High-resolution phylogenetic lineage recording with CRISPR base editors

Thesis by

Duncan Matthew Chadly

In Partial Fulfillment of the Requirements for

the degree of

Doctor of Philosophy, Bioengineering

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2026

(Defended September 8, 2025)

Duncan Matthew Chadly
ORCID: 0000-0002-8417-1522

ACKNOWLEDGEMENTS

There are far too many people to thank than I can hope to acknowledge here. But that won't stop me from trying.

I want to thank my scientific heroes:

Annette Erbse, for teaching me everything an experimentalist could ever want to know (starting with how to hold a pipette); Brian Ziemba, Pete Slivka, Keane Piasta for their advice and mentorship; Joseph Falke, for inviting me to do research as an excitable freshman;

Kelsey MacConaghy, for patiently explaining the Donnan potential to me innumerable times and having me along for part of her own PhD adventures; Joel Kaar and Mark Stoykovich for thoughtful discussions and support of my undergraduate research thesis;

Joseph Wood, for his constant support and for teaching me how to interact as part of a larger scientific team; Pauline Che and Adeyma Arroyo for kindly putting up with a nervous intern;

Akihiro Matsuoka, for teaching me about stem cells and medicine, giving me opportunities to grow in every way, always believing in me, and countless delightful chats and meetings; my fellow crewmates on the USS Spiral Ganglion: Zach Morrissey, Shun Kobayashi, Kevin Nella, Charles Miller, Jennifer Best, Cong Ran, Andrew Oleksijew, Kyle Coots, Jose Fernandez;

Erik Jue, my first mentor at Caltech, for your friendship and advice throughout the years; Rustem Ismagilov, for having me as a rotation student and for numerous chats;

Nico Pelaez, for taking me on as a rotation student and helping develop a project that would become (in?)famous among the Elowitz lab BMP subgroup; Mike Flynn and Jan Gregorowicz, who began and will end this journey alongside me; Amjad Askary and Alejadro Granados, for their friendship, mentorship, and countless hours debating the merits of varied lineage recording strategies; Grace Edmonds and Mark Budde, for friendship, ice cream, hot wings, advice, and fun chats; Kirsten Frieda and Chen Gui, for teaching me the ways of the microscope, fluorescence in situ hybridization, and how to layer codes on codes on codes; Martin Tran, for voyaging with me through the sea of MEMOIR, constantly insightful questions, and being a wonderful conference companion;

Felix Horns, for thoughtful feedback, exciting discussions, and your delightful optimism; Heidi Klumpe and Christina Su, my BMP role models; Rachael Kuintzle and Matthew Langley, for pushing me to do what I feel is right inside and outside of the lab; Yodai Takei, for creative conversations and collaborations; Roni Hadas, for his friendship and collegiality during my descent into PhD madness, as well as for knowing when to bring out the honey badger mode; Leslie Klock, for your constant enthusiasm, organization, and unwavering support of the projects described herein (they would not be what they are without you);

Yaron Antebi, Lucy Chong, Yitong Ma, Ron Zhu, Sheng Wang, Fangyuan Ding, Xiaoxing Gao, Pulin Li, Joseph Markson, Zibo Chen, Ben Emert, Rongrong Du, Kaiwen Luo, Sayaka Kozuki, Lukas Moeller (the finest foosball player I have ever seen), Evan Mun, Jacob Parres-Gold, Victoria Tobin, Andrew Lu, Maire Gavagan, Bo Gu, Dhiraj Indana, Dongyang Li, Gal Manella, Ethan Richman, Judy Shon, Shiyu Xia, Maria Cabrera, Megan Holmes, Michaela Ince, Brice Hendrickson, Jordan Lay, Ali Diaz, Haley Larsen for the pleasure of our interactions and the honor of working alongside you;

Leah Santat and James Linton, for everything you do- if Michael is the head of the lab, you are its heart and soul;

Jo Leonardo, for years of critical and underappreciated support behind the scenes; Rui Malinowski and Anthony Vasquez, for your encouragement and support in the last legs of my journey;

Inna Strazhnik, for helping develop beautiful figures and explaining the ins and outs of color and design;

Sisi Chen, Jeff Park, and Brian Williams for fruitful discussions and facilitating many of the sequencing-based aspects of this work;

Jeffrey Liang, Verona Yue, Margarita Artiukova, and Maggie Sui for selecting me as one of your mentors and for your hard work, some of which is reflected in this document. It was delightful getting to know you and a privilege to work with you;

Remco Bouckaert, for your patient assistance with these exceptionally long-term projects and being willing to read my million-page emails; Fantastic collaborators I have had the opportunity and pleasure to work with and learn from: Bryson Choy, Mary Vergel, Sunny Wang, Rong Lu, Jean-Paul Urenda, Carla Liaci, Giorgia Quadrato, Joaquin Navajas Acedo, Wei Chen, Carsten Tischbirek, Lynn Fang, Kate Cavanaugh, Hannah Greenfeld, Gianluca Amadei, Magda Zernicka-Goetz, Amy Chow, Joanna Jachowicz;

A special thanks to Yoav Mayshar, Ilana Taieb-Hagerty, Netta Reines, Markus Mittnenzweig, Yonatan Stelzer, and all the members of the Stelzer and Tanay labs for inviting me to visit, work with you, and share my research at the Weizmann Institute;

The members of my thesis committee: Long Cai, for insightful suggestions and pushing me to work towards a rewarding career; Carlos Lois, for creative solutions to unexpected problems and persistent optimism (a special thanks to you and Long for attending the MEMOIR meetings and offering your feedback throughout the years); Matt Thomson, for thoughtful questions and arriving to my committee meetings early to chat about concerts and music; Mitch Guttman, for your willingness to help in all ways and exciting lines of thought;

Finally, my advisor Michael Elowitz. I am eternally grateful for the opportunity to work with you, the infectious enthusiasm you have for scientific investigation, and your unending curiosity.

Thank you to my family for encouraging my interest in science, for taking me to soccer games and violin lessons, and for your loving support; to my Mom, Marcia Chadly, and stepmom, Eva Morrell, for encouraging a lifelong interest in cooking which is at least partially responsible for all of this; to my dad, Pete Chadly, for confusingly providing me with a second stepmom named Eva (and for teaching me about many of the important things in life); my second stepmom Eva Mart, for fun games of pinochle and teaching me about karjalanpiirakka; my sisters, Erin Chadly and Monica Morrell, for their companionship and kindness; to my grandmother, Shirley Miller, whose bravery and empathy inspire me every day; Steve and Rene' Chadly for card games and pleasant mini golf evenings; Tammy and Jon Jelen, for pool parties, pinochle, and trips to the farm; Chris Jelen, for many a New Years Eve watching all three Star Wars films and for being a brilliant role model from a young age; Eric Jelen, for countess hours of Heroes of Might and Magic and Guitar Hero; Steve Jelen, for always making sure everyone is doing well; Ashley and Jordan Emery, for youthful companionship and modern day Dungeons and Dragons;

To my grandparents who passed on; my grandfather, Max, who taught me to work hard, to be happy, to always think critically, to learn at least one thing from everyone you meet; my grandmother, Virginia, whose mischief and adventurous nature live in my heart. I miss you both constantly, I hope you would be proud.

To Gigi Lambert and Quinn Quintana, fast friends and kindred spirits; Tony Quintana, Valerie Nielsen, and Izzy Quintana, for good food and good company;

Thank you to Logan Wan, Jeffrey Won, Vignesh Muralidharan, and Brad Reynolds, for continued friendship, discussion, and debates (they say once you've been friends for 7 years, you'll probably be friends for a lifetime- I guess you guys are stuck with me for the next couple of those); Max Golden, for Costco runs, board games, enjoying my undercooked meats before the inevitable poisoning, plum wine, and a long line of prior escapades (going on four lifetimes for you); Sam Gansfield and Nate Marx, for many a night of whimsy, music, or both; David Wynne and John Daniel Drumheller, a fine couple of card-carrying plasma physicists; Sam Jiles, may we meet again soon; Brian Tamura and Eliott Farquharson, a couple of dreamers if ever I saw them; Adam Safadi, I eagerly await the fulfillment of your first grade promise to hire me as national scientific advisor in exchange for my vote in your forthcoming presidential campaign;

Thank you to my partner in crime, Mercedes Quintana. Whether it's going out for coffee, knitting together at home, or traveling to the nearest ostrich farm to discuss how kindly the emus are and how rude the 'strichs, I wouldn't trade our time together for the world. I eagerly await our next set of adventures.

Last but not least, to you, the one I have neglected to mention thus far. Come find me, I'd love to buy you a drink and catch up.

From the bottom of my heart, thank you. I wouldn't be here today without you.

ABSTRACT

Dividing and differentiating cells form exquisitely organized structures across every facet of multicellular life. If we could measure the complete history of cells as they divide, change transcriptional state, and move spatially, we could address critical questions about stem cell differentiation, development, and the onset of disease. However, determining cellular ontologies is challenging except in rare cases where continual optical access is possible. Base editing technology enables the generation of stochastic, heritable mutations into genomic DNA while cells grow and divide. Comparing mutation patterns between cells allows inference of their lineage relationships in a manner analogous to evolutionary phylogenetic reconstruction. Here, we present two phylogenetic recording systems that enable high resolution lineage reconstruction over long time scales. In the first system, termed baseMEMOIR, we introduce a multiplexed, genomically dispersed set of editable targets that can be read out by imaging in situ. This system preserves spatial organization of cells and is compatible with downstream transcriptional measurements. In the second system, which we term the hypercascade, we take advantage of the predictability of A-to-G base editing to create a system in which edits not only alter bases but also generate new editable target sites in synthetic sequences. This behavior linearizes the rate at which mutations accumulate, improving lineage reconstruction. These methods enable analysis of temporal dynamics in diverse biological contexts.

PUBLISHED CONTENT AND CONTRIBUTIONS

Chadly, D. M., et al. (2024). "Reconstructing cell histories in space with image-readable base editor recording". BioRxiv. doi: 10.1101/2024.01.03.573434.

D.M.C. participated in experimental ideation, design, execution, data analysis, and writing of the manuscript.

Tischbirek, C., et al. (2025). "Synaptic MEMOIR: mapping individual synapses of neurons with protein barcodes". BioRxiv. doi: 10.1101/2025.11.25.690442.

D.M.C. participated in data analysis and editing the manuscript.

Takei, Y., et al. (2025). "Genome-wide chromatin recording resolves dynamic cell state changes". BioRxiv. doi: 10.1101/2025.08.05.668773.

D.M.C. participated in sequencing data processing, analysis, and editing the manuscript.

Askary, A., et al. (2020). "In situ readout of DNA barcodes and single base edits facilitated by in vitro transcription". Nature Biotechnol. Doi: 10.1038/s41587-019-0299-4.

D.M.C. participated in sequencing data processing, analysis, and editing the manuscript.

Additionally, I would like to recognize the following people for significant contributions to the work described herein:

Kirsten Frieda – Designing the original baseMEMOIR implementation and engineering the initial cell line.

Ron Hadas – Performing all mouse work (with Shoma Nakagawa) and tissue sectioning; Choosing gene markers for in vivo experiments; Annotating cell types for tetraploid embryos; Hypercascade experiments beyond those described here.

Shoma Nakagawa – Performing all mouse work (with Ron Hadas).

Leslie Klock – Assistance with many of the experiments and analyses.

Martin Tran – Performing motif analysis for the baseMEMOIR project; Providing experimental support.

Remco Bouckaert – Developing the 'irreversible' model for BEAST2 and providing advice related to lineage reconstruction.

Jeffrey Liang – Assistance with development of 3D baseMEMOIR analysis scripts; Working with me on imaging experiments related to Supplemental Figure 3 and Figure 6.

Chen Gui – Early experimental support and ideation.

Jiahe (Verona) Yue – Hypercascade experiments beyond those described here.

Margaret Sui – Assistance with development of baseMEMOIR barcode analysis scripts.

Yodai Takei – Providing reagents, experimental support, and ideation.

Amjad Askary – Early experimental support and ideation related to the hypercascade project.

Alejandro Granados – Early ideation related to the hypercascade project, and for selecting endogeneous gene targets for the Patski study.

Inna Strahznik – Developing artistic and beautiful figures.

Carlos Lois – Experimental ideation and feedback.

Long Cai – Experimental ideation and feedback.

Michael Elowtiz – Experimental ideation and feedback, and for support writing the manuscript.

TABLE OF CONTENTS

Acknowledgements……...iii-vi
Abstract ……vii
Published Content and Contributions…….....viii-ix
Table of Contents…… x
List of Illustrations and/or Tables……vii
Chapter I: Introduction 1
Multicellular organisms arise from a lineage tracing back to a single cell.. 1
Some organisms have a deterministic lineage structure..... 1
Most animals exhibit complex developmental flexibility 3
Targeted DNA mutations enable lineage tracing between single cells..... 6
Summary 7
Chapter II: Reconstructing cell histories in space..... 9
Introduction..... 9
Base editing can enable lineage recording with spatial readout 12
Induction drives editing into diverse mutational states 15
Imaging recovers edited barcode sequences..... 17
Dynamic barcodes are accurately classified by Zombie-FISH..... 21
Simulations show that baseMEMOIR can accurately reconstruct detailed lineage trees 22
BaseMEMOIR reconstructs lineage trees in mESC colonies 24
BaseMEMOIR recovers lineage relationships, cell states, and spatial relationships in mESC colonies 29
BaseMEMOIR is portable to in vivo systems 31
Phylogenetic analysis reveals cellular relationships during embryogenesis 36
Phylogenetic and spatial distances correlate for some cell types at E7.5 .. 37
Discussion and conclusions..... 39
Methods 42
Supplemental information 67
Chapter III: Regenerative base editing enables deep lineage recording 87
Introduction..... 87
The hypercascade design allows sustained recording through generative editing 90
Simulations show that the hypercascade enables more accurate lineage reconstruction than a simple array 95
Different hypercascade sequences operate orthogonally with distinct kinetics 98
One-shot transfection of the hypercascade editing system reveals broad clonal features..... 102

Hypercascades have the potential to record chromatin transition dynamics 105
Discussion and conclusions 111
Methods 114
Supplemental information 128
Bibliography 142

Chapter 1

INTRODUCTION

Multicellular organisms arise from a bifurcating lineage dating back to a single cell

Most multicellular organisms begin as a single fertilized egg, which undergoes repeated rounds of cell division, transcriptional differentiation, and spatial organization to become a unified adult comprised of anywhere from several (for example, the four-celled algae *Tetrabaena*) up to quadrillions of cells (in the case of the blue whale). These cells cooperate to enable sophisticated functions and adaptation to environmental niches. Understanding the self-assembly process of multicellular life and the extent to which developmental blueprints are encoded into that initial cell are fundamental questions that have fascinated biologists since the conception of the field.

Some organisms have a deterministic lineage structure

Several invertebrate organisms, including ascidian[1–3] and nematode[4–7] species, have tightly controlled lineage relationships between individual cells. Cell divisions happen in a stereotyped manner across all individuals, generating adult animals with an identical number of cells that are positioned to form identical structures. Early studies took advantage of this feature, as well as the translucent nature of these embryos, to manually track cell division events either throughout the entire developmental process of the organism[4] or until visual tracking of individual cells became impossible[1,2].

In these cases, a combination of asymmetric cell division and specific intercellular interactions underlie the observed lineage determinism[8]. For example, PAR proteins in *C. elegans* generate spatial asymmetries within cells prior to division, leading to differences in transcriptional content among their cellular progeny[9,10]. This type of development has sometimes been called "mosaic", to contrast a "regulatory" form of development which is dominated by cell-cell interactions[11].

Even so, no known embryo truly represents the platonic ideal of either mosaic or regulative; all cases are instead a combination of these two archetypal modes of development[11,12]. For example, even in organisms with deterministic lineage relationships, extrinsic signaling can be critical to differentiation and correct spatial organization. Tail muscle specification in the ascidian embryo is driven by an asymmetric division partitioning the Macho-1 gene, which triggers a transcriptional cascade culminating in tail muscle fate[11]. However, differentiation of a subset of cells that contain Macho-1 is tempered by short range intercellular signaling mediated by FGF9[11]. Similarly, multiple intercellular signaling pathways interact to enable *C. elegans* vulval development, including EGF, LIN-12/NOTCH, and Wnt[8,13]. Even in these cases, due to tight control of the interactions at play, each organism exhibits a deterministic lineage. It is therefore important not to conflate a deterministic developmental lineage with purely cell intrinsic specification of fate.

Interestingly, deterministic lineage is not a feature of all nematode species. Related species, including *T. diversipapillatus* and species of order *Enoplida*, form a loosely-organized blastocele in early development with interindividual variability from the early stages, rather

than the stereotyped asymmetrical cell divisions characteristic of *C. elegans*[14–16]. This dramatic deviation in lineage structure nonetheless yields robust worms with body plans and organs similar to those of the deterministic nematodes. Formation of a blastocele prior to germ layer specification is more common across the animal kingdom, and the species described here are anciently related to nematodes that undergo deterministic development; it thus appears that stereotyped development is an evolutionary adaptation, potentially enabling more rapid embryo formation[15,16]. Further comparative study across nematode species could reveal genetic circuits controlling early developmental trajectories, with exciting implications for the growing field of synthetic morphology[17,18].

Most animals exhibit complex developmental flexibility

Vertebrate, including mammalian, development typically generates organisms with considerable variability in size and cell number across individuals within a species. From early stages, development is plastic and can accommodate dramatic perturbations. At the same time, healthy individuals form broadly similar structures throughout development and robustly accomplish similar functions. How are multicellular structures formed and maintained in this flexible context?

Part of the answer to this question is progressive restriction of cell fates, sometimes called cell-fate commitment. Initially, all cells in the early embryo have the potential to generate progeny of all final cell types in the adult organism. However, this developmental potential is progressively stymied, with progeny being able to take on smaller and smaller subsets of fates until they can take on no more, reaching a state of terminal differentiation. Waddington

likened this process to a ball rolling down a hill with channels carved into it[19]; at branch points, the ball will take one channel and be cut off from the others. The landscape itself, he supposed, was controlled by the expression levels of underlying networks of genes[19]. This is not far from modern understanding of the differentiation process, where cell states are understood as attractors in a dynamical gene expression space[20–22].

The paradigm of differentiation as generation of and movement between attractor states can explain some features of stochastic developmental processes. For example, mouse embryos begin as a blastocyst composed of an outer layer of cells which go on to become the placenta and an inner cell mass (ICM) which forms all other future cell types[23]. In the earliest fate decision within the preimplantation mouse embryo, ICM cells specify into either the epiblast fate, which will go on to populate the tissues of the adult organism, or the primitive endoderm, a supportive extraembryonic cell type. In contrast to the stark control mediated by asymmetric cell division that we see in *C. elegans*, fate choice in this context is driven by a mutually repressive interaction between two genes, Gata6 and Nanog, further modulated by extracellular Fgf4 signaling[23–25]. This genetic circuit serves to amplify small, stochastic initial differences in Gata6/Nanog, robustly generating populations of cells with the desired proportions by leverage heterogeneous noise.

Embryonic development additionally features stereotyped cell migration events, where populations of cells from one region travel across the embryo to populate different areas, often while simultaneously differentiating. In mouse development, gastrulation begins with epiblast cells generating a region known as the primitive streak. Cells at the primitive streak

then undergo an epithelial-to-mesenchymal transition and migrate large distances toward both the proximal and anterior portions of the embryo, a process known as ingression[26]. The timing of ingression is thought to dictate the fate of cells leaving the streak, with early cells generating extraembryonic mesoderm at the proximal end of the embryo and later cells generating more distal mesodermal fates and the definitive endoderm layer[26]. These decisions are at least partially the consequence of extracellular Nodal signaling during streak ingression[27].

How migration interacts with cellular lineages, and the extent to which cellular lineage plays a role at different phases of development, remain open questions. Seminal clonal tracing studies used injected labeling techniques to track the progeny of single cells labeled at approximately the onset of primitive streak formation (E6.7) through the onset of gastrulation, to the late streak stage (E7.5)[28,29]. Widespread mixing of cells was observed in the epiblast during this critical period, and intriguingly descendants of a single cell could be found widely spread and present in different germ layers. In fact, only 56% of epiblast cells at this stage had progeny confined to a single germ layer, suggesting a limited role for cell intrinsic bias at this stage[29]. More recent timelapse light-sheet microscopy has confirmed mixing and migration of the epiblast, primitive streak, and mesoderm during these critical stages[30,31].

Although we know broadly that cells progressively reduce their capacity to form diverse cell fates through development, encounter signals driving spatial positioning, and control their population size to meet the functional needs of the organism, it isn't clear how these

phenomena integrate with the bifurcating cellular lineage underlying each organism. It has not been possible thus far to capture complete lineage histories for opaque animals with billions to trillions of cells, such as mice or humans. This capability could allow us to understand the extent and timing of fate determination across development and to tease apart the effects of the intrinsic and extrinsic factors that cells use to assemble themselves into functional organisms.

Targeted DNA mutations enable lineage tracing between single cells

Recent advances in genomic editing enable a new strategy for recovering detailed cellular lineage relationships: cell divisions can be genetically recorded and recovered after the fact[32–45]. Much as mutations in DNA sequences enable recovery of the phylogenetic relationships between different organisms at the evolutionary level, heritable marks made on the timescale of cell division encode the lineage history between single cells.

Somatic mutations have potential in this regard[46–49]. However, these methods are limited by the ability to recover mutations at scale: somatic mutations occur sparsely across the entire genome, and thus the entire genome must be sequenced to identify mutated regions. Currently, whole genome sequencing of single cells is technically challenging and inefficient, especially when attempting to combine with other modalities like transcriptional profiling. Lineage tracing based on mitochondrial mutations[50–52] is also promising, especially since these are frequently recovered as a byproduct during single cell RNA sequencing[50], however inference is hampered by the fact that multiple mitochondria are present in the same cell and may have distinct genomic sequences (heteroplasmy).

To address these challenges, recent work makes use of exogenously introduced proteins, typically CRISPR effectors or integrases, to write mutations to relatively short pieces of DNA that can either be expressed on a single transcript or amplified directly from the genome using the polymerase chain reaction (PCR)[32–45]. Most methods to date have focused on barcodes that can be captured through sequencing[32–37,39–43], which requires sample dissociation and loss of spatial resolution. Using these techniques, researchers have explored clonal dynamics in processes ranging from cancer metastasis[34] to zebrafish embryogenesis[32].

Spatial relationships are critical to understanding a variety of processes, especially organismal development. To that end, some groups have constructed phylogenetic lineage tracing systems where mutations are designed to be recovered directly from microscopic imaging methods[38,44,45]. Early work was limited in the number of available mutable sites, but nonetheless was able to tease apart relative impacts of intrinsic and extrinsic cellular signals during drosophila brain development[45]. Recent work has built on these early examples to recover more detailed and accurate lineage trees by dramatically multiplexing the number of target sites present in the genome[38]. It is a matter of time until these systems are sufficiently powered to fully record lineage relationships across the entirety of mouse development and beyond.

Summary

Organismal development is variable amongst species and incompletely understood. Timelapse microscopy, clonal tracing, and transcriptional sequencing have shed light on specific cases, but we don’t have a broad picture of development except in limited contexts.

Phylogenetic recording strategies have the potential to record detailed lineage histories that can be recovered at scale after the fact in snapshot measurements, however they have been limited in depth and accuracy of reconstruction. To address these issues, we developed two new lineage tracing systems with unique capabilities. Chapter 2 describes the design of a highly multiplexed dinucleotide editing system, termed baseMEMOIR, which contains 792 bits of writable memory enabling detailed lineage reconstruction without disrupting the spatial context of cells. We demonstrate application of baseMEMOIR to in vivo mouse development, recording lineage relationships in gastrulating embryos. Chapter 3 describes the hypercascade, a strategy for linearizing the kinetics of phylogenetic recording through generative editing. We suggest through simulations that hypercascade editing has the potential to record not only lineage relationships, but also dynamic epigenetic transitions at the single cell level. We envision that both systems will have their place in addressing biological questions in the years to come.

Chapter 2

RECONSTRUCTING CELL HISTORIES IN SPACE WITH IMAGE-READABLE BASE EDITOR RECORDING

Adapted from:

Chadly, D. M., et al. (2024). "Reconstructing cell histories in space with image-readable base editor recording". BioRxiv. doi: 10.1101/2024.01.03.573434.

Introduction

Cells divide, differentiate, and move to form exquisitely organized structures. Reconstructing the dynamic histories of individual cells, particularly their lineage relationships, could enable researchers to understand how tissues form, analyze the roles of intrinsic and extrinsic determinants of cell fate decisions, and reveal how processes are dysregulated in disease[53]. Recent advances in single cell sequencing and spatial genomics now allow us to capture single cell states at specific moments in time[54–56]. However, with a few exceptions[57], the histories of those cells have largely remained hidden.

Researchers have sought to address this challenge through engineered recording systems, which progressively introduce stochastic edits in genomically integrated barcode sequences as cells proliferate. Systems such as GESTALT[32–34], CARLIN[35], LINNAEUS[36], SMALT[37], and the homing CRISPR barcoded mouse[39,40], use CRISPR/Cas9 or recombinases to edit designed target sequences, relying on next generation sequencing to read out edited barcodes. Alternative systems, including CAMERA, leverage CRISPR base editors to generate more

specific types of barcode diversity.[58,59] Further, prime editors introduced an additional paradigm for phylogenetic recording, sequentially inserting short nucleotide motifs for genomic information storage.[41–43] In all of these systems, lineage relationships between individual cells are reconstructed from each cell's unique pattern of target site edits, in a manner analogous to sequence-based phylogenetic reconstruction.[60,61]

These techniques can be powerful in their ability to recover lineage but disrupt spatial organization. A parallel set of methods were developed to allow barcode editing in ways that allow readout of edits and cell states by imaging[44,45]. For example, previous MEMOIR (Memory by Engineered Mutagenesis with Optical In situ Readout) systems showed that it is possible to stochastically and irreversibly edit engineered DNA barcodes, or 'scratchpads,' using CRISPR/Cas9[44] or an integrase[45], and then read out those edits by imaging. However, these methods were limited in accuracy by relatively low numbers of mutable target sites, which serve as memory in the genome, typically providing at most 16 bits of information storage[45]. Recently, a prime-editing-based lineage tracing system compatible with imaging readout demonstrated a system with roughly 30 mutable target sites per cell, increasing theoretical information storage to approximately 270 bits per cell by increasing the number of mutation outcomes[38]. Although impressive, limitations in barcode recovery during readout and issues of homoplasy during lineage reconstruction will require additional memory if we are to obtain lineage relationships with perfect accuracy at the whole-organism level[38].

Previously, we showed that in situ T7 transcription can amplify genomic DNA into localized RNA clusters, which can then be competitively probed to discriminate single base edits[62]. In

this strategy, termed “Zombie,” a genomic DNA of interest can be maintained without transcription in live cells, transcribed after fixation with the addition of T7 polymerase, and finally detected by RNA-FISH. Zombie transcription avoids silencing problems that occur when barcodes must be continuously transcribed in live cells, generates large quantities of RNA that spatially localize around the active site of transcription, can detect mutations at single nucleotide resolution, and is compatible with subsequent sequential rounds of FISH to detect endogenous gene transcripts. Zombie enables readout of dense editable memory arrays, expanding the capacity of MEMOIR approaches.

Here, we introduce “baseMEMOIR”, a multiplexed phylogenetic recording system which enables detailed recording of lineage relationships over time in a manner compatible with recovery of spatial position and gene expression patterns **(Figure 1A)**. We distributed mutable synthetic DNA sequences at high copy number randomly across the genome of mouse embryonic stem cells. These targets can be edited by the CRISPR A-to-G base editor (ABE), which complexes with guide RNAs to specifically mutate target sites within our synthetic sequences. For tight control of editing, we used two inducible systems to control the base editor (the TRE3G Tet-On system) and guide RNAs (either controlled by a Wnt responsive promoter or the TRE3G system) respectively. On induction, mutations occur at a rate commensurate with cell division and are passed down from parent cells to their progeny through DNA replication, creating lasting marks that link related cells to one another. We then recovered mutation states through microscopic imaging and applied Bayesian phylogenetic tools to infer lineage relationships as well as transcriptional state dynamics and spatiotemporal histories in a unified manner. By comparing the distinct pattern of mutations

in each cell after a series of divisions, we were able to reconstruct phylogenies and estimate uncertainty in both tree topology and the timing of past cell divisions.

To demonstrate the capabilities of baseMEMOIR, we applied this system to estimate state switching rates and probable past cell states along lineages of dividing mESCs grown in serum-LIF conditions in the presence of a Wnt agonist. We find that mESCs grown in these conditions undergo reversible transitions between formative and 2C-like states, with an intermediate naive state that can be broken up into three distinct subclusters. Each state and subcluster, in addition to being transcriptionally distinct, is further distinguished by a set of allowed state transitions. Spurred by these results, we applied baseMEMOIR to identify lineage relationships up to E7.5 during mouse development in vivo using tetraploid complementation. The baseMEMOIR cell lines and platform can be applied, and further scaled, in any model system that permits genetic engineering, opening up spatially resolved analysis of embryonic development and other processes.

Base editing can enable lineage recording with spatial readout

To ultimately capture detailed lineage relationships between cells while maintaining spatial context, we first sought to design an image-readable stochastic base editing system. One possible system would use designed target sequences that would be editable at a single base. For example, the A-to-G base editor ABE could target a set of defined sequences to stochastically edit each target site. However, this scheme is susceptible to convergent edits in unrelated cells, i.e. homoplasy. In the limit of complete editing, every editable A would be converted to a G in all cells, and no lineage information could be recovered.

To circumvent this issue, we used a modified design which takes advantage of the ability of the ABE to stochastically edit target sequences into one of multiple stable outcomes[63]. In our case, AA dinucleotide sequences in the target window are converted to any of three edited end-point states (GA, AG, or GG) (**Figure 1B**). Critically, because the GA and AG states each disrupt binding to the base editor gRNA, they are not expected to undergo further editing to GG. This dinucleotide editing scheme in principle reduces the likelihood of convergent editing and prevents the effective "erasure" of recorded information at long times.

Based on this principle, we designed a library of barcoded, editable target arrays that could be integrated into the genome (**Figure 1C**). Each target array contained 6 tandem editable target sites, with unique protospacer sequences outside of the AA dinucleotide, so that each of the six target sites required a distinct gRNA sequence for editing (**Figure 1C**). The arrays were flanked with piggyBac inverted terminal repeats, to enable high-copy-number genomic integration. To distinguish different integrations from one another, we also incorporated two static (non-editable) random barcodes: first, a 10bp barcode (10^6 variants), for compact readout by sequencing; second, a pair of static 80bp image-readable barcode sequences, each of which could take on one of 200 possible sequences, for a total diversity of $200^2 = \sim 10^4$ unique barcodes (see Methods for construction of plasmid libraries). Finally, to enable imaging-based "Zombie" readout of edits, the arrays incorporated a T7 promoter upstream of the editable array (**Figure 1C**)[62].

To mutate targets at a tunable rate, and create the potential for signal recording, we built constructs that allow inducible expression of the ABE and gRNAs. ABE expression was placed under doxycycline (dox) control using the Tet-ON system, by stably integrating the reverse tetracycline-controlled transactivator (rtTA) (**Methods**)[64]. gRNAs were also made Wnt-inducible by expressing them from the 3' UTR of an mTurquoise reporter gene under the control of a Wnt-responsive element[44,65] (WRE). This promoter is active only in the presence of Wnt signaling ligands, and can be driven by the small molecule GSK-3 inhibitor CHIR99021 (CHIR)[44,65]. To generate fully functional gRNAs after expression, we flanked each gRNA with previously described ribozyme sequences[66]. After stable co-integration in mESCs using piggyBac transposition, we identified a clone, termed baseMEM-01, which contained 66 genomically-dispersed array copies with diverse static barcodes (**Figure 1D, Supplemental Figure 1**). This cell line allowed recording in live cells with imaging-based readout of barcode base edits, static barcode sequences, and endogenous gene expression using multiple rounds of hybridization and imaging[54–56] (**Figure 1E**).

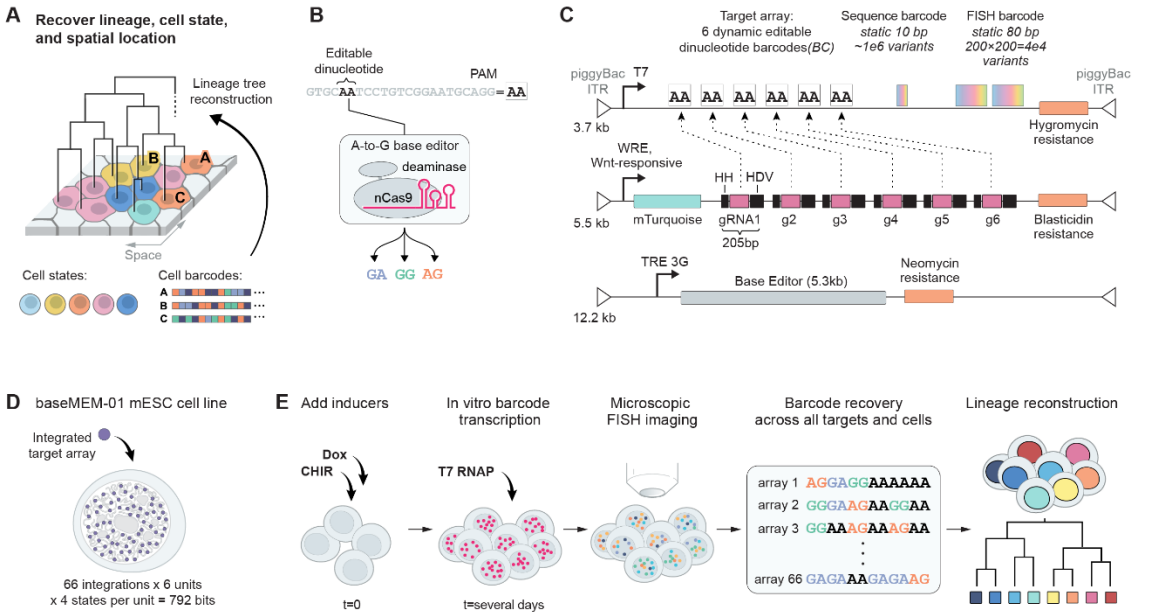


Figure 1: Multiplexed, genomically dispersed, editable barcodes enable detailed recording of lineages over many generations with in situ readout. (A) Detailed lineage trees can be measured alongside transcriptional cell states while maintaining spatial context through phylogenetic barcoding. **(B)** Predicted stochastic editing of AA dinucleotides results in one of three terminal outcomes. **(C)** An inducible barcode editing system can be integrated into cells at high copy number via piggyBac transposase. Target arrays (top) contain six AA dinucleotides flanked by unique protospacer sequences as well as sequencing and imaging-readable static barcodes which serve to uniquely mark different genomic integrations of the array. Editing is induced by expression of guide RNAs (middle), controlled by a Wnt-responsive element, and base editor (bottom), controlled by the TRE3G tet-on promoter. **(D)** We engineered a monoclonal mESC line containing 66 uniquely labeled target array copies (396 editable dinucleotides, or 792 bits of information) alongside the inducible editing machinery. **(E)** This cell line enables genomic recording and recovery through FISH imaging and phylogenetic tree inference.

Induction drives editing into diverse mutational states

Since lineage recording depends on edits being accumulated on the timescale of cell division, we next sought to analyze inducible base editing using the system. We exposed baseMEM-01 mESC cultures to the inducers CHIR, doxycycline, or both, over an 8-day period, a timescale long enough to allow multiple stem cell generations and, in an embryonic context, approach gastrulation. We collected samples at multiple time-points (**Figure 2A**). We then analyzed the editable barcodes by next generation sequencing.

Edits accumulated at all six target sites (**Figure 2B**). Without induction and in the absence of dox, some editing was observed. However, the fraction of such background edited sites generally remained constant during the time course, consistent with transient background editing during cell line construction, but minimal basal editing in stable clones. By contrast, in the presence of both CHIR and dox, edits accumulated rapidly at a rate that was well-fit by a model of editing with distinct edit rates at each site (**Figure 2B**, solid lines). Interestingly, in the presence of dox alone, editing still occurred, albeit at an attenuated rate. This non-zero edit rate could be due to basal transcriptional activity of the WRE promoter or to activation of gRNA expression by low levels of endogenous Wnt signaling in the mESCs. However, these results showed that doxycycline, with or without CHIR, could provide tight control of edit rate, making the line sufficient for lineage recording.

Next, we analyzed the distribution of edit outcomes at each site. Different sites exhibited distinct ratios of edit outcomes (**Figure 2C**). For example, site 6 exhibited a strong bias towards GA, and relatively little editing to AG. By contrast, sites 1 and 4 were more uniform in their outcomes. These differences in outcome bias across target sites are likely driven by differences in the sequences surrounding the dinucleotide for each target, which is known to impact CRISPR-Cas9 cleavage[67] and base editing[63]. Regardless of how edits were biased, however, all target sites exhibited constant biases over time, consistent with the notion that bias is an intrinsic feature of the sequence context (**Figure 2C**). Most importantly, this consistency indicates that the GA and AG edit outcomes are stable for at least 8 days and do not become further edited to GG over the timescales of these experiments. These results thus

validate the design goal of using inducible base editors to produce multiple distinct, individually stable states.

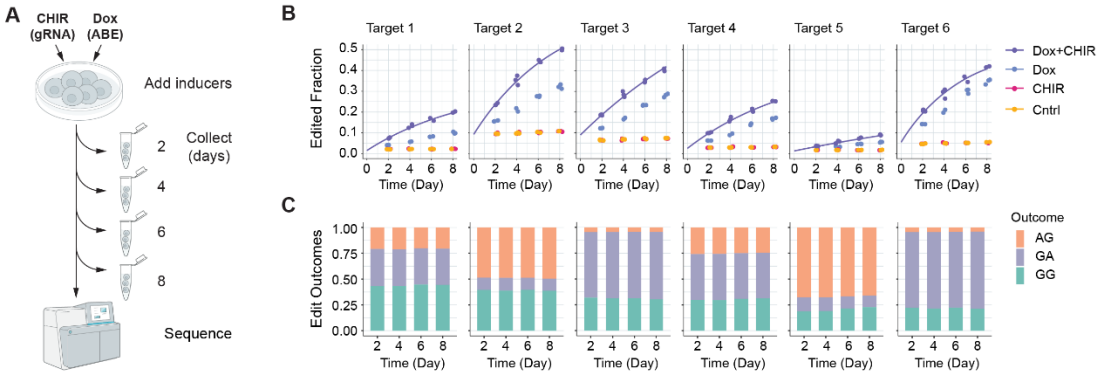


Figure 2: Dinucleotide targets accumulate edits over time in engineered mESCs. (A) Next generation amplicon sequencing quantified editing over time after induction of gRNAs and ABE. **(B)** All targets edited over time in the presence of the two inducers together, although at distinct rates **(B, purple)**. Dox induction alone drives editing at a slower rate **(B, blue)**. In the absence of dox, editing does not proceed at an appreciable rate **(B, red and gold)**. Three biological replicates were collected for each time point. The solid purple line shows the fit for a probabilistic model of editing over time (**Methods**). **(C)** Each target has a unique distribution of editing outcomes that remains constant as editing progresses.

Imaging recovers edited barcode sequences

We next turned from editing to imaging readout. In situ barcode readout creates the opportunity to assay lineage relationships without disrupting spatial relationships among cells. We developed an assay to readout barcode sequences through multiple rounds of single molecule RNA FISH (smFISH). We cultured BaseMEM-01 cells for 3 days — long enough for several cell generations — under editing conditions (3 μM CHIR, 1 μg/mL dox). We then fixed cells and performed in situ T7 transcription of barcodes using the previously described "Zombie" approach[62] (**Figure 3A**, upper and lower left panels). Next, we hybridized pools of 24 primary probes designed to bind to one of the four possible dinucleotide states in each of the six target sites (**Figure 3A, B**). During hybridization, we also included three primary

probes for each of the 400 different 80bp potential static barcode sequences, for a total of 1,200 probes altogether (**Figures 1C, 3A, B**). We also included additional primary probe sets to analyze 12 different endogenous mRNAs, known to distinguish mESC pluripotency states in serum-LIF media[54,68] (**Figures 3A, C**) (**Methods**).

After hybridizing all primary probes, we sequentially added sets of fluorescently-labeled secondary probes, designed to hybridize to corresponding "overhang" sequences engineered in the primary probes (**Figure 3A**, lower panels). For each set of secondary probes, we imaged cells in all channels and then stripped secondary probes to enable the next round of secondary hybridization and imaging. We used two orthogonal fluorescent channels to halve the number of rounds of hybridization required (**Methods**). We also labeled membranes with dye-conjugated wheat germ agglutinin to enable cell segmentation. All imaging was performed on a wide-field fluorescence microscope equipped with an automated fluid handling system similar to those described previously[54–56]. This procedure, similar to that used for seqFISH and related approaches[44,45,54–56,62,69–72], allowed us to systematically probe dynamic barcodes, static barcodes, and endogenous genes over a total of 50 rounds of hybridization and imaging.

