Imaging cell lineage with a synthetic digital recording system

Thesis by

KeHuan Kuo Edmonds

In Partial Fulfillment of the Requirements for

the degree of

Ph.D. in Biology

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2022

(Defended October 22, 2021)

KeHuan Kuo Edmonds
ORCID: 0000-0002-7317-2669

ACKNOWLEDGEMENTS

Up to this point, graduate school has been the biggest adventure of my lifetime. Like the fables of old, it was full of laughter and joy, sweat and tears. Not quite as much monster slaying, I admit, but definitely brimming with colorful challenges that pushed me to grow and expand my horizons. Looking back, I can barely recognize the timid student that stood next to the start sign of this journey. But if she asked, I would let her know, "Don't be scared to embark, for you will never be alone. Every step along the way you are blessed with friends that push and support you. Family that loves and believes in you. And mentors that care for and guide you. They make it possible. Go, and pursue the dream that you didn't know you had."

I am extremely fortunate to have too many individuals to thank, so allow me to begin by expressing my gratitude to all who crossed paths with me during my scientific journey. From UCLA (my alma mater) to Caltech and beyond, whether you are listed here or not, please know that I am honored to be a part of your community.

I would like to thank my entire family for cheering me on since day one of graduate school. I am especially grateful for my husband, Dennis, who — with no exaggeration — was the pillar that supported me as I traversed through the highs and lows of my Ph.D. pursuit, and the proud owner of the pair of shoulders that always lifted me up when I am down. I would also like to thank my mom, Chialam, who is a Caltech-trained biologist herself. She was my inspiration since childhood and the origin of my interest in science. It was through her hard work, support, and love that I could get to where I am today.

I would like to say thank you to my friends. Especially Mark, who doubled as a teacher at the start of my graduate career. I would also like to thank Christina, Darvin, Ding, Alejandro, Amjad, Shinae, Maria, Duncan, Martin, Heidi, Ron, Sandy, Lacra, and many, many more, for making cherished memories with me throughout the years. It is probably no coincidence that the list also overlaps significantly with the Elowitz lab roster. During my 7 years in the lab, its members have become almost like a second family. So thank you, to all its members, especially the MEMOIR group and the staff, Leah, James, and Jo, for making this experience possible.

I would also like to thank members of the first lab I joined in Caltech — the Patterson lab. Paul believed in me when I didn't believe in myself. And although my time in that lab was cut short by his unfortunate passing in 2014, my experience there has forever changed me for the better, and I've made lifelong friends like Wei-Li, Jan, Laura, and many more.

I also want to thank my committee members: Bruce, for all the fun scientific discussions throughout the years, especially those that happen over coffee and lunch; Paul, for all that you've taught me, first as the professor of my genetics course, and later as a member of my committee. Long, for your invaluable expertise, suggestions, and help throughout the MEMOIR projects; and Carlos, for your support throughout intMEMOIR, and for all the kindness and knowledge you've shared in both science and life; many of your words will continue to guide me for the future years to come.

Finally, I would like to express my deepest gratitude for my mentor and role model, Michael. Thank you for giving me the opportunity to learn from you. You have always encouraged and supported my curiosity. You've taught me scientific rigor, communication, and so much more. Your brilliance, creativity, and infectious passion for science are the inspiration and fuel for my desire to grow as a scientist.

Thank you.

ABSTRACT

In multicellular organisms, the lineage history and spatial organization of cells both play pivotal roles in cell fate determination during development, homeostasis, and disease. Investigating lineage relationships alongside cell state and space would provide a fundamental understanding of these biological processes. Current lineage tracking approaches rely on the progressive accumulation of either naturally-occurring somatic mutations or experimentally introduced markers. In most cases, these marks are then read out by sequencing, discarding the spatial information of the cells. To address this vital gap in our toolkit, we developed a new synthetic lineage tracking system that allows us to image single-cell lineage history. This system, termed integrase-editable memory by engineered mutagenesis with optical *in situ* readout (intMEMOIR), uses serine integrases to stochastically and irreversibly edit a synthetic memory array, generating up to 59,049 different outcomes that can be unambiguously distinguished by fluorescence *in situ* hybridization (FISH). We evaluated the reconstruction accuracy of our system in mouse embryonic stem (mES) cells and disentangled the relative contribution of lineage and space to cell fate determination in *Drosophila* brain development, establishing the foundation for an expandable synthetic microscopy-readable system. In this thesis, Chapter 1 introduces the importance of cell lineage and spatial organization to cell fate determination, and includes a brief history of the existing technologies of the lineage tracking field. Chapter 2 describes our characterization and demonstration of the intMEMOIR system. Finally, Chapter 3 discusses design principles for robust, serine-integrase-based recording systems and suggests future directions for intMEMOIR.

PUBLISHED CONTENT AND CONTRIBUTIONS

K.-H. K. Chow, M. W. Budde, A. A. Granados, M. Cabrera, S. Yoon, S. Cho, T.-H. Huang, N. Koulena, K. L. Frieda, L. Cai, C. Lois, M. B. Elowitz, Imaging cell lineage with a synthetic digital recording system. Science. 372 (2021), doi:10.1126/science.abb3099.

K.K.C. participated in the project conception, experiments, data analysis, and writing of the manuscript.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

1.1 Cell fate determination is a central topic in biology

Developed from a single zygote, each human contains trillions of cells with hundreds of cell types that perform specialized functions despite sharing mostly identical DNA (*1*, *2*). Cells need to transition from one cell state to another to acquire unique functions, a necessary developmental process that continues to play a central role in homeostasis, regeneration, and even disease progression throughout a person's lifetime (*3*, *4*). Uncovering the underlying mechanism behind these transitions would give biologists a fundamental understanding of both normal and abnormal development and lay the groundwork for developing systems to imitate and manipulate these processes in basic research and medicine (*5*, *6*).

Cell fate determination is a complex and context-dependent process that, in many cases, results from the integration of multiple intrinsic and extrinsic signals (*7*). Examples of factors known to affect cell fate decisions include cell lineage history, spatial organization, signaling pathways, cell cycle, and mechanical signals (*5*, *8–11*). To complicate matters further, the factors themselves are often interdependent on one another. Thus, to disentangle the underlying mechanism of cell fate determination, we need tools that will enable us to simultaneously analyze cell states and their relevant determinants.

We wish to contribute to this effort by focusing on two critical factors that contribute to cell fate determination: cell lineage and spatial organization. There is a current need for a robust and versatile recording system that could simultaneously capture single cell lineage, cell state, and spatial organization within the same tissue. To fulfill that need, we designed a synthetic recording system termed integrase-editable memory by engineered mutagenesis with optical *in situ* readout (intMEMOIR).

1.2 Cell lineage impacts cell fate determination

In 1855, Rudolf Virchow described the third cell theory — *omnis cellula e cellula* — all cells come from cells (*12*). This one sentence summarily points out that cell proliferation is involved in almost all biological systems of interest.

