokaryotic immune system has transformed genome engineering. In the simplified system, CRISPR proteins are directed to create double-stranded DNA breaks at genome positions matching a specified guide sequence (Wright et al., 2016). These double-stranded breaks are commonly repaired by a non-homologous end joining (NHEJ) pathway, leaving small insertions or deletions (indels) at the genomic target site. These resulting indels can be used to perturb endogenous gene function(Wang et al. 2017), encode information (McKenna et al., 2016), or characterize the function of a genomic sequence (Gasperini et al., 2016b; Liu et al., 2017).

Although CRISPR editing is highly-specific (Doench et al., 2014), not all guides function with the same efficiency or specificity. For instance, double-stranded breaks can occur at genomic locations ('targets') that are an imperfect match to the supplied guide sequence. To reduce the chance of such unintended genome editing, guide sequences can be chosen that contain less overlap with all possible alternate targets in the genome. The importance of specific differences in the guide sequence, the location of the target in the genome of interest, its chromatin landscape, and the method of guide delivery, all seem to have an effect on the prevalence of such off-targets (Haeussler et al., 2016).

To help users choose both specific and active guide sequences, the community has created a large number of CRISPR design tools, most of which are available as web applications (Labun et al., 2016; Haeussler et al., 2016). These web tools are convenient for those screening a small set of guides, or scanning a single genomic locus, such as an exon. Unfortunately, these tools require batched queries for large guide sets, which makes it more challenging to analyze and process whole-gene or whole-genome scans. Additionally, some guide screening tools rely upon genome-wide alignment tools to generate putative off-target lists for each guide. These aligners are generally designed for practical reasons to quickly discover only the most similar sequences with a limited number of mismatches in comparison to the guide (typically k 3 or less for the length of a guide). Importantly, some have been

shown to miss a subset of potential targets (Doench et al., 2016). Experimental efforts have also shown that some very active off-target sequences contain upwards of 6 mismatches when compared to the guide (Tsai et al., 2014). To address these issues, and to meet our needs for high-throughput guide selection from arbitrary genomic regions, we've created FlashFry, a command-line tool for discovery and characterization of CRISPR guide sequences.

2.4 Implementation

Given a reference genome of interest, FlashFry generates a block-compressed binary database of all potential target sequences (fig. 2.1). This database is CRISPR enzyme specific, and can be generated in a few hours on a standard computer (table 2.1). In this database, genome targets and their positions are encoded into a hierarchy of sorted prefix-bins. Given the inherent inefficiencies of high-mismatch searches, FlashFry uses a filtering approach to find candidate off-targets, precomputing a traversal over target bins with less than k mismatches to the each guide in the candidate set (such filtering approaches are well reviewed in Navarro et al. (Navarro, 2001)). FlashFry will switch over to a linear traversal of the database when a high threshold of bins will be visited, common with large guide screens or with a high k mismatch threshold. To further reduce search times, target sequences and their occurrences are stored as a bit-encoded value, allowing bit-parallelism comparisons when determining mismatches (fig. 2.1). FlashFry is currently compatible with target sequences of up to 24 bases in length, although it could be expanded to longer target sequences as the bins encode their prefix sequence.

Search times against the FlashFry database are proportional to the number of guides and the allowed number of mismatches (fig. 2.2), and compares favorably to similar CRISPR command line tools (Bae et al., 2014), or alternatively adapting FM-index based tools (fig. 2.3). Using simple bitwise operations, we can compute millions of Hamming guide-to-target distance comparisons per second (fig. 2.4). To further speed-up search, off-target discovery is halted for candidate guides that have exceeded a user defined number of off-target hits, saving compute time by eliminating poor candidates early from the putative guide pool.

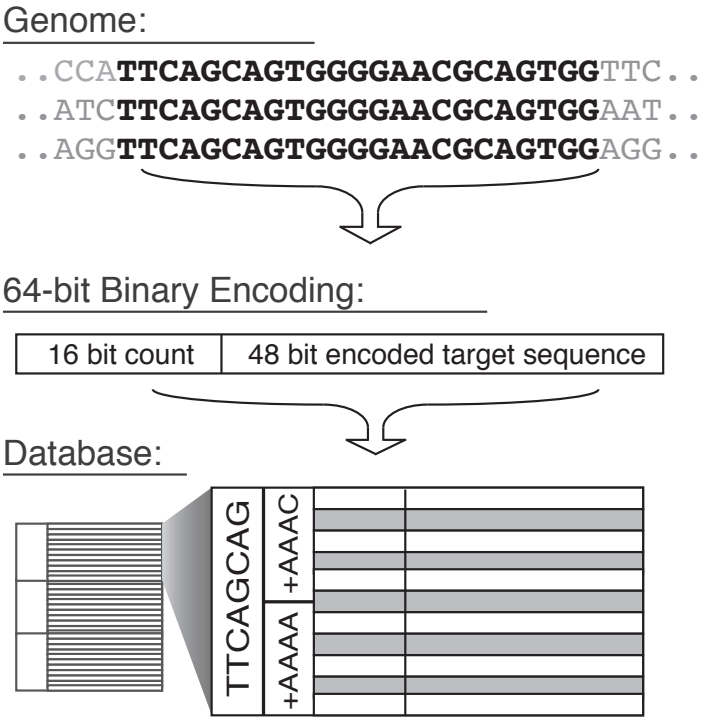


Figure 2.1: Schematic of target search and database creation. The genome of interest is scanned for target sequences of the specified length that are terminated with the CRISPR-specific PAM sequence. Identical target sequences are aggregated and encoded into 64-bit binary values, recording their sequence and number of occurrences in the genome. This compressed database uses a hierarchical filtering strategy to search for target sequences, and performs direct comparisons using bit-parallelism mismatch calculations.

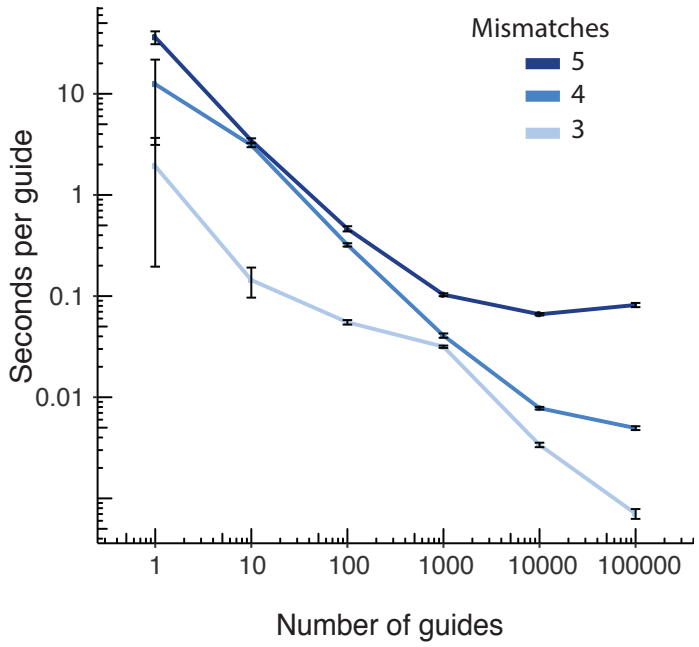


Figure 2.2: Average per-candidate runtime for increasing numbers of guides. Average runtime of twenty five replicates per combination of randomly chosen candidate guides with an increasing number of allowed mismatches. Plotted are the median runtime with median absolute deviation (MAD) bars for each set.

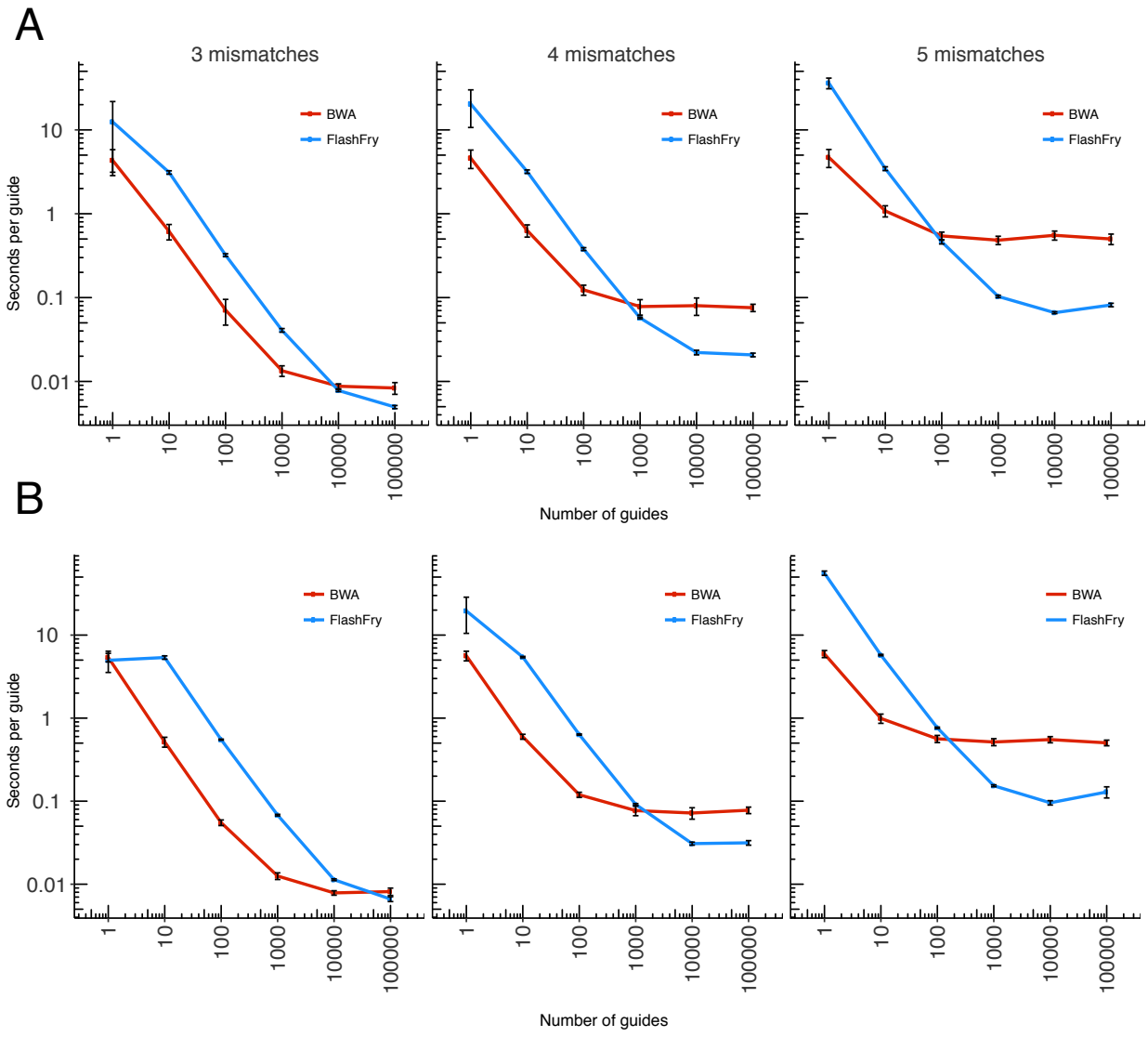
2.5 Guide characterization and scoring

The goal for most users is to pick some subset of highly active and specific guide sequences from a full list of candidate targets within a region of interest. In this light, we've implemented many commonly used scoring approaches for both on-target efficiency as well as off-target performance, including Cutting-frequency determination (CFD) (Doench et al., 2016), the Hsu et al. off-target scoring scheme (Hsu et al., 2013), the Moreno-Mateos and Vejnar et al. 2015, and the Doench et al. 2014 on-target metrics (Moreno-Mateos et al.; Doench et al., 2014). We have also included a set of basic design criteria filters, including high and low GC content, warnings for poly T tracts (which halt polIII transcription), and targets that have reciprocal off-targets within the region of interest (potentially leading to deletions of the intervening sequence). Lastly, regions can be annotated with information from external BED files, which may be useful for highlighting repetitive sequences or putative regulatory regions.

Here we demonstrate the utility of FlashFry for creating CRISPR libraries using two

---

Figure 2.3 *(following page)*: Runtime comparison of FlashFry and BWA. Comparison of runtimes for FlashFry and BWA version 0.7.13-r1126 (Li and Durbin, 2010) over an increasing number of guides and permitted mismatches. Twenty five random CRISPR guide sets were run for each guide-count (x-axis) and permitted mismatch level (2,777,775 potential guides per mismatch level). BWA runtime includes the initial alignment step (aln) and mapping to genomic coordinates (samse). Plotted are the median runtime with median absolute deviation (MAD) bars for each set of 25 runs. (A) Using the NGG motif for off-target selection, (B) using the NRG motif for off-target selection. FlashFry benefits from aggregating all guide-to-genome comparisons in one pass of the database, surpassing BWA's performance at hundreds of guides for five mismatches, and thousands of guides at 4 mismatches.


A
3 mismatches
4 mismatches
5 mismatches
BWA
FlashFry
Seconds per guide
10
1
0.1
0.01
1
10
100
1000
10000
100000
Number of guides
B
BWA
FlashFry
Seconds per guide
10
1
0.1
0.01
1
10
100
1000
10000
100000
Number of guides

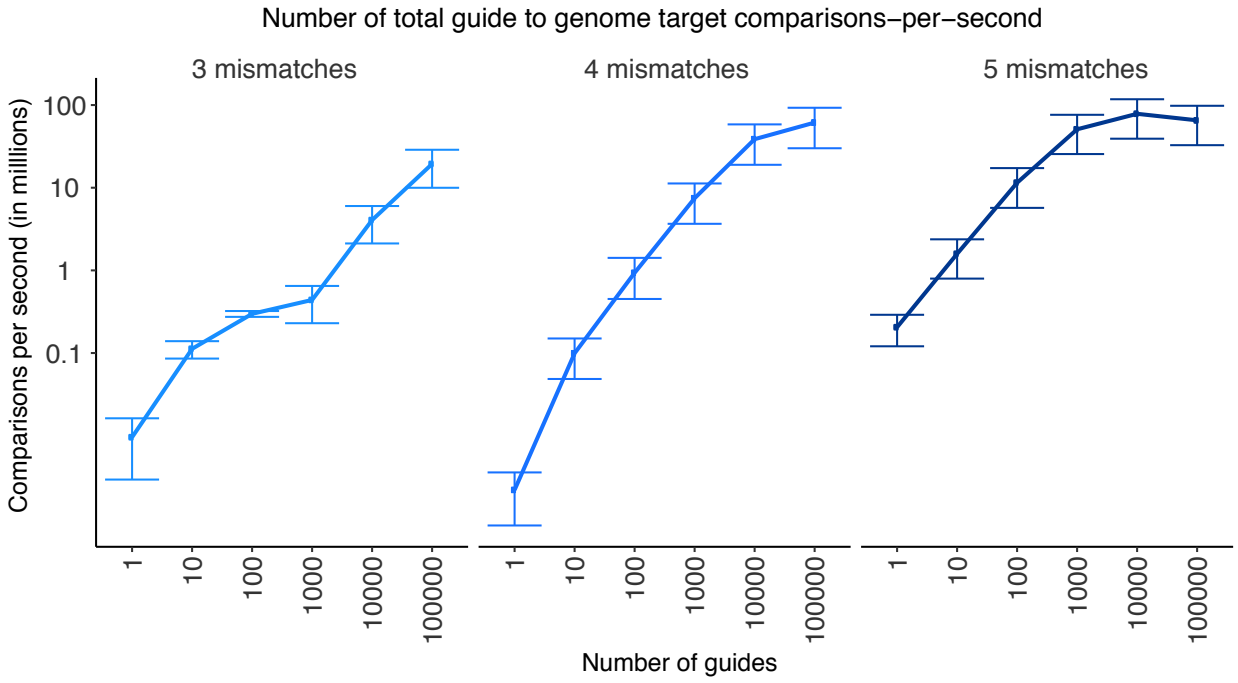


Figure 2.4: FlashFry off-target comparison speed. The number of guide to target comparisons per second, for 1, 3, and 5 allowed mismatches over an increasing number of candidate guides. Smaller numbers of guides and lower mismatch counts have lower rates as the initialization and output times are amortized over the whole run.

Genome / version	Cas9 (NGG)	Cas9 (NGG/NAG)	CPF1 (TTTN)
Caenorhabditis elegans - 235	0:3:21	0:6:03	0:5:35
Human - hg38	3:19:29	5:24:55	2:50:59
Mouse - mm10	2:36:53	4:36:03	2:11:35
Drosophila melanogaster- BDGP6	0:6:33	0:10:48	0:5:44

Table 2.1: Off-target database generation times. A sample of computational times (in hours) required to build a FlashFry database for versions of the *Caenorhabditis elegans*, human, mouse, and *Drosophila melanogaster* genomes for common CRISPR enzymes. This analysis was run on a disk-based network area storage (NAS) system.

genomic regions. First we scored 254,848 candidate SpCas9 target sequences (allowing both NGG and NAG PAMs) within 1 megabase of the human MYC gene. (2.5). The results were scored with two on-target and two off-target metrics from the literature, and intersected with a list of known repetitive elements. (Doench et al., 2016; Hsu et al., 2013; Moreno-Mateos et al.; Doench et al., 2014) The aggregate table could then be used to design a perturbation screen; for instance 8,038 candidate guide sequences had no other exact occurrence within the human genome, had a Hsu et al. score above 70, and did not intersect a repetitive element.

We then used FlashFry to design targets for a deletion scan of the hypoxanthine(-guanine) phosphoribosyltransferase (*HPRT1*) locus used in Gasperini et al. (fig. 2.6) (Gasperini et al., 2016a). In this *HPRT1* screen, guides were selected that occurred only once in the region on chrX from position 133,507,694 to 133,713,798. We then excluded protospacers that had a perfect sequence match elsewhere in the genome, and scored the remaining gRNAs for both on-target and off-target activity. We considered off-target sequences that had five or fewer mismatches to the putative gRNA, and calculated an aggregate off-target score using the method of Hsu et al. (Hsu et al., 2013). Final deletion pairs were matched using spacers

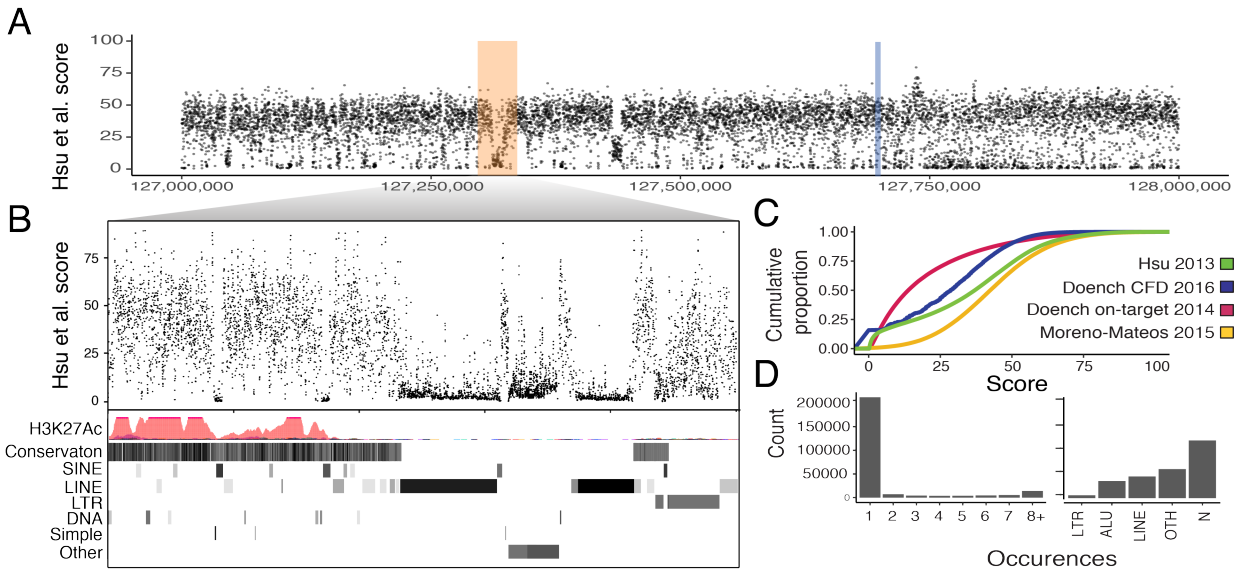


Figure 2.5: Characterizing CRISPR target sites in the *MYC* enhancer region. FlashFry was run for a 1Mb region flanking the human *MYC gene*, generating 254,848 candidate sites, which were scored for both on and off-target activity. **(A)** Average off-target specificity scores (Hsu et al., 2013) in 100 basepair windows. **(B)** Enlargement of the 25Kb region (chr8:127,300,000-127,325,000) highlighted in **A**. The observed off-target specificity tracks well with known repetitive elements. **(C)** Cumulative density function plot of on and off-target scoring metrics for all targets. **(D)** For each target, the number of times it is seen within the genome (left) and the overlap with repetitive elements (right). LTR = 'Long terminal repeat', OTH = 'Other repetitive element type', N = 'None'.

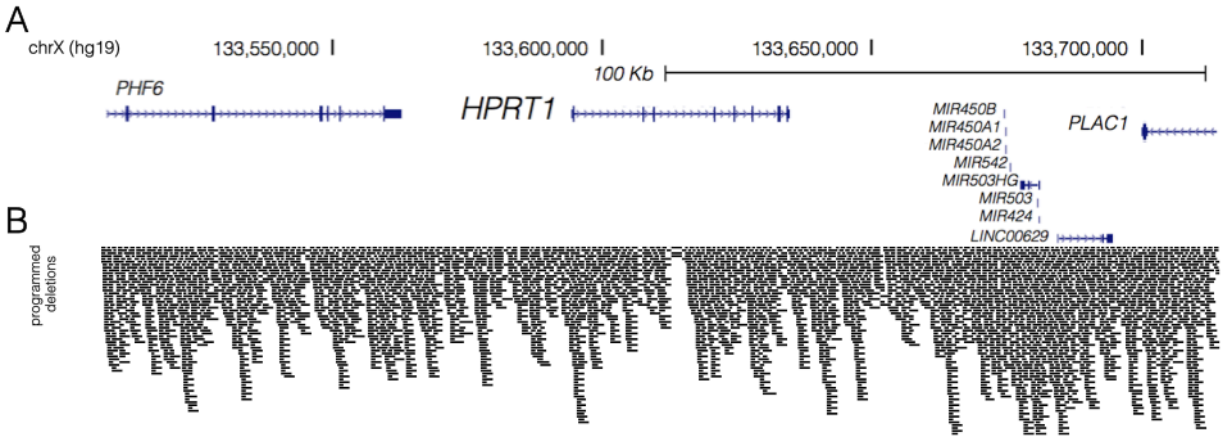


Figure 2.6: Design of an *HPRT1* deletion screen. **(A)**. Deletions were programmed across 206.1 Kb of the HPRT1 locus and its surrounding sequence (chrX:133,507,694-133,713,798, hg19, UCSC Genes track in blue). **(B)**. FlashFry was used to characterize the target endpoints for 4,342 overlapping 1 to 2 kilobase (Kb) deletions that tiled a 206 Kb region centered on *HPRT1*, tiling across the locus such that each base-pair was interrogated by a median of 27 independently programmed deletions.

that did not contain BsmBI restriction sites, were not predicted to have off-target hits in other 6TG resistance genes or in KBM7 essential genes (the HAP1 parental cell line), were greater than 25 bp apart, further than 50 bp from an exon, and passed on-target (above 10) and off-target (above 25) thresholds. Contrastingly, the individual gRNA library included all of the spacers targeting the same region, excluding those predicted to have 2,000 or more off-targets or to have off-targets with 4 or fewer mismatches within the targeted HPRT1 region.

2.6 Discussion

The needs of most biologists are well-met by many of the abundant CRISPR web applications. However, a more efficient and flexible toolset is required for genome-wide knockout studies,

noncoding deletion scans, and other large-scale studies or method development projects. Here we demonstrate that FlashFry fills this void, and can be used to rapidly discover and characterize tens to hundreds-of-thousands of guides from an arbitrary sequence quickly and with a low memory footprint. For methods developers, we also provide a simple interface for implementing additional scoring schemes, given the sequence context of a guide and it's off-target hits. FlashFry has no system dependencies outside of the JVM, and avoids many of the pitfalls and complexity of associated tools that rely on genome aligners to discover off-target sequences.

2.7 Acknowledgements and funding

We thank the authors of the CRISPOR paper for their excellent work aggregating and documenting many of the scoring methods, an invaluable resource for the community. For discussion and advice, we thank all the Shendure lab members, especially Molly Gasperini and Vikram Agarwal, and we thank Andrew Hill, Greg Findlay, and Vijay Ramani who motivated the construction of FlashFry. AM was supported by a fellowship from the NIH/NHLBI (T32HL007312).

Data set	Amazon EC2 (G2) instance	Local distributed cluster nodes
10000 guide (set A)	351.12	776.22
10000 guide (set B)	348.53	825.57
10000 guide (set C)	347.86	1397.88
10000 guide (set D)	344.26	783.44
10000 guide (set E)	347.3	776.44
mean	347.81	911.91
Standard deviation	2.46	272.47

Table 2.2: Effects of SSD storage on runtime. Off-target discovery times (seconds) for 10,000 random guide sequences on an G2 Amazon EC2 node with an SDD drive for off-target database storage, compared to a distributed node on a local cluster with a disk-based network area storage (NAS). Both jobs were run with identical parameter sets and memory allocations.

Chapter 3

WHOLE-ORGANISM LINEAGE TRACING BY COMBINATORIAL AND CUMULATIVE GENOME EDITING

This chapter has been adapted with changes from: McKenna A, Findlay GM, Gagnon JA, Horwitz MS, Schier AF, and Shendure J. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science* 29 July 2016

I contributed to all aspects of the paper except the zebrafish husbandry and breeding, and I did not contribute to figures 3.8 and 3.10.

3.1 Abstract

Multicellular systems develop from single cells through a lineage, but current lineage tracing approaches scale poorly to whole organisms. Here we use genome editing to progressively introduce and accumulate diverse mutations in a DNA barcode over multiple rounds of cell division. The barcode, an array of CRISPR/Cas9 target sites, records lineage relationships in the patterns of mutations shared between cells. In cell culture and zebrafish, we show that rates and patterns of editing are tunable, and that thousands of lineage-informative barcode alleles can be generated. We find that most cells in adult zebrafish organs derive from relatively few embryonic progenitors. Genome editing of synthetic target arrays for lineage tracing (GESTALT) will help generate large-scale maps of cell lineage in multicellular systems.

3.2 Introduction

The tracing of cell lineages was pioneered in nematodes by Charles Whitman in the 1870s, at a time of controversy surrounding Ernst Haeckel's theory of recapitulation (Stent, 1998). This line of work culminated a century later in the complete description of mitotic divisions in the roundworm *C. elegans* - a tour de force facilitated by its visual transparency as well as the modest size and invariant nature of its cell lineage (Sulston et al., 1983).

Over the past century, a variety of creative methods have been developed for tracing cell lineage in developmentally complex organisms (Kretzschmar and Watt, 2012). In general, subsets of cells are marked and their descendants followed as development progresses. The ways in which cell marking has been achieved include dyes and enzymes (Kimmel and Law, 1985; Keller, 1976; Weisblat et al., 1978), cross-species transplantation (Le Douarin and Teillet, 1974), recombinase-mediated activation of reporter gene expression (Dymecki and Tomasiewicz, 1998; Zinyk et al., 1998), insertion of foreign DNA (Walsh and Cepko, 1992; Lu et al., 2011; Porter et al., 2014), and naturally occurring somatic mutations (Salipante and Horwitz, 2006; Behjati et al., 2014; Lodato et al., 2015). However, despite many powerful applications, these methods have limitations for the large-scale reconstruction of cell lineages in multicellular systems. For example, dye and reporter gene-based cell marking are uninformative with respect to the lineage relationships between descendent cells. Furthermore, when two or more cells are independently but equivalently marked, the resulting multitude of clades cannot be readily distinguished from one another. Although these limitations can be overcome in part with combinatorial labeling systems (Livet et al., 2007; G. M. Church, N.d. Print.) or through the introduction of diverse DNA barcodes (Walsh and Cepko, 1992; Lu et al., 2011; Porter et al., 2014), these strategies fall short of a system for inferring lineage relationships throughout an organism and across developmental time. In contrast, methods based on somatic mutations have this potential, as they can identify lineages and sub-lineages within single organisms (Salipante and Horwitz, 2006; Carlson et al., 2011). However, somatic mutations are distributed throughout the genome, necessitating

whole genome sequencing (Behjati et al., 2014; Lodato et al., 2015), which is expensive to scale beyond small numbers of cells and not readily compatible with *in situ* readouts (Lee et al., 2014; Ke et al., 2013).

What are the requirements for a system for comprehensively tracing cell lineages in a complex multicellular system? First, it must uniquely and incrementally mark cells and their descendants over many divisions and in a way that does not interfere with normal development. Second, these unique marks must accumulate irreversibly over time, allowing the reconstruction of lineage trees. Finally, the full set of marks must be easily read out in each of many single cells.

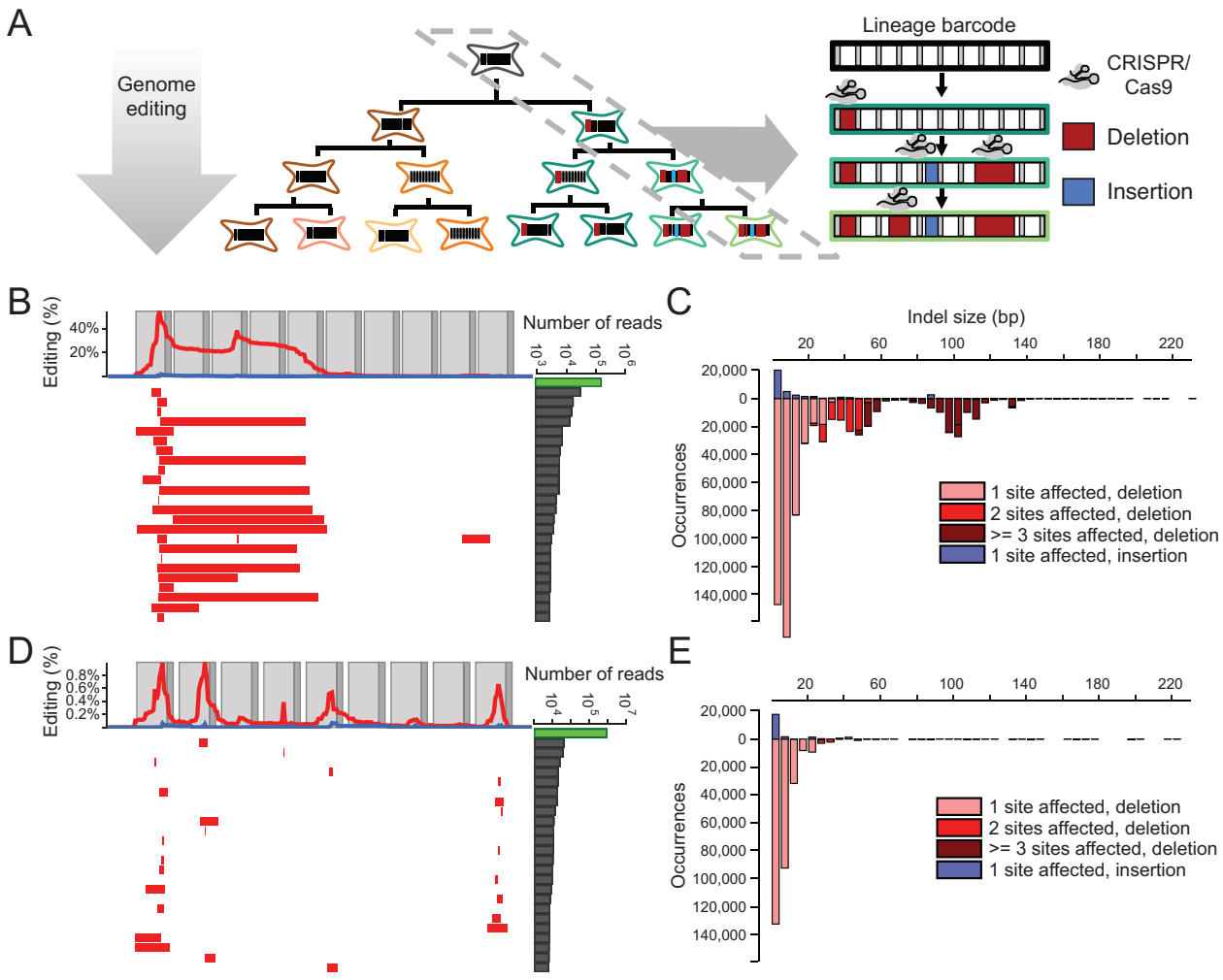
We hypothesized that genome editing, which introduces diverse, irreversible edits in a highly programmable fashion (Doudna and Charpentier, 2014), could be repurposed for cell lineage tracing in a way that realizes these requirements. To this end, we developed genome editing of synthetic target arrays for lineage tracing (GESTALT), a method that uses CRISPR/Cas9 genome editing to accumulate combinatorial sequence diversity to a compact, multi-target, densely informative barcode. Importantly, edited barcodes can be efficiently queried by a single sequencing read from each of many single cells (Fig. 3.1A). In both cell culture and in the zebrafish Danio rerio, we demonstrate the generation of thousands of uniquely edited barcodes that can be related to one another to reconstruct cell lineage relationships. In adult zebrafish, we observe that the majority of cells of each organ are derived from a small number of progenitor cells. Furthermore, ancestral progenitors, inferred on the basis of shared edits amongst subsets of derived alleles, make highly non-uniform contributions to germ layers and organ systems.

3.3 Results

3.3.1 Combinatorial and cumulative editing of a compact genomic barcode in cultured cells

To test whether genome editing can be used to generate a combinatorial diversity of mutations within a compact region, we synthesized a contiguous array of ten CRISPR/Cas9

Figure 3.1 *(following page)*: Genome editing of synthetic target arrays for lineage tracing (GESTALT).(A) An unmodified array of CRISPR/Cas9 target sites (i.e., a barcode) is engineered into a genome (gray cell). Editing reagents are introduced during expansion of cell culture or *in vivo* development of an organism, resulting in a unique pattern of insertions and deletions (right), and are stably accumulated in specific lineages (green cell lineage). The lineage relationships of alleles that differ in sequence can often be inferred on the basis of these accumulated edits. (B) The 25 most frequent alleles from the edited v1 barcode are shown. Each row corresponds to a unique sequence, with red bars indicating deleted regions and blue bars indicating insertion positions. Blue bars begin at the insertion site, with their width proportional to the size of the insertion, which will rarely obscure immediately adjacent deletions. The number of reads observed for each allele is plotted at the right (log10 scale; the green bar corresponds to the unedited allele). The frequency at which each base is deleted (red) or flanks an insertion (blue) is plotted at the top. Light gray boxes indicate the location of CRISPR protospacers while dark gray boxes indicate PAM sites. For the v1 array, inter-target deletions involving sites 1, 3 and 5, or focal (single target) edits of sites 1 and 3 were observed predominantly. (C) A histogram of the size distribution of insertion (top) and deletion (bottom) edits to the v1 array is shown. The colors indicate the number of target sites impacted. Although most edits are short and impact a single target, a substantial proportion of edits are inter-target deletions. (D) We tested three array designs in addition to v1, each comprising nine to ten weaker off-target sites for the same sgRNA (v2-v4) (22). Editing of the v2 array is shown with layout as described in panel (B). Editing of the v3 and v4 array are shown in Fig. 3.4, A and B. The weaker sites within these alternative designs exhibit lower rates of editing than the v1 array, but also a much lower proportion of inter-target deletions. (E) A histogram of the size distribution of insertion (top) and deletion (bottom) edits to the v2 array is shown. In contrast with the v1 array, almost all edits impact only a single target.


A
Genome editing
Lineage barcode
CRISPR/ Cas9
Deletion
Insertion
B
Editing (%)
40%
20%
Number of reads
C
Indel size (bp)
Occurrences
1 site affected, deletion
2 sites affected, deletion
>= 3 sites affected, deletion
1 site affected, insertion
D
Editing (%)
0.8%
0.6%
0.4%
0.2%
Number of reads
E
Occurrences
1 site affected, deletion
2 sites affected, deletion
>= 3 sites affected, deletion
1 site affected, insertion

targets (protospacers plus PAM sequences) separated by 3 base-pair (bp) linkers (total length of 257 bp). The first target perfectly matched one single guide RNA (sgRNA), while the remainder were off-target sites for the same sgRNA, ordered from highest to lowest activity (Tsai et al., 2014). This array of targets ('v1 barcode') was cloned downstream of an EGFP reporter in a lentiviral construct (Sancak et al., 2008). We then transduced HEK293T cells with lentivirus and used FACS to purify an EGFP-v1 positive population. To edit the barcode, we co-transfected these cells with a plasmid expressing Cas9 and the sgRNA and a vector expressing DsRed. Cells were sorted three days post-transfection for high DsRed expression, and genomic DNA (gDNA) was harvested on day 7. The v1 barcode was PCR amplified and the resulting amplicons subjected to deep sequencing.

To minimize confounding sequencing errors, which are primarily substitutions, we analyzed edited barcodes for only insertion-deletion changes relative to the 'wild-type' v1 barcode. In this first experiment, we observed 1,650 uniquely edited barcodes (each observed in $>$ 25 reads) with diverse edits concentrated at the expected Cas9 cleavage sites, predominantly inter-target deletions involving sites 1, 3 and 5, or focal edits of sites 1 and 3 (Fig. 3.1B and C, and FIX table S1). These results show that combinatorial editing of the barcode can give rise to a large number of unique sequences, i.e. 'alleles'.

To evaluate reproducibility, we transfected the same editing reagents to cultures expanded from three independent EGFP-v1 positive clones. Targeted RT-PCR and sequencing of EGFP-v1 RNA showed similar distributions of edits to the v1 barcode in the transcript pool, between replicates as well as in comparison to the previous experiment (Fig. 3.2). These results show that the observed editing patterns are largely independent of the site of integration and that edited barcodes can be queried from either RNA or DNA.

To evaluate how editing outcomes vary as a function of Cas9 expression, we co-transfected EGFP-v1 positive cells with a plasmid expressing Cas9 and the sgRNA as well as an DsRed vector, and after four days sorted cells into low, medium, and high DsRed bins and harvested gDNA. Overall editing rates matched DsRed expression (frequency of non-wild-type barcodes: low DsRed = 40%; medium DsRed = 69%; high DsRed = 91%). The profile

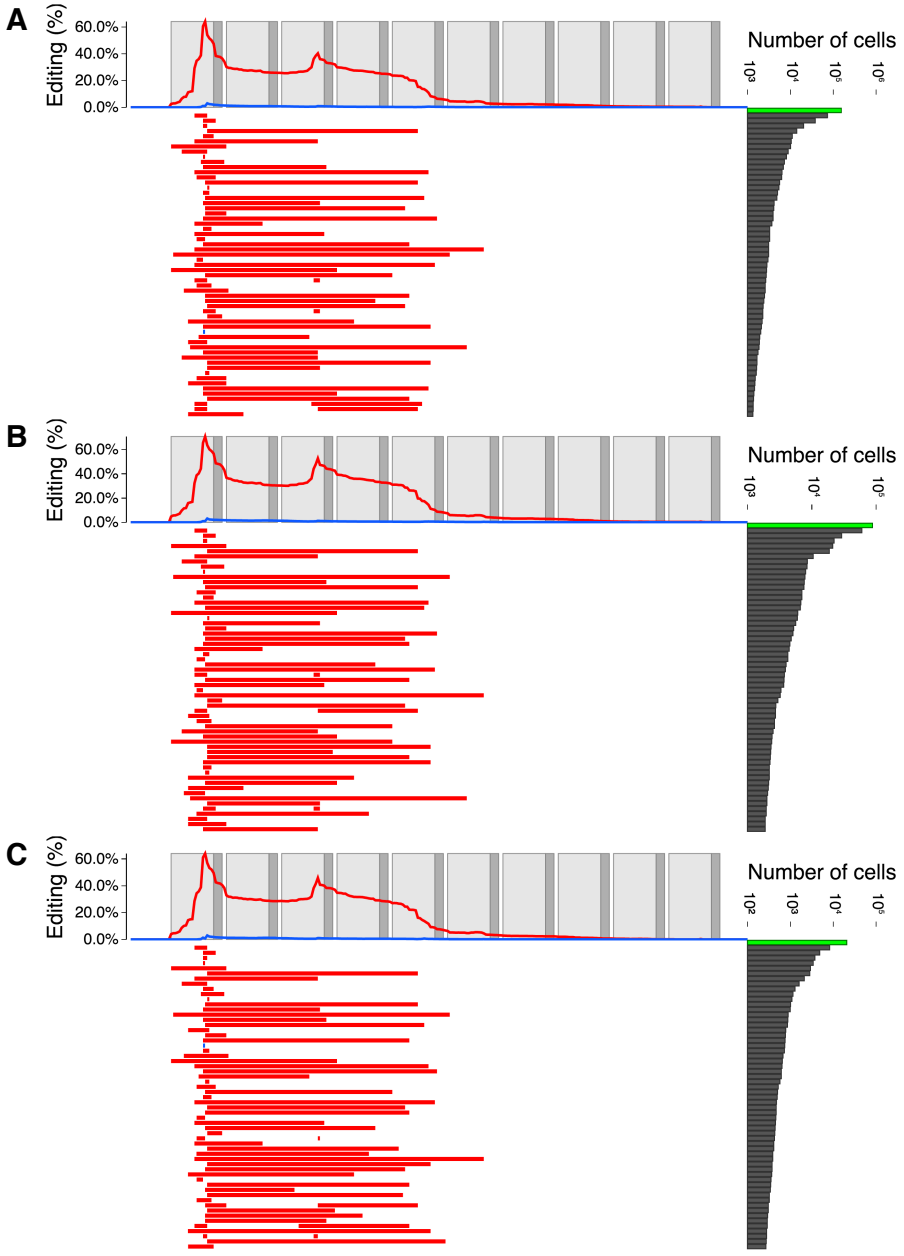


Figure 3.2: RNA-based readout of v1 barcode editing. (A-C) The 50 most frequent alleles of the v1 barcode are depicted, based on reverse transcription, amplification and sequencing from mRNA. Three biological replicates are shown for comparison, wherein each experiment was performed on a culture expanded from an independent v1+ clone. Layout is as described in the Fig. 3.1B legend.

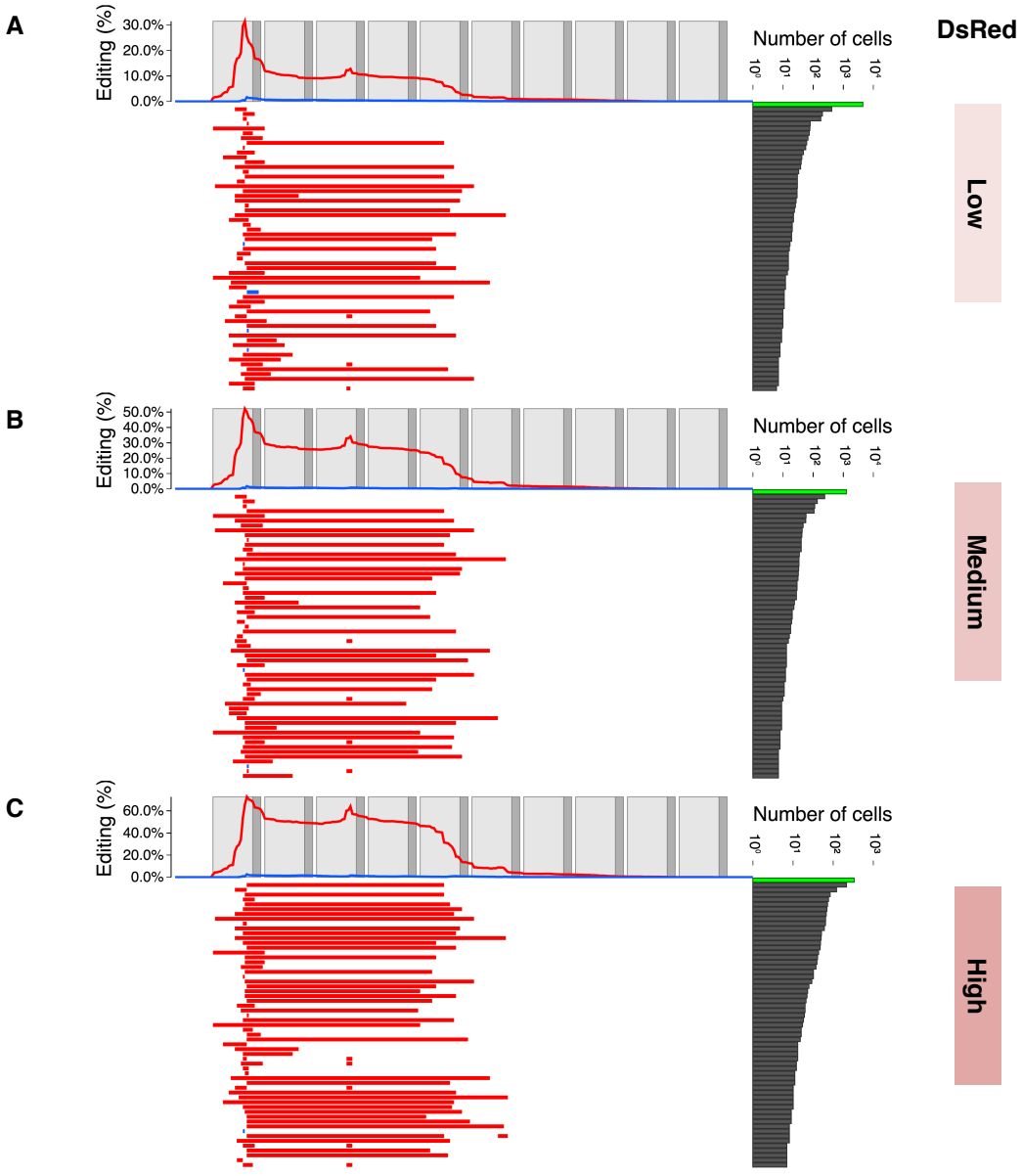
of edits observed remained similar, but there were fewer inter-target deletions in the lower DsRed bins (Fig. 3.3). These results show that adjusting expression levels of editing reagents can be used to modify the rates and patterns of barcode editing.

We also synthesized and tested three barcodes (v2-v4) with nine or ten weaker off-target sites for the same sgRNA as used for v1 (Tsai et al., 2014). Genome editing resulted in derivative barcodes with substantially fewer edits than seen with the v1 barcode, but a much greater proportion of these edits were to a single target site, i.e. fewer inter-target deletions were observed (Fig. 3.1, D and E, and Fig. 3.4, A and B). As only a few targets were substantially edited in designs v1-v4, we combined the most highly active targets to a new, twelve target barcode (v5). This barcode exhibited more uniform usage of constituent targets, but with relative activities still ranging over two orders of magnitude (Fig. 3.4 and table S1). These results illustrate the potential value of iterative barcode design.

To determine whether the means of editing reagent delivery influences patterns of barcode editing, we introduced a lentiviral vector expressing Cas9 and the same sgRNA to cells containing the v5 barcode (Sanjana et al., 2014). After two weeks of culturing a population bottlenecked to 200 cells by FACS, we observed diverse barcode alleles but with substantially

---

Figure 3.3 *(following page)*: Editing rates of the v1 barcode correlate with transfection efficiency. To evaluate how editing outcomes vary as a function of Cas9 expression, v1+ cells were co-transfected with pX330-v1 and a DsRed vector four days prior to sorting into bins based on DsRed expression and harvesting gDNA. The observed patterns of editing are shown for low (A), medium (B) and high (C) DsRed expression bins. Layout is as described in the Fig. 3.1B legend, and top 60 alleles in each experiment are shown. Overall editing rates correlated with DsRed expression (low: 40%; medium: 69%; high: 91%), presumably reflecting transfection efficiency. The overall profile of edits observed remained approximately similar, although the proportion of inter-target deletions correlated with DsRed expression.