In the resulting image sets, individual target arrays could be identified as bright spots across multiple rounds of imaging. Most dynamic barcodes and static barcodes could be uniquely identified by a pseudocolor or set of pseudocolors (rows in image grids, **Figure 3B**). We developed a computational pipeline to detect target array spots and classify the barcode states within each target array (**Supplementary Figure 2**). Across 8 mESC colonies, we were able

to detect roughly 50-80% of the 66 uniquely integrated target arrays in any given cell (**Figure 3D**). One colony had many fewer arrays detected than the others and was excluded from subsequent lineage analysis. The fractional detection of each unique barcode integration among all cells was broad but unimodally distributed, consistent with noisy detection efficiency in the absence of strong systematic differences among integration sites (**Figure 3E**). Overall, after applying quality controls, we detect roughly 200 high confidence dynamic characters, i.e. editable dinucleotides, per cell (**Figure 3F**). Of most direct relevance for lineage reconstruction, with these detection statistics, ~100 shared characters could be confidently recovered in both members of any cell pair.

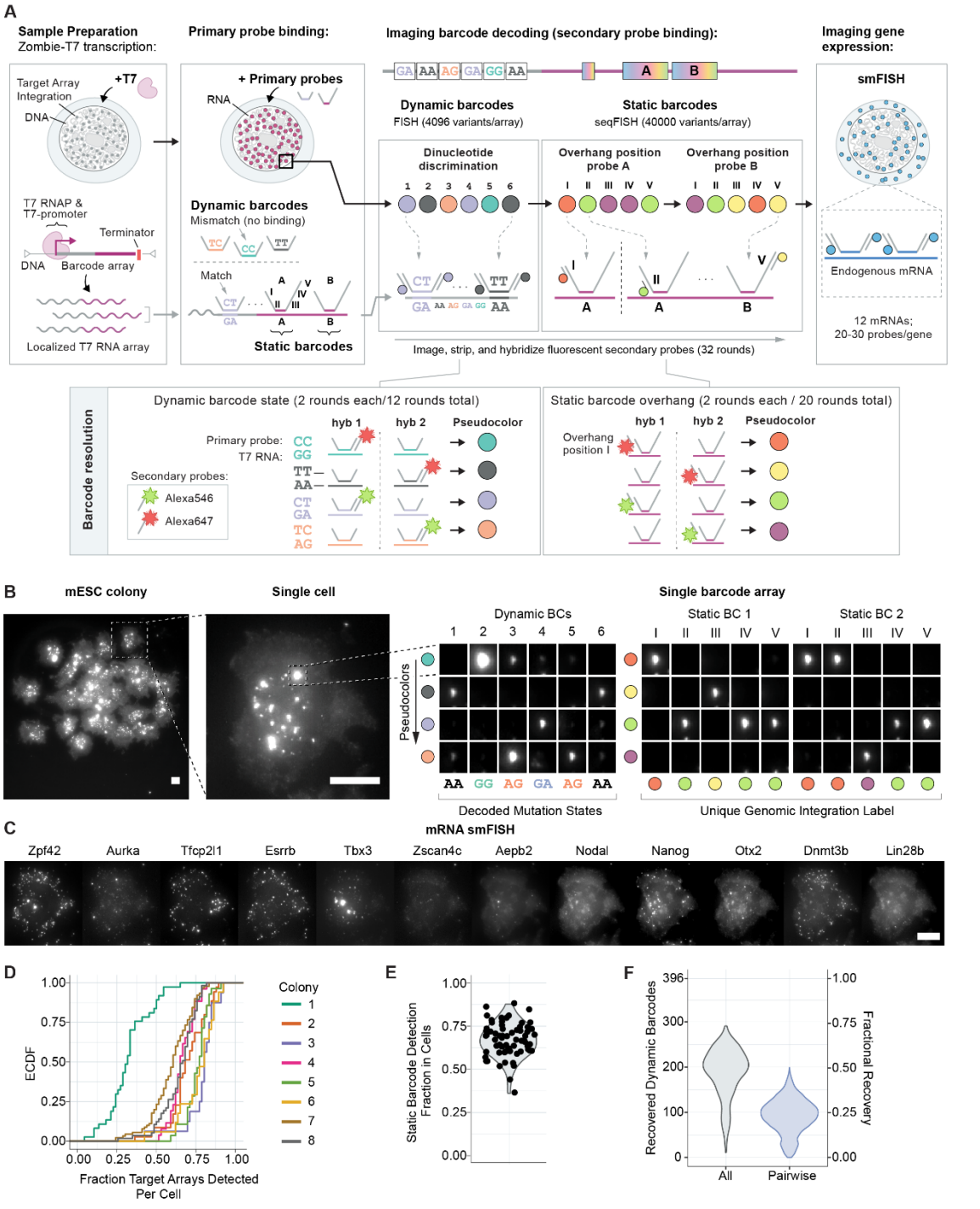

A
Sample Preparation
Zombie-T7 transcription:
Primary probe binding:
Imaging barcode decoding (secondary probe binding):
Imaging gene expression:
Target Array Integration
+T7
DNA
T7 RNAP & T7-promoter
Terminator
DNA Barcode array
Localized T7 RNA array
+ Primary probes
RNA
Dynamic barcodes
Mismatch (no binding)
Match
Static barcodes
Dynamic barcodes
FISH (4096 variants/array)
Static barcodes
seqFISH (40000 variants/array)
Dinucleotide discrimination
Overhang position probe A
Overhang position probe B
smFISH
Endogenous mRNA
12 mRNAs;
20-30 probes/gene
Image, strip, and hybridize fluorescent secondary probes (32 rounds)
Barcode resolution
Dynamic barcode state (2 rounds each/12 rounds total)
Static barcode overhang (2 rounds each / 20 rounds total)
hyb 1
hyb 2
Pseudocolor
Primary probe:
T7 RNA:
Secondary probes:
Alexa546
Alexa647
Overhang position I
B
mESC colony
Single cell
Single barcode array
Dynamic BCs
Static BC 1
Static BC 2
Pseudocolors
Decoded Mutation States
Unique Genomic Integration Label
C
mRNA smFISH
Zpf42
Aurka
Tfcp2l1
Esrrb
Tbx3
Zscan4c
Aepb2
Nodal
Nanog
Otx2
Dnmt3b
Lin28b
D
ECDF
Fraction Target Arrays Detected Per Cell
Colony
E
Static Barcode Detection Fraction in Cells
F
Recovered Dynamic Barcodes
Fractional Recovery
All
Pairwise

Figure 3: Multiple rounds of Zombie-FISH recover dynamic and static barcode states. (A) Barcode states can be recovered across multiple rounds of microscopic imaging. Ectopic application of T7 polymerase generates localized RNA clusters. Primary DNA probes are bound to the dynamic and static barcodes as well as to endogenous transcripts, competing primary probes against each other for binding to the different possible dynamic barcode variants. Each primary probe has an overhang sequence allowing for binding of one or more fluorescently labeled secondary probes, which are hybridized, imaged, and stripped away sequentially to recover barcoding **(B)** and transcriptional **(C)** information. **(D)** Across eight colonies, we recovered 50-80% of target arrays per cell. One colony had dramatically lower barcode recovery and was excluded from further analysis (**D**, colony 1). **(E)** Each unique target array is recovered in a similar fraction of cells. **(F)** We recovered approximately 200 dinucleotide dynamic targets with high confidence per cell, with around 100 of these measured jointly between any pair of cells. Scale bars are 20 μm.

Dynamic barcodes are accurately classified by Zombie-FISH

To verify the accuracy of our pipeline to recover all dinucleotide mutation patterns from Zombie-FISH images, we constructed four independent polyclonal cell lines with mock barcode arrays containing prescribed editing patterns (**Supplemental Figure 3A, Methods**). These arrays span the space of possible editing outcomes for each of the six ABE targets. We cultured these lines in a mixture and applied Zombie-FISH, probing for dynamic and static barcodes (**Supplemental Figure 3B**).

Cells of each type were easily distinguishable based on their static barcode content (**Supplemental Figure 3B**). To be extremely confident about the ground truth of all dynamic barcode states, we excluded cells with fewer than four observed barcode integrations (**Supplemental Figure 3B, right**). We then classified dynamic barcodes using our standard pipeline and asked how closely our classifications matched the actual ground truth arrays. We found that most classifications are 95-100% accurate, although there is a single outcome (the GG mutation on target 4, **Supplemental Figure 3C**) which is frequently misclassified to another character. Overall, we estimate very low experimental error rates as a function of barcode edit saturation given the observed distribution of edit outcomes measured by sequencing (**Supplemental Figures 2D, 3C**).

Simulations show that baseMEMOIR can accurately reconstruct detailed lineage trees

We next asked how the depth of tree reconstruction depends on the distribution of shared characters between cell pairs, as well as other parameters, such as the rate and uniformity of cell divisions, the rate of editing, and the duration of recording. To address these questions, we simulated barcode recording and recovery. During the recording phase, we simulated editing within a single initial cell and its progeny for up to 12 cell generations (**Figure 4A, left**), setting barcode edit rates based on our time course editing dataset (**Figure 2B, C, Supplemental Figure 4**). To simulate incomplete recovery of edit patterns, we stochastically subsampled the resulting barcode sequences corresponding to the observed empirical recovery distribution after hybridization and imaging (**Figure 3F**). As a result, different cell pairs shared different fractions of recovered barcodes (**Figure 4A, middle right**). Finally, we explored a filtering strategy to improve reconstruction accuracy (at the cost of reduced numbers of cells per tree) by restricting analysis to cell pairs with a minimum number of shared recovered barcodes (**Figure 4A, right**).

Next, we reconstructed lineage trees and compared them to the ground truth trees from the forward simulations (**Figure 4B**). As a metric of reconstruction accuracy, we used the normalized Robinson-Foulds distance, which quantifies the fraction of unmatched branches between the ground truth and reconstructed trees. For reconstruction, we adapted the Bayesian BEAST2 phylogenetic reconstruction framework, by incorporating a custom base editing model[73] (**Methods**). BEAST2 uses Markov Chain Monte Carlo (MCMC) sampling

to estimate the posterior probability distribution over different tree topologies and other system parameters. Briefly, it samples a forest of possible trees in proportion to their probability density (**Figure 4B**). As a Bayesian method, it allows for model-based inference, explicitly incorporates prior knowledge, and quantifies uncertainty in reconstruction.

In the ideal case of full recovery of all barcode edits in all cells, we obtained near perfect recovery of full lineage relationships for trees up to 12 cell divisions deep (**Figure 4C**). When ~50% of barcodes were lost, error rates for the same 12 generation tree increased to ~10%. However, this error rate could be reduced by restricting analysis to cells sharing at least 75 jointly measured barcode positions (**Figure 4C**). In contrast to the simulations, the experimental system could introduce additional factors such as errors in barcode readout, variability in mean edit rates between cells, and pre-existing edits in the ancestral (root) cell. Nevertheless, these simulations suggest that baseMEMOIR, with empirically observed error and edit rates, should be capable of reconstructing multi-generation lineage trees with cell cycle resolution at ~90% accuracy.

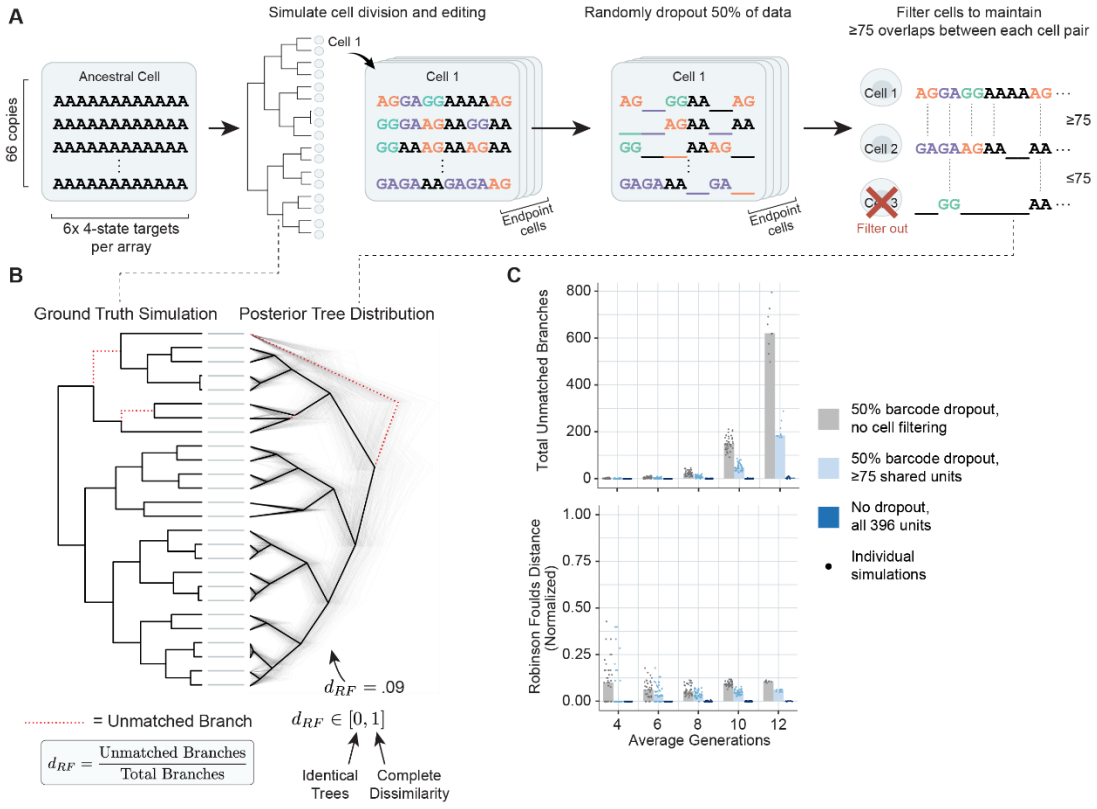


Figure 4: Lineage can be accurately reconstructed for at least 12 generations in simulation. (A) To estimate the expected accuracy of reconstruction, we simulated cell division and stochastic editing starting with unedited barcodes, represented as sets of AA dinucleotides (**left**) over time to produce heterogeneous edit patterns. We then either retained all sequences or dropped 50% of the data to represent random FISH detection losses, and filtered out cells that had few barcode characters overlapping with those measured in other cells (**right**). **(B)** Based on these ground truth simulations, we reconstructed lineage relationships and computed the Robinson-Foulds distance between the ground truth input (**left**) and reconstructed output (**right**) trees. **(C)** Reconstruction accuracy was nearly perfect without barcode dropout (**dark blue dots**). With dropout, we observed ~ 10% error rates with tree depths up to 12 cell generations (**C, gray dots**). In the presence of dropout, filtering cells with few shared units moderately improved the reconstructed tree (**C, light blue dots**).

BaseMEMOIR reconstructs lineage trees in mESC colonies

mESCs are known to undergo spontaneous reversible transitions among a set of molecularly and functionally distinct cell states, ranging from 2C-like to formative and primed epiblast-like states, in serum-LIF conditions[54,68,74–79]. A fundamental question about the mESC state-switching process is the structure of the transition graph, i.e. which transitions occur and at what rates, and how those rates are influenced by input signals. In particular, CHIR, a Wnt

pathway agonist that is often used to maintain pluripotent cells, could impact the observed cell states and their transitions. Although CHIR is known to promote expression of key pluripotency genes and self-renewal in this system[75,76,80–86], the effects of CHIR on single cell state transition dynamics remain unknown.

To study these dynamics, we grew mESC colonies over a three-day period in the presence of CHIR and Dox, which also serve to induce editing (**Figure 5A**). After three days, we imaged the colonies as described above and recovered barcode states for seven colonies out of eight total measured (**Figure 3D**). In addition to reading out barcode states, we also recovered gene expression levels for 12 pluripotency state markers, then clustered to identify 5 gene expression states (**Figure 5B, C and Supplemental Figure 5**). We identified 3 major cell states comprising the following: 2C-like cells with high expression of Zscan4C; naive cells expressing transcription factors Nanog, Esrrb, and Zfp42; and formative cells that express Otx2 and Dnmt3b in the absence of naive pluripotency factors and Zscan4C (**Figure 5B**). Naive cells exhibited a distribution of gene expression levels that varied from more 2C-like to more formative (**Figure 5B, C and Supplemental Figure 5**). These states largely correspond to those described in previous work[54,68,79], although we observed high and relatively homogeneous Tbx3 expression across all cell states, in contrast to observations of cells grown in serum/LIF without CHIR[54,68]. This difference is consistent with previous work showing that Tbx3 is significantly regulated by CHIR in the observed direction[86].

We next applied the BEAST2 system described above to reconstruct lineage trees for cells in these colonies. To incorporate information of cell state and spatial position, we extended

the underlying editing model (**Figure 4**) to represent these additional cellular properties, and thereby allowed simultaneous inference of cell state transitions and spatial movement alongside cell lineage (**Methods**). Applied to the mESC colonies, this approach reconstructed lineage relationships and division times with relatively low uncertainty in most cases, as indicated by the limited "fuzziness" of the reconstructed trees (**Fig. 5D and Supplemental Figure 6**). However, there were some ambiguities in reconstruction. For example, in Figure 5D, cell 23 is roughly equally likely to be a sister of cell 22 or cell 24. This illustrates the way in which uncertainties in the Bayesian reconstruction still provide specific alternative hypotheses rather than numerical confidence values. Additionally, in some cases, we did not capture all neighboring cells, which may introduce branch lengths longer than a single cell division (for example, see clades A and B of colony 7, **Supplemental Figure 6**). Together, these results demonstrate that the recording system allows precise lineage reconstruction of 3-day clonal mouse ESC colonies, with tree sizes of 30-50 cells.

The reconstructions also allowed estimation of cell state transition dynamics. To constrain the transition model, we treated transitions as a reversible, symmetric, continuous time Markov process (see **Methods** for discussion of these assumptions). Of the 10 possible symmetric interactions, posterior estimates suggested that ~5 occurred at appreciable rates during the growth of these colonies (**Figure 5F, G**). A sixth transition, between 2C-like and formative states, was suggested by a single event in colony 7 (**Supplemental Figure 6**). This event is unexpected given previous inference of chainlike dynamics, with these two states at opposite ends[68], however it could be a result of CHIR exposure, which was not present in previous work. Further, a substantial amount of posterior probability indicates a negligible

rate for this transition; more data would be needed to clarify this result (**Figure 5F, Supplemental Figure 6**). Overall, these reconstructions suggest frequent (median of 0.15 transitions per day across all colonies) transitions among Naive/2C-like, Naive, and Naive/Formative states, and frequent conversion between Naive/2C-like and 2C-like states (**Supplemental Figure 6**). Interestingly, the inferred transitions correlate with expectations based on transcriptional similarity among states, even though this information was not provided to the model (**Figure 5B, C and Supplemental Figure 5**).

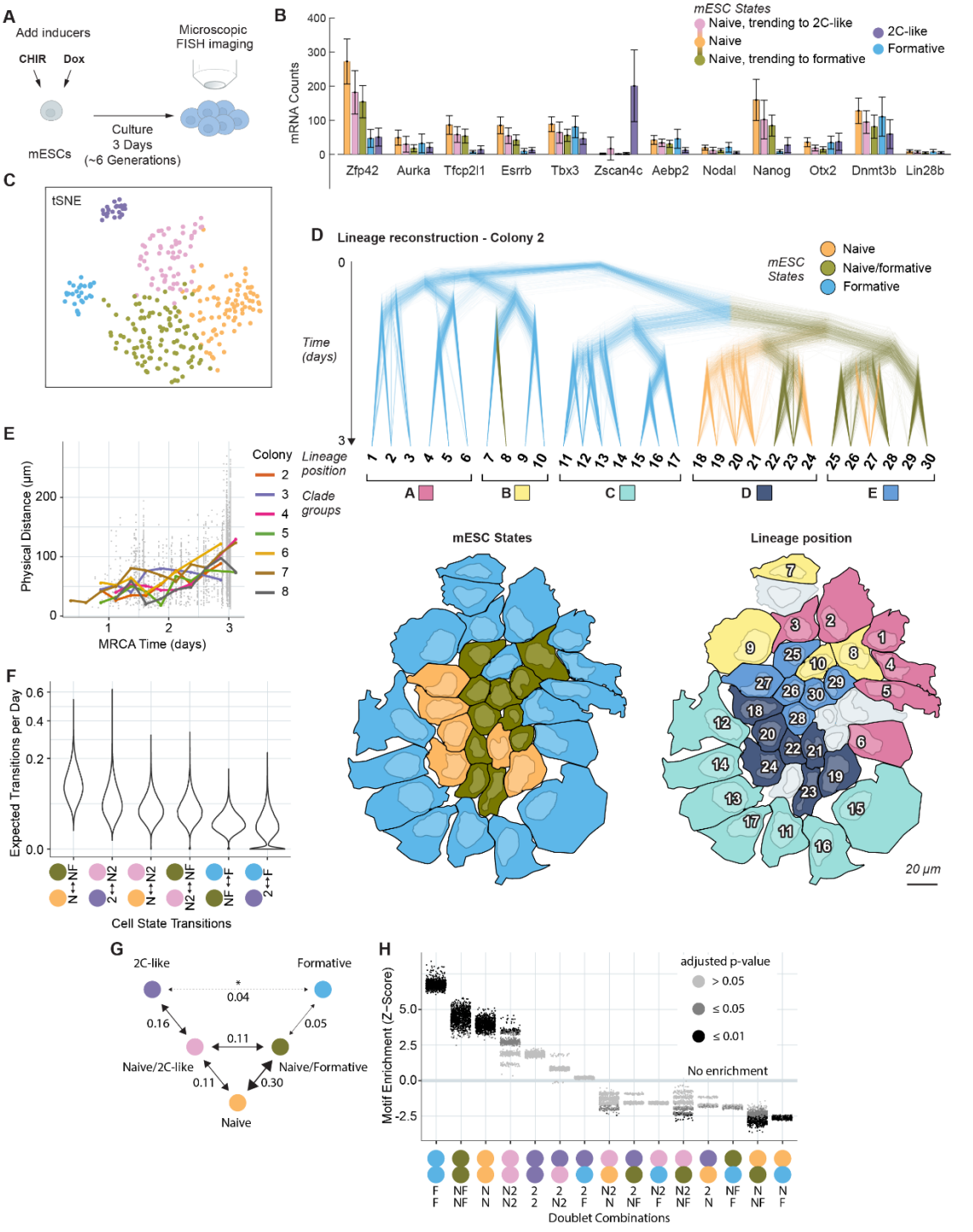

A
Add inducers
CHIR
Dox
mESCs
Culture
3 Days
(~6 Generations)
Microscopic
FISH imaging
B
mRNA Counts
mESC States
Naive, trending to 2C-like
Naive
Naive, trending to formative
2C-like
Formative
Zfp42
Aurka
Tfcp2l1
Esrrb
Tbx3
Zscan4c
Aebp2
Nodal
Nanog
Otx2
Dnmt3b
Lin28b
C
tSNE
D
Lineage reconstruction - Colony 2
mESC States
Naive
Naive/formative
Formative
Time (days)
A
B
C
D
E
E
Physical Distance (μm)
MRCA Time (days)
Colony
Lineage position
Clade groups
F
Expected Transitions per Day
Cell State Transitions
mESC States
Lineage position
20 μm
G
2C-like
Formative
Naive/2C-like
Naive/Formative
Naive
0.04
0.16
0.05
0.11
0.11
0.30
H
Motif Enrichment (Z-Score)
adjusted p-value
> 0.05
≤ 0.05
≤ 0.01
No enrichment
Doublet Combinations

Figure 5: Joint measurements of lineage, gene expression, and spatial position reveal cell state transition dynamics. **(A)** We recorded lineage relationships in mESC cells cultured in serum-LIF media over a 3-day period, inducing editing with 3 μM CHIR and 1 μM Dox. **(B, C)** Cells clustered into five states based on gene expression as measured by smFISH. Two clusters were well separated from the other groups while three clusters appeared continuously related and expressed different levels of key marker genes (**see Supplementary Figure 4**). **(D-G)** Lineage reconstruction infers topological lineage tree relationships, cell division timing, ancestral cell states, and transition rates between those states. Uncertainty in lineage tree measurements is visualized by overlaying trees sampled from the posterior distribution of trees generated by Markov chain Monte Carlo for each colony (**D, top; Supplemental Figure 5**). Cell states and clade groups from the lineage tree can be mapped to the spatial colony images to qualitatively inspect the relationships between cell state, lineage, and spatial location (**D, bottom; Supplemental Figure 5**). **(E)** Spatial distance is larger between cells with more distant common ancestors. **(F)** Several cell state transitions were inferred to have nonzero median values across all posterior samples. **(G)** These state transitions predict a restricted cell state transition graph. One transition (denoted by *) contained a high fraction of posterior samples with a transition rate of 0. Numbers indicate the median expected number of transitions per day for cells of the given type. **(H)** Several doublet motifs are significantly over or underrepresented across the lineage tree posterior samples. N: Naive; 2: 2C-like; F: Formative; N2: Naive, trending to 2C-like; NF: Naive, trending to formative; MRCA: Most recent common ancestor.

BaseMEMOIR recovers lineage relationships, cell states, and spatial relationships in mESC colonies

By mapping lineage trees back onto the original images, we were able to simultaneously observe spatial and lineage organization of colonies (**Figure 5D**, lower panels). This analysis revealed correlations between lineage history, spatial position, and cell fate. For example, in Figure 5d, the related D and E clades contain Naive and Naive/Formative cells, and were located towards the interior of the colony, while cells in clades A, B, and C were largely in the Formative state and located around the periphery. Individual cell morphologies also varied systematically, with cells in the periphery exhibiting larger sizes. Other colonies were less radially structured but still showed strong correlations between spatial positions and lineage relationships within each colony (**Figure 5E, Supplemental Figure 6**). These results show that it is possible to impose lineage relationships on spatial colonies with cell state information.

Lineage motifs provide a complementary approach to analyzing cell state transitions[87]. They are defined as statistically over-represented patterns of cell fates on lineage trees, which

reflect features of underlying stochastic cell fate control programs[87]. As a simple example, asymmetric division, in which sister cells acquire opposite fates, would be reflected in the enrichment of opposite fates among sibling pairs ("doublets"). In contrast to the inference of transition rates described above, lineage motifs can be identified with no assumptions about an underlying model. Applying Lineage Motif Analysis[87] (LMA) to 1000 samples from the Bayesian posterior tree distribution, we identified 4 overrepresented doublet pairs with an adjusted p-value less than 0.05 for a majority of the posterior samples (**Figure 5H**). These cases involve siblings in the same state, consistent with infrequent state transitions. Siblings in the formative state were the most overrepresented, mirroring results from the Bayesian Markov model, which predicts the slowest transitions to and from the formative state (**Figure 5F, H**).

We also observe two statistically underrepresented heterogeneous sibling pairs (**Figure 5H**). The most underrepresented pair, containing naive and formative cells, was also qualitatively consistent with predictions of the Bayesian Markov model, which identified a negligible transition rate between these states. Additionally, the naive and naive/formative sibling pair was also significantly underrepresented. This corresponds to the most rapid inferred transition rate in the dataset (**Figure 5F**), consistent with high rates of independent transitions out of either the naive or formative states. Together, these results demonstrate how baseMEMOIR's lineage reconstruction ability allows inference of lineage motifs.

Finally, we combined the lineage reconstruction with spatial and cell state dynamics to infer a property that would be difficult to analyze from sequencing-based readout or static

snapshots alone: the relative spatial mobilities of different cell states. The inferred histories of cell state and spatial position can be visualized (**Supplemental Movies**). These movies represent one possible history based on a simple model of cell diffusion, taking the highest credibility inference from BEAST2. Together, these results show how spatial position, cell state, and lineage can be analyzed and reconstructed together, and used to infer features of cell histories.

BaseMEMOIR is portable to in vivo systems

While there is a rich history of using engineered mESC lines to study embryonic development in vivo through generation of chimeric embryos and tetraploid complementation, many unresolved questions remain. For example, the origin of primordial germ cells (PGCs) is not completely determined: canonically, they are thought to originate from the extraembryonic mesoderm, but recent work suggests that there could be contributions from the epiblast and earlier mesodermal lineages, or even directly from the primitive streak[88]. While definitively addressing these questions is outside the scope of the current study, we wanted to demonstrate that our system has the potential to tackle them in an in vivo context. To that end, we applied baseMEMOIR to in vivo embryonic development through tetraploid complementation (**Methods**).

One issue with using baseMEM-01 directly is that gRNA expression is driven by Wnt. As the Wnt pathway is used extensively in development, constitutive overexpression leads to developmental defects and nonviable embryos in vivo. To that end, we reengineered a parental version of the baseMEM-01 cell line, which contained all constructs except the

gRNA expression vector, to harbor a redesigned gRNA expression vector under control of the TRE3G promoter (**Methods**). In this case, doxycycline induces expression of both base editor and gRNAs. After engineering, we selected monoclonal lines and characterized editing by next generation sequencing (**Supplemental Figure 7**). We used a single clone, baseMEM-02, for subsequent in vivo experiments.

Initially, we attempted to form chimeric embryos by injection of baseMEM-02 directly into E3.5 mouse blastocysts. We then reimplanted the resulting embryos into surrogate mothers, and allowed development through day 7.5 with administration of dox in the mother's food and water (**Methods**). We isolated and sectioned the embryos, then attempted to identify engineered cells through the presence of static barcode sequences by Zombie-FISH (**Methods**), however we were not able to identify contribution of the cell line to the embryo by this method.

We reasoned that lack of contribution could be related to the phenomenon of cell competition, where cells with higher overall fitness actively induce apoptosis in cells with lower fitness[89–93]. Since our cell line has undergone extensive engineering and expresses a number of transgenes, we speculated that they may not be able to compete with WT cells present in the developing blastocyst. Critically, outcompeted cells can be fully competent to form a normal embryo on their own in the absence of competition; for example, healthy WT cells can be "outcompeted" by WT cells engineered to express a high level of Myc[92].

Hence, we decided to apply tetraploid complementation instead[94–99]. In this assay, developing embryos undergo electrofusion at the 2-cell stage, and are subsequently grown

ex vivo until E3.5. This generates a blastocyst comprised of tetraploid cells, which are competent to form some extraembryonic tissues but not tissues of the embryo proper[94–99]. Injection of diploid cells at the E3.5 stage, followed by implantation into a surrogate mother, thus allows engineered cells to populate the entirety of the developing embryonic tissues.

Following this strategy, we generated and isolated chimeric tetraploid E7.5 embryos developed in the presence of dox fed to the mother through food and drink. In this case, we found that baseMEM-02 cells populated all embryonic tissues, with no recording cells identified in maternal tissues as expected (**Figure 6A**). Following barcode readout, we used single molecule FISH to characterize expression of a panel of 63 genes in the developing tissues (**Supplemental Data**). We observed patterns of gene expression characteristic of embryos developing at this time point, and resolved expected cell types based on marker gene expression (**Figure 6B, C, Supplemental Data**).

In tissue sections, we observed markedly lower recovery of barcodes than we observed in cell culture (**Supplemental Figure 8**). Barcode editing also occurred at a lower rate in this context (**Supplemental Figure 8**). We filtered cells with fewer than 75 measured barcode characters, leaving a total of 605 of the 1134 recording cells captured within the sample (**Figure 6D, E**). For this group of cells, we went on to reconstruct lineage relationships using BEAST2 (**Methods, Supplemental Data**). Due to the more limited editing and barcode recovery, tree distributions were more uncertain than in the case of cell culture, and branch lengths were highly variable across the posterior distribution of trees (**Supplemental Figure 8J**). We decided to focus further analysis on cladograms, ignoring branch lengths within the

tree. To assess support for specific clades, we determined the maximum clade credibility tree from the posterior distribution and computed the transfer score[100] for each resulting clade across the posterior samples. A number of clades scored above 70%, however many clades were less certain (**Figure 6D, Supplemental Figure 9**).

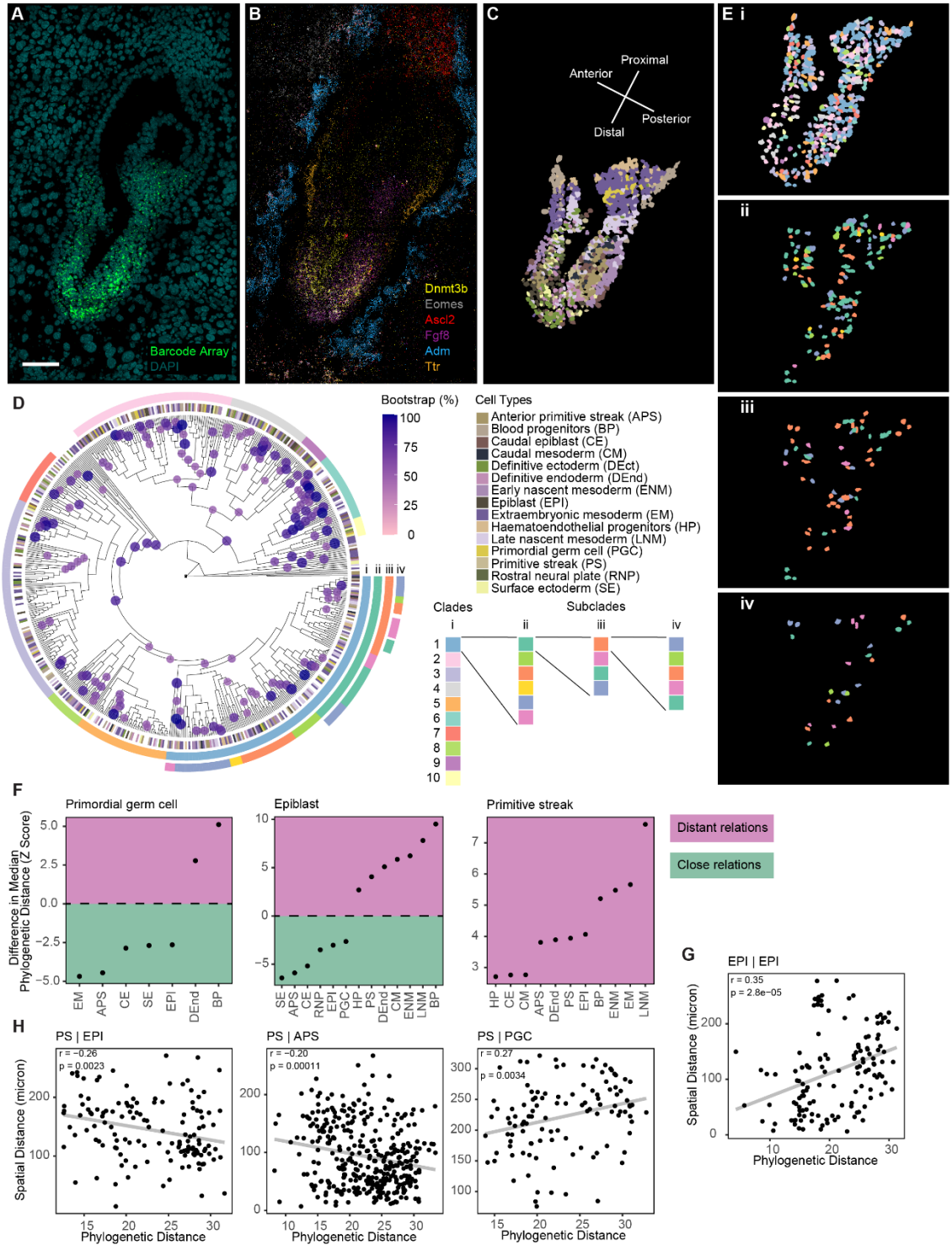


Figure 6: BaseMEMOIR captures mouse development in vivo. (A) We recorded lineage relationships in mouse embryos from E4.5-7.5 using tetraploid complementation. Barcode arrays are detected in the embryonic compartment at E7.5. Scale bar 100 micron. **(B)** We measured a panel of 63 genes to distinguish cell types at this developmental stage (several representative

marker genes shown). **(C)** Expected cell types were identified based on marker gene expression. **(D)** Genetic barcodes enabled inference of lineage relationships with uncertainty estimates for specific clades. A cladogram with smoothed branch lengths is shown, highlighting clades with transfer bootstrap score > 50%. Cell states for all taxa are visualized encircling the phylogeny (inner ring). Clades and subclades are denoted by the outer colored rings. **(E)** Clones and subclones identified in the cladogram show broad distributions across the tissue. **(F)** A variety of cell types showed significantly closer or more distant relationships to other cell types than would be expected by chance. Plots relay the difference in median phylogenetic distances across all pairwise cell comparisons for the indicated cell pairs relative to random cells sampled from the phylogeny. Only comparisons with significantly distinct median values (FDR corrected p-value < 0.05) are shown. For all significant comparisons, see Supplemental Figure 10. Significance was assessed using a permutation test (Methods). **(G)** Pairwise phylogenetic distances are positively correlated with spatial distances between cells for epiblast cells paired with other epiblast cells. **(H)** Primitive streak cells paired with either epiblast or anterior primitive streak have phylogenetic distances that anticorrelate with spatial distance, while they show positive correlation with primordial germ cells. Significance for panels G and H was assessed using the Pearson correlation test.

Phylogenetic analysis reveals cellular relationships during embryogenesis

We mapped clades from the resulting lineage, cut at various branch points, back to the original spatial coordinates. Qualitatively, we observed a large spatial spread for clones and subclones (**Figure 6E**). Previously reported cellular movement and mixing of clones at this stage could explain this result, although our tree is not sufficiently resolved to make conclusive statements about specific clades from the maximum credibility tree alone.

Instead, we looked at features of the entire posterior distribution to draw inferences about the system. We computed the average pairwise phylogenetic distance from the posterior cladogram distribution for all pairs of cells, defined as the total path length on the tree between a pair of taxa in units of cell divisions. We then asked, for specific cell types, whether the median phylogenetic distance between those cells was significantly higher or lower than the median phylogenetic distance for a random sample of cells (**Figure 6F, Supplemental Figure 10, Methods**).

PGCs are thought to arise from extraembryonic mesoderm, but their definitive origin remains to be elucidated; recent models suggest they could arise from multiple sources, including the late nascent mesoderm, early nascent mesoderm, the primitive streak, or directly from the

epiblast[88]. We find that PGCs are significantly more closely related to extraembryonic mesoderm cells, anterior primitive streak cells, and caudal epiblast cells than would be expected by chance (**Figure 6F, left**). Interestingly, we also find them to be closely related to surface ectoderm, for which there is some precedent based on transcriptional flow modeling[101]. We find them to be most distantly related to definitive endoderm and blood progenitors, which is also consistent with transcriptional predictions[101].

Relatively few epiblast cells remain by E7.5, the majority having ingressed through the primitive streak or taken on an ectodermal fate[101]. We find that those cells that remain are closely related to other epiblast or caudal epiblast cells, surface ectoderm, and anterior primitive streak (**Figure 6F, middle**). Surprisingly, the remaining epiblast population is comparatively distantly related to the primitive streak and mesodermal fates, perhaps implying that the streak forms from a clonally distinct population of epiblast cells earlier in development. By this stage, many primitive streak cells have also ingressed to take on various mesodermal fates[26,27]. Curiously, we find that the remaining primitive streak population tends to be more distantly related to mesodermal fates, anterior primitive streak, and other primitive streak cells, in addition to epiblast and a variety of other cell types (**Figure 6F, right**).

Phylogenetic and spatial distances correlate for some cell types at E7.5

Across all cells, correlation between spatial and phylogenetic distances is very low (Pearson's R = 0.014), in contrast to our observations of mESCs in culture (**Figure 5E**). This is consistent with the intensive cell migration and clonal mixing observed in previous

studies[30,31]. In contrast to this result, when stratified by cell type, several populations have significantly correlated (or anticorrelated) distances. One of the strongest observed correlations is among epiblast cells, perhaps indicating that cells remaining in the epiblast state by E7.5 undergo less mixing than in other compartments (**Figure 6G**, **Supplemental Table 1**). Similarly strong correlations were observed when comparing caudal mesoderm / early nascent mesoderm, definitive ectoderm / surface ectoderm, and surface ectoderm with itself (**Supplemental Table 1**).

The strongest occurrences of anticorrelation were observed between primitive streak / epiblast and primitive streak / anterior primitive streak (**Figure 6H, left panels**). This is consistent with the migratory nature of the primitive streak, where ingressing cells differentiate while moving large distances across the embryo[26,27]. Interestingly, PGCs and primitive streak cells exhibit positive distance correlations (**Figure 6H, right**), perhaps implying that some PGCs are direct descendants of the streak. This would be consistent with models suggested by recent investigation of transcriptional flows across development[88].