Cell state transitions can often occur concurrently with, or even depend directly on, cell divisions (*13*). Famous examples include asymmetrical division during early embryo development (*14*) and stem cell differentiation (*15*), as well as the common phenomenon of cell cycle exit in terminally differentiated cells (*16*). Cell fate commitment also occurs throughout development and homeostasis, where fate decisions made by progenitor cells restrict the possible fates of their progeny (*14*, *17*). These examples illustrate that cell lineage is fundamentally connected to cell fate decisions. Thus, a lineage tree record of all cell divisions involved in a biological process is an important variable to understanding fate determination. Furthermore, it is also an exceptional framework on which we could map all other determinants of cell fate, producing a "decorated lineage tree" that allows us to analyze the dynamics of cell state transition across time and cell divisions (Fig. 1.1) (*5*, *18*).

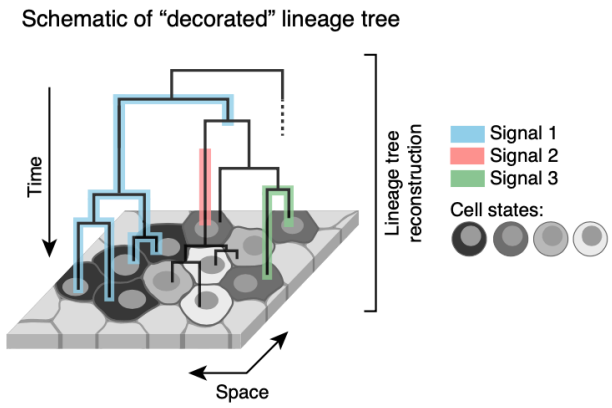

Fig. 1.1. Schematic of a "decorated" lineage tree.
A lineage tree can serve as the framework for mapping all determinants of cell fate. Depicted here are cell lineage, cell state, spatial organization, and signaling event history.

1.3 The lineage tracking field has an extensive history

Lineage tracking is a longstanding goal in biology. The field started during the 19th century with the study of developing invertebrate embryos (*19*). Decades later, the landmark work from Sulston mapped the complete development of the *Caenorhabditis elegans* embryo through direct observation

under the microscope, leading to numerous breakthrough insights on processes such as programmed cell death (*20*). However, the development of more complex organisms is often not tractable for direct observation. To address these challenges, researchers turned to other techniques such as transplantation and dye labeling for clonal tracing (*19*). The methods continued to evolve over the years, and recent advances in readout and genome editing technologies have led to another wave of new developments.

With new single-cell sequencing technology, naturally accumulating somatic mutations have become an invaluable source of lineage information (*21*). As cells divide, their otherwise identical DNA can gradually accumulate mutations that are inherited by their progenies. Shared mutations between the cells can then be used to reconstruct a lineage tree, similarly to how one would reconstruct a phylogenetic tree (*22*). Landmark experiments include the use of copy number variations (CNV) to infer tumor evolution (*23*), single-nucleotide variants (SNV) to reconstruct neuron lineage in human brains (*24*), and mutations in mitochondrial DNA to reconstruct cell lineage in human samples (*25*). More recently, a set of four studies also demonstrated lineage tracing in human development and homeostasis (*26–30*). Techniques such as these are often collectively referred to as "retrospective lineage tracking" (*3*, *22*), and can be used to directly study humans and other organisms in which genetic manipulation cannot be performed. However, they usually result in the loss of spatial information due to the use of sequencing for readout. Further, since somatic mutations are relatively rare and can occur anywhere within the genome, the single cell genomes often need to be amplified prior to readout, which can result in artifacts (*22*).

Inspired by the use of somatic mutations and enabled by the advances in genome editing technologies, others created systems to progressively generate mutations at engineered loci within cells to record lineage, collectively called "prospective lineage tracking" (*22*). Unlike somatic mutations, the location, rate, and types of edits are controlled, which simplifies the information recovery and allows tuning of the editing rates. Some methods use recombinases to generate DNA inversions and deletions as the lineage marker (*31*). Many others make use of CRISPR-Cas9 to target either synthetic recording units, with examples such as the genome editing of synthetic target arrays for lineage tracing (GESTALT) (*32*) and others (*33–35*), or to target their own gRNA loci with homing guides (*36*, *37*). A number of these techniques have been implemented to study normal and cancer development in zebrafish and mice (*31*, *32*, *35*, *36*, *38*, *39*). Since most rely on single-cell

RNA sequencing (scRNA-seq) to recover their edits, they are compatible with simultaneous analysis of cell states through the transcriptome. Like the retrospective lineage techniques, however, their readout also disrupts the spatial organization of the cells and is subject to the inherent limitations of sequencing technologies, such as sample dropout (*40*).

Although they do not reconstruct cell divisions in time, there is also a related class of techniques that use scRNA-seq datasets to order cells on state transition trajectories (*5*, *41*, *42*). These methods have been used to map the differentiation process in a number of different model organisms (*43–46*), and while the inferred trajectories do not always correspond to the mitotic lineage of the cells, combining the two could provide important insight into the interaction between lineage and cell fate (*5*, *44*).

Finally, it is worth noting that there is an ongoing effort in the scientific community to develop sequencing technologies that incorporate spatial information. This exciting branch of tool development may enable the *in situ* readouts of these mutations with single-cell resolution in the future (*8*). However, for the time being, recovering spatial information from these lineage tracking technologies remains a fundamental challenge (*3*).

1.4 Spatial organization impacts cell fate determination

Cells within multicellular organisms exist in the context of each other: organized and coordinated for development and homeostasis (*8*). Their spatial location plays a major role in dictating the extrinsic signals they receive from their environment as well as the functions that they perform (*47–49*). Examples range from morphogen gradients enabling essential pattern formation during development (*50*, *51*) to nutrient gradients creating spatially organized tumors (*52*). Thus, knowing the physical location of the cells is important for understanding both the cause and consequence of cell fate determination.

1.5 Some lineage tracking techniques recover single-cell spatial information

Given the importance of spatial organization, a number of lineage tracking technologies also allow retrieval of the spatial information of the recorded cells. These technologies can be generally divided into three categories. First, advances in microscopy techniques and computational analysis tools have

enabled researchers to track single cells and their divisions through live imaging. Recent examples include lineage tracking during the development of the *Parhyale hawaiensis* limb (*53*), mouse development at the pre-implantation (*54*) and post-implantation (*55*) stages, and cerebral organoid development (*56*). For the duration of imaging, these powerful approaches provide insight into the dynamics of development on a level that is not yet achievable through other methods. Compared to the other technologies, however, the size of their data and the experiments' hardware and software requirements often hinder their dissemination and usability (*57*). Most importantly, the fundamental challenge since Sulston's experiments remains: the sample needs to be accessible and relatively unperturbed by the imaging conditions needed to resolve and track living single cells. This challenge can only be partially overcome with highly specialized equipment, such as that built to sustain mouse embryo development *ex utero (55)*.