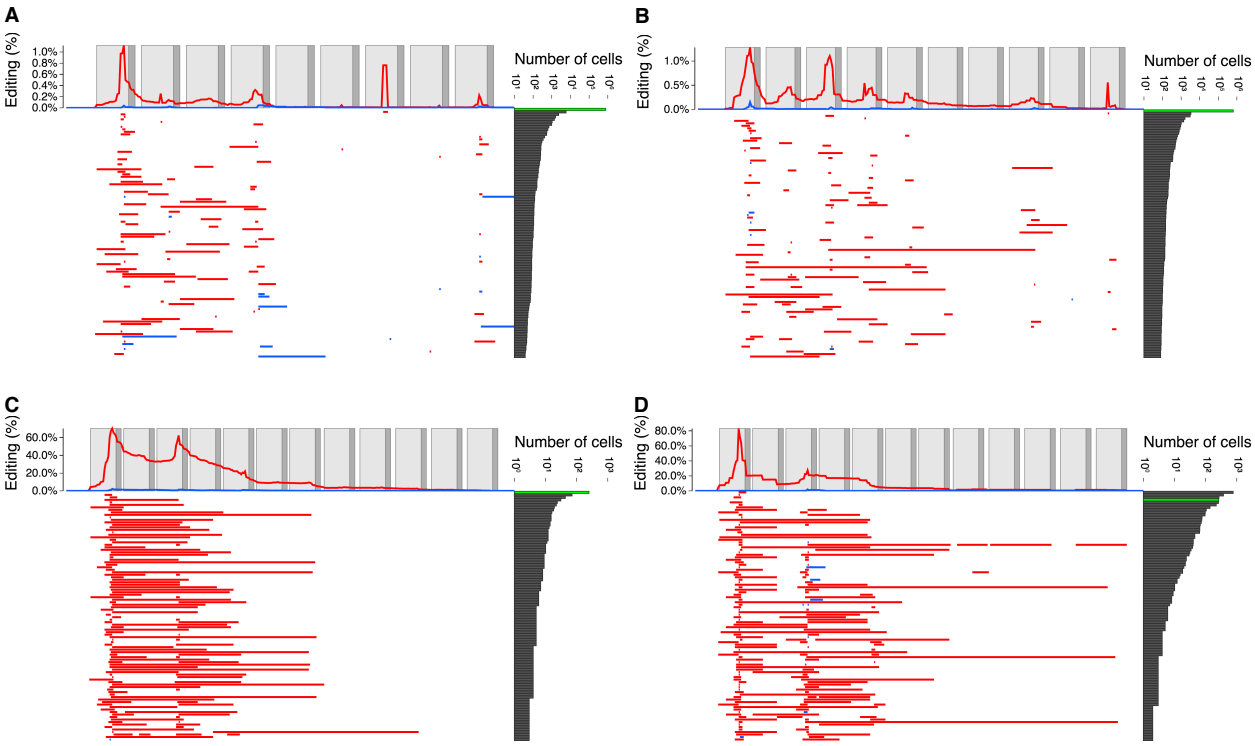

A
B
C
Editing (%)
30.0%
20.0%
10.0%
0.0%
50.0%
40.0%
60.0%
Number of cells
DsRed
Low
Medium
High

fewer inter-target deletions than with episomal delivery of editing reagents (Fig. 3.4D). This finding demonstrates that the allelic spectrum can also be modulated by the delivery mode of editing reagents.

Taken together, these results show that editing multiple target sites within a compact barcode can generate a combinatorial diversity of alleles, and also that these alleles can be read out by single sequencing reads derived from either DNA or RNA. Rates and patterns of barcode editing are tunable by using targets with different activities and/or off-target sequences, by iteratively recombining targets to new barcode designs, and by modulating

---

Figure 3.4 *(following page)*: Genome editing of alternative barcode designs. (A, B) In addition to v1, three additional barcode designs were tested, each with nine or ten weaker off-target sites for the same sgRNA (v2-v4). Editing observed in the v2 barcode is shown in Fig. 3.1D while editing observed in the v3 and v4 barcodes is shown above (panels (A) and (B), respectively; 100 alleles in each). Layout is as described in the Fig. 3.1B legend. The weaker off-targets within these alternative barcode designs exhibit lower rates of editing than the v1 barcode, but also a much lower proportion of inter-target deletions. (C) As only a subset of targets were substantially edited in designs v1-v4, the most highly active targets were combined to a new, twelve target design (v5), consisting, in order, of v1 targets 1-6; v3 target 1; v2 targets 1, 2 and 5; and v4 targets 1 and 3. Editing observed in the v5 barcode is shown in panel (C), with layout as described in the Fig. 3.1B legend. (D) To evaluate whether the means by which editing reagents are introduced impacts the rate and pattern of edits to the barcode, a lentiviral vector expressing Cas9 and the same sgRNA was introduced to cells prior to integration of the v5 barcode. After two weeks of culturing a population bottlenecked to 200 cells by FACS, diverse barcodes were observed but with substantially fewer inter-site deletions than with episomal delivery of editing reagents, i.e. (C): episomal expression of Cas9 and sgRNA vs. (D): lentiviral expression of Cas9 and sgRNA, both editing the v5 barcode.


A
Editing (%)
1.0%
0.8%
0.6%
0.4%
0.2%
0.0%
Number of cells
B
Editing (%)
1.0%
0.5%
0.0%
Number of cells
C
Editing (%)
60.0%
40.0%
20.0%
0.0%
Number of cells
D
Editing (%)
80.0%
60.0%
40.0%
20.0%
0.0%
Number of cells

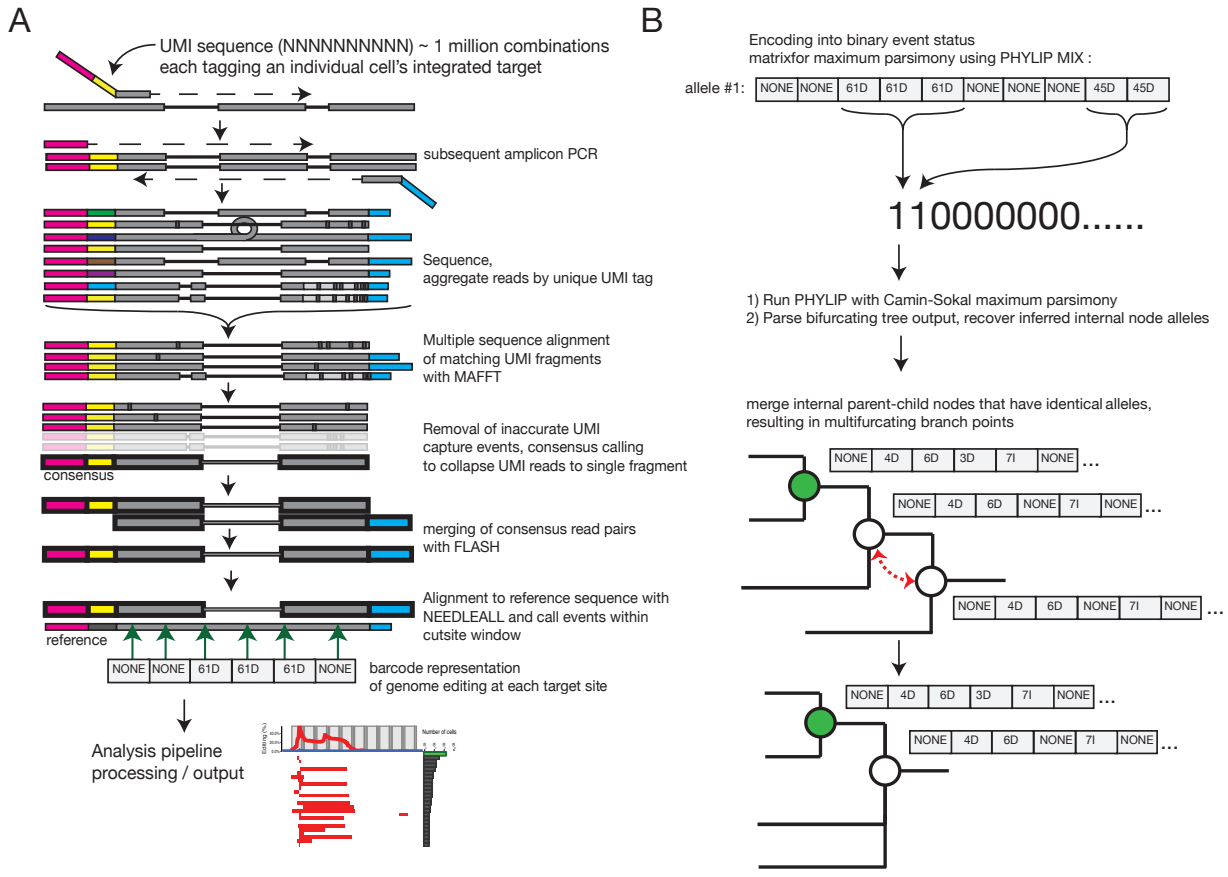
the concentration and means of delivery of editing reagents.

3.4 Reconstruction of lineage relationships in cultured cells

To determine whether GESTALT could be used to reconstruct lineage relationships, we applied it to a designed lineage in cell culture (Fig. 3.6). A monoclonal population of EGFP-v1 positive cells was transfected with editing reagents to induce a first round of mutations

---

Figure 3.5 *(following page)*: Counting edited barcode alleles with unique molecular identifiers (UMIs) and building lineage trees by maximum parsimony.(A) Single genomic copies of barcodes, each derived from a single cell, were tagged by performing either one or two polymerase extension cycles using a single primer with ten degenerate bases, i.e. a unique molecular identifier, or UMI. Amplicon sequencing reads were initially cleaned to remove low quality bases using the Trimmomatic software package. Sequencing reads were then aggregated using this UMI tag, and a consensus read was created by aligning matched UMI reads using the MAFFT aligner. These consensus reads were then merged using the FLASH bioinformatics tool. Both the consensus merged reads as well as any unmerged read pairs were aligned to the reference sequence, and insertions and deletions over target sites were called for each UMI-specific barcode. These barcode calls were used for downstream analysis. (B) Maximum parsimony was performed as described in the Methods. Briefly, individual alleles were converted to an indicator matrix, and maximum parsimony reconstruction was performed with the PHYLIP Mix program. The output was parsed to recover inferred ancestral alleles. To reduce the number of bifurcations in the tree, internal nodes with identical alleles were merged. Parent / child pairs internal to the tree that shared an inferred allele were first identified. When such a pair was found, the grandchildren nodes were moved to the parent node, and the child node was removed. This was repeated until no such pairs could be found. The resulting multifurcating tree was then visualized with custom scripts.


A
UMI sequence (NNNNNNNNNN) ~ 1 million combinations
each tagging an individual cell's integrated target
subsequent amplicon PCR
Sequence,
aggregate reads by unique UMI tag
Multiple sequence alignment
of matching UMI fragments
with MAFFT
Removal of inaccurate UMI
capture events, consensus calling
to collapse UMI reads to single fragment
consensus
merging of consensus read pairs
with FLASH
Alignment to reference sequence with
NEEDLEALL and call events within
cutsite window
reference
NONE NONE 61D 61D 61D NONE
barcode representation
of genome editing at each target site
Analysis pipeline
processing / output
B
Encoding into binary event status
matrixfor maximum parsimony using PHYLIP MIX :
allele #1:
NONE NONE 61D 61D 61D NONE NONE NONE 45D 45D
110000000.....
1) Run PHYLIP with Camin-Sokal maximum parsimony
2) Parse bifurcating tree output, recover inferred internal node alleles
merge internal parent-child nodes that have identical alleles,
resulting in multifurcating branch points
NONE 4D 6D 3D 7I NONE ...
NONE 4D 6D NONE 7I NONE ...
NONE 4D 6D NONE 7I NONE ...
NONE 4D 6D 3D 7I NONE ...
NONE 4D 6D NONE 7I NONE ...

in the v1 barcode. Clones derived from single cells were expanded, sampled, split, and re-transfected with editing reagents to induce a second round of mutations of the v1 barcode. For each clonal population, two 100-cell samples of the re-edited populations were expanded and harvested for gDNA. In these experiments, we began incorporating unique molecular identifiers (UMIs; 10 bp) during amplification of barcodes by a single round of polymerase extension (3.5A). Each UMI tags the single barcode present within each single cell, thereby allowing for correction of subsequent PCR amplification bias and enabling each UMI-barcode combination to be interpreted as deriving from a single cell (Miner et al., 2004).

Seven of twelve clonal populations we isolated contained mutations in the v1 barcode that were unambiguously introduced during the first round of editing (Fig. 3.6A). Additional edits accumulated in re-edited cells but generally did not disrupt the early edits (Fig. 3.6B and Fig. 3.7). We next sought to reconstruct the lineage relationships between all alleles observed in the experiment using a maximum parsimony approach (Fig. 3.5B)(26). The resultant tree contained major clades that were defined by the early edits present in each lineage (Fig. 3.6C). Four clonal populations (#3, #5, #7 and #8) were cleanly separated upon lineage reconstruction, with >99.7% of cells accurately placed into each lineage's major clade.

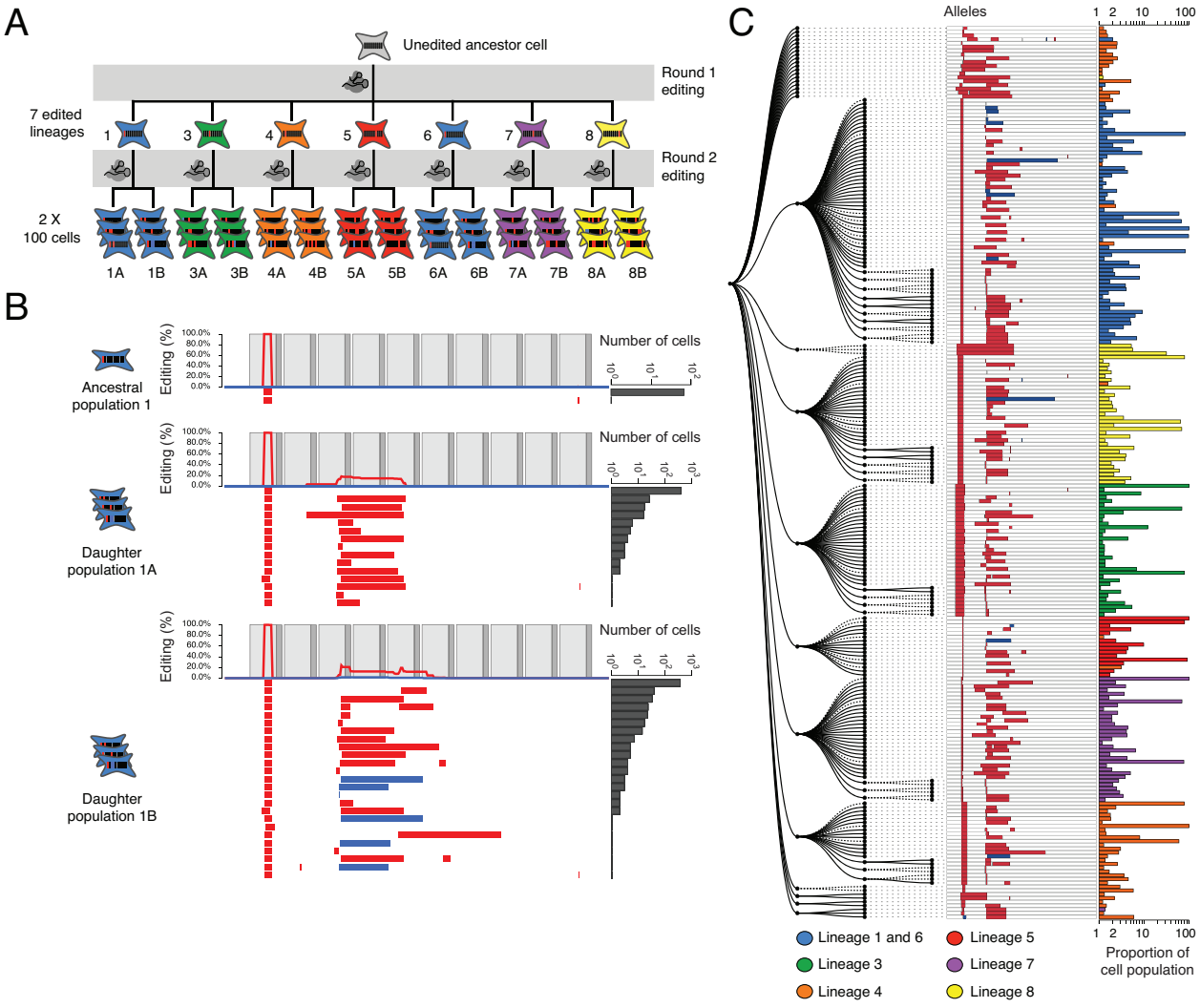
Two lineages (#1 and #6) were mixed because they shared identical mutations from the first round of editing. These most likely represent the recurrence of the same editing event across multiple lineages, but could also have been daughter cells subsequent to a single, early editing event prior to isolating clones. Consequently, 99.9% of cells of these two lineages were assigned to a single clade (Fig. 3.6C, blue). One clonal population (#4) appears to have derived from two independent cells, one of which harbored an unedited barcode. Later editing of these barcodes confounded the assignment of this lineage on the tree. Overall, however, these results demonstrate that GESTALT can be used to capture and reconstruct cell lineage relationships in cultured cells.

3.5 Combinatorial and cumulative editing of a compact genomic barcode in zebrafish

To test the potential of GESTALT for *in vivo* lineage tracing in a complex multicellular organism, we turned to the zebrafish Danio rerio. We designed two new barcodes, v6 and v7, each with ten sgRNA target sites that are absent from the zebrafish genome and predicted to be highly editable (methods). In contrast to v1-v5, in which the target sites are variably editable by one sgRNA, the targets within v6 or v7 are designed to be edited by distinct

---

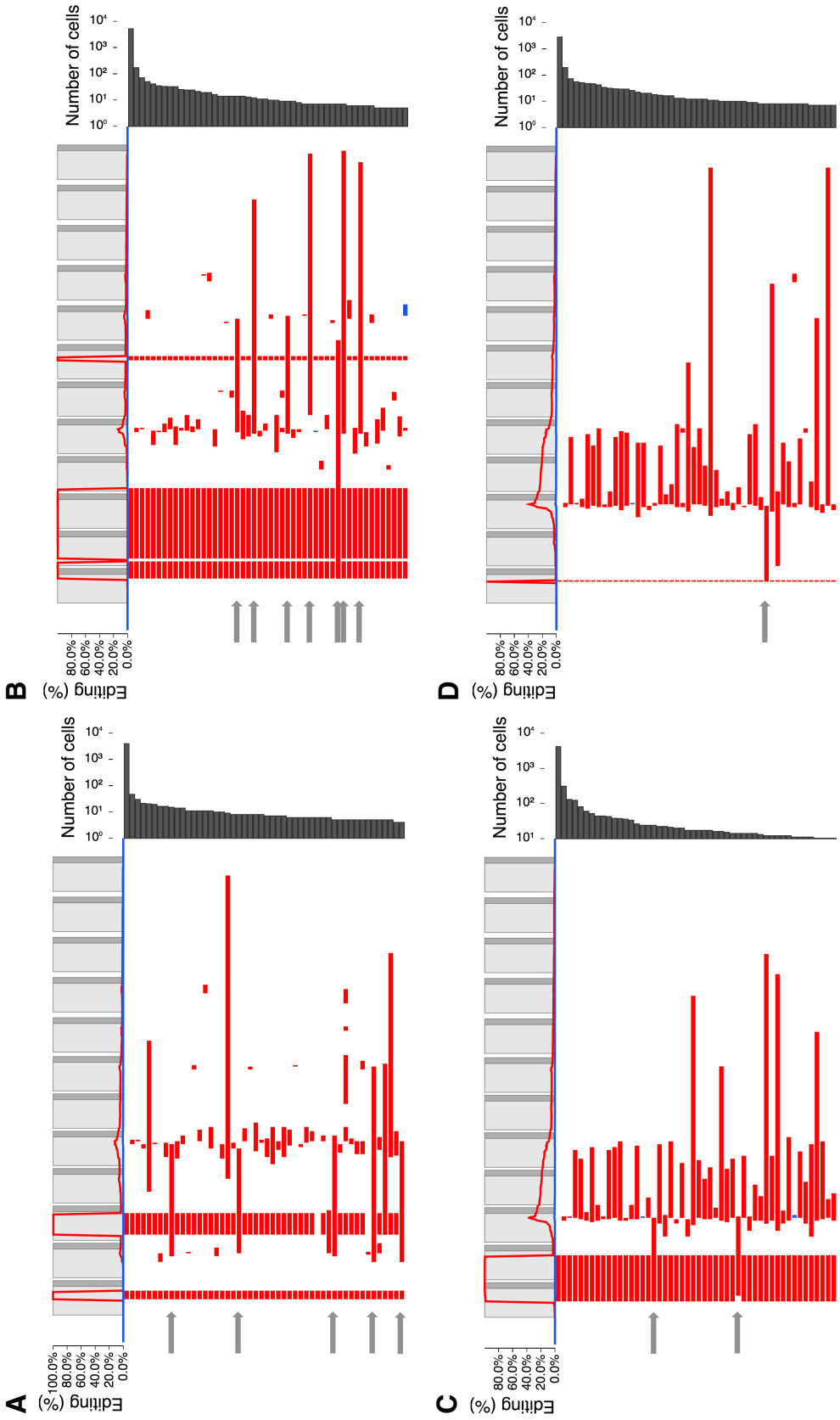
Figure 3.6 *(following page)*: Reconstruction of a synthetic lineage based on genome editing and targeted sequencing of edited barcodes. (A) A monoclonal population of cells was subjected to editing of the v1 array. Single cells were expanded, sampled (#1 to #12), re-transfected to induce a second round of barcode editing, and then expanded and sampled from 100-cell subpopulations (#1a, 1b to #12a, 12b). For clarity, the five clones where the original population was unedited are not shown. (B) Alleles observed in the synthetic lineage experiment are shown, with layout as described in the Fig. 3.1B legend. Cell population #1 represents sampling of cells that had been subjected to only the first round of editing; virtually all cells contain a shared edit to the first target. Populations #1a and #1b are derived from #1 but subjected to a second round of editing prior to sampling. These retain the edit to the first target, but subpopulations bear additional edits to other targets. (C) Maximum parsimony reconstruction using PHYLIP Mix (see Materials and Methods and Fig. 3.5B) from alleles seen two or more times in the seven cell lineages represented in panel (A). Lineage membership and abundance of each allele are shown on the right. Progenitor cell lineage #4 (orange) appears to be derived from two cells, one edited and the other wild-type: only 62% of lineage #4 falls into a single clade, consistent with the proportion (64%) of the lineage edited after the first round. We assume that cells unedited in the first round either accrued edits matching other lineages (thus causing mixing), or accrued different edits (thus remaining outside the major clades).


A
Unedited ancestor cell
Round 1 editing
Round 2 editing
7 edited lineages
2 X 100 cells
1 3 4 5 6 7 8
1A 1B 3A 3B 4A 4B 5A 5B 6A 6B 7A 7B 8A 8B
B
Ancestral population 1
Daughter population 1A
Daughter population 1B
Editing (%)
Number of cells
C
Alleles
Lineage 1 and 6
Lineage 3
Lineage 4
Lineage 5
Lineage 7
Lineage 8
Proportion of cell population

sgRNAs. We generated transgenic zebrafish that harbor each barcode in the 3' UTR of DsRed driven by the ubiquitin promoter (Kawakami, 2007; Porter et al., 2014) and a GFP marker that is expressed in the cardiomyocytes of the heart (Fig. 3.8) (Pan et al., 2013). To evaluate whether diverse alleles could be generated by *in vivo* genome editing, we injected Cas9 and ten different sgRNAs with perfect complementarity to the barcode target sites into

---

Figure 3.7 *(following page)*: Low frequency elimination of lineage-specific edits by re-editing of the v5 barcode in cell culture. A population of HEK293T cells bearing the unedited v5 barcode was subjected to initial editing by transfection with pX330-v1. Monoclonal populations containing edited v5 barcodes were then cultured and re-transfected twice to induce additional editing. The outcomes of re-editing are shown for each population with barcode editing plots (as in Fig. 3.1B) showing the top 50 alleles. In each, the top allele corresponds to the parental allele verified by sequencing each monoclonal line. We observed a variety of mechanisms in which an established edit was lost, examples of which are highlighted with gray arrows. (A) Loss of an existing deletion at site 3 occurred in 4.2% of cells (16.7% of cells with barcodes that were re-edited). Loss of this edit appears to have arisen from simultaneous Cas9 cleavage at site 2 and any of sites 4 through 10, thus forming a larger deletion spanning the original site 3 deletion. (B) An initial site 7 deletion was removed by re-editing in 3.0% of cells (12.7% of re-edited barcodes), likely due to the same mechanism describe in A. (C) A single deletion event spanning sites 1 and 2 was disrupted in 3.3% of cells (7.6% of re-edited barcodes). These alleles likely formed by deletions at sites 3 that extended as far as or beyond the ancestral deletion in sites 1 and 2. (D) Re-editing of a target site bearing a single 1 bp deletion occurred in 0.9% of cells (1.9% of re-edited barcodes), presumably consequential to residual Cas9 activity at the edited site. These experiments highlight examples of information loss in the v5 barcode, but understanding how often this occurs in different barcode designs and editing systems remains to be determined.


A
B
C
D
Number of cells
Editing (%)
100.0%
80.0%
60.0%
40.0%
20.0%
0.0%

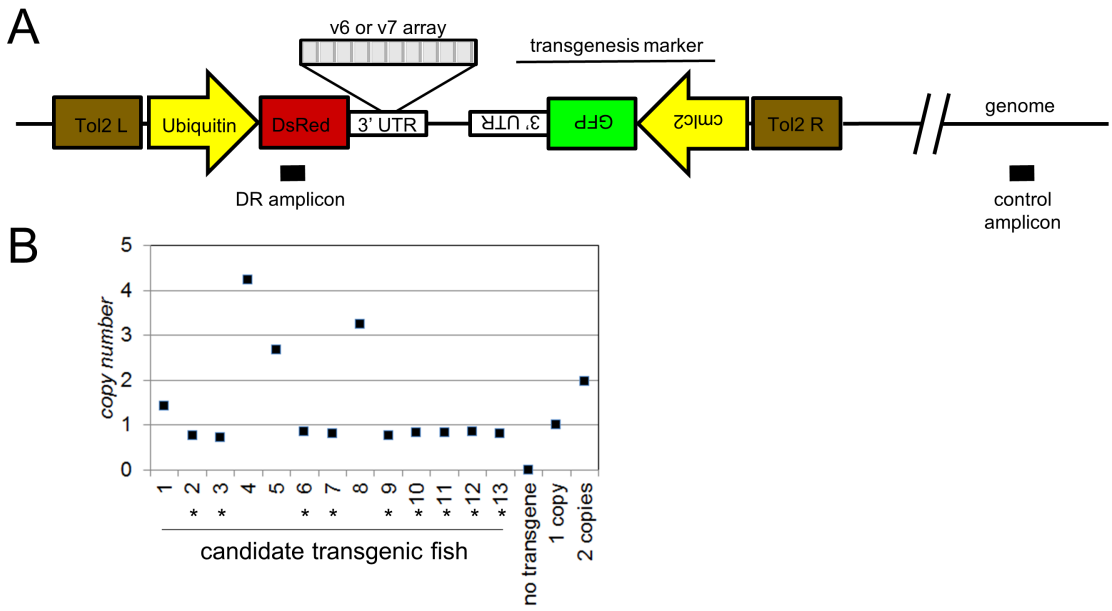


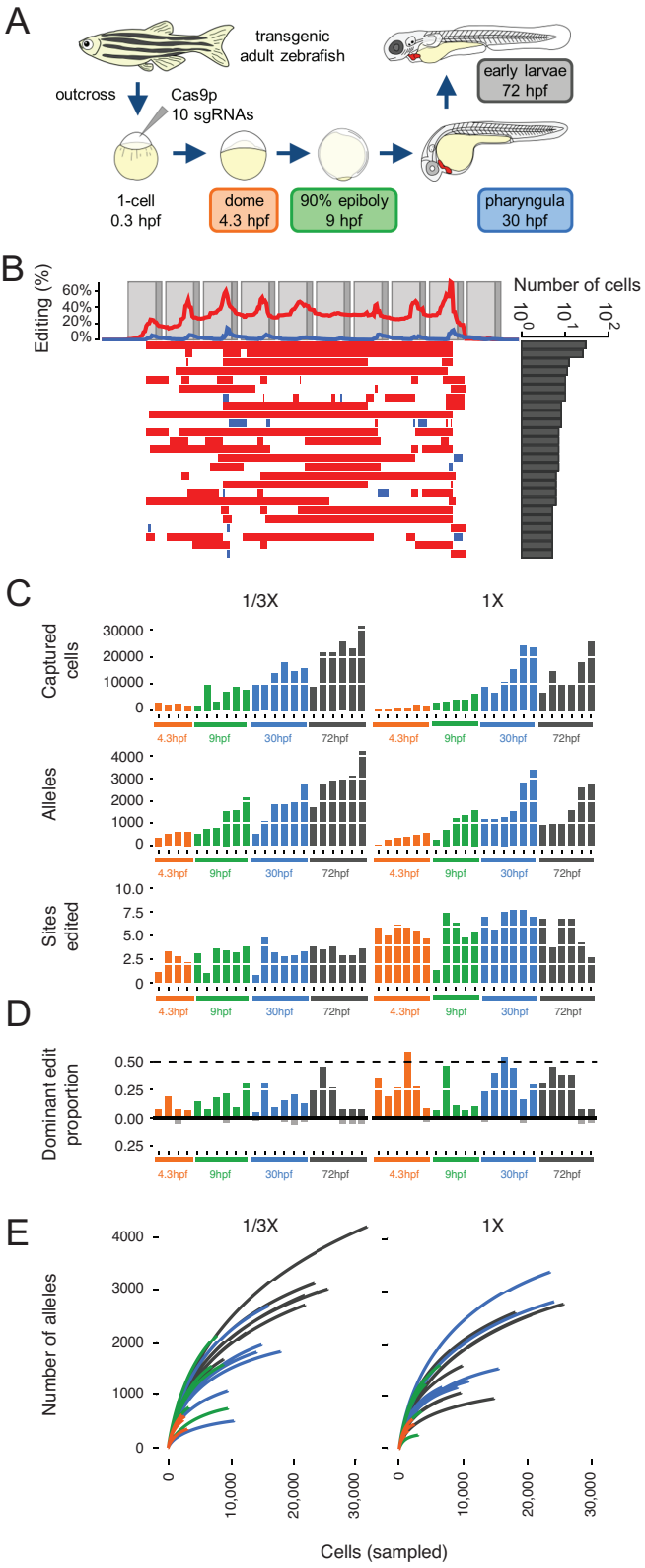
Figure 3.8: Generation of single copy transgenic v6 or v7 zebrafish. (A) Diagram of the barcode transgene with Tol2 integration arms, a Ubiquitin promoter upstream of DsRed with either the v6 or v7 barcode embedded in the 3' UTR, and a cmlc2:GFP transgenesis marker. (B) Quantification of transgene copy number by qPCR using DR amplicon and control amplicon as indicated in (A). Copy number was determined using the ddCt method and reference non-transgenic, 1-copy and 2-copy transgenic animals. Putative single copy individuals are indicated by asterisks (∗)

single-cell v6 embryos (Fig. 3.9A).

Editing of integrated barcodes had no noticeable effects on development (Fig. 3.10). To characterize barcode editing *in vivo*, we extracted gDNA from a series of single 30 hours

---

Figure 3.9 *(following page)*: Generating combinatorial barcode diversity in transgenic zebrafish. (A) One-cell zebrafish embryos were injected with complexed Cas9 ribonucleoproteins (RNPs) containing sgRNAs that matched each of the 10 targets in the array (v6 or v7). Embryos were collected at time points indicated. UMI-tagged barcodes were amplified and sequenced from genomic DNA. (B) Patterns of editing in alleles recovered from a 30 hpf v6 embryo, with layout as described in the Fig. 3.1B legend. (C) Bar plots show the number of cells sampled (top), unique alleles observed (middle) and proportion of sites edited (bottom) for 45 v7 embryos collected at four developmental time-points and two levels of Cas9 RNP (1/3x, 1x). Colors correspond to stages shown in panel (A). Although more alleles are observed with sampling of larger numbers of cells at later time points, the proportion of target sites edited remains relatively constant. (D) Bar plots show the proportion of edited barcodes containing the most common editing event in a given embryo. Six of 45 embryos had the most common edit in approximately 50% of cells (dashed line), consistent with this edit having occurred at the two-cell stage (see Fig. 3.12A for example). Colors correspond to stages shown in panel (A). These same edits are rarer or absent in other embryos (black bars below). (E) For each of the 45 v7 embryos, all barcodes observed were sampled without replacement. The cumulative number of unique alleles observed as a function of the number of cells sampled is shown (average of the 500 iterations shown per embryo; two levels of Cas9 RNP: 1/3x on left, 1x on right). The number of unique alleles observed, even in later developmental stages where we are sampling much larger numbers of cells, appears to saturate, and there is no consistent pattern supporting substantially greater diversity in later time-points, consistent with the bottom row of panel (C) in supporting the conclusion that the majority of editing occurs before dome stage.


A
transgenic
adult zebrafish
outcross
Cas9p
10 sgRNAs
early larvae
72 hpf
1-cell
0.3 hpf
dome
4.3 hpf
90% epiboly
9 hpf
pharyngula
30 hpf
B
Editing (%)
60%
40%
20%
0%
Number of cells
C
1/3X
1X
Captured cells
Alleles
Sites edited
4.3hpf
9hpf
30hpf
72hpf
D
Dominant edit proportion
E
Number of alleles
Cells (sampled)

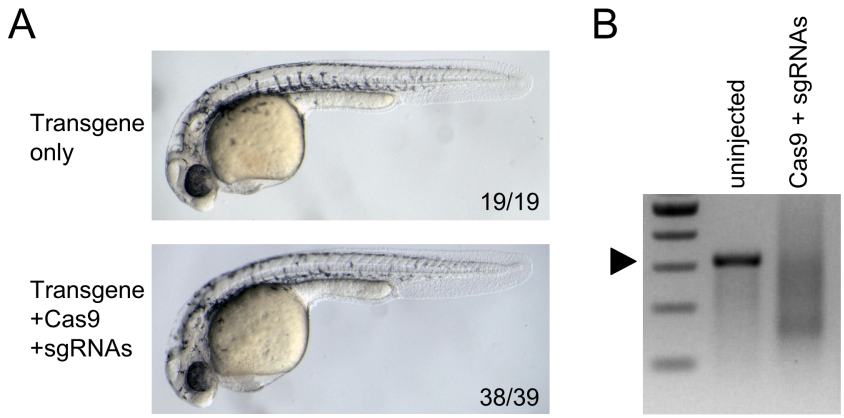


Figure 3.10: Barcode editing in transgenic zebrafish embryos is robust and does not affect development. (A) Representative v6 transgenic embryos uninjected or injected with Cas9 protein pre-complexed with a mix of 10 sgRNAs are shown, with phenotypic penetrance indicated. (B) Gel electrophoresis of v6 array PCR products from uninjected and injected embryos show extensive barcode editing as a smear of smaller molecules running predominantly below the expected size of the unedited barcode (arrowhead, 311 bp).

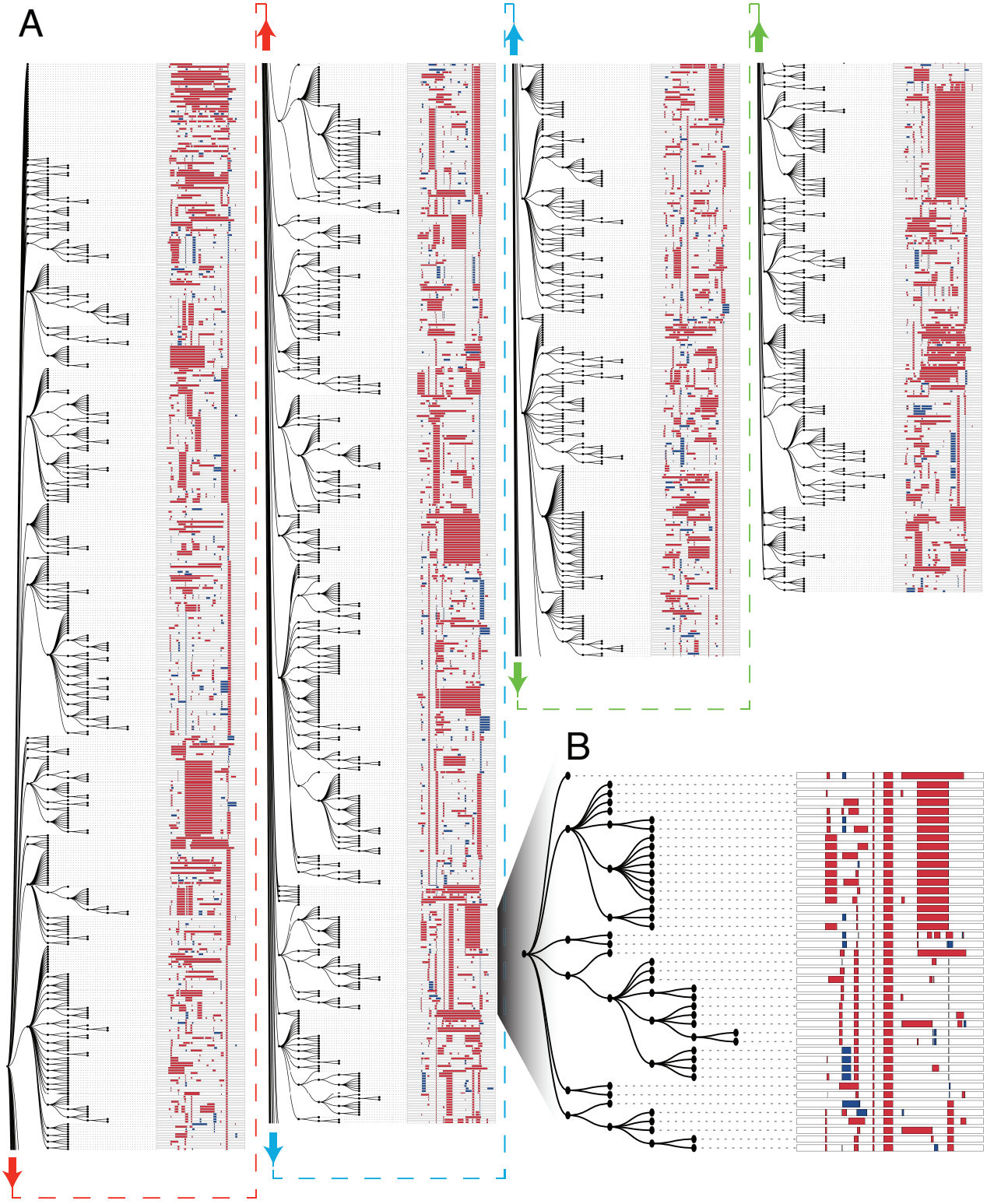
post fertilization (hpf) embryos, and UMI-tagged, amplified and sequenced the v6 barcode. In control embryos (Cas9-; n = 2), all 4,488 captured barcodes were unedited. In contrast, in edited embryos (Cas9+; n = 8), fewer than 1% of captured barcodes were unedited. We recovered barcodes from hundreds of cells per embryo (median 943; range 257-2,832) and identified dozens to hundreds of alleles per embryo (median 225; range 86-1,323). 41% +/- 10% of alleles were observed recurrently within single embryos, most likely reflecting alleles that were generated in a progenitor of two or more cells. Fewer than 0.01% of alleles were shared in pairwise comparisons of embryos, revealing the highly stochastic nature of editing in different embryos. These results demonstrate that GESTALT can generate very high allelic diversity *in vivo*.

3.6 Reconstruction of lineage relationships in embryos

To evaluate whether lineage relationships can be reconstructed using edited barcodes, we focused on the v6 embryo with the lowest rates of inter-target deletions and edited target

---

Figure 3.11 *(following page)*: Lineage reconstruction of an edited zebrafish embryo. (A) A lineage reconstruction of 1,323 alleles recovered from the v6 embryo also represented in Fig. 3.9B, generated by a maximum parsimony approach implemented in the PHYLIP Mix package (see Materials and Methods and Fig. 3.5). A dendrogram to the left of each column represents the lineage relationships, and the alleles are represented on the right. Each row represents a unique allele. Matched colored arrows and dashed lines connect subsections of the tree together. There are many large clades of alleles sharing specific edits, as well as sub-clades defined by 'dependent' edits. These dependent edits occur within a clade defined by a more frequent edit but are rare or absent elsewhere in the tree. (B) A portion of the tree is shown at higher resolution. Two edits are shared by all alleles in this clade. Six independent edits define descendent sub-clades within this clade, and further edits define additional sub-sub-clades within the clade.


A
B

sites (Fig. 3.9B; avg. 58% +/- 27% of target sites no longer a perfect match to the unedited target, compared to 87% +/- 21% for all other 30 hpf v6 embryos). Application of our parsimony approach (Fig. 3.5B) to the 1,961 cells in which we observed 1,323 distinct alleles generated the large tree shown in Fig. 3.11. 1,307 of the 1,323 (98%) alleles could be related to at least one other allele by one or more shared edits, 85% by two or more shared edits, and 56% by three or more shared edits. These results illustrate the principle of using patterns of shared edits between distinct barcode alleles to reconstruct their lineage relationships *in vivo*.

3.7 Developmental timing of barcode editing

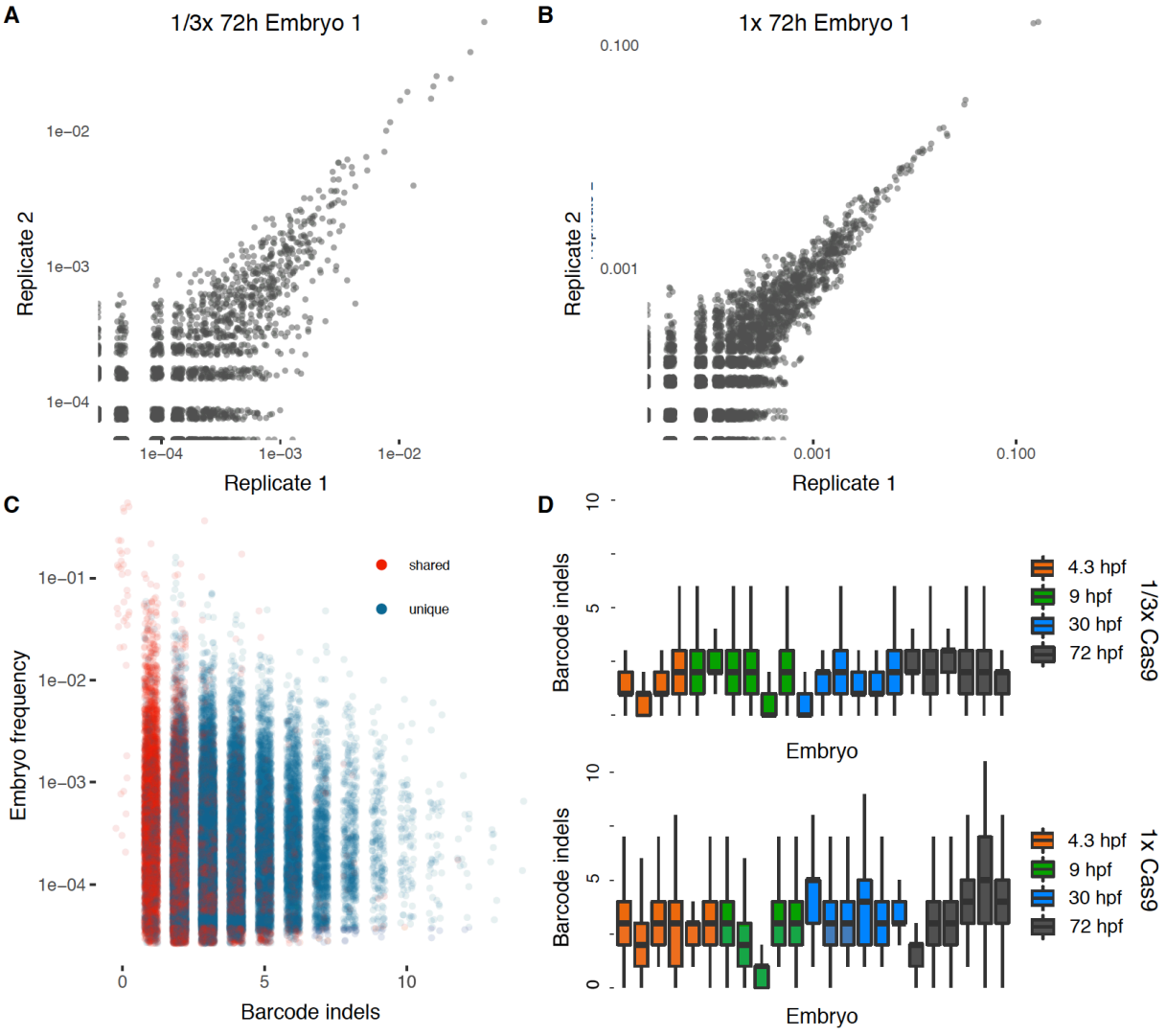
To determine the developmental timing of barcode editing, we injected Cas9 and ten sgRNAs into one-cell stage v7 transgenic embryos and harvested genomic DNA before gastrulation (dome stage, 4.3 hpf; n = 10 animals), after gastrulation (90% epiboly / bud stage, 9 hpf; n = 11 animals), at pharyngula stage (30 hpf; n = 12 animals), and from early larvae (72 hpf; n = 12 animals) (Fig. 3.9A). We recovered barcode sequences from a median of 8,785 cells per embryo (range 461-31,640; total of 45 embryos), comprising a median of 1,223 alleles per embryo (range 15-4,195) (Fig. 3.9C). Within single embryos, 65% +/- 6% of alleles were observed recurrently, whereas in pairwise comparisons of embryos only 2% +/- 5% of alleles were observed recurrently. The abundances of alleles were well-correlated between technical replicates for each of two 72 hpf embryos (Fig. 3.12A and B), and alleles containing many edits were more likely to be unique to an embryo than those with few edits (Fig. 3.12). To assess when editing begins, we analyzed the proportions of the most common editing events across all barcodes sequenced in a given embryo, reasoning that the earliest edits would be the most frequent. Across eight v6 and 45 v7 embryos, we never observed an edit that was present in 100% of cells. This observation indicates that no permanent edits were introduced at the one-cell stage. In nearly all embryos, we observe that the most common edit is present in >10% of cells, and in some cases in 50% of cells (Fig. 3.9D and Fig. 3.13). This observation also holds in 4,000-cell dome stage embryos, which result from approximately 12 rounds of

largely synchronous division unaccompanied by cell death. Most of these edits are rare or absent in other embryos, suggesting they are unlikely to have arisen recurrently within each lineage. These results suggest that the edits present in 50% of cells were introduced at the two-cell stage and that the edits present in >10% of cells were introduced before the 16-cell stage.

How long does barcode editing persist? Two aspects of the data suggest that it tapers relatively early in development. First, in dome stage embryos (4.3 hpf), we captured barcodes

---

Figure 3.12 *(following page)*: Characteristics of Cas9-mediated barcode editing across zebrafish embryos. After isolation of genomic DNA from each of two 72 hpf embryos injected with either (A) 1/3x volume or (B) 1x volume, the material was split and two separate amplification reactions performed, replicates 1 and 2. Unique Molecular Identifiers (UMIs) were used to tag genomic copies of embryo barcodes, such that each UMI's consensus sequencing call corresponds to a single cell's edited barcode. For each embryo, allele frequencies from UMI-tagging technical replicates are plotted against each other (i.e. technical replicates of one embryo in panel (A), and technical replicates of another embryo in panel (B)). Each point corresponds to a single barcode present in the union of the replicates. Pearson correlations: (A) = 0.96, (B) = 0.998. Spearman correlations: (A) = 0.42, (B) = 0.64. (C) To compare barcodes that were shared between embryos to those that were unique to a single embryo, we plotted each allele's indel count against its proportion on a per embryo basis (n = 45). Alleles that were shared had significantly fewer indels than those that were seen in only one embryo (2.01 mean indels per shared allele vs. 3.52 mean indels per embryo-specific allele (Wilcoxon Rank Sum (WRS); P << 0.00001)). The mean frequency within an embryo of a shared allele was modestly higher than a unique allele: 0.12% vs. 0.052%, respectively (WRS; P = 3.8x10-25). (D) Boxplots show the distribution of barcode indel events per embryo. (Bold line; median, box; 25th to 75th percentile; whiskers extend to the furthest point within 1.5x the interquartile range (IQR) of the box; outliers not shown)


A
1/3x 72h Embryo 1
Replicate 2
Replicate 1
1e−02
1e−03
1e−04
B
1x 72h Embryo 1
0.100
0.001
C
Embryo frequency
Barcode indels
shared
unique
D
Barcode indels
Embryo
4.3 hpf
9 hpf
30 hpf
72 hpf
1/3x Cas9
1x Cas9

from a median of 2,086 cells, in which a median of 4.8 targets were edited. Although the number of cells and alleles that we were able to sample increased at the later developmental stages, the proportion of edited sites appeared relatively stable (Fig. 3.9C). If editing were occurring throughout this time course, we would instead expect the proportion of edited sites to increase substantially. Second, the number of unique alleles appears to saturate early, never exceeding 4,200 (Fig. 3.9E). For example, only 4,195 alleles were observed in a 72 hpf embryo in which we sampled the highest number of cells (n = 31,639). These results suggest that the majority of editing events occurred before dome stage.