We urge readers to be cautious in interpreting these biological findings, given that we have only analyzed a single tree from a single section of a single embryo thus far. We see these results as suitable for generating hypotheses, but not for definitively addressing the biological system. Future work will explore these topics more deeply by examining more embryos and optimizing barcode recovery and editing for improved lineage inference.

Discussion and conclusions

A long-standing dream in biology is to image a tissue or organism and visualize not only its current state, but also its past history. Previous work has approached this ideal in different ways, including lineage recording by accumulation of irreversible recombination events and reconstruction of small trees, however these efforts were limited in the amount and scalability of memory storage[44,45,62]. Here, we introduce a new approach, baseMEMOIR, which provides much larger memory sizes and allows for deeper, more accurate lineage tree reconstruction, while preserving spatial structure.

To achieve this, baseMEMOIR introduces several key innovations. First, it uses base editors to introduce stochastic, but precise, edits at dense target arrays (**Figure 1C**). Second, it uses dinucleotide editable target sites, each of which can be edited to any of three permanent end states (**Figures 1B, 2C**). Third, to discriminate between those states we expanded the Zombie readout system[62] to allow 4-way probe competition (**Figure 3A, B**). Fourth, baseMEMOIR massively expands the amount of memory accessible in single cells by incorporating 66 unique statically barcoded target arrays, collectively providing 792 bits of editable information in the baseMEM-01 cell line. Theoretically, this number could be readily increased with additional target array integrations without modifying other components of the system. Fifth, baseMEMOIR achieves high density recording, while maintaining compatibility with FISH-based readout of endogenous genes (**Figure 3C**). Finally, to address the challenge of lineage reconstruction from stochastic edits, we adapted the BEAST2 framework for Bayesian tree inference, both by adding a new mutation model and taking

advantage of its phylogeographical and discrete trait models[73,102] (**Methods**). We anticipate that this probabilistic framework should be applicable for a broad variety of lineage recording methods.

To demonstrate these capabilities, we applied baseMEMOIR to stem cells undergoing interconversion among transcriptional states[54,68,74–77,86]. This allowed us to reconstruct lineage trees for 7 colonies totaling 197 cells, with as many as 4-7 cell generations per colony (**Figures 5D, Supplemental Figure 6**). Further, we were able to infer transition rates for specific pairs of states. These rates were consistent with a role for Wnt (through CHIR) in influencing state dynamics relative to similar cultures in the absence of Wnt[54,68,86] (**Figure 5F, G**). Beyond cell culture, we demonstrate preliminary use of the system in vivo, although we emphasize that biological conclusions should not be conclusively drawn at this time (**Figure 6**). Future work could use baseMEMOIR to systematically compare the effects of different signals and perturbations on cell state dynamics. While probabilistic inference is not equivalent to direct time-lapse observation, it nevertheless is beginning to yield related insights that would ordinarily be concealed from any static endpoint measurements (**Supplemental Movies**). Extrapolating from the capabilities of this system to future implementations, such as those containing either more memory or linking signaling pathway activity to recording machinery[41,44], it should become possible to infer increasingly detailed views of earlier dynamic events in complex multicellular settings, effectively "decorating" lineage trees with events, such as changes in cell state or even movements in space. baseMEMOIR should also allow one to infer state-switching dynamics and developmental

programs using approaches such as Kin Correlation Analysis and Lineage Motif Analysis that exploit lineage tree information[68,87].

While powerful, baseMEMOIR has some limitations. Because it does not directly probe the states of cells at earlier time points, it cannot directly detect earlier states that do not appear in the endpoint measurement. Analyzing systems at multiple timepoints could help to avoid missing transient states. Additionally, cells that die or migrate away prior to measurement will be omitted from the tree and could confound estimates of variation in cell cycle durations in different lineages. Similarly, failure to recover sufficient barcodes from an individual cell could make it difficult to classify. This is especially apparent in tissue sections, however we are optimistic that further technical improvement to barcode imaging could alleviate these issues, for example by applying tissue clearing methods developed in related work[38].

baseMEM-01 and baseMEM-02 can immediately be used to explore stem cell differentiation and early mouse embryogenesis, among other phenomena. Looking ahead, baseMEMOIR should be readily adaptable to diverse developmental and physiological processes. The constructs and system can be transplanted to additional cell types using standard methods, and potentially combined with readout of additional "multi-omics" information such as chromatin accessibility[54]. One can therefore anticipate augmenting spatial cell atlases with lineage information[53], and using baseMEMOIR to investigate the role of lineage, signaling, and differentiation in disease progression.

Methods

Dynamic barcoding strategy

Dynamic barcode sequences consist of 20bp CRISPR target sites with 3bp downstream NGG PAM sequences. These were chosen by designing sequences with AA nucleotides at the location predicted to be edited by the ABE (positions 5-6 in the protospacer sequence)[103], then screening them for significant, varied editing of the AA sites. Six unique target sites are arrayed sequentially downstream of a T7 promoter sequence to enable imaging-based readout as described below (**Figure 1C**).

Static barcoding strategy

Static barcodes consist of two variable 80bp sequences downstream of the six dynamic barcode targets (**Figure 1C**). A pooled plasmid library was formed by generating constructs with 200 variants at each of the two 80bp regions, for a total of 40,000 unique sequences (**Supplemental Data**). Each sequence contains three unique primary probe binding sites for signal amplification during FISH readout (**Supplemental Data**).

Plasmid construction

Plasmids were constructed in piggyBac backbones for later transposase mediated integration into the genome. The inducible ABE plasmid was made by integrating a tet-responsive promoter (TRE3G, Takara Bio) and ABE 7.10[103] (Addgene #102919) into a piggyBac plasmid[104] with neomycin resistance. The Tet-On 3G protein gene used to activate the ABE

in a doxycycline dependent fashion was supplied as a piggyBac plasmid with a pEF promoter and puromycin resistance.

The dynamic and static barcode arrays were constructed in a piggyBac vector containing hygromycin resistance and double T7-T3 promoter sites followed by the dynamic barcode array, which was synthesized by Integrated DNA Technologies (IDT). The static barcode was then integrated 3' of the dynamic barcode array. The static barcode was composed of two sites of 80 bp each, with 200 possible sequences for each of the two sites to give an overall possible barcode diversity of up to 40,000 unique sequences. The static barcode sequences were synthesized by Twist Bioscience and amplified with the appropriate cloning ends by PCR. The 5' primer for the first static barcode site had a set of 10 random nucleotides to provide a further NGS-readable ID to each barcode. A mix of Gibson and sticky end cloning were used for plasmid construction.

The plasmid library containing static barcodes was generated by transforming high-efficiency competent cells (NEB C3019), then plating them onto a large surface area of LB-agar (~30 10-cm petri dishes) to generate a large number of colonies. These were scraped and pooled into a single liquid culture. Subsequently, plasmid DNA was collected using multiple DNA Miniprep columns (Qiagen 27104) and pooled.

An array of six gRNAs targeting the six sites of the dynamic barcode were integrated in the 3' UTR of an NLS-mTurquoise gene. Each gRNA sequence was flanked by the hammerhead and HDV ribozyme sequences on upstream and downstream sides, respectively, in order to excise the gRNA from the transcript. These gRNA-ribozyme sequences were each

synthesized as gBlocks by IDT and combined by assembly of unique sticky end junctions into the piggyBac plasmid. A Wnt-responsive promoter was integrated to drive expression of the mTurquoise-gRNAs construct. This plasmid included blasticidin resistance for subsequent mammalian selection. For the development of baseMEM-02, we replaced the Wnt-responsive promoter in this construct with the tet-responsive promoter used in the ABE vector (TRE3G, Takara Bio).

To develop constructs used in generating control cell lines (**Supplemental Figure 7**), the static barcode plasmid library was transformed into competent cells (NEB 10-Beta) and plated on standard LB-agar plates. Individual colonies were selected and plasmids were purified on DNA Miniprep columns as described above. Resulting constructs were screened by Sanger sequencing (Laragen) to identify resulting static barcode sequences for each plasmid. Four constructs were selected and further modified commercially to include the desired dynamic barcode states (Genscript).

Primary probe library construction

Primary probes for dynamic barcode readout were purchased from IDT as individual sequences. The primary probe library, containing 1200 probes targeting all static barcode variants across both regions (3 probes per variant, 200 variants per region, 2 regions), was ordered as an oligoarray pool from Twist Bioscience. Each probe was assembled with a 35-nucleotide sequence complementary to the static barcode sequence, five 15-nucleotide readout sequences uniquely labeling each variant separated by a 2-nucleotide spacer, and two flanking primer sequences to allow for PCR amplification of the probe library (structure 5’-

(primer 1)-(readout 1)-(readout 2)-(probe)-(readout 3)-(readout 4)-(readout 5)-(primer 2)-3'). The probe library was amplified following an established protocol[54].

Endogenous marker genes were selected based on previous work.[54,68] Probes for non-barcoded sequential smFISH of gene markers were a kind gift from Long Cai, generated as described previously[54], using a single readout sequence repeated four times in place of a unique barcode (structure 5'-(primer 1)-(readout 1)-(readout 1)-(probe)-(readout 1)-(readout 1)-(primer 2)-3').

For control cell line and in vivo experiments, similarly designed primary probe libraries were constructed to recover barcode and endogenous gene sequences with several key differences. The barcode library pool was reduced to include probes for only the static barcodes identified in the system in earlier experiments (**Supplemental Figure 1, Supplemental Data**), as well as an additional four static barcode sequences corresponding to fixed array control lines (**Supplemental Figure 3**).

The static barcode readout strategy was additionally redesigned to reduce the number of imaging rounds. In the first implementation, the two FISH-readable static barcode sequences are decoded over 5 rounds of 4-pseudocolor imaging each (10 total rounds, **Figure 3A**), where each round is designed to have at most one detected pseudocolor. In the redesigned library, the readout rounds for barcode positions 4 and 5 on the first static barcode probe were designed to be redundant (using the same readout binding sites in different combinations) as rounds 4 and 5 on the second static barcode sequence, reducing the required total rounds of readout to 8. Due to the relatively small number of insertion sequences relative

to the encoding space (66 total integrations), static barcode integrations can still be uniquely assigned even with this redundancy. In this case, some array integrations are expected to have up to 2 pseudocolors recovered per 4-pseudocolor imaging round (**Supplemental Figure 8C**).

Endogeneous marker genes for in vivo experiments were selected based on a recent atlas of transcription during mouse embryogenesis[101]. Primary probes were designed using PaintSHOP[105] to identify appropriate binding sequences, then manually appending secondary readout binding motifs, alongside a T7 promoter for probe library amplification and a bridge ligation binding motif as discussed in previous work[54] (**Supplemental Data**).

Readout probe synthesis

Fluorescently conjugated secondary readout probes 15-nt in length were designed as in previous work[54,72]. Probe sequences were ordered conjugated to AlexaFluor 546 or 647 from IDT as indicated (**Supplemental Data**).

Coverslip functionalization

24 x 60 mm coverslips were functionalized prior to cell culture. Coverslips were first rinsed in 100% ethanol, then dried and functionalized using a plasma cleaner on the high setting for 5 minutes. Coverslips were subsequently immersed in 1% bind-silane (GE, 17-1330-13) solution (1% bind silane, 10 mM acetic acid in 90% ethanol) for 1 hr at room temperature. Coverslips were rinsed in 100% ethanol then heat dried in an oven at 90 C for 30 minutes before being treated with 100 ug/mL Poly-D-Lysine in water overnight. The following day,

slides were rinsed with nuclease free water and air dried. Slides were stored for up to 2 weeks at 4 C prior to use. These coverslips were either used directly for sectioning tetraploid embryos or further treated with laminin for cell culture experiments as described below.

Just before cell attachment, coverslips were treated with UV in a biosafety cabinet for 5 minutes, then the surface was treated with 10 ug/mL laminin (Biolaminin 511 LN, Biolamina) at 37 C for 90 minutes. Laminin was removed, then cell suspension was added directly to the surface for attachment.

Cell culture

E14 mES cells (ATCC cat. No. CRL-1821) were cultured in medium containing GMEM (Sigma), 15% ES cell qualified FBS (Gibco), 1x MEM non-essential amino acids (Thermo Fisher Scientific), 1 mM sodium pyruvate (Thermo Fisher Scientific), 100 µM B-mercaptoethanol (Thermo Fisher Scientific), 1x penicillin-streptomycin-L-glutamine (Thermo Fisher Scientific), and 1000 U/mL leukemia inhibitory factor (Millipore). For cell engineering and standard culture, cells were maintained on polystyrene (Falcon) plates coated with 0.1% gelatin (Sigma) at 37 C and 5% CO2.

Cell line engineering

Sequences of all integrated constructs are reported as Supplementary Data. BaseMEMOIR components were integrated over several rounds of transfection and selection. For all transfection steps, mESCs were cultured in 24 well plates, then cotransfected with the plasmid(s) to be integrated as well as piggyBac transposase plasmid with HD FuGENE

transfection reagent. First, cells were cotransfected with ABE and Tet3G activator plasmids. The cells were allowed to recover for a day, passaged, and then underwent selection with 400 ug/mL neomycin followed by 500 ug/mL geneticin. Cells were plated sparsely in a 10 cm dish to grow monoclonal colonies, and then the monoclones were selected and grown in 96 well plates. Clones were screened for ABE expression after dox induction by qPCR, then subsequently by FISH to identify clones with homogenous expression among single cells.

Barcode target plasmids were integrated into the parental line containing inducible ABE by a second round of transfection, then selected with 100 ng/mL hygromycin as previously described. Monoclonal colonies were selected as previously, then screened by qPCR for high relative copy number. Zombie-FISH (described below) was used to screen promising candidates and select the clone with the highest visible integration number.

Finally, gRNAs and additional ABE plasmid were integrated into the most promising line from the previous step. Cells were selected with 15 ng/mL blasticidin, then monoclonal lines were generated as described above. Clones with a clear mTurquoise expression upon addition of 3 μM CHIR, which indicated expression of the gRNA construct, were kept for further analysis.

The best clones were tested for array targeting by adding 1 ug/mL doxycycline and 3 μM CHIR for multiple days followed by Sanger sequencing. Editing resulted in mixed peaks at the edited bases. One of the clones (baseMEM-01) was identified to have the most editing via this approach and was used for subsequent in vitro experiments.

To develop baseMEM-02, we began with an intermediate parental line retained from baseMEM-01 development, containing all vectors except gRNAs. To this line, we integrated a modified gRNA cassette containing a TET-on promoter in place of the Wnt-responsive element present in baseMEM-01. Cells were selected with 15 ng/mL blasticidin, then monoclonal lines were generated as described above and screened for editing activity as for baseMEM-01.

To engineer cell lines with known barcode states, four separate 24wp wells with WT mESCs were transfected as described above with one of the four control plasmid vectors each, then selected with 15 ng/mL blasticidin. Polyclonal lines were used directly in Zombie imaging experiments.

Next generation sequencing.

Genomic DNA was extracted from cells using the DNeasy Blood and Tissue Kit (Qiagen) according to manufacturer instructions. Amplicon libraries containing the dynamic barcode sequences and short NGS static barcodes were generated with a two-step PCR protocol to add Illumina adapters and Nextera i5 and i7 combinatorial indices. Indexed amplicons were pooled and sequenced on the Illumina MiSeq platform with a 600-cycle, v3 reagent kit (Illumina, MS-102-3003). Raw FASTQ files were aligned to a FASTA-format reference file containing the expected amplicon sequences. Alignment was performed using the Burrows-Wheeler alignment tool (bwa-mem[106]). Subsequent analysis and data visualization were performed in the R statistical computing platform, v 4.1.1[107] (**Supplemental Data**).

Edit accumulation model

Edit accumulation at each target site was modeled by fitting Equation 1:

$$E = \frac{p}{\ln(1-p)}[(1-p)^{t+d} - 1].$$

Here, edit accumulation, E, is a function of time, t, with parameters p, the probability of editing per unit time, and d, the duration of time during which edits accumulated prior to the zero time point, which accounts for empirically observed background edits (**Figure 2B**). This relation can be derived by assuming a probability p of a target site being editing per unit time t in a long string of target sites. After a unit of time t, we expect to see p edited targets and $(1-p)$ unedited targets. By the same logic, after another time step we expect $p + p(1-p)$ edited targets and $(1-p)^2$ unedited targets. After T time steps we would expect to see

$$p\sum_{t=0}^{T-1}(1-p)^t$$

edited targets. Taking the limit of a discrete time step dt approaching zero, this sum can be approximated by the integral

$$p\int_0^T (1-p)^t dt$$

which simplifies to **Equation 1**.

Parameters were fit to editing time course data (**Figure 2**) to determine the empirical edit accumulation rate for each target using the "nls" function from the "stats" package[107] in R (**Supplemental Data**).

Stochastic simulations

Barcode editing was simulated in R using the Gillespie method(Gillespie 1977) (**Supplemental Data**). Separate propensities were estimated for each editing outcome and target site by multiplying the edit accumulation parameter p (**Equation 1**) for each target site by the observed mean outcome proportion at each target site across time (**Figure 2C**). This stochastic simulation method recapitulates both the edit accumulation model fit and the empirical target state outcome distribution (**Supplemental Figure 4**).

Cell division was modeled by allowing editing until a predetermined cell division time, after which barcodes were duplicated before allowing editing to continue. Cell division waiting times were drawn from a distribution derived from Eyring-Stover survival theory that has been shown to model cell division times more accurately than the exponential distribution[108].

Lineage relationships were reconstructed based on the resulting barcodes using BEAST2 software as described below, considering only barcode data (see BEAST2 XML files for complete modeling information, available at https://doi.org/10.22002/327t7-ke088). Reconstructed trees were compared to simulated ground truth trees by computing the normalized Robinson-Foulds distance as implemented in the "RF.dist" function from the R package "phangorn"[109].

Zombie preparation

For Zombie and subsequent RNA-FISH, cells were plated on treated coverslips as described above. After culture and editing, coverslips were washed with 1 mL PBS with calcium and magnesium (PBS +/+) then fixed with a 1:1 solution of Methanol : Acetic Acid (MAA) for 20 minutes. RNase-free reagents were used for all subsequent steps to minimize RNA degradation. MAA was removed, then coverslips were transferred to a 100 mm petri dish and covered with 70% ethanol. Petri dishes were parafilmed and stored at -20 C to await imaging.

Immediately prior to imaging, coverslips were removed from cold storage and brought to room temperature. 70% ethanol was removed and replaced with a fresh solution of MAA, then incubated for 2 hrs at room temperature. The sample was washed twice with PBS +/+, incubating for 2-3 min between each wash. The final wash solution was removed and the sample was dried until all liquid had just evaporated. A custom fluidic cell, built to interface with a custom designed liquid handling system, was affixed to the coverslip surface[54–56]. Subsequent washes took place within the flow cell, manually adding reagents into the inlet of the cell and removing them from the outlet using a standard micropipette. The cells were washed with nuclease free water once, then replaced with T7 RNAP mix (New England Biolabs E2040S). The sample was incubated at 37 C overnight in a humidified tupperware.

The following morning, the T7 RNAP mix was removed and replaced with fresh T7 RNAP mix, then incubated for 1 hr at 37 C in the humidified tupperware. The mix was removed, then the sample was immediately fixed with 4% paraformaldehyde for 10 min. This solution

was removed and the sample was washed three times with PBS +/+, then washed with 30% formamide probe wash buffer (30% formamide in 5x SSC with 9 mM citric acid, 0.1% Tween-20, and 50 μg/mL heparin, pH 6.0) for an additional 5 min. The wash buffer was replaced with primary probe hybridization mix, then incubated overnight at 37 C.

FISH imaging (Figures 3 and 5)

Images were collected across multiple rounds of fluorescence hybridization to identify barcode and cell states. Formamide wash buffers and secondary probe hybridization mixes were generated immediately prior to imaging. A custom-built, automated liquid handling system was used to perform sequential rounds of in situ hybridization as previously described[54–56]. Briefly, the sample was connected to an automated fluidics system attached to a widefield fluorescence Nikon Eclipse Ti microscope. The custom-made automated fluid sampler was used to transfer readout probes in hybridization buffer from a 2.0 mL 96 well plate through a fluidic valve (IDEX Health & Science EZ1213-820-4) to the custom-made flow cell using a syringe pump (Hamilton Company 63133-01). Fluidics and imaging were integrated using a custom script controlling uManager. Eleven fields of view (FOVs), capturing eight well separated regions of cell growth, were selected based on the DAPI signal. For each FOV, images were acquired with 0.5-micron z steps for twenty total slices. Integration of the automated fluidics system and imaging was controlled by a custom script written in uManager[110].

First, 12 hybridization rounds were imaged to capture all dynamic barcode states. The hybridization buffer for each round included two unique 15-nucleotide readout probes

(**Supplemental Data**) conjugated to either Alexa Fluor 647 (50 nM) or Alexa Fluor 546 (50 nM) in EC buffer (10% ethylene carbonate, 10% low molecular weight dextran sulfate, 4x SSC).

Probes were allowed to hybridize for 15 minutes. Excess probes were washed away with 10% wash buffer (10% formamide, 0.1% Triton X-100 in 2x SSC) incubating for 1 minute. Nuclei were re-stained with DAPI solution (5 ug/mL DAPI in 4xSSC) incubating for 2 minutes. The sample was washed with 4x SSC then imaged in antibleaching buffer (50 mM Tris-HCl pH 8.0, 300 mM NaCl, 2xSSC, 3 mM trolox, 0.8% D-glucose, 1,000-fold diluted catalase, 0.5 mg/mL glucose oxidase). After imaging, readout probes were stripped off using 35% wash buffer (35% formamide, 0.1% Triton X-100 in 2x SSC). Although 55% formamide is typical for stripping readout probes, we used a lower amount to avoid stripping primary probes and losing signal, as our primary probes are shorter than normal for dynamic barcode rounds (only 20-nucleotides compared to 28). Images were collected after probe stripping to verify loss of signal. Due to occasional technical issues such as loss of focus during automated imaging, these 12 rounds were repeated a second time to collect backup images for each dynamic barcode round.

Static barcode sequences were captured by a similar scheme over 20 additional rounds of hybridization (see **Supplemental Data** for probe sequences), except using 55% formamide wash buffer to strip the readout probes. An additional six rounds of hybridization were used to capture the 12 gene markers described above. A final round of hybridization with wheat

germ agglutinin (WGA) conjugated to Alexa Fluor 647 was used to stain cell membranes for downstream segmentation.

FISH Imaging (Figure 6, Supplemental Figure 3)

Imaging of control cell lines and tissue sections was performed similarly, but using a confocal microscope to capture 3D optical sections. Images were acquired on a Nikon Eclipse TI microscope equipped with a confocal scanner unit (Yokogawa CSU-W1), a sCMOS camper (Andor Zyla 4.2), 60x objective lens (Nikon 1.42 NA), and a motorized stage (ASI MS2000). Lasers were controlled using a LUN-F XL 7-line laser unit (Nikon, channels 405 nm, 440 nm, 488 nm, 514 nm, 594 nm, 640 nm). This microscope setup was coupled to an identical liquid handling system for FISH automation. For control samples, barcode information was recovered across 28 rounds of imaging. For in vivo tissue sections, three-color imaging was used to enable capture of gene expression and barcode information across 46 rounds of imaging. Imaging rounds were occasionally repeated due to loss of focus.

Image processing

Images were processed using custom Matlab scripts (**Supplemental Data**). DAPI signal was measured in each round of imaging and used to register images across each hybridization round. After registration, z-stacks were projected by their maximum intensity to yield one image per channel per hybridization round for each colony.

Transcribed barcodes form dots of variable intensity around the active site of T7 transcription. Dots were segmented using a combination of Laplacian of Gaussian filtering

and watershed, requiring a maximum eccentricity of 0.8 to reduce noise. Loose parameters were chosen to detect all real dots at the expense of accepting some background noise. Binary images for each hybridization/channel were summed together to create a single mask, where pixel values represent the number of times a pixel was identified across all imaging rounds, termed “analog mask”.

Since each real dot should appear across all hybridization rounds in at least one channel, we further reduced noise by thresholding this image. We determined a threshold for each colony individually based on the elbow method. Frequently, we observed segmentation errors where the watershed algorithm was unable to separate adjacent dots from one another. We manually corrected these errors based on the analog mask, yielding a final binary segmentation mask of all detected barcode dots for each field of view.

We further generated DAPI segmentation masks using Ilastik to isolate individual cell nuclei[111]. Masks were manually corrected using ImageJ to separate nuclei that were segmented together. Cells that intersected the border of the image were excluded. Any Zombie dots identified outside of cell nuclei were filtered. Dots may not be completely captured by the binary mask within any given round of imaging. A K-nearest neighbors classifier was used to partition all pixels belonging to each cell to the nearest dot in the segmentation mask so that intensity values could be extracted.

Raw images were background subtracted to improve signal to noise. First, a tophat filter was used to globally reduce background. To further correct for variable intensity across images,

local background correction was applied on a cell-by-cell basis by subtracting the median pixel intensity, excluding dilated segmented Zombie dots.

We extracted several features for each dot across all channels and hybridization rounds based on the background-subtracted raw images (total intensity; median intensity; 90^{th} percentile pixel intensity; pixel count; background median intensity; and intensity variance), taking the log + 1 of all intensity values.

We used a supervised machine learning approach to classify barcode states across each hybridization round. The barcode state is reflected in higher intensity fluorescence of probes that outcompete other possible binders (**Figure 3**). We manually classified approximately 1,000 randomly sampled barcode spots for each image based on their pseudocolor intensities, then used this sample to train a support vector machine (SVM) classifier in Matlab (**Supplemental Data**). Some spots were ambiguous; these were omitted in manual classification. Ten-fold cross validation was used to evaluate model generalization and control for overfitting (**Supplemental Figure 2**).

For dynamic barcode sites, we estimated the posterior probability for each spot belonging to each class under the SVM model. Many barcodes could be classified with high accuracy (>70% posterior probability, **Figure 3F** and **Supplemental Figure 2B**). For static barcode sites, class assignments were compared to the whitelist of possible barcode sequences. We filtered out Zombie spots with a character distance greater than two from an expected sequence and those which did not unambiguously correspond to a whitelisted static barcode, leaving 79.3% of all detected spots after filtering.

In many cases, duplicated barcodes were observed, where the same static barcode was identified multiple times in a single cell. These duplicates tended to be spatially localized and may be explained by either DNA replication or over-segmentation errors during analysis. For duplicated barcodes, classification probabilities were averaged at dynamic barcode sites and the most confident state was used for downstream lineage reconstruction.

Membrane masks were manually generated based on WGA staining images, then gene markers were identified using the bigFISH package dot detection method.[112] Thresholds for dot detection were manually determined for each gene. Segmented spots, corresponding to mRNA molecules, were tallied within each cell as defined by the membrane segmentation mask. To validate the consistency of this method, we plotted the detection frequency for each gene across all cells that were measured in multiple images (**Supplemental Figure 11**). The measures were highly correlated.

A similar process was used to analyze 3D images of tetraploid sections (**Figure 6**) and fixed barcode control cells (**Supplemental Figure 3**) collected through confocal imaging, with some modifications (**Supplemental Data**). Z-stacks were not projected to their maximum intensity but rather treated in 3D throughout analysis. Analog masks with Zombie dot detections were corrected using Cellpose-SAM[113] (v4.0.6) rather than using watershed and manual correction. Cellpose-SAM was additionally used to create cell segmentation masks based on both DAPI and WGA signal. Masks were manually inspected and corrected as needed in the Cellpose GUI. To decode static barcode rounds where either 1 or 2

pseudocolors are expected, a second SVM model was trained to classify resulting images (**Supplemental Figure 8C, Supplemental Data**).

Cell type analysis

Cell types in vitro (**Figure 5**) were determined by using k-means clustering on log transformed mRNA counts with five centers to group the most distinct sets of cells in the dataset. Dimensionality reduction by the tSNE method visualizes three groups as well separated and three of the identified cell states (Naive/2C-like, Naive, and Naive/Formative) as potentially a continuous distribution, although we note that dimensionality reduction techniques can obscure the true distances between cells and clusters in the higher dimensional transcriptome space. Most importantly, we identify unique allowed and forbidden transitions between each purported cell state through subsequent lineage analysis that is agnostic to the underlying transcriptional information, bolstering the claim that these five clusters of cells should be treated as distinct populations.

Cell types in vivo (**Figure 6**) were determined through Leiden clustering of gene expression information using Squidpy[114] (**Supplemental Data**). Cell types were expertly annotated based on marker gene expression and physical location of cells within the embryo section.

Lineage reconstruction and Bayesian modeling

We used a Bayesian model under the BEAST2[73] v2.7 framework that takes barcode information, end point cell labels, and cell centroid positions as input to jointly estimate lineage relationships, cell state transition dynamics, and cell motility. An XML file

specifying all modeling information is provided as supplemental data and modeling choices are briefly described below.

Barcode information for each dinucleotide was extracted using Matlab and R scripts (**Supplemental Data**). Each of the four dinucleotide states (AA, AG, GA, and GG) was encoded in a single character (A, T, C, or G). Characters that were not recovered during imaging were marked as missing data by the "?" character. Cell division for each tree was modeled as a pure birth process (the Yule model) with birth rate estimated.

Barcode character mutation in our system is irreversible. With few exceptions[115], existing BEAST2 packages only model reversible character transitions because these make computing tree likelihoods more computationally efficient. We developed a new irreversible character substitution model to better capture the evolutionary process that generated our data (available as the 'irreversible' package for BEAST2, with source code available from https://github.com/rbouckaert/irreversible). Under this model, each possible transition (AA to AG, GA, or GG) can take a unique rate value, which we assume is constant along the tree. Stationary frequencies, which are used at the root of the tree to calculate the tree likelihood, are set at 1 for the AA state, and 0 for the others, reflecting our knowledge that every state is AA at the root of the tree. Since we know that targets can edit at different rates and into different outcomes, we allow the rate to vary across sites through the gamma site heterogeneity model, partitioning the allowable rates into four categories[116]. This model is shared across all trees. Furthermore, we use a strict molecular clock since we do not expect significant rate variation per branch.

Cell state transitions are modeled as a continuous time Markov chain with symmetrical transition rates possible between each state. These rates are assumed to be constant across time with an associated strict clock model. Transition rates are shared between all trees, so a single unified cell state transition model was estimated across all colonies. We assume symmetric transition rates based on previous work, which identified most transitions as roughly symmetric in this system[68]. In principle, the assumption can be relaxed, although it greatly increases the number of parameters in the model, thus increasing susceptibility to overfitting.

Cell motility was modeled as single parameter 2D diffusion along the surface of a sphere as implemented in previous phylogeographical work[102]. Spherical diffusion is a good approximation of diffusion in a 2D plane for small patches of the surface[102] and its implementation is efficient. Accordingly, cell positions were mapped to geographical coordinates falling within two latitude and longitude degrees. The diffusion parameter describing motility was allowed to take unique values along each branch of the tree under a relaxed clock model.

Notably, all seven colonies in this dataset were analyzed simultaneously under a single model. This allowed us to infer barcode character substitution and cell state transition models that are shared across all the data, reflecting our belief that all colonies are representative of the same underlying barcode mutation and cell state transition processes. We think this is reasonable given that all colonies are generated from a monoclonal culture grown in identical culture conditions over the same time period.

We chose priors to be uninformative with the exception of the root height, since we have strong prior information that experiments lasted three days. An uninformative but improper uniform distribution across all possible rates was chosen for barcode mutation rate, although this is not expected to affect the resulting analysis or MCMC mixing. Detailed prior information is recorded in the supplemental XML file for Figure 5 specifying all modeling choices.

Supplementary movies were generated by creating inferred still images of the maximum a posteriori histories of cells over time, incorporating inferred ancestral cell states, positions, and cell division timings (**Supplemental Data**). These still images were compiled into movies using the open-source video editor Shotcut (Meltytech).

For in vivo lineage reconstruction, we used BEAST2 as described above but reconstructing only based on genetic barcode information. Convergence of the MCMC chain was monitored using Tracer v1.7.2 (https://github.com/beast-dev/tracer/releases/tag/v1.7.2). The resulting distribution was summarized as a consensus maximum clade credibility tree using the TreeAnnotator tool packaged with BEAST2. Further analysis was performed using custom R scripts (**Supplemental Data**).

Lineage motif analysis

The posterior baseMEMOIR trees were analyzed using Lineage Motif Analysis (LMA) as described previously[87], using the resample_trees_doublets, resample_trees_triplets, and resample_trees_quartets functions with 1,000 resamples. These functions are available in the

publicly available "linmo" package for Python (https://github.com/labowitz/linmo). To generate a z-score and adjusted p-value for all cell fate patterns across the entire posterior distribution of each tree dataset, 1,000 synthetic datasets were generated by randomly drawing one tree from the posterior distribution of each tree dataset. Each synthetic dataset therefore contains seven total trees. LMA was then performed on each synthetic dataset, and the distribution of z-scores and adjusted p-values was plotted for each cell fate pattern.

Median phylogenetic distance comparison (Figure 6F and Supplemental Figure 10)

To identify significantly closer or more distantly related cells than expected by chance, pairwise phylogenetic distances were computed across all cell pairs that were captured in the phylogeny with a measured cell type (**Figure 6D**). The phylogenetic distance was defined as the number of splits in the tree along the shortest path between the pair of cells. This distance was computed for 800 samples of the posterior distribution taken from the output of BEAST2 after removing burn in, then averaged for each cell pair.

We then focused on specific cell type comparisons. One or two cell types were chosen, and posterior-averaged phylogenetic distances for all pairs matching the comparison of interest were extracted. The median of these distances was computed. One thousand samples of cells, sized equal to the number of cells present in the comparison of interest, were randomly chosen from the entire phylogeny, and the same procedure was repeated. We then computed a z-score for the experimental median relative to the medians of the randomly sampled control distributions. From this z-score we computed p-values assuming a two-tailed test, and corrected these for false discovery rate via the Benjamini-Hochberg procedure[117].

Statistically significant z-scores (adjusted p-value < 0.05) were plotted (**Figure 6F and Supplemental Figure 10**).

In vivo experiments

For the generation of chimeric embryos, doxycycline-induced baseMEM-02 recording cells were injected into host wild-type nascent blastocysts (E3.5). Briefly, 6-12 week old B6D2F1 (C57BL/6J x DBA/2J, Jackson Laboratory) females were hormonally primed by intraperitoneal injection of pregnant mare serum gonadotropin (PMSG, Prospec Tany Technogene), followed 48 hours later by human chorionic gonadotropin (hCG, Patterson Veterinary). Embryos were collected at the zygote stage (E0.5) and cultured to the blastocyst stage. On the day of injection, 8-12 cells were introduced into the blastocoel cavity using a 16-micron-diameter injection pipette (Biomedical Instruments).

For tetraploid complementation, two-cell embryos were fused into one-cell embryos using a ECM 2001 (BTX) with a single DC square pulse of 150 V for 60 microseconds and 1-2 V AC, in 0.3 M mannitol solution containing BSA. Approximately 15 cells were injected into tetraploid blastocysts. Roughly 15 embryos were transferred into each CD1 recipient female (Crl:CD1 (ICR), Charles River Laboratories). The day of injection was considered E2.5. Pseudo-pregnant mice were administered doxycycline via drinking water and food, provided *ad libitum* from E3.5 until euthanasia at E7.5. All mice were handled in accordance with Caltech's institutional guidelines, and all procedures were approved by the Institutional Animal Care and Use Committee (IACUC protocol IA23-1742).

Embryo implantation sites were dissected at E7.5, embedded in O.C.T. (Tissue-Tek) in a dry ice ethanol bath, and cryopreserved at -80 C. Twenty-micron thick sections were taken from the embedded tissues and placed on Poly-D-Lysine (Gibco) coated coverslips on dry ice for subsequent automated imaging experiments.

Samples were prepared for imaging following the procedures used for in vitro cultured cells with several modifications. The slide was allowed to come to room temperature and air dry for 10 minutes prior to permeabilization with MAA. After MAA fixation, sections were permeabilized with 1% Triton X-100 (Sigma-Aldrich, 93443) in PBS for 1 hour at room temperature. Subsequently, the section was washed three times with PBS and treated for 15 minutes at room temperature with 0.1 M freshly prepared hydrochloric acid. Samples were then washed with nuclease free water, briefly washed with 0.1% Triton X 100, then incubated for 48 hours with T3 RNAP (MEGAscript, ThermoFisher AM1338) prepared as for the in vitro experiments.

After transcription, samples were fixed for 10 minutes in 4% PFA as for in vitro samples, then permeabilized in 8% SDS dissolved in PBS for 10 minutes at room temperature. Samples were washed five times with PBS and two times with 2x SSC buffer, then once with 30% formamide wash buffer for 30 minutes at 37 C. Samples were then incubated with primary probes for 48 hours at 37 C.

On the fifth day, samples were washed 3x with 30% formamide wash buffer, then returned to 37 C for 1 hour. Samples were then washed three times with 2x SSC and incubated with a primer to enable bridge ligation[54] (**Supplemental Data**) for 4 hours at 37 C. Samples were

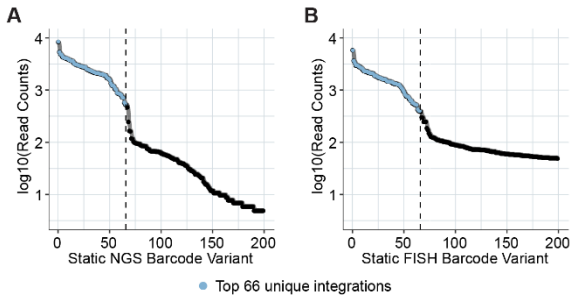
then washed once with 12.5% formamide wash buffer, then three times with 4x SSC at room temperature, then once with a 1:1 mixture of nuclease free water to Quick Ligase Reaction Buffer (NEB M2200L). Bridge ligation was allowed to occur overnight using the Quick Ligation kit.

On the sixth day, samples were washed again with 12.5% formamide wash buffer for 5 minutes at room temperature, then three times with 4xSSC. Samples were treated with Label-IT (Mirus Bio, MIR3925) at 37 C for 30 minutes, washed three times with PBS, then fixed with 1.5 mM BS(PEG)5 (ThermoFisher A35396) for 1 hour at room temperature. Finally, samples were washed three times with 1 M Tris-HCl, pH 7.4, then with 55% formamide wash buffer, then three times with 4x SSC buffer. At this point, samples were ready for imaging as with in vitro samples.

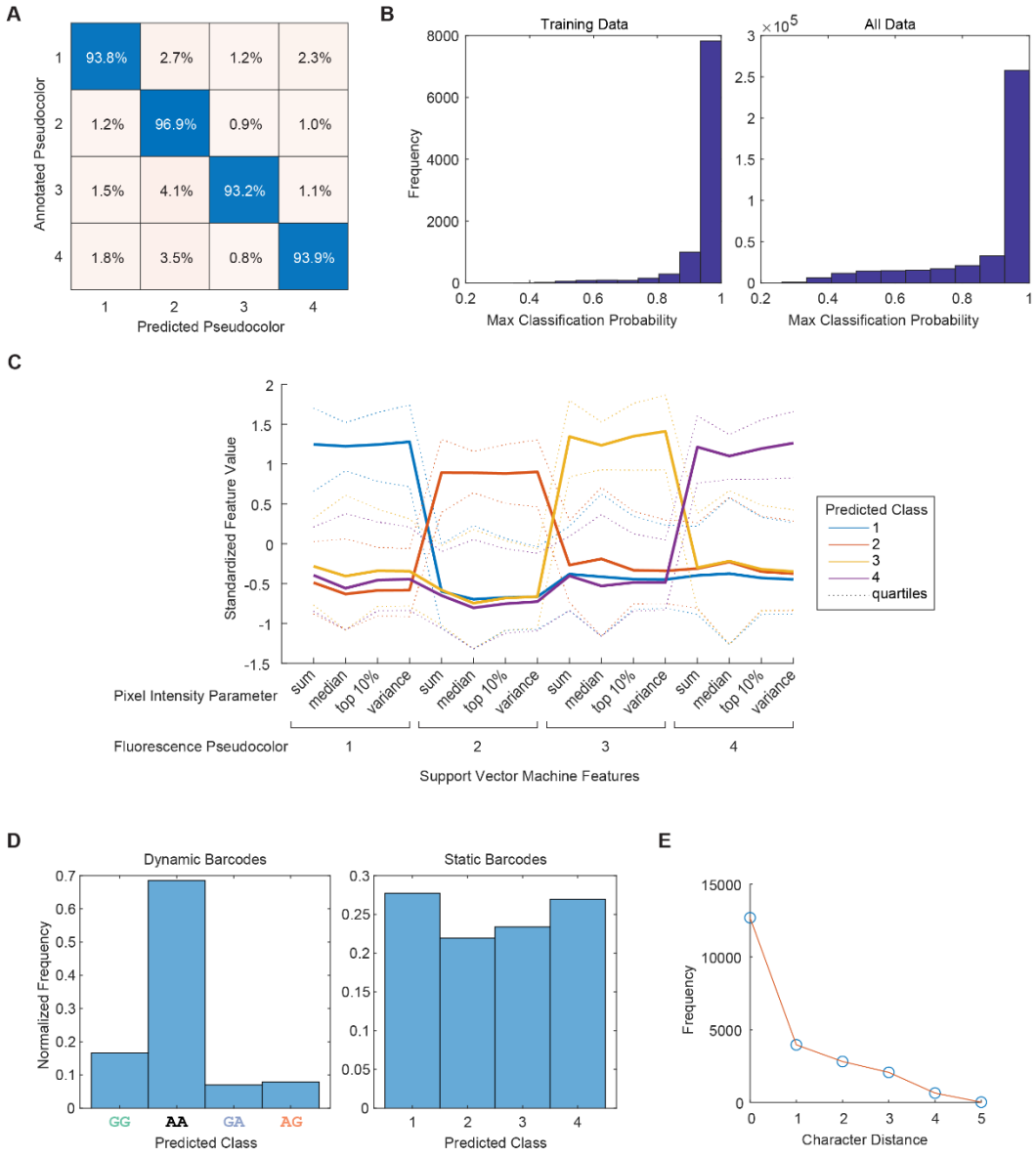
Data availability

Raw image data are available at data.caltech.edu, DOIs: 10.22002/d15ek-0dx91, 10.22002/q33zp-z2k11, and 10.22002/qjm4z-jzc10. All analysis scripts, amplicon sequencing data, max projected image data, and additional supplementary files are available at data.caltech.edu, DOIs: 10.22002/327t7-ke088, 10.22002/3n0t0-jvn33, 10.22002/sreyy-rky20, and 10.22002/4gvvq-hek65.