A second set of lineage tracking technologies uses DNA recombination to label cells with random combinations of fluorescent proteins, which are inherited by their progenies to enable clonal tracing. Early predecessors to these techniques include mosaic analysis with a repressible cell marker (MARCM) and its mouse version, mosaic analysis with double markers (MADM), which were used to study brain development in their respective model organisms (*58–60*). These techniques were limited in the number of clones they could simultaneously follow. In 2007, Brainbow revolutionized the field by using the Cre-recombinase to invert and delete a cassette of up to four fluorescent proteins flanked by loxP sites, resulting in the random expression of one. By introducing multiple copies of the cassette into the same cell, one could now stochastically generate a significantly larger number of colors through the random combination of fluorescent proteins expressing from each array: a heritable marker that can be used for clonal tracing (*61*). Since its initial demonstration in the mouse brain, the Brainbow technology has been improved and adapted by multiple groups, resulting in higher multiplexibility and use in numerous other model organisms (*62–64*). The system's diverse color combinations drastically increased the number of clones that could be simultaneously analyzed in parallel, but the number of colors that could be distinguished in practice (approximately 100) still remains relatively low (*65*). Further, because the edit outcomes cannot be assigned to the individual cassettes, Brainbow is not suitable for lineage reconstruction across multiple generations. In addition, because the system uses multiple fluorescent channels, it is

challenging to perform subsequent gene expression analysis with *in situ* methods such as fluorescence *in situ* hybridization (FISH) or fluorescent reporters (*22*).

The third category contains the genetic lineage recording methods that generate progressive mutations, like those described in Chapter 1.3, that are compatible with *in situ* readouts. Very few technologies fall under this category. One of them is cell lineage access driven by an edition sequence (CLADES), which couples a gRNA cascade to fluorescent reporters in such a way that Cas9 editings would progressively change the expressed fluorescent protein in the cell (*66*). We developed the other system, termed memory by engineered mutagenesis with optical in situ readout (MEMOIR), with the goal to create a recording system that can simultaneously analyze cell lineage, state, and space (*67*). The system uses CRISPR-Cas9 to progressively and stochastically delete genomically distributed, barcoded recording units called scratchpads. We then used the accumulated deletions to reconstruct cell lineage based on the shared edit patterns of each cell. Each barcoded scratchpad was transcribed, and the edit states were read out by sequential rounds of single-molecule FISH (smFISH) (*68*). Thus, the system can be used alongside smFISH readout of endogenous genes to decipher cell state and, most importantly, ensures that we retain the spatial information of the cells.

MEMOIR is a proof-of-principle that demonstrates the possibility to extract recorded multigenerational lineage information *in situ*. However, its reconstruction depth and accuracy are limited, and its designs are difficult to implement in a germline transmissible manner *in vivo*. Thus, we sought to create new systems that retain the strengths of the original MEMOIR while addressing its weaknesses. With that goal in mind, the next generation of technologies, Zombie is optical measurement of barcodes by *in situ* expression (Zombie) (*69*) and intMEMOIR, were developed (*70*). The prior uses *in vitro* transcription to distinguish DNA edits as small as a single base pair, and the latter uses serine integrase to progressively edit memory units, with the ability to generate up to 59,049 distinct outcomes that can be unambiguously distinguished using FISH. We implemented intMEMOIR in mouse embryonic stem cells and *Drosophila melanogaster* lines, and the results are more extensively discussed in Chapter 2.

1.6 Summary

Uncovering the mechanism behind cell fate determination will advance our understanding and ability to manipulate development, homeostasis, and disease. Since fate decisions are often influenced by many intrinsic and extrinsic factors, simultaneous readout of cell state alongside these influences will enable us to disentangle their relative contributions.

Cell lineage and spatial organization are two dominant factors in cell fate determination. Most modern lineage tracking technologies are able to analyze cell state alongside lineage; however, those that also incorporate spatial information are relatively limited in the number of clones they can discriminate and their reconstruction depth and accuracy. To address this important gap in our toolkit, we developed intMEMOIR: a robust recording system that enables us to simultaneously analyze cell lineage, cell state, and spatial organization within the same tissue. Chapter 2 describes the design of the system, characterizes its performance *in vitro*, and illustrates its application *in vivo*. This system can also serve as a foundation for future recorders to capture other determinants of cell fate, such as the signaling history of the cell. To facilitate that goal, Chapter 3 discusses important design principles of intMEMOIR-based systems alongside suggestions for future directions.

1.7 References

1. E. Bianconi, A. Piovesan, F. Facchin, A. Beraudi, R. Casadei, F. Frabetti, L. Vitale, M. C. Pelleri, S. Tassani, F. Piva, S. Perez-Amodio, P. Strippoli, S. Canaider, An estimation of the number of cells in the human body. *Ann. Hum. Biol.* **40**, 463–471 (2013).

2. A. Abyzov, F. M. Vaccarino, Cell Lineage Tracing and Cellular Diversity in Humans. *Annu. Rev. Genomics Hum. Genet.* **21**, 101–116 (2020).

3. C. S. Baron, A. van Oudenaarden, Unravelling cellular relationships during development and regeneration using genetic lineage tracing. *Nat. Rev. Mol. Cell Biol.* **20**, 753–765 (2019).

4. P. B. Gupta, C. M. Fillmore, G. Jiang, S. D. Shapira, K. Tao, C. Kuperwasser, E. S. Lander, Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell*. **146**, 633–644 (2011).

5. D. E. Wagner, A. M. Klein, Lineage tracing meets single-cell omics: opportunities and challenges. *Nat. Rev. Genet.* **21**, 410–427 (2020).

6. V. Tabar, L. Studer, Pluripotent stem cells in regenerative medicine: challenges and recent progress. *Nat. Rev. Genet.* **15**, 82–92 (2014).

7. C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J.

Livak, T. S. Mikkelsen, J. L. Rinn, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).

8. M. Asp, J. Bergenstråhle, J. Lundeberg, Spatially Resolved Transcriptomes-Next Generation Tools for Tissue Exploration. *Bioessays*. **42**, e1900221 (2020).

9. N. Perrimon, C. Pitsouli, B.-Z. Shilo, Signaling mechanisms controlling cell fate and embryonic patterning. *Cold Spring Harb. Perspect. Biol.* **4**, a005975 (2012).

10. S.-W. Gao, F. Liu, Novel insights into cell cycle regulation of cell fate determination. *J. Zhejiang Univ. Sci. B*. **20**, 467–475 (2019).

11. C. J. Chan, C.-P. Heisenberg, T. Hiiragi, Coordination of Morphogenesis and Cell-Fate Specification in Development. *Curr. Biol.* **27**, R1024–R1035 (2017).

12. D. Ribatti, An historical note on the cell theory. *Exp. Cell Res.* **364**, 1–4 (2018).

13. S. Hormoz, Z. S. Singer, J. M. Linton, Y. E. Antebi, B. I. Shraiman, M. B. Elowitz, Inferring Cell-State Transition Dynamics from Lineage Trees and Endpoint Single-Cell Measurements. *Cell Syst*. **3**, 419–433.e8 (2016).

14. S. J. Arnold, E. J. Robertson, Making a commitment: cell lineage allocation and axis patterning in the early mouse embryo. *Nat. Rev. Mol. Cell Biol.* **10**, 91–103 (2009).

15. J. A. Knoblich, Mechanisms of asymmetric stem cell division. *Cell*. **132**, 583–597 (2008).

16. A. Soufi, S. Dalton, Cycling through developmental decisions: how cell cycle dynamics control pluripotency, differentiation and reprogramming. *Development*. **143**, 4301–4311 (2016).

17. Y. Zhang, S. Gao, J. Xia, F. Liu, Hematopoietic Hierarchy – An Updated Roadmap. *Trends in Cell Biology*. **28** (2018), pp. 976–986.