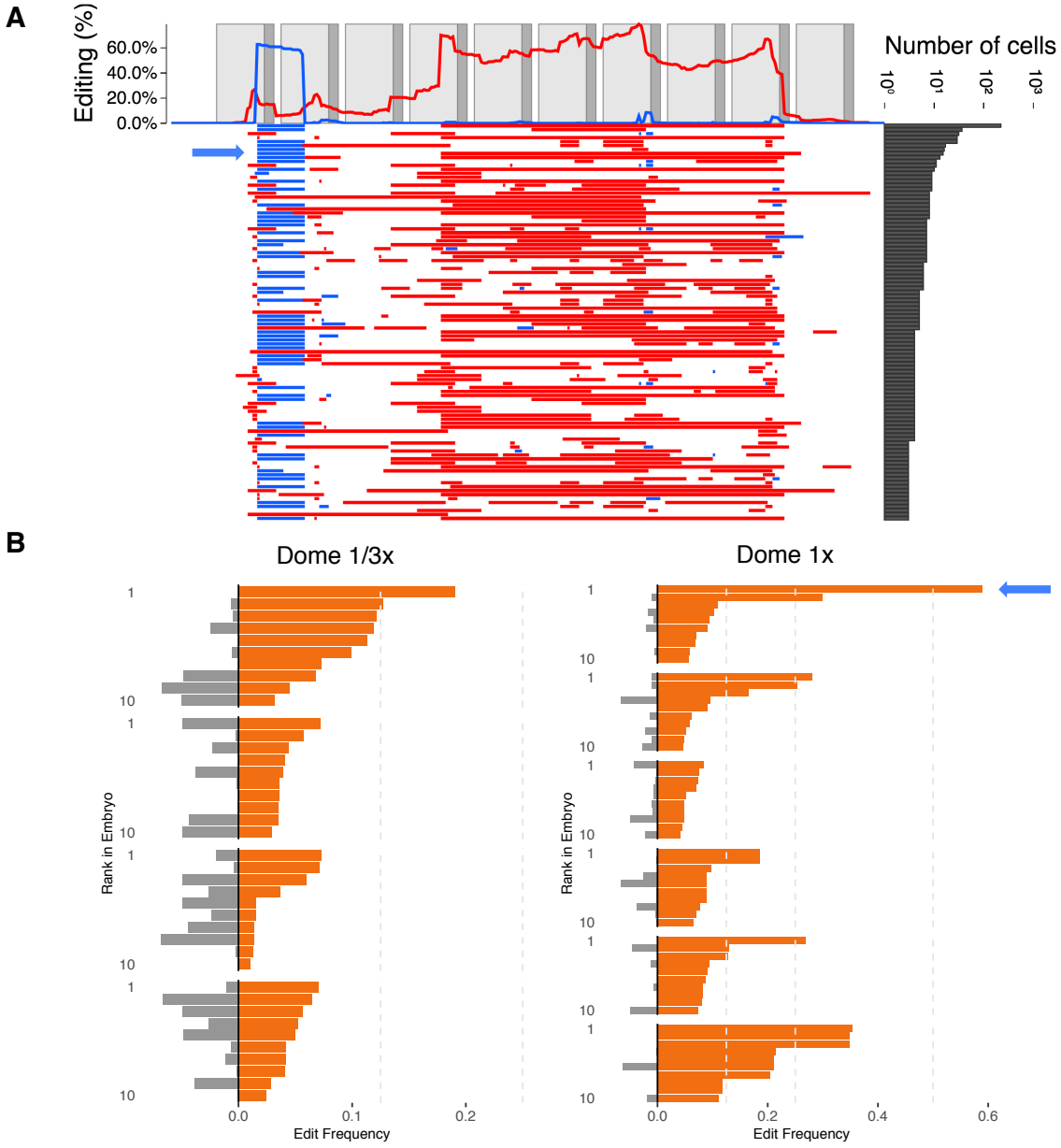
3.8 Editing diversity in adult organs

To evaluate whether barcodes edited during embryogenesis can be recovered in adults, we dissected two edited 4-month old v7 transgenic zebrafish (ADR1 and ADR2) (Fig. 3.14A). We collected organs representing all germ layers - the brain and both eyes (ectodermal), the intestinal bulb and posterior intestine (endodermal), the heart and blood (mesodermal), and the gills (neural crest, with contributions from other germ layers). We further divided the

---

Figure 3.13 *(following page)*: Abundances of the most common editing events in each embryo often reflect the onset of editing. (A) A barcode editing plot showing the top 50 alleles from one dome stage embryo (1x #1) exemplifies a high frequency editing event - in this case a 20 bp insertion at the first target site (blue bar indicated by arrow). The event is seen in 59.0% of barcodes and was absent from all other embryos, suggesting it derived from editing at the two-cell stage. (B) The frequencies of the top ten indels from each of ten dome stage embryos are shown in (orange), plotted next to the average frequency of that indel in all other embryos (gray). Cases in which the edit is common in one embryo and rare in others strongly suggest the edit's abundance resulted from occurring early in development, and not from stereotypical double-strand break repair outcomes during to barcode editing. The arrow indicates the event corresponding to the common insertion shown in (A).

[p!]

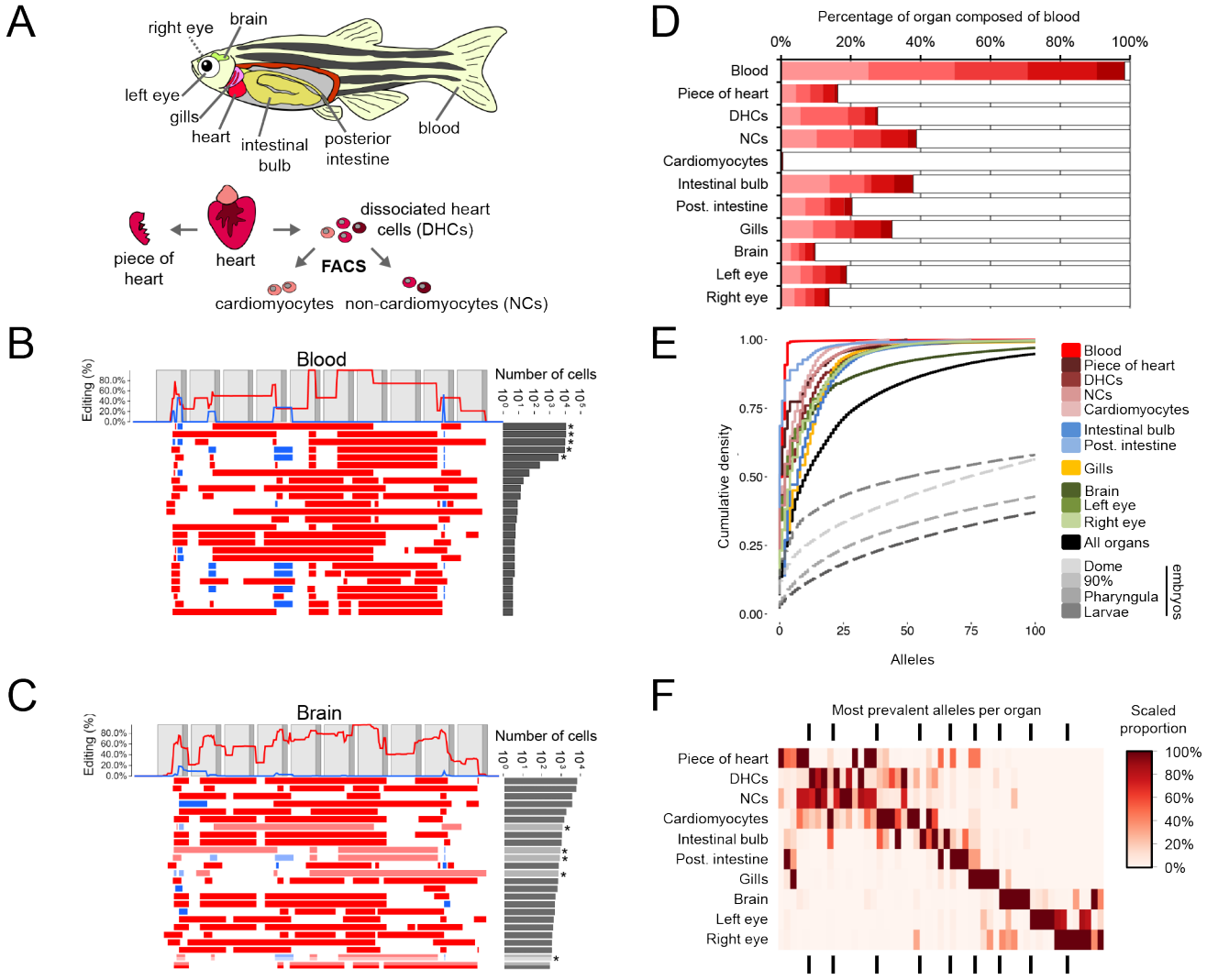

A
Editing (%)
60.0%
40.0%
20.0%
0.0%
Number of cells
10⁰
10¹
10²
10³
B
Dome 1/3x
Dome 1x
Rank in Embryo
Edit Frequency
0.0
0.1
0.2
0.4
0.6

heart into four samples: a piece of heart tissue, dissociated unsorted cells (DHCs), FACS-sorted GFP+ cardiomyocytes, and non-cardiomyocyte heart cells (NCs) (Fig. 3.15).

We isolated genomic DNA from each sample, amplified and sequenced edited barcodes with high technical reproducibility (Fig. 3.16), and observed barcode editing rates akin to

---

Figure 3.14 *(following page)*: Organ-specific progenitor cell dominance. (A) The indicated organs were dissected from a single adult v7 transgenic edited zebrafish (ADR1). A blood sample was collected as described in the Methods. The heart was further split into the four samples shown (Fig. 3.15). (B) Patterns of editing in the most prevalent 25 alleles (out of 135 total) recovered from the blood sample. Layout as described in the Fig 3.1B legend. The most prevalent 5 alleles (indicated by asterisks) comprise >98% of observed cells. (C) Patterns of editing in the most prevalent 25 alleles (out of 399 total) recovered from brain. Layout as described in the Fig. 3.1B legend. Alleles that have identical editing patterns compared to the most prevalent blood alleles are indicated by asterisks and light shading. (D) The five dominant blood alleles (shades of red) are present in varying proportions (10-40%) in all intact organs except the FACS-sorted cardiomyocyte population (0.5%). All other alleles are summed in grey. (E) The cumulative proportion of cells (y-axis) represented by the most frequent alleles (x-axis) for each adult organ of ADR1 is shown, as well as the adult organs in aggregate. In all adult organs except blood, the five dominant blood alleles are excluded. All organs exhibit dominance of sampled cells by a small number of progenitors, with fewer than 7 alleles comprising the majority of cells. For comparison, a similar plot for the median embryo (dashed) from each time-point of the developmental time course experiment is also shown. (F) The distribution of the most prevalent alleles for each organ, after removal of the five dominant blood alleles, across all organs. The most prevalent alleles were defined as being at >5% abundance in a given organ (median 5 alleles, range 4-7). Organ proportions were normalized by column and colored as shown in legend. Underlying data presented in table S2.


A
right eye
brain
left eye
gills
heart
intestinal bulb
posterior intestine
blood
piece of heart
heart
dissociated heart cells (DHCs)
FACS
cardiomyocytes
non-cardiomyocytes (NCs)
B
Blood
Editing (%)
Number of cells
C
Brain
D
Percentage of organ composed of blood
0%
20%
40%
60%
80%
100%
Blood
Piece of heart
DHCs
NCs
Cardiomyocytes
Intestinal bulb
Post. intestine
Gills
Brain
Left eye
Right eye
E
Cumulative density
Alleles
Blood
Piece of heart
DHCs
NCs
Cardiomyocytes
Intestinal bulb
Post. intestine
Gills
Brain
Left eye
Right eye
All organs
Dome
90%
Pharyngula
Larvae
embryos
F
Most prevalent alleles per organ
Scaled proportion
100%
80%
60%
40%
20%
0%

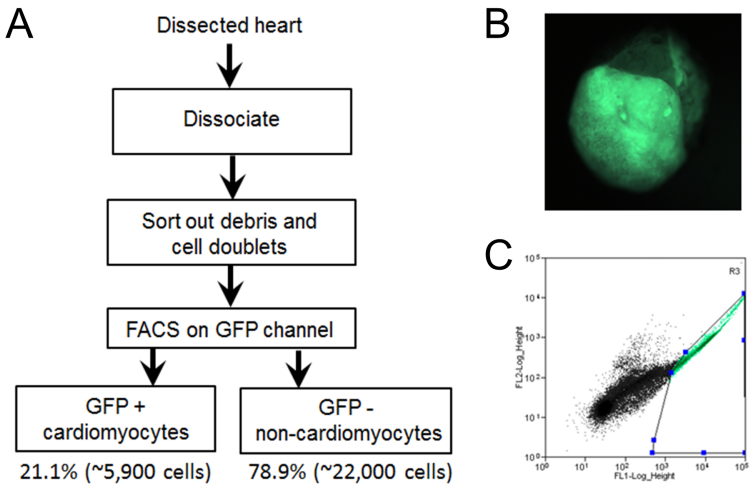


Figure 3.15: FACS sorting of cardiomyocytes and non-cardiomyocyte heart cells]. (A) Schematic. Adult transgenic hearts (example in B) were dissected, dissociated, and sorted via FACS. Gates were applied to remove cellular debris (exclude high SSC-H, low FSC-H, keep 15.6% of events as cells) and cell doublets (exclude high SSC-W, keep 97.2% of events as single cells) before gating on the ratio of GFP:RFP fluorescence (C) to sort GFP+ cardiomyocytes from GFP- non-cardiomyocyte heart cells (21.1% GFP+, 78.9% GFP-). Percentages provided are for ADR1.

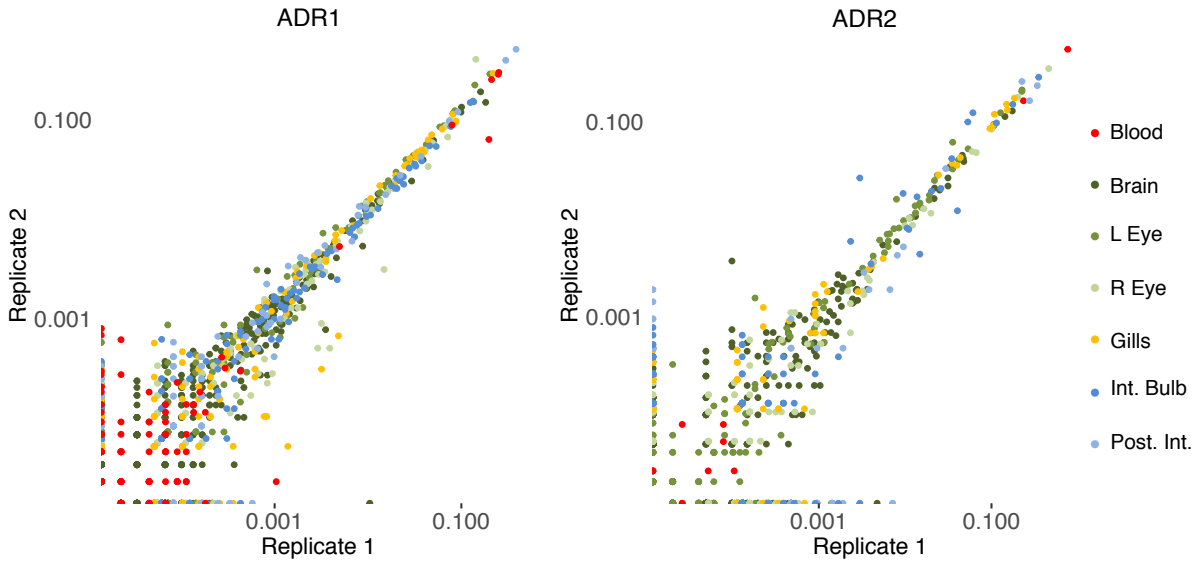


Figure 3.16: Reproducibility of barcode sampling from adult zebrafish organs. After isolation of genomic DNA, two separate amplification reactions (Replicate 1 and Replicate 2) were performed for each of seven organs from ADR1 and ADR2. UMIs were used to tag genomic copies of the barcodes. Replicate samples were sequenced in separate runs, and UMI consensus read thresholds were set proportional to sequencing depth. For each organ in each fish, allele frequencies for each replicate are plotted against each other. Each point corresponds to a single barcode present in the union of the two replicate samplings, colored by organ. Pearson correlations calculated from log10-transformed values are: ADR1 = 0.90, ADR2 = 0.85. For this analysis, the top five (ADR1) or two (ADR2) blood alleles were computationally removed from non-blood samples.

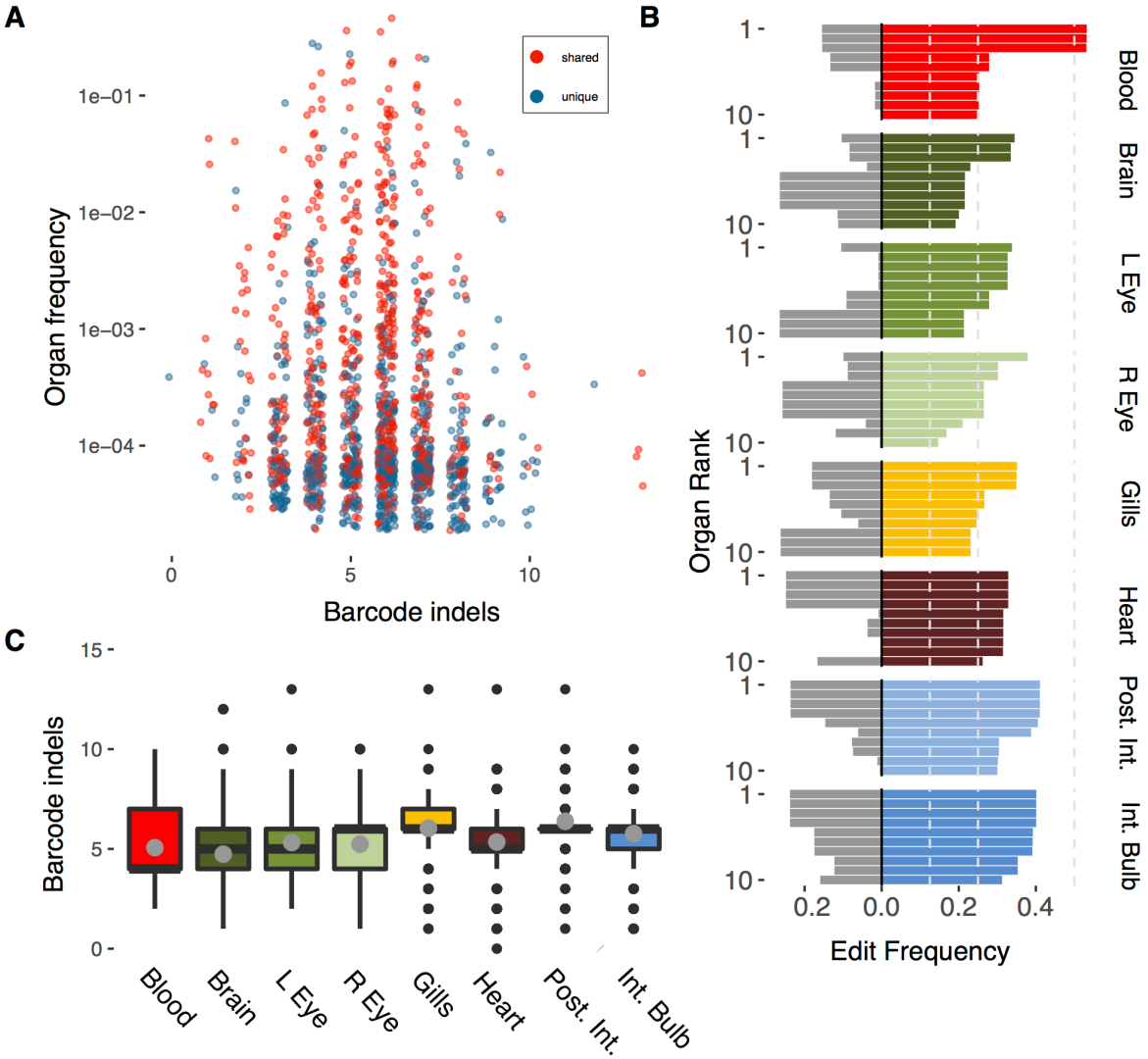
those in embryos (Fig. 3.17). For zebrafish ADR1, we captured barcodes from between 776 and 44,239 cells from each tissue sample (median 17,335), corresponding to a total of 197,461 cells and 1,138 alleles. For zebrafish ADR2, we captured barcodes from between 84 and 52,984 cells from each tissue sample (median 20,973), corresponding to a total of 217,763 cells and 2,016 alleles. These results show that edits introduced to the barcode during embryogenesis are inherited through development and tissue homeostasis and can be detected in adult organs.

3.9 Differential contribution of embryonic progenitors to adult organs

To analyze the contribution of diverse alleles to different organs, we compared the frequency of edited barcodes within and between organs. We first examined blood (of note, zebrafish erythrocytes are nucleated (Thisse and Zon, 2002)). Only 5 alleles defined over 98% of cells in the ADR1 blood sample (Fig. 3.14B), suggesting highly clonal origins of the adult

---

Figure 3.17 *(following page)*: Barcode editing characteristics in organs from adult zebrafish ADR1. For these analyses, the top five blood alleles were computationally removed from non-blood samples. (A) Sharing of alleles across organs as a function of frequency within organs and the number of indels. Compared to barcodes shared between unrelated embryos by chance recurrence (Fig. 3.12), alleles contributing to multiple organs (red) harbor more indels and are seen at higher proportions compared to alleles restricted to a single organ (blue), suggesting sharing is largely explained by way of developmental lineages as opposed to by chance. (B) Frequencies of the top ten indel events per organ (colored bars on right) plotted against the average frequency of that edit in all other organs (gray bars on left). The most common edits within organs are also seen frequently in other organs. (C) Boxplots show the distribution of barcode indel events within each organ from ADR1. (Bold line; median, box; 25th to 75th percentile; whiskers extend to the furthest point within 1.5x the IQR from the box, gray dot; mean value, outliers not shown)


A
shared
unique
1e−01
1e−02
1e−03
1e−04
Organ frequency
0
5
10
Barcode indels
B
Organ Rank
1
10
Blood
Brain
L Eye
R Eye
Gills
Heart
Post. Int.
Int. Bulb
0.2
0.0
0.2
0.4
Edit Frequency
C
Barcode indels
15
10
5
0
Blood
Brain
L Eye
R Eye
Gills
Heart
Post. Int.
Int. Bulb

zebrafish blood system from a few embryonic progenitors. Consistent with the presence of blood in all dissected organs, these common blood alleles were also observed in all organs (10-40%; Fig. 3.14C) but largely absent from cardiomyocytes isolated by flow sorting (0.5%). Furthermore, the relative proportions of these five alleles remained constant in all dissected organs, suggesting that they primarily mark the blood and do not substantially contribute to non-blood lineages (Fig. 3.14D). In performing similar analyses of clonality across all organs (while excluding the five most common blood alleles), we observed that a small subset of alleles dominates each organ (Fig. 3.14E). Indeed, for all dissected organs, fewer than 7 alleles comprised >50% of cells (median 4, range 2-6), and, with the exception of the brain, fewer than 25 alleles comprised >90% of cells (median 19, range 4-38). Most of these dominant alleles were organ-specific, i.e. although they were found rarely in other organs, they tended to be dominant in only one organ (Fig. 3.14F). For example, the most frequent allele observed in the intestinal bulb comprised 13.6% of captured non-blood cells observed in that organ, but <0.01% of cells observed in any other organ. There are exceptions, however. For example, one allele is observed in 24.7% of sorted cardiomyocytes, 13.4% of the intestinal bulb, and at lower abundances in all other organs. Similar results were observed in ADR2 (Fig. 3.18). These results indicate that the majority of cells in diverse adult organs are descended from a few differentially edited embryonic precursors.

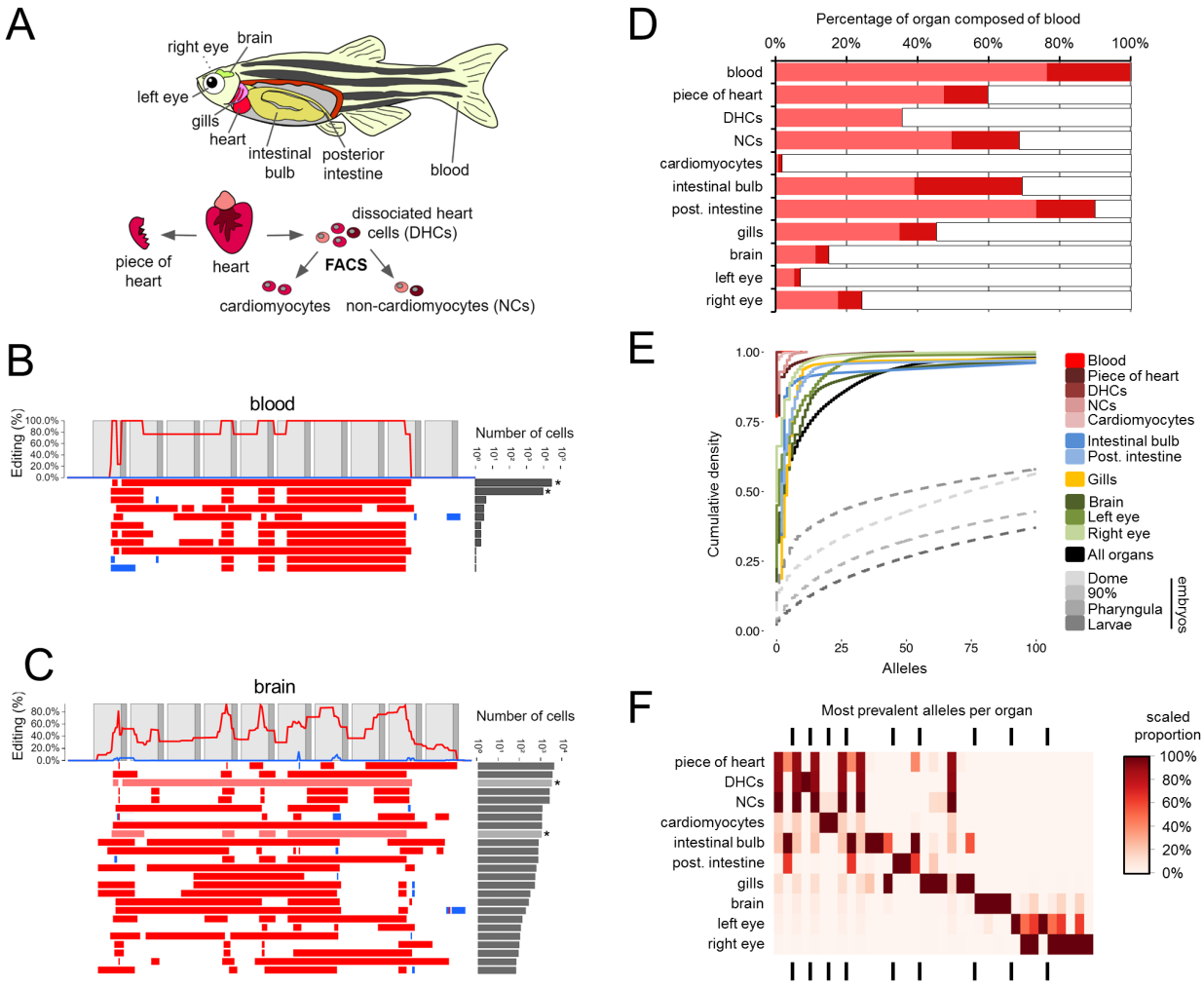
3.10 Reconstructing lineage relationships in adult organs

To reconstruct the lineage relationships between cells both within and across organs on the basis of shared edits, we again relied on maximum parsimony methods (Fig. 3.5B). The resulting trees for ADR1 and ADR2 are shown in Fig. 3.22 and Fig. 3.19, respectively. We observed clades of alleles that shared specific edits. For example, ADR1 had 8 major clades, each defined by 'ancestral' edits that are shared by all captured cells assigned to that clade (Fig. 3.24A; also indicated by colors in the tree shown in Fig. 3.22). Collectively, these clades comprised 49% of alleles and 90% of the 197,461 cells sampled from ADR1 (Fig. 3.24A). Blood was contributed to by 3 major clades (#3, #6, #7) (Fig. 3.24B).

After re-allocating the 5 dominant blood alleles from the composition of individual organs back to blood (Fig. 3.14B and Fig. 3.20), we observed that all major clades made highly non-uniform contributions across organs. For example, clade #3 contributed almost exclusively to mesodermal and endodermal organs, while clade #5 contributed almost exclusively to ectodermal organs. These results reveal that GESTALT can be used to infer the contributions

---

Figure 3.18 *(following page)*: Organ-specific progenitor cell dominance in ADR2. (A) The indicated organs were dissected from a single adult v7 transgenic edited zebrafish (ADR1). A blood sample was collected as described in the Methods. The heart was further split into the four samples shown (Fig 3.18). (B) Patterns of editing in the 11 alleles recovered from the blood sample. Layout as described in the Fig 3.1B legend. The most prevalent 2 alleles (indicated by asterisks) comprise >98% of observed cells. (C) Patterns of editing in the most prevalent 25 alleles (out of 699 total) recovered from brain. Layout as described in the Fig 3.1B legend. Alleles that are identical in sequence to the most prevalent blood alleles are indicated by asterisks and light shading. (D) The five dominant blood alleles (shades of red) are present in varying proportions (7-90%) in all intact organs except the FACS-sorted cardiomyocyte population (2%). All other alleles are summed in grey. (E) The cumulative proportion of cells (y-axis) represented by the most frequent alleles (x-axis) for each adult organ of ADR1 is shown, as well as the adult organs in aggregate. In all adult organs except blood, the five dominant blood alleles are excluded. All organs exhibit dominance of sampled cells by a small number of progenitors, with fewer than 5 alleles comprising the majority of cells. For comparison, a similar plot for the median embryo (dashed) from each time-point of the developmental time course experiment is also shown. (F) The distribution of the most prevalent alleles for each organ, after removal of the five dominant blood alleles, across all organs. The most prevalent alleles were defined as being at >5% abundance in a given organ (median 4 alleles, range 4-6). Organ proportions were normalized by column and colored as shown in legend. Underlying data presented in table S2.


A
right eye
brain
left eye
gills
heart
intestinal bulb
posterior intestine
blood
piece of heart
heart
dissociated heart cells (DHCs)
FACS
cardiomyocytes
non-cardiomyocytes (NCs)
B
blood
Editing (%)
Number of cells
C
brain
D
Percentage of organ composed of blood
blood
piece of heart
DHCs
NCs
cardiomyocytes
intestinal bulb
post. intestine
gills
brain
left eye
right eye
E
Cumulative density
Alleles
Blood
Piece of heart
DHCs
NCs
Cardiomyocytes
Intestinal bulb
Post. intestine
Gills
Brain
Left eye
Right eye
All organs
Dome
90%
Pharyngula
Larvae
embryos
F
Most prevalent alleles per organ
scaled proportion

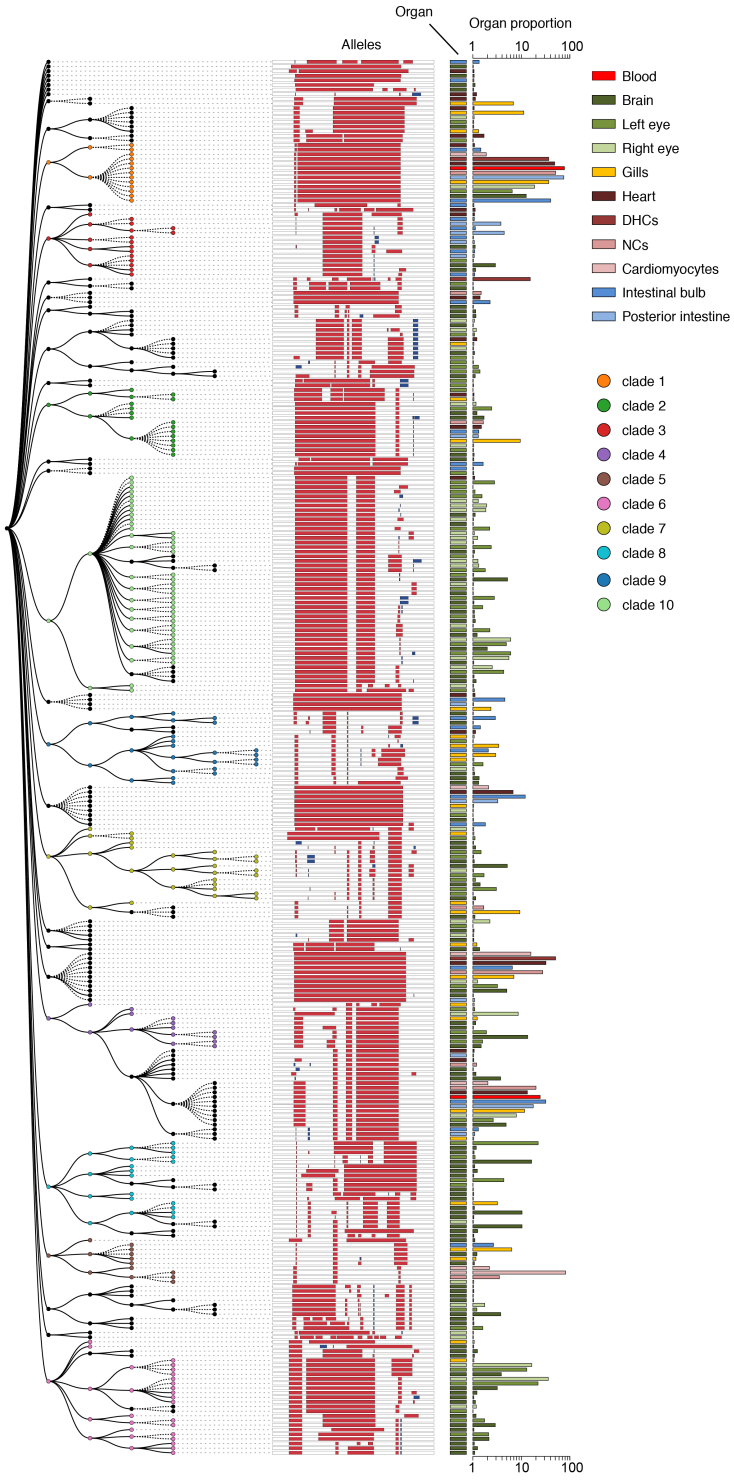
of inferred ancestral progenitors to adult organs.

Although some ancestral clades appear to contribute to all germ layers, we find that subclades, defined by additional shared edits within a clade, exhibit greater specificity. For example, while clade #1 contributes substantially to all organs except blood, additional edits divide clade #1 into three subclades with greater tissue restriction (Fig. 3.24C and D). The #1+A subclade primarily contributes to mesendodermal organs (heart, both gastrointestinal organs) while the #1+C subclade primarily contributes to neuroectodermal organs (brain, left eye, and gills). Similar patterns are observed for clade #2 (Fig. 3.24E and F), where the #2+A subclade contributes primarily to mesendodermal organs, the #2+B subclade to the heart, and the #2+C clade to neuroectodermal organs. Additional edits divide these subclades into further tissue-specific sub-subclades.

For example, while the #2+A subclade is predominantly mesendoderm, additional edits define #2+A+D (heart, primarily cardiomyocytes), #2+A+E (heart and posterior intestine), and #2+A+F (intestinal bulb). All of the major clades exhibit similar patterns of increasing restriction with additional edits (Fig. 3.24C-F and Fig. 3.21). Similar observations were made in fish ADR2 (Fig. 3.23). These results indicate that GESTALT can record lin-

---

Figure 3.19 *(following page)*: Lineage reconstruction for adult zebrafish ADR2. Unique alleles sequenced from adult zebrafish organs can be related to one another using a maximum parsimony approach into a multifurcating lineage tree. For reasons of space, we show a tree reconstructed from the 302 ADR2 alleles observed at least 5 times in individual organs. Ten major clades are displayed with colored nodes, each defined by 'ancestral' edits that are shared by all alleles assigned to that clade (shown in Fig. 3.20). Editing patterns in individual alleles are represented as shown previously. Alleles in multiple organs are plotted on separate lines per organ and these nodes connected with stippled branches. Two sets of bars outside the alleles identify the organ in which the allele was observed and the proportion of cells in that organ represented by that allele (log scale).


Organ
Organ proportion
Alleles
1
10
100
Blood
Brain
Left eye
Right eye
Gills
Heart
DHCs
NCs
Cardiomyocytes
Intestinal bulb
Posterior intestine
clade 1
clade 2
clade 3
clade 4
clade 5
clade 6
clade 7
clade 8
clade 9
clade 10

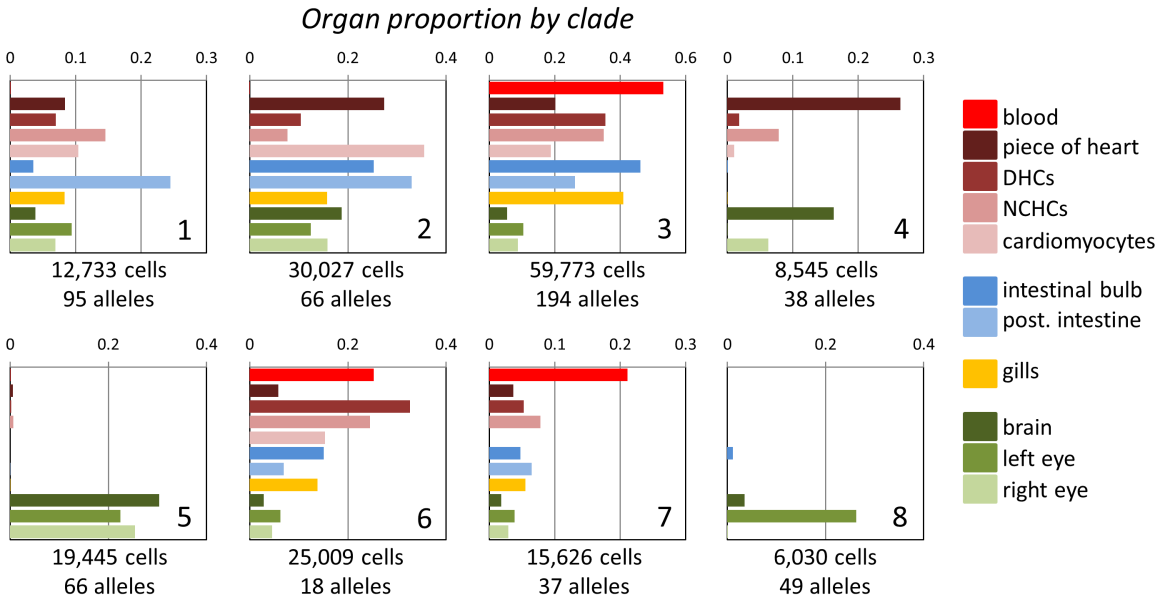


Figure 3.20: Contributions of the eight major clades within ADR1 to each organ, prior to the reassignment of the most prevalent blood alleles. The total number of cells and unique alleles within a given major clade are listed below. Fig. 3.24 shows similar information, but after the reassignment of the dominant blood alleles from all organs to blood. For heart subsamples, 'piece of heart' = a piece of heart tissue, 'DHCs' = dissociated unsorted cells; 'cardiomyocytes' = FACS-sorted GFP+ cardiomyocytes; and 'NCHCs' = non-cardiomyocyte heart cells.

eage relationships across many cell divisions and capture information both before and during tissue restriction.

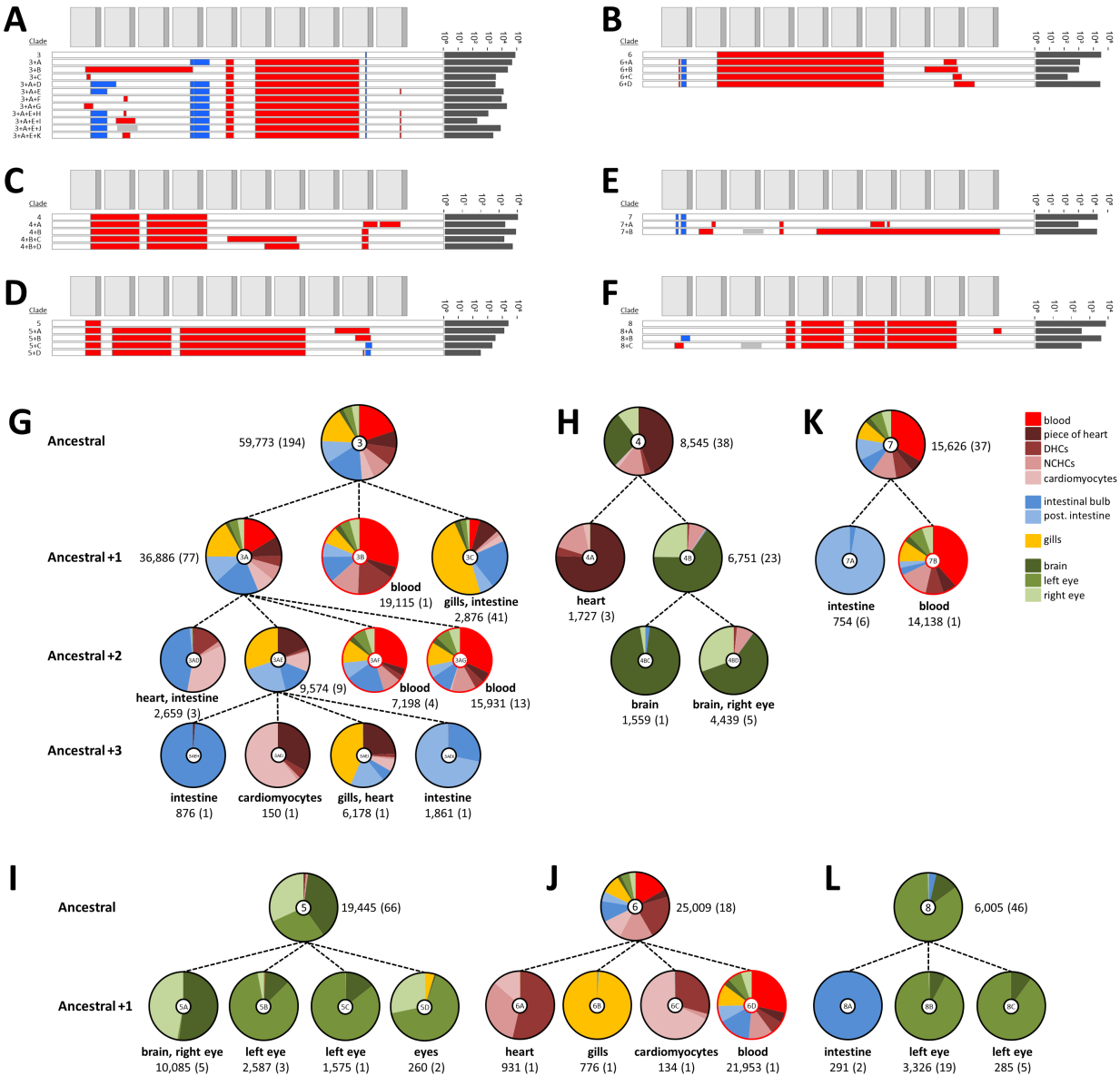
3.11 Discussion

We describe a new method, GESTALT, which uses combinatorial and cumulative genome editing to record cell lineage information in a highly multiplexed fashion. We successfully applied this method to both artificial lineages (cell culture) as well as to whole organisms (zebrafish).

The strengths of GESTALT include: 1) the combinatorial diversity of mutations that can be generated within a dense array of CRISPR/Cas9 target sites; 2) the potential for informative mutations to accumulate across many cell divisions and throughout an organism's developmental history; 3) the ability to scalably query lineage information from at least

---

Figure 3.21 *(following page)*: Tracing lineage through editing patterns within additional ADR1 clades. (A-E) Edits that define subclades of clades #3 (A), #4 (B), #5 (C), #6 (D), #7 (E) and #8 (F), with the total number of cells in which these are observed indicated on the right. A grey box indicates an unedited site or sites, distinguishing it from related alleles that contain an edit at this location. (G-L) Lineage trees corresponding to subclades of #3 (G), #4 (H), #5 (I), #6 (J), #7 (K) and #8 (L) that show how dependent edits are associated with increasing lineage restriction. The pie chart at each node indicates the organ distribution within a clade or subclade. Ratios of cell proportions are plotted, a normalization which accounts for differential depth of sampling between organs. Labels in the center of each pie chart correspond to the subclade labels in (c/e). Alleles present in a clade but not assigned to a descendent subclade (either they have no additional lineage restriction or are at low abundance) are not plotted for clarity. The number of cells (and the number of unique alleles) are also listed, and terminal nodes also list major organ restriction(s), i.e. those comprising >25% of a subclade by proportion.


A
B
C
D
E
F
G
Ancestral
59,773 (194)
Ancestral +1
36,886 (77)
blood
19,115 (1)
gills, intestine
2,876 (41)
Ancestral +2
heart, intestine
2,659 (3)
9,574 (9)
blood
7,198 (4)
blood
15,931 (13)
Ancestral +3
intestine
876 (1)
cardiomyocytes
150 (1)
gills, heart
6,178 (1)
intestine
1,861 (1)
H
8,545 (38)
heart
1,727 (3)
6,751 (23)
brain
1,559 (1)
brain, right eye
4,439 (5)
K
15,626 (37)
intestine
754 (6)
blood
14,138 (1)
blood
piece of heart
DHCs
NCHCs
cardiomyocytes
intestinal bulb
post. intestine
gills
brain
left eye
right eye
I
Ancestral
19,445 (66)
Ancestral +1
brain, right eye
10,085 (5)
left eye
2,587 (3)
left eye
1,575 (1)
eyes
260 (2)
J
25,009 (18)
heart
931 (1)
gills
776 (1)
cardiomyocytes
134 (1)
blood
21,953 (1)
L
6,005 (46)
intestine
291 (2)
left eye
3,326 (19)
left eye
285 (5)

hundreds of thousands of cells and with a single sequencing read per single cell; 4) the likely applicability of GESTALT to any organism, from bacteria and plants to vertebrates, that allows genome editing, as well as human cells (e.g. tumor xenografts). Even in organisms in which transgenesis is not established, lineage tracing by genome editing may be feasible by expressing editing reagents to densely mutate an endogenous, non-essential genomic sequence.

Our experiments also highlight several remaining technical challenges. Chief amongst these are: 1) the chance recurrence of identical edits or similar patterns of edits in distantly related cells can confound lineage inference; 2) non-uniform editing efficiencies and inter-target deletions within the barcode contribute to suboptimal sequence diversity and loss of information, respectively; 3) the transient means by which Cas9 and sgRNAs are introduced likely restrict editing to early embryogenesis; 4) the computational challenge of precisely defining the multiple editing events that give rise to different alleles complicates the unequivocal reconstruction of lineage trees; and 5) the difficulty of isolating tissues without contamination by blood and other cells can hinder the assignment of alleles to specific organs.

Figure 3.22 *(following page)*: Lineage reconstruction for adult zebrafish ADR1. Unique alleles sequenced from adult zebrafish organs can be related to one another using a maximum parsimony approach implemented in the PHYLIP Mix package (see Materials and Methods and Fig. 3.5B). For reasons of space, we show a tree reconstructed from the 601 ADR1 alleles observed at least five times in individual organs. Eight major clades are displayed with colored nodes, each defined by 'ancestral' edits that are shared by all alleles assigned to that clade (shown in Fig. 3.24A). Editing patterns in individual alleles are represented as shown previously. Alleles observed in multiple organs are plotted on separate lines per organ and are connected with stippled branches. Two sets of bars outside the alleles identify the organ in which the allele was observed and the proportion of cells in that organ represented by that allele (log scale).