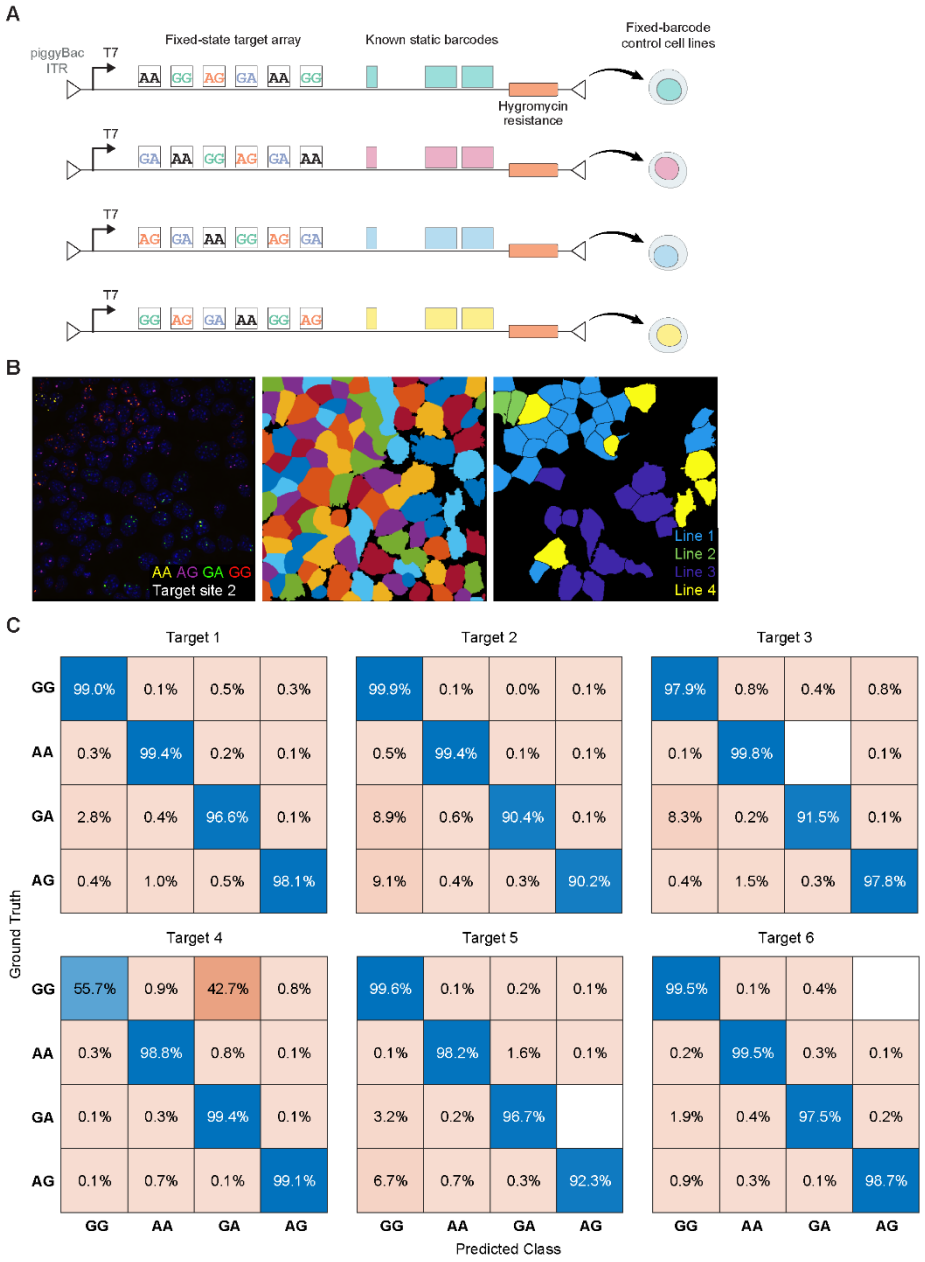
Supplemental information



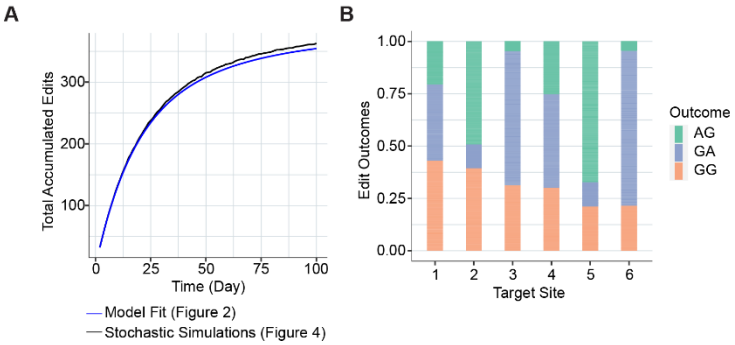
Supplemental Figure 1: 66 unique integrations are detected in the baseMEM-01 cell line. 66 barcode integrations were identified by next generation sequencing of target arrays amplified from genomic DNA. We quantified the number of reads corresponding to unique sequenceable (A) and image readable (B) static barcodes, identifying approximately 66 variants in each case. The top 200 most frequent variants are shown; we separated true variants from noise heuristically by identifying the "knee of the curve" (dashed vertical lines). Importantly, these 66 variants are all also identified in FISH experiments (Figure 3E).



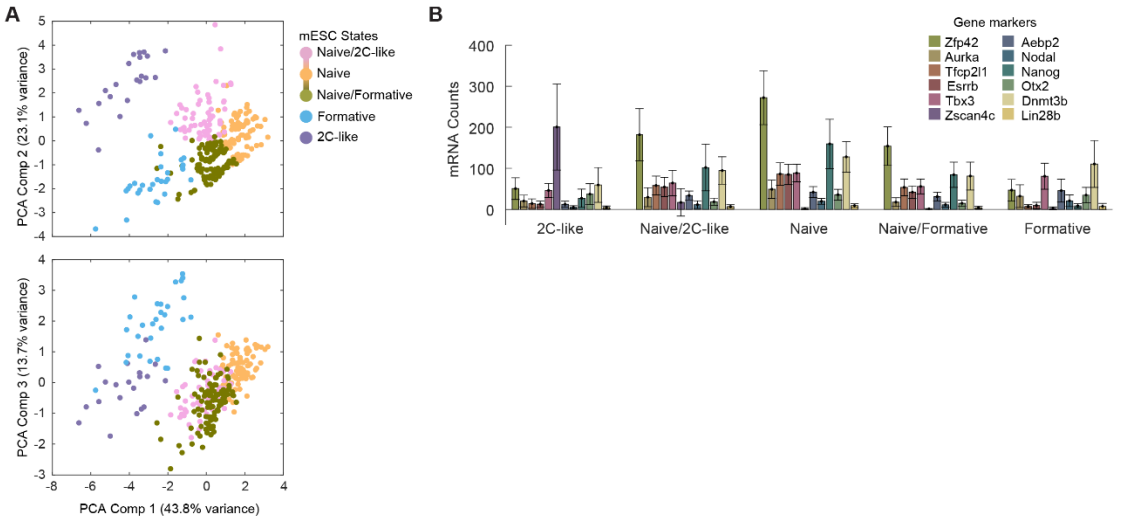
Supplemental Figure 2: A support vector machine classifies barcode states based on fluorescence measurements. (A) Manually annotated barcodes are correctly classified by a quadratic kernel support vector machine (SVM) approximately 94% of the time. **(B)** Classification probability estimates are very high within the training dataset **(B, left)**. Outside of the training sample, most classification probabilities are still high but with a subset of predictions that are less certain **(B, right)**. The support vector machine predicts classes based on 16 fluorescence measurements corresponding to each pseudocolor as defined in **Figure 3B**. **(C)** Each class is well separated based on these features. **(D)** After 3 days of editing induction, many dynamic barcodes are identified as class 2, corresponding to the unedited state **(D, left)**. Static barcode classifications are more evenly distributed, as anticipated **(D, right)**. **(E)** Static barcodes decoded by FISH typically perfectly match the 66 image readable barcode sequences identified by sequencing (**Supplemental Figure 1B**), although a fraction of barcodes are recovered with one or more character differences relative to their closest match.



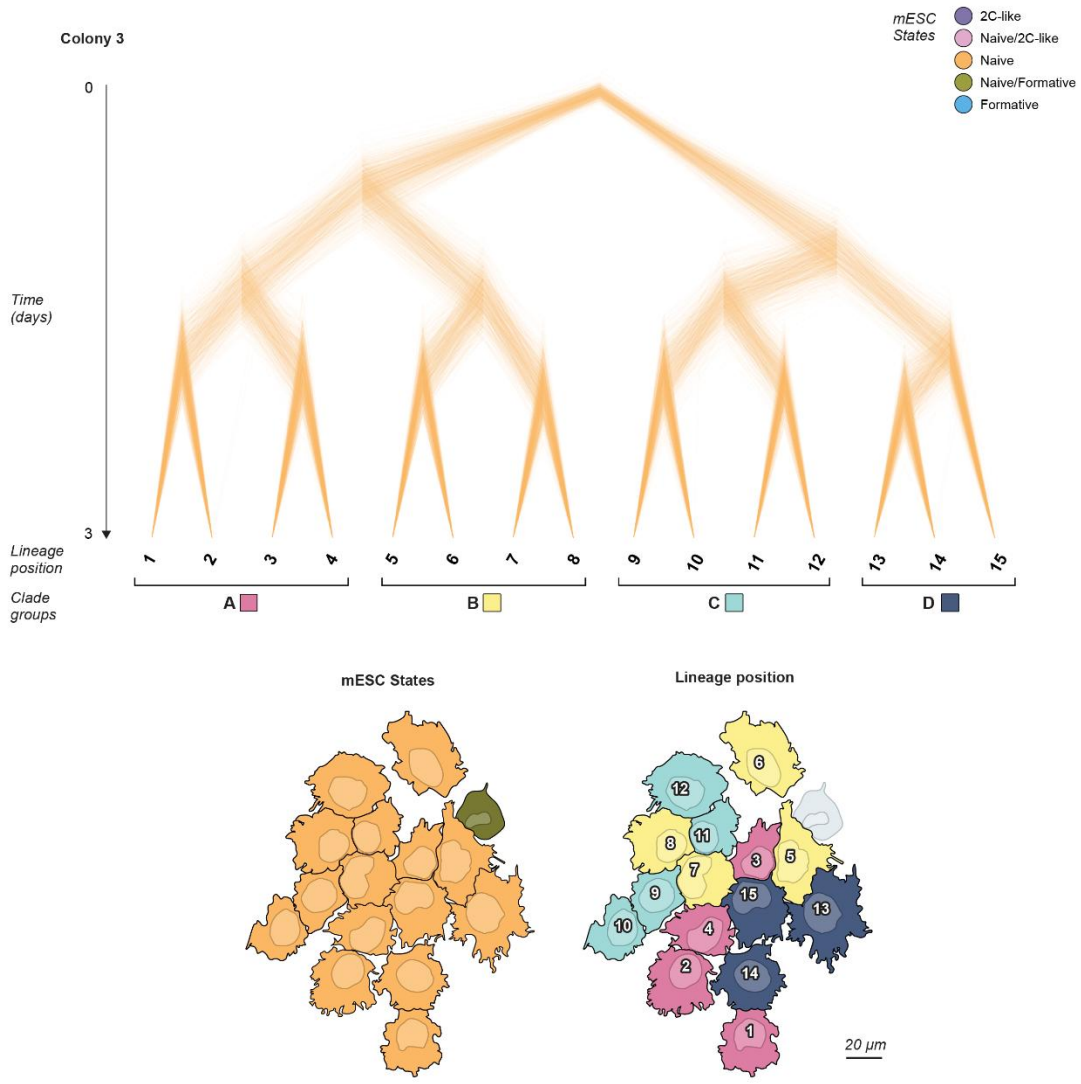
Supplemental Figure 3: Control cell lines reveal accurate dynamic barcode calling. (A) Four polyclonal cell lines were engineered with mock barcode arrays, each with a prescribed edit state and static barcode. **(B)** Cells were mixed, plated on a coverslip, and subjected to our standard FISH pipeline. Cells of each type were easily identifiable in each round of imaging (two representative hybridization rounds visualizing the state of target site 2 are depicted, left). We segmented cells (middle) and filtered to retain cells containing at least four identical array integrations, ensuring accuracy of array ground truth assignments (right). 64 total positions were imaged and treated in this way, measuring 3,311 total cells and 19,358 total barcode arrays for downstream accuracy analysis. **(C)** Dynamic barcode inference was typically very accurate, with a single edit outcome on one target (1/24) having lower accuracy.

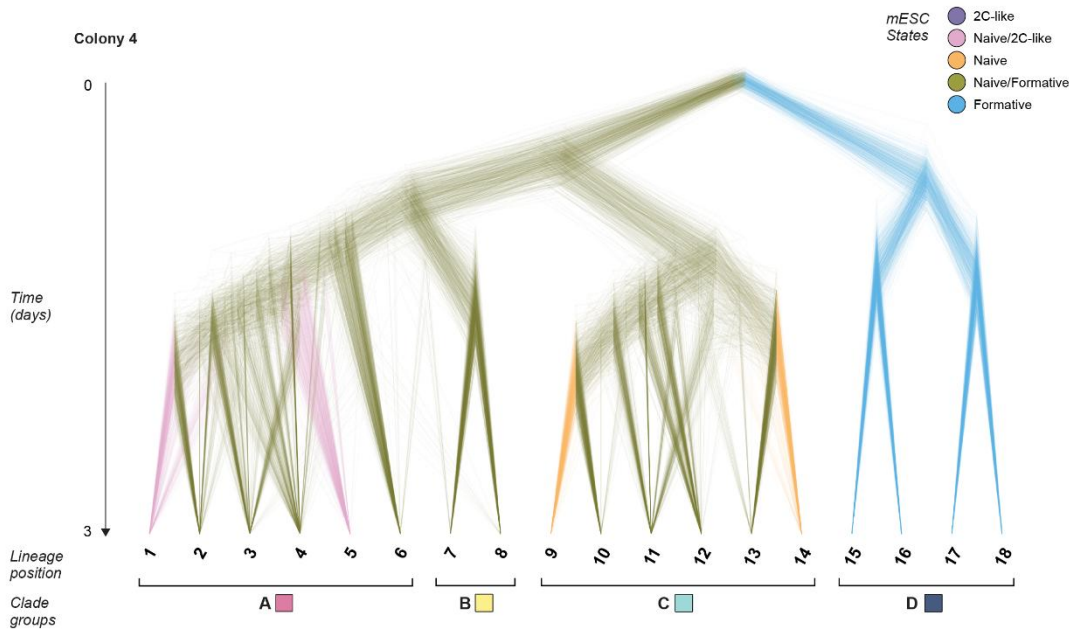


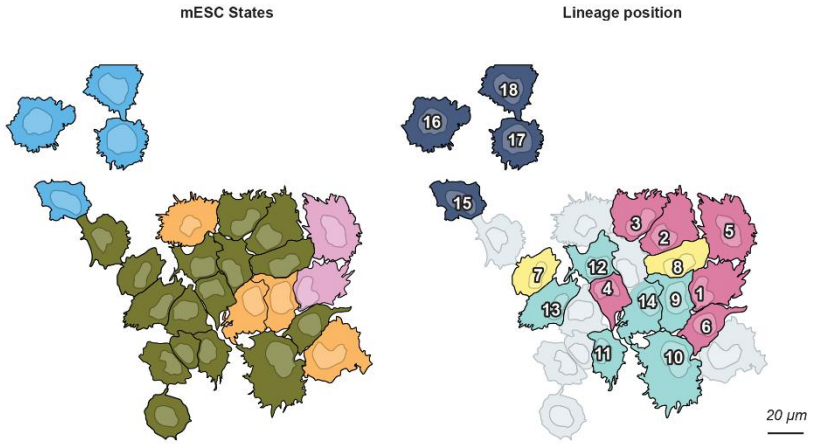
Supplemental Figure 4: Stochastic simulations closely recapitulate the empirical editing process. (A) We developed a stochastic editing simulator based on the Gillespie algorithm that closely recapitulates the average edit accumulation model developed in **Figure 2B** (**Methods**). **(B)** The simulated edit outcome distributions for each target site match the observed distributions from **Figure 2C**.

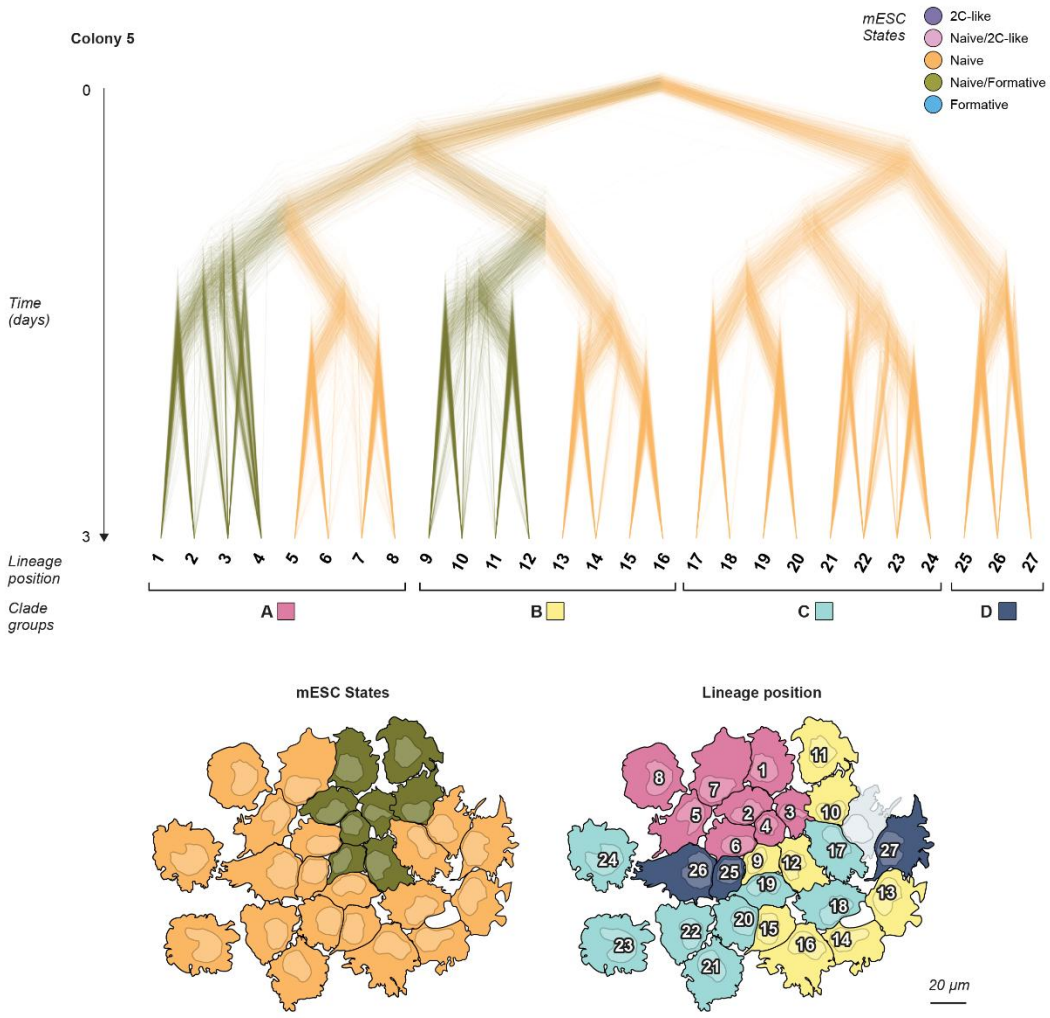


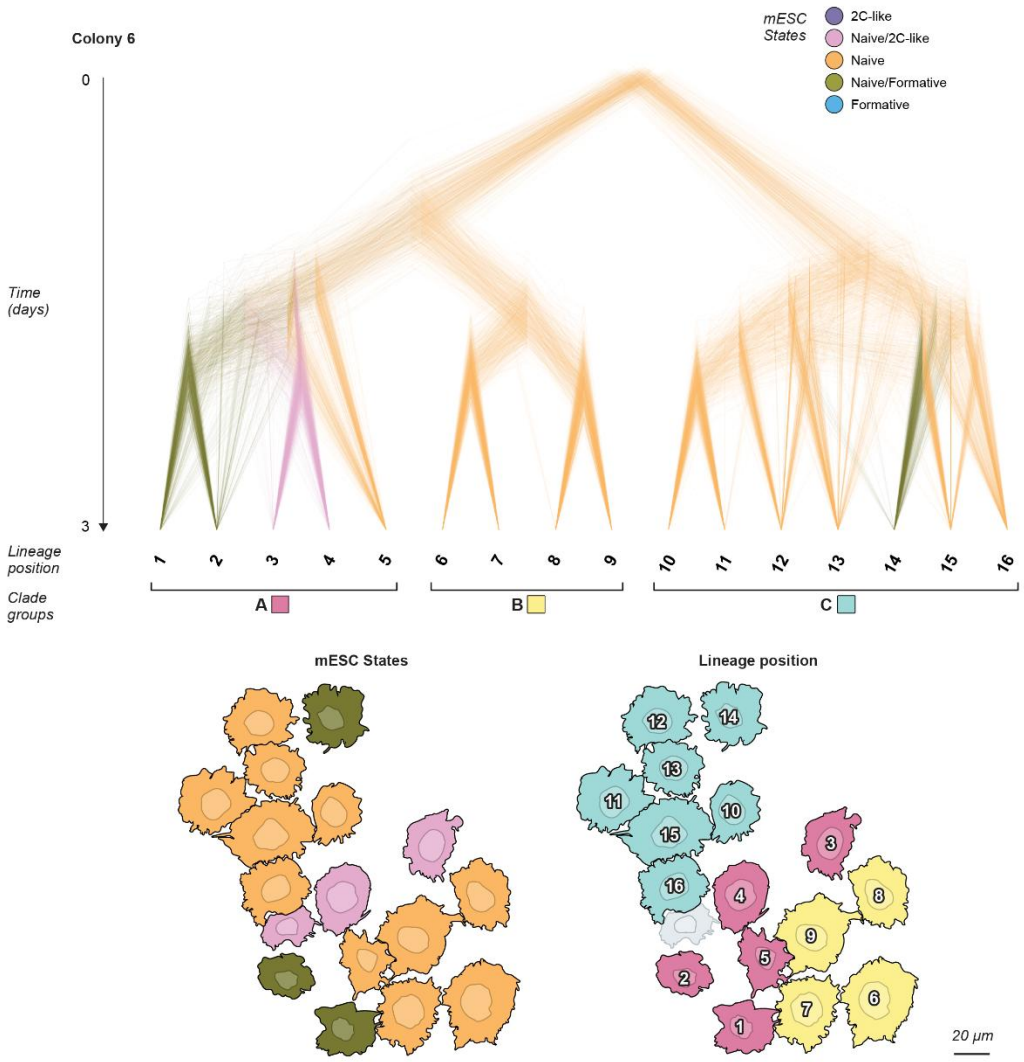
Supplemental Figure 5: mESC gene expression clustering. (A) Principal component analysis is largely in agreement with nonlinear dimensionality reduction, with separation between major clusters observed along the first three components. The naive states also appear continuously related in this view. **(B)** Clusters have distinct marker gene expression patterns, with some similarity between the Naive/2C-like, Naive, and Naive/Formative states.

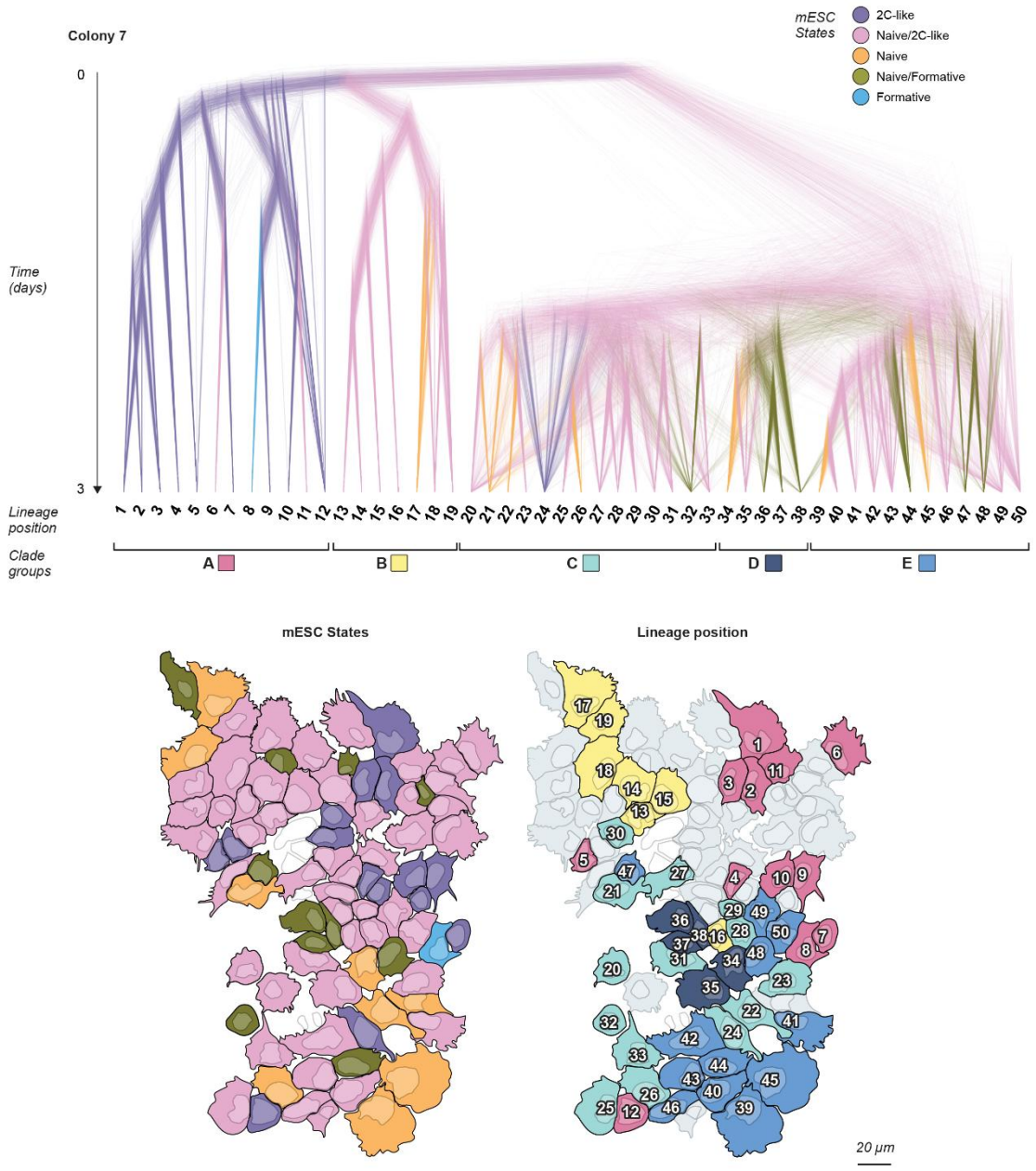

Colony 3
mESC States
2C-like
Naive/2C-like
Naive
Naive/Formative
Formative
0
Time (days)
3
Lineage position
Clade groups
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
A
B
C
D
mESC States
Lineage position
20 μm

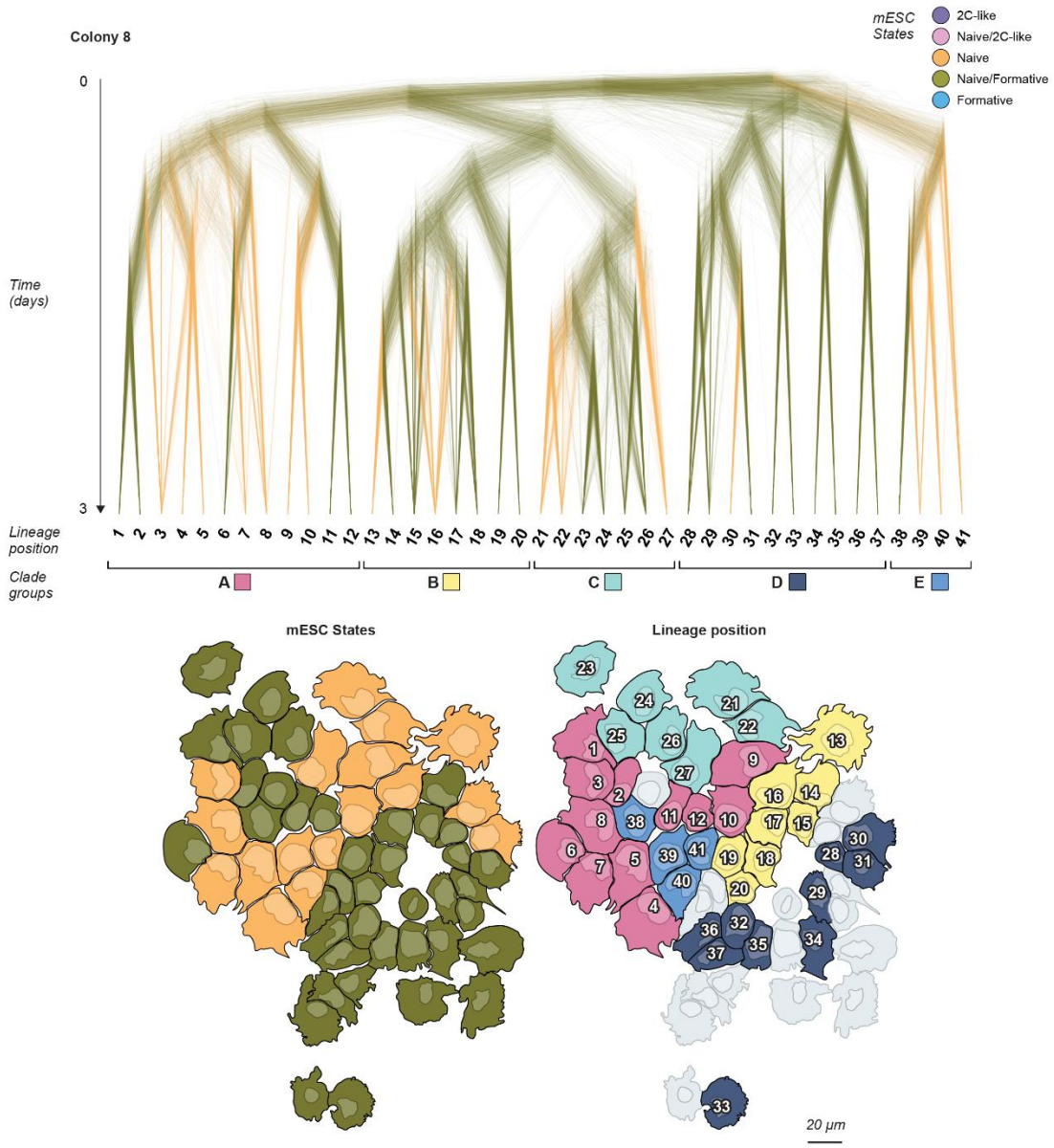

Colony 4
mESC States
2C-like
Naive/2C-like
Naive
Naive/Formative
Formative
0
3
Time (days)
Lineage position
Clade groups
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
A
B
C
D


mESC States
Lineage position
16
18
17
15
3
2
5
7
12
8
4
14
9
1
13
6
11
10
20 µm

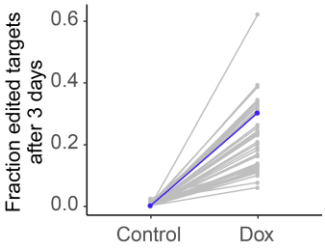

Colony 5
mESC States
2C-like
Naive/2C-like
Naive
Naive/Formative
Formative
0
3
Time (days)
Lineage position
Clade groups
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
A
B
C
D
mESC States
Lineage position
20 µm


Colony 6
mESC States
2C-like
Naive/2C-like
Naive
Naive/Formative
Formative
0
3
Time (days)
Lineage position
Clade groups
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
A
B
C
mESC States
Lineage position
20 µm

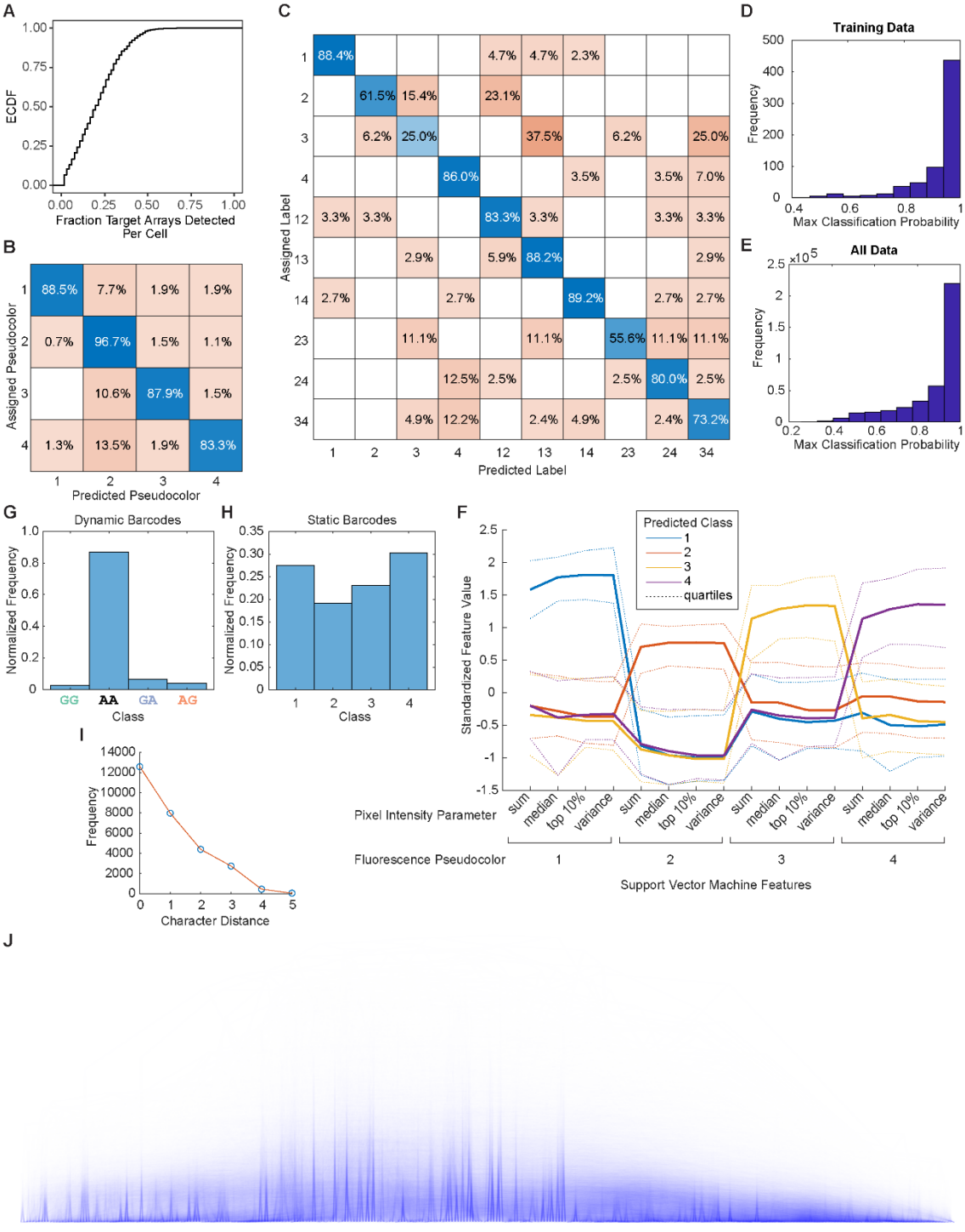

Colony 7
mESC States
2C-like
Naive/2C-like
Naive
Naive/Formative
Formative
0
3
Time (days)
Lineage position
Clade groups
A
B
C
D
E
mESC States
Lineage position
20 μm



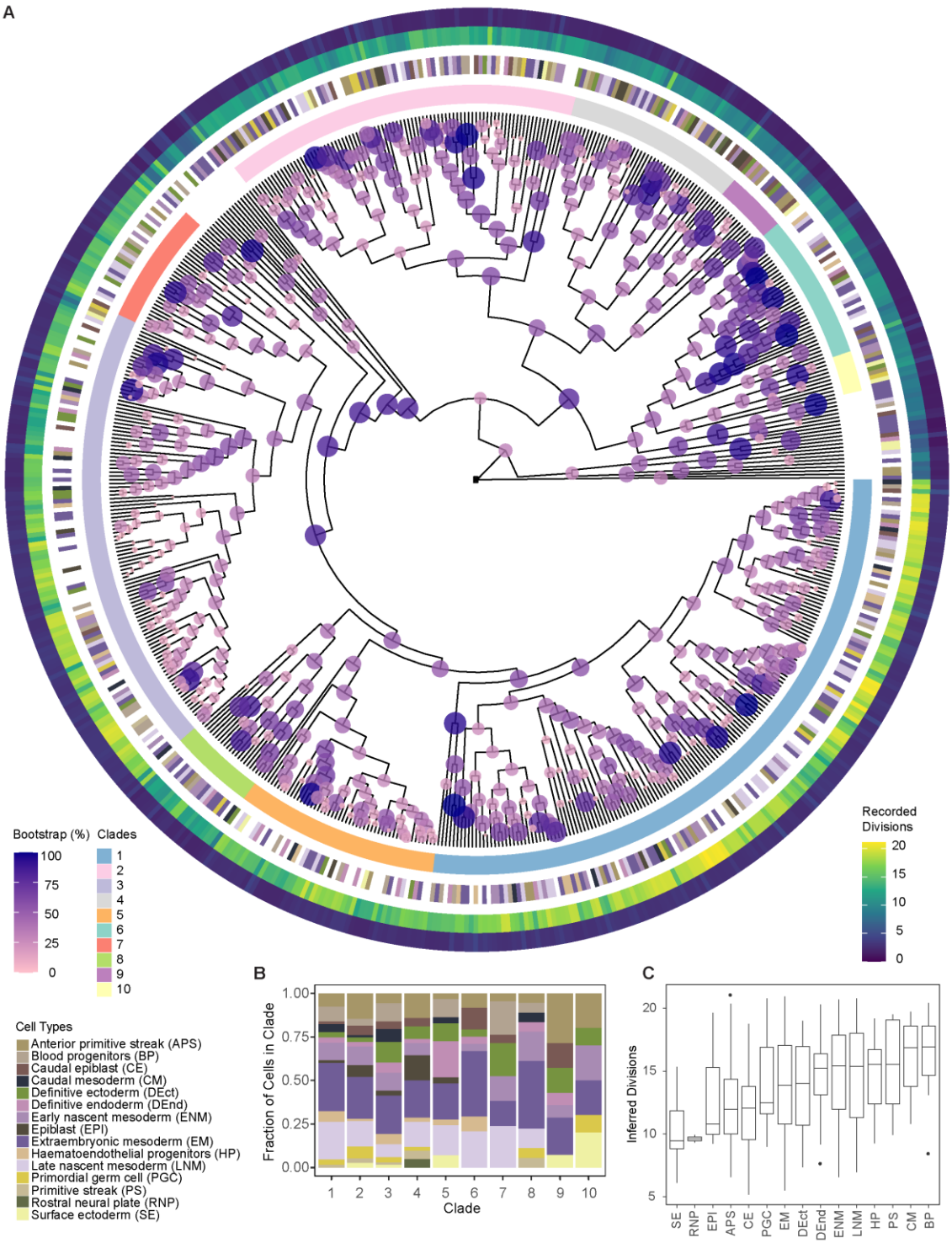
Supplemental Figure 6: BaseMEMOIR reveals lineage relationships, cell states, and spatial positions across multiple colonies. Posterior tree distributions are visualized and mapped back to illustrations of each colony as in **Figure 5D**.