18. V. E. Papaioannou, Concepts of Cell Lineage in Mammalian Embryos. *Curr. Top. Dev. Biol.* **117**, 185–197 (2016).

19. K. Kretzschmar, F. M. Watt, Lineage tracing. *Cell*. **148**, 33–45 (2012).

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

21. D. Frumkin, A. Wasserstrom, S. Kaplan, U. Feige, E. Shapiro, Genomic variability within an organism exposes its cell lineage tree. *PLoS Comput. Biol.* **1**, e50 (2005).

22. M. B. Woodworth, K. M. Girskis, C. A. Walsh, Building a lineage from single cells: genetic techniques for cell lineage tracking. *Nat. Rev. Genet.* **18**, 230–244 (2017).

23. N. Navin, J. Kendall, J. Troge, P. Andrews, L. Rodgers, J. McIndoo, K. Cook, A. Stepansky, D. Levy, D. Esposito, L. Muthuswamy, A. Krasnitz, W. R. McCombie, J. Hicks, M. Wigler, Tumour evolution inferred by single-cell sequencing. *Nature*. **472**, 90–94 (2011).

24. M. A. Lodato, M. B. Woodworth, S. Lee, G. D. Evrony, B. K. Mehta, A. Karger, S. Lee, T. W. Chittenden, A. M. D'Gama, X. Cai, L. J. Luquette, E. Lee, P. J. Park, C. A. Walsh, Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science*. **350**, 94–98 (2015).

25. L. S. Ludwig, C. A. Lareau, J. C. Ulirsch, E. Christian, C. Muus, L. H. Li, K. Pelka, W. Ge, Y. Oren, A. Brack, T. Law, C. Rodman, J. H. Chen, G. M. Boland, N. Hacohen, O. Rozenblatt-Rosen, M. J. Aryee, J. D. Buenrostro, A. Regev, V. G. Sankaran, Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. *Cell*. **176**, 1325–1339.e22 (2019).

26. K. Naxerova, Mutation fingerprints encode cellular histories. *Nature*. **597** (2021), pp. 334–336.

27. S. Park, N. M. Mali, R. Kim, J.-W. Choi, J. Lee, J. Lim, J. M. Park, J. W. Park, D. Kim, T. Kim, K. Yi, J. H. Choi, S. G. Kwon, J. H. Hong, J. Youk, Y. An, S. Y. Kim, S. A. Oh, Y. Kwon, D. Hong, M. Kim, D. S. Kim, J. Y. Park, J. W. Oh, Y. S. Ju, Clonal dynamics in early human embryogenesis inferred from somatic mutation. *Nature*. **597**, 393–397 (2021).

28. T. H. H. Coorens, L. Moore, P. S. Robinson, R. Sanghvi, J. Christopher, J. Hewinson, M. J. Przybilla, A. R. J. Lawson, M. Spencer Chapman, A. Cagan, T. R. W. Oliver, M. D. C. Neville, Y. Hooks, A. Noorani, T. J. Mitchell, R. C. Fitzgerald, P. J. Campbell, I. Martincorena, R. Rahbari, M. R. Stratton, Extensive phylogenies of human development inferred from somatic mutations. *Nature*. **597**, 387–392 (2021).

29. R. Li, L. Di, J. Li, W. Fan, Y. Liu, W. Guo, W. Liu, L. Liu, Q. Li, L. Chen, Y. Chen, C. Miao, H. Liu, Y. Wang, Y. Ma, D. Xu, D. Lin, Y. Huang, J. Wang, F. Bai, C. Wu, A body map of somatic mutagenesis in morphologically normal human tissues. *Nature*. **597**, 398–403 (2021).

30. L. Moore, A. Cagan, T. H. H. Coorens, M. D. C. Neville, R. Sanghvi, M. A. Sanders, T. R. W. Oliver, D. Leongamornlert, P. Ellis, A. Noorani, T. J. Mitchell, T. M. Butler, Y. Hooks, A. Y. Warren, M. Jorgensen, K. J. Dawson, A. Menzies, L. O'Neill, C. Latimer, M. Teng, R. van Boxtel, C. A. Iacobuzio-Donahue, I. Martincorena, R. Heer, P. J. Campbell, R. C. Fitzgerald, M. R. Stratton, R. Rahbari, The mutational landscape of human somatic and germline cells. *Nature*. **597**, 381–386 (2021).

31. W. Pei, T. B. Feyerabend, J. Rössler, X. Wang, D. Postrach, K. Busch, I. Rode, K. Klapproth, N. Dietlein, C. Quedenau, W. Chen, S. Sauer, S. Wolf, T. Höfer, H.-R. Rodewald, Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. *Nature*. **548**, 456–460 (2017).

32. A. McKenna, G. M. Findlay, J. A. Gagnon, M. S. Horwitz, A. F. Schier, J. Shendure, Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*. **353**, aaf7907 (2016).

33. A. Alemany, M. Florescu, C. S. Baron, J. Peterson-Maduro, A. van Oudenaarden, Whole-organism clone tracing using single-cell sequencing. *Nature*. **556**, 108–112 (2018).

34. B. Spanjaard, B. Hu, N. Mitic, P. Olivares-Chauvet, S. Janjuha, N. Ninov, J. P. Junker, Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic

scars. *Nat. Biotechnol.* **36**, 469–473 (2018).

35. M. M. Chan, Z. D. Smith, S. Grosswendt, H. Kretzmer, T. M. Norman, B. Adamson, M. Jost, J. J. Quinn, D. Yang, M. G. Jones, A. Khodaverdian, N. Yosef, A. Meissner, J. S. Weissman, Molecular recording of mammalian embryogenesis. *Nature*. **570**, 77–82 (2019).

36. R. Kalhor, K. Kalhor, L. Mejia, K. Leeper, A. Graveline, P. Mali, G. M. Church, Developmental barcoding of whole mouse via homing CRISPR. *Science*. **361** (2018), doi:10.1126/science.aat9804.

37. T. B. Loveless, J. H. Grotts, M. W. Schechter, E. Forouzmand, C. K. Carlson, B. S. Agahi, G. Liang, M. Ficht, B. Liu, X. Xie, C. C. Liu, Lineage tracing and analog recording in mammalian cells by single-site DNA writing. *Nat. Chem. Biol.* **17**, 739–747 (2021).

38. K. P. Simeonov, C. N. Byrns, M. L. Clark, R. J. Norgard, B. Martin, B. Z. Stanger, J. Shendure, A. McKenna, C. J. Lengner, Single-cell lineage tracing of metastatic cancer reveals selection of hybrid EMT states. *Cancer Cell*. **39**, 1150–1162.e9 (2021).

39. J. J. Quinn, M. G. Jones, R. A. Okimoto, S. Nanjo, M. M. Chan, N. Yosef, T. G. Bivona, J. S. Weissman, Single-cell lineages reveal the rates, routes, and drivers of metastasis in cancer xenografts, , doi:10.1101/2020.04.16.045245.

40. I. Espinosa-Medina, J. Garcia-Marques, C. Cepko, T. Lee, High-throughput dense reconstruction of cell lineages. *Open Biol.* **9**, 190229 (2019).