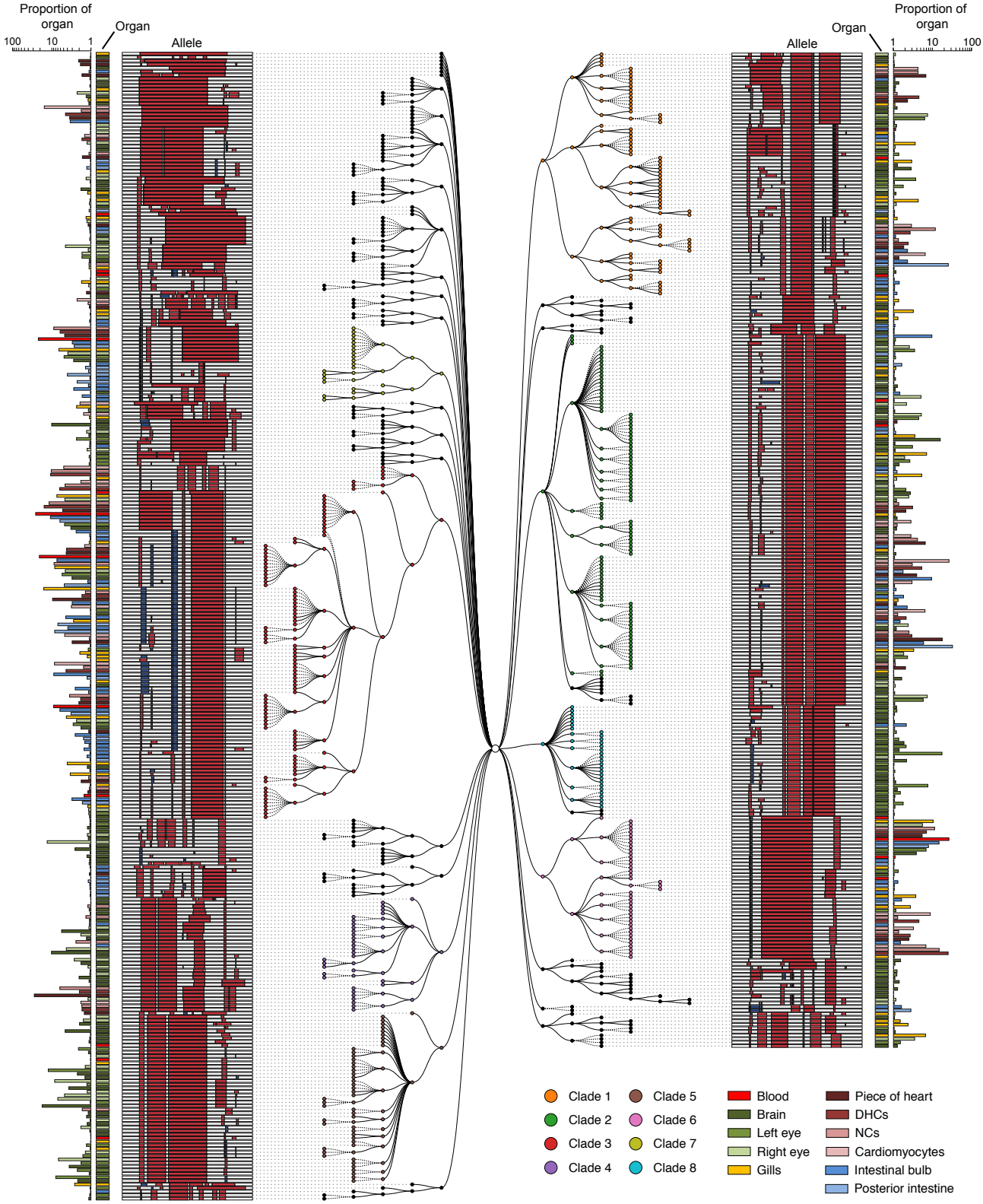
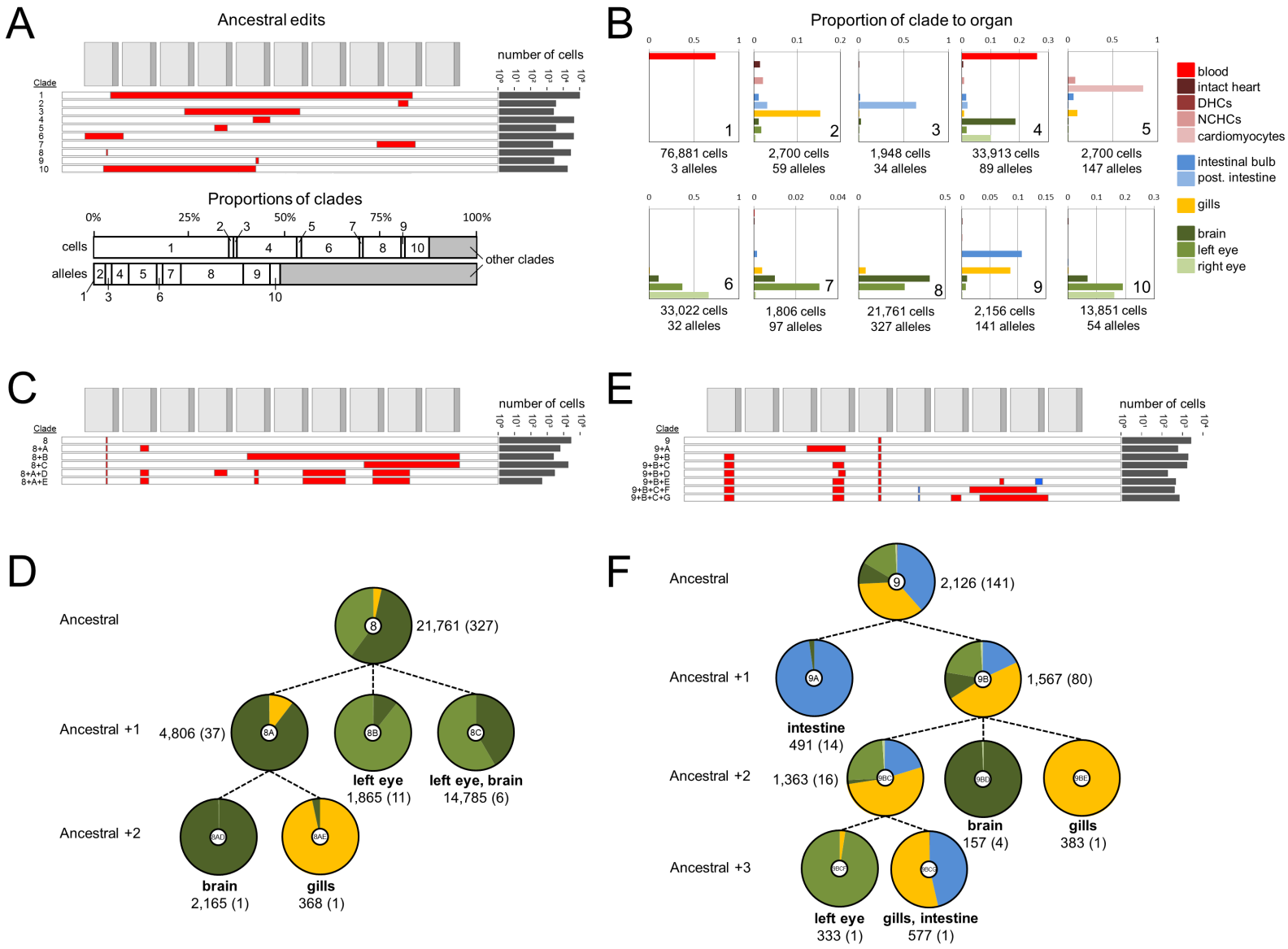

Proportion of organ
100
10
1
Organ
Allele
Allele
Organ
Proportion of organ
1
10
100
Clade 1
Clade 2
Clade 3
Clade 4
Clade 5
Clade 6
Clade 7
Clade 8
Blood
Brain
Left eye
Right eye
Gills
Piece of heart
DHCs
NCs
Cardiomyocytes
Intestinal bulb
Posterior intestine

Figure 3.23 *(following page)*: Clades and subclades corresponding to inferred progenitors exhibit increasing levels of organ restriction in ADR2. (A) Top panel: The inferred ancestral edits that define ten major clades of ADR1, as determined by parsimony, are shown, with the total number of cells in which these are observed indicated on the right. Bottom panel: Contributions of the ten major clades to all cells or all alleles. 126 alleles (out of 2,016 total) that contained ancestral edits from more than one clade were excluded from assignment to any clade, and any further lineage analysis. (B) Contributions of each of the ten major clades to each organ, displayed as a proportion of each organ. To accurately display the contributions of the ten major clades to each organ, we first re-assigned the two dominant blood alleles from other organs back to the blood. The total number of cells and alleles within a given major clade are listed below. The clade contributions of all clades and subclades are presented in table S3. For heart subsamples, 'piece of heart' = a piece of heart tissue, 'DHCs' = dissociated unsorted cells; 'cardiomyocytes' = FACS-sorted GFP+ cardiomyocytes; and 'NCHCs' = non-cardiomyocyte heart cells. (C/E) Edits that define subclades of clade #8 (C) and clade #9 (E), with the total number of cells in which these are observed indicated on the right. (D/F) Lineage trees corresponding to subclades of clade #8 (D) and clade #9 (F) that show how dependent edits are associated with increasing lineage restriction. The pie chart at each node indicates the organ distribution within a clade or subclade. Ratios of cell proportions are plotted, a normalization which accounts for differential depth of sampling between organs. Labels in the center of each pie chart correspond to the subclade labels in (c/e). Alleles present in a clade but not assigned to a descendent subclade (either they have no additional lineage restriction or are at low abundance) are not plotted for clarity. The number of cells (and the number of unique alleles) are also listed, and terminal nodes also list major organ restriction(s), i.e. those comprising >25% of a subclade by proportion.

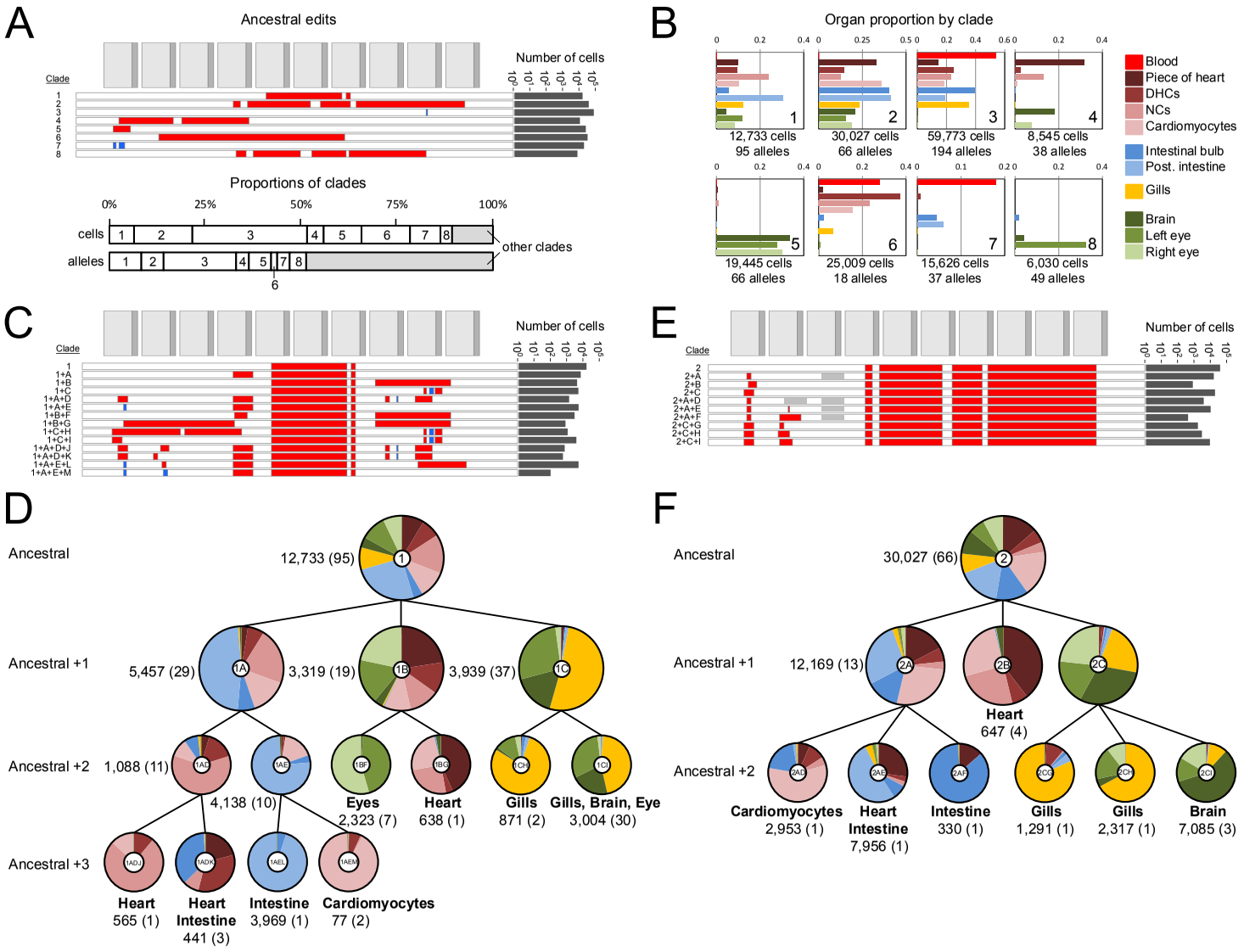

A
Ancestral edits
number of cells
Clade
Proportions of clades
cells
alleles
other clades
B
Proportion of clade to organ
76,881 cells
3 alleles
2,700 cells
59 alleles
1,948 cells
34 alleles
33,913 cells
89 alleles
2,700 cells
147 alleles
33,022 cells
32 alleles
1,806 cells
97 alleles
21,761 cells
327 alleles
2,156 cells
141 alleles
13,851 cells
54 alleles
blood
intact heart
DHCs
NCHCs
cardiomyocytes
intestinal bulb
post. intestine
gills
brain
left eye
right eye
C
E
D
Ancestral
21,761 (327)
Ancestral +1
4,806 (37)
left eye
1,865 (11)
left eye, brain
14,785 (6)
Ancestral +2
brain
2,165 (1)
gills
368 (1)
F
Ancestral
2,126 (141)
Ancestral +1
intestine
491 (14)
1,567 (80)
Ancestral +2
1,363 (16)
brain
157 (4)
gills
383 (1)
Ancestral +3
left eye
333 (1)
gills, intestine
577 (1)

A broader set of challenges includes the lack of information about the precise anatomical location and exact cell type of each queried cell, the fact that genome editing events are not directly coupled to the cell cycle, and the failure to recover all cells. These challenges currently hinder the reconstruction of a lineage tree as complete and precise as the one that Sulston and colleagues described for *C. elegans*. Despite these limitations, our proof-of-principle study shows that GESTALT can inform developmental biology by richly defining lineage relationships among vast numbers of cells recovered from an organism.

The current challenges highlight the need for further optimization of the design of targets and arrays, as well as the delivery of editing reagents. For example, an array containing twice as many targets as used here could fit within a single read on contemporary sequencing platforms, thus yielding more lineage information per cell without sacrificing throughput. Also, as we have shown, adjustments to the target sequences and dosages of editing reagents can be used to fine-tune mutation rates and to minimize undesirable inter-target deletions. Finally, sgRNA sequences and lengths (Fu et al., 2014), Cas9 cleavage activity and target preferences (Kleinstiver et al., 2015; Slaymaker et al., 2016), and the means by which Cas9 and sgRNA(s) are expressed (e.g. transient, constitutive (Platt et al., 2014), or induced (Ablain et al., 2015; Yin et al., 2015)), can be altered to control the pace, temporal window and tissue(s) at which the barcodes are mutated. For example, coupling editing to cell cycle progression might enable higher resolution reconstruction of lineage relationships throughout development.

Our application of GESTALT to a vertebrate model organism, zebrafish, demonstrates its potential to yield insights into developmental biology. First, our results suggest that relatively few embryonic progenitor cells give rise to the majority of cells of many adult zebrafish organs, reminiscent of clonal dominance (Gupta and Poss, 2012; Snippert et al., 2010). For example, only 5 of the 1,138 alleles observed in ADR1 gave rise to >98% of blood cells, and for all dissected organs, fewer than 7 alleles comprised >50% of cells. There are several mechanisms by which such dominance can emerge, e.g. by uneven starting populations in the embryo, drift, competition, interference, unequal cell proliferation or death, or

Figure 3.24 *(following page)*: Clades and subclades corresponding to inferred progenitors exhibit increasing levels of organ restriction. (A) Top panel: The parsimony inferred ancestral edits that define eight major clades of ADR1 are shown, with the total number of cells in which these are observed indicated on the right. Bottom panel: Contributions of the eight major clades to all cells or all alleles. 19 alleles (out of 1,138 total) that contained ancestral edits from more than one clade were excluded from assignment to any clade, and any further lineage analysis. (B) Contributions of each of the eight major clades to each organ, displayed as a proportion of each organ. To accurately display the contributions of the eight major clades to each organ, we first re-assigned the five dominant blood alleles from other organs back to the blood. The total number of cells and alleles within a given major clade are listed below. The clade contributions of all clades and subclades are presented in table S3. For heart subsamples, 'piece of heart' = a piece of heart tissue, 'DHCs' = dissociated unsorted cells; 'cardiomyocytes' = FACS-sorted GFP+ cardiomyocytes; and 'NCs' = non-cardiomyocyte heart cells. (C) and (E) Edits that define subclades of clade #1 (C) and clade #2 (E), with the total number of cells in which these are observed indicated on the right. A grey box indicates an unedited site or sites, distinguishing it from related alleles that contain an edit at this location. (D) and (F) Lineage trees corresponding to subclades of clade #1 (D) and clade #2 (F) that show how dependent edits are associated with increasing lineage restriction. The pie chart at each node indicates the organ distribution within a clade or subclade. Ratios of cell proportions are plotted, a normalization that accounts for differential depth of sampling between organs. Labels in the center of each pie chart correspond to the subclade labels in (C) and (E). Alleles present in a clade but not assigned to a descendent subclade (either they have no additional lineage restriction or are at low abundance) are not plotted for clarity. The number of cells (and the number of unique alleles) are also listed, and terminal nodes also list major organ restriction(s), i.e. those comprising >25% of a subclade by proportion.


A
Ancestral edits
Clade
Number of cells
Proportions of clades
0%
25%
50%
75%
100%
cells
alleles
other clades
B
Organ proportion by clade
12,733 cells
95 alleles
30,027 cells
66 alleles
59,773 cells
194 alleles
8,545 cells
38 alleles
19,445 cells
66 alleles
25,009 cells
18 alleles
15,626 cells
37 alleles
6,030 cells
49 alleles
Blood
Piece of heart
DHCs
NCs
Cardiomyocytes
Intestinal bulb
Post. intestine
Gills
Brain
Left eye
Right eye
C
Number of cells
Clade
E
Number of cells
Clade
D
Ancestral
Ancestral +1
Ancestral +2
Ancestral +3
12,733 (95)
5,457 (29)
3,319 (19)
3,939 (37)
1,088 (11)
4,138 (10)
Eyes
2,323 (7)
Heart
638 (1)
Gills
871 (2)
Gills, Brain, Eye
3,004 (30)
Heart
565 (1)
Heart
Intestine
441 (3)
Intestine
3,969 (1)
Cardiomyocytes
77 (2)
F
Ancestral
Ancestral +1
Ancestral +2
30,027 (66)
12,169 (13)
Heart
647 (4)
Cardiomyocytes
2,953 (1)
Heart
Intestine
7,956 (1)
Intestine
330 (1)
Gills
1,291 (1)
Gills
2,317 (1)
Brain
7,085 (3)

a combination of these mechanisms (Klein and Simons, 2011; Blanpain and Simons, 2013; Henson and Hume, 2006; Pellettieri and Sánchez Alvarado, 2007). Controlling the temporal and spatial induction of edits and isolating defined cell types from diverse organs should help resolve the mechanisms by which different embryonic progenitors come to dominate different adult organs.

Second, we show that GESTALT can inform the lineage relationships amongst thousands of differentiated cells. For example, following the accumulation of edits from ancestral to more complex reveals the progressive restriction of progenitors to germ layers and then organs. Cells within an organ can both share and differ in their alleles, revealing additional information about organ development. Future studies will need to determine whether such lineages reflect distinct cell fates (e.g., blood sub-lineages or neuronal subpopulations), because the anatomical resolution at which we queried alleles was restricted to grossly dissected organs and tissues. Because edited barcodes are expressed as RNA, we envision that combining our system with other platforms will permit much greater levels of anatomical resolution without sacrificing throughput. For example, *in situ* RNA sequencing of barcodes would provide explicit spatial and histological context to lineage reconstructions (Lee et al., 2014; Ke et al., 2013). Also, capturing richly informative lineage markers in single cell RNA-seq or ATAC-seq datasets may inform the interpretation of those molecular phenotypes, while also adding cell type resolution to studies of lineage (Satija et al., 2015; Cusanovich et al., 2015). Such integration may be particularly relevant to efforts to build comprehensive atlases of cell types. Because these single cell methods generate many reads per single cell, this would also facilitate using multiple, unlinked target arrays. In principle, the combined diversity of the barcodes queried from single cells could be engineered to uniquely identify every cell in a complex organism. In addition, orthogonal imaging-based lineage tracing approaches in fixed and live samples (e.g., Brainbow and related methods (Livet et al., 2007; Pan et al., 2013)) and longitudinal whole animal imaging approaches (Megason and Fraser, 2007; Liu and Keller, 2016) might be leveraged in parallel to validate and complement lineages resolved by GESTALT.

Although further work is required to optimize GESTALT towards enabling spatiotemporally complete maps of cell lineage, our proof-of-principle experiments show that using multiplex *in vivo* genome editing to record lineage information to a compact barcode at an organism-wide scale will be a powerful tool for developmental biology. This approach is not limited to normal development but can also be applied to animal models of developmental disorders, as well as to investigate the origins and progression of cancer. Our study also supports the notion that whereas its most widespread application has been to modify endogenous biological circuits, genome editing can also be used to stably record biological information (Farzadfard and Lu, 2014), analogous to recombinase-based memories but with considerably greater flexibility and scalability. For example, coupling editing activity to external stimuli or physiological changes could record the history of exposure to intrinsic or extrinsic signals. In the long term, we envision that rich, systematically generated maps of organismal development, wherein lineage, epigenetic, transcriptional and positional information are concurrently captured at single cell resolution, will advance our understanding of normal development, inherited diseases, and cancer.

3.12 Materials and Methods

3.12.1 Design of synthetic target arrays

Barcodes were designed as arrays of nine to twelve sense-oriented CRISPR/Cas9 target sites (23 bp, protospacer plus PAM sequences) separated by 3-5 bp linker sequences. Four initial designs (barcodes v1-v4) comprised of target sites for the sgRNA spacer sequence: 5'-GGCACTGCGGCTGGAGGTGG. The v1 barcode was comprised of ten targets arrayed in order of decreasing activity as measured with the GUIDE-seq assay performed in human cells (Tsai et al., 2014), starting with the target perfectly matching the sgRNA spacer sequence. The v2-v4 barcodes comprised of nine to ten non-overlapping target sets, all with activities less than half the perfectly matching target in the GUIDE-seq assay. To reduce repetitive subsequences within each barcode, protospacers were chosen such that no 8 bp

sequence was present in more than one protospacer within each barcode. After testing activities of targets in the v1-v4 barcodes in cell culture, the v5 barcode was designed to contain twelve targets that showed greater than 1% editing activity, including v1 targets 1-6, v3 target 1, v2 targets 1, 2 and 5, and v4 targets 1 and 3.

Two new barcodes, v6 and v7, were designed for use in zebrafish, each with ten CRISPR target sites not found in the D. rerio genome. Candidate target sequences were screened to remove any homopolymer runs, outside of the NGG of the protospacer, and were selected for editing activity [http://crispr.mit.edu]. The v6 and v7 barcodes were constructed as a series of 10 protospacer sequences meeting these criteria, with 4 bp linkers.

Each barcode was ordered as a gBlock (IDT) with ends compatible for In-Fusion cloning (ClonTech) into the 3' UTR of the EGFP gene in the lentiviral construct pLJM1-EGFP (Addgene #19319).The sequences of all barcodes (v1 through v7) are provided in table S4.

3.12.2 Generation of cell lines containing synthetic target arrays

To generate cell lines harboring single copies of barcodes, lentiviral particles were produced in HEK 293T cells transfected with lentivirus V2 packaging plasmids and barcode constructs. Viral supernatant harvested three days post transfection was used at low MOI to transduce 293T cells (MOI < 0.2). Successfully transduced cells were selected using puromycin (2 ug/ml), yielding polyclonal, barcode+ populations for barcodes v1-v5. Three monoclonal lines each harboring barcode v1 were generated by single-cell FACS, and used experimentally to compare editing rates across different integration sites. One of these was used as the parent line for cell culture lineages derived using barcode v1.

3.12.3 Editing of barcodes in cell lines

293T populations bearing barcodes v1-v5 were grown to 50-90% confluency in a 6-well dish. Cells were co-transfected using Lipofectamine 3000 (Life Technologies) according to protocol with 2g pX330-v1 and 0.5g pDsRed in a 6-well dish. One to three days post transfection, the cells were sorted on an Aria III FACS machine for DsRed fluorescence (as a marker

transfection). As indicated, either DsRed low, DsRed high, or total DsRed populations were sorted and cultured. At 7 days post-transfection, cells were harvested for gDNA preparation using the Qiagen DNeasy kit.

To stably deliver Cas9 and the sgRNA via lentivirus, the spacer sequence was cloned into the plasmid LentiCrispr v2 (Zhang lab, Addgene #52961) and virus was produced in 293T cells in the same manner described above. Wild-type 293T cells were transduced with pLenti-Crispr-V2-HMID.v1 and selected with puromycin, and then transduced with lentivirus bearing barcode v5. To impose a bottleneck, 200 GFP+ cells were sorted from this population and expanded under puromycin selection for two weeks prior to sampling gDNA.

3.12.4 Cell culture lineage experiments

Twelve lineages were established from a monoclonal barcode v1 293T cell line by transfecting cells as described above, and sorting single DsRed-low cells into a 96-well plate (DsRed low cells were used to limit Cas9 delivery and thus potential saturation of possible edits in this initial editing round). Cell sorting was performed seven days post-transfection, to reduce the likelihood that additional edits would arise after lineages were separated. Single cell-derived populations were expanded in culture for 3 weeks. A sample of cells from each lineage was pelleted and frozen. Next, each of the twelve lineages were transfected a second time, to induce another round of editing. Two 100-cell DsRed-low populations from each lineage were sorted 4-days post-transfection, and cultured to confluence in 96-well plates before harvesting gDNA.

Four additional monoclonal populations bearing v5 barcodes edited via transfection of pX330-v1 were also isolated by single-cell sorting. Re-editing of each population was achieved by two successive rounds of transfection with pX330-v1 (3 days apart). Cells were harvested for gDNA one week after the second transfection.

3.12.5 Barcode amplification and sequencing protocols

Kapa High Fidelity Polymerase was used for all barcode amplification steps. Gradient PCRs were performed to optimize annealing temperatures for amplification from gDNA. For experiments performed without UMIs, up to 250 ng of gDNA was loaded into a single 50 ul PCR reaction and amplified using primers immediately flanking the barcode (see table S4 for oligo sequences). If there was less than 250 ng from a sample, all of it was used in a single reaction. For experiments performed with UMIs, a primer with a sequencing adapter and 10 nt of fully degenerate sequence 5' to the barcode-flanking sequence was used for a single prolonged extension step, in which the temperature was ramped between annealing and extending for five cycles (without a denaturing step to prevent re-sampling of gDNA barcodes). All cell culture experiments and v6 zebrafish embryos received a single extension to incorporate UMIs, whereas v7 embryo time-course experiments and all ADR1 tissues (also v7) received 2 UMI incorporation cycles due to having low gDNA consequent to fewer cells being present in early embryo and sorted heart samples. To minimize repetitive amplification of the same barcode, no reverse primer was included in UMI-tagging reactions. DNA was then purified using AMPure beads (Agencourt), and loaded into a PCR primed from the sequencing adapter flanking each UMI and a site immediately 3' of the barcode.

For all experiments, two ensuing qPCRs were performed prior to sequencing to incorporate sequencing adapters, sample indexes, and flow cell adapters. AMPure beads were used to purify PCR products after each reaction.

Paired-end sequencing was performed on an Illumina MiSeq using 500- or 600-cycle kits for all cell culture experiments. Zebrafish experiments were sequenced on an Illumina NextSeq using 300-cycle kits. All sequencing generated adequate depth to sample each barcode present in a given sample to an average of greater than 10x coverage. To minimize contributions from sequencing error a read threshold was used for calling unique barcodes. This was conservatively set by dividing the number of reads from a sample by the number of expected barcode copies to be present in the amount of gDNA loaded into each PCR based

on the assumption that each cell contributed a single barcode.

Sequencing data for all samples was processed in a custom pipeline available on GitHub (https://github.com/shendurelab/Cas9FateMapping). Briefly, amplicon sequencing reads were first processed with the Trimmomatic software package to remove low quality bases (fig. 3.5A) (Bolger et al., 2014). The resulting reads were then grouped by their UMI tag. A raw read count threshold was set for each experiment based on sequencing depth, such that only UMIs observed in at least that many reads were analyzed to minimize contributions from sequencing error. For each UMI, a consensus sequence was called by jointly aligning all UMI-matched reads using the MAFFT (Rice et al., 2000) multiple sequence aligner. These reads were merged using the FLASH (Magoč and Salzberg, 2011)read merging tool, and both merged and unmerged reads were aligned to the amplicon reference using the NEEDLEALL (Rice et al., 2000) aligner with a gap open penalty of 10 and a gap extension penalty of 0.5. To capture read-through, UMI degenerate bases and adapter sequences were included in the reference amplicon sequence, and mismatches to Ns in the degenerate bases were set to a penalty of 0. To eliminate off-target sequencing reads, aligned sequences were required to match greater than 85% of bases at non-indel positions, to have correct PCR primer sequences on both the 5' and 3' ends, and to match at least 50 bases of the reference sequence (including primer sequences). Target sites were deemed edited if there was an insertion or deletion event present within 3 bases of the predicted Cas9 cut site (3 nucleotides 5' of each PAM), or if a deletion spanned the site entirely. Sites were marked as disrupted if there was not perfect alignment of the barcode over the entirety of the reference target sequence. An edited barcode was then defined as the complete list of insertion and deletion events (i.e. 'editing events') within the consensus sequence for a given UMI.

3.12.6 Maximum parsimony lineage reconstruction

For lineage reconstruction (fig. 3.5B), recurrently observed barcode alleles within a single organ or cell population were reduced to a single representative entry. We then used Camin-Sokal maximum parsimony to reconstruct lineages, as implemented in the PHYLIP Mix

software package (Felsenstein, 1989). Camin-Sokal maximum parsimony assumes that the initial cell or zygote is unedited, and that editing is irreversible. To run Mix, a matrix was created where each row corresponded to an allele, and each column corresponds to a unique editing event. Each entry in this matrix is an indicator variable of presence or absence of a specific edit in that allele (1 or 0). Events were also weighted by their log-abundance and scaled to the range allowed in Mix (0-Z). Mix was run with both the indicator data matrix as well as the weights file (selecting run options P, W, 4, and 5), and the output was parsed to recover the edit state of ancestral (internal) nodes. When Mix discovered multiple equal-scoring trees, we took the tree in the highest proportion. If two trees tied for highest proportion, we took the last highest scoring tree. To eliminate unsupported internal branching, we pruned internal parent-child nodes that had identical alleles. When a parent node and child node share the same allele, and neither node was a leaf, the grandchildren nodes were transferred to the parent and the child node was removed, creating multifurcating parent nodes. The resulting tree was converted to an annotated JSON tree compatible with our visualization tools. All code is available on the Shendure lab github website: https://github.com/shendurelab/Cas9FateMapping.

3.12.7 Zebrafish husbandry

All vertebrate animal work was performed at the facilities of Harvard University, Faculty of Arts & Sciences (HU/FAS). This study was approved by the Harvard University/Faculty of Arts & Sciences Standing Committee on the Use of Animals in Research & Teaching under Protocol No. 2508. The HU/FAS animal care and use program maintains full AAALAC accreditation, is assured with OLAW (A3593-01), and is currently registered with the USDA.

3.12.8 Cloning transgenesis vector

The transgenesis vectors pTol2-DRv6 and pTol2-DRv7 were constructed as follows. The v6 or v7 array was cloned into the 3' UTR of a DsRed coding sequence under control of the ubiquitin promoter (Mosimann et al., 2011). This cassette was placed in a Tol2

transgenesis vector containing a cmlc2:GFP marker, which drives expression of GFP in the cardiomyocytes of the heart from 24 hpf to adulthood (Huang et al., 2003). Plasmids are available from Addgene.

3.12.9 Generating transgenic zebrafish

To generate founder fish, 1-cell embryos were injected with zebrafish codon optimized Tol2 mRNA and pTol2-DR1v6 or pTol2-DR1v7 vector. Potential founder fish were screened for heart GFP expression at 30 hpf and grown to adulthood. Adult founder transgenic fish were identified by outcrossing to wild type and screening clutches of embryos for heart GFP expression at 30 hpf.

3.12.10 Transgene copy number quantification

To identify single copy Tol2 transgenics, copy number was quantified using qPCR (Pan et al., 2013). Briefly, genomic DNA was extracted from candidate embryos or fin-clips of adult fish using the HotSHOT method (53) and subjected to qPCR using a set of primers targeting DsRed and a set targeting a diploid conserved region of the genome (table S4) and compared to reference non-transgenic, 1-copy and 2-copy transgenic animals using the ddCt method.

3.12.11 Generation and delivery of editing reagents

sgRNAs specific to each site of the v6 or v7 array were generated as previously described (Gagnon et al., 2014), except that sgRNAs were isolated after transcription by column purification (Zymogen). 1-cell embryos resulting from an outcross of a transgenic founder were injected with two different volumes (0.5 nl, 1/3x or 1.5nl, 1x) of Cas9 protein (NEB) and sgRNAs in salt solution (8 mM Cas9, 100 ng/ul pooled sgRNAs, 50 mM KCl, 3 mM MgCl2, 5 mM Tris HCl pH 8.0, 0.05% phenol red). Transgenic embryos were collected at the time points indicated in the text and genomic DNA extracted as described below. To confirm editing, PCR was conducted on a subset of samples using primers flanking the v6 or v7 array (table

S4), and amplicons were loaded on a 2% agarose gel for electrophoresis.

3.12.12 Imaging

Embryos were anaesthetized and manually dechorionated in MS222, mounted in methylcellulose and imaged using a Leica upright fluorescence microscope.

3.12.13 Organ Dissection

Adult edited single copy transgenic fish were isolated without food for one day to reduce food particles in the gastrointestinal system, then anaesthetized in MS222 and euthanized on ice. Before dissection, blood was collected using a centrifugation method (Babaei et al., 2013). This collection method greatly enriches for blood cells, particularly red blood cells, but also results in contamination from skin or other tissues. The fish were pinned on a silicon mat and surgery was conducted using sterile tools to remove organs as in (Gupta and Mullins, 2010). Organs were washed in PBS and, with the exception of the heart, frozen in tubes on dry ice. A piece of heart tissue was collected before the remainder of the heart was dissociated following manufacturer's instructions (Miltenyi # 130-098-373). After dissociation, a sample of dissociated heart cells was collected (DHCs), and the remaining cells sorted using a Beckman Coulter MoFlo XDP Cell Sorter through a series of three gates to minimize debris and cell doublets, and then split into two additional populations: GFP+ cardiomyocytes and GFP- non-cardiomyocyte heart cells (NCs, fig. 3.15).

3.12.14 Genomic DNA preparation from zebrafish embryos and organs

Zebrafish embryo and adult organ gDNA was prepared using the Qiagen DNeasy kit. For heart samples from cell sorting experiments, 1 ul of poly-dT carrier DNA (25 uM) was added prior to gDNA preparation. Digestion with proteinase K at 56o C was performed overnight for intact organs (brain, eyes, gills, intestinal bulb, posterior intestine, and piece of heart) and for 30 minutes for blood samples, dissociated heart cells and embryos. gDNA was eluted

in 100 ul, then concentrated using an Eppendorf Vacufuge for samples yielding less than 1ng.

Experiment	Sample	Reference	UMI sample
2016_04_04_Fish_15_17	17_Brain	target1	TRUE
2016_04_04_Fish_15_17	17_Eye1	target1	TRUE
2016_04_04_Fish_15_17	17_Eye2	target1	TRUE
2016_04_04_Fish_15_17	17_Gills	target1	TRUE
2016_04_04_Fish_15_17	17_Intestine	target1	TRUE
2016_04_04_Fish_15_17	17_Upper_GI	target1	TRUE
2016_04_04_Fish_15_17	17_Blood	target1	TRUE
2016_04_04_Fish_15_17	17_Heart_chunk	target1	TRUE
2016_04_04_Fish_15_17	17_Heart_diss	target1	TRUE
2016_04_04_Fish_15_17	17_Heart_GFP-	target1	TRUE
2016_04_04_Fish_15_17	17_Heart_GFP+	target1	TRUE
2016_04_08_Adult_Fish_7_9_12	7B_Brain	target1	TRUE
2016_04_08_Adult_Fish_7_9_12	7B_Eye1	target1	TRUE
2016_04_08_Adult_Fish_7_9_12	7B_Eye2	target1	TRUE
2016_04_08_Adult_Fish_7_9_12	7B_Gills	target1	TRUE
2016_04_08_Adult_Fish_7_9_12	7B_Intestine	target1	TRUE
2016_04_08_Adult_Fish_7_9_12	7B_Upper_GI	target1	TRUE
2016_04_08_Adult_Fish_7_9_12	7B_Blood	target1	TRUE
2016_04_08_Adult_Fish_7_9_12	7B_Heart_chunk	target1	TRUE
2016_04_08_Adult_Fish_7_9_12	7B_Heart_diss	target1	TRUE
2016_04_08_Adult_Fish_7_9_12	7B_Heart_GFP-	target1	TRUE
2016_04_08_Adult_Fish_7_9_12	7B_Heart_GFP+	target1	TRUE
2016_04_08_Adult_Fish_7_9_12	7B_1_to_20_blood	target1	TRUE
2016_04_08_Adult_Fish_7_9_12	7B_1_to_100_blood	target1	TRUE
2016_04_08_Adult_Fish_7_9_12	7B_1_to_500_blood	target1	TRUE

Table 3.1: GESTALT barcode performance numbers for adult zebrafish

Experiment	Passing HMIDs	Unique HMIDs	Edited HMID prop.
2016_04_04_Fish_15_17	23840	699	1
2016_04_04_Fish_15_17	52984	524	1
2016_04_04_Fish_15_17	25060	111	1
2016_04_04_Fish_15_17	16099	318	1
2016_04_04_Fish_15_17	20973	182	1
2016_04_04_Fish_15_17	25005	332	1
2016_04_04_Fish_15_17	38202	11	1
2016_04_04_Fish_15_17	11063	54	0.999005695
2016_04_04_Fish_15_17	84	3	1
2016_04_04_Fish_15_17	3103	13	1
2016_04_04_Fish_15_17	1350	10	1
2016_04_08_Adult_Fish_7_9_12	33004	399	1
2016_04_08_Adult_Fish_7_9_12	17335	169	1
2016_04_08_Adult_Fish_7_9_12	21354	168	1
2016_04_08_Adult_Fish_7_9_12	28556	261	1
2016_04_08_Adult_Fish_7_9_12	15115	138	1
2016_04_08_Adult_Fish_7_9_12	25759	190	1
2016_04_08_Adult_Fish_7_9_12	44239	135	1
2016_04_08_Adult_Fish_7_9_12	5184	116	0.999614198
2016_04_08_Adult_Fish_7_9_12	776	51	1
2016_04_08_Adult_Fish_7_9_12	4994	76	1
2016_04_08_Adult_Fish_7_9_12	1145	44	1
2016_04_08_Adult_Fish_7_9_12	2656	24	1
2016_04_08_Adult_Fish_7_9_12	524	9	1
2016_04_08_Adult_Fish_7_9_12	78	5	1

Experiment	Mean Sites Edited	Mean Cut sites	Mean events
2016_04_04_Fish_15_17	7.584689597	5.365939597	4.285528523
2016_04_04_Fish_15_17	8.114430772	4.662709497	3.1089008
2016_04_04_Fish_15_17	8.323503591	4.293096568	2.915722267
2016_04_04_Fish_15_17	7.582458538	3.626374309	2.709422946
2016_04_04_Fish_15_17	8.480617937	3.423306156	2.427883469
2016_04_04_Fish_15_17	8.168126375	3.599320136	2.608478304
2016_04_04_Fish_15_17	8.529893723	3.471074813	2.470917753
2016_04_04_Fish_15_17	8.714272801	2.924342403	1.912229956
2016_04_04_Fish_15_17	8.857142857	3.214285714	2.214285714
2016_04_04_Fish_15_17	8.436351918	3.136641959	2.162101192
2016_04_04_Fish_15_17	4.043703704	2.865185185	2.688148148
2016_04_08_Adult_Fish_7_9_12	8.335080596	6.720397528	4.662768149
2016_04_08_Adult_Fish_7_9_12	8.043207384	6.737006057	5.226420536
2016_04_08_Adult_Fish_7_9_12	8.397302613	6.841153882	5.078205488
2016_04_08_Adult_Fish_7_9_12	8.089403278	6.469043283	5.11920437
2016_04_08_Adult_Fish_7_9_12	8.333972875	7.266887198	5.683493219
2016_04_08_Adult_Fish_7_9_12	8.179510074	6.386622151	4.92476416
2016_04_08_Adult_Fish_7_9_12	8.459029363	5.569994801	4.319650083
2016_04_08_Adult_Fish_7_9_12	8.735339506	7.035686728	5.018132716
2016_04_08_Adult_Fish_7_9_12	8.06443299	5.62371134	4.395618557
2016_04_08_Adult_Fish_7_9_12	8.10432519	6.088706448	4.7917501
2016_04_08_Adult_Fish_7_9_12	8.219213974	5.946724891	4.513537118
2016_04_08_Adult_Fish_7_9_12	8.458584337	5.640060241	4.366340361
2016_04_08_Adult_Fish_7_9_12	8.505725191	5.610687023	4.311068702
2016_04_08_Adult_Fish_7_9_12	8.615384615	5.756410256	4.384615385

Experiment	Mean intact sites	Editing for target 1	Unique events target1
2016_04_04_Fish_15_17	1.800377517	0.969253356	113
2016_04_04_Fish_15_17	1.711271327	0.971179979	127
2016_04_04_Fish_15_17	1.588347965	0.983918595	43
2016_04_04_Fish_15_17	1.637865706	0.997266911	51
2016_04_04_Fish_15_17	1.295665856	0.996328613	36
2016_04_04_Fish_15_17	1.304579084	0.991761648	60
2016_04_04_Fish_15_17	1.231401497	1	7
2016_04_04_Fish_15_17	0.832504746	0.994124559	34
2016_04_04_Fish_15_17	0.642857143	1	3
2016_04_04_Fish_15_17	0.986464712	1	10
2016_04_04_Fish_15_17	2.493333333	0.999259259	5
2016_04_08_Adult_Fish_7_9_12	1.302054296	0.998333535	86
2016_04_08_Adult_Fish_7_9_12	1.534179406	0.999307759	51
2016_04_08_Adult_Fish_7_9_12	1.135150323	0.999859511	34
2016_04_08_Adult_Fish_7_9_12	1.52570388	0.990649951	61
2016_04_08_Adult_Fish_7_9_12	1.207542177	0.999206087	41
2016_04_08_Adult_Fish_7_9_12	1.574750573	0.985519624	52
2016_04_08_Adult_Fish_7_9_12	1.072628224	0.999841769	22
2016_04_08_Adult_Fish_7_9_12	1.092399691	0.998842593	41
2016_04_08_Adult_Fish_7_9_12	1.576030928	0.99871134	22
2016_04_08_Adult_Fish_7_9_12	1.508009612	1	27
2016_04_08_Adult_Fish_7_9_12	1.689082969	1	19
2016_04_08_Adult_Fish_7_9_12	1.109563253	0.999246988	9
2016_04_08_Adult_Fish_7_9_12	1.057251908	1	7
2016_04_08_Adult_Fish_7_9_12	1.012820513	1	5

Experiment	Intact prop target 1	Editing for target 2	Unique events target2
2016_04_04_Fish_15_17	0.023112416	0.516107383	74
2016_04_04_Fish_15_17	0.006077306	0.678016005	62
2016_04_04_Fish_15_17	0.015921788	0.91292897	41
2016_04_04_Fish_15_17	0.001552891	0.536989875	28
2016_04_04_Fish_15_17	0.001192009	0.76488819	35
2016_04_04_Fish_15_17	0.006718656	0.648430314	33
2016_04_04_Fish_15_17	0	0.765012303	4
2016_04_04_Fish_15_17	0.005875441	0.867034258	31
2016_04_04_Fish_15_17	0	1	3
2016_04_04_Fish_15_17	0	0.777634547	8
2016_04_04_Fish_15_17	0.000740741	0.166666667	3
2016_04_08_Adult_Fish_7_9_12	0.001181675	0.839898194	114
2016_04_08_Adult_Fish_7_9_12	0.000692241	0.791173926	67
2016_04_08_Adult_Fish_7_9_12	0.000140489	0.940760513	58
2016_04_08_Adult_Fish_7_9_12	0.009280011	0.697086427	69
2016_04_08_Adult_Fish_7_9_12	0.000661594	0.914918955	48
2016_04_08_Adult_Fish_7_9_12	0.014363912	0.738266237	61
2016_04_08_Adult_Fish_7_9_12	0.000158231	0.998756753	46
2016_04_08_Adult_Fish_7_9_12	0.000964506	0.863040123	57
2016_04_08_Adult_Fish_7_9_12	0	0.791237113	28
2016_04_08_Adult_Fish_7_9_12	0	0.873448138	33
2016_04_08_Adult_Fish_7_9_12	0	0.613100437	23
2016_04_08_Adult_Fish_7_9_12	0.000753012	0.998870482	10
2016_04_08_Adult_Fish_7_9_12	0	1	8
2016_04_08_Adult_Fish_7_9_12	0	1	5

Experiment	Intact prop target 2	Editing for target 3	Unique events target3
2016_04_04_Fish_15_17	0.260360738	0.37420302	53
2016_04_04_Fish_15_17	0.261626151	0.66918315	72
2016_04_04_Fish_15_17	0.016999202	0.915482841	41
2016_04_04_Fish_15_17	0.156779924	0.51158457	17
2016_04_04_Fish_15_17	0.064320793	0.829495065	45
2016_04_04_Fish_15_17	0.024915017	0.654389122	37
2016_04_04_Fish_15_17	0.000104707	0.76495995	4
2016_04_04_Fish_15_17	0.007864051	0.868570912	28
2016_04_04_Fish_15_17	0	0.857142857	2
2016_04_04_Fish_15_17	0	0.775378666	6
2016_04_04_Fish_15_17	0.000740741	0.166666667	3
2016_04_08_Adult_Fish_7_9_12	0.154769119	0.564234638	54
2016_04_08_Adult_Fish_7_9_12	0.045399481	0.357023363	35
2016_04_08_Adult_Fish_7_9_12	0.048702819	0.617214573	35
2016_04_08_Adult_Fish_7_9_12	0.281446981	0.34448102	37
2016_04_08_Adult_Fish_7_9_12	0.083824016	0.128944757	28
2016_04_08_Adult_Fish_7_9_12	0.255522342	0.283784308	23
2016_04_08_Adult_Fish_7_9_12	0.001085015	0.503085513	17
2016_04_08_Adult_Fish_7_9_12	0.126157407	0.504822531	34
2016_04_08_Adult_Fish_7_9_12	0.170103093	0.585051546	17
2016_04_08_Adult_Fish_7_9_12	0.122547056	0.488386063	19
2016_04_08_Adult_Fish_7_9_12	0.384279476	0.379912664	13
2016_04_08_Adult_Fish_7_9_12	0.001129518	0.510542169	7
2016_04_08_Adult_Fish_7_9_12	0	0.528625954	4
2016_04_08_Adult_Fish_7_9_12	0	0.5	2

Experiment	Intact prop target 3	Editing for target 4	Unique events target4
2016_04_04_Fish_15_17	0.622315436	0.948238255	151
2016_04_04_Fish_15_17	0.328004681	0.968141326	124
2016_04_04_Fish_15_17	0.083838787	0.999441341	49
2016_04_04_Fish_15_17	0.485309647	0.888999317	53
2016_04_04_Fish_15_17	0.167691794	0.999570877	81
2016_04_04_Fish_15_17	0.342531494	0.999840032	104
2016_04_04_Fish_15_17	0.233914455	1	5
2016_04_04_Fish_15_17	0.130073217	0.999005695	38
2016_04_04_Fish_15_17	0.142857143	1	3
2016_04_04_Fish_15_17	0.222365453	0.993232356	7
2016_04_04_Fish_15_17	0.832592593	1	4
2016_04_08_Adult_Fish_7_9_12	0.411041086	0.975639316	74
2016_04_08_Adult_Fish_7_9_12	0.533141044	0.990423998	50
2016_04_08_Adult_Fish_7_9_12	0.35042615	0.998079985	51
2016_04_08_Adult_Fish_7_9_12	0.597142457	0.991035159	70
2016_04_08_Adult_Fish_7_9_12	0.542441283	0.981276877	42
2016_04_08_Adult_Fish_7_9_12	0.644357312	0.976202492	44
2016_04_08_Adult_Fish_7_9_12	0.284952192	0.999773955	39
2016_04_08_Adult_Fish_7_9_12	0.45158179	0.999035494	46
2016_04_08_Adult_Fish_7_9_12	0.355670103	0.858247423	23
2016_04_08_Adult_Fish_7_9_12	0.43071686	0.871245495	30
2016_04_08_Adult_Fish_7_9_12	0.617467249	0.951091703	18
2016_04_08_Adult_Fish_7_9_12	0.300828313	1	13
2016_04_08_Adult_Fish_7_9_12	0.265267176	1	7
2016_04_08_Adult_Fish_7_9_12	0.256410256	1	4