Supplemental Figure 7: BaseMEM-02 clones edit in the presence of doxycycline. Fraction of edited target sites observed in control and dox-treated (1 ug/mL) samples after 3 days in culture is plotted for clonal mESC lines (**Methods**). The blue highlighted clone was used for subsequent experiments (**Figure 6**).

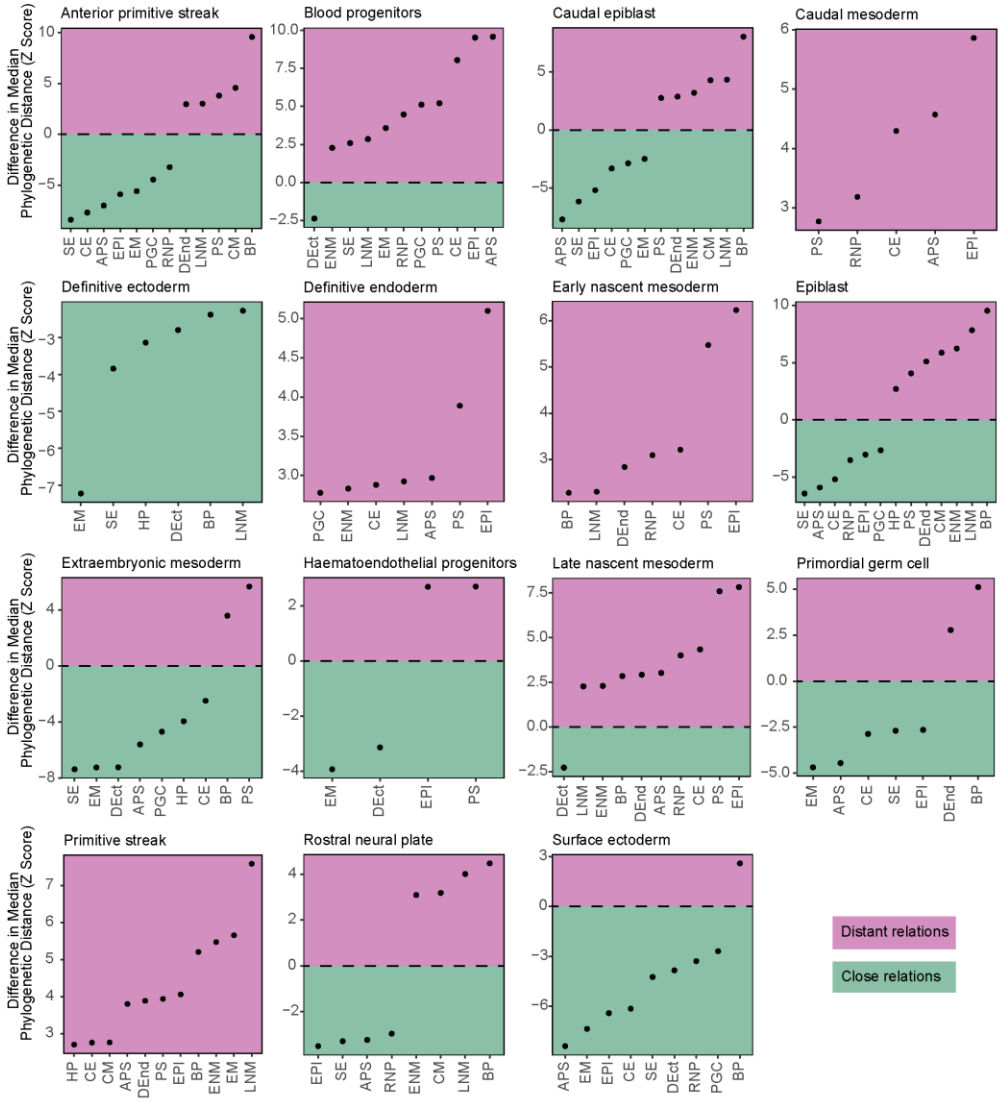

A
ECDF
Fraction Target Arrays Detected Per Cell
B
Assigned Pseudocolor
Predicted Pseudocolor
C
Assigned Label
Predicted Label
D
Training Data
Frequency
Max Classification Probability
E
All Data
Frequency
Max Classification Probability
F
Predicted Class
quartiles
Standardized Feature Value
Pixel Intensity Parameter
Fluorescence Pseudocolor
Support Vector Machine Features
G
Dynamic Barcodes
Normalized Frequency
Class
H
Static Barcodes
Normalized Frequency
Class
I
Frequency
Character Distance
J

Supplemental Figure 8: BaseMEMOIR barcode array recovery from E7.5 tetraploid embryo section. (A) Fewer target arrays were detected per cell compared to mESCs grown in culture (**Figure 3D**). **(B)** Manually annotated barcodes were typically accurately predicted using a support vector machine with quadratic kernel as in Supplemental Figure 2. **(C)** In this experiment, two static barcode rounds were allowed to take on double labels across the four measured pseudocolors (**Methods**). A separate support vector machine was used to predict outcomes for this classification task. Accuracy was lower than for single pseudocolor classification, however errors typically represented one of the two simultaneously measured channels; thus, overall errors to static barcode determination are reduced relative to the possibility of random errors. As in Supplemental Figure 2, classification probabilities remain high within the training set **(D)** and are typically high across the entire distribution **(E)**. **(F)** Pseudocolor classes are well separated based on spot intensity features. **(G)** After 3 days of editing by doxycycline administration in vivo (**Methods**), approximately 14% of barcodes were edited using the baseMEM-02 line. **(H)** Static barcode pseudocolor class predictions are more evenly spread, as anticipated. **(I)** Static barcodes decoded by FISH typically match one of the 66 image readable barcode sequences perfectly. Barcodes that could be uniquely classified and contained three or fewer mismatches relative to the barcode whitelist were carried on for further analysis. **(J)** Compiled barcodes were used to generate a posterior phylogenetic tree distribution in BEAST2. Lineage relationships were less certain than observed for mESCs in culture, likely due to lower barcode recovery and decreased editing in vivo. Even so, confident clades and consistent features of the posterior distribution could be identified (**Figure 6**).

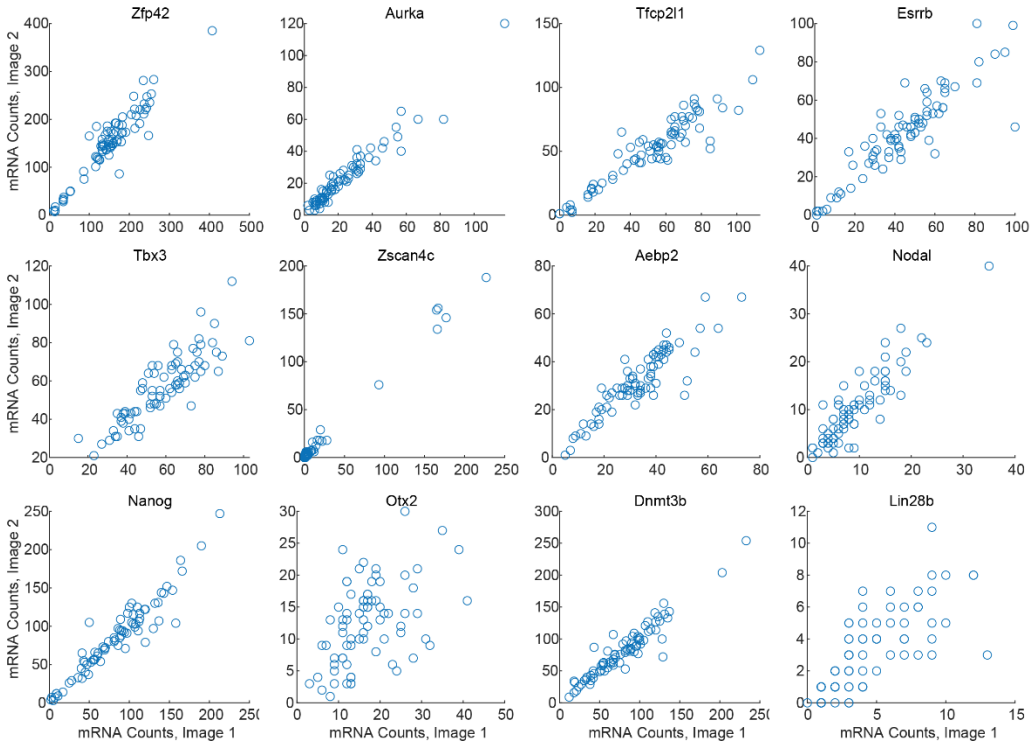


Supplemental Figure 9: Lineage relationships were resolved among hundreds of cells in the developing E7.5 mouse embryo. (A) A larger image of the full tree with transfer bootstrap scores on each clade is depicted. The central ring defines different clades as referred to in panel **(B)**. The middle ring displays cell types corresponding to each taxon, where white indicates missing cell type annotations. The outer two rings give the recorded divisions for each taxon as mean root-to-tip depth (inner ring) and the standard deviation (outer ring) across 800 BEAST2 posterior samples. **(B)** Cell types were widely mixed across clades. This may be expected, given known cell type mixing within embryos at this stage and the fact that we

began recording with approximately 15 founder cells at E4.5, which may all contribute to a variety of tissues. **(C)** Different cell types tended to have different tip depth distributions. Although interesting, the number of inferred divisions may be inflated in some cases since we began with 15 founder cells, which likely each form a separate larger clade on the tree. Future work will separately label founder cells to avoid this confounding factor.



Supplemental Figure 10: Many cell types are more or less closely related to one another than expected by chance. The median phylogenetic distances between pairwise cell comparisons (**Figure 6D**) were computed, then compared to a null distribution sampled from the full lineage tree to determine z-scores. Results were filtered for significance, defined by a false discovery rate corrected p-value of < 0.05 (**Methods**).



Supplemental Figure 11: Gene detection is consistent across images. Stochastic simulations closely recapitulate the empirical editing process. Gene counts as quantified from FISH images by the bigFISH package[112] are correlated for cells that were measured in multiple images.

Supplemental Table 1: Statistically significant correlations between phylogenetic and spatial distances observed for different pairwise cell type comparisons in the E7.5 tetraploid mouse embryo lineage. Significance was assessed using the Pearson correlation test.

Cell type comparison	Pearson Correlation	p-value	FDR adjusted p-value
Rostral neural plate | Surface ectoderm	0.55	1.2E-02	2.7E-02
Surface ectoderm | Surface ectoderm	0.46	1.3E-03	4.1E-03
Epiblast | Epiblast	0.35	2.8E-05	1.5E-04
Early nascent mesoderm | Rostral neural plate	0.34	1.7E-03	4.9E-03
Definitive ectoderm | Surface ectoderm	0.34	2.1E-09	3.1E-08
Caudal mesoderm | Early nascent mesoderm	0.31	1.6E-14	1.9E-12
PGC | Primitive streak	0.27	3.4E-03	9.1E-03
Definitive ectoderm | Epiblast	0.27	1.1E-09	2.2E-08
Early nascent mesoderm | Epiblast	0.26	1.5E-12	5.8E-11
Epiblast | Surface ectoderm	0.26	6.9E-04	2.5E-03
Definitive endoderm | Epiblast	0.22	1.4E-05	8.5E-05
Definitive ectoderm | Definitive ectoderm	0.22	3.6E-06	2.6E-05
Caudal mesoderm | Epiblast	0.21	1.1E-03	3.5E-03
Caudal epiblast | Surface ectoderm	0.21	3.9E-03	1.0E-02
Epiblast | Haematoendothelial progenitors	0.21	3.9E-04	1.5E-03
Early nascent mesoderm | Haematoendothelial progenitors	0.20	6.6E-08	7.1E-07
Caudal epiblast | Definitive ectoderm	0.19	2.7E-06	2.0E-05
Early nascent mesoderm | Early nascent mesoderm	0.19	1.5E-08	1.8E-07
Definitive endoderm | Surface ectoderm	0.19	5.9E-03	1.5E-02
Blood progenitors | Early nascent mesoderm	0.18	1.7E-10	5.1E-09
Anterior primitive streak | Epiblast	0.17	1.6E-06	1.3E-05
Definitive ectoderm | Definitive endoderm	0.17	1.2E-05	7.7E-05
Anterior primitive streak | Definitive ectoderm	0.17	1.3E-09	2.2E-08
Anterior primitive streak | Surface ectoderm	0.16	1.0E-03	3.5E-03
Blood progenitors | Epiblast	0.15	4.1E-04	1.6E-03
Anterior primitive streak | Caudal mesoderm	0.15	2.2E-04	9.1E-04
Caudal epiblast | Definitive endoderm	0.12	1.2E-02	2.7E-02
Anterior primitive streak | Anterior primitive streak	0.11	4.9E-04	1.8E-03
Blood progenitors | Haematoendothelial progenitors	0.11	1.3E-02	2.9E-02
Early nascent mesoderm | Late nascent mesoderm	0.11	6.1E-09	8.0E-08
Late nascent mesoderm | PGC	0.10	1.4E-03	4.3E-03
Late nascent mesoderm | Late nascent mesoderm	0.10	7.0E-07	6.3E-06
Caudal mesoderm | Late nascent mesoderm	0.10	2.1E-03	6.0E-03
Anterior primitive streak | Definitive endoderm	0.08	1.6E-02	3.2E-02
Extraembryonic Mesoderm | Late nascent mesoderm	0.08	8.4E-13	5.0E-11

Anterior primitive streak | Blood progenitors	0.07	1.2E-02	2.7E-02
Extraembryonic Mesoderm | PGC	0.06	9.3E-03	2.2E-02
Anterior primitive streak | Early nascent mesoderm	0.06	8.6E-03	2.1E-02
Extraembryonic Mesoderm | Extraembryonic Mesoderm	0.05	1.8E-05	1.0E-04
Blood progenitors | Extraembryonic Mesoderm	-0.05	5.4E-03	1.4E-02
Anterior primitive streak | Extraembryonic Mesoderm	-0.05	1.9E-04	8.1E-04
Caudal epiblast | Extraembryonic Mesoderm	-0.08	7.9E-05	3.9E-04
Late nascent mesoderm | Surface ectoderm	-0.09	1.4E-02	3.0E-02
Definitive endoderm | Extraembryonic Mesoderm	-0.10	6.9E-07	6.3E-06
Blood progenitors | Definitive ectoderm	-0.10	3.4E-03	9.1E-03
Definitive ectoderm | Late nascent mesoderm	-0.10	9.6E-06	6.3E-05
Definitive endoderm | Late nascent mesoderm	-0.10	9.7E-05	4.6E-04
Blood progenitors | Caudal epiblast	-0.10	1.4E-02	2.9E-02
Definitive ectoderm | Extraembryonic Mesoderm	-0.10	4.0E-10	9.6E-09
Definitive ectoderm | Haematoendothelial progenitors	-0.11	1.5E-02	3.2E-02
Extraembryonic Mesoderm | Haematoendothelial progenitors	-0.11	7.4E-07	6.3E-06
Extraembryonic Mesoderm | Surface ectoderm	-0.11	7.9E-05	3.9E-04
Blood progenitors | Definitive endoderm	-0.15	1.3E-04	5.7E-04
Haematoendothelial progenitors | PGC	-0.15	2.0E-02	4.0E-02
Blood progenitors | Primitive streak	-0.16	9.8E-03	2.3E-02
Definitive endoderm | Haematoendothelial progenitors	-0.17	1.0E-03	3.5E-03
Blood progenitors | Surface ectoderm	-0.18	1.3E-03	4.1E-03
Anterior primitive streak | Primitive streak	-0.20	1.1E-04	5.0E-04
Epiblast | Primitive streak	-0.26	2.3E-03	6.5E-03

Chapter 3

REGENERATIVE BASE EDITING ENABLES DEEP LINEAGE RECORDING

Introduction

Multicellular organisms are built from a single cell that repeatedly undergoes division and differentiation, producing structure and function. Taken together, the set of division events form a binary lineage tree, with phenotypic changes decorating the branches. In most organisms and contexts, the interplay between lineage and phenotype is incompletely understood. While individuals from a species are similar in form, there can be substantial variability in the underlying cellular lineage process[53]. At the same time, lineage constrains fate decisions through the progressive restriction of transcriptional plasticity[19,118]. Diseases manifest as a disruption of the lineage process, and can generate pathological cell division patterns[34]. Recovering lineage relationships can answer basic biological questions about multicellular development, the evolutionary conservation of these mechanisms across species, and reveal pathological effects of disease. Therefore, identifying the lineage relationships between cells has been a fundamental goal across both developmental biology and medicine.

In classic work, researchers were able to fully map the cellular lineage tree of the *C. elegans* nematode. This heroic effort unexpectedly revealed that each cell in the organism originates through a unique deterministic lineage that is identical across individuals[4]. However, this feat could not be easily replicated in most other organisms, as they present

greater challenges: first, they are optically opaque, preventing direct time lapse imaging. Second, they are often larger, comprising many thousands or even millions of cells in a tissue or context of interest. Third, the lineages that form tissues are often variable and seemingly stochastic, making it more difficult to integrate observations from multiple individuals.

To address these challenges, several groups have developed synthetic lineage recording systems, including MEMOIR[44,45], GESTALT[32–34], CARLIN[35], LINNAEUS[36], SMALT[37], homing CRISPR barcoding[39,40], and others[38,119,120]. These systems use different methods, including recombinases[45], CRISPR nucleases[32–36,39,40,44,119,120], and CRISPR base editors[37] to dynamically edit (mutate) specific, heritable target sequences as cells proliferate. Each cell lineage thereby accumulates a unique edit pattern. Edits can subsequently be read out in bulk or, ideally, in individual cells, either by sequencing[32–37,39,40,119,120] or microscopic imaging[44,45]. Finally, computational algorithms allow reconstruction of lineage relationships from measured endpoint edit patterns, in a manner loosely analogous to phylogenetic reconstruction. Most recently, barcode labeling by incorporating temporally ordered sequences using Prime Editing technology has been introduced as a new paradigm for phylogenetic recording.[38,41–43]

In this work, we sought to improve two key difficulties with current lineage recording methodologies that constrain the depth (number of cell generations) of recording. First, a limitation of most recording systems is that edit rates are generally proportional to the number of unedited sites that can be targeted. This causes the fraction of unedited sites to

decay exponentially over time, rather than at a constant rate[121]. Frontloading mutations in this way uses the majority of memory capacity early, leading to little information being stored later in the process. Second, the total amount of memory available in most recording systems is small (on the order of tens of editable target sites). This further limits the potential duration of recording and resolution of resulting trees.

To address these critical issues, we designed a new generative recording system termed the hypercascade. This system takes advantage of A-to-G base editors[103], which can make specific, single base pair mutations with high fidelity and allow dense packing of target sites. The defining feature of the hypercascade is that editing existing target sites generates new editable target sites in an approximately 1:1 ratio. Edits can therefore accumulate over time without diminishing the number of remaining editable target sites for an extended period. Simulations reveal that this generative property linearizes the rate of editing over time and improves the accuracy of lineage reconstruction. At the same time, this strategy enables dense packing of target sites, approaching a density of one target site every five base pairs of sequence, allowing vast amounts of memory to be integrated into single cells for recording.

Experimental analysis in mouse and human cells showed that the generative editing property works as designed. One-shot engineering of the system into human induced pluripotent stem cells enabled reconstruction of gross clonal features, although was not sufficient to confidently resolve detailed lineage relationships. Finally, we used simulations to assess the potential of hypercascades to systematically record chromatin

state transitions across time. Future work will aim to develop multiplexed single-cell methods to fully take advantage of the strengths of hypercascade recording.

The hypercascade design allows sustained recording through generative editing

In contrast to some editing mechanisms, CRISPR A-to-G base editors (ABEs) are distinguished by their ability to generate precise and predictable single base mutations at defined target sites (**Figure 1A**). Further, tandem arrays of base editor target sites allow compact genetic encoding of multiple bits of recorded information. Edits can accumulate over multiple cell generations and are stably inherited by daughter cells (**Figure 1B**). Consequently, readout of edits by sequencing can in principle be used to reconstruct cell lineage relationships (**Figure 1B**).

Even if it densely encodes multiple bits, a simple tandem array of base editable target sites faces a fundamental limit in its recording capacity: if each individual site has a constant probability of editing per unit time, as has been observed[121], then the number of available unedited sites decays exponentially over time. Early generations receive many edits while later generations receive few or none (**Figure 1C, D**).

One way to circumvent this problem is to generate new target sites as old ones are consumed by editing. As long as new sites can be generated, this scheme could keep the total number of unedited sites, and therefore the edit rate, approximately constant. For example, imagine a set of target sites organized into 4 logical "layers" (**Figure 1C**, right panel) such that only layer 1 target sites (upper row) are initially accessible to editing.

Further, assume that editing of layer 1 sites generates editable layer 2 sites (second row). Similarly, editing of layer 2 sites could generate layer 3 sites (third row), and layer 3 edits could generate layer 4 sites (fourth row). In such a scheme, editing would consume and generate sites at roughly the same rate, keeping the total edit rate constant over time, at least until nearly all potential edits have occurred (**Figure 1D**).

We designed a base editable target sequence, termed the hypercascade, that exploits unique properties of ABEs to implement such a scheme, and simultaneously provide high density memory encoding. The ABE has two requirements for efficient function. One is a 20 bp homology sequence (encoded by the gRNA it complexes). The other is a 3 bp protospacer adjacent motif (PAM), which is NGG for the most commonly used Cas9 homolog derived from *S. pyogenes*.[103] When these two requirements are met, an A in the fifth or sixth position of the target sequence is mutated to a G (**Figure 1A**).[103] The predictability of this edit outcome makes it possible for editing of available target sites to repair engineered mismatches that otherwise prevent editing by distinct gRNAs at other target sites (**Figure 1E**).

Taking advantage of this edit-repair principle, we designed a cascading system that packs four target sites, representing four logical "layers," into one tandemly repeatable 20bp sequence. These target sites can be edited using four corresponding gRNAs. However, only the layer 1 target site starts with a perfect match to its cognate gRNA and a functional PAM. Each of the others is initially prevented from editing by at least two mismatches in its protospacer sequence and PAM (**Figure 1F**).

The final feature of this design is intended to enforce an ordered sequence of edits (**Figure 1F**). Editing of a layer two site only becomes possible after two surrounding layer one edits. Similarly, editing of a layer 3 site requires previous editing of two surrounding layer 2 sites, and a previously edited layer 1 site. Finally, editing of each layer 4 site requires previous editing of two surrounding layer 3 sites, as well as previous editing of the layer 1 and 2 sites necessary for those layer 3 edits. We created an animated video to explain and visualize the operation of the system at youtu.be/GRVMbn-dElc.

Conceptually, this system comprises a 2-dimensional array of linked edit cascades, generalizing the concept; analogous to the way a hypersphere generalizes the concept of a sphere, we term this structure a hypercascade (**Figure 1E, F**). The hypercascade design allows dense packing of editable bases while requiring only four distinct gRNAs for operation. The specification for this scheme leaves 11 base pairs unconstrained in the 20bp repeating element, yielding a total of approximately 4 million possible designs.

To test whether the hypercascade could address the exponential memory decay problem, we simulated a stochastic edit process using the Gillespie algorithm(Gillespie 1977), and compared a hypercascade of 20 repeating units (74 target sites total in a 403bp sequence) to two simpler independent control editing schemes. The first contained 20 edit sites, one per 20bp unit, comprising a similar overall DNA length (403bp), which is convenient for standard Illumina sequencing readouts. The second control scheme contained 74 non-interacting target sites, all of which were initially available for editing, which could be

encoded in a longer stretch of 1.48kb of DNA. Thus, one control scheme maintained the same DNA length while the other contained the same number of target sites.

With either control scheme, the number of subunits available for editing decreased exponentially over time, as expected, such that few edits occurred at later time points (**Figure 1G**). By contrast, the simulated hypercascade showed an approximately linear accumulation of edits. For a given edit rate, the hypercascade extends the duration and linearity of recording relative to independent recorders with the same total DNA length, which saturate earlier regardless of how many total target sites they contain. Thus, in simulations, the hypercascade successfully extended editing.

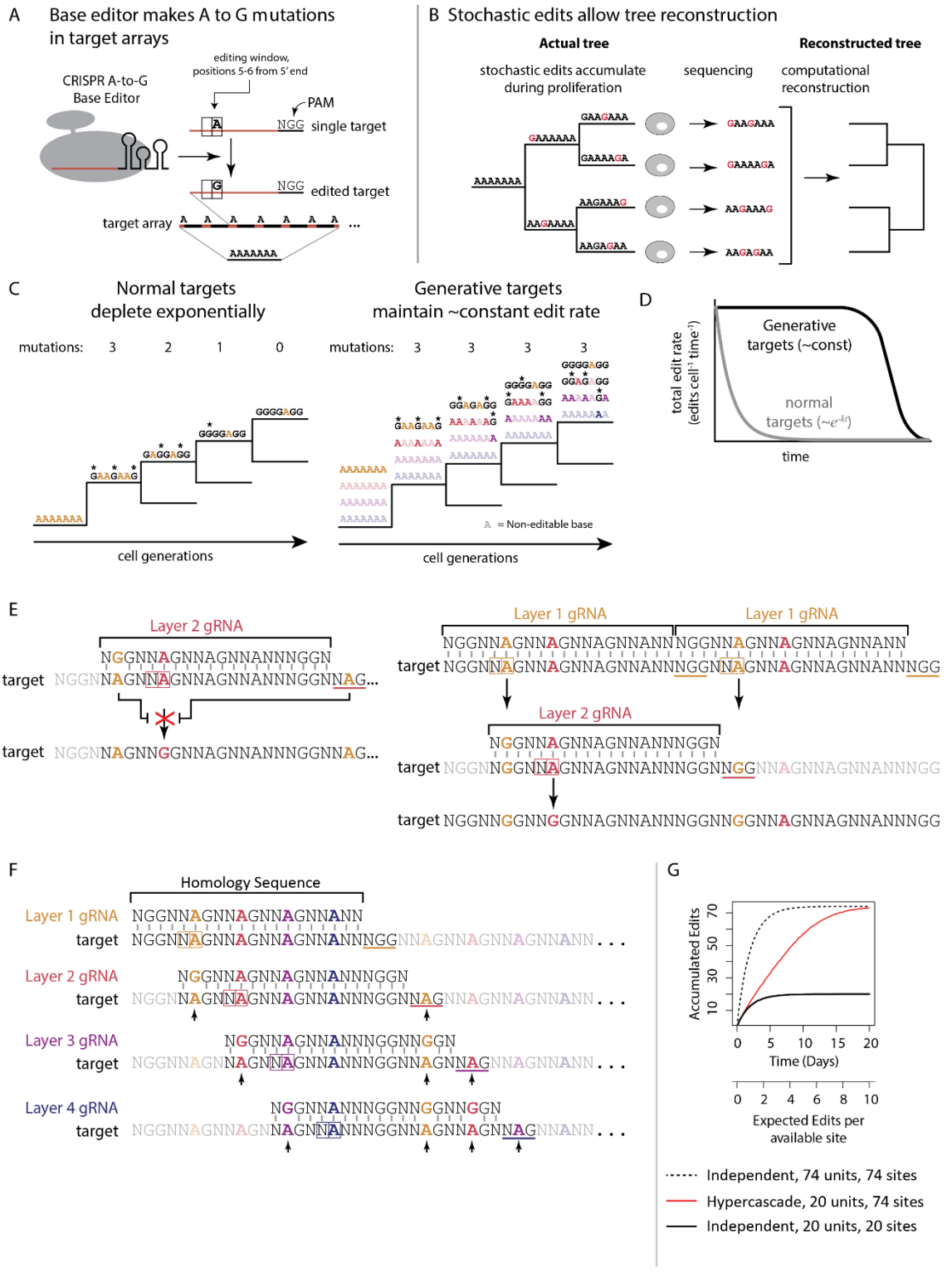

A
Base editor makes A to G mutations in target arrays
CRISPR A-to-G Base Editor
editing window, positions 5-6 from 5' end
PAM
NGG
single target
edited target
target array
B
Stochastic edits allow tree reconstruction
Actual tree
Reconstructed tree
stochastic edits accumulate during proliferation
sequencing
computational reconstruction
C
Normal targets deplete exponentially
Generative targets maintain ~constant edit rate
mutations: 3 2 1 0
mutations: 3 3 3 3
= Non-editable base
cell generations
cell generations
D
total edit rate (edits cell⁻¹ time⁻¹)
Generative targets (~const)
normal targets (~e^-kt)
time
E
Layer 2 gRNA
target
Layer 1 gRNA
Layer 1 gRNA
Layer 2 gRNA
F
Homology Sequence
Layer 1 gRNA
Layer 2 gRNA
Layer 3 gRNA
Layer 4 gRNA
target
G
Accumulated Edits
Time (Days)
Expected Edits per available site
Independent, 74 units, 74 sites
Hypercascade, 20 units, 74 sites
Independent, 20 units, 20 sites

Figure 1: The hypercascade system linearizes edit rate over time and densely packs mutable target sites. (A) The CRISPR A-to-G base editor makes predictable mutations at defined target sites. **(B)** Reading out heritable mutations enables reconstruction of cellular lineage relationships. **(C)** Arrays of independently edited targets are exponentially lost over time; in contrast, a system that generates new targets over the course of editing would maintain constant edit rate over an extended time scale. **(D)** Generative targets are predicted to maintain approximately constant edit rate until the generative mechanism fails (schematic). **(E)** New targets can be generated by repairing protospacer and PAM mismatches through A-to-G mutations. **(F)** This concept can extend to multiple unlocking layers of densely packed target sites. This sequence consists of a tandem repeating 20-mer which is acted on by four unique guide RNAs. Mismatches that are repaired to generate new targets are indicated with arrows. **(G)** Simulations of editing in this scheme reveal linear edit accumulation relative to arrays of independent targets.

Simulations show that the hypercascade enables more accurate lineage reconstruction than a simple array

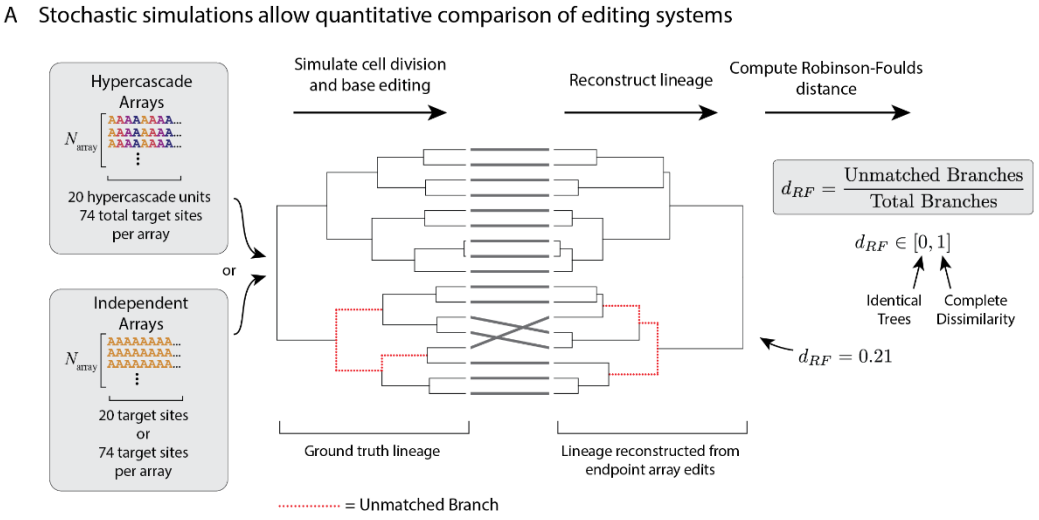
In this context, the hypercascade is intended to function as a lineage recorder. Therefore, it is critical to test not only whether edits occur linearly over time but also whether the typical patterns of edits produced can allow accurate lineage reconstruction. To address this question, we simulated a population of cells undergoing repeated rounds of division, while also stochastically editing either the 74 site hypercascade or one of the two independent control target arrays described above (**Figure 2A**). Stochastic editing was simulated in both systems across a range of edit rates, array copy numbers, and lineage tree depths. After simulating trees in the forward direction, endpoint barcode states were used to reconstruct cell lineage relationships using the Unweighted Pair Group Method with Arithmetic mean (UPGMA)[123]. Finally, reconstructed trees were compared to ground truth trees using the normalized Robinson-Foulds distance, where 0 represents a perfect topological match between the ground truth and reconstructed trees, and 1 represents two maximally different trees (**Figure 2A**).

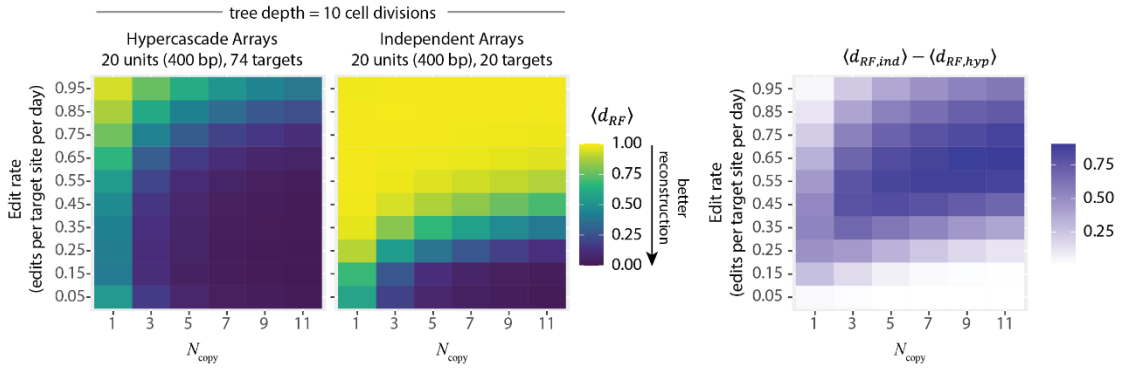
We first compared the hypercascade to an independent array of equal DNA length. We fixed the mean tree depth at 10 cell divisions and varied both edit rate and array copy

number (**Figure 2B**). At low edit rates, the hypercascade performed comparably to a 20-unit array. This is to be expected, as the first layer of the hypercascade is essentially equivalent to 20 independent units. At higher edit rates, two main benefits of the hypercascade became clear. First, high reconstruction accuracy was attainable for fewer integrations of the hypercascade compared to the 20-unit array (**Figure 2B**). Second, the hypercascade was relatively insensitive to edit rate. This feature is enabled by the generation of new target sites as old sites are depleted, linearizing the loss of targets over time and preventing the exponential loss of memory characteristic of independent units (**Figure 1D**). Similar results were obtained using a simple model of Cas9 editing that produces multiple terminal edit outcomes at each of the 20 target sites (**Supplemental Figure 1**). These results suggest that, for the same total DNA length and edit rate, the hypercascade is either similar to or outperforms the independent target arrays.

We next asked how the hypercascade would compare to an independent editable system with the same total number of target sites (**Figure 2C**). Because fully independent targets are uncorrelated, they have the potential to contain more information (higher entropy) than hypercascade targets, where some barcode states are never realized due to the ordered nature of editing. We fixed the target copy number to 3 and varied both edit rate and mean tree depth in both simulated systems. At very low edit rates, the independent array modestly outperformed the hypercascade. However, in regimes of higher edit rate or tree depth, where independent arrays saturate, the hypercascade dominated due its extended editing dynamics (**Figure 2C**). We expect that the hypercascade would also

outperform larger independent arrays for deeper trees, even when edit rates are low, due to earlier saturation in the independent case.





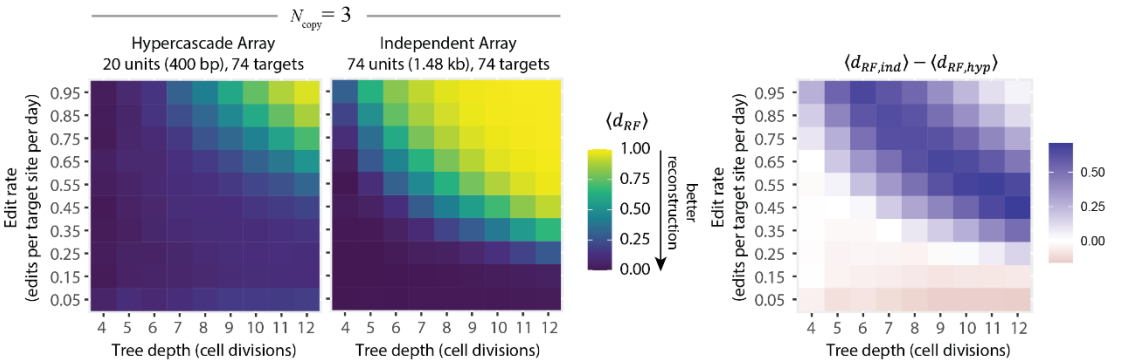


Figure 2: Hypercascades allow lineage recording and reconstruction across a broader range of target copy numbers and edit rates compared to arrays of independent sites. (A) Stochastic simulations of target editing and cell division together enable comparison of lineage reconstruction accuracy between independent and hypercascading systems. **(B)** Holding tree depth constant, the hypercascade system outperforms an independent array of targets with comparable sequence length over a variety of edit rates and copy numbers. **(C)** The situation is more nuanced comparing the hypercascade to independent arrays with the same total number of targets.

Different hypercascade sequences operate orthogonally with distinct kinetics

To experimentally assess the functionality of the hypercascade in cells, we constructed three different hypercascade sequence arrays, all based on the same design principles, but differing at unconstrained sequence positions (**Figure 3A**). Each sequence contained 19 tandem repeats of its 20bp unit, plus an additional partial unit containing a PAM site, for a total length of ~400bp, which is short enough to fit on a single sequencing read. To choose bases at the unconstrained positions, we used a machine learning model of Cas9 edit rates to select bases predicted to allow high edit rates in all four layers[67] (**Figure 3A**, **Methods**).

We then integrated each of the barcode arrays together with ABE in a mouse embryonic kidney fibroblast cell line[124] via piggyBac transposition. We selected stable polyclonal lines, each containing ABE and one or more copies of the integrated hypercascade. We then initiated editing by transfecting piggyBac transposons containing gRNA expression constructs.

To test the orthogonality of the gRNA-target site pairs, we examined all pairwise combinations of gRNAs and target designs. We allowed editing to continue for up to 44 days after gRNA transfection, periodically harvesting cells, and sequencing their hypercascades. In these experiments, editing occurred only for the expected gRNA-

hypercascade match, as expected (**Figure 3B**). Further, the fraction of edited sites increased linearly over time for all three target sequence designs (**Figure 3C**).

Previous studies with wild type Cas9 have shown that the nuclease still edits targets with a small number of mismatches, although at a reduced rate, especially at targets with functional PAM sites.[38,67,125] This mismatch editing efficiency has been found to depend on both the sequence of the protospacer and the distance of the mismatches from the PAM site, with higher tolerance of mismatches at sites further from the PAM.[38,67,125] We therefore sought to detect potential out-of-order editing, and quantify the rate at which it occurs. Edit rates were highest for gRNAs with perfect homology with their target. Consistent with previous findings, the editing rate typically decreased with more mismatches and was especially suppressed when the PAM site was broken (**Supplemental Figure 4**).

Finally, we asked whether edits accumulated sequentially through the designed mechanism (**Figure 1E**). We compiled reads from all time points, grouped them by the total edits per read and by their layer identity, and plotted the distribution of edits for each layer (**Figure 3D**). Each site was parameterized by a layer-specific edit rate, with multiplicative edit rate penalties for each possible mismatch, disfavoring premature edits. These plots reveal an ordered structure, with edits generally occurring in the expected sequence. Further, the patterns of accumulated edits could be well fit by a stochastic model of editing incorporating all edit rate parameters, including the low rates of editing at mismatched target sites (**Figure 3D**, solid lines and **Figure 3E**; **Supplemental Figure**

4). Taken together, these results indicate that the hypercascade successfully generates new targets during editing, as designed, leading to the intended linear and sequential accumulation of edits.