41. L. Kester, A. van Oudenaarden, Single-Cell Transcriptomics Meets Lineage Tracing. *Cell Stem Cell*. **23**, 166–179 (2018).

42. W. Saelens, R. Cannoodt, H. Todorov, Y. Saeys, A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* **37**, 547–554 (2019).

43. J. A. Briggs, C. Weinreb, D. E. Wagner, S. Megason, L. Peshkin, M. W. Kirschner, A. M. Klein, The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science*. **360** (2018), doi:10.1126/science.aar5780.

44. D. E. Wagner, C. Weinreb, Z. M. Collins, J. A. Briggs, S. G. Megason, A. M. Klein, Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*. **360**, 981–987 (2018).

45. J. A. Farrell, Y. Wang, S. J. Riesenfeld, K. Shekhar, A. Regev, A. F. Schier, Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* (2018) (available at https://science.sciencemag.org/content/360/6392/eaar3131.abstract).

46. E. Llorens-Bobadilla, J. M. Chell, P. Le Merre, Y. Wu, M. Zamboni, J. Bergenstråhle, M. Stenudd, E. Sopova, J. Lundeberg, O. Shupliakov, M. Carlén, J. Frisén, A latent lineage potential in resident neural stem cells enables spinal cord repair. *Science*. **370** (2020), doi:10.1126/science.abb8795.

47. C. M. Nelson, M. J. Bissell, Of extracellular matrix, scaffolds, and signaling: tissue architecture regulates development, homeostasis, and cancer. *Annu. Rev. Cell Dev. Biol.* **22**, 287–309 (2006).

48. M. Quante, T. C. Wang, Stem cells in gastroenterology and hepatology. *Nat. Rev. Gastroenterol. Hepatol.* **6**, 724–737 (2009).

49. H. Tekin, J. G. Sanchez, C. Landeros, K. Dubbin, R. Langer, A. Khademhosseini, Controlling spatial organization of multiple cell types in defined 3D geometries. *Adv. Mater.* **24**, 5543–7, 5542 (2012).

50. J. B. Gurdon, P. Y. Bourillot, Morphogen gradient interpretation. *Nature*. **413**, 797–803 (2001).

51. O. Wartlick, A. Kicheva, M. González-Gaitán, Morphogen gradient formation. *Cold Spring Harb. Perspect. Biol.* **1**, a001255 (2009).

52. C. Carmona-Fontaine, M. Deforet, L. Akkari, C. B. Thompson, J. A. Joyce, J. B. Xavier, Metabolic origins of spatial organization in the tumor microenvironment. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 2934–2939 (2017).

53. C. Wolff, J.-Y. Tinevez, T. Pietzsch, E. Stamataki, B. Harich, L. Guignard, S. Preibisch, S. Shorte, P. J. Keller, P. Tomancak, A. Pavlopoulos, Multi-view light-sheet imaging and tracking with the MaMuT software reveals the cell lineage of a direct developing arthropod limb. *Elife*. **7** (2018), doi:10.7554/eLife.34410.

54. M. Welling, M. A. Mohr, A. Ponti, L. R. Sabater, A. Boni, Y. K. Kawamura, P. Liberali, A. H. Peters, P. Pelczar, P. Pantazis, Primed Track, high-fidelity lineage tracing in mouse pre-implantation embryos using primed conversion of photoconvertible proteins. *eLife*. **8** (2019), , doi:10.7554/elife.44491.

55. K. McDole, L. Guignard, F. Amat, A. Berger, G. Malandain, L. A. Royer, S. C. Turaga, K. Branson, P. J. Keller, In Toto Imaging and Reconstruction of Post-Implantation Mouse Development at the Single-Cell Level. *Cell*. **175**, 859–876.e33 (2018).

56. Z. He, T. Gerber, A. Maynard, A. Jain, R. Petri, M. Santel, Lineage recording reveals dynamics of cerebral organoid regionalization. *bioRxiv* (2020) (available at https://www.biorxiv.org/content/10.1101/2020.06.19.162032v1.abstract).

57. S. Wolf, Y. Wan, K. McDole, Current approaches to fate mapping and lineage tracing using image data. *Development*. **148** (2021), doi:10.1242/dev.198994.

58. T. Lee, C. Winter, S. S. Marticke, A. Lee, L. Luo, Essential roles of Drosophila RhoA in the regulation of neuroblast proliferation and dendritic but not axonal morphogenesis. *Neuron*. **25**, 307–316 (2000).

59. H. Zong, J. S. Espinosa, H. H. Su, M. D. Muzumdar, L. Luo, Mosaic analysis with double markers in mice. *Cell*. **121**, 479–492 (2005).

60. L. Luo, Fly MARCM and mouse MADM: genetic methods of labeling and manipulating single neurons. *Brain Res. Rev.* **55**, 220–227 (2007).

61. J. Livet, T. A. Weissman, H. Kang, R. W. Draft, J. Lu, R. A. Bennis, J. R. Sanes, J. W. Lichtman, Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature*. **450**, 56–62 (2007).

62. Y. A. Pan, T. Freundlich, T. A. Weissman, D. Schoppik, X. C. Wang, S. Zimmerman, B. Ciruna, J. R. Sanes, J. W. Lichtman, A. F. Schier, Zebrabow: multispectral cell labeling for cell tracing and lineage analysis in zebrafish. *Development*. **140**, 2835–2846 (2013).

63. K. Loulier, R. Barry, P. Mahou, Y. Le Franc, W. Supatto, K. S. Matho, S. Ieng, S. Fouquet, E. Dupin, R. Benosman, A. Chédotal, E. Beaurepaire, X. Morin, J. Livet, Multiplex cell and lineage tracking with combinatorial labels. *Neuron*. **81**, 505–520 (2014).

64. O. Kanca, E. Caussinus, A. S. Denes, A. Percival-Smith, M. Affolter, Raeppli: a whole-tissue labeling tool for live imaging of Drosophila development. *Development*. **141**, 472–480 (2014).

65. T. A. Weissman, Y. A. Pan, Brainbow: new resources and emerging biological applications for multicolor genetic labeling and analysis. *Genetics*. **199**, 293–306 (2015).

66. J. Garcia-Marques, I. Espinosa-Medina, K.-Y. Ku, C.-P. Yang, M. Koyama, H.-H. Yu, T. Lee, A programmable sequence of reporters for lineage analysis. *Nat. Neurosci.* **23**, 1618–1628 (2020).

67. K. L. Frieda, J. M. Linton, S. Hormoz, J. Choi, K.-H. K. Chow, Z. S. Singer, M. W. Budde, M. B. Elowitz, L. Cai, Synthetic recording and in situ readout of lineage information in single cells. *Nature*. **541**, 107–111 (2017).

68. E. Lubeck, A. F. Coskun, T. Zhiyentayev, M. Ahmad, L. Cai, Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods*. **11** (2014), pp. 360–361.

69. A. Askary, L. Sanchez-Guardado, J. M. Linton, D. M. Chadly, M. W. Budde, L. Cai, C. Lois, M. B. Elowitz, In situ readout of DNA barcodes and single base edits facilitated by in vitro transcription. *Nat. Biotechnol.* **38**, 66–75 (2020).