Experiment	Intact prop target 4	Editing for target 5	Unique events target5
2016_04_04_Fish_15_17	0.003271812	0.98317953	148
2016_04_04_Fish_15_17	7.55E-05	0.995055111	117
2016_04_04_Fish_15_17	0	0.999281724	47
2016_04_04_Fish_15_17	0.108081247	0.936269333	47
2016_04_04_Fish_15_17	0	0.99723454	70
2016_04_04_Fish_15_17	0	0.978804239	61
2016_04_04_Fish_15_17	0	1	5
2016_04_04_Fish_15_17	0.000994305	0.996565127	31
2016_04_04_Fish_15_17	0	1	3
2016_04_04_Fish_15_17	0.006767644	0.974540767	8
2016_04_04_Fish_15_17	0	0.177037037	4
2016_04_08_Adult_Fish_7_9_12	0.004181311	0.855381166	57
2016_04_08_Adult_Fish_7_9_12	0.001153735	0.993654456	36
2016_04_08_Adult_Fish_7_9_12	0	0.923480378	49
2016_04_08_Adult_Fish_7_9_12	3.50E-05	0.916795069	61
2016_04_08_Adult_Fish_7_9_12	0	0.981872312	44
2016_04_08_Adult_Fish_7_9_12	0	0.997593074	51
2016_04_08_Adult_Fish_7_9_12	0	0.999773955	35
2016_04_08_Adult_Fish_7_9_12	0.000385802	0.998842593	39
2016_04_08_Adult_Fish_7_9_12	0	0.993556701	20
2016_04_08_Adult_Fish_7_9_12	0	0.979975971	29
2016_04_08_Adult_Fish_7_9_12	0	0.998253275	19
2016_04_08_Adult_Fish_7_9_12	0	1	13
2016_04_08_Adult_Fish_7_9_12	0	1	5
2016_04_08_Adult_Fish_7_9_12	0	1	3

Experiment	Intact prop target 5	Editing for target 6	Unique events target6
2016_04_04_Fish_15_17	8.39E-05	0.794840604	142
2016_04_04_Fish_15_17	0	0.99588555	128
2016_04_04_Fish_15_17	0	0.998762969	50
2016_04_04_Fish_15_17	0	0.890117399	75
2016_04_04_Fish_15_17	0	0.99589949	50
2016_04_04_Fish_15_17	0	0.956808638	65
2016_04_04_Fish_15_17	0	1	5
2016_04_04_Fish_15_17	0	0.996836301	33
2016_04_04_Fish_15_17	0	1	3
2016_04_04_Fish_15_17	0	0.974218498	8
2016_04_04_Fish_15_17	0	0.177037037	4
2016_04_08_Adult_Fish_7_9_12	0	0.980729609	72
2016_04_08_Adult_Fish_7_9_12	0	0.999711566	39
2016_04_08_Adult_Fish_7_9_12	0	0.999484874	53
2016_04_08_Adult_Fish_7_9_12	0	0.999614792	65
2016_04_08_Adult_Fish_7_9_12	0	0.983526298	41
2016_04_08_Adult_Fish_7_9_12	0	0.998563609	50
2016_04_08_Adult_Fish_7_9_12	0	0.999954791	38
2016_04_08_Adult_Fish_7_9_12	0.000385802	0.999421296	37
2016_04_08_Adult_Fish_7_9_12	0	1	22
2016_04_08_Adult_Fish_7_9_12	0	0.999599519	37
2016_04_08_Adult_Fish_7_9_12	0	0.999126638	22
2016_04_08_Adult_Fish_7_9_12	0	1	16
2016_04_08_Adult_Fish_7_9_12	0	1	6
2016_04_08_Adult_Fish_7_9_12	0	1	3

Experiment	Intact prop target 6	Editing for target 7	Unique events target7
2016_04_04_Fish_15_17	0.000209732	0.983473154	240
2016_04_04_Fish_15_17	0	0.989676129	159
2016_04_04_Fish_15_17	0	0.99557063	47
2016_04_04_Fish_15_17	6.21E-05	0.829554631	99
2016_04_04_Fish_15_17	0	0.969007772	74
2016_04_04_Fish_15_17	8.00E-05	0.95404919	95
2016_04_04_Fish_15_17	0	0.99992147	4
2016_04_04_Fish_15_17	0.000271174	0.992768688	31
2016_04_04_Fish_15_17	0	1	3
2016_04_04_Fish_15_17	0	0.967773123	8
2016_04_04_Fish_15_17	0	0.18	5
2016_04_08_Adult_Fish_7_9_12	0	0.972972973	71
2016_04_08_Adult_Fish_7_9_12	0	0.998961638	40
2016_04_08_Adult_Fish_7_9_12	0	0.876603915	50
2016_04_08_Adult_Fish_7_9_12	0	0.998634263	67
2016_04_08_Adult_Fish_7_9_12	0	0.999933841	45
2016_04_08_Adult_Fish_7_9_12	0	1	50
2016_04_08_Adult_Fish_7_9_12	0	0.999954791	36
2016_04_08_Adult_Fish_7_9_12	0.000192901	0.999614198	39
2016_04_08_Adult_Fish_7_9_12	0	1	22
2016_04_08_Adult_Fish_7_9_12	0.00020024	0.998398078	33
2016_04_08_Adult_Fish_7_9_12	0	1	22
2016_04_08_Adult_Fish_7_9_12	0	1	15
2016_04_08_Adult_Fish_7_9_12	0	1	6
2016_04_08_Adult_Fish_7_9_12	0	1	3

Experiment	Intact prop target 7	Editing for target 8	Unique events target8
2016_04_04_Fish_15_17	4.19E-05	0.7625	105
2016_04_04_Fish_15_17	0	0.413917409	63
2016_04_04_Fish_15_17	7.98E-05	0.34792498	31
2016_04_04_Fish_15_17	0	0.852537425	85
2016_04_04_Fish_15_17	0.028131407	0.927239784	17
2016_04_04_Fish_15_17	0.000719856	0.95204959	55
2016_04_04_Fish_15_17	0	0.99992147	4
2016_04_04_Fish_15_17	0.000994305	0.989243424	27
2016_04_04_Fish_15_17	0	1	3
2016_04_04_Fish_15_17	0	0.967450854	6
2016_04_04_Fish_15_17	0	0.177037037	4
2016_04_08_Adult_Fish_7_9_12	0	0.633559569	50
2016_04_08_Adult_Fish_7_9_12	0	0.665820594	25
2016_04_08_Adult_Fish_7_9_12	0	0.640207924	31
2016_04_08_Adult_Fish_7_9_12	0	0.711829388	45
2016_04_08_Adult_Fish_7_9_12	0	0.619649355	34
2016_04_08_Adult_Fish_7_9_12	0	0.795915991	38
2016_04_08_Adult_Fish_7_9_12	0	0.745609078	30
2016_04_08_Adult_Fish_7_9_12	0.000385802	0.663580247	32
2016_04_08_Adult_Fish_7_9_12	0	0.62757732	14
2016_04_08_Adult_Fish_7_9_12	0	0.665398478	26
2016_04_08_Adult_Fish_7_9_12	0	0.777292576	18
2016_04_08_Adult_Fish_7_9_12	0	0.761295181	9
2016_04_08_Adult_Fish_7_9_12	0	0.769083969	3
2016_04_08_Adult_Fish_7_9_12	0	0.871794872	2

Experiment	Intact prop target 8	Editing for target 9	Unique events target9
2016_04_04_Fish_15_17	0.230369128	0.969001678	157
2016_04_04_Fish_15_17	0.583006191	0.981182999	174
2016_04_04_Fish_15_17	0.650079808	0.994892259	54
2016_04_04_Fish_15_17	0.143611404	0.996148829	120
2016_04_04_Fish_15_17	0.067276975	0.997806704	33
2016_04_04_Fish_15_17	0.04435113	0.997480504	80
2016_04_04_Fish_15_17	7.85E-05	1	5
2016_04_04_Fish_15_17	0.009852662	0.997649824	34
2016_04_04_Fish_15_17	0	1	3
2016_04_04_Fish_15_17	0.032549146	0.999677731	8
2016_04_04_Fish_15_17	0.817777778	1	7
2016_04_08_Adult_Fish_7_9_12	0.326839171	0.998818325	100
2016_04_08_Adult_Fish_7_9_12	0.327314681	0.999077012	50
2016_04_08_Adult_Fish_7_9_12	0.333380163	0.999250726	59
2016_04_08_Adult_Fish_7_9_12	0.230144278	1	71
2016_04_08_Adult_Fish_7_9_12	0.378167383	0.994773404	49
2016_04_08_Adult_Fish_7_9_12	0.203734617	0.999262394	48
2016_04_08_Adult_Fish_7_9_12	0.253803205	1	35
2016_04_08_Adult_Fish_7_9_12	0.333719136	0.999421296	50
2016_04_08_Adult_Fish_7_9_12	0.356958763	1	20
2016_04_08_Adult_Fish_7_9_12	0.329195034	1	34
2016_04_08_Adult_Fish_7_9_12	0.221834061	0.998253275	21
2016_04_08_Adult_Fish_7_9_12	0.236822289	1	11
2016_04_08_Adult_Fish_7_9_12	0.229007634	1	5
2016_04_08_Adult_Fish_7_9_12	0.128205128	1	3

Experiment	Intact prop target 9	Editing for target 10	Unique events target10
2016_04_04_Fish_15_17	0.021770134	0.283892617	58
2016_04_04_Fish_15_17	0.018118677	0.452193115	66
2016_04_04_Fish_15_17	0.004748603	0.175299282	15
2016_04_04_Fish_15_17	0.002174048	0.142990248	9
2016_04_04_Fish_15_17	0.001573452	0.003146903	5
2016_04_04_Fish_15_17	0.000319936	0.034513097	10
2016_04_04_Fish_15_17	0	7.85E-05	1
2016_04_04_Fish_15_17	0.001627045	0.012474012	6
2016_04_04_Fish_15_17	0	0	0
2016_04_04_Fish_15_17	0.000322269	0.006445375	1
2016_04_04_Fish_15_17	0	0	0
2016_04_08_Adult_Fish_7_9_12	0.000242395	0.515513271	35
2016_04_08_Adult_Fish_7_9_12	0	0.248053072	16
2016_04_08_Adult_Fish_7_9_12	0.000655615	0.402360214	20
2016_04_08_Adult_Fish_7_9_12	0	0.43927721	28
2016_04_08_Adult_Fish_7_9_12	0.002117102	0.729870989	22
2016_04_08_Adult_Fish_7_9_12	0	0.404402345	23
2016_04_08_Adult_Fish_7_9_12	0	0.212278759	6
2016_04_08_Adult_Fish_7_9_12	0.000385802	0.708719136	22
2016_04_08_Adult_Fish_7_9_12	0	0.210051546	9
2016_04_08_Adult_Fish_7_9_12	0	0.227873448	13
2016_04_08_Adult_Fish_7_9_12	0	0.502183406	10
2016_04_08_Adult_Fish_7_9_12	0	0.188629518	4
2016_04_08_Adult_Fish_7_9_12	0	0.208015267	2
2016_04_08_Adult_Fish_7_9_12	0	0.243589744	1

Chapter 4

CONCLUDING THOUGHTS

In this chapter I present my thoughts on the future of whole-organism lineage tracing, and suggest directions for follow-on work. There are many opportunities to increase the information capacity of GESTALT and related technologies to more aptly match the requirements of lineage tracing in mammalian systems. Some of these improvements are straightforward and can be implemented immediately. Others are more far reaching, and as we have learned in the development of GESTALT, may require rounds of *in vivo* optimization. In the short term, the goal is to increase the information content and sensitivity of the system to more closely match the number of cell division cycles in complex organisms. In a longer view, such improvements in lineage tracing will be coupled with nascent single-cell technologies to sample from the many cellular dimensions. Many experimental, technical, and computational challenges will be faced, but it's reasonable to believe we'll be capable of recording fine-grained and richly annotated temporal maps of mammalian development in the foreseeable future.

4.1 Near-term improvements to the GESTALT technology

The proof of concept system presented in the last chapters has many obvious shortcomings. The most obvious of which is the reliance on zygotic injection to deliver the Cas9 protein and associated guide sequences. In the zebrafish zygote, large enough to be visible with the naked eye, direct injections are a routine way to deliver various perturbations (Rosen et al., 2009). In other model organisms this is either a more complicated endeavour or is outright impossible due to biological or technical constraints. It's also reasonable to assume our injection strategy introduces more variability into editing outcomes: from slight changes

in reagent concentrations, imprecision in the injection machinery, or even the caffeine level of the operator. All these make reproducibility more challenging. It would be advantageous if the whole GESTALT system was integrated into the genome of interest to both reduce this variation and to eventually tie expression into the regulatory landscape of the cell.

The first hurdle to a fully endogenous system is introducing the expression of many individual guides into the genome. In our zebrafish system each target sequence had a perfectly matched guide RNA – ten guides for ten targets. We did this to hopefully maximize diversity and editing efficiency of the barcode. When moving to an endogenous system our choice is either to use multiple integrations of expressed targets, or a single transcript that is processed into individual guides sequences. The first approach would require expression of all guides individually off of their own polymerase III promoter such as U6 or H1 (Dumay-Odelot et al., 2010). Although different variations of these promoters exist, their repetitive nature complicates cloning, or if done as independent genome integrations they would require substantial breeding to complete the crosses into a single animal.

Instead, we've had some initial success adapting a single expressed transcript approach developed by the drosophila community (Port and Bullock, 2016). Here Cas9 (or any gene of interest) is expressed by a traditional polymerase II promoter system. A series of tRNA sequences separate individual guides within the 3' UTR of that gene before any polyadenylation signal sequence. When the transcript is expressed, individual tRNA sequences are cleaved by endogenous splicing machinery and the guide sequences are then free to associate with the Cas9 protein. To save space, the common RNA backbone shared by all guides is expressed elsewhere. Such a system would allow for single transgenic lines to be made with a large set of individual guides. The full details are beyond the scope of this perspective, but initial efforts in cell culture demonstrate such a system can be effective (fig. 4.1), and would ameliorate many of the challenges of switching to new model organisms.

In conjunction with the integration of the guides into a single construct, a clear practical improvement to GESTALT is the expansion of the target cassette size. Our initial constructs had either 9, 10 or 12 target sequences. We feared that we would be confounded by a

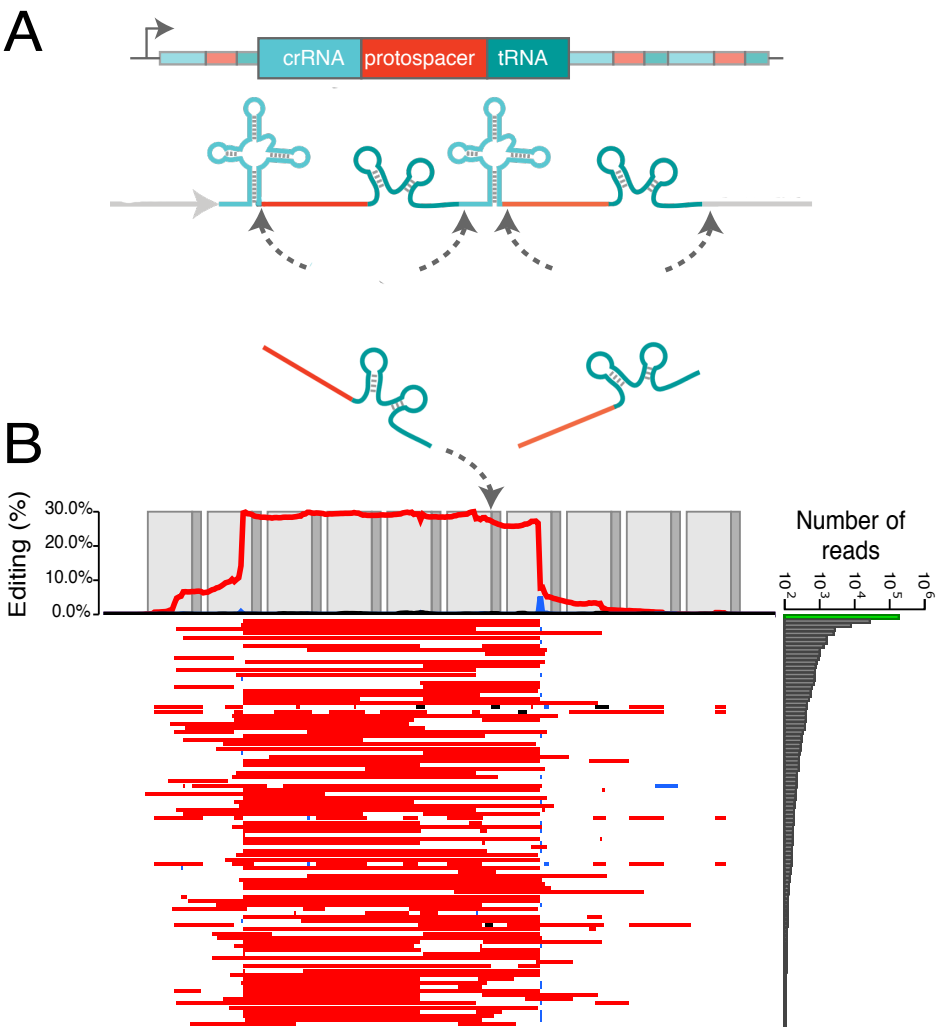


Figure 4.1: Expression of tRNA-guide arrays to drive multisite editing. **(A)** An array of tRNAs separating target-specific protospacers and crRNA fragments. This array is placed in the 3' end of a transcribed marker gene such as GFP. When expressed, endogenous tRNA splicing machinery frees the guide sequences to complex with Cas9 and the RNA 'tracer' backbone (expressed elsewhere). **(B)** tRNA array editing of a V7 GESTALT target in human 293T cells, with high rates of editing across the barcode. An array containing guides corresponding to each target was expressed. Due to construction issues sites 1, 8, 9, and 10 weren't properly loaded into a Cas9 complex.

large number of insertions or deletions and the lower read quality typical of longer Illumina sequencing. By limiting the cassette to a maximum of 300 basepairs, we covered the majority of the barcode redundantly with the forward and reverse reads. In the end we didn't need this redundancy, and we could easily increase the barcode capacity to 15 targets without significantly changing the approach. We could also take advantage of technology developed in the viral lineage tracing approaches and add a static DNA tag to the front of the barcode cassette. Single cells would then have multiple barcodes recording simultaneously, each with it's own unique identifiable tag. This approach could also be used in single-cell RNA sequencing, where all transcripts can be linked by a unifying cell ID, an advantage we didn't have in the previous paper.

4.2 Computational improvements

There are clear next-step computational improvements to be made. Alignment of the barcodes to the reference sequence has been challenging due the dynamic nature of editing. Simple dynamic programming methods such as Needleman-Wunsch guarantee the optimal solution for a given set of parameters (Needleman and Wunsch, 1970), but without prior information these parameters are assumed to be uniform across the alignment (for instance the cost of beginning a deletion is equal at every base). For GESTALT it's clear that such a flat prior is incorrect: we know where the double strand break will occur, and any insertion or deletion should either include or overlap this site. Hidden Markov Models or Dynamic Bayesian Networks could be used to both learn these site specific parameters and improve alignments (Eddy, 1998).

Such parameterized probabilistic models could then be leveraged to improve our phylogenetic reconstruction methods. In the initial work, we used the Camin-Sokal maximum parsimony method, borrowed from phylogenetics, to reconstruct the relationship between our recovered barcodes (Felsenstein, 1989). Although this model was appropriate, improvements could be made by adopting maximum likelihood approaches to further refine the recovered lineages. Likelihood methods also allow us to infer both branch lengths and assign con-

fidence intervals to various aspects of our tree. Such methods are more compatible with techniques that combine information for multiple samplings of the same species, such as tree graphs (Smith et al., 2013).

4.3 Moving into additional model organisms

The zebrafish is widely used to characterize development, and progress has been made towards zebrafish as a cancer model (White et al., 2013). Zebrafish also have the unique ability to regenerate critical organs, such as the heart, an interest to medical science. Although there is much to learn in this system, we should not preclude the use of other organisms. Other like *Mus musculus*, the house mouse, are the canonical systems for various disease states and developmental milestones. Also as mammals, they serve as our most familiar model for human biology. Additionally many resources are available in mice and other model organisms such as *Drosophila melanogaster* (fly) that are not as developed in zebrafish, including more refined reference sequences, and in the case of the fly, much shorter generation times.

In this vein we've made progress incorporating GESTALT into both mouse and fly systems. We've incorporated our GESTALT V5 system into the fly, where one guide targets a series of off-target sites. In such an approach the barcode system is integrated into one fly line, the expression of Cas9 and the guide into another, and the two lines are crossed to activate lineage tracing. Initial experiments show that we've successfully integrated our target, expressed the editing components, and recovered GESTALT barcodes. Interestingly, the editing patterns reflect more activity than we saw using the same target in cell culture systems, further stressing the need to test and validate such systems *in vivo* (fig. 4.2).

Such a demonstration is a stepping stone into a much larger effort to characterize global developmental patterns in flies. We have the advantages of an enormous wealth of experimental knowledge of both early and late fly development, as well as emerging single-cell efforts, within our lab and elsewhere. With the short generation times and amenable genome integration systems, flies can serve as a testbed for *in vivo* development of improved GESTALT systems. We will also release this GESTALT fly line, and hopefully this rapid prototyping

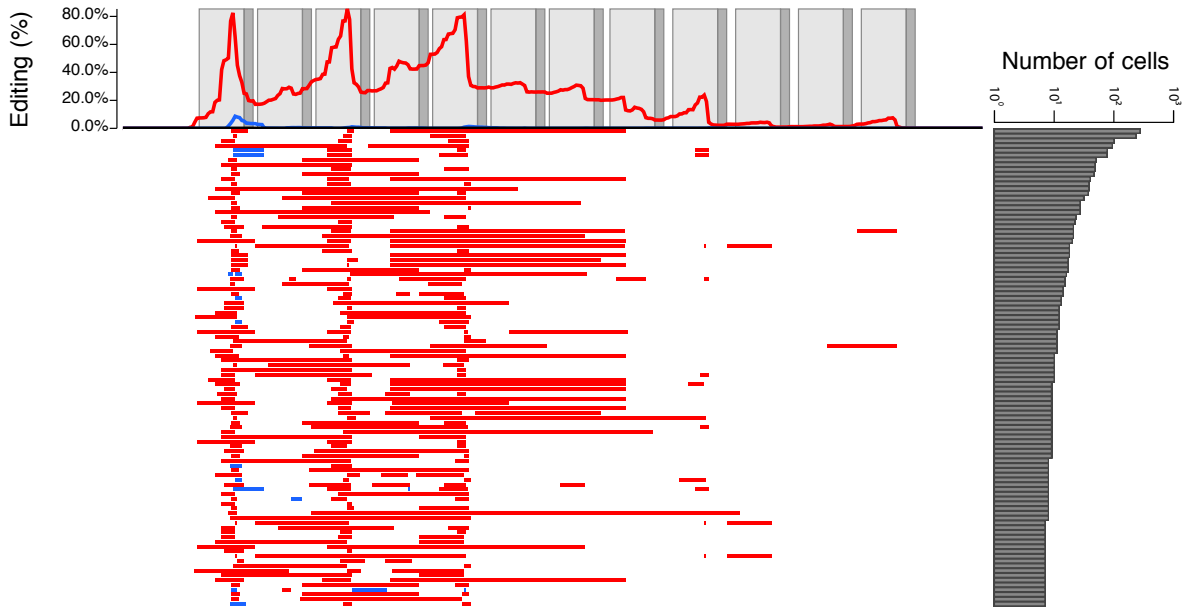


Figure 4.2: GESTALT lineage recording in *Drosophila melanogaster*. The V5 barcode from (McKenna et al., 2016), CRISPR/Cas9, and the corresponding guide sequence were introduced to the germline of *Drosophila melanogaster*. Adult flies were harvested, V5 integrants tagged with UMIs, and sequenced. We see increased editing when compared to human cell culture lines containing the same V5 construct. Cell division rates, cell size, differences in repair enzymes, and other endogenous factors may affect the overall editing efficiency.

and technology advancement will enable new discoveries in the *Drosophila* field.

In the end, the most desirable organism for GESTALT integration is the mouse. In mice we have access to a large number of disease models, including cancer; mice also serve as the foremost testbed for experimental discoveries in mammalian development. The mouse model also has the advantage of established techniques and protocols such as *in situ* sequencing as well as practical considerations like well trodden antibodies and safe-harbor loci for genomic integration. We've taken initial steps towards such a mouse model. As an initial testbed for GESTALT in mice, we've generated a mouse line that contains the same target cassette used in our zebrafish experiments. In mice this is spliced into the ROSA26 transcript, which is ubiquitously expressed across tissues (Maryann Giel-Moloney, 2007). Using 3T3 mouse cell culture we were able to demonstrate effective capture of GESTALT barcode as expressed RNA molecules (fig. 4.3). In the short-term our experimental plan will be to inject the pronucleus of these mice with Cas9 and the guides, in a similar fashion to our zebrafish experiments. In an independent line of work, we're establishing all components of the GESTALT system into the mouse genome, with each component either tied to an inducible promoter, or only activated on exposure to site-specific recombinases. These gates will allow for a more fine-grained control of the GESTALT components, and will pave the way for efforts targeting the lineage of specific tissues, by breeding to other community developed mouse models.

4.4 Enhancements to our knowledge of genome editing outcomes

GESTALT relies upon the activation of endogenous repair pathways to resolve double-stranded breaks in an error-prone fashion, resulting in the measured insertions and deletions (indels) we use to infer lineage. There is some discussion in the field around the specifics of this repair process (Thyme and Schier, 2016), but there are clear indications that local sequence context has a large role in determining the type, size, and location of the repair outcome (van Overbeek et al., 2016). Although experimental results are limited, a large-scale screen of guides and their associated target sequences would determine the importance of

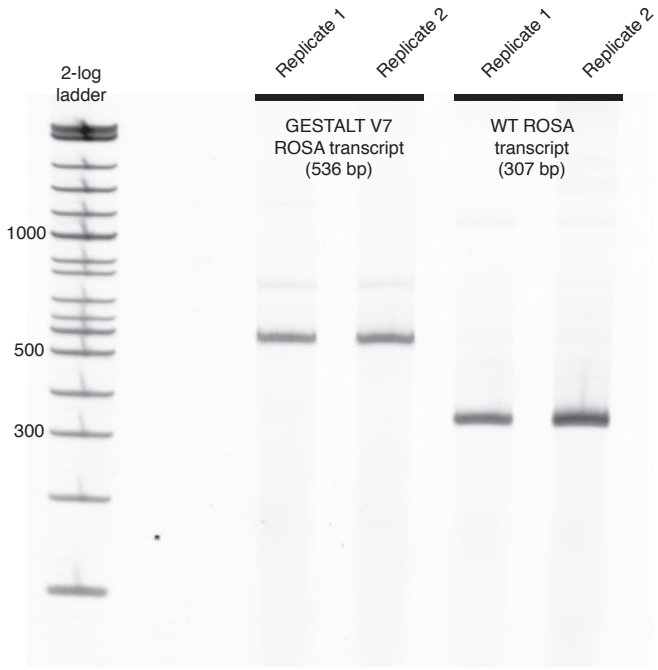


Figure 4.3: Endogenous expression of GESTALT constructs in mouse cell lines. To generate a stably-integrated GESTALT mouse, we've created a construct with the V7 target region flanked by sequence homologous to an intron of the ROSA26 locus. This V7 target plasmid contains a splice acceptor and a polyadenylation signal sequence, and is spliced into and terminates the endogenous transcript. Here 3T3 mouse cells were co-transfected with the V7-ROSA plasmid and a CRISPR integration plasmid to cut the ROSA26 site. RNA was reverse transcribed, and amplified with RNA specific primers for the ROSA-V7 transcript (left) and the native transcript (right).

such factors, and will allow us to potentially predict Cas9 editing outcomes from sequence alone.

Such a predictive framework would also allow us to choose target sequences that repair into diverse outcomes, maximizing the information content of our barcode. Conversely, approaches such as *in situ* sequencing would benefit from the opposite: target sequences that resolve into one and only one indel outcome, greatly simplifying the recovery of the barcode status with hybridization probes. This knowledge has applications outside of lineage tracing. Enormous resources and effort are being directed at therapeutic applications of CRISPR. If you could predict the predominant indel outcome from the sequence context alone, gene targets could be selected that produce out-of-frame indels for gene knockouts, avoid long deletions that could obscure detection by PCR, or even generate in-frame deletions for gene tailoring *in situ*.

4.5 Recording cellular perturbations

This chapter so far has addressed improvements to the underlying technology of GESTALT; be it larger barcodes, more active editing, or even the development of target sequences with deterministic indel profiles. These developments will help improve the resolution of the lineage recording, but there is a great opportunity to capture permanent DNA memories of other transient cell signals or states. Both Frieda et al. and Perli et al. have prototyped such an approach, recording Wnt signaling and LPS stimulation using their respective promoter sequences (Frieda et al., 2017; Perli et al., 2016). Given GESTALT's DNA sequencing read-out, an interesting path forward would be to combine lineage information from GESTALT barcodes with the self-targeting guide system of Perli et al. Individual self-targeting guides would need to be placed behind promoter or enhancer elements, targeted by the signaling pathway of interest. Much like the tRNA and guide systems detailed previously, a tRNA sequence or other self-cleaving fragment such as the 3' MALAT1 triple helix could be used (Wilusz and Sharp, 2012) to separate this self-targeting guide from a marker gene. This would allow the guide to complex with Cas9 and retain the polyadenylation needed for

capture in single-cell RNA sequencing.

There are many applications of such recordings. First, we could use this approach to link terminal cell states to historical expression of key lineage branching factors, many of which are not retained after their use in early embryogenesis. For instance, the early transient expression of *Liz* has been shown to epigenetically prime regions that play an important role later in development (Greenberg et al., 2017). These signals could be multiplexed for a large number of developmental factors, once suitable expression systems are determined. With self-targeting guides such an approach would also be semi-quantitative, as more edited sequences would indicate higher expression or longer exposure to the signal of interest. These recordings could also be used as a concrete way to bridge the gap to other experimental datasets. For instance, say we were going to generate single-cell RNA-Seq for whole adult flies, and we had a previous dataset of embryonic single-cell ATAC-Seq. We could use cell-type specific hypersensitivity regions discovered in the ATAC-Seq as promoters, linking the two datasets. We could then not only infer the lineage relationships between cells, but also relate the current expression to its generative chromatin landscape earlier in development.

4.6 Applications in the single-cell era

Lineage tracing provides a developmental scaffold on which to anchor other biological findings. As we build and improve these scaffolds, it's important to think about lineage tracing's role in redefining our set notions of biology. We are witnessing a sea change in the static, categorical view of cell typing. Historically biologists had bunched the 37 million or so cells of the human body (Bianconi et al., 2013) into surprisingly limited number distinct classes, types, and subtypes. Citable numbers are hard to come by, but many authors give a total cell-type count of approximately 200. These cell classifications originate from countless studies into the form, function, behavior, or location of specific cells and tissues. With the advent of single-cell technologies we're seeing a more nuanced picture emerge, including large variations within canonical types, stochastic transitions between types, gradients thereof, and novel cell types previously unknown to biology (Macosko et al., 2015).

We first need a rich catalog of individual cells. Single-cell technologies have made enormous strides over the last half-decade; we're now able to capture whole transcriptional profiles using microfluidic approaches, single-cell sorting, or combinatorial indexing (Cao et al., 2017; Macosko et al., 2015; Macaulay et al., 2016). In parallel, methods to measure chromatin state (single-cell ATAC-Seq), three dimensional genome structure (single-cell Hi-C), and DNA variation have become available (Ramani et al., 2017; Cusanovich et al., 2015; Chen et al., 2017). These techniques provide an unprecedented view of the variation between cells, and allow for the interrogation of individual branch points in development, such as the differentiation of terminal lung cells (Treutlein et al., 2014).

The challenge is then to reconstitute larger developmental vignettes, spanning more than a single tree bifurcation, using these high-dimensional measurements of individual cells. This effort faces many challenges. Often data for rapid cell transitions states is not captured, and much of captured data suffers from collection bias. To quote Michael Yaffe: "biomedical scientists tend to look under the sequencing lamppost where the 'light is brightest' -that is, where the most data can be obtained as quickly as possible"Yaffe (2013). Such challenges aren't confined to experimental design. Transcriptional programs can overlap, be reused in development, or be regulated by orthogonal or even opposing signal cascades (Trapnell, 2015). These single-cell measurements are also high dimensional in nature, and have increased technical variability compared to their bulk counterparts (Grun et al., 2014). Lastly individual cells are also dynamic, responding to local signals in their environment, and activating various programs in response (Shaffer et al., 2017).

Lineage information can provide a path through this darkness, attaching diffuse single-cell states onto the scaffold of development. Even sparse lineage measurements will allow connections over difficult-to-capture sections of the developmental landscape. GESTALT and other CRISPR/Cas9 lineage technologies are amenable to direct integration with so many of these single-cell approaches. As the technologies mature and lineage information become richer, it's foreseeable that most single-cell experiments will be done on the background of lineage tracing animals.

4.7 Cancer evolution

Portions of this section have been adapted from a previous F31 application

A cancer's lineage begins much like our own, with a single transformed cell. This cell begins to divide, acquiring new mutations as it grows. Each of these mutations face selective pressures for growth and immune evasion. When this collection of cells has reached the detection threshold of modern medicine, it is a complex and heterogenous mixure, and in many cases, has hierarchical structure and interdependencies with proximal tissue (Greaves and Maley, 2012b). With the increased resolution of prospective lineage tracing technologies and advances in sequencing, we are in a prime moment to deepen our understanding of cancer biology by characterizing the lineage of individual tumors.

Second generation sequencing allows for the direct interrogation of cancer's mutational landscape at an unprecedented scale. It is now tractable with existing technologies to describe the full spectrum of driver mutations in most common cancer types(Lawrence et al., 2014). Additionally, the digital nature of sequencing allows for the identification of mutations that are present only in small subset of cells(Cibulskis et al., 2013). Our increased resolution into cancer's mutational landscape reinforces the theory that tumors are molecularly heterogeneous, and that subclonal populations thrive and may even be necessary for the tumor's development (Greaves and Maley, 2012a; Caldas, 2012). Although different degrees of tumor heterogeneity have been described for decades, it is unclear how this standing variation is spatially organized or how it modulates important clinical outcomes such as disease progression (Landau et al., 2013), resistance to chemotherapeutics(Magrangeas et al., 2013), and eventual metastasis(Klco et al., 2014). New molecular approaches are needed to leverage this abundant sequencing data to resolve clonal evolution, heterogeneity, and spatial interactions.

Histopathology has long been the primary source for tumor classifications, but in the last decade biomarker and molecular classifications have increasingly been used to assign tumor

subtypes. For instance, breast cancers are classified as either HER2, triple negative, luminal-A, or luminal-B(Fan et al., 2006). In breast cancer it has been thought that these classifications correlate with distinct patterns of mutations in specific genes and clear differences in survival rates. These classifications have even led to personalized cancer treatments, with a modicum of success. Recent studies have begun to break down this single-tumor/single-type hypothesis, showing that even in well-defined subtypes the mutational landscape can be highly heterogeneous and contain only partially stereotyped driver mutations. Studies such as this raise obvious questions. Do cancer cells differentiate to fulfill specific functional needs of the tumor community? What are the temporal and spatial components to tumor evolution?

Recent advances in genotyping technologies have enabled researchers to more deeply characterize these aspects of the clonal evolution and heterogeneity of tumors, offering tantalizing clues to these questions. Ding et al. used whole genome sequencing to capture clonal evolution before and after treatment in acute myeloid leukemia (AML) (Ding et al., 2012). The authors showed that the mutational landscape constantly evolves, and that treatment acts as a selective pressure on clonal diversity, culling the majority of clones that respond to chemotherapy. In a similar finding, Landau at al. used sequencing and array data to show that clonal diversity may provide an equilibrium in Chronic Lymphocytic Leukemia (CLL), which when disrupted leads to more aggressive tumor growth and a worsened prognosis (Landau et al., 2013). These two vignettes suggest a complex evolutionary process within tumors that we're only beginning to understand, and which will require new approaches to characterize fully.

One such approach would be tumor lineage tracing. Clearly labeling human tumors *in situ* is out of the question. Given such limitations, there are two indirect approaches commonly used to characterize cancer development: xenografts, and induced or predisposed animal models. Each has it's own advantages and drawbacks. Patient derived xenograft models of cancer have been shown to retain much of the heterogeneity of the tumor itself (Hidalgo et al., 2014), and engraftment rates are high in experienced hands. The challenge here is

two-fold: cells would have to be tagged pre-engraftment, each with a unique identifier. There are reasonable viral tagging approaches available, but experiments would have to be carefully monitored for tagging rates and integration-dependent growth. The other pitfall is that such an approach doesn't account for the role of human stromal tissue in the maintenance and selection on the tumor cell population.

Alternatively, predisposed or induced mouse cancer line could be used for lineage tracing. A large collection of cancer types have been modeled with this approach (Frese and Tuveson, 2007). In one potential approach, a lineage cassette would record normal development, while another lineage recording cassette would be inhibited by an inline floxed stop site. A homozygous version of this mouse could then be bred to mice which induce an oncogenic transformation using activation of the recombinase. This would ensure recording of normal development, while enabling a separate lineage recording at the birth of the tumor. Many other approaches are possible. The important point is that such a two-tiered system would put tumor development into the context of whole-organism development. As more credence is given to the relationship between tumor evolution and normal developmental patterns, systems that can provide a concrete link between development and disease will become increasing valuable.

4.8 Closing remarks

We are entering a new atomic age of biology. Single-cell technologies can now measure transcriptomes, genomes, regulatory DNA, and 3D structures of individual cells (Ramani et al., 2017; Cao et al., 2017; Cusanovich et al., 2015; Chen et al., 2017). Other complementary measurements of single-cell state are sure to follow. Importantly, these technologies aren't a cell-by-cell affair, but instead scale to thousands or even millions of cells. A vast space of previously unexplored biology is now open to us. Many challenges accompany this new space. The data produced is high dimensional, and sophisticated new techniques must be devised to integrate across platforms. Batch effects still confound analysis (Tung et al., 2017). Importantly such single-cell measurements cannot be made for every cell over the continuous

landscape of development. The solution will involve carefully designed experiments leveraging known biology, with inference methods that use lineage information to reconstruct the full cellular trajectories from zygote to adult.

There is a long-standing precedent for such inferences in biology, and the branching structure of the tree has been the central pattern. With great foresight, Darwin wrote "The affinities of all the beings of the same class have sometimes been represented by a great tree. I believe this simile largely speaks the truth. The green and budding twigs may represent existing species; and those produced during each former year may represent the long succession of extinct species." (Darwin, 1909). Darwin's evolutionary tree of life provided a scaffold for understanding speciation and evolution. Mendelian pedigrees do the same for genetic inheritance over generations. In the future, lineage trees are poised to transform our knowledge of developmental biology and the pathology of disease.

In his Nobel lecture, John Sulston discusses the decision he and the rest of the worm team faced at the completion of the *C. elegans* lineage. Their tree provided a scaffold onto which they could contextualize biological findings, but there was a scarcity of these biological 'decorations' available. Each investigator took a different tangent to enrich this tree. Robert Horvitz's solution was "Heavy duty molecular biology" (Sulston, 2003); Sulston explored the genome. Our challenge is quite the opposite: not one of paucity but of richness. How do we develop cell lineage information rich enough to exploit the ongoing single-cell revolution? How do we integrate these signals into a cohesive view of development? Despite the enormous challenges, advances in lineage tracing will play a critical role in scaffolding diverse biological measurements into a cohesive view of development and disease.

Chapter 5

APPENDIX

5.1 Source code

All source code used in this work is archived on the GitHub website:

https://github.com/aaronmck/

5.2 Appendix A: Whole-organism lineage tracing by combinatorial and cumulative genome editing

The following tables represent the remainder of supplementary table 1 from *Whole-organism lineage tracing by combinatorial and cumulative genome editing.* McKenna A, Findlay GM, Gagnon JA, Horwitz MS, Schier AF, and Shendure J.Science 29 JUL 2016. A portion of this table has been added to Chapter 3. The remaining supplementary tables and figures were too large to supply here, and are attached to the associated publication.

experiment	sample	refName	umi	passHMIDs	uniqueHMIDs
2016_05_04_V1_V2	1	HEK	TRUE	4448	3
2016_05_04_V1_V2	2	HEK	TRUE	7494	698
2016_05_04_V1_V2	3	HEK	TRUE	4099	818
2016_05_04_V1_V2	4	HEK	TRUE	3665	1089
2016_05_04_V1_V2	5	HEK	TRUE	4812	7
2016_05_04_V1_V2	6	HEK	TRUE	7101	1006
2016_05_04_V1_V2	7	HEK	TRUE	4079	1247
2016_05_04_V1_V2	8	HEK	TRUE	2500	1305
2016_05_04_V1_V2	9	HEK	TRUE	5219	7
2016_05_04_V1_V2	10	HEK	TRUE	4933	272
2016_05_04_V1_V2	11	HEK	TRUE	1711	475
2016_05_04_V1_V2	12	HEK	TRUE	309	96
2016_05_04_V1_V2	13	HEK	TRUE	117	23
2016_05_04_HEK4_exp3_redo	142	HEK4_t1_20	FALSE	345099	851
2016_05_04_HEK4_exp3_redo	143	HEK4_t1_20	FALSE	30401	203
2016_05_04_HEK4_exp3_redo	144	HEK4_t1_20	FALSE	231858	701
2016_05_04_HEK4_exp3_redo	D4	HEK4_t1_20	FALSE	47319	856
2016_05_04_HEK4_exp3_redo	D5	HEK4_t1_20	FALSE	2531	214
2016_05_04_HEK4_exp3_redo	D6	HEK4_t1_20	FALSE	8494	698
2016_05_04_HEK4_exp3_redo	D7	HEK4_t1_20	FALSE	3866	813
2016_05_04_HEK4_exp3_redo	D8	HEK4_t1_20	FALSE	14368	1539
2016_05_04_HEK4_exp3_redo	D9	HEK4_t1_20	FALSE	248250	8402
2016_05_04_HEK4_exp3_redo	R1	HEK4_t1_20	FALSE	861614	1311
2016_05_04_HEK4_exp3_redo	R2	HEK4_t1_20	FALSE	909416	1267
2016_05_04_HEK4_exp3_redo	R3	HEK4_t1_20	FALSE	745641	1206
2016_05_04_HEK4_exp3_redo	R7	HEK4_t1_20	FALSE	864350	23796
2016_05_04_HEK4_exp3_redo	R8	HEK4_t1_20	FALSE	145094	5411
2016_05_04_HEK4_exp3_redo	R9	HEK4_t1_20	FALSE	1011692	40715
2016_05_04_Rerun_HEK_analysis	r10control	HEK4_Rev10_20	FALSE	981536	513
2016_05_04_Rerun_HEK_analysis	r10high	HEK4_Rev10_20	FALSE	801579	834
2016_05_04_Rerun_HEK_analysis	r10low	HEK4_Rev10_20	FALSE	1261199	700
2016_05_04_Rerun_HEK_analysis	r50control	HEK4_Rev50_20	FALSE	1052283	437
2016_05_04_Rerun_HEK_analysis	r50high	HEK4_Rev50_20	FALSE	1258195	1346
2016_05_04_Rerun_HEK_analysis	r50low	HEK4_Rev50_20	FALSE	1190405	649
2016_05_04_Rerun_HEK_analysis	t1control	HEK4_t1_20	FALSE	765511	1479
2016_05_04_Rerun_HEK_analysis	t1high	HEK4_t1_20	FALSE	656859	11580
2016_05_04_Rerun_HEK_analysis	t1low	HEK4_t1_20	FALSE	897448	7986
2016_05_04_Rerun_HEK_analysis	u500control	HEK4_U500_20	FALSE	929754	933
2016_05_04_Rerun_HEK_analysis	u500high	HEK4_U500_20	FALSE	665570	916
2016_05_04_Rerun_HEK_analysis	u500low	HEK4_U500_20	FALSE	943134	668