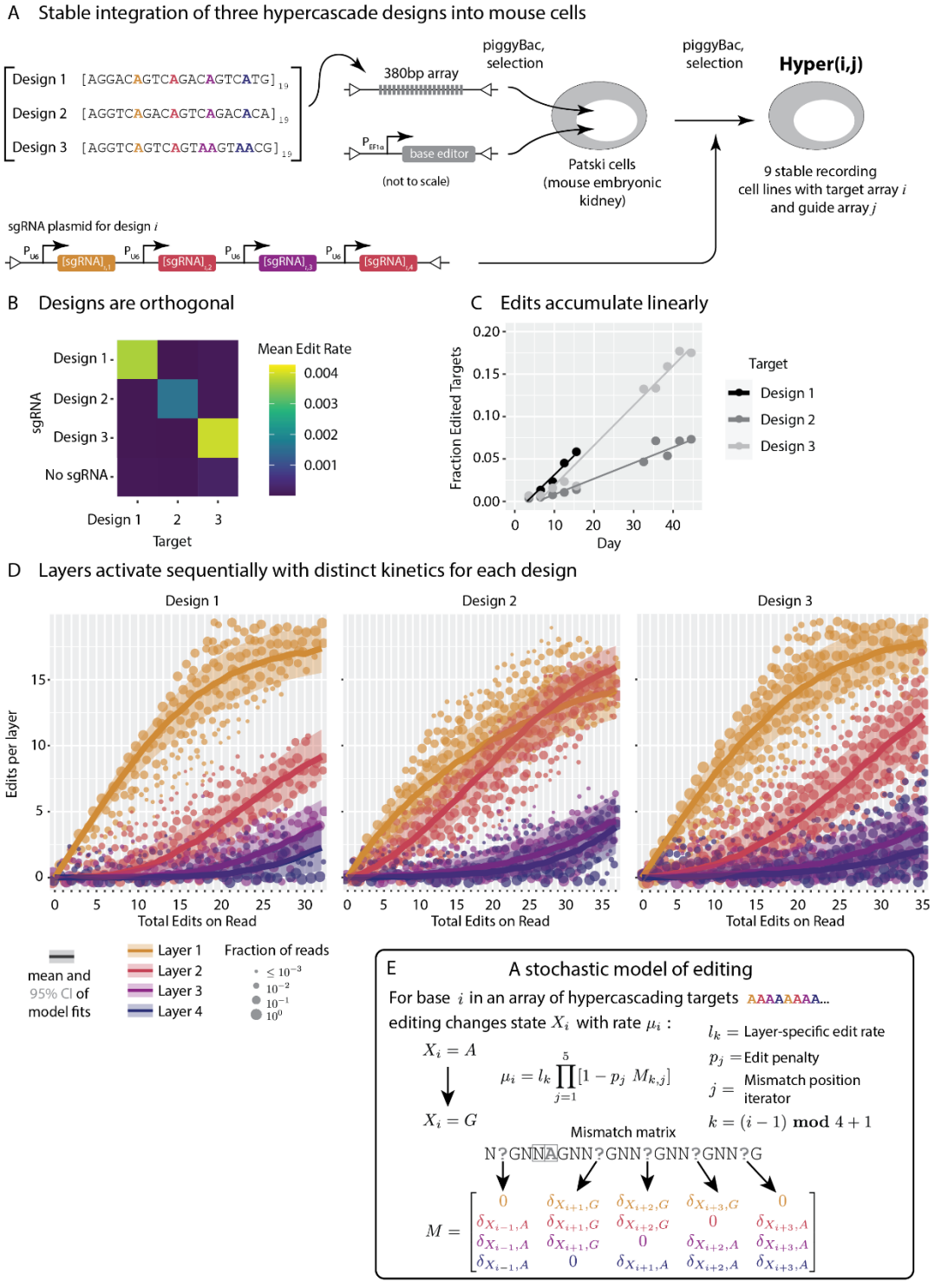

A Stable integration of three hypercascade designs into mouse cells
Design 1 [AGGACAGTCAGACAGTCATG]19
Design 2 [AGGTCAGACAGTCAGACACA]19
Design 3 [AGGTCAGTCAGTAAGTAACG]19
380bp array
piggyBac, selection
base editor
(not to scale)
Patski cells (mouse embryonic kidney)
piggyBac, selection
Hyper(i,j)
9 stable recording cell lines with target array i and guide array j
sgRNA plasmid for design i
B Designs are orthogonal
Mean Edit Rate
0.004
0.003
0.002
0.001
sgRNA
Design 1
Design 2
Design 3
No sgRNA
Target
C Edits accumulate linearly
Fraction Edited Targets
Day
Target
Design 1
Design 2
Design 3
D Layers activate sequentially with distinct kinetics for each design
Edits per layer
Total Edits on Read
mean and 95% CI of model fits
Layer 1
Layer 2
Layer 3
Layer 4
Fraction of reads
E A stochastic model of editing
For base i in an array of hypercascading targets AAAAAAAA...
editing changes state X_i with rate μ_i :
$X_i = A$
$X_i = G$
$\mu_i = l_k \prod_{j=1}^{5} [1 - p_j M_{k,j}]$
l_k = Layer-specific edit rate
p_j = Edit penalty
j = Mismatch position iterator
$k = (i-1) \bmod 4 + 1$
Mismatch matrix
N?GNNAGNN?GNN?GNN?GNN?G
$M = \begin{bmatrix} 0 & \delta_{X_{i+1},G} & \delta_{X_{i+2},G} & \delta_{X_{i+3},G} & 0 \\ \delta_{X_{i-1},A} & \delta_{X_{i+1},G} & \delta_{X_{i+2},G} & 0 & \delta_{X_{i+3},A} \\ \delta_{X_{i-1},A} & \delta_{X_{i+1},G} & 0 & \delta_{X_{i+2},A} & \delta_{X_{i+3},A} \\ \delta_{X_{i-1},A} & 0 & \delta_{X_{i+1},A} & \delta_{X_{i+2},A} & \delta_{X_{i+3},A} \end{bmatrix}$

Figure 3: The hypercascade exhibits sequential editing in living cells. (A) Three different implementations of the hypercascade were stably integrated into a mouse embryonic kidney line with either on- or off-target gRNAs. **(B)** Editing occurs only in the presence of on-target gRNAs. **(C)** Edits accumulate linearly over a 44-day period. **(D)** Layers of targets sequentially activate and are well fit by a stochastic model of editing **(E)**.

One-shot transfection of the hypercascade editing system reveals broad clonal features

A practical challenge for any recording system is integrating it into cells. In most systems, integration requires laboriously engineering cells to express multiple components, screening monoclones, and potentially repeating the whole process until cell lines with the desired characteristics are generated. We decided to explore whether, with the hypercascade, one could simultaneously integrate all components in a single step through piggyBac transposition. Resulting clones could then be selected in a single antibiotic selection step prior to downstream experiments being immediately performed on the resulting polyclonal cell line. This approach could enable recording in systems that are traditionally difficult to engineer or prone to silencing, including primary cells and human induced pluripotent stem cells (hiPSCs).

We investigated this approach using hypercascade target sequence 1 by simultaneously transfecting all components into hiPSCs, selecting with antibiotic for 1 day, then continuing culture of the cells and collecting samples every passage for genomic DNA extraction once they recovered (**Figure 4A**). Edits accumulated over time after an initial delay period (**Figure 4B**). This delay may reflect excess episomal plasmid arrays persisting from the initial transfection, slowly diluting out over time, and competing with genomically integrated target sites for editing and amplification. After the initial delay, edits accumulated through day 23 post transfection (**Figure 4B**). As in the previous

experiment (**Figure 3D**), the distribution of edits into different layers was consistent with sequential activation of layers (**Figure 4C**). These results indicate that generative hypercascade editing can be achieved with a single genome engineering step.

Interestingly, the frequency of reads with different total numbers of edits naturally grouped into several distinct distributions (**Figure 4D**). These likely represent individual clones or groups of clones that received specific copy numbers and integration sites for the editing component plasmids, leading to different edit rates. Distributions shifted to the right over time, as would be expected in the context of genomic edit accumulation.

We identified unique edit patterns within the data and reconstructed lineage relationships, leaving out targets with few edits that contain relatively little recorded information (**Figure 4E**). We find that medium- and high-edit rate cells cluster together into clades, a feature that is not retained in a scrambled barcode control (**Figure 4F**). The mean depth of the reconstructed tree is similar for reads generated from both medium and high edit rates, and is only slightly smaller than the number of generations expected for hiPSCs over 23 days of recording, where we would anticipate roughly 23 cell divisions (**Figure 4G**). Based on our simulations, however, we do not expect that individual clade bifurcations are resolved with perfect accuracy given only a single hypercascade copy (**Figure 2B**). Together, these results indicate that one-shot transfection has the potential to identify broad clonal features within a population of cells but is likely insufficient to capture detailed lineage relationships at the single cell level with the current design.

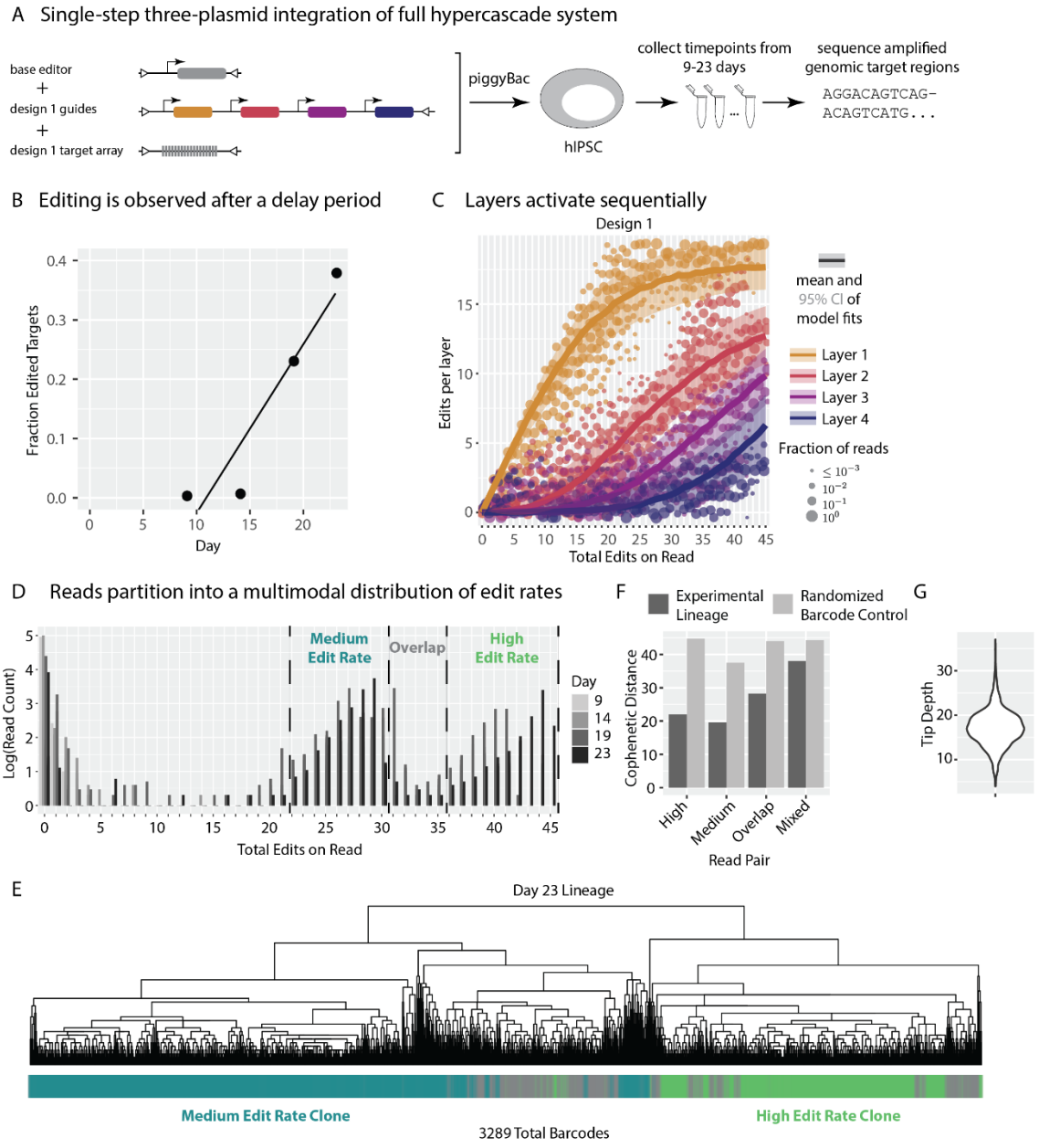


Figure 4: The hypercascade can be integrated in a single step and used for subsequent lineage recording. (A) All components of the system were simultaneously integrated into hiPSCs. **(B)** Edits accrued on integrated targets over a 23-day time course. **(C)** Layered targets sequentially activated. **(D)** Read counts binned by total edits present a multimodal distribution with peaks that shift right over time. **(E)** Hierarchical clustering reconstructed lineage relationships between 3,289 unique barcode sequences observed with at least 22 accumulated edits. **(F)** The mean cophenetic distance between pairs of cells from distinct edit rate groups is lower than in a randomized barcode control, identifying edit rate as a clonal feature. **(G)** The average tip depth is estimated to be approximately 17 generations.

Hypercascades have the potential to record chromatin transition dynamics

In addition to recovering lineage relationships, the recording field has long sought to record other biologically meaningful signals[41,44]. By making the rate of editing dependent on a signal of interest, one can, in appropriate parameter regimes, simultaneously reconstruct lineage and signal intensity over time based on the endpoint edit pattern (**Figure 5A**). This works because once edit patterns have been used to reconstruct a lineage tree, it is possible to assign individual edits to the point on the tree at which they most likely arose. The density of edits on each branch of the tree is then, ideally, proportional to the signal intensity, as represented by edit rate, during the corresponding time interval or cell cycle. One of the most biologically interesting signals to record is the accessibility of different chromatin loci, which can change dynamically during differentiation and other processes.

Previous work demonstrated that Cas9[126–128] and prime[129,130] editing are dependent on chromatin context. We reasoned that this effect could extend to the ABE, which is itself a Cas9 fusion protein. To test this hypothesis, we employed a female mouse embryonic kidney cell line derived from two mouse species, *M. musculus* and *M. spretus* (the Patski line, **Figure 5B**)[124,131,132]. In this line, one X chromosome comes from each species, which can be distinguished by the presence of single nucleotide polymorphisms (SNPs) present roughly every 80 bp along the sequences (**Figure 5B**). These cells have already undergone X inactivation, with the silent X in this line associated exclusively with the *M. musculus* allele, while the active X derives from *M. spretus*.

Patski cells present an ideal system for assaying the effects of chromatin silencing on base editing, since identical sequences present on the active and inactive X chromosomes can be simultaneously targeted by a single ABE-gRNA pair. Further, a recent study of the Patski line characterized differences in chromatin accessibility along the active and silent X chromosomes by ATAC-seq[131]. We reasoned that chromatin accessibility might be most tightly controlled near the transcription start site (TSS) of genes, specifically in regions where ATAC peaks were present on the active X but not the silent X.

To investigate differential base editing at these sites, we engineered a stable, polyclonal population of Patski cells to constitutively express the ABE. We then transiently transfected sgRNAs targeting endogenous loci into the cells and performed amplicon sequencing at multiple time points (**Figure 5C**). We grouped the reads based on their chromatin context, then determined the edited fraction at each target site. If chromatin context impacts the edit rate, we expect to see a disproportionate amount of editing on reads associated with open chromatin.

We designed gRNAs targeting 12 sites that contained differential ATAC peaks in a 2 kb window upstream of the TSS in 9 separate genomic loci (**Figure 5C**, **Supplemental Figure 6A**). Almost all target sites exhibited differential editing between X alleles, in some cases with an over 40-fold ratio of active to silent editing. Similar results were obtained targeting 10 sites within a single locus (**Supplemental Figure 6B**).

We hypothesized that the high density of target sites packed into a single genomic locus by the hypercascade could enable recording of chromatin state transitions if the

chromatin context of hypercascades shifts between open and closed during lineage recording. We used stochastic simulations to investigate this possibility, considering a model in which a cell contains 20 distinguishable copies of the hypercascade array integrated at different genomic locations. We aim to model differentiation of a multipotent cell into terminal fates, where some genomic loci transition from open to closed or closed to open chromatin. In our simulations, 15 of the integrated barcodes are inserted into statically open chromatin, which never changes (**Figure 5E**). The remaining five barcodes are inserted into dynamic chromatin, which will change state along some lineage trajectories. We modeled the transition from open to closed chromatin by a four-fold change in editing rate, as observed in the Patski experiments.

To investigate what types of trajectories through epigenetic space could be recovered, we contrived a differentiation process marked by progressive state commitment over time (**Figure 5F**). At every cell division event, cells were allowed to transition between differentiation trajectories at a defined rate (**Figure 5F**, right). The simulation was divided into 5 epochs. Transitioning between each epoch led to progressively more restricted mixing between trajectories, until each differentiation trajectory became canalized into its own path. Two of the trajectories (**Figure 5F**, blue and pink) shared a common progenitor across the first three epochs before diverging to become distinguishable terminal fates. In contrast, the other two trajectories (**Figure 5F**, orange and green) had distinct trajectories from the outset, but converged on the same final fate, modeling convergent differentiation known to occur in some natural systems.

To reconstruct chromatin histories based on the simulated hypercascade edit patterns in these cells, we first reconstructed lineage histories across 200 simulated trees using the UPGMA method. We then estimated the edit rate along every branch of the tree based only on the terminal barcode sequences and inferred lineage using a custom R script. This yields an estimate of edit rate across time for each cell (**Figure 5G**).

Even with the abundant memory of the hypercascade, the estimate of chromatin trajectory for any individual cell is inherently noisy. However, it is possible to get a more accurate picture of chromatin dynamics by averaging over many cells that experience the same trajectory. An initial challenge is to identify which cells in a pool belong to a given trajectory. We accomplished this by fitting fourth degree polynomial curves to the dynamic curve for each integration in each simulated cell, then clustering the cells based on their fitting parameters. This method accurately groups simulated cells by trajectory for sufficiently high edit rates (**Figure 5H**).

Finally, we fit smoothed curves across the cells belonging to each cluster to identify the edit rate across time for each integration locus, mirroring the chromatin accessibility at that locus (**Figure 5I**). Promisingly, all four trajectories yielded curves that reflect the underlying differentiation process we defined. Although we simulated sharp transitions, these are blurred out across time in the resulting estimates (**Figure 5I**). More sophisticated modeling, and potentially incorporation into a Bayesian framework, could enable even better reconstruction of chromatin dynamics. Together, these results suggest

that hypercascades have the potential to be used not only to reconstruct lineage but also to recover chromatin dynamics, a possibility that could be explored in future work.

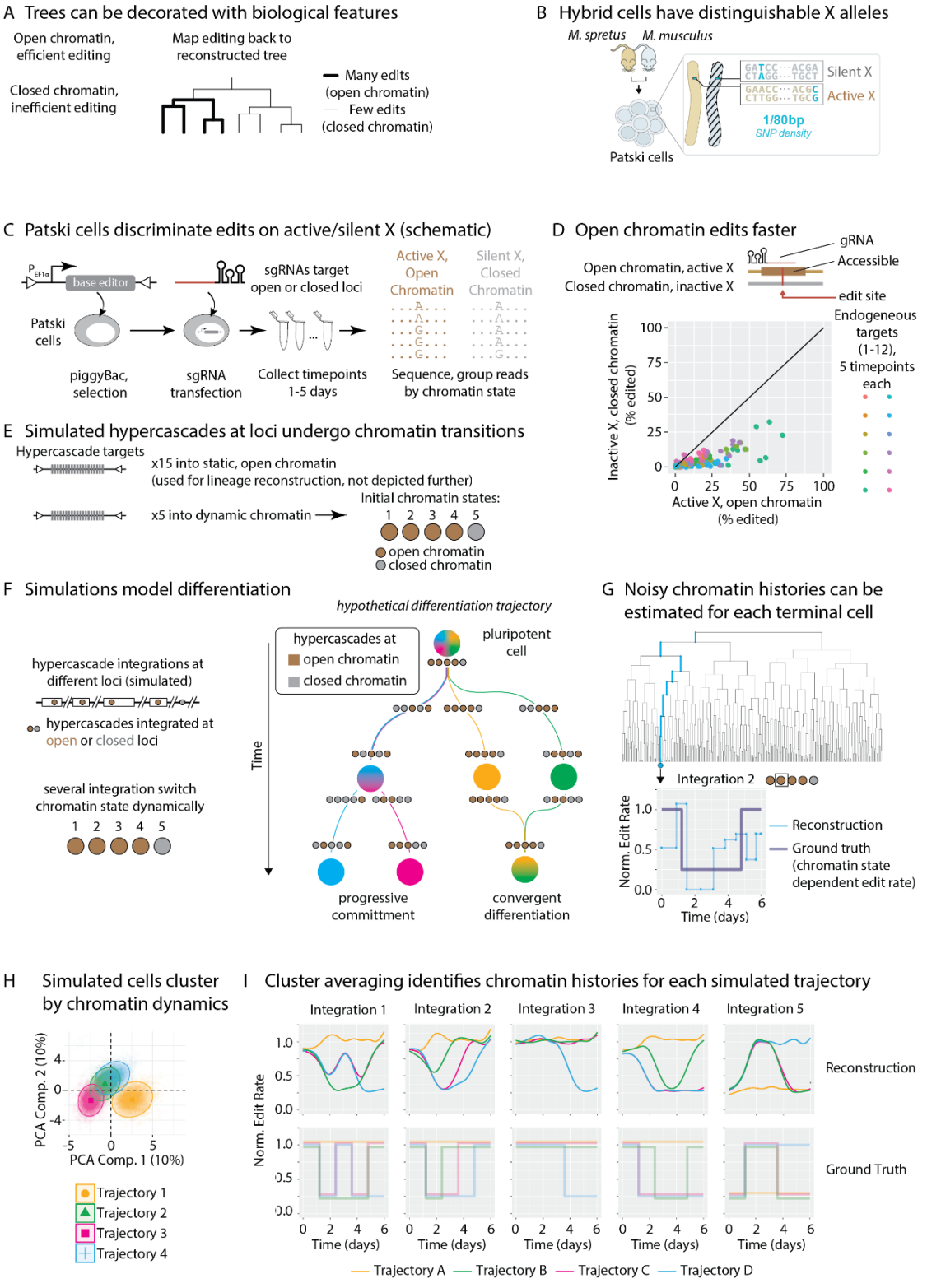

A Trees can be decorated with biological features
Open chromatin, efficient editing
Closed chromatin, inefficient editing
Map editing back to reconstructed tree
Many edits (open chromatin)
Few edits (closed chromatin)
B Hybrid cells have distinguishable X alleles
M. spretus
M. musculus
Silent X
Active X
1/80bp
SNP density
Patski cells
C Patski cells discriminate edits on active/silent X (schematic)
base editor
Patski cells
piggyBac, selection
sgRNAs target open or closed loci
sgRNA transfection
Collect timepoints 1-5 days
Active X, Open Chromatin
Silent X, Closed Chromatin
Sequence, group reads by chromatin state
D Open chromatin edits faster
Open chromatin, active X
Closed chromatin, inactive X
gRNA
Accessible
edit site
Endogeneous targets (1-12), 5 timepoints each
Inactive X, closed chromatin (% edited)
Active X, open chromatin (% edited)
E Simulated hypercascades at loci undergo chromatin transitions
Hypercascade targets
x15 into static, open chromatin
(used for lineage reconstruction, not depicted further)
x5 into dynamic chromatin
Initial chromatin states:
1 2 3 4 5
open chromatin
closed chromatin
F Simulations model differentiation
hypothetical differentiation trajectory
hypercascade integrations at different loci (simulated)
hypercascades integrated at open or closed loci
several integration switch chromatin state dynamically
1 2 3 4 5
hypercascades at
open chromatin
closed chromatin
pluripotent cell
Time
progressive committment
convergent differentiation
G Noisy chromatin histories can be estimated for each terminal cell
Integration 2
Norm. Edit Rate
Reconstruction
Ground truth (chromatin state dependent edit rate)
Time (days)
H Simulated cells cluster by chromatin dynamics
PCA Comp. 2 (10%)
PCA Comp. 1 (10%)
Trajectory 1
Trajectory 2
Trajectory 3
Trajectory 4
I Cluster averaging identifies chromatin histories for each simulated trajectory
Integration 1
Integration 2
Integration 3
Integration 4
Integration 5
Norm. Edit Rate
Reconstruction
Ground Truth
Time (days)
Trajectory A
Trajectory B
Trajectory C
Trajectory D

Figure 5: ABE editing is modulated by chromatin state, which has the potential to allow dynamic chromatin recording. (A) Linking biological signals, such as chromatin state, to edit rate can enable dynamic inference of the signal in single cells. **(B)** Patski cells, a female hybrid mouse cell line, contain a diploid genome with one allele representing each parental species, rendering alleles distinguishable through single nucleotide polymorphisms. **(C)** One X allele is constitutively silenced, allowing internally controlled investigation of editing at silent and active DNA sequences. **(D)** Across multiple target sites, we estimate a 4-fold increase in editing at open chromatin compared to closed. **(E)** Hypercascades integrated at high copy number may be able to recover both lineage and dynamic chromatin context information at their insertion locus. **(F)** We simulated a hypothetical heterogeneous differentiation, where cells transition through different states (marked by differences in chromatin state across 5 hypercascade-containing loci). **(G)** Lineage reconstruction enables noisy inference of dynamic edit rate modulation (a proxy for chromatin state) over time for each cell. **(H)** Single cells cluster into discrete groups based on their inferred chromatin transition history. **(I)** Averaging across all cells in a cluster gives qualitatively accurate estimates of the dynamic chromatin history experienced by each group.

Discussion and conclusions

Recording lineage relationships over long timescales and in many cells is necessary to understand organismal development but is in general a difficult problem. Most CRISPR recorders to date are limited by a small number of mutable sites and an editing mechanism that makes exponentially greater mutations at early points in time. We developed the hypercascade to address both challenges by densely packing editable targets in short synthetic DNA sequences in such a way as to create on average nearly 1 new target site for every site used, linearizing the rate of edit accumulation over time (**Figure 1**). Simulations reveal that this strategy can dramatically improve the ability to reconstruct lineage relationships in an edit-rate dependent manner (**Figure 2**). Many sequences exist that can linearize editing by our design; we show that three examples all produce the designed behavior (**Figure 3**), although with occasional out-of-order edits (**Supplemental Figure 1-4**). In addition to traditional cell engineering strategies, the hypercascade system can be added to cells in a single step. We demonstrated this by integrating the components into hiPSCs and recording over a 23-day period (**Figure 4**). Using the genomic records, we identified edit rate as a clonal feature and generated a partially resolved tree containing thousands of cells from only a single barcode. Going

forward, simulations indicate that genomically multiplexing hypercascade barcodes will enable near-perfect reconstruction of developmental processes (**Figure 2**).

The information rich datasets from polyclonal editing experiments (**Figures 3** and **4**) allow us to estimate the rates at which out-of-order editing occurs, where the ABE edits targets that have mismatches compared to the gRNA sequence (**Supplemental Figure 1**). As anticipated, we find that edit rates are highest when gRNAs have perfect homology with their target. Consistent with previous findings, editing rate typically decreases with more mismatches and is especially suppressed when the PAM site is broken (**Supplemental Figure 4**). We used a simplified model of mismatch editing to generate stochastic editing models for each specific hypercascade sequence (**Figure 3D, E** and **4C**, solid lines).

Hypercascade recording has some limitations in its current form. At long time scales it is not clear whether barcodes continue to edit effectively. For one of the three targets investigated, full length targets could no longer be recovered by PCR from samples of the population at days 32 and later. Instead, a much shorter product is formed by PCR (**Supplemental Figure 5**). Notably, this target was observed to edit more quickly than the other two targets during the first 15 days, at a rate approximately 5-fold faster. Control samples with the same targets but with mismatched gRNAs were still able to be amplified and sequenced; full length targets were observed in the resulting reads (**Figure 3A** and **Supplemental Figure 5**). The other two barcode samples were also able to be recovered and sequenced by PCR amplification (**Supplemental Figure 5**). This could

indicate a low rate of barcode collapse due to the repetitive nature of the sequences, which may set an upper bound on the length of time for system recording.

Future work will bring the hypercascade to the single-cell level, facilitating joint measurement of the transcriptome alongside lineage barcodes, and address issues of barcode collapse. We speculate that different base editors, in particular those that rely on catalytically dead Cas9 rather than the Cas9 nickase, could reduce collapse incidence by creating less local damage to DNA and prompting different repair pathways. In principle, edits read out at the single cell level can further be used to reconstruct the dynamics of a given signaling pathway or phenotypic feature, so long as that feature is linked to edit rate.[41,44] We find that ABE edit rate can be modulated not only transcriptionally, but also epigenetically based on the chromatin context of the target site (**Figure 5**). Simulations show that hypercascades integrated at high copy number have the potential to capture chromatin state transitions at each insertion locus (**Figure 5**), potentially yielding a temporal recording of chromatin state that can be recovered with an endpoint measure. We envision many applications that can benefit from hypercascade technology as we continue to build understanding in how cells coordinate to form and maintain complex structures.

Methods

Cell culture

Patski cells were a generous gift from Christine Disteche and Mitch Guttman. For routine culture, Patski cells were maintained in DMEM with 4.5 g/L glucose, L-glutamine, and sodium pyruvate supplemented with 10% Tet-approved FBS (Clontech) and 500 ug/mL Pen-Strep. When cells reached roughly 80% confluence they were split at a ratio of 1:6 by dissociation with StemPro Accutase (ThermoFisher Scientific). Cells were frozen as necessary in culture medium with 10% DMSO.

Human iPSCs (line WTC-11 / GM25256, NIGMS Human Genetic Cell Repository) were cultured as colonies on embryonic stem cell qualified Matrigel (Corning) with mTeSR1 Complete media (Stem Cell Technologies), changing media daily during culture. Plates were coated according to manufacturer instructions. Cells were passaged every 4-5 days as small clumps using ReLeSR gentle passaging reagent (Stem Cell Technologies).

Cloning

PiggyBac vectors for ABE integration were developed by Gibson cloning a linear PCR fragment containing ABE7.10 (generated from pCMV-ABE7.10, a gift from David Liu; Addgene plasmid # 102919) into a linearized piggyBac[104] insertion vector backbone containing the human Ef1a promoter and puromycin resistance (final vector PB-EF1a-ABE-puro).

gRNA expression plasmids were cloned using the GoldenGate[133] method to simultaneously integrate multiple U6-gRNA fragments in tandem into a piggyBac backbone containing inverted terminal repeats (ITRs) and a blasticidin resistance gene (final vectors H01-4xgRNA-blast, H05-4xgRNA-blast, H10-4xgRNA-blast).

Hypercascade target plasmids, containing repetitive target sites, primer sites, and piggyBac ITRs were synthesized by Genewiz and used directly (final vectors H01, H05, and H10). We elected to use rearranged ITR sequences described previously[134] to minimize the amount of genomically integrated DNA. In final hypercascade barcode designs, the free 11bp (**Figure 1E,F**) were chosen based on a machine learning algorithm (Azimuth 2.0[67,135]) that predicts the editing activity of Cas9 target sites. Each target contained 20 repeats of the 20-mer repeat sequence for a total of 70 editable target sites per barcode (as no PAM site was included for the 20th repeat). Azimuth 2.0 on-target targeting score predictions were chosen to meet the following criteria: layer 1 > 0.7; layer 2 > 0.65; layer 3 > 0.65; layer 4 > 0.6. Final target sequences used in experiments, as well as several further sequences that passed the design specifications, are listed in **Supplemental Table 1**. All plasmids were verified by Sanger sequencing prior to use. Sequence maps are available as supplemental data.

Generation of stably integrated lines

The Super PiggyBac Transposase system (System Biosciences) was used to generate Patski cells expressing ABE. Patski cells were transfected with PB-EF1a-ABE-puro along with a vector encoding Super piggyBac Transposase[104] at a mass ratio of 4:1 using

FuGENE HG transfection reagent following the manufacturer's protocol in a six-well plate (Promega). A negative control was performed by transfecting a plasmid encoding GFP without a selection marker. Polyclonal lines were selected with antibiotic (1 ug/mL puromycin) for a period of 14 days, until all cells had died in the negative controls and cells recovered under selection. Lines were expanded in standard media and frozen down for use as needed. Hypercascade targets were integrated into this cell line by transfecting a 10:1 mass ratio mixture of hypercascade target to PB-PGK-ABE-neo vector along with a vector encoding piggyBac transposase using FuGene as described above, then selected with geneticin (500 ug/mL). To start hypercascade editing in Patski cells, a third piggyBac transfection of a vector containing the four gRNAs for hypercascade editing and blast resistance was transfected into the polyclonal hypercascade target lines along with piggyBac transposase using the FuGene reagent as above. All combinations of targets and gRNAs were tested individually. These lines were selected with blasticidin (10 ug/mL) for the first 15 days in culture.

For hiPSCs, cells were single suspended 24 hours prior to transfection, counted, and plated in 24 well plates at a concentration of 26.3k cells per cm2. Rock inhibitor Y-27632 (10 uM) was included in the culture media from this point until the end of the antibiotic selection to facilitate cell survival (StemCell Technologies). hiPSCs were transfected with mixtures of 100 ng transposase, 167 ng EF1a-ABE-puro, 167 ng H01, and 167 ng H01-4xgRNA-blast vectors. LipofectSTEM lipid transfection reagent was included in transfection mixtures to add a final volume of 2 uL to each transfected well of a 24 well plate as recommended by the manufacturer (ThermoFisher). For editing negative control

samples, the H01-4xgRNA-blast plasmid was exchanged for 167 ng of ePB-CAG-H2B-Cerulean. As a negative transfection control, parallel samples were transfected with no vector. Transfected cells were allowed to recover for 48 hrs after transfection, then were selected with puromycin (1 ug/mL) for 24 hrs.

In each case, samples were expanded and split with each passage to save cells for sequencing. Collected cells were pelleted at 800xg for 5 min, the supernatant was aspirated, then frozen at -80 C to await genomic DNA extraction.

Endogenous editing in Patski cells

To initiate editing of endogenous targets in Patski cells, gRNA was transfected directly into polyclonal cells selected to express PB-EF1a-ABE-puro as described above. Custom crRNA-XT reagents were purchased and used following manufacturer instructions (Integrated DNA Technologies, homology sequences listed in **Supplemental Table 3**). gRNA complexes were transfected by lipofection using the RNAiMAX reagent (ThermoFisher Scientific) at either 15 or 30 nM final gRNA concentration, using a 2:1 ratio of transfection reagent to gRNA in either 6 or 24 well plates. Cells were passaged the day after transfection. Subsequently, samples were split with each passage to save cells for sequencing. Cells for sequencing were centrifuged at 800 RPM for 5 minutes, the supernatant was aspirated, and pellets were frozen at -80 C to await genomic DNA extraction. This sample collection and passaging process was repeated to collect cells at multiple timepoints.

Targets were selected based on a publicly available dataset containing genome-wide ATAC peaks for the Patski cell line focused only on the X chromosomes[131]. The goal was to identify regions that showed differential accessibility between the two X alleles. Since we are interested in target sites that can be informative of cellular state, we decided to focus on promoters. We retrieved 2,000 bp length regions upstream of all promoters on the X chromosome and asked whether they contain differential ATAC peaks between the two chromosomes. To classify peaks, we used the ATAC differential score, which ranges from -0.5 (no differential accessibility) to 0.5 (maximum differential accessibility) as defined by Bonora et al[131]. We counted the number of peaks within each promoter region, summed their score and finally selected promoter regions with the highest score. For the X chromosome, the dataset contained in total 1605 ATAC peaks, of which only 204 have 0.5 scores. Therefore, even the promoters with the highest score showed 1 or 2 peaks at most. We finally inspected the selected regions using the UCSC Genome Browser to confirm that the regions did not overlap with transcribed regions, and chose 12 of the resulting ATAC peaks for further analysis.

Library preparation

Genomic DNA was extracted using the Qiagen DNEasy Blood and Tissue kit, following manufacturer instructions for cultured cells. In some cases, the process was automated using a QiaCube (Qiagen). DNA concentration was measured by Nanodrop. If the concentration was below the limit of detection for the Nanodrop (5 ng/uL) or samples did not exhibit characteristic 260/280nm absorption ratios, samples were further concentrated

and purified using the Zymo Clean and Concentrator kit following manufacturer instructions for genomic DNA.

Amplicons were generated for sequencing by a 2- or 3-step PCR approach based on the Illumina 16S amplicon sequencing protocol. Region specific primers were designed for each target site with added Illumina adapter sequences (**Supplementary Table 2**). PCR reactions were setup using Phusion HF 2x Master Mix with gDNA template concentration 5-20 ng/uL and primer concentration 200 nM for the forward and reverse primers. Reactions were denatured for 30 sec at 98 C, then cycled for 35 rounds with 10 sec at 98 C, 30 sec at primer-set specific annealing temperatures chosen using the NEB Tm calculator tool (https://tmcalculator.neb.com/), and 30 sec at the extension temperature of 72 C. Finally, reactions were incubated at the extension temperature for 10 min and stored at 4 C for future processing.

Products were briefly centrifuged then purified with AMPureXP magnetic beads using a 1:1 ratio of sample:beads and following manufacturer instructions (Beckman Coulter). The purified round 1 PCR product served as the template for a second PCR reaction. Primers for the reaction contained Illumina adapter sequences and Nextera XT i5 and i7 combinatorial indices to allow for sample multiplexing. Reactions were cycled 35 times as above with an annealing temperature of 60 C. Samples were briefly centrifuged and purified using AMPure XP beads with a bead-to-sample volume ratio of 1.2:1.

For the samples used to generate Supplemental Figure 5B, unique molecular identifiers (UMIs) were incorporated to control for potential PCR bias. In this scheme, an additional

linear PCR step was included prior to round 1 PCR using only a region-specific forward primer containing the Illumina adapter sequence and a random sequence of 10 bases. The reaction was cycled between annealing and extension temperatures for 5 rounds, denaturing DNA template only once at the beginning of the reaction to ensure that each extended primer bound to only one DNA target. In this way, target sequences are linearly amplified with one UMI being attributable to each genomic DNA target present in the original sample. The products of this reaction were purified with AMPureXP beads using a bead-to-sample-volume ratio of 1:1 as described above. This reaction was used as the template for the round 1 exponential PCR reaction, priming with a locus specific reverse primer and a forward primer designed to bind the Illumina adapter overhang of the linearly extended fragments. From this point, samples were processed as described for the non-UMI samples above.

Samples were run out on 2% agarose gels either independently or in mixtures to assess product purity, and concentrations were measured by Nanodrop. Samples were pooled into 4 nM libraries at an equimolar ratio. These libraries were denatured by addition of 0.2 M NaOH, then mixed with similarly denatured PhiX control DNA to increase library diversity (5% PhiX). Samples were run pair end with either 2x250 cycle v2 or 2x300 cycle v3 MiSeq reagent kits following manufacturer instructions. Final libraries were loaded at a concentration of 4 pM.

Next generation sequencing and data analysis

Hypercascade amplicon sequences were aligned to reference files using the Burrows-Wheeler bwa-mem algorithm[106]. Only reads mapping to the correct target for each line were considered for further analysis. A custom R-script was used to extract base calls and quality scores for each editing target position in the reads. Base calls with quality less than 10 were treated as "N". The fraction of edited targets over time was computed as the fraction of G base calls divided by the sum of G and A base calls for a time point (**Figure 3C** and **Figure 4B**). For layer distribution analysis and mismatch edit rate estimation, reads containing "N" were excluded (**Figure 3D, Figure 4C-E, Supplemental Figures 3** and **4**). Reads were then grouped by total number of edits (G base calls) on the target sites. Groups with sample size less than 30 were excluded.

A Kallisto[136]-based pipline was applied to analyze data for Supplemental Figure 6A. Raw FASTQ files were directly aligned to a manually generated reference file containing, for each amplicon, four sequences: the unedited amplicon sequence from the mm10 genome (UCSC), the unedited sequence from the *Mus spretus* genome (Ensembl), and versions of each sequence containing the predicted edit.

For the data presented in Supplemental Figure 6B, sequencing of relatively long amplicons (~570 bp) was performed with 2x300 cycle paired end Illumina sequencing in order to obtain information for all edited positions on each individual read. Stringent quality filtering was applied to retain only base reads that had a quality score greater than 20 at the edited positions as described in the following paragraphs.