70. K.-H. K. Chow, M. W. Budde, A. A. Granados, M. Cabrera, S. Yoon, S. Cho, T.-H. Huang, N. Koulena, K. L. Frieda, L. Cai, C. Lois, M. B. Elowitz, Imaging cell lineage with a synthetic digital recording system. *Science*. **372** (2021), doi:10.1126/science.abb3099.

Chapter 2

IMAGING CELL LINEAGE WITH A SYNTHETIC DIGITAL RECORDING SYSTEM

2.1 Abstract

During multicellular development, spatial position and lineage history play powerful roles in controlling cell fate decisions. Using a serine integrase-based recording system, we engineered cells to record lineage information in a format that can be read out in situ. The system, termed integrase-editable memory by engineered mutagenesis with optical in situ readout (intMEMOIR), allowed in situ reconstruction of lineage relationships in cultured mouse cells and flies. intMEMOIR uses an array of independent three-state genetic memory elements that can recombine stochastically and irreversibly, allowing up to 59,049 distinct digital states. It reconstructed lineage trees in stem cells and enabled simultaneous analysis of single cell clonal history, spatial position, and gene expression in Drosophila brain sections. These results establish a foundation for microscopy-readable lineage recording and analysis in diverse systems.

2.2 Introduction

Cell lineage plays pivotal roles in cell fate determination in development, homeostasis, and disease (1–5). The ability to visualize lineage relationships directly within their native tissue context provides insight into the roles of intrinsic and extrinsic factors in cell fate specification. Inspired by the recovery of lineage information from naturally occurring somatic mutations (2, 3, 5–12), engineered lineage recording systems actively generate stochastic, heritable mutations at defined genomic target sites, and then identify those edits in individual cells to reconstruct their lineage (3, 11, 13–23). Currently, most of these methods require readout by sequencing, which disrupts spatial organization. Another approach, memory by engineered mutagenesis with optical *in situ* readout (MEMOIR) uses single-molecule fluorescence *in situ* hybridization (smFISH) to allow readout by imaging. However, this method relies on genomically distributed deletion edits that do not permit extended recording and germline transmission (23). Thus, there is a need for a broadly useful, digital, image-readable recording system.

2.3 Results

2.3.1 Serine integrases enable a FISH-readable three-state memory element design

One mode of lineage reconstruction is clonal tracing, which identifies cells descended from a common ancestor from distinct, heritable labels (e.g. sequence barcodes). A more complete lineage reconstruction provides the tree, or pedigree, of multiple divisions through which cells are related (Fig. 2.1A). To access both regimes, a recording system should be able to produce and preserve as much molecular diversity as possible to maximize the number of distinguishable clones, and accumulate that diversity over multiple cell generations to enable tree reconstruction. In systems with two-state memory elements (bits), extended recording durations eventually edit all memory elements, producing a noninformative homogeneous end state, and effectively erasing recorded information. By contrast, three-state memory elements, or trits, that start in an initial state and irreversibly switch to one of two potential end states, provide additional information per element and preserve recorded information. As a result, the use of trits improves the accuracy of multi-generation reconstruction and the multiplexibility of clonal classification in simulations, and allows the system to function across a broader range of edit rates compared to bit-based memory (Fig. 2.1, B and C) (*24*).

Phage serine integrases provide an ideal basis for engineering trits. They mediate irreversible recombination between directional *attP* and *attB* target sites, deleting or inverting the intervening sequence depending on relative site orientation (*25–29*). Serine integrases such as Bxb1 do not rely on endogenous repair mechanisms to generate edits, and they function across species, including mammalian cells (*29–31*). To create a trit, we flanked a barcode sequence by an inverted pair of *attP* sites on one end and an *attB* site on the other such that Bxb1-mediated recombination produces either irreversible barcode deletion or inversion (Fig. 2.1D). A strong polymerase II (Pol II) promoter drives transcription of the trit, allowing *in situ* readout by FISH methods. Before editing, the promoter expresses the forward barcode, whereas after recombination, it expresses either no transcript (deletion) or the reverse complement barcode (inversion), enabling digital discrimination of trit states by FISH.

Concatenating multiple trits in a compact array and integrating it at a safe harbor locus (*32–34*) increase the amount of memory while facilitating germline transmissibility (*35*). Edit strategies that rely on double-stranded breaks and endogenous DNA repair machinery are prone to information loss in arrays through interunit deletions (*24*, *36*). By contrast, integrases allow recombinational isolation between distinct trits in the same array. Within each *att* site, a central dinucleotide confers both specificity and directionality of recombination (*31*, *37*) (Fig. 2.2A). In principle, 10 distinct dinucleotide variants can be used orthogonally (fig. S2.1), enabling 10 corresponding independent memory units in a single array, for a theoretical diversity of 3^{10} (59,049) states.

To validate the trit design in mouse embryonic stem (mES) cells, we constructed a prototype Bxb1 trit that expressed no fluorescent protein, citrine, or mCherry in the unedited, inverted, and deleted states, respectively, allowing rapid characterization by flow cytometry (*30*) (Fig. 2.2B). Recombination occurred more efficiently between matching compared to mismatching dinucleotides, indicating that distinct dinucleotides operate in a largely orthogonal manner, as intended (Fig. 2.2C). When the *attB* and inverted *attPs* all contained the same dinucleotide, such as GT, inversion and deletion both occurred efficiently (Fig. 2.2D). Further, *att* sites made from palindromic dinucleotides such as AT, which lack directionality, enable a simplified design in which a single *attP/B* pair mediates both inversion and deletion (Fig. 2.2E). A control construct in which one site is inverted (AT') performed similarly to the uninverted counterpart, permitting the use of palindromic sites in either orientation (Fig. 2.2E). This design could also be generalized to other integrases (fig. S2.2). Although deletion crosstalk did occur at low rates (Fig. 2.2C), consistent with observations using phiC31 integrase (*38*), these results indicate that dinucleotide variants permit orthogonal recombination, enabling construction of compact 10-unit trit arrays.