Table 5.1: Cell culture statistics for samples not used in the synthetic lineage

experiment	edited.HMID.prop	meanSitesEdited	meanCutSites	meanEvents	meanIntactSites
2016_05_04_V1_V2	0.00044964	0.00044964	0.00044964	0.00044964	9.789343525
2016_05_04_V1_V2	0.403923139	0.845209501	0.584067254	0.476114225	8.805844676
2016_05_04_V1_V2	0.686264943	1.864601122	1.164674311	0.877287143	7.721395462
2016_05_04_V1_V2	0.909140518	3.136425648	1.816643929	1.268758527	6.350068213
2016_05_04_V1_V2	0.010806318	0.048212801	0.021197007	0.013923525	10.7861596
2016_05_04_V1_V2	0.421630756	0.856217434	0.703422053	0.568370652	9.913814956
2016_05_04_V1_V2	0.704094141	1.755087031	1.381466046	1.071341015	8.906594754
2016_05_04_V1_V2	0.9016	3.0308	2.2044	1.6012	7.5364
2016_05_04_V1_V2	0.006706266	0.037555087	0.013220924	0.006706266	10.7955547
2016_05_04_V1_V2	0.147374823	0.30853436	0.221771741	0.167443746	10.49604703
2016_05_04_V1_V2	0.549970777	1.475745178	1.081823495	0.787258913	9.212741087
2016_05_04_V1_V2	0.54368932	1.537216828	1.103559871	0.812297735	9.213592233
2016_05_04_V1_V2	0.376068376	1.162393162	0.854700855	0.666666667	9.547008547
2016_05_04_HEK4_exp3_redo	0.033164396	0.037548645	0.035734673	0.034645131	8.860515388
2016_05_04_HEK4_exp3_redo	0.050195717	0.060688793	0.057761258	0.055590277	8.817769152
2016_05_04_HEK4_exp3_redo	0.035034374	0.040843965	0.03830793	0.036811324	8.901525934
2016_05_04_HEK4_exp3_redo	0.272343033	0.363934149	0.328176842	0.300069739	8.032206936
2016_05_04_HEK4_exp3_redo	0.397076254	0.884630581	0.634926906	0.506914263	7.322402213
2016_05_04_HEK4_exp3_redo	0.398045679	0.730868849	0.57240405	0.469390158	8.228278785
2016_05_04_HEK4_exp3_redo	0.67563373	1.896016555	1.302121055	0.952146922	6.785049146
2016_05_04_HEK4_exp3_redo	0.767608575	2.21875	1.559507238	1.114768931	6.393095768
2016_05_04_HEK4_exp3_redo	0.829176234	2.706670695	1.750735146	1.176378651	6.116483384
2016_05_04_HEK4_exp3_redo	0.014876731	0.027450807	0.01921278	0.01549998	9.170903676
2016_05_04_HEK4_exp3_redo	0.014758922	0.028549091	0.019232123	0.015413188	9.185221065
2016_05_04_HEK4_exp3_redo	0.014983082	0.030745359	0.020102167	0.015665716	9.172455646
2016_05_04_HEK4_exp3_redo	0.821310812	2.132603691	1.545383236	1.15212009	6.884851044
2016_05_04_HEK4_exp3_redo	0.858781204	2.319303348	1.687375081	1.241574428	6.710484238
2016_05_04_HEK4_exp3_redo	0.913354064	2.574363541	1.881676439	1.375457155	6.399661162
2016_05_04_Rerun_HEK_analysis	0.010952222	0.011670484	0.011378085	0.011229339	6.160442409
2016_05_04_Rerun_HEK_analysis	0.034607943	0.037923898	0.037362506	0.036322059	6.165963679
2016_05_04_Rerun_HEK_analysis	0.019090564	0.020853172	0.020416286	0.019820028	6.194480808
2016_05_04_Rerun_HEK_analysis	0.008912051	0.01005528	0.009505998	0.009257966	1.43E-05
2016_05_04_Rerun_HEK_analysis	0.047142931	0.052090495	0.050703587	0.049339729	3.50E-05
2016_05_04_Rerun_HEK_analysis	0.016501947	0.01822741	0.017410881	0.017073181	1.68E-06
2016_05_04_Rerun_HEK_analysis	0.08674859	0.296951971	0.16224326	0.091302411	4.883729953
2016_05_04_Rerun_HEK_analysis	0.745676013	1.923817745	1.470271398	1.123874378	6.18082724
2016_05_04_Rerun_HEK_analysis	0.279303091	0.813911224	0.542174031	0.365816181	6.534623733
2016_05_04_Rerun_HEK_analysis	0.01224625	0.014524272	0.013468079	0.013073351	0.000236622
2016_05_04_Rerun_HEK_analysis	0.052627072	0.062786784	0.058226783	0.056042189	9.32E-05
2016_05_04_Rerun_HEK_analysis	0.022061552	0.025770463	0.0239616	0.023284072	4.24E-06

experiment	targetEditProp1	uniqueEventsTarget1	intactProp1	targetEditProp2	uniqueEventsTarget2
2016_05_04_V1_V2	0.00044964	2	0.989208633	0	0
2016_05_04_V1_V2	0.402188417	395	0.583533493	0.104883907	199
2016_05_04_V1_V2	0.682361552	384	0.304464504	0.279824347	256
2016_05_04_V1_V2	0.903956344	459	0.082673943	0.521691678	348
2016_05_04_V1_V2	0.00166251	2	0.997298421	0.00166251	2
2016_05_04_V1_V2	0.416701873	397	0.575693564	0.100690044	218
2016_05_04_V1_V2	0.697474871	392	0.292473646	0.221622947	247
2016_05_04_V1_V2	0.8944	396	0.0956	0.4196	301
2016_05_04_V1_V2	0.000574823	2	0.998083924	0.00076643	3
2016_05_04_V1_V2	0.142104196	211	0.85586864	0.048043787	105
2016_05_04_V1_V2	0.542372881	259	0.452367037	0.237872589	166
2016_05_04_V1_V2	0.54368932	72	0.453074434	0.249190939	41
2016_05_04_V1_V2	0.367521368	20	0.632478632	0.162393162	12
2016_05_04_HEK4_exp3_redo	0.002903515	167	0.95046349	0.002155903	192
2016_05_04_HEK4_exp3_redo	0.004473537	72	0.95292918	0.003453834	57
2016_05_04_HEK4_exp3_redo	0.003230426	164	0.952423466	0.002367829	167
2016_05_04_HEK4_exp3_redo	0.258691012	423	0.702170376	0.023436674	167
2016_05_04_HEK4_exp3_redo	0.386408534	137	0.560252864	0.102331094	84
2016_05_04_HEK4_exp3_redo	0.381681187	370	0.582999765	0.091358606	206
2016_05_04_HEK4_exp3_redo	0.660889809	361	0.306001035	0.276254527	267
2016_05_04_HEK4_exp3_redo	0.758212695	532	0.220002784	0.332614143	395
2016_05_04_HEK4_exp3_redo	0.819899295	1756	0.164704935	0.421711984	1177
2016_05_04_HEK4_exp3_redo	0.003318191	210	0.953855207	0.003373901	302
2016_05_04_HEK4_exp3_redo	0.003236143	193	0.956066311	0.00341208	253
2016_05_04_HEK4_exp3_redo	0.003501685	185	0.954181704	0.003898659	263
2016_05_04_HEK4_exp3_redo	0.812757563	4565	0.169441777	0.286633887	2216
2016_05_04_HEK4_exp3_redo	0.852536976	1289	0.134292252	0.313038444	828
2016_05_04_HEK4_exp3_redo	0.904510464	6845	0.080754815	0.341340052	2949
2016_05_04_Rerun_HEK_analysis	0.000212932	68	0.923349729	0.001862387	105
2016_05_04_Rerun_HEK_analysis	0.015182533	134	0.914514976	0.003816218	145
2016_05_04_Rerun_HEK_analysis	0.005273553	106	0.923621887	0.002247068	136
2016_05_04_Rerun_HEK_analysis	0.00214296	61	0	0.00036112	114
2016_05_04_Rerun_HEK_analysis	0.013875433	200	3.34E-05	0.011541931	274
2016_05_04_Rerun_HEK_analysis	0.005761905	122	0	0.003360201	180
2016_05_04_Rerun_HEK_analysis	0.007012309	160	0.95271002	0.072641673	291
2016_05_04_Rerun_HEK_analysis	0.727154534	1526	0.253578013	0.250184591	1322
2016_05_04_Rerun_HEK_analysis	0.208880069	857	0.745565203	0.142437222	1067
2016_05_04_Rerun_HEK_analysis	0.00073783	78	0	0.00090454	159
2016_05_04_Rerun_HEK_analysis	0.019541145	136	7.81E-05	0.005722914	128
2016_05_04_Rerun_HEK_analysis	0.005722411	95	0	0.001911711	116

experiment	intactProp2	targetEditProp3	uniqueEventsTarget3	intactProp3	targetEditProp4
2016_05_04_V1_V2	0.978417266	0	0	0.98471223	0
2016_05_04_V1_V2	0.853749666	0.141446491	253	0.817053643	0.096744062
2016_05_04_V1_V2	0.668455721	0.372773847	337	0.57794584	0.266650403
2016_05_04_V1_V2	0.425920873	0.690040928	462	0.278035471	0.512414734
2016_05_04_V1_V2	0.996674979	0.000207814	1	0.990232751	0.005818786
2016_05_04_V1_V2	0.866779327	0.171384312	322	0.817349669	0.067455288
2016_05_04_V1_V2	0.724442265	0.383917627	376	0.592547193	0.17013974
2016_05_04_V1_V2	0.5088	0.7008	434	0.2708	0.354
2016_05_04_V1_V2	0.995401418	0.000383215	1	0.991952481	0.00479019
2016_05_04_V1_V2	0.938374214	0.048246503	112	0.939387796	0.03040746
2016_05_04_V1_V2	0.7106955	0.323202805	211	0.650496786	0.158971362
2016_05_04_V1_V2	0.692556634	0.355987055	55	0.621359223	0.187702265
2016_05_04_V1_V2	0.837606838	0.247863248	15	0.700854701	0.111111111
2016_05_04_HEK4_exp3_redo	0.889759171	0.001214144	222	0.942349876	0.002804992
2016_05_04_HEK4_exp3_redo	0.888918128	0.002861748	72	0.938784908	0.004144601
2016_05_04_HEK4_exp3_redo	0.888565415	0.001699316	217	0.940269475	0.003182983
2016_05_04_HEK4_exp3_redo	0.843276485	0.028614299	231	0.899173693	0.018956445
2016_05_04_HEK4_exp3_redo	0.762149348	0.150533386	107	0.755432635	0.113393915
2016_05_04_HEK4_exp3_redo	0.789027549	0.116317401	273	0.818224629	0.066399812
2016_05_04_HEK4_exp3_redo	0.579668908	0.403000517	356	0.527677186	0.267977237
2016_05_04_HEK4_exp3_redo	0.512110245	0.509813474	518	0.437778396	0.358992205
2016_05_04_HEK4_exp3_redo	0.48469285	0.575814703	1637	0.368636455	0.435859013
2016_05_04_HEK4_exp3_redo	0.888352557	0.002676372	371	0.943902954	0.00460415
2016_05_04_HEK4_exp3_redo	0.890257044	0.002876571	326	0.945243981	0.004608452
2016_05_04_HEK4_exp3_redo	0.888513373	0.003492297	338	0.941856738	0.005174072
2016_05_04_HEK4_exp3_redo	0.588226991	0.463591138	3241	0.48022676	0.277385318
2016_05_04_HEK4_exp3_redo	0.572167009	0.527547659	1174	0.421078749	0.318393593
2016_05_04_HEK4_exp3_redo	0.528045097	0.596007481	4630	0.351399438	0.363001783
2016_05_04_Rerun_HEK_analysis	0.916550183	0.000236364	66	0.893480219	0.000372885
2016_05_04_Rerun_HEK_analysis	0.916880307	0.001787721	86	0.898026271	0.004816743
2016_05_04_Rerun_HEK_analysis	0.922759216	0.00077466	84	0.894222878	0.00197114
2016_05_04_Rerun_HEK_analysis	1.90E-06	0.000323107	94	0	0.004795288
2016_05_04_Rerun_HEK_analysis	1.59E-06	0.001860602	161	0	0.004742508
2016_05_04_Rerun_HEK_analysis	0	0.000603996	120	8.40E-07	0.00401964
2016_05_04_Rerun_HEK_analysis	0.857923661	0.067796544	338	0.875301596	0.065957249
2016_05_04_Rerun_HEK_analysis	0.648764803	0.445812572	1688	0.493844189	0.286534858
2016_05_04_Rerun_HEK_analysis	0.744542302	0.175781772	1367	0.741403402	0.140432649
2016_05_04_Rerun_HEK_analysis	0	0.000716319	158	1.08E-06	0.003273984
2016_05_04_Rerun_HEK_analysis	7.51E-06	0.012799555	160	1.50E-06	0.008047238
2016_05_04_Rerun_HEK_analysis	0	0.004073652	126	0	0.004637729

experiment	uniqueEventsTarget4	intactProp4	targetEditProp5	uniqueEventsTarget5	intactProp5
2016_05_04_V1_V2	0	0.913669065	0	0	0.986061151
2016_05_04_V1_V2	188	0.771550574	0.066986923	156	0.887910328
2016_05_04_V1_V2	241	0.638692364	0.185655038	207	0.713100756
2016_05_04_V1_V2	356	0.404638472	0.350341064	313	0.46521146
2016_05_04_V1_V2	3	0.987531172	0.005403159	2	0.992310889
2016_05_04_V1_V2	170	0.909308548	0.051401211	148	0.923250246
2016_05_04_V1_V2	225	0.781564109	0.134591812	200	0.800931601
2016_05_04_V1_V2	277	0.5624	0.302	262	0.5796
2016_05_04_V1_V2	2	0.989269975	0.00479019	2	0.99444338
2016_05_04_V1_V2	64	0.961078451	0.011149402	28	0.973038719
2016_05_04_V1_V2	119	0.802454705	0.098188194	93	0.845704267
2016_05_04_V1_V2	32	0.770226537	0.084142395	17	0.86407767
2016_05_04_V1_V2	9	0.854700855	0.136752137	11	0.863247863
2016_05_04_HEK4_exp3_redo	147	0.932607744	0.000776589	140	0.929191334
2016_05_04_HEK4_exp3_redo	49	0.93042992	0.001644683	41	0.928291833
2016_05_04_HEK4_exp3_redo	145	0.931807399	0.000944544	125	0.929064341
2016_05_04_HEK4_exp3_redo	127	0.906020837	0.012362899	100	0.904858513
2016_05_04_HEK4_exp3_redo	66	0.779533781	0.095219281	46	0.792572106
2016_05_04_HEK4_exp3_redo	171	0.855898281	0.037320462	113	0.864139393
2016_05_04_HEK4_exp3_redo	235	0.631143301	0.155457838	174	0.669684428
2016_05_04_HEK4_exp3_redo	373	0.547466592	0.187152004	296	0.573705457
2016_05_04_HEK4_exp3_redo	1177	0.479931521	0.273011078	919	0.52858006
2016_05_04_HEK4_exp3_redo	268	0.932614837	0.002707709	239	0.929901325
2016_05_04_HEK4_exp3_redo	259	0.935467377	0.002846882	259	0.931475804
2016_05_04_HEK4_exp3_redo	255	0.932464819	0.003190543	238	0.929420458
2016_05_04_HEK4_exp3_redo	2092	0.623105224	0.18140221	1898	0.656307052
2016_05_04_HEK4_exp3_redo	757	0.579596675	0.19405351	610	0.622954774
2016_05_04_HEK4_exp3_redo	2948	0.528498792	0.232916738	2802	0.572378748
2016_05_04_Rerun_HEK_analysis	64	0.810844432	0.000229233	64	0.004217879
2016_05_04_Rerun_HEK_analysis	138	0.810921943	0.000276953	76	0.004173014
2016_05_04_Rerun_HEK_analysis	98	0.817766268	0.000433714	90	0.004174599
2016_05_04_Rerun_HEK_analysis	139	9.50E-07	0.000235678	79	0
2016_05_04_Rerun_HEK_analysis	172	0	0.007066472	325	0
2016_05_04_Rerun_HEK_analysis	132	0	0.001108866	93	0
2016_05_04_Rerun_HEK_analysis	262	0.878478559	0.062401455	276	0.41499469
2016_05_04_Rerun_HEK_analysis	1250	0.624773353	0.158472366	1059	0.391522382
2016_05_04_Rerun_HEK_analysis	1084	0.772884891	0.10179197	1020	0.375285253
2016_05_04_Rerun_HEK_analysis	90	0	0.000847536	99	0.000233395
2016_05_04_Rerun_HEK_analysis	116	1.50E-06	0.004436799	100	4.51E-06
2016_05_04_Rerun_HEK_analysis	87	0	0.00165618	70	3.18E-06

experiment	targetEditProp6	uniqueEventsTarget6	intactProp6	targetEditProp7	uniqueEventsTarget7
2016_05_04_V1_V2	0	0	0.997077338	0	0
2016_05_04_V1_V2	0.018548172	75	0.967707499	0.008406725	45
2016_05_04_V1_V2	0.050012198	105	0.919736521	0.016101488	51
2016_05_04_V1_V2	0.103956344	175	0.837380628	0.032742156	83
2016_05_04_V1_V2	0.005403159	2	0.973399834	0.008935993	4
2016_05_04_V1_V2	0.012815096	59	0.950429517	0.017180679	66
2016_05_04_V1_V2	0.038734984	88	0.899975484	0.045109095	92
2016_05_04_V1_V2	0.1044	137	0.8276	0.1144	135
2016_05_04_V1_V2	0.00479019	2	0.97374976	0.005939835	3
2016_05_04_V1_V2	0.007095074	14	0.97182242	0.008514089	17
2016_05_04_V1_V2	0.045002922	49	0.914669784	0.036236119	37
2016_05_04_V1_V2	0.042071197	9	0.944983819	0.038834951	9
2016_05_04_V1_V2	0.042735043	4	0.871794872	0.034188034	3
2016_05_04_HEK4_exp3_redo	0.000672271	101	0.935302623	0.000608521	98
2016_05_04_HEK4_exp3_redo	0.000888129	26	0.934048222	0.000526298	16
2016_05_04_HEK4_exp3_redo	0.000776337	78	0.93266137	0.000698704	91
2016_05_04_HEK4_exp3_redo	0.001310256	39	0.91546736	0.000908726	25
2016_05_04_HEK4_exp3_redo	0.008297116	17	0.883050178	0.003951008	10
2016_05_04_HEK4_exp3_redo	0.007770191	43	0.898987521	0.003178714	20
2016_05_04_HEK4_exp3_redo	0.042162442	63	0.836523539	0.019141231	29
2016_05_04_HEK4_exp3_redo	0.03591314	127	0.851684298	0.008908686	54
2016_05_04_HEK4_exp3_redo	0.063186304	424	0.796656596	0.020237664	214
2016_05_04_HEK4_exp3_redo	0.002308458	202	0.933716258	0.001826804	192
2016_05_04_HEK4_exp3_redo	0.002435629	193	0.935156188	0.001964997	202
2016_05_04_HEK4_exp3_redo	0.002678233	205	0.933923966	0.002072043	192
2016_05_04_HEK4_exp3_redo	0.052763348	949	0.842711864	0.025210852	583
2016_05_04_HEK4_exp3_redo	0.064103271	289	0.835720292	0.025810854	153
2016_05_04_HEK4_exp3_redo	0.069228579	1469	0.81781906	0.033207735	869
2016_05_04_Rerun_HEK_analysis	0.000605174	93	0.754435905	0.007237636	85
2016_05_04_Rerun_HEK_analysis	0.000562639	73	0.760051099	0.007809586	68
2016_05_04_Rerun_HEK_analysis	0.000677926	95	0.763550399	0.007695058	92
2016_05_04_Rerun_HEK_analysis	0.00027179	71	0	0.000323107	61
2016_05_04_Rerun_HEK_analysis	0.001212849	125	0	0.001847885	124
2016_05_04_Rerun_HEK_analysis	0.000260416	71	0	0.00099546	85
2016_05_04_Rerun_HEK_analysis	0.004875175	228	0.246926563	0.004626975	224
2016_05_04_Rerun_HEK_analysis	0.030120619	472	0.564687399	0.005355792	265
2016_05_04_Rerun_HEK_analysis	0.018597178	600	0.360791934	0.010759398	524
2016_05_04_Rerun_HEK_analysis	0.000454959	114	0	0.000742132	230
2016_05_04_Rerun_HEK_analysis	0.001244046	76	0	0.001021681	97
2016_05_04_Rerun_HEK_analysis	0.000628755	64	1.06E-06	0.000661624	111

experiment	intactProp7	targetEditProp8	uniqueEventsTarget8	intactProp8	targetEditProp9
2016_05_04_V1_V2	0.98471223	0	0	0.977293165	0
2016_05_04_V1_V2	0.975713904	0.005070723	29	0.971310382	0.0006672
2016_05_04_V1_V2	0.962185899	0.008294706	31	0.963161747	0.002195657
2016_05_04_V1_V2	0.945156889	0.015006821	37	0.949795362	0.004638472
2016_05_04_V1_V2	0.894222776	0.005610973	2	0.982128013	0.008312552
2016_05_04_V1_V2	0.918884664	0.006900437	23	0.980002817	0.006759611
2016_05_04_V1_V2	0.90438833	0.022554548	47	0.960039225	0.019122334
2016_05_04_V1_V2	0.8512	0.0464	71	0.9364	0.0432
2016_05_04_V1_V2	0.886568308	0.00555662	2	0.986204254	0.00479019
2016_05_04_V1_V2	0.896006487	0.004459761	8	0.985201703	0.004257044
2016_05_04_V1_V2	0.892460549	0.014026885	17	0.978375219	0.011104617
2016_05_04_V1_V2	0.915857605	0.01618123	4	0.980582524	0.009708738
2016_05_04_V1_V2	0.905982906	0.025641026	2	0.957264957	0.025641026
2016_05_04_HEK4_exp3_redo	0.915114214	0.002190676	109	0.891413768	0.00119386
2016_05_04_HEK4_exp3_redo	0.914673859	0.003453834	27	0.889641788	0.001578895
2016_05_04_HEK4_exp3_redo	0.915137714	0.001673438	80	0.891955421	0.001237827
2016_05_04_HEK4_exp3_redo	0.909211944	0.002028783	34	0.87768127	0.001901984
2016_05_04_HEK4_exp3_redo	0.868826551	0.007506914	17	0.826945871	0.006321612
2016_05_04_HEK4_exp3_redo	0.896515187	0.004473746	26	0.874381917	0.004709206
2016_05_04_HEK4_exp3_redo	0.860320745	0.017330574	24	0.833936886	0.01758924
2016_05_04_HEK4_exp3_redo	0.8905902	0.005567929	37	0.86247216	0.004175947
2016_05_04_HEK4_exp3_redo	0.878940584	0.015585096	146	0.873985901	0.027838872
2016_05_04_HEK4_exp3_redo	0.914598649	0.002100709	183	0.89178797	0.001200073
2016_05_04_HEK4_exp3_redo	0.916212163	0.002231102	176	0.893516279	0.001296436
2016_05_04_HEK4_exp3_redo	0.914316675	0.002314787	152	0.892155877	0.001271389
2016_05_04_HEK4_exp3_redo	0.882147278	0.018150055	403	0.864650894	0.006552901
2016_05_04_HEK4_exp3_redo	0.884212993	0.01395647	117	0.870490854	0.00555502
2016_05_04_HEK4_exp3_redo	0.873664119	0.019439711	572	0.859985055	0.007143478
2016_05_04_Rerun_HEK_analysis	0.816518192	0.000658152	66	0.125053997	0.000255722
2016_05_04_Rerun_HEK_analysis	0.819990294	0.000704859	60	0.122410892	0.002966645
2016_05_04_Rerun_HEK_analysis	0.822021743	0.000660483	75	0.123737015	0.00111957
2016_05_04_Rerun_HEK_analysis	0	0.000129243	55	1.14E-05	0.001472988
2016_05_04_Rerun_HEK_analysis	0	0.000510255	88	0	0.00943256
2016_05_04_Rerun_HEK_analysis	8.40E-07	0.000110887	41	0	0.00200604
2016_05_04_Rerun_HEK_analysis	0.367207003	0.002671418	159	0.118076683	0.001579337
2016_05_04_Rerun_HEK_analysis	0.735311536	0.003433005	198	0.692049588	0.007395803
2016_05_04_Rerun_HEK_analysis	0.579572298	0.002939446	326	0.490952122	0.002902675
2016_05_04_Rerun_HEK_analysis	2.15E-06	0.000635652	163	0	0.001009945
2016_05_04_Rerun_HEK_analysis	0	0.0029178	100	0	0.000670102
2016_05_04_Rerun_HEK_analysis	0	0.000732664	66	0	0.000462288

experiment	uniqueEventsTarget9	intactProp9	targetEditProp10	uniqueEventsTarget10	intactProp10
2016_05_04_V1_V2	0	0.980440647	0	0	0.997751799
2016_05_04_V1_V2	5	0.981451828	0.00026688	2	0.995863357
2016_05_04_V1_V2	9	0.976579654	0.000731886	3	0.997072457
2016_05_04_V1_V2	17	0.969986357	0.001637108	6	0.991268759
2016_05_04_V1_V2	2	0.986076475	0.005195345	1	0
2016_05_04_V1_V2	16	0.984650049	0.003379806	10	0
2016_05_04_V1_V2	33	0.971071341	0.014464329	30	0
2016_05_04_V1_V2	66	0.9464	0.0288	47	0
2016_05_04_V1_V2	2	0.986970684	0.00479019	2	0
2016_05_04_V1_V2	5	0.987837016	0.003243462	3	0
2016_05_04_V1_V2	11	0.974868498	0.007013442	6	0
2016_05_04_V1_V2	2	0.98381877	0.006472492	2	0
2016_05_04_V1_V2	2	0.923076923	0.008547009	1	0
2016_05_04_HEK4_exp3_redo	114	0.885519807	0.023028175	99	0.588793361
2016_05_04_HEK4_exp3_redo	26	0.878753988	0.037663235	34	0.561297326
2016_05_04_HEK4_exp3_redo	84	0.887215451	0.025032563	70	0.632425881
2016_05_04_HEK4_exp3_redo	35	0.881569771	0.015723071	37	0.192776686
2016_05_04_HEK4_exp3_redo	12	0.82615567	0.01066772	14	0.267483208
2016_05_04_HEK4_exp3_redo	27	0.874381917	0.017659524	25	0.773722628
2016_05_04_HEK4_exp3_redo	26	0.852819452	0.03621314	18	0.687273668
2016_05_04_HEK4_exp3_redo	32	0.88592706	0.017399777	27	0.611358575
2016_05_04_HEK4_exp3_redo	111	0.873389728	0.053526687	87	0.666964753
2016_05_04_HEK4_exp3_redo	158	0.88803571	0.00333444	85	0.89413821
2016_05_04_HEK4_exp3_redo	161	0.888370119	0.003640798	97	0.8934558
2016_05_04_HEK4_exp3_redo	147	0.888950581	0.003151651	94	0.896671454
2016_05_04_HEK4_exp3_redo	228	0.887941228	0.008156418	135	0.890091977
2016_05_04_HEK4_exp3_redo	70	0.895536687	0.004307552	34	0.894433953
2016_05_04_HEK4_exp3_redo	302	0.891137817	0.007567521	164	0.895978223
2016_05_04_Rerun_HEK_analysis	45	0.915991874	NA	NA	NA
2016_05_04_Rerun_HEK_analysis	99	0.918994884	NA	NA	NA
2016_05_04_Rerun_HEK_analysis	69	0.922626802	NA	NA	NA
2016_05_04_Rerun_HEK_analysis	32	0	NA	NA	NA
2016_05_04_Rerun_HEK_analysis	104	0	NA	NA	NA
2016_05_04_Rerun_HEK_analysis	48	0	NA	NA	NA
2016_05_04_Rerun_HEK_analysis	168	0.087980447	0.007389835	91	0.084130731
2016_05_04_Rerun_HEK_analysis	205	0.880657797	0.009353606	245	0.895638181
2016_05_04_Rerun_HEK_analysis	441	0.869182393	0.009388845	214	0.854443934
2016_05_04_Rerun_HEK_analysis	136	0	0.005201376	51	0
2016_05_04_Rerun_HEK_analysis	70	0	0.006385504	39	0
2016_05_04_Rerun_HEK_analysis	71	0	0.005283449	25	0

experiment	targetEditProp11	uniqueEventsTarget11	intactProp11	containsProp11	targetEditProp12
2016_05_04_V1_V2	NA	NA	NA	NA	NA
2016_05_04_V1_V2	NA	NA	NA	NA	NA
2016_05_04_V1_V2	NA	NA	NA	NA	NA
2016_05_04_V1_V2	NA	NA	NA	NA	NA
2016_05_04_V1_V2	0	0	0.995428096	0.995428096	0
2016_05_04_V1_V2	0.000563301	4	0.996338544	0.996479369	0.000985777
2016_05_04_V1_V2	0.004903163	15	0.990683991	0.990929149	0.002451581
2016_05_04_V1_V2	0.0172	35	0.976	0.976	0.0056
2016_05_04_V1_V2	0.000383215	1	0.997317494	0.997509101	0
2016_05_04_V1_V2	0.000810866	4	0.996351105	0.996959254	0.000202716
2016_05_04_V1_V2	0.001753361	3	0.997662186	0.997662186	0
2016_05_04_V1_V2	0.003236246	1	0.987055016	0.987055016	0
2016_05_04_V1_V2	0	0	1	1	0
2016_05_04_HEK4_exp3_redo	NA	NA	NA	NA	NA
2016_05_04_HEK4_exp3_redo	NA	NA	NA	NA	NA
2016_05_04_HEK4_exp3_redo	NA	NA	NA	NA	NA
2016_05_04_HEK4_exp3_redo	NA	NA	NA	NA	NA
2016_05_04_HEK4_exp3_redo	NA	NA	NA	NA	NA
2016_05_04_HEK4_exp3_redo	NA	NA	NA	NA	NA
2016_05_04_HEK4_exp3_redo	NA	NA	NA	NA	NA
2016_05_04_HEK4_exp3_redo	NA	NA	NA	NA	NA
2016_05_04_HEK4_exp3_redo	NA	NA	NA	NA	NA
2016_05_04_HEK4_exp3_redo	NA	NA	NA	NA	NA
2016_05_04_HEK4_exp3_redo	NA	NA	NA	NA	NA
2016_05_04_HEK4_exp3_redo	NA	NA	NA	NA	NA
2016_05_04_HEK4_exp3_redo	NA	NA	NA	NA	NA
2016_05_04_HEK4_exp3_redo	NA	NA	NA	NA	NA
2016_05_04_HEK4_exp3_redo	NA	NA	NA	NA	NA
2016_05_04_Rerun_HEK_analysis	NA	NA	NA	NA	NA
2016_05_04_Rerun_HEK_analysis	NA	NA	NA	NA	NA
2016_05_04_Rerun_HEK_analysis	NA	NA	NA	NA	NA
2016_05_04_Rerun_HEK_analysis	NA	NA	NA	NA	NA
2016_05_04_Rerun_HEK_analysis	NA	NA	NA	NA	NA
2016_05_04_Rerun_HEK_analysis	NA	NA	NA	NA	NA
2016_05_04_Rerun_HEK_analysis	NA	NA	NA	NA	NA
2016_05_04_Rerun_HEK_analysis	NA	NA	NA	NA	NA
2016_05_04_Rerun_HEK_analysis	NA	NA	NA	NA	NA
2016_05_04_Rerun_HEK_analysis	NA	NA	NA	NA	NA
2016_05_04_Rerun_HEK_analysis	NA	NA	NA	NA	NA
2016_05_04_Rerun_HEK_analysis	NA	NA	NA	NA	NA

experiment	sample	refName	umi	passHMIDs	uniqueHMIDs
2016_05_04_Cell_Culture_Lineage	1	HEK4_t1_20	TRUE	67	2
2016_05_04_Cell_Culture_Lineage	2	HEK4_t1_20	TRUE	262	3
2016_05_04_Cell_Culture_Lineage	3	HEK4_t1_20	TRUE	179	2
2016_05_04_Cell_Culture_Lineage	4	HEK4_t1_20	TRUE	184	5
2016_05_04_Cell_Culture_Lineage	5	HEK4_t1_20	TRUE	892	4
2016_05_04_Cell_Culture_Lineage	6	HEK4_t1_20	TRUE	395	4
2016_05_04_Cell_Culture_Lineage	7	HEK4_t1_20	TRUE	306	2
2016_05_04_Cell_Culture_Lineage	8	HEK4_t1_20	TRUE	532	3
2016_05_04_Cell_Culture_Lineage	9	HEK4_t1_20	TRUE	576	2
2016_05_04_Cell_Culture_Lineage	10	HEK4_t1_20	TRUE	1272	1
2016_05_04_Cell_Culture_Lineage	11	HEK4_t1_20	TRUE	1467	3
2016_05_04_Cell_Culture_Lineage	12	HEK4_t1_20	TRUE	1603	3
2016_05_04_Cell_Culture_Lineage	13	HEK4_t1_20	TRUE	484	15
2016_05_04_Cell_Culture_Lineage	14	HEK4_t1_20	TRUE	565	25
2016_05_04_Cell_Culture_Lineage	15	HEK4_t1_20	TRUE	1388	36
2016_05_04_Cell_Culture_Lineage	16	HEK4_t1_20	TRUE	1863	40
2016_05_04_Cell_Culture_Lineage	17	HEK4_t1_20	TRUE	656	16
2016_05_04_Cell_Culture_Lineage	18	HEK4_t1_20	TRUE	2446	25
2016_05_04_Cell_Culture_Lineage	19	HEK4_t1_20	TRUE	2340	37
2016_05_04_Cell_Culture_Lineage	20	HEK4_t1_20	TRUE	1493	31
2016_05_04_Cell_Culture_Lineage	21	HEK4_t1_20	TRUE	948	15
2016_05_04_Cell_Culture_Lineage	22	HEK4_t1_20	TRUE	940	5
2016_05_04_Cell_Culture_Lineage	23	HEK4_t1_20	TRUE	976	22
2016_05_04_Cell_Culture_Lineage	24	HEK4_t1_20	TRUE	948	16
2016_05_04_Cell_Culture_Lineage	25	HEK4_t1_20	TRUE	1062	21
2016_05_04_Cell_Culture_Lineage	26	HEK4_t1_20	TRUE	752	22
2016_05_04_Cell_Culture_Lineage	27	HEK4_t1_20	TRUE	876	19
2016_05_04_Cell_Culture_Lineage	28	HEK4_t1_20	TRUE	924	28
2016_05_04_Cell_Culture_Lineage	29	HEK4_t1_20	TRUE	827	28
2016_05_04_Cell_Culture_Lineage	30	HEK4_t1_20	TRUE	806	28
2016_05_04_Cell_Culture_Lineage	31	HEK4_t1_20	TRUE	1747	44
2016_05_04_Cell_Culture_Lineage	32	HEK4_t1_20	TRUE	2175	31
2016_05_04_Cell_Culture_Lineage	33	HEK4_t1_20	TRUE	1980	42
2016_05_04_Cell_Culture_Lineage	34	HEK4_t1_20	TRUE	1291	46
2016_05_04_Cell_Culture_Lineage	35	HEK4_t1_20	TRUE	1659	42
2016_05_04_Cell_Culture_Lineage	36	HEK4_t1_20	TRUE	1772	44

Table 5.2: Cell culture statistics from synthetic lineage experiment

experiment	edited.HMID.prop	meanSitesEdited	meanCutSites	meanEvents	meanIntactSites
2016__05__04__Cell__Culture__Lineage	1	1.014925373	1.014925373	1.014925373	8.626865672
2016__05__04__Cell__Culture__Lineage	0.007633588	0.041984733	0.041984733	0.026717557	9.801526718
2016__05__04__Cell__Culture__Lineage	1	1.005586592	1.005586592	1.005586592	8.687150838
2016__05__04__Cell__Culture__Lineage	0.641304348	0.695652174	0.695652174	0.679347826	9.005434783
2016__05__04__Cell__Culture__Lineage	1	1.003363229	1.003363229	1.003363229	8.682735426
2016__05__04__Cell__Culture__Lineage	0.994936709	1.040506329	1.040506329	1.040506329	8.675949367
2016__05__04__Cell__Culture__Lineage	1	1	1	1	8.774509804
2016__05__04__Cell__Culture__Lineage	1	2.236842105	1.312030075	1.003759399	7.537593985
2016__05__04__Cell__Culture__Lineage	0.006944444	0.006944444	0.006944444	0.006944444	9.734375
2016__05__04__Cell__Culture__Lineage	0	0	0	0	9.808176101
2016__05__04__Cell__Culture__Lineage	0.001363327	0.001363327	0.001363327	0.001363327	9.773687798
2016__05__04__Cell__Culture__Lineage	0.001871491	0.001871491	0.001871491	0.001871491	9.72489083
2016__05__04__Cell__Culture__Lineage	1	1.417355372	1.328512397	1.183884298	8.183884298
2016__05__04__Cell__Culture__Lineage	1	1.546902655	1.518584071	1.391150442	7.922123894
2016__05__04__Cell__Culture__Lineage	0.633285303	1.948126801	1.257925072	0.80259366	7.79610951
2016__05__04__Cell__Culture__Lineage	0.364465915	0.826623725	0.552871712	0.40633387	8.862050456
2016__05__04__Cell__Culture__Lineage	1	1.4375	1.416158537	1.31097561	8.105182927
2016__05__04__Cell__Culture__Lineage	1	1.311937858	1.276778414	1.235077678	8.191741619
2016__05__04__Cell__Culture__Lineage	0.811538462	1.348290598	1.16965812	0.985042735	8.313675214
2016__05__04__Cell__Culture__Lineage	0.731413262	1.034159411	0.957803081	0.869390489	8.6316142
2016__05__04__Cell__Culture__Lineage	0.998945148	1.321729958	1.270042194	1.213080169	8.228902954
2016__05__04__Cell__Culture__Lineage	1	1.142553191	1.142553191	1.111702128	8.359574468
2016__05__04__Cell__Culture__Lineage	0.99897541	1.741803279	1.579918033	1.414959016	7.678278689
2016__05__04__Cell__Culture__Lineage	1	1.324894515	1.32278481	1.214135021	8.091772152
2016__05__04__Cell__Culture__Lineage	0.99905838	1.548022599	1.417137476	1.241996234	7.928436911
2016__05__04__Cell__Culture__Lineage	1	1.706117021	1.543882979	1.320478723	7.776595745
2016__05__04__Cell__Culture__Lineage	1	1.45890411	1.311643836	1.165525114	8.029680365
2016__05__04__Cell__Culture__Lineage	1	1.742424242	1.596320346	1.37012987	7.767316017
2016__05__04__Cell__Culture__Lineage	0.596130593	1.822249093	1.018137848	0.663845224	7.823458283
2016__05__04__Cell__Culture__Lineage	0.612903226	1.710918114	1.094292804	0.776674938	7.818858561
2016__05__04__Cell__Culture__Lineage	0.447052089	0.903262736	0.769891242	0.573554665	8.617057813
2016__05__04__Cell__Culture__Lineage	0.431724138	1.051494253	0.76045977	0.497931034	8.619310345
2016__05__04__Cell__Culture__Lineage	0.431313131	0.789393939	0.615656566	0.465656566	8.738888889
2016__05__04__Cell__Culture__Lineage	0.569326104	1.411309063	0.99535244	0.766847405	8.231603408
2016__05__04__Cell__Culture__Lineage	0.4026522	0.797468354	0.622061483	0.50331525	8.833031947
2016__05__04__Cell__Culture__Lineage	0.459367946	0.961625282	0.761851016	0.59255079	8.563205418

experiment	targetEditProp1	uniqueEventsTarget1	intactProp1	targetEditProp2	uniqueEventsTarget2
2016_05_04_Cell_Culture_Lineage	1	1	0	0	0
2016_05_04_Cell_Culture_Lineage	0.003816794	1	0.988549618	0.003816794	1
2016_05_04_Cell_Culture_Lineage	1	1	0	0	0
2016_05_04_Cell_Culture_Lineage	0.635869565	2	0.364130435	0.005434783	1
2016_05_04_Cell_Culture_Lineage	1	1	0	0.001121076	1
2016_05_04_Cell_Culture_Lineage	0.984810127	1	0.005063291	0.010126582	1
2016_05_04_Cell_Culture_Lineage	1	2	0	0	0
2016_05_04_Cell_Culture_Lineage	1	2	0	0.308270677	1
2016_05_04_Cell_Culture_Lineage	0	0	0.998263889	0	0
2016_05_04_Cell_Culture_Lineage	0	0	0.997641509	0	0
2016_05_04_Cell_Culture_Lineage	0	0	0.99795501	0	0
2016_05_04_Cell_Culture_Lineage	0	0	0.997504679	0	0
2016_05_04_Cell_Culture_Lineage	1	2	0	0.033057851	1
2016_05_04_Cell_Culture_Lineage	1	3	0	0.001769912	1
2016_05_04_Cell_Culture_Lineage	0.633285303	30	0.365994236	0.298991354	13
2016_05_04_Cell_Culture_Lineage	0.363929147	32	0.634997316	0.113794954	15
2016_05_04_Cell_Culture_Lineage	1	1	0	0.013719512	2
2016_05_04_Cell_Culture_Lineage	1	2	0	0.000408831	1
2016_05_04_Cell_Culture_Lineage	0.811111111	21	0.188461538	0.07008547	10
2016_05_04_Cell_Culture_Lineage	0.731413262	17	0.268586738	0.032819826	8
2016_05_04_Cell_Culture_Lineage	0.998945148	2	0.001054852	0.001054852	1
2016_05_04_Cell_Culture_Lineage	1	1	0	0	0
2016_05_04_Cell_Culture_Lineage	0.99897541	3	0	0.008196721	2
2016_05_04_Cell_Culture_Lineage	0.998945148	1	0.001054852	0.007383966	2
2016_05_04_Cell_Culture_Lineage	0.99905838	3	0.00094162	0.063088512	3
2016_05_04_Cell_Culture_Lineage	1	4	0	0.046542553	2
2016_05_04_Cell_Culture_Lineage	1	3	0	0.049086758	2
2016_05_04_Cell_Culture_Lineage	1	5	0	0.045454545	3
2016_05_04_Cell_Culture_Lineage	0.596130593	22	0.403869407	0.347037485	15
2016_05_04_Cell_Culture_Lineage	0.611662531	21	0.370967742	0.351116625	10
2016_05_04_Cell_Culture_Lineage	0.44590727	22	0.553520321	0.120778477	10
2016_05_04_Cell_Culture_Lineage	0.431264368	25	0.567816092	0.23816092	11
2016_05_04_Cell_Culture_Lineage	0.423232323	35	0.563636364	0.123232323	13
2016_05_04_Cell_Culture_Lineage	0.567776917	37	0.430673896	0.206041828	12
2016_05_04_Cell_Culture_Lineage	0.4026522	34	0.596745027	0.08981314	15
2016_05_04_Cell_Culture_Lineage	0.458803612	33	0.540632054	0.075620767	10

experiment	intactProp2	targetEditProp3	uniqueEventsTarget3	intactProp3	targetEditProp4
2016_05_04_Cell_Culture_Lineage	1	0	0	0.850746269	0
2016_05_04_Cell_Culture_Lineage	0.996183206	0.003816794	1	0.938931298	0.003816794
2016_05_04_Cell_Culture_Lineage	1	0	0	0.759776536	0
2016_05_04_Cell_Culture_Lineage	0.994565217	0.005434783	1	0.815217391	0.005434783
2016_05_04_Cell_Culture_Lineage	0.998878924	0	0	0.746636771	0
2016_05_04_Cell_Culture_Lineage	0.989873418	0	0	0.830379747	0.010126582
2016_05_04_Cell_Culture_Lineage	0.996732026	0	0	0.85620915	0
2016_05_04_Cell_Culture_Lineage	0.691729323	0.308270677	1	0.541353383	0.308270677
2016_05_04_Cell_Culture_Lineage	1	0	0	0.809027778	0
2016_05_04_Cell_Culture_Lineage	0.999213836	0	0	0.867138365	0
2016_05_04_Cell_Culture_Lineage	0.999318337	0	0	0.824130879	0
2016_05_04_Cell_Culture_Lineage	0.99937617	0	0	0.782907049	0
2016_05_04_Cell_Culture_Lineage	0.95661157	0.181818182	12	0.619834711	0.144628099
2016_05_04_Cell_Culture_Lineage	0.991150442	0.267256637	16	0.499115044	0.125663717
2016_05_04_Cell_Culture_Lineage	0.690201729	0.40778098	19	0.495677233	0.347262248
2016_05_04_Cell_Culture_Lineage	0.858293076	0.147611379	23	0.693505099	0.130971551
2016_05_04_Cell_Culture_Lineage	0.981707317	0.192073171	12	0.50152439	0.105182927
2016_05_04_Cell_Culture_Lineage	0.996320523	0.136549469	22	0.405560098	0.041700736
2016_05_04_Cell_Culture_Lineage	0.921367521	0.242307692	23	0.527350427	0.15982906
2016_05_04_Cell_Culture_Lineage	0.958472873	0.125251172	19	0.591426658	0.070997991
2016_05_04_Cell_Culture_Lineage	0.996835443	0.200421941	11	0.419831224	0.054852321
2016_05_04_Cell_Culture_Lineage	0.996808511	0.030851064	1	0.54787234	0.030851064
2016_05_04_Cell_Culture_Lineage	0.986680328	0.404713115	18	0.301229508	0.161885246
2016_05_04_Cell_Culture_Lineage	0.989451477	0.165611814	14	0.379746835	0.102320675
2016_05_04_Cell_Culture_Lineage	0.921845574	0.209981168	14	0.38606403	0.131826742
2016_05_04_Cell_Culture_Lineage	0.946808511	0.291223404	15	0.33643617	0.214095745
2016_05_04_Cell_Culture_Lineage	0.941780822	0.194063927	15	0.423515982	0.134703196
2016_05_04_Cell_Culture_Lineage	0.946969697	0.301948052	19	0.335497835	0.226190476
2016_05_04_Cell_Culture_Lineage	0.620314389	0.348246675	17	0.420798065	0.291414752
2016_05_04_Cell_Culture_Lineage	0.614143921	0.308933002	15	0.49751861	0.287841191
2016_05_04_Cell_Culture_Lineage	0.800228964	0.178019462	23	0.616485404	0.144819691
2016_05_04_Cell_Culture_Lineage	0.740689655	0.227126437	14	0.60137931	0.139310345
2016_05_04_Cell_Culture_Lineage	0.870707071	0.111111111	16	0.556060606	0.099494949
2016_05_04_Cell_Culture_Lineage	0.784663052	0.327652982	22	0.44771495	0.138652208
2016_05_04_Cell_Culture_Lineage	0.902350814	0.150090416	27	0.56238698	0.088004822
2016_05_04_Cell_Culture_Lineage	0.889954853	0.183408578	19	0.526523702	0.156884876

experiment	uniqueEventsTarget4	intactProp4	targetEditProp5	uniqueEventsTarget5	intactProp5
2016_05_04_Cell_Culture_Lineage	0	0.850746269	0	0	1
2016_05_04_Cell_Culture_Lineage	1	0.942748092	0.003816794	1	0.992366412
2016_05_04_Cell_Culture_Lineage	0	0.94972067	0	0	0.994413408
2016_05_04_Cell_Culture_Lineage	1	0.956521739	0.005434783	1	0.994565217
2016_05_04_Cell_Culture_Lineage	0	0.955156951	0	0	0.996636771
2016_05_04_Cell_Culture_Lineage	1	0.929113924	0	0	0.984810127
2016_05_04_Cell_Culture_Lineage	0	0.947712418	0	0	1
2016_05_04_Cell_Culture_Lineage	1	0.633458647	0.308270677	1	0.687969925
2016_05_04_Cell_Culture_Lineage	0	0.953125	0	0	0.998263889
2016_05_04_Cell_Culture_Lineage	0	0.964622642	0	0	0.999213836
2016_05_04_Cell_Culture_Lineage	0	0.97068848	0	0	0.999318337
2016_05_04_Cell_Culture_Lineage	0	0.954460387	0	0	1
2016_05_04_Cell_Culture_Lineage	6	0.77892562	0.055785124	1	0.853305785
2016_05_04_Cell_Culture_Lineage	6	0.796460177	0.12920354	5	0.775221239
2016_05_04_Cell_Culture_Lineage	10	0.614553314	0.258645533	8	0.64481268
2016_05_04_Cell_Culture_Lineage	17	0.815351583	0.059044552	7	0.898013956
2016_05_04_Cell_Culture_Lineage	4	0.858231707	0.12347561	4	0.785060976
2016_05_04_Cell_Culture_Lineage	9	0.93949305	0.127555192	7	0.865085854
2016_05_04_Cell_Culture_Lineage	11	0.803418803	0.021794872	4	0.925213675
2016_05_04_Cell_Culture_Lineage	12	0.898861353	0.07367716	9	0.921634293
2016_05_04_Cell_Culture_Lineage	6	0.907172996	0.060126582	5	0.935654008
2016_05_04_Cell_Culture_Lineage	1	0.940425532	0.04893617	2	0.919148936
2016_05_04_Cell_Culture_Lineage	8	0.68545082	0.159836066	7	0.766393443
2016_05_04_Cell_Culture_Lineage	7	0.815400844	0.025316456	2	0.943037975
2016_05_04_Cell_Culture_Lineage	8	0.806967985	0.104519774	8	0.875706215
2016_05_04_Cell_Culture_Lineage	10	0.730053191	0.136968085	5	0.784574468
2016_05_04_Cell_Culture_Lineage	10	0.831050228	0.065068493	3	0.882420091
2016_05_04_Cell_Culture_Lineage	14	0.745670996	0.159090909	7	0.756493506
2016_05_04_Cell_Culture_Lineage	9	0.683192261	0.239419589	6	0.708585248
2016_05_04_Cell_Culture_Lineage	10	0.657568238	0.130272953	4	0.714640199
2016_05_04_Cell_Culture_Lineage	8	0.81625644	0.012593017	4	0.872352604
2016_05_04_Cell_Culture_Lineage	8	0.830804598	0.014252874	3	0.895172414
2016_05_04_Cell_Culture_Lineage	9	0.879292929	0.021717172	4	0.903030303
2016_05_04_Cell_Culture_Lineage	9	0.793958172	0.162664601	9	0.814872192
2016_05_04_Cell_Culture_Lineage	14	0.876431585	0.047016275	9	0.929475588
2016_05_04_Cell_Culture_Lineage	13	0.804740406	0.070541761	8	0.836343115

experiment	targetEditProp6	uniqueEventsTarget6	intactProp6	targetEditProp7	uniqueEventsTarget7
2016_05_04_Cell_Culture_Lineage	0	0	1	0	0
2016_05_04_Cell_Culture_Lineage	0.003816794	1	0.996183206	0.003816794	1
2016_05_04_Cell_Culture_Lineage	0	0	1	0	0
2016_05_04_Cell_Culture_Lineage	0.005434783	1	0.994565217	0.005434783	1
2016_05_04_Cell_Culture_Lineage	0	0	1	0	0
2016_05_04_Cell_Culture_Lineage	0.010126582	1	0.989873418	0	0
2016_05_04_Cell_Culture_Lineage	0	0	1	0	0
2016_05_04_Cell_Culture_Lineage	0	0	1	0	0
2016_05_04_Cell_Culture_Lineage	0	0	1	0	0
2016_05_04_Cell_Culture_Lineage	0	0	0.999213836	0	0
2016_05_04_Cell_Culture_Lineage	0	0	0.99795501	0	0
2016_05_04_Cell_Culture_Lineage	0	0	1	0	0
2016_05_04_Cell_Culture_Lineage	0	0	0.997933884	0	0
2016_05_04_Cell_Culture_Lineage	0.019469027	4	0.87079646	0.001769912	1
2016_05_04_Cell_Culture_Lineage	0.000720461	1	0.996397695	0.000720461	1
2016_05_04_Cell_Culture_Lineage	0.008588298	3	0.986044015	0.001073537	1
2016_05_04_Cell_Culture_Lineage	0	0	0.99695122	0	0
2016_05_04_Cell_Culture_Lineage	0.002861815	1	0.996320523	0.002861815	1
2016_05_04_Cell_Culture_Lineage	0.014529915	2	0.983760684	0.014102564	1
2016_05_04_Cell_Culture_Lineage	0	0	0.996651038	0	0
2016_05_04_Cell_Culture_Lineage	0.002109705	2	0.984177215	0.001054852	1
2016_05_04_Cell_Culture_Lineage	0.030851064	1	0.968085106	0	0
2016_05_04_Cell_Culture_Lineage	0.00204918	2	0.960040984	0.00204918	2
2016_05_04_Cell_Culture_Lineage	0.025316456	2	0.970464135	0	0
2016_05_04_Cell_Culture_Lineage	0.025423729	3	0.972693032	0.014124294	1
2016_05_04_Cell_Culture_Lineage	0.015957447	1	0.984042553	0	0
2016_05_04_Cell_Culture_Lineage	0.015981735	2	0.974885845	0	0
2016_05_04_Cell_Culture_Lineage	0.008658009	1	0.99025974	0	0
2016_05_04_Cell_Culture_Lineage	0	0	1	0	0
2016_05_04_Cell_Culture_Lineage	0.004962779	3	0.986352357	0.003722084	1
2016_05_04_Cell_Culture_Lineage	0.00114482	2	0.989696623	0	0
2016_05_04_Cell_Culture_Lineage	0	0	0.996781609	0	0
2016_05_04_Cell_Culture_Lineage	0.002525253	1	0.983838384	0.003030303	2
2016_05_04_Cell_Culture_Lineage	0.003872967	2	0.97443842	0.000774593	1
2016_05_04_Cell_Culture_Lineage	0.015069319	2	0.983725136	0.004822182	1
2016_05_04_Cell_Culture_Lineage	0.015801354	1	0.983069977	0	0