Raw paired-end FASTQ files were merged using BBMerge using the xloose setting (Department of Energy Joint Genome Institute). Where applicable, random 10-mer UMIs were extracted from reads using UMI-tools[137]. Merged reads were aligned to a reference FASTA file containing only the unedited *Mus musculus* and *Mus spretus* sequences using GraphMap, an aligner built to work well with error-prone reads[138]. The reference file was manually generated by combining short genomic sequences containing the amplicons of interest from both the *spretus* (Ensembl) and mm10 (UCSC) genome assemblies. SAM files were read into R and reads were queried for base calls and quality scores at editing target positions and species SNP locations.

Reads were deduplicated based on their UMI with UMI-tools[137]. Each read contains five SNPs that can be used to distinguish the X alleles. Alleles were classified by considering reads that only contained perfect matches to all five SNP positions for one species or the other. Editing was assessed at each target site individually, considering only reads that had a quality score greater than 20 at the target site. Samples were discarded if there were fewer than 30 reads remaining after UMI demultiplexing and quality filtering.

Mismatch edit rate estimation

To estimate mismatch edit rates, we assume, as in previous work[44,61,121,139], that there is a constant per-target edit rate for each possible gRNA/target combination (referred to here as an "edit pattern"). We assume that the rate of edit accumulation for an edit pattern is given by

$$\frac{dG}{dt} = k\phi s,$$

where G is the number of edits observed for that edit pattern, t is time, k is the per-target, per-unit-time editing rate constant (the parameter we are interested in), ϕ is a cell line intrinsic parameter reflecting the expression levels of ABE and gRNAs in a cell, and s is the number of target sites available for editing. Notably, since we are sampling from a polyclonal population, our measurements are taken from cells with a broad distribution of values. It is assumed that k and ϕ are independent of time.

Rearranging and integrating, p can be estimated by evaluating

$$k = \frac{G_T}{\phi \int_0^T s(t)dt}$$

at time T, where G_T is the number of edits measured at time T. Every read has a G_T associated with it. We also know the total number of edits accumulated on each read and the time when it was measured. We use the overall edit rate per read as an estimate of ϕ, normalizing out variance due to heterogeneous ABE and gRNA expression.

It is not possible to assume a functional form for $s(t)$ a priori due to the interdependent nature of editing. The functional form varies widely for different edit patterns. From the data, we can extract s as a function of the total number of edits per read, taking the mean value at each point. Assuming the total number of edits on a read increases linearly with time (an assumption that seems reasonable given the design of the hypercascade and data

in **Figures 3** and **4**), we can locate a read along the trajectory of s based on the total number of edits on that read, rescale the "edit time" into real time based on the time point when the read was collected, and integrate the mean s curve in real time for every read (**Supplemental Figures 2-4**). The mean and standard error of the estimates of k over all reads for each edit pattern are plotted in Supplemental Figure 4.

Stochastic simulations

Hypercascade editing was simulated in R using the Gillespie method(Gillespie 1977), where each target site is assigned a propensity to edit in a given time period. Briefly, waiting times between edits are assumed to be exponential. To simulate editing, a waiting time is drawn based on the total propensity to edit any site, then the specific site to be edited is randomly selected, weighting the choice in proportion to each unedited target's propensity. To investigate theoretical performance of the system, editing was constrained to only allow edits with perfect homology between gRNAs and protospacer sequences as well as completely intact PAM sites (**Figure 2**). For this part of the study, each site was assigned an identical propensity. Cell divisions were modeled by duplicating barcode sequences with waiting times drawn from a distribution derived from Eyring-Stover survival theory that has been shown to better model cell division times than the exponential distribution[108]. We chose to parameterize this distribution using $\tau = 24$ and $\alpha = 0.9$ to model hiPSCs, which typically divide approximately once per day. We screened a variety of overall edit rates, total experiment times, and target array copy numbers, averaging over 30 independent replicates for each parameter set (**Figure 2**).

Independently edited barcodes were simulated in a similar fashion, except without any attenuation of editing propensity based on the states of neighboring targets. To model Cas9-based editing, we augmented the independent model to incorporate nine different editing outcomes at each target site (all with equal propensities), mirroring the 3.2 bits of information identified on average in Cas9 cut sites[140].

To better capture editing dynamics in the real system, editing in the presence of mismatches and broken PAM sequences was allowed at an attenuated rate. Further, we allowed each layer its own protospacer-specific edit rate. Mismatch editing was modeled with a simplified, multiplicative model, where each position within the target where mismatches could occur was penalized (**Figure 3D, E**). We applied non-linear least squares fitting to polyclonal editing data to estimate parameters l_k and p_j. All simulation and analysis R scripts are available as supplemental data.

Chromatin state inference simulations

Hypercascade editing was simulated alongside cell divisions as described above, with each cell assigned 20 total barcode arrays, each containing 20 repeating subunits for a total of 74 editable sites per array. Fifteen of the sites were assigned a constant rate, simulating insertions into statically open chromatin. Five of the sites had variable edit rate, defined by the chromatin context of the site at different points in time during the simulation (**Figure 5F**). Editing and cell divisions were allowed to proceed for six total days, with open chromatin editing at a rate of two edits per barcode array per day and

closed chromatin editing 3.98-fold more slowly (based on empirical measurements from **Figure 5D**).

Simulations were divided into five evenly spaced epochs (**Figure 5F**). Four cell-state trajectories were defined over the course of the epochs, with dynamic chromatin trajectories assigned to each state. In some epochs, trajectories had redundant states (for example, in the first epoch all trajectories had the pattern open-open-open-open-closed, as visualized in **Figure 5E, F**). At each cell division, progeny had a chance to switch trajectories based on a predetermined cell state transition matrix, which itself changed in each epoch. Transitions were only allowed between trajectories while they had identical chromatin patterns.

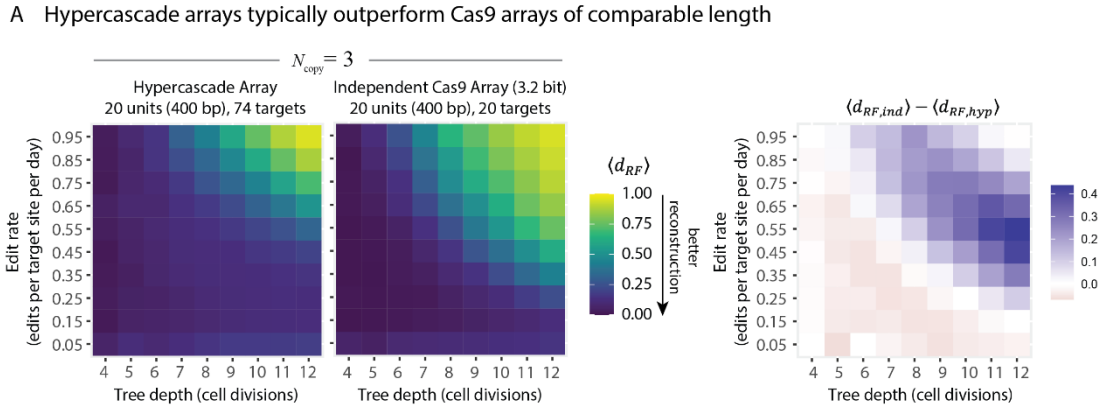
At the end of the 6-day period, lineages were reconstructed based on the resulting barcodes for each cell using UPGMA. Ancestral barcodes were inferred by taking the intersection of edit patterns for each child, starting at the leaves and working back to the root of the tree. Edit rates for each edge were estimated from these ancestral sequences, yielding noisy estimates of edit rate across time for each cell (**Figure 5G**).

To denoise these estimates, edit rate data for each cell was fit to a fourth-degree polynomial. Fitting coefficients were clustered using a four-center Gaussian Mixture Model (Mclust v6.1.1)[141]. Averaged trajectories across each group qualitatively recapitulated the ground truth dynamic chromatin time course (**Figure 5I**).

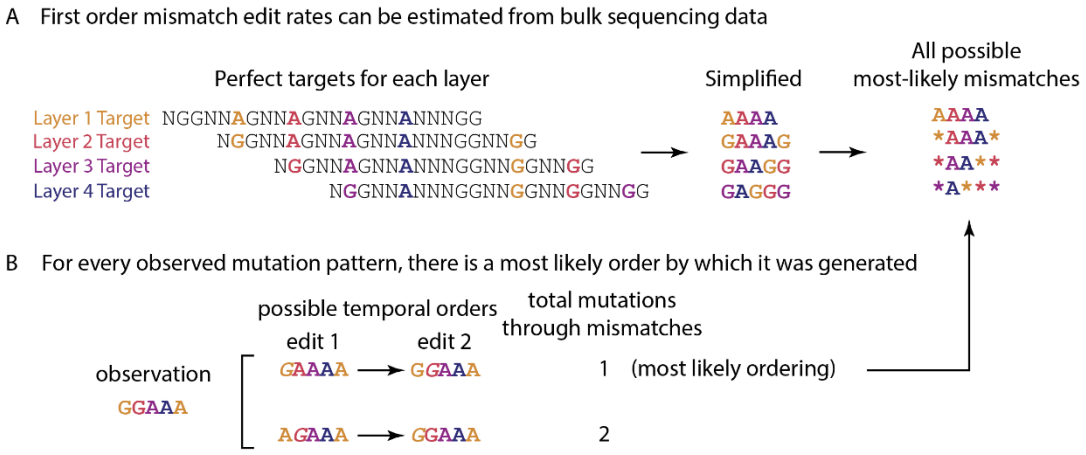
Data availability

All next generation sequencing data are available at the National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO), accession number GSE314635. All analysis scripts, custom code, and additional supplementary files are available at data.caltech.edu, DOI: 10.22002/d15ek-0dx91.

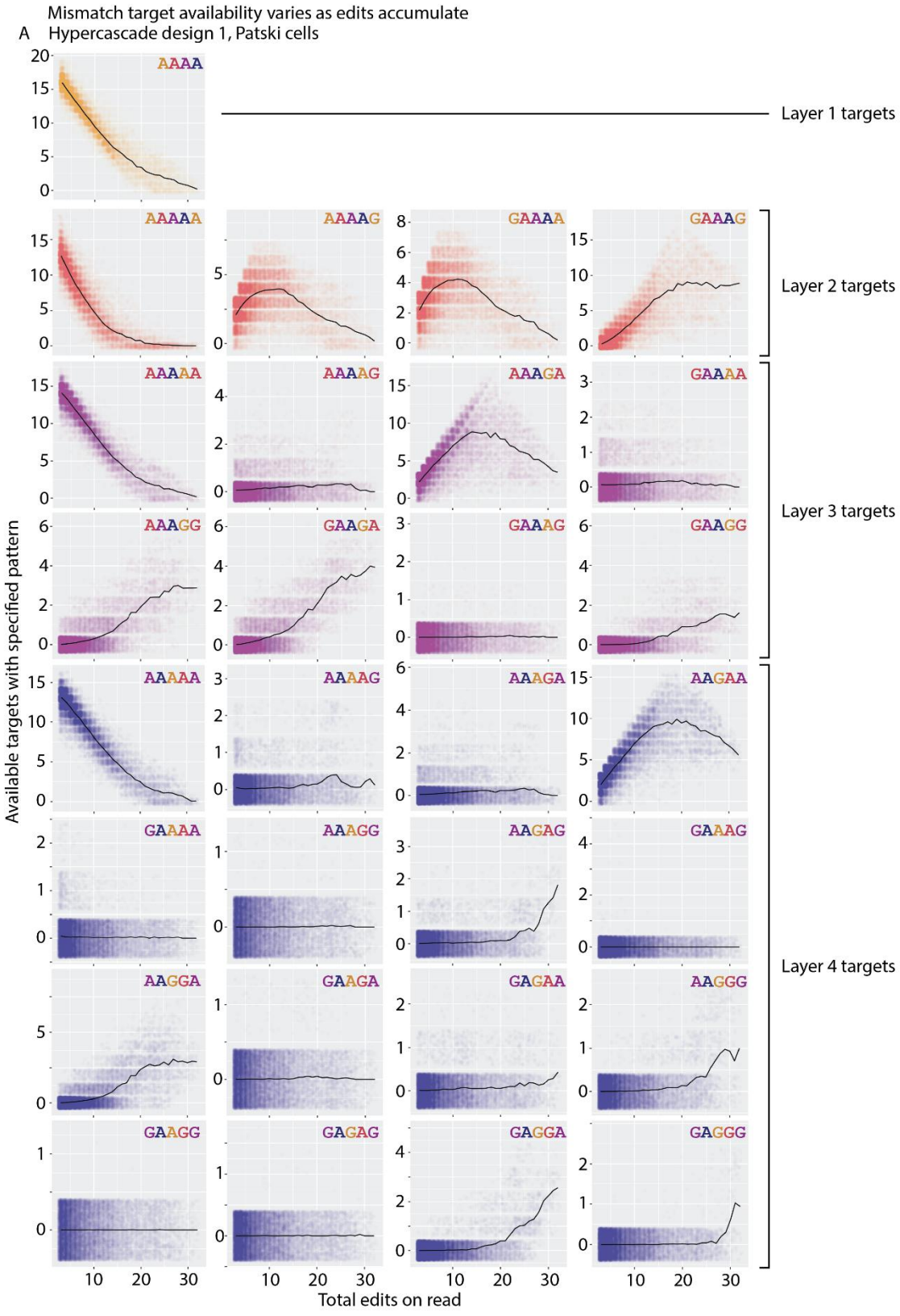
Supplemental information

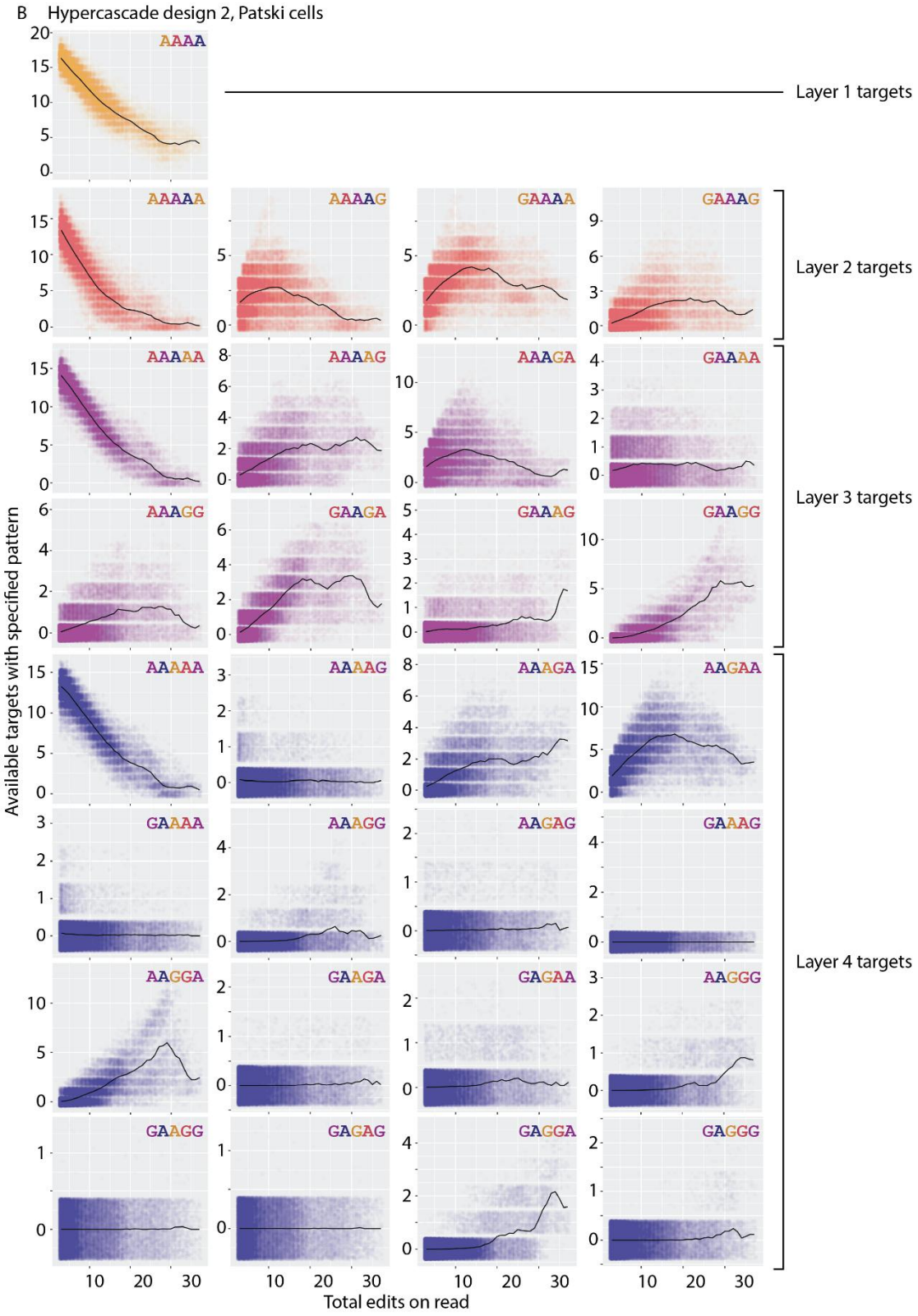


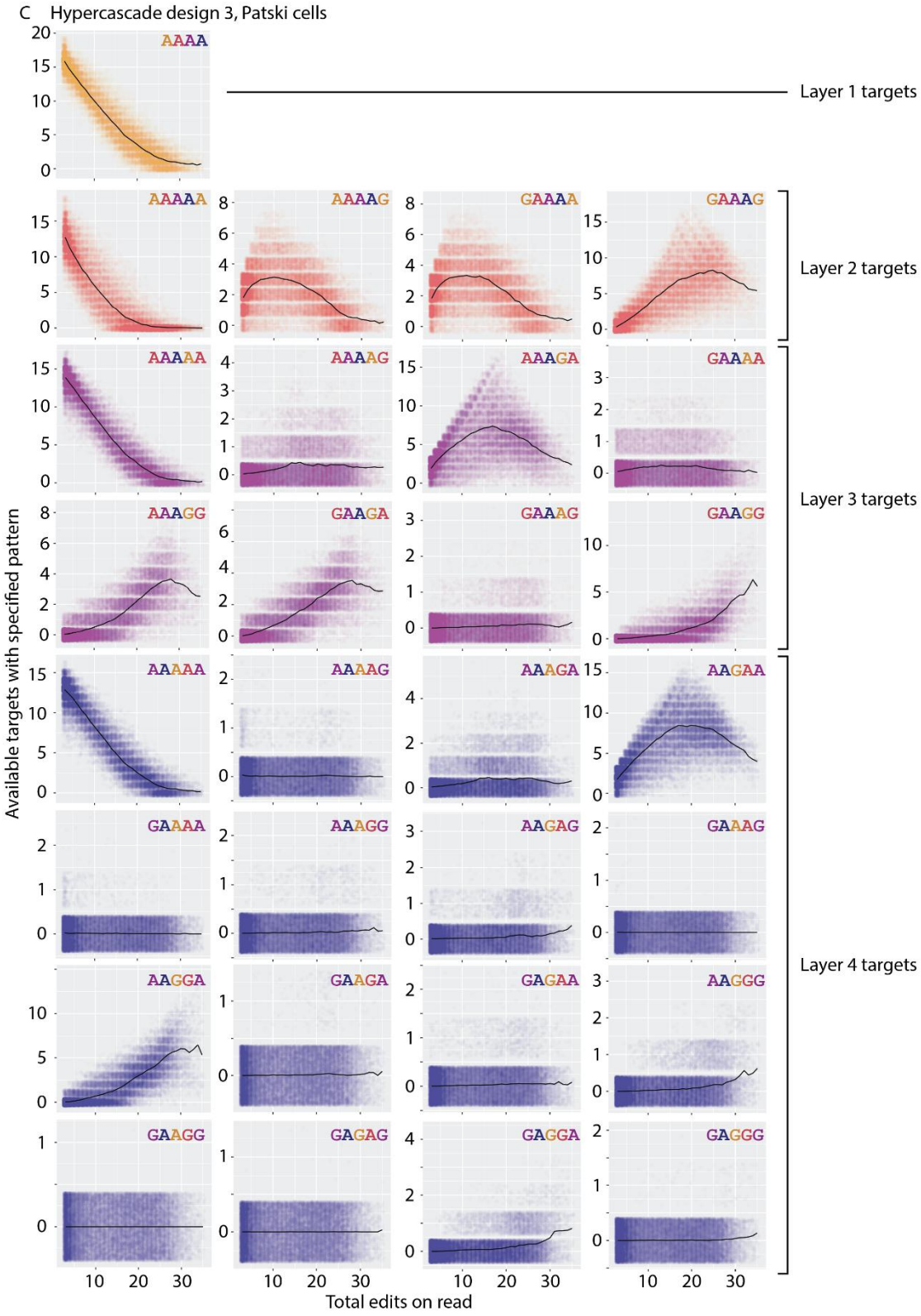
Supplemental Figure 1: The hypercascade typically outperforms Cas9 based systems of equal length in simulations. Cas9 can produce multiple editing outcomes where base editors produce only a single A-to-G transition. We used a simplified model of Cas9 editing by allowing for each target state to take one of eight different editing outcomes (see methods). Holding tree depth constant, the hypercascade system outperforms a simulated independent Cas9 target array with comparable sequence length over a variety of edit rates and copy numbers but underperforms slightly for low edit rates and tree depths.

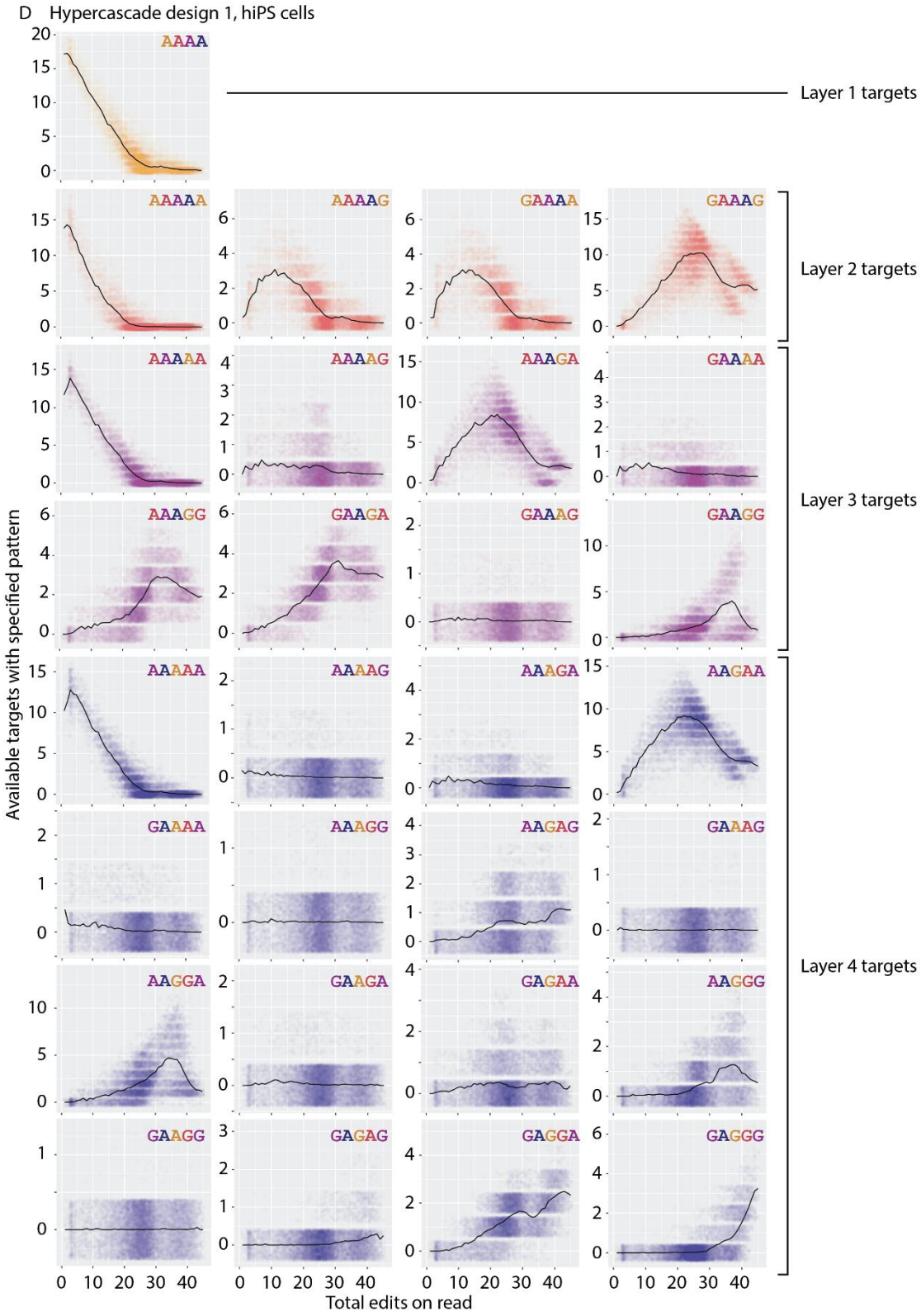


Supplemental Figure 2: Mismatch edit rates can be estimated. Bulk sequencing data can be used to estimate edit rates through either protospacer or PAM site mismatches under several assumptions **(A)**. This is possible by making the assumption that any edit pattern was generated in the most likely ordering of events **(B)**.


Mismatch target availability varies as edits accumulate
A Hypercascade design 1, Patski cells
Layer 1 targets
Layer 2 targets
Layer 3 targets
Layer 4 targets
Available targets with specified pattern
Total edits on read
AAAA
AAAAA
AAAAG
GAAAA
GAAAG
AAAAA
AAAAG
AAAGA
GAAAA
AAAGG
GAAGA
GAAAG
GAAGG
AAAAA
AAAAG
AAAGA
AAGAA
GAAAA
AAAGG
AAGAG
GAAAG
AAGGA
GAAGA
GAGAA
AAGGG
GAAGG
GAGAG
GAGGA
GAGGG

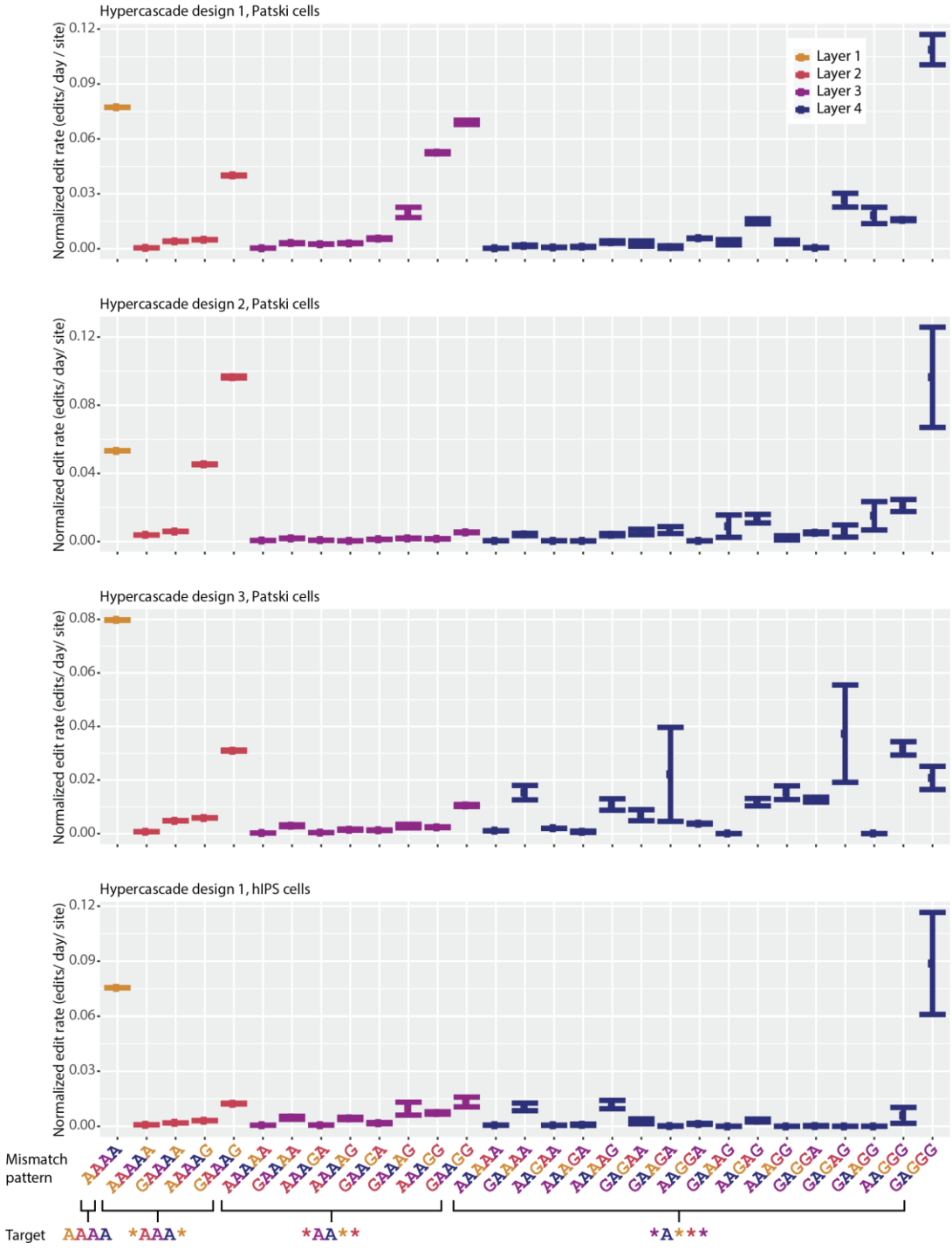

B Hypercascade design 2, Patski cells
Layer 1 targets
Layer 2 targets
Layer 3 targets
Layer 4 targets
Available targets with specified pattern
Total edits on read
AAAA
AAAAA
AAAAG
GAAAA
GAAAG
AAAAAA
AAAAAG
AAAGA
GAAAA
AAAGG
GAAGA
GAAAG
GAAGG
AAAAAA
AAAAAG
AAAGA
AAGAA
GAAAA
AAAGG
AAGAG
GAAAG
AAGGA
GAAGA
GAGAA
AAGGG
GAAGG
GAGAG
GAGGA
GAGGG


C Hypercascade design 3, Patski cells
Layer 1 targets
Layer 2 targets
Layer 3 targets
Layer 4 targets
Available targets with specified pattern
Total edits on read
AAAA
AAAAA
AAAAG
GAAAA
GAAAG
AAAAA
AAAAG
AAAGA
GAAAA
AAAGG
GAAGA
GAAAG
GAAGG
AAAAA
AAAAG
AAAGA
AAGAA
GAAAA
AAAGG
AAGAG
GAAAG
AAGGA
GAAGA
GAGAA
AAGGG
GAAGG
GAGAG
GAGGA
GAGGG

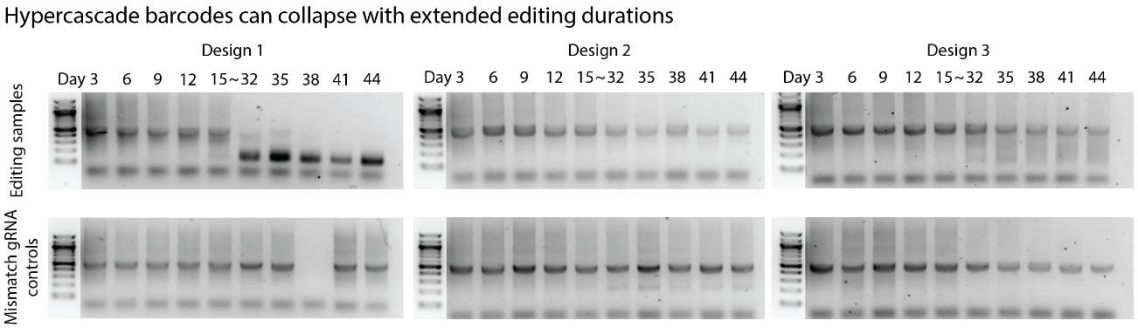

D Hypercascade design 1, hiPS cells
Layer 1 targets
Layer 2 targets
Layer 3 targets
Layer 4 targets
Available targets with specified pattern
Total edits on read
AAAA
AAAAA
AAAAG
GAAAA
GAAAG
AAAAA
AAAAG
AAAGA
GAAAA
AAAGG
GAAGA
GAAAG
GAAGG
AAAAA
AAAAG
AAAGA
AAGAA
GAAAA
AAAGG
AAGAG
GAAAG
AAGGA
GAAGA
GAGAA
AAGGG
GAAGG
GAGAG
GAGGA
GAGGG

Supplemental Figure 3: The editing process generates potential mismatch targets dynamically. To estimate mutation rates for a given sequence context, we need to know the number of target sites seen by the editor over time. This can be extracted from bulk sequencing data, both in Patski cells for designs 1 **(A)**, 2 **(B)**, and 3 **(C)**, as well as for design 1 in hiPSCs **(D)**.

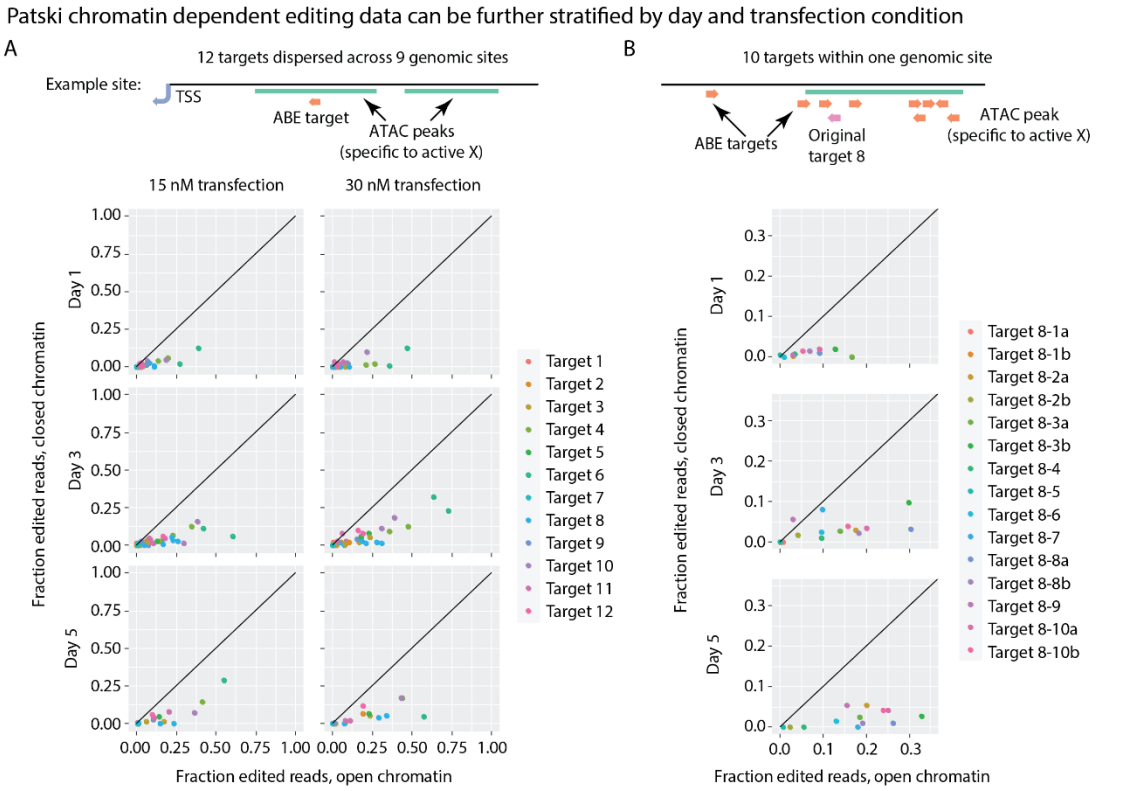
A Editing rate is reduced to varying extent bymismatches between the gRNA and target site



Supplemental Figure 4: Mismatches decrease the rate of editing across all layers. Mismatch edit rates for many sequence contexts can be estimated from bulk sequencing data. Typically gRNA-protospacer mismatches are estimated to decrease edit rate, with multiple mismatches and mismatches proximal to the PAM site having greater effects. Error bars represent standard error.



Supplemental Figure 5: Editing over long times can lead to target array collapse. Hypercascade sequences are amplified from bulk genomic DNA and expected to produce a 500 bp product.



Supplemental Figure 6: Open chromatin edits more rapidly than closed chromatin. (A) 12 targets dispersed across the X chromosome were targeted for base editing over a 1-5 day time course, transfecting either 15 or 30 nM of gRNA into cells constitutively expressing ABE. Targets were selected to have differential chromatin accessibility on the active and inactive X alleles based on existing ATAC sequencing data. Editing was quantified for each allele to determine whether chromatin context impacts ABE edit rate. **(B)** An additional 10 targets were selected and targeted for editing within and around a single differential ATAC peak from panel A. Cells were transfected with 30 nM gRNA only in this case.

Supplemental Table 1: Hypercascade target sequences investigated in this study.

	20mer Repeat	Layer 1 gRNA	Layer 2 gRNA	Layer 3 gRNA	Layer 4 gRNA
Target 1	AGGACAGTCAG ACAGTCATG…	AGGACAGTCA GACAGTCATG	CGGTCAGACA GTCATGAGGA	CGGACAGTCAT GAGGACGGT	CGGTCATGAG GACGGTCGGA
Target 2	AGGTCAGACAG TCAGACACA…	AGGTCAGACA GTCAGACACA	CGGACAGTCA GACACAAGGT	CGGTCAGACAC AAGGTCGGA	CGGACACAAG GTCGGACGGT
Target 3	AGGTCAGTCAG TAAGTAACG…	AGGTCAGTCA GTAAGTAACG	CGGTCAGTAA GTAACGAGGT	CGGTAAGTAAC GAGGTCGGT	AGGTAACGAG GTCGGTCGGT

Supplemental Table 2: Primer sequences used in this study.