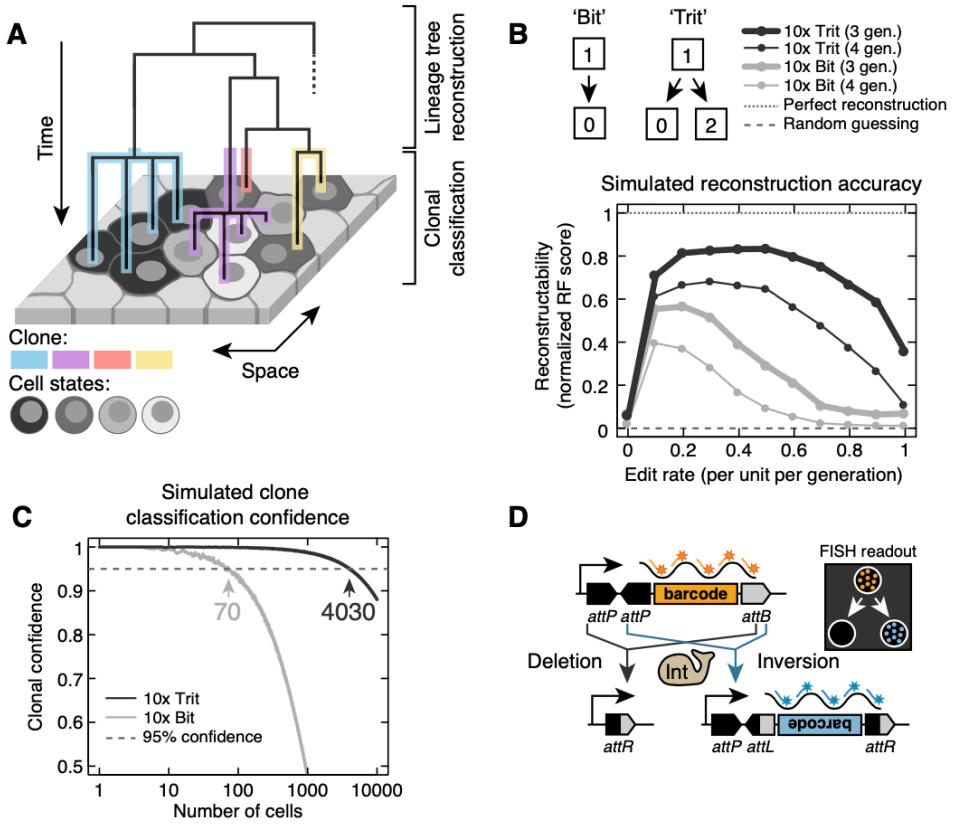

Fig. 2.1. Three-state memory elements (trits) enable *in situ* developmental lineage reconstruction.

(**A**) An ideal recording system connects the spatial information, gene expression, and lineage history of single cells (schematic). Lineage information comprises multi-generation reconstruction of cell division trees (B) as well as multiplexed clone classification (C). (**B** and **C**) Simulations of editing and reconstruction for systems with 10 irreversible recording units. For lineage tree reconstruction in (B), trit recording units improve tree reconstructability compared with two-state bits across a wide range of edit rates and retain information when edited to completion. Reconstructability is defined as the normalized Robinson-Foulds score obtained by comparing the reconstructed tree to ground truth simulated lineage. For clonal classification in (C), trits enable simultaneous tracing of a large number of clones in the same organism, potentially distinguishing up to two orders of magnitude more clones with >95% confidence than its bit counterpart (4030 and 70 clones, respectively). Clonal confidence is defined as the probability that two randomly selected cells with identical edit patterns are from the same clone. (**D**) Serine integrases enable trit designs compatible with FISH readout methods. Transcribed barcodes are flanked by two *attP*s and one *attB*. Recombination results in either an inverted or deleted barcode, which can be distinguished by fluorescent probes (colored lines and asterisks) directed against either strand.

2.3.2 Design and characterization of the intMEM1 recording cell line

On the basis of these results, we engineered a cell line, intMEM1, capable of inducible autonomous recording. We constructed an array of 10 trits, each with 500 base pairs of distinct barcode flanked by *att* sites, and site-specifically integrated it at the ROSA26 locus in mES cells (Fig. 2.2F). We also site-specifically integrated an inducible Bxb1 cassette at the TIGRE safe-harbor locus (*39*) and introduced the Tet3G doxycycline-dependent activator via piggyBac. In this cell line, doxycycline controls Bxb1 transcription and trimethoprim (TMP) stabilizes the protein by inhibiting a fused ecDHFR degron sequence (*40*). These complementary, redundant control systems together ensure tight regulation of integrase activity.

To quantify editing rates and outcomes, we co-cultured a low density of intMEM1 cells together with an excess of unengineered parental cells to support their growth. We induced recording for 36 hours by addition of doxycycline and TMP, terminated recording by washing out both inducers with fresh media, and then continued growth for 18 additional hours without inducers. Afterwards, we fixed the cells with formaldehyde (Fig. 2.2G). We then used five sequential rounds of hybridization chain reaction FISH (HCR-FISH) (fig. S2.3A) (*41*, *42*) to read each trit's edit state. Further, we subsequently imaged the same cells by immunofluorescence with antibodies against membrane proteins E-cadherin and β-catenin to facilitate segmentation of adjacent cells in images.

The state of the entire array was determined with five rounds of imaging using four fluorophores and 20 probe sets, one for each orientation of each trit. For example (Fig. 2.2H), in one cell, the first imaging round revealed a signal for the inverted orientation of trit 3 and no signal for the inverted orientations of trits 1 and 4, nor for the unedited orientation of trit 2. The second hybridization probed the opposite states of the same four trits, revealing inversion of trit 2 and deletion of trits 1 and 4. Automating most of this analysis, we determined the full array of editing outcomes in 1487 array-expressing cells (figs. S2.4 and S2.5). Most trits were deleted and inverted at similar rates (Fig. 2.2I). However, trits 6 and 10 underwent deletions but rarely inverted. DNA sequencing revealed that both trits had acquired truncation mutations in their *attP* sites, likely during cloning. Most notably, mutual information analysis showed that trits within the same array were edited independently (Fig. 2.2J). Taken together, these results demonstrate that trits can be combined into a compact recording array of independent units.