experiment	intactProp7	targetEditProp8	uniqueEventsTarget8	intactProp8	targetEditProp9
2016__05__04__Cell__Culture__Lineage	1	0	0	1	0
2016__05__04__Cell__Culture__Lineage	0.996183206	0.003816794	1	0.996183206	0.003816794
2016__05__04__Cell__Culture__Lineage	1	0	0	1	0
2016__05__04__Cell__Culture__Lineage	0.994565217	0.005434783	1	0.994565217	0.005434783
2016__05__04__Cell__Culture__Lineage	0.998878924	0	0	1	0.001121076
2016__05__04__Cell__Culture__Lineage	0.987341772	0.010126582	1	0.997468354	0.010126582
2016__05__04__Cell__Culture__Lineage	0.996732026	0	0	0.996732026	0
2016__05__04__Cell__Culture__Lineage	0.996240602	0	0	0.998120301	0
2016__05__04__Cell__Culture__Lineage	0.998263889	0	0	0.996527778	0
2016__05__04__Cell__Culture__Lineage	0.998427673	0	0	0.998427673	0
2016__05__04__Cell__Culture__Lineage	0.998636673	0	0	0.998636673	0
2016__05__04__Cell__Culture__Lineage	0.99937617	0	0	0.99937617	0
2016__05__04__Cell__Culture__Lineage	0.997933884	0	0	0.997933884	0
2016__05__04__Cell__Culture__Lineage	0.996460177	0	0	0.996460177	0
2016__05__04__Cell__Culture__Lineage	0.999279539	0.000720461	1	0.998559078	0
2016__05__04__Cell__Culture__Lineage	0.994095545	0.001073537	1	0.996242619	0
2016__05__04__Cell__Culture__Lineage	0.99847561	0	0	0.99847561	0
2016__05__04__Cell__Culture__Lineage	0.995094031	0	0	0.999591169	0
2016__05__04__Cell__Culture__Lineage	0.985042735	0.014102564	1	0.985042735	0
2016__05__04__Cell__Culture__Lineage	0.999330208	0	0	1	0
2016__05__04__Cell__Culture__Lineage	0.997890295	0.001054852	1	0.994725738	0.001054852
2016__05__04__Cell__Culture__Lineage	0.99893617	0	0	0.99787234	0
2016__05__04__Cell__Culture__Lineage	0.994877049	0.00204918	2	0.993852459	0.00102459
2016__05__04__Cell__Culture__Lineage	0.997890295	0	0	0.997890295	0
2016__05__04__Cell__Culture__Lineage	0.973634652	0	0	0.998116761	0
2016__05__04__Cell__Culture__Lineage	0.998670213	0	0	0.998670213	0
2016__05__04__Cell__Culture__Lineage	0.985159817	0	0	1	0
2016__05__04__Cell__Culture__Lineage	0.998917749	0	0	0.998917749	0.001082251
2016__05__04__Cell__Culture__Lineage	0.99758162	0	0	0.99758162	0
2016__05__04__Cell__Culture__Lineage	0.996277916	0.003722084	1	0.992555831	0.003722084
2016__05__04__Cell__Culture__Lineage	0.99942759	0	0	0.99885518	0
2016__05__04__Cell__Culture__Lineage	0.99816092	0	0	0.996781609	0
2016__05__04__Cell__Culture__Lineage	0.996464646	0.003030303	2	0.995959596	0.000505051
2016__05__04__Cell__Culture__Lineage	0.996127033	0.000774593	1	0.996901627	0.000774593
2016__05__04__Cell__Culture__Lineage	0.994575045	0	0	0.998191682	0
2016__05__04__Cell__Culture__Lineage	0.996613995	0	0	0.99717833	0

experiment	uniqueEventsTarget9	intactProp9	targetEditProp10	uniqueEventsTarget10	intactProp10
2016_05_04_Cell_Culture_Lineage	0	1	0.014925373	1	0.925373134
2016_05_04_Cell_Culture_Lineage	1	0.996183206	0.007633588	2	0.958015267
2016_05_04_Cell_Culture_Lineage	0	1	0.005586592	1	0.983240223
2016_05_04_Cell_Culture_Lineage	1	0.994565217	0.016304348	2	0.902173913
2016_05_04_Cell_Culture_Lineage	1	0.998878924	0.001121076	1	0.987668161
2016_05_04_Cell_Culture_Lineage	1	0.989873418	0.005063291	1	0.972151899
2016_05_04_Cell_Culture_Lineage	0	1	0	0	0.980392157
2016_05_04_Cell_Culture_Lineage	0	1	0.003759398	1	0.988721805
2016_05_04_Cell_Culture_Lineage	0	1	0.006944444	1	0.980902778
2016_05_04_Cell_Culture_Lineage	0	0.997641509	0	0	0.98663522
2016_05_04_Cell_Culture_Lineage	0	0.999318337	0.001363327	2	0.987730061
2016_05_04_Cell_Culture_Lineage	0	0.99937617	0.001871491	2	0.992514036
2016_05_04_Cell_Culture_Lineage	0	0.995867769	0.002066116	1	0.98553719
2016_05_04_Cell_Culture_Lineage	0	1	0.001769912	1	0.996460177
2016_05_04_Cell_Culture_Lineage	0	0.998559078	0	0	0.992074928
2016_05_04_Cell_Culture_Lineage	0	0.994632313	0.000536769	1	0.990874933
2016_05_04_Cell_Culture_Lineage	0	0.993902439	0.00304878	1	0.990853659
2016_05_04_Cell_Culture_Lineage	0	0.999182339	0	0	0.995094031
2016_05_04_Cell_Culture_Lineage	0	0.998717949	0.00042735	1	0.995299145
2016_05_04_Cell_Culture_Lineage	0	0.998660415	0	0	0.997990623
2016_05_04_Cell_Culture_Lineage	1	0.995780591	0.001054852	1	0.995780591
2016_05_04_Cell_Culture_Lineage	0	0.995744681	0.00106383	1	0.994680851
2016_05_04_Cell_Culture_Lineage	1	0.99692623	0.00102459	1	0.992827869
2016_05_04_Cell_Culture_Lineage	0	0.997890295	0	0	0.998945148
2016_05_04_Cell_Culture_Lineage	0	0.998116761	0	0	0.994350282
2016_05_04_Cell_Culture_Lineage	0	0.998670213	0.001329787	1	0.998670213
2016_05_04_Cell_Culture_Lineage	0	0.996575342	0	0	0.994292237
2016_05_04_Cell_Culture_Lineage	1	0.996753247	0	0	0.997835498
2016_05_04_Cell_Culture_Lineage	0	0.99879081	0	0	0.992744861
2016_05_04_Cell_Culture_Lineage	1	0.995037221	0.004962779	2	0.993796526
2016_05_04_Cell_Culture_Lineage	0	0.975386377	0	0	0.994848311
2016_05_04_Cell_Culture_Lineage	0	0.996781609	0.00137931	1	0.994942529
2016_05_04_Cell_Culture_Lineage	1	0.994444444	0.001515152	2	0.995454545
2016_05_04_Cell_Culture_Lineage	1	0.998450813	0.00232378	2	0.993803253
2016_05_04_Cell_Culture_Lineage	0	0.998794454	0	0	0.990355636
2016_05_04_Cell_Culture_Lineage	0	0.994920993	0.000564334	1	0.993227991

experiment	targetEditProp11	uniqueEventsTarget11	intactProp11	containsProp11	targetEditProp12
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA
2016_05_04_Cell_Culture_Lineage	NA	NA	NA	NA	NA

experiment	sample	refName	umi	passHMIDs	uniqueHMIDs
2016__05__04__embryo__rerun	epi90__12__0.3x	target1	TRUE	6876	1544
2016__05__04__embryo__rerun	30hr__1__1x	target1	TRUE	10697	1282
2016__05__04__embryo__rerun	30hr__2__1x	target1	TRUE	6633	1169
2016__05__04__embryo__rerun	30hr__3__1x	target1	TRUE	24023	2784
2016__05__04__embryo__rerun	30hr__4__1x	target1	TRUE	9006	1161
2016__05__04__embryo__rerun	30hr__5__1x	target1	TRUE	23385	3362
2016__05__04__embryo__rerun	30hr__6__1x	target1	TRUE	15414	1520
2016__05__04__embryo__rerun	30hr__1__0.3x	target1	TRUE	10429	507
2016__05__04__embryo__rerun	30hr__2__0.3x	target1	TRUE	9430	1075
2016__05__04__embryo__rerun	30hr__3__0.3x	target1	TRUE	14059	1815
2016__05__04__embryo__rerun	30hr__4__0.3x	target1	TRUE	14839	1954
2016__05__04__embryo__rerun	30hr__5__0.3x	target1	TRUE	17878	1826
2016__05__04__embryo__rerun	30hr__6__0.3x	target1	TRUE	15910	2693
2016__05__04__embryo__rerun	3d__1__1x	target1	TRUE	25584	2746
2016__05__04__embryo__rerun	3d__2__1x	target1	TRUE	14691	947
2016__05__04__embryo__rerun	3d__3__1x	target1	TRUE	17984	2584
2016__05__04__embryo__rerun	3d__4__1x	target1	TRUE	6574	919
2016__05__04__embryo__rerun	3d__5__1x	target1	TRUE	9541	1053
2016__05__04__embryo__rerun	3d__6__1x	target1	TRUE	9752	1574
2016__05__04__embryo__rerun	3d__1__0.3x	target1	TRUE	21720	2893
2016__05__04__embryo__rerun	3d__2__0.3x	target1	TRUE	25480	3014
2016__05__04__embryo__rerun	3d__3__0.3x	target1	TRUE	21756	2709
2016__05__04__embryo__rerun	3d__4__0.3x	target1	TRUE	23258	3116
2016__05__04__embryo__rerun	3d__5__0.3x	target1	TRUE	8785	1678
2016__05__04__embryo__rerun	3d__6__0.3x	target1	TRUE	31639	4195
2016__05__04__embryo__rerun	3d__1b__1x	target1	TRUE	20475	2536
2016__05__04__embryo__rerun	3d__1b__0.3x	target1	TRUE	12382	2189
2016__05__04__embryo__rerun	Dome__7__0.3x	target1	TRUE	2137	621
2016__05__04__embryo__rerun	Dome__10__0.3x	target1	TRUE	2323	520

Table 5.3: V7 embryo data - part 1

experiment	edited.HMID.prop	meanSitesEdited	meanCutSites	meanEvents	meanIntactSites
2016_05_04_embryo_rerun	0.948225713	3.442844677	2.702588714	2.124054683	5.299883653
2016_05_04_embryo_rerun	1	7.473777695	5.635505282	4.277274002	1.8064878
2016_05_04_embryo_rerun	0.999547716	5.643750942	4.061209106	3.039499472	3.547565204
2016_05_04_embryo_rerun	1	7.716188653	4.284019481	3.169504225	1.61091454
2016_05_04_embryo_rerun	1	6.909837886	4.799578059	3.582722629	2.421718854
2016_05_04_embryo_rerun	1	6.933760958	4.534958307	3.397134916	2.56968142
2016_05_04_embryo_rerun	0.99967562	7.735565071	4.861813935	3.18684313	1.799013883
2016_05_04_embryo_rerun	0.39227155	0.758941413	0.60092051	0.48221306	7.369834116
2016_05_04_embryo_rerun	0.967550371	4.725344645	2.72173913	1.841145281	4.478154825
2016_05_04_embryo_rerun	0.893662423	3.193328117	2.272138843	1.705384451	5.571520023
2016_05_04_embryo_rerun	0.891704293	2.891704293	2.158905587	1.597816564	5.778960846
2016_05_04_embryo_rerun	0.798691129	2.749804229	2.058507663	1.55906701	5.746783757
2016_05_04_embryo_rerun	0.917913262	3.310685104	2.509679447	1.966499057	5.365619107
2016_05_04_embryo_rerun	0.840408068	2.643370857	2.106042839	1.619136961	5.896106942
2016_05_04_embryo_rerun	0.999727724	3.686134368	3.601865087	3.044176707	4.843713838
2016_05_04_embryo_rerun	0.981094306	4.20457073	3.474477313	2.805660587	4.53080516
2016_05_04_embryo_rerun	1	6.751749315	5.1670216	4.097809553	2.356708245
2016_05_04_embryo_rerun	0.999790378	6.749292527	5.308248611	4.473116026	2.456765538
2016_05_04_embryo_rerun	0.998872026	6.72292863	4.720980312	3.582342084	2.711033634
2016_05_04_embryo_rerun	0.981721915	4.032504604	3.112200737	2.468508287	4.7446593
2016_05_04_embryo_rerun	0.850078493	2.897841444	2.302354788	1.768720565	5.725392465
2016_05_04_embryo_rerun	0.998115462	3.521143593	3.111785255	2.653888582	4.836275051
2016_05_04_embryo_rerun	0.880815203	2.859962164	2.38021326	1.872044028	5.640381804
2016_05_04_embryo_rerun	0.963574274	3.932043256	3.013318156	2.37381901	4.651337507
2016_05_04_embryo_rerun	0.90072379	3.618730048	2.447675337	1.79951958	5.224122128
2016_05_04_embryo_rerun	0.831648352	2.529279609	2.052698413	1.602100122	5.957069597
2016_05_04_embryo_rerun	0.978921014	3.700048457	3.051122597	2.476901954	5.010902924
2016_05_04_embryo_rerun	0.793635938	2.193729527	1.864763687	1.499298081	6.190453907
2016_05_04_embryo_rerun	0.953077916	3.330176496	2.639259578	1.96427034	5.340077486

experiment	targetEditProp1	uniqueEventsTarget1	intactProp1	targetEditProp2	uniqueEventsTarget2
2016_05_04_embryo_rerun	0.753635835	178	0.24272833	0.115474113	177
2016_05_04_embryo_rerun	0.868654763	198	0.053846873	0.624754604	274
2016_05_04_embryo_rerun	0.738278305	211	0.253882105	0.502035278	207
2016_05_04_embryo_rerun	0.942471798	349	0.040419598	0.736502518	577
2016_05_04_embryo_rerun	0.880857206	172	0.088940706	0.529424828	240
2016_05_04_embryo_rerun	0.806243318	486	0.152405388	0.470429763	565
2016_05_04_embryo_rerun	0.963215259	241	0.024782665	0.825483327	302
2016_05_04_embryo_rerun	0.132131556	100	0.864320644	0.04564196	66
2016_05_04_embryo_rerun	0.614846235	186	0.380275716	0.429162248	128
2016_05_04_embryo_rerun	0.325912227	278	0.668112953	0.142826659	236
2016_05_04_embryo_rerun	0.251095087	288	0.740076825	0.120358515	216
2016_05_04_embryo_rerun	0.404687325	288	0.584796957	0.213502629	220
2016_05_04_embryo_rerun	0.394217473	371	0.596228787	0.193086109	330
2016_05_04_embryo_rerun	0.410217323	331	0.576962164	0.148452158	256
2016_05_04_embryo_rerun	0.335035055	231	0.654346198	0.086787829	123
2016_05_04_embryo_rerun	0.498832295	359	0.483763345	0.324621886	274
2016_05_04_embryo_rerun	0.89017341	173	0.09111652	0.534377852	155
2016_05_04_embryo_rerun	0.927051672	140	0.050728435	0.55193376	211
2016_05_04_embryo_rerun	0.951497129	167	0.038556194	0.492821985	301
2016_05_04_embryo_rerun	0.498066298	359	0.49106814	0.200644567	281
2016_05_04_embryo_rerun	0.401923077	394	0.581279435	0.219897959	335
2016_05_04_embryo_rerun	0.457253172	364	0.522936202	0.132193418	271
2016_05_04_embryo_rerun	0.416931808	398	0.573437097	0.154011523	296
2016_05_04_embryo_rerun	0.601593625	249	0.381673307	0.233921457	228
2016_05_04_embryo_rerun	0.472960587	496	0.515977117	0.270141281	408
2016_05_04_embryo_rerun	0.408253968	318	0.577924298	0.139291819	243
2016_05_04_embryo_rerun	0.498546277	295	0.490308512	0.17048942	194
2016_05_04_embryo_rerun	0.426298549	121	0.555919513	0.135236313	74
2016_05_04_embryo_rerun	0.44640551	105	0.550150667	0.175204477	72

experiment	intactProp2	targetEditProp3	uniqueEventsTarget3	intactProp3	targetEditProp4
2016_05_04_embryo_rerun	0.725276323	0.18397324	147	0.801628854	0.358202443
2016_05_04_embryo_rerun	0.325511826	0.43563616	154	0.520893709	0.896980462
2016_05_04_embryo_rerun	0.454997739	0.366350068	175	0.62777024	0.642544852
2016_05_04_embryo_rerun	0.21383674	0.700203971	549	0.281605128	0.940848354
2016_05_04_embryo_rerun	0.400843882	0.484787919	166	0.462580502	0.829558072
2016_05_04_embryo_rerun	0.482617062	0.643660466	454	0.335984606	0.89198204
2016_05_04_embryo_rerun	0.145127806	0.727715064	230	0.20954976	0.98027767
2016_05_04_embryo_rerun	0.943043437	0.030683671	50	0.963275482	0.06980535
2016_05_04_embryo_rerun	0.551431601	0.406044539	119	0.579533404	0.517815483
2016_05_04_embryo_rerun	0.828579558	0.148445835	204	0.841951775	0.542072694
2016_05_04_embryo_rerun	0.869802547	0.172383584	182	0.820270908	0.57092796
2016_05_04_embryo_rerun	0.770723795	0.098445016	158	0.893388522	0.287000783
2016_05_04_embryo_rerun	0.786297926	0.188560654	264	0.795788812	0.444311754
2016_05_04_embryo_rerun	0.839899937	0.135748906	207	0.855847405	0.281113196
2016_05_04_embryo_rerun	0.900959771	0.054659315	89	0.932679872	0.634402015
2016_05_04_embryo_rerun	0.651301157	0.200122331	237	0.7667371	0.509341637
2016_05_04_embryo_rerun	0.347733496	0.366139337	107	0.57864314	0.885762093
2016_05_04_embryo_rerun	0.392097264	0.314013206	138	0.660832198	0.914683995
2016_05_04_embryo_rerun	0.493437244	0.387510254	212	0.575266612	0.767227235
2016_05_04_embryo_rerun	0.767771639	0.169751381	205	0.807688766	0.431353591
2016_05_04_embryo_rerun	0.767032967	0.171742543	249	0.816091052	0.390855573
2016_05_04_embryo_rerun	0.790356683	0.09146902	208	0.896074646	0.615048722
2016_05_04_embryo_rerun	0.782397455	0.122151518	235	0.862713905	0.347149368
2016_05_04_embryo_rerun	0.746613546	0.265793967	183	0.721001707	0.506431417
2016_05_04_embryo_rerun	0.704763109	0.266411707	370	0.721957078	0.446221436
2016_05_04_embryo_rerun	0.849035409	0.122588523	184	0.869304029	0.27032967
2016_05_04_embryo_rerun	0.797286383	0.120820546	139	0.857615894	0.349539654
2016_05_04_embryo_rerun	0.848385587	0.065512401	52	0.925128685	0.221338325
2016_05_04_embryo_rerun	0.808437365	0.149806285	51	0.836848902	0.294446836

experiment	uniqueEventsTarget4	intactProp4	targetEditProp5	uniqueEventsTarget5	intactProp5
2016_05_04_embryo_rerun	246	0.620564282	0.38336242	273	0.187754508
2016_05_04_embryo_rerun	264	0.038235019	0.802187529	245	0.019444704
2016_05_04_embryo_rerun	267	0.328961254	0.58857229	252	0.136288256
2016_05_04_embryo_rerun	664	0.025517213	0.888773259	615	0.018607168
2016_05_04_embryo_rerun	273	0.10937153	0.771929825	268	0.070841661
2016_05_04_embryo_rerun	606	0.081334189	0.855291854	598	0.037759247
2016_05_04_embryo_rerun	294	0.007395874	0.890294537	292	0.01226158
2016_05_04_embryo_rerun	95	0.916578771	0.046313165	75	0.036916291
2016_05_04_embryo_rerun	188	0.454718982	0.504347826	177	0.086320255
2016_05_04_embryo_rerun	306	0.438153496	0.408066008	292	0.07198236
2016_05_04_embryo_rerun	308	0.404002965	0.366871083	335	0.065705236
2016_05_04_embryo_rerun	292	0.680613044	0.331580714	264	0.044747735
2016_05_04_embryo_rerun	448	0.534632307	0.34318039	446	0.083343809
2016_05_04_embryo_rerun	356	0.66764384	0.229752971	352	0.12703252
2016_05_04_embryo_rerun	155	0.317337145	0.342862977	172	0.145122864
2016_05_04_embryo_rerun	381	0.394072509	0.417815836	384	0.105260231
2016_05_04_embryo_rerun	174	0.066778217	0.803011865	175	0.018253727
2016_05_04_embryo_rerun	168	0.036474164	0.647940467	203	0.024735353
2016_05_04_embryo_rerun	352	0.175246103	0.693601313	363	0.071575062
2016_05_04_embryo_rerun	352	0.486832413	0.594198895	301	0.009438306
2016_05_04_embryo_rerun	422	0.573508634	0.263618524	402	0.085518053
2016_05_04_embryo_rerun	330	0.340779555	0.26751241	340	0.058420666
2016_05_04_embryo_rerun	397	0.606887953	0.275647089	410	0.101685442
2016_05_04_embryo_rerun	302	0.458622652	0.402048947	293	0.119408082
2016_05_04_embryo_rerun	559	0.521034167	0.382597427	586	0.11618572
2016_05_04_embryo_rerun	321	0.674969475	0.212356532	324	0.117655678
2016_05_04_embryo_rerun	272	0.549345825	0.561944759	237	0.004361169
2016_05_04_embryo_rerun	98	0.749181095	0.158633599	107	0.132428638
2016_05_04_embryo_rerun	98	0.692208351	0.373224279	95	0.107619458

experiment	targetEditProp6	uniqueEventsTarget6	intactProp6	targetEditProp7	uniqueEventsTarget7
2016_05_04_embryo_rerun	0.406340896	281	0.112856312	0.623182083	338
2016_05_04_embryo_rerun	0.93699168	205	0.054501262	0.999439095	174
2016_05_04_embryo_rerun	0.611789537	238	0.075983718	0.95627921	199
2016_05_04_embryo_rerun	0.942471798	580	0.010656454	0.986929193	543
2016_05_04_embryo_rerun	0.717965801	276	0.135243171	0.97712636	237
2016_05_04_embryo_rerun	0.914218516	571	0.027111396	0.977506949	518
2016_05_04_embryo_rerun	0.897430907	316	0.044180615	0.977228494	283
2016_05_04_embryo_rerun	0.075750312	69	0.040176431	0.202224566	128
2016_05_04_embryo_rerun	0.74369035	181	0.04931071	0.862672322	225
2016_05_04_embryo_rerun	0.504232164	271	0.095454869	0.609716196	361
2016_05_04_embryo_rerun	0.305209246	313	0.102567558	0.564188962	390
2016_05_04_embryo_rerun	0.408602752	262	0.048551292	0.558004251	346
2016_05_04_embryo_rerun	0.353614079	410	0.168824639	0.621935889	537
2016_05_04_embryo_rerun	0.264970294	338	0.168503752	0.529549719	453
2016_05_04_embryo_rerun	0.096657818	90	0.467701314	0.566401198	43
2016_05_04_embryo_rerun	0.462744662	357	0.162088523	0.884008007	420
2016_05_04_embryo_rerun	0.903559477	149	0.032552479	0.993459081	131
2016_05_04_embryo_rerun	0.922230374	218	0.005240541	0.921391888	213
2016_05_04_embryo_rerun	0.905455291	379	0.028506973	0.977645611	382
2016_05_04_embryo_rerun	0.518554328	356	0.076933702	0.89825046	416
2016_05_04_embryo_rerun	0.336734694	392	0.169740973	0.595368917	526
2016_05_04_embryo_rerun	0.362934363	327	0.068900533	0.704357419	434
2016_05_04_embryo_rerun	0.334594548	417	0.118926821	0.605942041	497
2016_05_04_embryo_rerun	0.528742174	297	0.100056915	0.790210586	363
2016_05_04_embryo_rerun	0.464331995	588	0.134169854	0.648977528	702
2016_05_04_embryo_rerun	0.245274725	309	0.159462759	0.516336996	419
2016_05_04_embryo_rerun	0.460184138	266	0.094088193	0.884025198	313
2016_05_04_embryo_rerun	0.217126813	111	0.103416004	0.483387927	172
2016_05_04_embryo_rerun	0.418424451	87	0.131726216	0.602668963	115

experiment	intactProp7	targetEditProp8	uniqueEventsTarget8	intactProp8	targetEditProp9
2016_05_04_embryo_rerun	0.34045957	0.209569517	191	0.755671902	0.385107621
2016_05_04_embryo_rerun	0.000467421	0.772366084	129	0.000186968	0.973357016
2016_05_04_embryo_rerun	0.001658375	0.446856626	150	0.542137796	0.708276798
2016_05_04_embryo_rerun	0.00245598	0.54768347	330	0.281563502	0.909878034
2016_05_04_embryo_rerun	0.010437486	0.663224517	152	0.286919831	0.929047302
2016_05_04_embryo_rerun	0.003420996	0.448663673	385	0.501731879	0.857900363
2016_05_04_embryo_rerun	0.003114052	0.467691709	235	0.45043467	0.922278448
2016_05_04_embryo_rerun	0.787803241	0.05216224	63	0.932016492	0.081982932
2016_05_04_embryo_rerun	0.125980912	0.158112407	129	0.803075292	0.458748674
2016_05_04_embryo_rerun	0.28316381	0.16494772	183	0.757166228	0.327974963
2016_05_04_embryo_rerun	0.408383314	0.225149943	210	0.71305344	0.302311477
2016_05_04_embryo_rerun	0.35669538	0.146828504	164	0.797348697	0.289685647
2016_05_04_embryo_rerun	0.306222502	0.307353865	250	0.636077938	0.444814582
2016_05_04_embryo_rerun	0.376289869	0.196060038	227	0.770129769	0.427493746
2016_05_04_embryo_rerun	0.000544551	0.562453203	29	0.435640869	0.998842829
2016_05_04_embryo_rerun	0.053158363	0.290591637	213	0.616436833	0.559108096
2016_05_04_embryo_rerun	0	0.486309705	117	0.266504411	0.803163979
2016_05_04_embryo_rerun	0.000209622	0.519337596	140	0.385284561	0.942563673
2016_05_04_embryo_rerun	0.004511895	0.533941756	240	0.376538146	0.887510254
2016_05_04_embryo_rerun	0.076197053	0.329097606	257	0.645764273	0.352348066
2016_05_04_embryo_rerun	0.382378336	0.161420722	244	0.759379906	0.339521193
2016_05_04_embryo_rerun	0.183029969	0.117622725	160	0.774912668	0.735934915
2016_05_04_embryo_rerun	0.317009201	0.201737037	225	0.730673317	0.378450426
2016_05_04_embryo_rerun	0.183494593	0.212976665	191	0.490153671	0.368810472
2016_05_04_embryo_rerun	0.272480167	0.220898258	353	0.723821865	0.423433105
2016_05_04_embryo_rerun	0.392185592	0.181050061	213	0.784957265	0.412258852
2016_05_04_embryo_rerun	0.090696172	0.273380714	199	0.694718139	0.340655791
2016_05_04_embryo_rerun	0.480112307	0.134768367	75	0.798315395	0.33224146
2016_05_04_embryo_rerun	0.196728368	0.353852777	50	0.580284115	0.506672406

experiment	uniqueEventsTarget9	intactProp9	targetEditProp10	uniqueEventsTarget10	intactProp10
2016_05_04_embryo_rerun	299	0.591477603	0.02399651	54	0.921465969
2016_05_04_embryo_rerun	157	0.016359727	0.163410302	91	0.777040292
2016_05_04_embryo_rerun	230	0.258254184	0.082767978	52	0.867631539
2016_05_04_embryo_rerun	488	0.047787537	0.120426258	131	0.688465221
2016_05_04_embryo_rerun	228	0.052520542	0.125916056	52	0.804019543
2016_05_04_embryo_rerun	525	0.100748343	0.067864015	110	0.846568313
2016_05_04_embryo_rerun	393	0.061826911	0.083949656	99	0.840339951
2016_05_04_embryo_rerun	112	0.913798063	0.022245661	13	0.971905264
2016_05_04_embryo_rerun	223	0.531389183	0.02990456	29	0.91611877
2016_05_04_embryo_rerun	273	0.652393485	0.019133651	50	0.934561491
2016_05_04_embryo_rerun	300	0.688456095	0.013208437	41	0.966641957
2016_05_04_embryo_rerun	256	0.600178991	0.011466607	32	0.969739344
2016_05_04_embryo_rerun	354	0.4902577	0.019610308	53	0.967944689
2016_05_04_embryo_rerun	337	0.56300813	0.020012508	47	0.950789556
2016_05_04_embryo_rerun	42	0.001157171	0.008032129	17	0.988224083
2016_05_04_embryo_rerun	337	0.415146797	0.057384342	46	0.882840302
2016_05_04_embryo_rerun	208	0.09355035	0.085792516	43	0.861575905
2016_05_04_embryo_rerun	232	0.050938057	0.088145897	45	0.850225343
2016_05_04_embryo_rerun	324	0.085520919	0.125717801	72	0.861874487
2016_05_04_embryo_rerun	386	0.460174954	0.040239411	69	0.922790055
2016_05_04_embryo_rerun	356	0.637872841	0.016758242	44	0.952590267
2016_05_04_embryo_rerun	268	0.253998897	0.03681743	61	0.946865233
2016_05_04_embryo_rerun	359	0.596052971	0.023346805	38	0.950597644
2016_05_04_embryo_rerun	257	0.503244166	0.021513944	35	0.947068867
2016_05_04_embryo_rerun	513	0.563228926	0.022756724	76	0.950504125
2016_05_04_embryo_rerun	324	0.577777778	0.021538462	52	0.953797314
2016_05_04_embryo_rerun	312	0.504603457	0.040461961	60	0.927879179
2016_05_04_embryo_rerun	132	0.653252223	0.019185774	13	0.94431446
2016_05_04_embryo_rerun	102	0.474386569	0.009470512	8	0.961687473

experiment	targetEditProp11	uniqueEventsTarget11	intactProp11	containsProp11	targetEditProp12
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA

experiment	sample	refName	umi	passHMIDs	uniqueHMIDs
2016_05_04_embryo_rerun	Dome_1_1x	target1	TRUE	1262	371
2016_05_04_embryo_rerun	Dome_3_1x	target1	TRUE	2036	555
2016_05_04_embryo_rerun	Dome_4_1x	target1	TRUE	126	32
2016_05_04_embryo_rerun	Dome_5_1x	target1	TRUE	861	242
2016_05_04_embryo_rerun	Dome_8_1x	target1	TRUE	1385	337
2016_05_04_embryo_rerun	Dome_9_1x	target1	TRUE	461	15
2016_05_04_embryo_rerun	Dome_10_1x	target1	TRUE	2258	459
2016_05_04_embryo_rerun	Dome_5_0.3x	target1	TRUE	2571	588
2016_05_04_embryo_rerun	Dome_6_0.3x	target1	TRUE	3028	353
2016_05_04_embryo_rerun	epi90_2_1x	target1	TRUE	4053	1223
2016_05_04_embryo_rerun	epi90_3_1x	target1	TRUE	10	7
2016_05_04_embryo_rerun	epi90_4_1x	target1	TRUE	9	5
2016_05_04_embryo_rerun	epi90_5_1x	target1	TRUE	4014	1328
2016_05_04_embryo_rerun	epi90_6_1x	target1	TRUE	7	5
2016_05_04_embryo_rerun	epi90_7_1x	target1	TRUE	3	3
2016_05_04_embryo_rerun	epi90_8_1x	target1	TRUE	2901	258
2016_05_04_embryo_rerun	epi90_9_1x	target1	TRUE	3415	705
2016_05_04_embryo_rerun	epi90_10_1x	target1	TRUE	3	3
2016_05_04_embryo_rerun	epi90_11_1x	target1	TRUE	7	4
2016_05_04_embryo_rerun	epi90_12_1x	target1	TRUE	6259	1586
2016_05_04_embryo_rerun	epi90_1_0.3x	target1	TRUE	2087	519
2016_05_04_embryo_rerun	epi90_2_0.3x	target1	TRUE	7686	2132
2016_05_04_embryo_rerun	epi90_3_0.3x	target1	TRUE	3239	765
2016_05_04_embryo_rerun	epi90_4_0.3x	target1	TRUE	13	8
2016_05_04_embryo_rerun	epi90_5_0.3x	target1	TRUE	15	6
2016_05_04_embryo_rerun	epi90_6_0.3x	target1	TRUE	12	6
2016_05_04_embryo_rerun	epi90_8_0.3x	target1	TRUE	38	6
2016_05_04_embryo_rerun	epi90_9_0.3x	target1	TRUE	8776	1578
2016_05_04_embryo_rerun	epi90_10_0.3x	target1	TRUE	9480	745
2016_05_04_embryo_rerun	epi90_11_0.3x	target1	TRUE	5	4

Table 5.4: V7 Embryo data - part 2

experiment	edited.HMID.prop	meanSitesEdited	meanCutSites	meanEvents	meanIntactSites
2016_05_04_embryo_rerun	0.999207607	5.914421553	4.1022187	3.201267829	3.183835182
2016_05_04_embryo_rerun	0.94891945	4.634577603	3.052554028	2.186149312	4.286345776
2016_05_04_embryo_rerun	1	7.76984127	4.698412698	3.349206349	1.912698413
2016_05_04_embryo_rerun	1	4.974448316	3.615563298	2.833914053	3.802555168
2016_05_04_embryo_rerun	0.999277978	6.108303249	4.020938628	2.983393502	2.686642599
2016_05_04_embryo_rerun	0.802603037	6	3.275488069	2.494577007	3.26681128
2016_05_04_embryo_rerun	0.9574845	5.448184234	3.687333924	2.838352524	3.662976085
2016_05_04_embryo_rerun	0.771295216	2.770517308	1.976662777	1.479968884	5.899649942
2016_05_04_embryo_rerun	0.510237781	1.079590489	0.885733157	0.691215324	7.154227213
2016_05_04_embryo_rerun	0.98494942	6.355292376	3.756476684	2.6797434	3.023192697
2016_05_04_embryo_rerun	1	5.5	4.6	3.9	2.9
2016_05_04_embryo_rerun	0.444444444	2.222222222	1	0.777777778	6.222222222
2016_05_04_embryo_rerun	0.954160438	4.729446936	2.996512207	2.178126557	4.277279522
2016_05_04_embryo_rerun	1	6.428571429	3.714285714	2.857142857	2.285714286
2016_05_04_embryo_rerun	1	8	3.666666667	2.666666667	1.333333333
2016_05_04_embryo_rerun	0.514305412	1.271285764	0.963805584	0.722854188	7.011375388
2016_05_04_embryo_rerun	1	7.4	4.174524158	2.886090776	1.907174231
2016_05_04_embryo_rerun	1	6.333333333	3	2	3
2016_05_04_embryo_rerun	1	5.857142857	3	2.285714286	3.142857143
2016_05_04_embryo_rerun	0.950471321	5.367470842	3.859082921	2.878095542	3.689407254
2016_05_04_embryo_rerun	0.861523718	3.124580738	2.349784379	1.799233349	5.518447532
2016_05_04_embryo_rerun	0.964480874	3.925578975	3.174993495	2.480223784	4.786364819
2016_05_04_embryo_rerun	0.92096326	3.656066687	2.549861068	1.816918802	5.367397345
2016_05_04_embryo_rerun	0.769230769	3.538461538	2.230769231	1.615384615	4.923076923
2016_05_04_embryo_rerun	0.8	5.2	2.266666667	1.466666667	3.866666667
2016_05_04_embryo_rerun	1	5.083333333	4.083333333	3.166666667	3.666666667
2016_05_04_embryo_rerun	0.526315789	2.184210526	1.684210526	1.289473684	6.131578947
2016_05_04_embryo_rerun	0.879899727	3.15872835	2.436189608	1.732566089	5.61588423
2016_05_04_embryo_rerun	0.460126582	0.973839662	0.808860759	0.642616034	7.106751055
2016_05_04_embryo_rerun	1	7	3.8	2.8	2.2

experiment	targetEditProp1	uniqueEventsTarget1	intactProp1	targetEditProp2	uniqueEventsTarget2
2016_05_04_embryo_rerun	0.961172742	40	0.027733756	0.318541997	77
2016_05_04_embryo_rerun	0.64194499	119	0.307956778	0.279960707	90
2016_05_04_embryo_rerun	0.984126984	11	0.007936508	0.507936508	5
2016_05_04_embryo_rerun	0.735191638	56	0.264808362	0.303135889	48
2016_05_04_embryo_rerun	0.615162455	88	0.127797834	0.339350181	71
2016_05_04_embryo_rerun	0.798264642	8	0.199566161	0.47505423	5
2016_05_04_embryo_rerun	0.758635961	111	0.231620903	0.259521701	100
2016_05_04_embryo_rerun	0.364060677	112	0.628549203	0.168805912	82
2016_05_04_embryo_rerun	0.182959049	74	0.812087186	0.05317041	43
2016_05_04_embryo_rerun	0.765852455	260	0.22353812	0.503577597	272
2016_05_04_embryo_rerun	1	5	0	0.4	3
2016_05_04_embryo_rerun	0.333333333	3	0.666666667	0.222222222	2
2016_05_04_embryo_rerun	0.599152965	267	0.379671151	0.333582461	235
2016_05_04_embryo_rerun	1	5	0	0.142857143	1
2016_05_04_embryo_rerun	1	3	0	0.333333333	1
2016_05_04_embryo_rerun	0.180282661	56	0.811789038	0.063426405	35
2016_05_04_embryo_rerun	0.934407028	154	0.061493411	0.609370425	175
2016_05_04_embryo_rerun	0.666666667	2	0.333333333	0.333333333	1
2016_05_04_embryo_rerun	0.714285714	3	0.285714286	0.428571429	1
2016_05_04_embryo_rerun	0.665921074	295	0.299089311	0.368109922	269
2016_05_04_embryo_rerun	0.55055103	92	0.439865836	0.138476282	82
2016_05_04_embryo_rerun	0.677205308	276	0.31459797	0.228727557	259
2016_05_04_embryo_rerun	0.468045693	146	0.521765977	0.160852115	119
2016_05_04_embryo_rerun	0.384615385	3	0.615384615	0.076923077	1
2016_05_04_embryo_rerun	0.4	4	0.6	0.2	2
2016_05_04_embryo_rerun	0.666666667	4	0.333333333	0.25	2
2016_05_04_embryo_rerun	0.263157895	3	0.736842105	0.131578947	1
2016_05_04_embryo_rerun	0.343436645	243	0.648587056	0.184366454	167
2016_05_04_embryo_rerun	0.099050633	141	0.893459916	0.041877637	92
2016_05_04_embryo_rerun	1	4	0	0.4	2

experiment	intactProp2	targetEditProp3	uniqueEventsTarget3	intactProp3	targetEditProp4
2016_05_04_embryo_rerun	0.677496038	0.238510301	46	0.742472266	0.740095087
2016_05_04_embryo_rerun	0.70481336	0.282907662	82	0.7043222	0.455795678
2016_05_04_embryo_rerun	0.253968254	0.507936508	5	0.492063492	1
2016_05_04_embryo_rerun	0.643437863	0.303135889	36	0.68757259	0.573751452
2016_05_04_embryo_rerun	0.37833935	0.334296029	50	0.564620939	0.742960289
2016_05_04_embryo_rerun	0.305856833	0.581344902	5	0.414316703	0.787418655
2016_05_04_embryo_rerun	0.518600531	0.236492471	72	0.748007086	0.728963685
2016_05_04_embryo_rerun	0.819136523	0.171917542	66	0.817969662	0.325943213
2016_05_04_embryo_rerun	0.937252312	0.033685601	30	0.958058124	0.177014531
2016_05_04_embryo_rerun	0.469035283	0.597828769	237	0.383913151	0.775968418
2016_05_04_embryo_rerun	0.2	0.2	2	0.8	0.7
2016_05_04_embryo_rerun	0.777777778	0.222222222	2	0.777777778	0.222222222
2016_05_04_embryo_rerun	0.631788739	0.332835077	200	0.656701545	0.574987544
2016_05_04_embryo_rerun	0.571428571	0.142857143	1	0.857142857	0.857142857
2016_05_04_embryo_rerun	0	0.333333333	1	0.666666667	1
2016_05_04_embryo_rerun	0.92175112	0.055498104	27	0.934505343	0.184763875
2016_05_04_embryo_rerun	0.344948755	0.555490483	117	0.426939971	0.895754026
2016_05_04_embryo_rerun	0.333333333	0.333333333	1	0.666666667	0.666666667
2016_05_04_embryo_rerun	0.285714286	0.428571429	1	0.571428571	0.571428571
2016_05_04_embryo_rerun	0.564307397	0.408371944	227	0.58156255	0.632369388
2016_05_04_embryo_rerun	0.853378055	0.103497844	62	0.882606612	0.381887877
2016_05_04_embryo_rerun	0.737574811	0.176294562	201	0.799505595	0.368722352
2016_05_04_embryo_rerun	0.823402285	0.18400741	97	0.798703303	0.454769991
2016_05_04_embryo_rerun	0.615384615	0.076923077	1	0.923076923	0.307692308
2016_05_04_embryo_rerun	0.333333333	0.6	3	0.4	0.8
2016_05_04_embryo_rerun	0.333333333	0.166666667	1	0.833333333	0.333333333
2016_05_04_embryo_rerun	0.684210526	0	0	1	0.263157895
2016_05_04_embryo_rerun	0.793185962	0.281677302	143	0.696330902	0.368846855
2016_05_04_embryo_rerun	0.948417722	0.045042194	71	0.949050633	0.098734177
2016_05_04_embryo_rerun	0.2	0.4	2	0.6	0.6

experiment	uniqueEventsTarget4	intactProp4	targetEditProp5	uniqueEventsTarget5	intactProp5
2016_05_04_embryo_rerun	74	0.032488114	0.621236133	91	0.084786054
2016_05_04_embryo_rerun	133	0.503929273	0.512278978	147	0.144891945
2016_05_04_embryo_rerun	8	0	1	13	0
2016_05_04_embryo_rerun	63	0.404181185	0.50058072	53	0.082462253
2016_05_04_embryo_rerun	90	0.233212996	0.716967509	84	0.042599278
2016_05_04_embryo_rerun	5	0.199566161	0.79175705	5	0
2016_05_04_embryo_rerun	124	0.255535872	0.691762622	125	0.048272808
2016_05_04_embryo_rerun	118	0.650719564	0.27849086	124	0.085958771
2016_05_04_embryo_rerun	67	0.813077939	0.072655218	65	0.051188904
2016_05_04_embryo_rerun	343	0.192450037	0.771527264	339	0.038983469
2016_05_04_embryo_rerun	4	0.3	0.8	4	0
2016_05_04_embryo_rerun	2	0.777777778	0.222222222	2	0
2016_05_04_embryo_rerun	329	0.378425511	0.539860488	349	0.075984056
2016_05_04_embryo_rerun	3	0.142857143	0.857142857	4	0
2016_05_04_embryo_rerun	3	0	1	3	0
2016_05_04_embryo_rerun	54	0.802481903	0.100654947	50	0.062736987
2016_05_04_embryo_rerun	182	0.096339678	0.956661786	104	0.00556369
2016_05_04_embryo_rerun	2	0.333333333	0.666666667	2	0
2016_05_04_embryo_rerun	2	0.428571429	0.714285714	3	0
2016_05_04_embryo_rerun	340	0.315066305	0.625019971	335	0.063907973
2016_05_04_embryo_rerun	116	0.576425491	0.339722089	105	0.074748443
2016_05_04_embryo_rerun	329	0.578324226	0.369372886	375	0.190866511
2016_05_04_embryo_rerun	182	0.496449521	0.526705773	171	0.080580426
2016_05_04_embryo_rerun	3	0.615384615	0.461538462	4	0
2016_05_04_embryo_rerun	5	0.2	0.8	5	0
2016_05_04_embryo_rerun	3	0.666666667	0.666666667	4	0.25
2016_05_04_embryo_rerun	3	0.736842105	0.263157895	3	0
2016_05_04_embryo_rerun	250	0.516066545	0.330446673	274	0.198837739
2016_05_04_embryo_rerun	126	0.886814346	0.063607595	112	0.060654008
2016_05_04_embryo_rerun	3	0.4	1	4	0

experiment	targetEditProp6	uniqueEventsTarget6	intactProp6	targetEditProp7	uniqueEventsTarget7
2016_05_04_embryo_rerun	0.786053883	93	0.061014263	0.919175911	124
2016_05_04_embryo_rerun	0.615913556	138	0.1021611	0.752946955	161
2016_05_04_embryo_rerun	1	13	0	1	4
2016_05_04_embryo_rerun	0.567944251	63	0.06155633	0.698025552	72
2016_05_04_embryo_rerun	0.82166065	90	0.03898917	0.981949458	91
2016_05_04_embryo_rerun	0.787418655	7	0	0.670281996	3
2016_05_04_embryo_rerun	0.676705049	95	0.080602303	0.815766165	134
2016_05_04_embryo_rerun	0.354336834	126	0.114352392	0.565149747	195
2016_05_04_embryo_rerun	0.132760898	65	0.06010568	0.258916777	122
2016_05_04_embryo_rerun	0.760917839	338	0.113249445	0.920552677	348
2016_05_04_embryo_rerun	0.4	4	0	0.5	4
2016_05_04_embryo_rerun	0.222222222	2	0	0.333333333	3
2016_05_04_embryo_rerun	0.609616343	334	0.07922272	0.783258595	371
2016_05_04_embryo_rerun	0.857142857	4	0	0.714285714	3
2016_05_04_embryo_rerun	1	3	0	1	3
2016_05_04_embryo_rerun	0.162702516	50	0.073767666	0.287142365	93
2016_05_04_embryo_rerun	0.955197657	102	0.013469985	0.978623719	96
2016_05_04_embryo_rerun	1	3	0	1	3
2016_05_04_embryo_rerun	0.714285714	3	0	0.857142857	3
2016_05_04_embryo_rerun	0.649624541	351	0.09186771	0.808276082	409
2016_05_04_embryo_rerun	0.439865836	122	0.073310973	0.603737422	169
2016_05_04_embryo_rerun	0.31459797	365	0.2529274	0.769190736	417
2016_05_04_embryo_rerun	0.565297931	171	0.073170732	0.725223835	222
2016_05_04_embryo_rerun	0.538461538	5	0	0.615384615	5
2016_05_04_embryo_rerun	0.8	5	0	0.8	5
2016_05_04_embryo_rerun	0.666666667	4	0.25	0.583333333	4
2016_05_04_embryo_rerun	0.263157895	3	0	0.394736842	4
2016_05_04_embryo_rerun	0.415793072	266	0.138787603	0.62579763	329
2016_05_04_embryo_rerun	0.097362869	118	0.087025316	0.229852321	192
2016_05_04_embryo_rerun	1	4	0	0.6	3

experiment	intactProp7	targetEditProp8	uniqueEventsTarget8	intactProp8	targetEditProp9
2016_05_04_embryo_rerun	0.039619651	0.519017433	68	0.416798732	0.770206022
2016_05_04_embryo_rerun	0.106090373	0.411591356	99	0.478880157	0.631139489
2016_05_04_embryo_rerun	0	0.698412698	6	0.301587302	0.976190476
2016_05_04_embryo_rerun	0.042973287	0.341463415	30	0.616724739	0.933797909
2016_05_04_embryo_rerun	0	0.558844765	70	0.423826715	0.858483755
2016_05_04_embryo_rerun	0.299349241	0.321041215	2	0.659436009	0.787418655
2016_05_04_embryo_rerun	0.137732507	0.411426041	58	0.545172719	0.832595217
2016_05_04_embryo_rerun	0.412679891	0.211590821	101	0.743290548	0.315830416
2016_05_04_embryo_rerun	0.730515192	0.070343461	52	0.906208719	0.090488771
2016_05_04_embryo_rerun	0.050826548	0.490994325	218	0.445595855	0.705403405
2016_05_04_embryo_rerun	0.2	0.8	6	0.2	0.7
2016_05_04_embryo_rerun	0.666666667	0.111111111	1	0.888888889	0.222222222
2016_05_04_embryo_rerun	0.166417539	0.382162431	213	0.553562531	0.54857997
2016_05_04_embryo_rerun	0	0.714285714	3	0.285714286	1
2016_05_04_embryo_rerun	0	1	3	0	1
2016_05_04_embryo_rerun	0.699069286	0.10513616	47	0.86797656	0.123061013
2016_05_04_embryo_rerun	0.010834553	0.63806735	76	0.342898975	0.855929722
2016_05_04_embryo_rerun	0	0.666666667	2	0.333333333	1
2016_05_04_embryo_rerun	0	0.714285714	3	0.285714286	0.714285714
2016_05_04_embryo_rerun	0.143313628	0.406294935	257	0.521968366	0.722799169
2016_05_04_embryo_rerun	0.292285577	0.16914231	88	0.777192142	0.368950647
2016_05_04_embryo_rerun	0.169398907	0.397215717	234	0.41946396	0.590944575
2016_05_04_embryo_rerun	0.23927138	0.206853967	105	0.755171349	0.346403211
2016_05_04_embryo_rerun	0.230769231	0.461538462	4	0.538461538	0.461538462
2016_05_04_embryo_rerun	0.2	0.266666667	3	0.733333333	0.4
2016_05_04_embryo_rerun	0	0.75	4	0.166666667	0.75
2016_05_04_embryo_rerun	0.605263158	0.263157895	3	0.736842105	0.263157895
2016_05_04_embryo_rerun	0.34286691	0.16943938	171	0.752962625	0.421718323
2016_05_04_embryo_rerun	0.729535865	0.081962025	90	0.890506329	0.209177215
2016_05_04_embryo_rerun	0	0.8	3	0.2	1

experiment	uniqueEventsTarget9	intactProp9	targetEditProp10	uniqueEventsTarget10	intactProp10
2016_05_04_embryo_rerun	121	0.198098257	0.040412044	23	0.903328051
2016_05_04_embryo_rerun	134	0.33497053	0.050098232	24	0.898330059
2016_05_04_embryo_rerun	17	0.023809524	0.095238095	5	0.833333333
2016_05_04_embryo_rerun	64	0.06271777	0.017421603	8	0.93612079
2016_05_04_embryo_rerun	127	0.115523466	0.138628159	17	0.761732852
2016_05_04_embryo_rerun	6	0.193058568	0	0	0.995661605
2016_05_04_embryo_rerun	94	0.162976085	0.036315323	19	0.93445527
2016_05_04_embryo_rerun	146	0.669778296	0.014391287	15	0.957215091
2016_05_04_embryo_rerun	81	0.901585205	0.007595773	11	0.984147952
2016_05_04_embryo_rerun	293	0.233654083	0.062669627	62	0.871946706
2016_05_04_embryo_rerun	5	0.2	0	0	1
2016_05_04_embryo_rerun	2	0.777777778	0.111111111	1	0.888888889
2016_05_04_embryo_rerun	295	0.425510713	0.025411061	36	0.929995017
2016_05_04_embryo_rerun	5	0	0.142857143	1	0.428571429
2016_05_04_embryo_rerun	3	0	0.333333333	1	0.666666667
2016_05_04_embryo_rerun	68	0.865563599	0.008617718	10	0.971733885
2016_05_04_embryo_rerun	132	0.133821376	0.020497804	17	0.470863836
2016_05_04_embryo_rerun	3	0	0	0	1
2016_05_04_embryo_rerun	3	0.285714286	0	0	1
2016_05_04_embryo_rerun	376	0.254034191	0.080683815	64	0.854289823
2016_05_04_embryo_rerun	141	0.616195496	0.028749401	18	0.932438908
2016_05_04_embryo_rerun	358	0.390580276	0.033307312	58	0.933125163
2016_05_04_embryo_rerun	166	0.632602655	0.017906761	19	0.946279716
2016_05_04_embryo_rerun	4	0.538461538	0.153846154	1	0.846153846
2016_05_04_embryo_rerun	4	0.6	0.133333333	1	0.8
2016_05_04_embryo_rerun	4	0.25	0.25	1	0.583333333
2016_05_04_embryo_rerun	3	0.736842105	0.078947368	1	0.894736842
2016_05_04_embryo_rerun	263	0.567570647	0.017206016	37	0.960688241
2016_05_04_embryo_rerun	160	0.78407173	0.007172996	19	0.87721519
2016_05_04_embryo_rerun	4	0	0.2	1	0.8

experiment	targetEditProp11	uniqueEventsTarget11	intactProp11	containsProp11	targetEditProp12
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA
2016_05_04_embryo_rerun	NA	NA	NA	NA	NA

experiment	sample	refName	umi	passHMIDs	uniqueHMIDs
2016__05__04__Early__embryo__target__6__and__7	embryos__1__1	target3	TRUE	2832	542
2016__05__04__Early__embryo__target__6__and__7	embryos__1__2	target3	TRUE	1286	286
2016__05__04__Early__embryo__target__6__and__7	embryos__1__3	target3	TRUE	1058	428
2016__05__04__Early__embryo__target__6__and__7	embryos__1__4	target3	TRUE	405	163
2016__05__04__Early__embryo__target__6__and__7	embryos__1__5	target3	TRUE	397	86
2016__05__04__Early__embryo__target__6__and__7	embryos__1__6	target3	TRUE	827	142
2016__05__04__Early__embryo__target__6__and__7	embryos__1__7	target3	TRUE	1961	1323
2016__05__04__Early__embryo__target__6__and__7	embryos__1__8	target3	TRUE	257	148
2016__05__04__Early__embryo__target__6__and__7	embryos__3__13	target3	TRUE	1579	1
2016__05__04__Early__embryo__target__6__and__7	embryos__3__14	target3	TRUE	2909	1