Amplicon	Primer Sequence
g01 (Supplemental Figure 6A), Forward, with Illumina Adapters	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGTAACCAAAGGCTGGAGCTC
g01 (Supplemental Figure 6A), Reverse, with Illumina Adapters	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCAGCGCCCACTATTAAAGTACC
g02 (Supplemental Figure 6A), Forward, with Illumina Adapters	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTGTGATGTGCATCCAGTAATGTTCC
g02 (Supplemental Figure 6A), Reverse, with Illumina Adapters	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGAGTTGTCCAAGGCCACTGAG
g03 (Supplemental Figure 6A), Forward, with Illumina Adapters	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCGTTTCAGCCAAAGTACTCCTCAC
g03 (Supplemental Figure 6A), Reverse, with Illumina Adapters	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGCTAGCTCCTGACTGATCAG
g04 (Supplemental Figure 6A), Forward, with Illumina Adapters	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGTACGTGCCCCAAGATAACC
g04 (Supplemental Figure 6A), Reverse, with Illumina Adapters	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTTGCCCTGCTCTAGCCTTG
g05 (Supplemental Figure 6A), Forward, with Illumina Adapters	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTGGCTAATGCCAAGGACTCTG
g05 (Supplemental Figure 6A), Reverse, with Illumina Adapters	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGCTGGCCAAAGGACCATC
g06 (Supplemental Figure 6A), Forward, with Illumina Adapters	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTCAAAGAAGGCCAAAGCCAAG
g06 (Supplemental Figure 6A), Reverse, with Illumina Adapters	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTCATCCGGATCCTCACCAATC
g07 (Supplemental Figure 6A), Forward, with Illumina Adapters	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGATAAACCTAGACAATGCTTCCAGG
g07 (Supplemental Figure 6A), Reverse, with Illumina Adapters	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCACTGAGCTACATTCTGAGTCC
g08 (Supplemental Figure 6A), Forward, with Illumina Adapters	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCGAGCAAAACTAGGAGTCTTTCC
g08 (Supplemental Figure 6A), Reverse, with Illumina Adapters	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCTCTGGCTTCTGTAGTGGG
g09 (Supplemental Figure 6A), Forward, with Illumina Adapters	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTCTGTTTATGCACAGTGCCG
g09 (Supplemental Figure 6A), Reverse, with Illumina Adapters	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTTGACGACACCCACTGGC
g10 (Supplemental Figure 6A), Forward, with Illumina Adapters	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAGGCAAGAAACTTCTGCCAC
g10 (Supplemental Figure 6A), Reverse, with Illumina Adapters	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCTGTGATTGGTATATGAGGCAGG
g11 (Supplemental Figure 6A), Forward, with Illumina Adapters	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTGCATTGCATACCCAGAGTTTC
g11 (Supplemental Figure 6A), Reverse, with Illumina Adapters	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCTCAGGGTAACACACATGAGC

g12 (Supplemental Figure 6A), Forward, with Illumina Adapters	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGCTCCGAGGTACTGGAAAGG
g12 (Supplemental Figure 6A), Reverse, with Illumina Adapters	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGAAGAGTCAAAGCTGGGACACAG
Long Amplicon for Region 8 with UMI (Supplemental Figure 6B), Linear PCR Primer, with Illumina Adapters	CAGCGTCAGATGTGTATAAGAGACAGNNNNNNNNNNGCCACATGATGGCTCACAAC
Long Amplicon for Region 8 with UMI (Supplemental Figure 6B), Exponential PCR Forward Primer, with Illumina Adapters	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG
Hypercascade target, Forward, with Illumina Adapters	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAGTGCAGTGCTTGATAACAGG
Hypercascade target, Reverse, with Illumina Adapters	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGATAGTCTGCGTAAAATTGACGCATG

Supplemental Table 3: Homology sequences for Patski guide RNA targets (Figure 5D and Supplemental Figure 6).

gRNA Target	**Homology Sequence**
g01 (Supplemental Figure 6A)	CGTCGACTGCTGGTCACGTG
g02 (Supplemental Figure 6A)	TGTCCATCTCTGGACACAGG
g03 (Supplemental Figure 6A)	TCCAAAATGAGTAGAGACCG
g04 (Supplemental Figure 6A)	TAGGACTGTGTTACAAACAG
g05 (Supplemental Figure 6A)	TACCCAGCTCTTTGTCACAG
g06 (Supplemental Figure 6A)	ACACACCAAGAGATAGAGAG
g07 (Supplemental Figure 6A)	AAGTCAATTAAATGGCTGCA
g08 (Supplemental Figure 6A)	CGGAACACTGACCATGGTCA
g09 (Supplemental Figure 6A)	GCCGACTTTGAAATTCGAGG
g10 (Supplemental Figure 6A)	ACCCCAGTGAGCAATGACAG
g11 (Supplemental Figure 6A)	TTGGAGAAAGCACGTCACTG
g12 (Supplemental Figure 6A)	ACTGTACCGAGCTAGCTCGG
gRNA 1 (Supplemental Figure 6B)	TAGTAAACATTATGCACATA
gRNA 2 (Supplemental Figure 6B)	CATCAAACAACCTTAATAAG
gRNA 3 (Supplemental Figure 6B)	TTTGAAGTGAGCCCTGACCA
gRNA 4 (Supplemental Figure 6B)	Same as g08 (Supplemental Figure 6A)
gRNA 5 (Supplemental Figure 6B)	CCTTCATAACTTGCCAACAA
gRNA 6 (Supplemental Figure 6B)	TTGCCACGTCTGGATGCTCA
gRNA 7 (Supplemental Figure 6B)	TTACCATGAGCATCCAGACG
gRNA 8 (Supplemental Figure 6B)	GTAAAAGAGATTAGTAACCC
gRNA 9 (Supplemental Figure 6B)	GGAGGAGGCACAGTCATCCT
gRNA 10 (Supplemental Figure 6B)	AGGGAAGTAGGCTATGCAGG

BIBLIOGRAPHY

Ai, Zhiying, Jingjing Shao, Yongyan Wu, Mengying Yu, Juan Du, Xiaoyan Shi, Xinglong Shi, Yong Zhang, and Zekun Guo. 2016. "CHIR99021 Enhances Klf4 Expression through β-Catenin Signaling and MiR-7a Regulation in J1 Mouse Embryonic Stem Cells." *PloS One* 11 (3): e0150936.

Arbab, Mandana, Max W. Shen, Beverly Mok, Christopher Wilson, Żaneta Matuszek, Christopher A. Cassa, and David R. Liu. 2020. "Determinants of Base Editing Outcomes from Target Library Analysis and Machine Learning." *Cell* 182 (2): 463-480.e30.

Askary, Amjad, Luis Sanchez-Guardado, James M. Linton, Duncan M. Chadly, Mark W. Budde, Long Cai, Carlos Lois, and Michael B. Elowitz. 2020. "In Situ Readout of DNA Barcodes and Single Base Edits Facilitated by in Vitro Transcription." *Nature Biotechnology* 38 (1): 66–75.

Behjati, Sam, Meritxell Huch, Ruben van Boxtel, Wouter Karthaus, David C. Wedge, Asif U. Tamuri, Inigo Martincorena, et al. 2014. "Genome Sequencing of Normal Cells Reveals Developmental Lineages and Mutational Processes." *Nature* 513 (7518): 422–25.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 57 (1): 289–300.

Berg, Stuart, Dominik Kutra, Thorben Kröger, Christoph-Nikolas Straehle, Bernhard Kausler, Carsten Haubold, Martin Schiegg, et al. n.d. *Ilastik: Interactive Machine Learning for (Bio)Image Analysis*. Universitätsbibliothek Heidelberg.

Berge, Derk ten, Dorota Kurek, Tim Blauwkamp, Wouter Koole, Alex Maas, Elif Eroglu, Ronald K. Siu, and Roel Nusse. 2011. "Embryonic Stem Cells Require Wnt Proteins to Prevent Differentiation to Epiblast Stem Cells." *Nature Cell Biology* 13 (9): 1070–75.

Bessonnard, Sylvain, Laurane De Mot, Didier Gonze, Manon Barriol, Cynthia Dennis, Albert Goldbeter, Geneviève Dupont, and Claire Chazaud. 2014. "Gata6, Nanog and Erk Signaling Control Cell Fate in the Inner Cell Mass through a Tristable Regulatory Network." *Development (Cambridge, England)* 141 (19): 3637–48.

Bonora, G., X. Deng, H. Fang, V. Ramani, R. Qiu, J. B. Berletch, G. N. Filippova, et al. 2018. "Orientation-Dependent Dxz4 Contacts Shape the 3D Structure of the Inactive X Chromosome." *Nature Communications* 9 (1): 1445.

Bouckaert, Remco. 2016. "Phylogeography by Diffusion on a Sphere: Whole World Phylogeography." *PeerJ* 4 (e2406): e2406.

Bouckaert, Remco, Timothy G. Vaughan, Joëlle Barido-Sottani, Sebastián Duchêne, Mathieu Fourment, Alexandra Gavryushkina, Joseph Heled, et al. 2019. "BEAST 2.5: An Advanced Software Platform for Bayesian Evolutionary Analysis." *PLoS Computational Biology* 15 (4): e1006650.

Bowling, Sarah, Duluxan Sritharan, Fernando G. Osorio, Maximilian Nguyen, Priscilla Cheung, Alejo Rodriguez-Fraticelli, Sachin Patel, et al. 2020. "An Engineered CRISPR-Cas9 Mouse Line for Simultaneous Readout of Lineage Histories and Gene Expression Profiles in Single Cells." *Cell* 181 (7): 1693–94.

Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. “Near-Optimal Probabilistic RNA-Seq Quantification.” *Nature Biotechnology* 34 (5): 525–27.

Casey, Michael J., Patrick S. Stumpf, and Ben D. MacArthur. 2020. “Theory of Cell Fate.” *Wiley Interdisciplinary Reviews. Systems Biology and Medicine* 12 (2): e1471.

Chen, Kok Hao, Alistair N. Boettiger, Jeffrey R. Moffitt, Siyuan Wang, and Xiaowei Zhuang. 2015. “RNA Imaging. Spatially Resolved, Highly Multiplexed RNA Profiling in Single Cells.” *Science (New York, N.Y.)* 348 (6233): aaa6090.

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

Chen, Wei, Aaron McKenna, Jacob Schreiber, Maximilian Haeussler, Yi Yin, Vikram Agarwal, William Stafford Noble, and Jay Shendure. 2019. “Massively Parallel Profiling and Predictive Modeling of the Outcomes of CRISPR/Cas9-Mediated Double-Strand Break Repair.” *Nucleic Acids Research* 47 (15): 7989–8003.

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

Chow, Ke-Huan K., Mark W. Budde, Alejandro A. Granados, Maria Cabrera, Shinae Yoon, Soomin Cho, Ting-Hao Huang, et al. 2021. “Imaging Cell Lineage with a Synthetic Digital Recording System.” *Science (New York, N.Y.)* 372 (6538): eabb3099.

Clavería, Cristina, Giovanna Giovinazzo, Rocío Sierra, and Miguel Torres. 2013. “Myc-Driven Endogenous Cell Competition in the Early Mammalian Embryo.” *Nature* 500 (7460): 39–44.

Conklin, Edwin Grant. 1905. *The Organization and Cell-Lineage of the Ascidian Egg / by Edwin G. Conklin*. Philadelphia : [Academy of Natural Sciences],.

Das, Atze T., Liliane Tenenbaum, and Ben Berkhout. 2016. “Tet-on Systems for Doxycycline-Inducible Gene Expression.” *Current Gene Therapy* 16 (3): 156–67.

Davies, Jamie A. 2008. “Synthetic Morphology: Prospects for Engineered, Self-Constructing Anatomies.” *Journal of Anatomy* 212 (6): 707–19.

Davies, Jamie, and Michael Levin. 2023. “Synthetic Morphology with Agential Materials.” *Nature Reviews Bioengineering* 1 (1): 46–59.

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

Domcke, Silvia, and Jay Shendure. 2023. “A Reference Cell Tree Will Serve Science Better than a Reference Cell Atlas.” *Cell* 186 (6): 1103–14.

Eckersley-Maslin, Mélanie A., Valentine Svensson, Christel Krueger, Thomas M. Stubbs, Pascal Giehr, Felix Krueger, Ricardo J. Miragaia, et al. 2016. “MERVL/Zscan4 Network Activation Results in Transient Genome-Wide DNA Demethylation of MESCs.” *Cell Reports* 17 (1): 179–92.

Edelstein, Arthur, Nenad Amodaj, Karl Hoover, Ron Vale, and Nico Stuurman. 2010. “Computer Control of Microscopes Using MManager.” *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]* Chapter 14 (October): Unit14.20.

Ellis, Stephanie J., Nicholas C. Gomez, John Levorse, Aaron F. Mertz, Yejing Ge, and Elaine Fuchs. 2019. “Distinct Modes of Cell Competition Shape Mammalian Tissue Morphogenesis.” *Nature* 569 (7757): 497–502.

Emert, Benjamin L., Christopher J. Cote, Eduardo A. Torre, Ian P. Dardani, Connie L. Jiang, Naveen Jain, Sydney M. Shaffer, and Arjun Raj. 2021. "Variability within Rare Cell States Enables Multiple Paths toward Drug Resistance." *Nature Biotechnology* 39 (7): 865–76.

Eng, Chee-Huat Linus, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulena, Yodai Takei, Jina Yun, et al. 2019. "Transcriptome-Scale Super-Resolved Imaging in Tissues by RNA SeqFISH." *Nature* 568 (7751): 235–39.

Engler, Carola, Romy Kandzia, and Sylvestre Marillonnet. 2008. "A One Pot, One Step, Precision Cloning Method with High Throughput Capability." *PloS One* 3 (11): e3647.

Ferrell, James E., Jr. 2012. "Bistability, Bifurcations, and Waddington's Epigenetic Landscape." *Current Biology: CB* 22 (11): R458-66.

Frieda, Kirsten L., James M. Linton, Sahand Hormoz, Joonhyuk Choi, Ke-Huan K. Chow, Zakary S. Singer, Mark W. Budde, Michael B. Elowitz, and Long Cai. 2017. "Synthetic Recording and in Situ Readout of Lineage Information in Single Cells." *Nature* 541 (7635): 107–11.

Gao, Yangbin, and Yunde Zhao. 2014. "Self-Processing of Ribozyme-Flanked RNAs into Guide RNAs in Vitro and in Vivo for CRISPR-Mediated Genome Editing." *Journal of Integrative Plant Biology* 56 (4): 343–49.

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

Gillespie, Daniel T. 1977. "Exact Stochastic Simulation of Coupled Chemical Reactions." *The Journal of Physical Chemistry* 81 (25): 2340–61.

Gong, Wuming, Alejandro A. Granados, Jingyuan Hu, Matthew G. Jones, Ofir Raz, Irepan Salvador-Martínez, Hanrui Zhang, et al. 2021. "Benchmarked Approaches for Reconstruction of in Vitro Cell Lineages and in Silico Models of C. Elegans and M. Musculus Developmental Trees." *Cell Systems* 12 (8): 810-826.e4.

Goto, Yuji, Junko Matsui, and Nobuo Takagi. 2002. "Developmental Potential of Mouse Tetraploid Cells in Diploid Tetraploid Chimeric Embryos." *The International Journal of Developmental Biology* 46 (5): 741–45.

Hadas, Ron, Hernan Rubinstein, Markus Mittnenzweig, Yoav Mayshar, Raz Ben-Yair, Saifeng Cheng, Alejandro Aguilera-Castrejon, et al. 2024. "Temporal BMP4 Effects on Mouse Embryonic and Extraembryonic Development." *Nature* 634 (8034): 652–61.

Hershberg, Elliot A., Conor K. Camplisson, Jennie L. Close, Sahar Attar, Ryan Chern, Yuzhen Liu, Shreeram Akilesh, Philip R. Nicovich, and Brian J. Beliveau. 2021. "PaintSHOP Enables the Interactive Design of Transcriptome- and Genome-Scale Oligonucleotide FISH Experiments." *Nature Methods* 18 (8): 937–44.

Hormoz, Sahand, Zakary S. Singer, James M. Linton, Yaron E. Antebi, Boris I. Shraiman, and Michael B. Elowitz. 2016. "Inferring Cell-State Transition Dynamics from Lineage Trees and Endpoint Single-Cell Measurements." *Cell Systems* 3 (5): 419-433.e8.

Hsu, Patrick D., David A. Scott, Joshua A. Weinstein, F. Ann Ran, Silvana Konermann, Vineeta Agarwala, Yinqing Li, et al. 2013. "DNA Targeting Specificity of RNA-Guided Cas9 Nucleases." *Nature Biotechnology* 31 (9): 827–32.

Hwang, Byungjin, Wookjae Lee, Soo-Young Yum, Yujin Jeon, Namjin Cho, Goo Jang, and Duhee Bang. 2019. "Lineage Tracing Using a Cas9-Deaminase Barcoding System Targeting Endogenous L1 Elements." *Nature Communications* 10 (1): 1234.

Ichikawa, Takehiko, Kenichi Nakazato, Philipp J. Keller, Hiroko Kajiura-Kobayashi, Ernst H. K. Stelzer, Atsushi Mochizuki, and Shigenori Nonaka. 2013. “Live Imaging of Whole Mouse Embryos during Gastrulation: Migration Analyses of Epiblast and Mesodermal Cells.” *PloS One* 8 (7): e64506.

Imbert, Arthur, Wei Ouyang, Adham Safieddine, Emeline Coleno, Christophe Zimmer, Edouard Bertrand, Thomas Walter, and Florian Mueller. 2022. “FISH-Quant v2: A Scalable and Modular Tool for SmFISH Image Analysis.” *RNA (New York, N.Y.)* 28 (6): 786–95.

James, R. M., A. H. Klerkx, M. Keighren, J. H. Flockhart, and J. D. West. 1995. “Restricted Distribution of Tetraploid Cells in Mouse Tetraploiddiploid Chimaeras.” *Developmental Biology* 167 (1): 213–26.

Ju, Young Seok, Inigo Martincorena, Moritz Gerstung, Mia Petljak, Ludmil B. Alexandrov, Raheleh Rahbari, David C. Wedge, et al. 2017. “Somatic Mutations Reveal Asymmetric Cellular Dynamics in the Early Human Embryo.” *Nature* 543 (7647): 714–18.

Kalhor, Reza, Kian Kalhor, Leo Mejia, Kathleen Leeper, Amanda Graveline, Prashant Mali, and George M. Church. 2018. “Developmental Barcoding of Whole Mouse via Homing CRISPR.” *Science (New York, N.Y.)* 361 (6405): eaat9804.

Kallimasioti-Pazi, Eirini M., Keerthi Thelakkad Chathoth, Gillian C. Taylor, Alison Meynert, Tracy Ballinger, Martijn J. E. Kelder, Sébastien Lalevée, Ildem Sanli, Robert Feil, and Andrew J. Wood. 2018. “Heterochromatin Delays CRISPR-Cas9 Mutagenesis but Does Not Influence the Outcome of Mutagenic DNA Repair.” *PLoS Biology* 16 (12): e2005595.

Kemphues, K. J., J. R. Priess, D. G. Morton, and N. S. Cheng. 1988. “Identification of Genes Required for Cytoplasmic Localization in Early C. Elegans Embryos.” *Cell* 52 (3): 311–20.

Kim, Somang, Jimmy B. Yuan, Wendy S. Woods, Destry A. Newton, Pablo Perez-Pinera, and Jun S. Song. 2023. “Chromatin Structure and Context-Dependent Sequence Features Control Prime Editing Efficiency.” *BioRxiv : The Preprint Server for Biology*, April. https://doi.org/10.1101/2023.04.15.536944.

Koblan, Luke W., Kathryn E. Yost, Pu Zheng, William N. Colgan, Matthew G. Jones, Dian Yang, Arhan Kumar, et al. 2025. “High-Resolution Spatial Mapping of Cell State and Lineage Dynamics in Vivo with PEtracer.” *Science (New York, N.Y.)*, no. eadx3800 (July): eadx3800.

Kolodziejczyk, Aleksandra A., Jong Kyoung Kim, Jason C. H. Tsang, Tomislav Ilicic, Johan Henriksson, Kedar N. Natarajan, Alex C. Tuck, et al. 2015. “Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation.” *Cell Stem Cell* 17 (4): 471–85.

Kumar, Roshan M., Patrick Cahan, Alex K. Shalek, Rahul Satija, Ajay DaleyKeyser, Hu Li, Jin Zhang, et al. 2014. “Deconstructing Transcriptional Heterogeneity in Pluripotent Stem Cells.” *Nature* 516 (7529): 56–61.

Lang, Charles F., and Edwin Munro. 2017. “The PAR Proteins: From Molecular Circuits to Dynamic Self-Stabilizing Cell Polarity.” *Development (Cambridge, England)* 144 (19): 3405–16.

Lareau, Caleb A., Leif S. Ludwig, Christoph Muus, Satyen H. Gohil, Tongtong Zhao, Zachary Chiang, Karin Pelka, et al. 2021. “Massively Parallel Single-Cell Mitochondrial DNA Genotyping and Chromatin Profiling.” *Nature Biotechnology* 39 (4): 451–61.

Lawrence, Peter A., and Michael Levine. 2006. “Mosaic and Regulative Development: Two Faces of One Coin.” *Current Biology: CB* 16 (7): R236-9.

Lawson, K. A. 1999. “Fate Mapping the Mouse Embryo.” *The International Journal of Developmental Biology* 43 (7): 773–75.

Lawson, K. A., J. J. Meneses, and R. A. Pedersen. 1991. “Clonal Analysis of Epiblast Fate during Germ Layer Formation in the Mouse Embryo.” *Development (Cambridge, England)* 113 (3): 891–911.

Leeper, Kathleen, Kian Kalhor, Andyna Vernet, Amanda Graveline, George M. Church, Prashant Mali, and Reza Kalhor. 2021. “Lineage Barcoding in Mice with Homing CRISPR.” *Nature Protocols* 16 (4): 2088–2108.

Lee-Six, Henry, Nina Friesgaard Øbro, Mairi S. Shepherd, Sebastian Grossmann, Kevin Dawson, Miriam Belmonte, Robert J. Osborne, et al. 2018. “Population Dynamics of Normal Human Blood Inferred from Somatic Mutations.” *Nature* 561 (7724): 473–78.

Lemoine, F., J-B Domelevo Entfellner, E. Wilkinson, D. Correia, M. Dávila Felipe, T. De Oliveira, and O. Gascuel. 2018. “Renewing Felsenstein’s Phylogenetic Bootstrap in the Era of Big Data.” *Nature* 556 (7702): 452–56.

Li, Heng. 2013. “Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM.” *ArXiv [q-Bio.GN]*. arXiv. http://arxiv.org/abs/1303.3997.

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

Lingenfelter, P. A., D. A. Adler, D. Poslinski, S. Thomas, R. W. Elliott, V. M. Chapman, and C. M. Disteche. 1998. “Escape from X Inactivation of Smcx Is Preceded by Silencing during Mouse Development.” *Nature Genetics* 18 (3): 212–13.

Listgarten, Jennifer, Michael Weinstein, Benjamin P. Kleinstiver, Alexander A. Sousa, J. Keith Joung, Jake Crawford, Kevin Gao, et al. 2018. “Prediction of Off-Target Activities for the End-to-End Design of CRISPR Guide RNAs.” *Nature Biomedical Engineering* 2 (1): 38–47.

Liu, Kehui, Shanjun Deng, Chang Ye, Zeqi Yao, Jianguo Wang, Han Gong, Li Liu, and Xionglei He. 2021. “Mapping Single-Cell-Resolution Cell Phylogeny Reveals Cell Population Dynamics during Organ Development.” *Nature Methods* 18 (12): 1506–14.

Lodato, Michael A., Mollie B. Woodworth, Semin Lee, Gilad D. Evrony, Bhaven K. Mehta, Amir Karger, Soohyun Lee, et al. 2015. “Somatic Mutation in Single Human Neurons Tracks Developmental and Transcriptional History.” *Science (New York, N.Y.)* 350 (6256): 94–98.

Loveless, Theresa B., Courtney K. Carlson, Catalina A. Dentzel Helmy, Vincent J. Hu, Sara K. Ross, Matt C. Demelo, Ali Murtaza, et al. 2025. “Open-Ended Molecular Recording of Sequential Cellular Events into DNA.” *Nature Chemical Biology* 21 (4): 512–21.

Lubeck, Eric, Ahmet F. Coskun, Timur Zhiyentayev, Mubhij Ahmad, and Long Cai. 2014. “Single-Cell in Situ RNA Profiling by Sequential Hybridization.” *Nature Methods*. Springer Science and Business Media LLC.

Ludwig, Leif S., Caleb A. Lareau, Jacob C. Ulirsch, Elena Christian, Christoph Muus, Lauren H. Li, Karin Pelka, et al. 2019. “Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics.” *Cell* 176 (6): 1325-1339.e22.

Madan, Esha, António M. Palma, Vignesh Vudatha, Amit Kumar, Praveen Bhoopathi, Jochen Wilhelm, Tytus Bernas, et al. 2024. “Ovarian Tumor Cells Gain Competitive Advantage by Actively Reducing the Cellular Fitness of Microenvironment Cells.” *Nature Biotechnology*, December. https://doi.org/10.1038/s41587-024-02453-3.

Madan, Esha, Christopher J. Pelham, Masaki Nagane, Taylor M. Parker, Rita Canas-Marques, Kimberly Fazio, Kranti Shaik, et al. 2019. “Flower Isoforms Promote Competitive Growth in Cancer.” *Nature* 572 (7768): 260–64.

McDole, Katie, Léo Guignard, Fernando Amat, Andrew Berger, Grégoire Malandain, Loïc A. Royer, Srinivas C. Turaga, Kristin Branson, and Philipp J. Keller. 2018. “In Toto Imaging and Reconstruction of Post-Implantation Mouse Development at the Single-Cell Level.” *Cell* 175 (3): 859-876.e33.

McKenna, Aaron, Gregory M. Findlay, James A. Gagnon, Marshall S. Horwitz, Alexander F. Schier, and Jay Shendure. 2016. “Whole-Organism Lineage Tracing by Combinatorial and Cumulative Genome Editing.” *Science (New York, N.Y.)* 353 (6298): aaf7907.

Mittnenzweig, Markus, Yoav Mayshar, Saifeng Cheng, Raz Ben-Yair, Ron Hadas, Yoach Rais, Elad Chomsky, et al. 2021. “A Single-Embryo, Single-Cell Time-Resolved Model for Mouse Gastrulation.” *Cell* 184 (11): 2825-2842.e22.

Moffitt, Jeffrey R., Junjie Hao, Guiping Wang, Kok Hao Chen, Hazen P. Babcock, and Xiaowei Zhuang. 2016. “High-Throughput Single-Cell Gene-Expression Profiling with Multiplexed Error-Robust Fluorescence in Situ Hybridization.” *Proceedings of the National Academy of Sciences of the United States of America* 113 (39): 11046–51.

Molenaar, Miranda, Marc van de Wetering, Mariette Oosterwegel, Josi Peterson-Maduro, Susan Godsave, Vladimir Korinek, Jeroen Roose, Olivier Destrée, and Hans Clevers. 1996. “XTcf-3 Transcription Factor Mediates β-Catenin-Induced Axis Formation in Xenopus Embryos.” *Cell* 86 (3): 391–99.

Murphy, J. S., F. R. Landsberger, T. Kikuchi, and I. Tamm. 1984. “Occurrence of Cell Division Is Not Exponentially Distributed: Differences in the Generation Times of Sister Cells Can Be Derived from the Theory of Survival of Populations.” *Proceedings of the National Academy of Sciences of the United States of America* 81 (8): 2379–83.

Nagy, A., E. Gócza, E. M. Diaz, V. R. Prideaux, E. Iványi, M. Markkula, and J. Rossant. 1990. “Embryonic Stem Cells Alone Are Able to Support Fetal Development in the Mouse.” *Development (Cambridge, England)* 110 (3): 815–21.

Nagy, A., J. Rossant, R. Nagy, W. Abramow-Newerly, and J. C. Roder. 1993. “Derivation of Completely Cell Culture-Derived Mice from Early-Passage Embryonic Stem Cells.” *Proceedings of the National Academy of Sciences of the United States of America* 90 (18): 8424–28.

Nishida, H. 1987. “Cell Lineage Analysis in Ascidian Embryos by Intracellular Injection of a Tracer Enzyme. III. Up to the Tissue Restricted Stage.” *Developmental Biology* 121 (2): 526–41.

———. 1997. “Cell Lineage and Timing of Fate Restriction, Determination and Gene Expression in Ascidian Embryos.” *Seminars in Cell & Developmental Biology* 8 (4): 359–65.

Pachitariu, Marius, Michael Rariden, and Carsen Stringer. 2025. “Cellpose-SAM: Superhuman Generalization for Cellular Segmentation.” *BioRxiv*. https://doi.org/10.1101/2025.04.28.651001.

Palla, Giovanni, Hannah Spitzer, Michal Klein, David Fischer, Anna Christina Schaar, Louis Benedikt Kuemmerle, Sergei Rybakov, et al. 2022. "Squidpy: A Scalable Framework for Spatial Omics Analysis." *Nature Methods* 19 (2): 171–78.

Park, Jihye, Jung Min Lim, Inkyung Jung, Seok-Jae Heo, Jinman Park, Yoojin Chang, Hui Kwon Kim, et al. 2021. "Recording of Elapsed Time and Temporal Information about Biological Events Using Cas9." *Cell* 184 (4): 1047-1063.e23.

Price, Christopher J., Dylan Stavish, Paul J. Gokhale, Ben A. Stevenson, Samantha Sargeant, Joanne Lacey, Tristan A. Rodriguez, and Ivana Barbaric. 2021. "Genetically Variant Human Pluripotent Stem Cells Selectively Eliminate Wild-Type Counterparts through YAP-Mediated Cell Competition." *Developmental Cell* 56 (17): 2455-2470.e10.

Qiu, Dongbo, Shoudong Ye, Bryan Ruiz, Xingliang Zhou, Dahai Liu, Qi Zhang, and Qi-Long Ying. 2015. "Klf2 and Tfcp2l1, Two Wnt/β-Catenin Targets, Act Synergistically to Induce and Maintain Naive Pluripotency." *Stem Cell Reports* 5 (3): 314–22.

Quinn, Jeffrey J., Matthew G. Jones, Ross A. Okimoto, Shigeki Nanjo, Michelle M. Chan, Nir Yosef, Trever G. Bivona, and Jonathan S. Weissman. 2021. "Single-Cell Lineages Reveal the Rates, Routes, and Drivers of Metastasis in Cancer Xenografts." *Science* 371 (6532). https://doi.org/10.1126/science.abc1944.

R Core Team, R., and Others. 2013. "R: A Language and Environment for Statistical Computing." https://apps.dtic.mil/sti/citations/AD1039033.

Raj, Arjun, Patrick van den Bogaard, Scott A. Rifkin, Alexander van Oudenaarden, and Sanjay Tyagi. 2008. "Imaging Individual MRNA Molecules Using Multiple Singly Labeled Probes." *Nature Methods* 5 (10): 877–79.

Raj, Bushra, Daniel E. Wagner, Aaron McKenna, Shristi Pandey, Allon M. Klein, Jay Shendure, James A. Gagnon, and Alexander F. Schier. 2018. "Simultaneous Single-Cell Profiling of Lineages and Cell Types in the Vertebrate Brain." *Nature Biotechnology* 36 (5): 442–50.

Riemann, Franz, V. V. Malakhov, G. V. Bentz, and W. D. Hope. 1994. "Nematodes: Structure, Development, Classification, and Phylogeny." *Transactions of the American Microscopical Society* 113 (2): 199.

Rivera-Pérez, Jaime A., and Anna-Katerina Hadjantonakis. 2014. "The Dynamics of Morphogenesis in the Early Mouse Embryo." *Cold Spring Harbor Perspectives in Biology* 7 (11): a015867.

Rodriguez-Terrones, Diego, Xavier Gaume, Takashi Ishiuchi, Amélie Weiss, Arnaud Kopp, Kai Kruse, Audrey Penning, Juan M. Vaquerizas, Laurent Brino, and Maria-Elena Torres-Padilla. 2018. "A Molecular Roadmap for the Emergence of Early-Embryonic-like Cells in Culture." *Nature Genetics* 50 (1): 106–19.

Salvador-Martínez, Irepan, Marco Grillo, Michalis Averof, and Maximilian J. Telford. 2019. "Is It Possible to Reconstruct an Accurate Cell Lineage Using CRISPR Recorders?" *ELife* 8 (January). https://doi.org/10.7554/eLife.40292.

Schierenberg, Einhard. 2005. "Unusual Cleavage and Gastrulation in a Freshwater Nematode: Developmental and Phylogenetic Implications." *Development Genes and Evolution* 215 (2): 103–8.

———. 2006. "Embryological Variation during Nematode Development." *WormBook: The Online Review of C. Elegans Biology*, January, 1–13.

Schliep, Klaus Peter. 2011. "Phangorn: Phylogenetic Analysis in R." *Bioinformatics* 27 (4): 592–93.

Schrode, Nadine, Néstor Saiz, Stefano Di Talia, and Anna-Katerina Hadjantonakis. 2014. "GATA6 Levels Modulate Primitive Endoderm Cell Fate Choice and Timing in the Mouse Blastocyst." *Developmental Cell* 29 (4): 454–67.

Scrucca, Luca, Chris Fraley, T. Brendan Murphy, and Adrian E. Raftery. 2023. *Model-Based Clustering, Classification, and Density Estimation Using Mclust in R*. Chapman & Hall/CRC The R Series. Philadelphia, PA: Chapman & Hall/CRC.

Seidel, Sophie, and Tanja Stadler. 2022. "TiDeTree: A Bayesian Phylogenetic Framework to Estimate Single-Cell Trees and Population Dynamic Parameters from Genetic Lineage Tracing Data." *Proceedings. Biological Sciences / The Royal Society* 289 (1986): 20221844.

Shah, Sheel, Yodai Takei, Wen Zhou, Eric Lubeck, Jina Yun, Chee-Huat Linus Eng, Noushin Koulena, et al. 2018. "Dynamics and Spatial Genomics of the Nascent Transcriptome by Intron SeqFISH." *Cell* 174 (2): 363-376.e16.

Sim, Ye-Ji, Min-Seong Kim, Abeer Nayfeh, Ye-Jin Yun, Su-Jin Kim, Kyung-Tae Park, Chang-Hoon Kim, and Kye-Seong Kim. 2017. "2i Maintains a Naive Ground State in ESCs through Two Distinct Epigenetic Mechanisms." *Stem Cell Reports* 8 (5): 1312–28.

Simeonov, Kamen P., China N. Byrns, Megan L. Clark, Robert J. Norgard, Beth Martin, Ben Z. Stanger, Jay Shendure, Aaron McKenna, and Christopher J. Lengner. 2021. "Single-Cell Lineage Tracing of Metastatic Cancer Reveals Selection of Hybrid EMT States." *Cancer Cell* 39 (8): 1150-1162.e9.

Simon, Claire S., Anna-Katerina Hadjantonakis, and Christian Schröter. 2018. "Making Lineage Decisions with Biological Noise: Lessons from the Early Mouse Embryo." *Wiley Interdisciplinary Reviews. Developmental Biology* 7 (4): e319.

Singer, Zakary S., John Yong, Julia Tischler, Jamie A. Hackett, Alphan Altinok, M. Azim Surani, Long Cai, and Michael B. Elowitz. 2014. "Dynamic Heterogeneity and DNA Methylation in Embryonic Stem Cells." *Molecular Cell* 55 (2): 319–31.

Smith, Austin. 2017. "Formative Pluripotency: The Executive Phase in a Developmental Continuum." *Development (Cambridge, England)* 144 (3): 365–73.

Smith, Tom, Andreas Heger, and Ian Sudbery. 2017. "UMI-Tools: Modeling Sequencing Errors in Unique Molecular Identifiers to Improve Quantification Accuracy." *Genome Research* 27 (3): 491–99.

Sokal, R. R., and C. D. Michener. n.d. "A Statistical Method for Evaluating Systematic Relationships." *Multivariate Statistical Methods, among-Groups*.

Sommer, Ralf J. 2005. "Evolution of Development in Nematodes Related to C. Elegans." *WormBook: The Online Review of C. Elegans Biology*, December, 1–17.

Sović, Ivan, Mile Šikić, Andreas Wilm, Shannon Nicole Fenlon, Swaine Chen, and Niranjan Nagarajan. 2016. "Fast and Sensitive Mapping of Nanopore Sequencing Reads with GraphMap." *Nature Communications* 7 (1): 11307.

Spanjaard, Bastiaan, Bo Hu, Nina Mitic, Pedro Olivares-Chauvet, Sharan Janjuha, Nikolay Ninov, and Jan Philipp Junker. 2018. "Simultaneous Lineage Tracing and Cell-Type Identification Using CRISPR-Cas9-Induced Genetic Scars." *Nature Biotechnology* 36 (5): 469–73.

Sternberg, P. W., and M. A. Félix. 1997. "Evolution of Cell Lineage." *Current Opinion in Genetics & Development* 7 (4): 543–50.

Sternberg, P. W., and H. R. Horvitz. 1981. "Gonadal Cell Lineages of the Nematode Panagrellus Redivivus and Implications for Evolution by the Modification of Cell Lineage." *Developmental Biology* 88 (1): 147–66.

———. 1982. “Postembryonic Nongonadal Cell Lineages of the Nematode Panagrellus Redivivus: Description and Comparison with Those of Caenorhabditis Elegans.” *Developmental Biology* 93 (1): 181–205.

Strogatz, Steven H. 2024. *Nonlinear Dynamics and Chaos*. 3rd ed. Philadelphia, PA: Chapman & Hall/CRC.

Sugino, Ken, Jorge Garcia-Marques, Isabel Espinosa-Medina, and Tzumin Lee. 2019. “Theoretical Modeling on CRISPR-Coded Cell Lineages: Efficient Encoding and Optimal Reconstruction.” *BioRxiv*. bioRxiv. https://doi.org/10.1101/538488.

Sulston, J. E., and H. R. Horvitz. 1977. “Post-Embryonic Cell Lineages of the Nematode, Caenorhabditis Elegans.” *Developmental Biology* 56 (1): 110–56.

Takei, Yodai, Jina Yun, Shiwei Zheng, Noah Ollikainen, Nico Pierson, Jonathan White, Sheel Shah, et al. 2021. “Integrated Spatial Genomics Reveals Global Architecture of Single Nuclei.” *Nature* 590 (7845): 344–50.

Tam, P. P., and R. R. Behringer. 1997. “Mouse Gastrulation: The Formation of a Mammalian Body Plan.” *Mechanisms of Development* 68 (1–2): 3–25.

Tang, Weixin, and David R. Liu. 2018. “Rewritable Multi-Event Analog Recording in Bacterial and Mammalian Cells.” *Science (New York, N.Y.)* 360 (6385): eaap8992.

Tran, Martin, Amjad Askary, and Michael B. Elowitz. 2024. “Lineage Motifs as Developmental Modules for Control of Cell Type Proportions.” *Developmental Cell* 59 (6): 812-826.e3.

Troyanovsky, Boris, Vira Bitko, Viktor Pastukh, Brian Fouty, and Victor Solodushko. 2016. “The Functionality of Minimal PiggyBac Transposons in Mammalian Cells.” *Molecular Therapy. Nucleic Acids* 5 (10): e369.

Verkuijl, Sebald A. N., and Marianne G. Rots. 2019. “The Influence of Eukaryotic Chromatin State on CRISPR–Cas9 Editing Efficiencies.” *Current Opinion in Biotechnology* 55 (February): 68–73.

Waddington, C. H. 1956. *Principles of Embryology*. St Leonards, NSW, Australia: Allen & Unwin.

———. 2014. *The Strategy of the Genes*. Routledge.

Wang, Z. Q., F. Kiefer, P. Urbánek, and E. F. Wagner. 1997. “Generation of Completely Embryonic Stem Cell-Derived Mutant Mice Using Tetraploid Blastocyst Injection.” *Mechanisms of Development* 62 (2): 137–45.

Wen, Duancheng, Nestor Saiz, Zev Rosenwaks, Anna-Katerina Hadjantonakis, and Shahin Rafii. 2014. “Completely ES Cell-Derived Mice Produced by Tetraploid Complementation Using Inner Cell Mass (ICM) Deficient Blastocysts.” *PloS One* 9 (4): e94730.

Wu, Yongyan, Zhiying Ai, Kezhen Yao, Lixia Cao, Juan Du, Xiaoyan Shi, Zekun Guo, and Yong Zhang. 2013. “CHIR99021 Promotes Self-Renewal of Mouse Embryonic Stem Cells by Modulation of Protein-Encoding Gene and Long Intergenic Non-Coding RNA Expression.” *Experimental Cell Research* 319 (17): 2684–99.

Xu, Jin, Kevin Nuno, Ulrike M. Litzenburger, Yanyan Qi, M. Ryan Corces, Ravindra Majeti, and Howard Y. Chang. 2019. “Single-Cell Lineage Tracing by Endogenous Mutations Enriched in Transposase Accessible Mitochondrial DNA.” *ELife* 8 (April). https://doi.org/10.7554/eLife.45105.

Yang, Dian, Matthew G. Jones, Santiago Naranjo, William M. Rideout 3rd, Kyung Hoi Joseph Min, Raymond Ho, Wei Wu, et al. 2022. “Lineage Tracing Reveals the Phylodynamics, Plasticity, and Paths of Tumor Evolution.” *Cell* 185 (11): 1905-1923.e25.

Yang, Fan, Tomas Babak, Jay Shendure, and Christine M. Disteche. 2010. "Global Survey of Escape from X Inactivation by RNA-Sequencing in Mouse." *Genome Research* 20 (5): 614–22.

Yang, Z. 1994. "Maximum Likelihood Phylogenetic Estimation from DNA Sequences with Variable Rates over Sites: Approximate Methods." *Journal of Molecular Evolution* 39 (3): 306–14.

Yarrington, Robert M., Surbhi Verma, Shaina Schwartz, Jonathan K. Trautman, and Dana Carroll. 2018. "Nucleosomes Inhibit Target Cleavage by CRISPR-Cas9 in Vivo." *Proceedings of the National Academy of Sciences of the United States of America* 115 (38): 9351–58.

Ye, Shoudong, Ping Li, Chang Tong, and Qi-Long Ying. 2013. "Embryonic Stem Cell Self-Renewal Pathways Converge on the Transcription Factor Tfcp2l1." *The EMBO Journal* 32 (19): 2548–60.

Ying, Qi-Long, Jason Wray, Jennifer Nichols, Laura Batlle-Morera, Bradley Doble, James Woodgett, Philip Cohen, and Austin Smith. 2008. "The Ground State of Embryonic Stem Cell Self-Renewal." *Nature* 453 (7194): 519–23.

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

Zhao, Zhongying, Thomas J. Boyle, Zhirong Bao, John I. Murray, Barbara Mericle, and Robert H. Waterston. 2008. "Comparative Analysis of Embryonic Cell Lineage between Caenorhabditis Briggsae and Caenorhabditis Elegans." *Developmental Biology* 314 (1): 93–99.

Zhu, Ronghui, Jesus M. Del Rio-Salgado, Jordi Garcia-Ojalvo, and Michael B. Elowitz. 2022. "Synthetic Multistability in Mammalian Cells." *Science (New York, N.Y.)* 375 (6578): eabg9765.