Table 5.5: V6 embryo data

experiment	edited.HMID.prop	meanSitesEdited	meanCutSites	meanEvents	meanIntactSites
2016_05_04_Early_embryo_target_6_and_7	1	8.796610169	6.188206215	4.811793785	0.603460452
2016_05_04_Early_embryo_target_6_and_7	1	9.047433904	6.51088647	4.910575428	0.612752722
2016_05_04_Early_embryo_target_6_and_7	1	8.699432892	6.849716446	5.329867675	0.821361059
2016_05_04_Early_embryo_target_6_and_7	1	8.730864198	6.538271605	5.2	0.683950617
2016_05_04_Early_embryo_target_6_and_7	0.987405542	8.488664987	6.816120907	5.375314861	0.639798489
2016_05_04_Early_embryo_target_6_and_7	1	9.146311971	6.296251511	5.252720677	0.830713422
2016_05_04_Early_embryo_target_6_and_7	1	5.84599694	4.745028047	4.071902091	3.75063743
2016_05_04_Early_embryo_target_6_and_7	1	8.194552529	7.260700389	6.60311284	1.373540856
2016_05_04_Early_embryo_target_6_and_7	0	0	0	0	9.675744142
2016_05_04_Early_embryo_target_6_and_7	0	0	0	0	9.64352011

experiment	targetEditProp1	uniqueEventsTarget1	intactProp1	targetEditProp2	uniqueEventsTarget2
2016_05_04_Early_embryo_target_6_and_7	0.959392655	92	0.004237288	0.98269774	107
2016_05_04_Early_embryo_target_6_and_7	0.982892691	85	0.005443235	0.961897356	85
2016_05_04_Early_embryo_target_6_and_7	0.90831758	87	0.059546314	0.904536862	121
2016_05_04_Early_embryo_target_6_and_7	0.97037037	67	0.014814815	0.869135802	61
2016_05_04_Early_embryo_target_6_and_7	0.987405542	22	0.012594458	0.921914358	25
2016_05_04_Early_embryo_target_6_and_7	0.99637243	25	0.00241838	1	20
2016_05_04_Early_embryo_target_6_and_7	0.410504844	179	0.570117287	0.670066293	237
2016_05_04_Early_embryo_target_6_and_7	0.770428016	41	0.171206226	0.968871595	38
2016_05_04_Early_embryo_target_6_and_7	0	0	0.987967068	0	0
2016_05_04_Early_embryo_target_6_and_7	0	0	0.987280853	0	0

experiment	intactProp2	targetEditProp3	uniqueEventsTarget3	intactProp3	targetEditProp4
2016_05_04_Early_embryo_target_6_and_7	0.000353107	0.99329096	99	0	0.972810734
2016_05_04_Early_embryo_target_6_and_7	0.00777605	0.83281493	47	0	0.99844479
2016_05_04_Early_embryo_target_6_and_7	0.04536862	0.867674858	83	0	0.974480151
2016_05_04_Early_embryo_target_6_and_7	0.019753086	0.95308642	54	0	0.987654321
2016_05_04_Early_embryo_target_6_and_7	0.020151134	0.98488665	22	0.012594458	0.83627204
2016_05_04_Early_embryo_target_6_and_7	0	0.993954051	29	0	0.99637243
2016_05_04_Early_embryo_target_6_and_7	0.297807241	0.871494136	315	0.106068332	0.723610403
2016_05_04_Early_embryo_target_6_and_7	0.027237354	0.856031128	40	0	0.964980545
2016_05_04_Early_embryo_target_6_and_7	0.989233692	0	0	0.987967068	0
2016_05_04_Early_embryo_target_6_and_7	0.985218288	0	0	0.987624613	0

experiment	uniqueEventsTarget4	intactProp4	targetEditProp5	uniqueEventsTarget5	intactProp5
2016_05_04_Early_embryo_target_6_and_7	87	0	0.990112994	86	0.001059322
2016_05_04_Early_embryo_target_6_and_7	52	0	0.99844479	58	0.00155521
2016_05_04_Early_embryo_target_6_and_7	85	0	0.973534972	88	0.017958412
2016_05_04_Early_embryo_target_6_and_7	58	0	0.958024691	59	0.007407407
2016_05_04_Early_embryo_target_6_and_7	22	0	0.624685139	22	0.012594458
2016_05_04_Early_embryo_target_6_and_7	41	0	1	31	0
2016_05_04_Early_embryo_target_6_and_7	299	0.203977562	0.582865885	274	0.39163692
2016_05_04_Early_embryo_target_6_and_7	33	0.007782101	0.93385214	42	0.058365759
2016_05_04_Early_embryo_target_6_and_7	0	0.769474351	0	0	0.991133629
2016_05_04_Early_embryo_target_6_and_7	0	0.746304572	0	0	0.987280853

experiment	targetEditProp6	uniqueEventsTarget6	intactProp6	targetEditProp7	uniqueEventsTarget7
2016_05_04_Early_embryo_target_6_and_7	0.843220339	98	0.016242938	0.985169492	115
2016_05_04_Early_embryo_target_6_and_7	0.917573872	57	0.00777605	0.99066874	62
2016_05_04_Early_embryo_target_6_and_7	0.941398866	103	0.037807183	0.971644612	102
2016_05_04_Early_embryo_target_6_and_7	0.935802469	59	0.061728395	0.982716049	46
2016_05_04_Early_embryo_target_6_and_7	0.901763224	26	0.098236776	0.982367758	15
2016_05_04_Early_embryo_target_6_and_7	1	27	0	0.99637243	35
2016_05_04_Early_embryo_target_6_and_7	0.37072922	203	0.582355941	0.593574707	260
2016_05_04_Early_embryo_target_6_and_7	0.782101167	44	0.210116732	0.754863813	41
2016_05_04_Early_embryo_target_6_and_7	0	0	0.996200127	0	0
2016_05_04_Early_embryo_target_6_and_7	0	0	0.992437264	0	0

experiment	intactProp7	targetEditProp8	uniqueEventsTarget8	intactProp8	targetEditProp9
2016_05_04_Early_embryo_target_6_and_7	0.004237288	0.955861582	82	0.000706215	0.873234463
2016_05_04_Early_embryo_target_6_and_7	0.000777605	0.976671851	63	0	1
2016_05_04_Early_embryo_target_6_and_7	0.017958412	0.962192817	92	0.00094518	0.993383743
2016_05_04_Early_embryo_target_6_and_7	0.009876543	0.701234568	43	0	1
2016_05_04_Early_embryo_target_6_and_7	0.012594458	0.987405542	21	0.012594458	0.987405542
2016_05_04_Early_embryo_target_6_and_7	0.00241838	0.99879081	31	0	1
2016_05_04_Early_embryo_target_6_and_7	0.34982152	0.598164202	192	0.376848547	0.996940337
2016_05_04_Early_embryo_target_6_and_7	0.159533074	0.949416342	37	0.042801556	0.980544747
2016_05_04_Early_embryo_target_6_and_7	0.986067131	0	0	0.989233692	0
2016_05_04_Early_embryo_target_6_and_7	0.990030938	0	0	0.988655895	0

experiment	uniqueEventsTarget9	intactProp9	targetEditProp10	uniqueEventsTarget10	intactProp10
2016_05_04_Early_embryo_target_6_and_7	50	0	0.240819209	104	0.576624294
2016_05_04_Early_embryo_target_6_and_7	41	0	0.388024883	43	0.589424572
2016_05_04_Early_embryo_target_6_and_7	56	0	0.202268431	44	0.641776938
2016_05_04_Early_embryo_target_6_and_7	32	0	0.372839506	27	0.57037037
2016_05_04_Early_embryo_target_6_and_7	14	0.010075567	0.274559194	27	0.44836272
2016_05_04_Early_embryo_target_6_and_7	21	0	0.164449819	45	0.825876663
2016_05_04_Early_embryo_target_6_and_7	144	0.000509944	0.028046915	24	0.871494136
2016_05_04_Early_embryo_target_6_and_7	17	0	0.233463035	9	0.696498054
2016_05_04_Early_embryo_target_6_and_7	0	0.992400253	0	0	0.986067131
2016_05_04_Early_embryo_target_6_and_7	0	0.991062221	0	0	0.987624613

experiment	targetEditProp11	uniqueEventsTarget11	intactProp11	containsProp11	targetEditProp12
2016_05_04_Early_embryo_target_6_and_7	NA	NA	NA	NA	NA
2016_05_04_Early_embryo_target_6_and_7	NA	NA	NA	NA	NA
2016_05_04_Early_embryo_target_6_and_7	NA	NA	NA	NA	NA
2016_05_04_Early_embryo_target_6_and_7	NA	NA	NA	NA	NA
2016_05_04_Early_embryo_target_6_and_7	NA	NA	NA	NA	NA
2016_05_04_Early_embryo_target_6_and_7	NA	NA	NA	NA	NA
2016_05_04_Early_embryo_target_6_and_7	NA	NA	NA	NA	NA
2016_05_04_Early_embryo_target_6_and_7	NA	NA	NA	NA	NA
2016_05_04_Early_embryo_target_6_and_7	NA	NA	NA	NA	NA
2016_05_04_Early_embryo_target_6_and_7	NA	NA	NA	NA	NA

BIBLIOGRAPHY

Julien Ablain, Ellen M Durand, Song Yang, Yi Zhou, and Leonard I Zon. A CRISPR/Cas9 vector system for tissue-specific gene disruption in zebrafish. *Dev. Cell*, 32(6):756–764, 23 March 2015.

Brad Amos. Lessons from the history of light microscopy. *Nat Cell Biol*, 2(8):E151–E152, 08 2000. URL `http://dx.doi.org/10.1038/35019639`.

Fatemeh Babaei, Rajkumar Ramalingam, Amy Tavendale, Yimin Liang, Leo So Kin Yan, Paul Ajuh, Shuk Han Cheng, and Yun Wah Lam. Novel blood collection method allows plasma proteome analysis from single zebrafish. *J. Proteome Res.*, 12(4):1580–1590, 5 April 2013.

Sangsu Bae, Jeongbin Park, and Jin-Soo Kim. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of cas9 RNA-guided endonucleases. *Bioinformatics*, 30(10):1473–1475, 15 May 2014.

Sam Behjati, Meritxell Huch, Ruben van Boxtel, Wouter Karthaus, David C Wedge, Asif U Tamuri, Iñigo Martincorena, Mia Petljak, Ludmil B Alexandrov, Gunes Gundem, Patrick S Tarpey, Sophie Roerink, Joyce Blokker, Mark Maddison, Laura Mudie, Ben Robinson, Serena Nik-Zainal, Peter Campbell, Nick Goldman, Marc van de Wetering, Edwin Cuppen, Hans Clevers, and Michael R Stratton. Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature*, 513(7518):422–425, 18 September 2014.

Eva Bianconi, Allison Piovesan, Federica Facchin, Alina Beraudi, Raffaella Casadei, Flavia Frabetti, Lorenza Vitale, Maria Chiara Pelleri, Simone Tassani, Francesco Piva, Soledad Perez-Amodio, Pierluigi Strippoli, and Silvia Canaider. An estimation of the number of cells in the human body. *Ann. Hum. Biol.*, 40(6):463–471, November 2013.

Cédric Blanpain and Benjamin D Simons. Unravelling stem cell dynamics by lineage tracing. *Nat. Rev. Mol. Cell Biol.*, 14(8):489–502, August 2013.

Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, Aug 2014. ISSN 1367-4811 (Electronic); 1367-4803 (Linking). doi: 10.1093/bioinformatics/btu170.

Carlos Caldas. Cancer sequencing unravels clonal evolution. *Nat. Biotechnol.*, 30(5): 408–410, May 2012.

Junyue Cao, Jonathan S Packer, Vijay Ramani, Darren A Cusanovich, Chau Huynh, Riza Daza, Xiaojie Qiu, Choli Lee, Scott N Furlan, Frank J Steemers, Andrew Adey, Robert H Waterston, Cole Trapnell, and Jay Shendure. Comprehensive single cell transcriptional profiling of a multicellular organism by combinatorial indexing. 2 February 2017.

Cheryl A Carlson, Arnold Kas, Robert Kirkwood, Laura E Hays, Bradley D Preston, Stephen J Salipante, and Marshall S Horwitz. Decoding cell lineage from acquired mutations using arbitrary deep sequencing. *Nat. Methods*, 9(1):78–80, 27 November 2011.

Scott L Carter, Kristian Cibulskis, Elena Helman, Aaron McKenna, Hui Shen, Travis Zack, Peter W Laird, Robert C Onofrio, Wendy Winckler, Barbara A Weir, Rameen Beroukhim, David Pellman, Douglas A Levine, Eric S Lander, Matthew Meyerson, and Gad Getz. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.*, 30(5):413–421, 29 April 2012.

Chongyi Chen, Dong Xing, Longzhi Tan, Heng Li, Guangyu Zhou, Lei Huang, and X Sunney Xie. Single-cell whole-genome analyses by linear amplification via transposon insertion (LIANTI). *Science*, 356 (6334):189–194, 14 April 2017.

Kristian Cibulskis, Michael S Lawrence, Scott L Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S Lander, and Gad Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.*, 31 (3):213–219, March 2013.

Edwin G. Conklin. The organization and cell-lineage of the ascidian egg. *Journal of the Academy of Natural Sciences of Philadelphia*, 8, 1905.

Darren A Cusanovich, Riza Daza, Andrew Adey, Hannah A Pliner, Lena Christiansen, Kevin L Gunderson, Frank J Steemers, Cole Trapnell, and Jay Shendure. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science*, 348(6237):910–914, 22 May 2015.

C. Darwin. *The Origin of Species*. P. F. Collier & Son, 1909. URL `https://books.google.com/books?id=YY4EAAAAYAAJ`.

Li Ding, Timothy J Ley, David E Larson, Christopher A Miller, Daniel C Koboldt, John S Welch, Julie K Ritchey, Margaret A Young, Tamara Lamprecht, Michael D McLellan, Joshua F McMichael, John W Wallis, Charles Lu, Dong Shen, Christopher C Harris, David J Dooling, Robert S Fulton, Lucinda L Fulton, Ken Chen, Heather Schmidt, Joelle Kalicki-Veizer, Vincent J Magrini, Lisa Cook, Sean D McGrath, Tammi L Vickery, Michael C Wendl, Sharon Heath, Mark A Watson, Daniel C Link, Michael H Tomasson, William D Shannon, Jacqueline E Payton, Shashikant Kulkarni, Peter Westervelt, Matthew J Walter, Timothy A Graubert, Elaine R Mardis, Richard K Wilson, and John F DiPersio. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature*, 481(7382):506–510, 26 January 2012.

John G Doench, Ella Hartenian, Daniel B Graham, Zuzana Tothova, Mudra Hegde, Ian Smith, Meagan Sullender, Benjamin L Ebert, Ramnik J Xavier, and David E Root. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol.*, 32(12):1262–1267, December 2014.

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

Jennifer A Doudna and Emmanuelle Charpentier. Genome editing. the new frontier of genome engineering with CRISPR-Cas9. *Science*, 346(6213):1258096, 28 November 2014.

Helene Dumay-Odelot, Stephanie Durrieu-Gaillard, Daniel Da Silva, Robert G Roeder, and Martin Teichmann. Cell growth- and differentiation-dependent regulation of rna polymerase iii transcription. *Cell Cycle*, 9(18):3687–3699, Sep 2010. ISSN 1551-4005 (Electronic); 1551-4005 (Linking). doi: 10.4161/cc.9.18.13203.

S M Dymecki and H Tomasiewicz. Using flp-recombinase to characterize expansion of wnt1-expressing neural progenitors in the mouse. *Dev. Biol.*, 201(1):57–65, 1 September 1998.

S R Eddy. Profile hidden markov models. *Bioinformatics*, 14(9):755–763, 1998. ISSN 1367-4803 (Print); 1367-4803 (Linking).

Cheng Fan, Daniel S Oh, Lodewyk Wessels, Britta Weigelt, Dimitry S A Nuyten, Andrew B Nobel, Laura J van't Veer, and Charles M Perou. Concordance among Gene-Expression– based predictors for breast cancer. *N. Engl. J. Med.*, 355: 560–569, 2006.

Fahim Farzadfard and Timothy K Lu. Synthetic biology. genomically encoded analog memory with precise in vivo DNA writing in living cell populations. *Science*, 346 (6211):1256272, 14 November 2014.

Joseph Felsenstein. PHYLIP - phylogeny inference package (version 3.2). *Cladistics*, 5: 164–166, 1989.

Kristopher K Frese and David A Tuveson. Maximizing mouse cancer models. *Nat. Rev. Cancer*, 7(9):645–658, September 2007.

Kirsten L Frieda, James M Linton, Sahand Hormoz, Joonhyuk Choi, Ke-Huan K Chow, Zakary S Singer, Mark W Budde, Michael B Elowitz, and Long Cai. Synthetic recording and in situ readout of lineage information in single cells. *Nature*, 541 (7635):107–111, 5 January 2017.

Dan Frumkin, Adam Wasserstrom, Shai Kaplan, Uriel Feige, and Ehud Shapiro. Genomic variability within an organism exposes its cell lineage tree. *PLoS Comput. Biol.*, 1(5):e50, October 2005.

Yanfang Fu, Jeffry D Sander, Deepak Reyon, Vincent M Cascio, and J Keith Joung. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat. Biotechnol.*, 32(3):279–284, March 2014.

J. Shendure G. M. Church. Nucleic acid memory device. *Patent US20030228611*, N.d. Print.

James A Gagnon, Eivind Valen, Summer B Thyme, Peng Huang, Laila Ahkmetova, Andrea Pauli, Tessa G Montague, Steven Zimmerman, Constance Richter, and Alexander F Schier. Efficient mutagenesis by cas9 Protein-Mediated oligonucleotide insertion and Large-Scale assessment of Single-Guide RNAs. *PLoS One*, 9 (5):e98186, 29 May 2014.

Molly Gasperini, Gregory M Findlay, Aaron McKenna, Jennifer H Milbank, Choli Lee, Melissa D Zhang, Darren A Cusanovich, and Jay Shendure. Paired crispr/cas9 guide-rnas enable high-throughput deletion scanning (scandel) of a mendelian disease locus for functionally critical noncoding elements. *bioRxiv*, 2016a. doi: 10.1101/092445. URL `http://biorxiv.org/content/early/2016/12/08/092445`.

Molly Gasperini, Gregory M Findlay, Aaron McKenna, Jennifer H Milbank, Choli Lee, Melissa D Zhang, Darren A Cusanovich, and Jay Shendure. Paired CRISPR/Cas9 guide-RNAs enable high-throughput deletion scanning (ScanDel) of a mendelian disease locus for functionally critical noncoding elements. 8 December 2016b.

Mel Greaves and Carlo C Maley. Clonal evolution in cancer. *Nature*, 481(7381):306–313, 19 January 2012a.

Mel Greaves and Carlo C. Maley. Clonal evolution in cancer. *Nature*, 481(7381):306–313, 01 2012b. URL `http://dx.doi.org/10.1038/nature10762`.

Maxim V C Greenberg, Juliane Glaser, Máté Borsos, Fatima El Marjou, Marius Walter, Aurélie Teissandier, and Déborah Bourc'his. Transient transcription in the early embryo sets an epigenetic state that programs postnatal growth. *Nat. Genet.*, 49(1):110–118, January 2017.

Nigel D F Grindley, Katrine L Whiteson, and Phoebe A Rice. Mechanisms of site-specific recombination. *Annu. Rev. Biochem.*, 75: 567–605, 2006.

Dominic Grun, Lennart Kester, and Alexander van Oudenaarden. Validation of noise models for single-cell transcriptomics. *Nat Meth*, 11(6):637–640, 06 2014. URL `http://dx.doi.org/10.1038/nmeth.2930`.

Tripti Gupta and Mary C Mullins. Dissection of organs from the adult zebrafish. *J. Vis. Exp.*, (37), 4 March 2010.

Vikas Gupta and Kenneth D Poss. Clonally dominant cardiomyocytes direct heart morphogenesis. *Nature*, 484(7395):479–484, 25 April 2012.

Maximilian Haeussler, Kai Schönig, Hélène Eckert, Alexis Eschstruth, Joffrey Mianné, Jean-Baptiste Renaud, Sylvie Schneider-Maunoury, Alena Shkumatava, Lydia Teboul, Jim Kent, Jean-Stephane Joly, and Jean-Paul Concordet. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol.*, 17(1): 148, 5 July 2016.

Peter M Henson and David A Hume. Apoptotic cell removal in development and tissue homeostasis. *Trends Immunol.*, 27(5): 244–250, May 2006.

Gaelen T Hess, Laure Frésard, Kyuho Han, Cameron H Lee, Amy Li, Karlene A Cimprich, Stephen B Montgomery, and Michael C Bassik. Directed evolution using dcas9-targeted somatic hypermutation in mammalian cells. *Nat. Methods*, 13(12): 1036–1042, December 2016.

Manuel Hidalgo, Frederic Amant, Andrew V Biankin, Eva Budinská, Annette T Byrne, Carlos Caldas, Robert B Clarke, Steven de Jong, Jos Jonkers, Gunhild Mari Mælandsmo, Sergio Roman-Roman, Joan Seoane, Livio Trusolino, and Alberto Villanueva. Patient-derived xenograft models: an emerging platform for translational cancer research. *Cancer Discov.*, 4(9):998–1013, September 2014.

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

Chiu-Ju Huang, Chi-Tang Tu, Chung-Der Hsiao, Fong-Jou Hsieh, and Huai-Jen Tsai. Germ-line transmission of a myocardium-specific GFP transgene reveals critical regulatory elements in the cardiac myosin

light chain 2 promoter of zebrafish. *Dev. Dyn.*, 228(1):30–40, September 2003.

Young Seok Ju, Inigo Martincorena, Moritz Gerstung, Mia Petljak, Ludmil B Alexandrov, Raheleh Rahbari, David C Wedge, Helen R Davies, Manasa Ramakrishna, Anthony Fullam, Sancha Martin, Christopher Alder, Nikita Patel, Steve Gamble, Sarah O'Meara, Dilip D Giri, Torril Sauer, Sarah E Pinder, Colin A Purdie, Åke Borg, Henk Stunnenberg, Marc van de Vijver, Benita K T Tan, Carlos Caldas, Andrew Tutt, Naoto T Ueno, Laura J van 't Veer, John W M Martens, Christos Sotiriou, Stian Knappskog, Paul N Span, Sunil R Lakhani, Jórunn Erla Eyfjörd, Anne-Lise Børresen-Dale, Andrea Richardson, Alastair M Thompson, Alain Viari, Matthew E Hurles, Serena Nik-Zainal, Peter J Campbell, and Michael R Stratton. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature*, 543 (7647):714–718, 30 March 2017.

Jan Philipp Junker, Bastiaan Spanjaard, Josi Peterson-Maduro, Anna Alemany, Bo Hu, Maria Florescu, and Alexander van Oudenaarden. Massively parallel clonal analysis using CRISPR/Cas9 induced genetic scars. 4 January 2017.

Reza Kalhor, Prashant Mali, and George M Church. Rapidly evolving homing CRISPR barcodes. 1 December 2016.

Koichi Kawakami. Tol2: a versatile gene transfer vector in vertebrates. *Genome Biol.*, 8 Suppl 1:S7, 2007.

Rongqin Ke, Marco Mignardi, Alexandra Pacureanu, Jessica Svedlund, Johan Botling, Carolina Wählby, and Mats Nilsson. In situ sequencing for RNA analysis in preserved tissue and cells. *Nat. Methods*, 10(9):857–860, September 2013.

R E Keller. Vital dye mapping of the gastrula and neurula of xenopus laevis. II. prospective areas and morphogenetic movements of the deep layer. *Dev. Biol.*, 51(1):118–137, 1 July 1976.

Charles B Kimmel and Robert D Law. Cell lineage of zebrafish blastomeres. *Dev. Biol.*, 108(1):94–101, 1 March 1985.

Jeffery M Klco, David H Spencer, Christopher A Miller, Malachi Griffith, Tamara L Lamprecht, Michelle O'Laughlin, Catrina

Fronick, Vincent Magrini, Ryan T Demeter, Robert S Fulton, William C Eades, Daniel C Link, Timothy A Graubert, Matthew J Walter, Elaine R Mardis, John F Dipersio, Richard K Wilson, and Timothy J Ley. Functional heterogeneity of genetically defined subclones in acute myeloid leukemia. *Cancer Cell*, 25(3):379–392, 17 March 2014.

Allon M Klein and Benjamin D Simons. Universal patterns of stem cell fate in cycling adult tissues. *Development*, 138(15):3103–3111, August 2011.

Benjamin P Kleinstiver, Michelle S Prew, Shengdar Q Tsai, Ved V Topkar, Nhu T Nguyen, Zongli Zheng, Andrew P W Gonzales, Zhuyun Li, Randall T Peterson, Jing-Ruey Joanna Yeh, Martin J Aryee, and J Keith Joung. Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature*, 523(7561):481–485, 23 July 2015.

Kai Kretzschmar and Fiona M Watt. Lineage tracing. *Cell*, 148(1-2):33–45, 20 January 2012.

Kornel Labun, Tessa G Montague, James A Gagnon, Summer B Thyme, and Eivind Valen. CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering. *Nucleic Acids Res.*, 44(W1): W272–6, 8 July 2016.

Dan A Landau, Scott L Carter, Petar Stojanov, Aaron McKenna, Kristen Stevenson, Michael S Lawrence, Carrie Sougnez, Chip Stewart, Andrey Sivachenko, Lili Wang, Youzhong Wan, Wandi Zhang, Sachet A Shukla, Alexander Vartanov, Stacey M Fernandes, Gordon Saksena, Kristian Cibulskis, Bethany Tesar, Stacey Gabriel, Nir Hacohen, Matthew Meyerson, Eric S Lander, Donna Neuberg, Jennifer R Brown, Gad Getz, and Catherine J Wu. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell*, 152(4):714–726, 14 February 2013.

Michael S Lawrence, Petar Stojanov, Craig H Mermel, James T Robinson, Levi A Garraway, Todd R Golub, Matthew Meyerson, Stacey B Gabriel, Eric S Lander, and Gad Getz. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484):495–501, 23 January 2014.

N M Le Douarin and M A Teillet. Experimen-

tal analysis of the migration and differentiation of neuroblasts of the autonomic nervous system and of neurectodermal mesenchymal derivatives, using a biological cell marking technique. *Dev. Biol.*, 41(1):162–184, November 1974.

Je Hyuk Lee, Evan R Daugharthy, Jonathan Scheiman, Reza Kalhor, Joyce L Yang, Thomas C Ferrante, Richard Terry, Sauveur S F Jeanty, Chao Li, Ryoji Amamoto, Derek T Peters, Brian M Turczyk, Adam H Marblestone, Samuel A Inverso, Amy Bernard, Prashant Mali, Xavier Rios, John Aach, and George M Church. Highly multiplexed subcellular RNA sequencing in situ. *Science*, 343 (6177):1360–1363, 21 March 2014.

Heng Li and Richard Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5): 589–595, 1 March 2010.

D Linder and S M Gartler. Glucose-6-phosphate dehydrogenase mosaicism: utilization as a cell marker in the study of leiomyomas. *Science*, 150(3692):67–69, 1 October 1965.

S John Liu, Max A Horlbeck, Seung Woo Cho, Harjus S Birk, Martina Malatesta, Daniel He, Frank J Attenello, Jacqueline E Villalta, Min Y Cho, Yuwen Chen, Mohammad A Mandegar, Michael P Olvera, Luke A Gilbert, Bruce R Conklin, Howard Y Chang, Jonathan S Weissman, and Daniel A Lim. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science*, 355(6320), 6 January 2017.

Zhe Liu and Philipp J Keller. Emerging imaging and genomic tools for developmental systems biology. *Dev. Cell*, 36(6):597–610, 21 March 2016.

Jean Livet, Tamily A Weissman, Hyuno Kang, Ryan W Draft, Ju Lu, Robyn A Bennis, Joshua R Sanes, and Jeff W Lichtman. Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature*, 450(7166):56–62, 1 November 2007.

Michael A Lodato, Mollie B Woodworth, Semin Lee, Gilad D Evrony, Bhaven K Mehta, Amir Karger, Soohyun Lee, Thomas W Chittenden, Alissa M D'Gama, Xuyu Cai, Lovelace J Luquette, Eunjung

Lee, Peter J Park, and Christopher A Walsh. Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science*, 350(6256):94–98, 2 October 2015.

Rong Lu, Norma F Neff, Stephen R Quake, and Irving L Weissman. Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. *Nat Biotechnol*, 29 (10):928–933, Oct 2011. ISSN 1546-1696 (Electronic); 1087-0156 (Linking). doi: 10.1038/nbt.1977.

Iain C Macaulay, Valentine Svensson, Charlotte Labalette, Lauren Ferreira, Fiona Hamey, Thierry Voet, Sarah A Teichmann, and Ana Cvejic. Single-Cell RNA-Sequencing reveals a continuous spectrum of differentiation in hematopoietic cells. *Cell Rep.*, 14(4):966–977, 2 February 2016.

Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, John J Trombetta, David A Weitz, Joshua R Sanes, Alex K Shalek, Aviv Regev, and Steven A McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 21 May 2015.

Tanja Magoč and Steven L Salzberg. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27(21):2957–2963, 1 November 2011.

F Magrangeas, H Avet-Loiseau, W Gouraud, L Lodé, O Decaux, P Godmer, L Garderet, L Voillat, T Facon, A M Stoppa, G Marit, C Hulin, P Casassus, M Tiab, E Voog, E Randriamalala, K C Anderson, P Moreau, N C Munshi, and S Minvielle. Minor clone provides a reservoir for relapse in multiple myeloma. *Leukemia*, 27(2):473–481, February 2013.

Gang Chen-Richard A. Van Etten Andrew B. Leiter Maryann Giel-Moloney, Daniela S. Krause. Ubiquitous and uniform in vivo fluorescence in rosa26-egfp bac transgenic mice. *Genesis*, 45(2):83–89, Feb 2007.

Aaron McKenna, Gregory M Findlay, James A Gagnon, Marshall S Horwitz, Alexander F Schier, and Jay Shendure. Whole-organism lineage tracing by com-

binatorial and cumulative genome editing. *Science*, 353(6298):aaf7907, 29 July 2016.

Sean G Megason and Scott E Fraser. Imaging in systems biology. *Cell*, 130(5):784–795, 7 September 2007.

Brooks E Miner, Reinhard J Stöger, Alice F Burden, Charles D Laird, and R Scott Hansen. Molecular barcodes detect redundancy and contamination in hairpin-bisulfite PCR. *Nucleic Acids Res.*, 32(17): e135, 30 September 2004.

John Alexander Moore. *Science as a way of knowing : the foundations of modern biology*. Harvard University Press, 1993. ISBN 067479480X.

John Alexander Moore. *Science as a Way of Knowing: The Foundations of Modern Biology*. Harvard University Press, 1999.

Miguel A Moreno-Mateos, Charles E Vejnar, Jean-Denis Beaudoin, Juan P Fernandez, Emily K Mis, Mustafa K Khokha, and Antonio J Giraldez. cris prscan: designing highly efficient sgrn as for cris pr-cas9 targeting in vivo.

Christian Mosimann, Charles K Kaufman, Pulin Li, Emily K Pugach, Owen J Tamplin, and Leonard I Zon. Ubiquitous transgene expression and cre-based recombination driven by the ubiquitin promoter in zebrafish. *Development*, 138(1):169–177, January 2011.

A Nagy. Cre recombinase: the universal reagent for genome tailoring. *Genesis*, 26 (2):99–109, February 2000.

Gonzalo Navarro. A guided tour to approximate string matching. *ACM Comput. Surv.*, 33(1):31–88, March 2001.

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

Y Albert Pan, Tom Freundlich, Tamily A Weissman, David Schoppik, X Cindy Wang, Steve Zimmerman, Brian Ciruna, Joshua R Sanes, Jeff W Lichtman, and Alexander F Schier. Zebrabow: multispectral cell labeling for cell tracing and lineage analysis in zebrafish. *Development*, 140(13):2835–2846, July 2013.

Ian D Peikon, Diana I Gizatullina, and An-

thony M Zador. In vivo generation of DNA sequence diversity for cellular barcoding. *Nucleic Acids Res.*, 42(16):e127, 10 July 2014.

Jason Pellettieri and Alejandro Sánchez Alvarado. Cell turnover and adult tissue homeostasis: from humans to planarians. *Annu. Rev. Genet.*, 41:83–105, 2007.

Samuel D Perli, Cheryl H Cui, and Timothy K Lu. Continuous genetic recording with self-targeting CRISPR-Cas in human cells. *Science*, 353(6304), 9 September 2016.

Randall J Platt, Sidi Chen, Yang Zhou, Michael J Yim, Lukasz Swiech, Hannah R Kempton, James E Dahlman, Oren Parnas, Thomas M Eisenhaure, Marko Jovanovic, Daniel B Graham, Siddharth Jhunjhunwala, Matthias Heidenreich, Ramnik J Xavier, Robert Langer, Daniel G Anderson, Nir Hacohen, Aviv Regev, Guoping Feng, Phillip A Sharp, and Feng Zhang. CRISPR-Cas9 knockin mice for genome editing and cancer modeling. *Cell*, 159(2): 440–455, 9 October 2014.

Fillip Port and Simon L Bullock. Augmenting CRISPR applications in drosophila with tRNA-flanked sgRNAs. *Nat. Methods*, 13 (10):852–854, October 2016.

Shaina N Porter, Lee C Baker, David Mittelman, and Matthew H Porteus. Lentiviral and targeted cellular barcoding reveals ongoing clonal dynamics of cell lines in vitro and in vivo. *Genome Biol.*, 15(5):R75, 30 May 2014.

Vijay Ramani, Xinxian Deng, Ruolan Qiu, Kevin L Gunderson, Frank J Steemers, Christine M Disteche, William S Noble, Zhijun Duan, and Jay Shendure. Massively multiplex single-cell Hi-C. *Nat. Methods*, 14(3):263–266, March 2017.

Yitzhak Reizel, Noa Chapal-Ilani, Rivka Adar, Shalev Itzkovitz, Judith Elbaz, Yosef E Maruvka, Elad Segev, Liran I Shlush, Nava Dekel, and Ehud Shapiro. Colon stem cell and crypt dynamics exposed by cell lineage reconstruction. *PLoS Genet.*, 7(7):e1002192, July 2011.

P Rice, I Longden, and A Bleasby. EMBOSS: the european molecular biology open software suite. *Trends Genet.*, 16(6):276–277, June 2000.

Jonathan N Rosen, Michael F Sweeney, and John D Mably. Microinjection of zebrafish embryos to analyze gene function. *J Vis Exp*, (25), Mar 2009. ISSN 1940-087X (Electronic); 1940-087X (Linking). doi: 10.3791/1115.

Stephen J Salipante and Marshall S Horwitz. Phylogenetic fate mapping. *Proc. Natl. Acad. Sci. U. S. A.*, 103(14):5448–5453, 4 April 2006.

Stephen J Salipante, Arnold Kas, Eva McMonagle, and Marshall S Horwitz. Phylogenetic analysis of developmental and postnatal mouse cell lineages. *Evol. Dev.*, 12(1): 84–94, January 2010.

Yasemin Sancak, Timothy R Peterson, Yoav D Shaul, Robert A Lindquist, Carson C Thoreen, Liron Bar-Peled, and David M Sabatini. The rag GTPases bind raptor and mediate amino acid signaling to mTORC1. *Science*, 320(5882):1496–1501, 13 June 2008.

Jeffry D Sander and J Keith Joung. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat. Biotechnol.*, 32(4): 347–355, April 2014.

Neville E Sanjana, Ophir Shalem, and Feng Zhang. Improved vectors and genome-wide libraries for CRISPR screening. *Nat. Methods*, 11(8):783–784, August 2014.

Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.*, 33(5): 495–502, May 2015.

Stephanie Tzouanas Schmidt, Stephanie M Zimmerman, Jianbin Wang, Stuart K Kim, and Stephen R Quake. Quantitative analysis of synthetic cell lineage tracing using nuclease barcoding. *ACS Synth. Biol.*, 10 March 2017.

Sydney M. Shaffer, Margaret C. Dunagin, Stefan R. Torborg, Eduardo A. Torre, Benjamin Emert, Clemens Krepler, Marilda Beqiri, Katrin Sproesser, Patricia A. Brafford, Min Xiao, Elliott Eggan, Ioannis N. Anastopoulos, Cesar A. Vargas-Garcia, Abhyudai Singh, Katherine L. Nathanson, Meenhard Herlyn, and Arjun Raj. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature*, 546(7658):431–435, 06

2017. URL http://dx.doi.org/10.1038/nature22794.

Sheel Shah, Eric Lubeck, Wen Zhou, and Long Cai. seqfish accurately detects transcripts in single cells and reveals robust spatial organization in the hippocampus. *Neuron*, 94(4):752–758, May 2017. ISSN 1097-4199 (Electronic); 0896-6273 (Linking). doi: 10.1016/j.neuron.2017.05.008.

Ian M Slaymaker, Linyi Gao, Bernd Zetsche, David A Scott, Winston X Yan, and Feng Zhang. Rationally engineered cas9 nucleases with improved specificity. *Science*, 351 (6268):84–88, 1 January 2016.

Stephen A Smith, Joseph W Brown, and Cody E Hinchliff. Analyzing and synthesizing phylogenies using tree alignment graphs. *PLoS Comput. Biol.*, 9(9): e1003223, 26 September 2013.

Hugo J Snippert, Laurens G van der Flier, Toshiro Sato, Johan H van Es, Maaike van den Born, Carla Kroon-Veenboer, Nick Barker, Allon M Klein, Jacco van Rheenen, Benjamin D Simons, and Hans Clevers. Intestinal crypt homeostasis results from neutral competition between symmetrically dividing lgr5 stem cells. *Cell*, 143(1): 134–144, 1 October 2010.

G S Stent. Developmental cell lineage. *Int. J. Dev. Biol.*, 42(3):237–241, 1998.

J E Sulston, E Schierenberg, J G White, and J N Thomson. The embryonic cell lineage of the nematode caenorhabditis elegans. *Dev. Biol.*, 100(1):64–119, November 1983.

John E Sulston. Caenorhabditis elegans: the cell lineage and beyond (nobel lecture). *Chembiochem*, 4(8):688–696, 4 August 2003.

Christine Thisse and Leonard I Zon. Organogenesis–heart and blood formation from the zebrafish point of view. *Science*, 295(5554):457–462, 18 January 2002.

Summer B Thyme and Alexander F Schier. Polq-Mediated end joining is essential for surviving DNA Double-Strand breaks during early zebrafish development. *Cell Rep.*, 13 April 2016.

Cole Trapnell. Defining cell types and states with single-cell genomics. *Genome Res.*, 25 (10):1491–1498, October 2015.

Barbara Treutlein, Doug G Brownfield, Angela R Wu, Norma F Neff, Gary L Mantalas, F Hernan Espinoza, Tushar J Desai, Mark A Krasnow, and Stephen R Quake. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature*, 509(7500):371–375, 15 May 2014.

Shengdar Q Tsai, Zongli Zheng, Nhu T Nguyen, Matthew Liebers, Ved V Topkar, Vishal Thapar, Nicolas Wyvekens, Cyd Khayter, A John Iafrate, Long P Le, Martin J Aryee, and J Keith Joung. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.*, 33(2):187–197, 16 December 2014.

Po-Yuan Tung, John D Blischak, Chiaowen Joyce Hsiao, David A Knowles, Jonathan E Burnett, Jonathan K Pritchard, and Yoav Gilad. Batch effects and the effective design of single-cell gene expression studies. *Sci. Rep.*, 7: 39921, 3 January 2017.

Megan van Overbeek, Daniel Capurso, Matthew M Carter, Matthew S Thompson, Elizabeth Frias, Carsten Russ, John S Reece-Hoyes, Christopher Nye, Scott Gradia, Bastien Vidal, Jiashun Zheng, Gregory R Hoffman, Christopher K Fuller, and Andrew P May. DNA repair profiling reveals nonrandom outcomes at Cas9-Mediated breaks. *Mol. Cell*, 63(4):633–646, 18 August 2016.

C Walsh and C L Cepko. Widespread dispersion of neuronal clones across functional regions of the cerebral cortex. *Science*, 255 (5043):434–440, 24 January 1992.

D A Weisblat, R T Sawyer, and G S Stent. Cell lineage analysis by intracellular injection of a tracer enzyme. *Science*, 202(4374): 1295–1298, 22 December 1978.

Richard White, Kristin Rose, and Leonard Zon. Zebrafish cancer: the state of the art and the path forward. *Nat. Rev. Cancer*, 13(9):624–636, September 2013.

JnBaptiste C.K. Lu L.Y. Kuhn C.-D. Joshua-Tor L. Wilusz, J.E. and P.A. Sharp. A triple helix stabilizes the 3' ends of long noncoding rnas that lack poly(a) tails. *Genes Dev.*, 26:2392–2407, 2012.

Addison V Wright, James K Nuñez, and Jennifer A Doudna. Biology and applications

of CRISPR systems: Harnessing nature's toolbox for genome engineering. *Cell*, 164 (1-2):29–44, 14 January 2016.

Michael B. Yaffe. The scientific drunk and the lamppost: Massive sequencing efforts in cancer discovery and treatment. *Science Signaling*, 6(269):pe13–pe13, 2013. ISSN 1945-0877. doi: 10.1126/scisignal.2003684. URL `http://stke.sciencemag.org/content/6/269/pe13`.

Linlin Yin, Lisette A Maddison, Mingyu Li, Nergis Kara, Matthew C LaFave, Gaurav K Varshney, Shawn M Burgess, James G Patton, and Wenbiao Chen. Multiplex conditional mutagenesis using transgenic expression of cas9 and sgRNAs. *Genetics*, 200(2):431–441, June 2015.

Ting Zheng, Yingzi Hou, Pingjing Zhang, Zhenxi Zhang, Ying Xu, Letian Zhang, Leilei Niu, Yi Yang, Da Liang, Fan Yi, Wei Peng, Wenjian Feng, Ying Yang, Jianxin Chen, York Yuanyuan Zhu, Li-He Zhang, and Quan Du. Profiling single-guide RNA specificity reveals a mismatch sensitive core sequence. *Sci. Rep.*, 7:40638, 18 January 2017.

Dawn L Zinyk, Eric H Mercer, Esther Harris, David J Anderson, and Alexandra L Joyner. Fate mapping of the mouse midbrain–hindbrain constriction using a site-specific recombination system. *Curr. Biol.*, 8(11):665–672, 21 May 1998.

5.3 Vita

Aaron McKenna was born into a loving family in Bitburg, Germany, where his father was stationed as a psychiatrist. After a brief stop in Mississippi, the rest of his childhood was spent in northern Vermont. After high school he attended Worcester Polytechnic Institute (WPI), a small and very old engineering school in Worcester, Massachusetts. At WPI he applied himself sporadically, but made great friends, some of whom were critical in the GESTALT project. After graduating from WPI, he took an engineering job for an intelligence contractor in Boston, which was not very exciting, but it paid well and he got to see the Red Sox win a World Series. To keep himself busy, Aaron got a masters degree in bioinformatics from Boston University. This degree got him in the door at the Broad Institute of Harvard and MIT, where he fell in love with computational biology and genetics. At the Broad he applied himself a bit too much, but it was an exciting time and place, and he had a lot of fun.

When Aaron's not having fun at work, you can find him wrestling with his dog, hanging out with friends, having a beer with his wife, or crying gently on a rock climb somewhere in the Cascades.