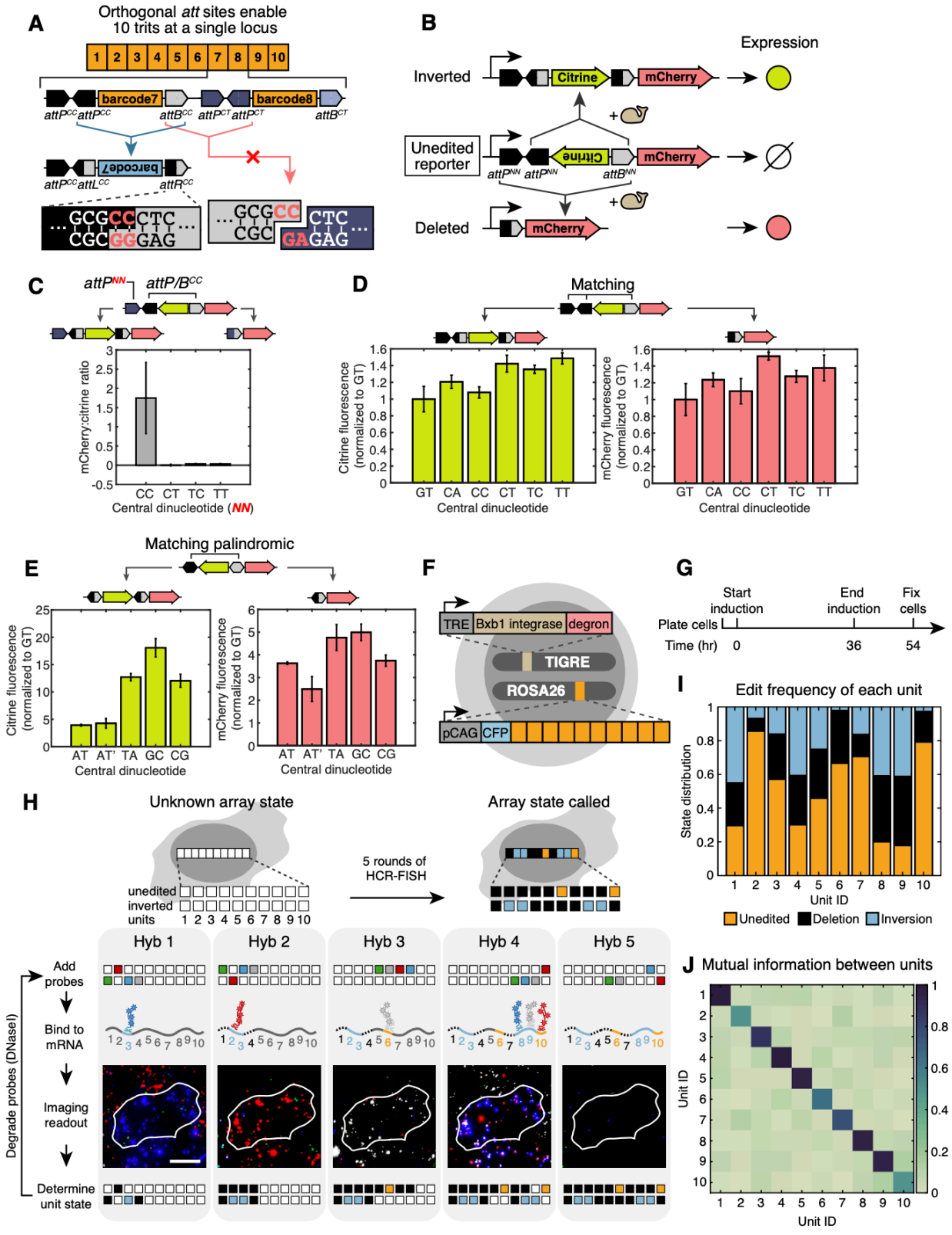

Fig. 2.2. Trits can be independently edited within recording arrays.
(*Figure legend continued on next page*)

(**A**) Ten trits (orange numbered rectangles) can be concatenated into a 10-unit array using *attP*/*attB* pairs with orthogonal central dinucleotides (red letters). (**B**) Fluorescent reporter assay enables rapid characterization of the recording units. The trit reporter construct (middle) produces no fluorescence until an integrase inverts it to express Citrine (top) or deletes it to express mCherry (bottom). In subsequent panels, cells transfected with reporter constructs were analyzed by flow cytometry after gating on a co-transfection marker (*41*). Plots show the mean ± SEM of the median fluorescent values from three independent experiments. (**C**) Bxb1 does not efficiently recombine *attP* and *attB* with mismatched dinucleotides. (**D**) All six non-palindromic dinucleotides mediate inversion (left) and deletion (right) of *attP/B* pairs. (**E**) With palindromic dinucleotides, one pair of matching *attP* and *attB* is sufficient for inversion and deletion. Bxb1 is agnostic to the relative orientation of these *att* sites, as demonstrated by the comparable edit efficiencies between $attP/B^{AT}$ when the two sites are arranged in the opposite (AT) and same (AT') orientations. (**F**) intMEM1 is a stable mES cell line with the 10-unit array integrated at the ROSA26 locus and an inducible Bxb1 integrated at the TIGRE locus. The 10-unit array is constitutively expressed, whereas Bxb1 can be activated by the combination of doxycycline for transcription and trimethoprim (TMP) for protein stabilization. (**G**) Over 36 hours of growth with Bxb1 induction, cells progressively accumulate edits for multigenerational lineage reconstruction. Induction is then stopped, and array edits are inherited by daughter cells over an 18-hour expansion period, enabling clonal classification. (**H**) Five rounds of HCR-FISH read out all possible states of the recording array *in situ* (scale bar, 10 μm). (**I**) The relative frequency with which each trit is observed in its unedited, deleted, or inverted state after 36 hours of Bxb1 induction. (**J**) The low mutual information between any given pair of units illustrates functional independence of each of the 10 units in the array.

2.3.3 intMEMOIR reconstructs lineage relationships

To quantify integrase-editable MEMOIR's (intMEMOIR's) lineage reconstruction ability, we obtained ground-truth lineages from time-lapse movies and compared them with lineage relationships reconstructed from array edits in the same cells (Fig. 2.3A and movies S1 to S3). We induced Bxb1 expression for 36 hours (Fig. 2.2G) to achieve ~3 generations of recording, followed by an additional ~1 or 2 generations of clonal expansion without Bxb1 induction, resulting in colonies of 13.7±7.7 cells (mean±SD). We then read out the state of the array using HCR-FISH, classified the cells into clones corresponding to distinct edit patterns, and performed multi-generation lineage tree reconstruction.

To assess clonal accuracy, we quantified the number of distinct edit states that could be detected in each colony and the fidelity with which they reflected ground truth clonal relationships (fig. S2.6) (*41*). We focused on the 76% of intMEM1 cells within colonies that showed the strongest array expression (Fig. 2.3B, orange versus green cells) and more than one array state. These colonies

exhibited 8.4±4.7 (mean±SD) distinct array states. Across 1453 cells spanning 105 colonies, 318 array edit patterns appeared in two or more cells in the same colony, with most clones (289) comprising two to four cells. Most clones were classified perfectly (median accuracy = 100%), and the average percentage of correctly classified cells per reconstructed clone was 85%, exceeding results from a negative control analysis in which barcode-cell relationships were scrambled (Fig. 2.3, C and D). Errors reflected false negative ambiguities due to subsets of cells within a clone undergoing additional edits, and false positive events in which distantly related cells convergently edited to identical patterns (Fig. 2.3C and fig. S2.6). On average, these errors occurred in <10% of cells per clone, and more than half of the clones had an error rate of 0% (Fig. 2.3D). Thus, intMEMOIR performed accurate clonal classification.

Next, we assessed the ability to reconstruct lineage trees (Fig. 2.3A, "lineage tree reconstruction"). We used a maximum likelihood approach that incorporates the empirically determined recording parameters (Fig. 2.2I), assuming a constant edit rate per unedited site (fig. S2.7) (*41*). Using this framework, we computed the probability of observing each array state after *G* generations starting from an unedited array; the conditional probability of observing any two specific array states as a pair of sister cells; and, from these probabilities, the relative likelihood of observing a given pair of array states for two sister cells compared to two unrelated cells. This likelihood provided a pairwise distance metric, which we then used to reconstruct a hierarchical lineage tree (Fig. 2.3, E to G, and table S2) (*41*). We analyzed 93 colonies, omitting those with three or fewer states that are trivial to reconstruct. Reconstructed trees for the classified clones were often identical (Fig. 2.3, E and F) or at least markedly similar (Fig. 2.3G) to corresponding ground truth lineage trees. In some cases, reconstruction errors could be attributed to multi-trit deletions at the 3' end of the array (e.g. Fig. 2.3G, cells 16 and 17).

To quantify reconstruction fidelity, we used the Robinson-Foulds (RF) metric, defined as the fraction of lineage partitions (clades) that are shared between the reconstruction and the ground truth (*43*). We defined a normalized RF score (*41*) that ranges from 0 (complete disagreement) to 1 (perfect agreement). Stochastic simulations provided an upper bound to the possible accuracy of the system under ideal conditions, given the memory capacity of a single array, the empirically measured editing rates, and the observed set of ground truth trees (Fig. 2.3H, cyan line). In parallel, we repeated this analysis, randomizing the cell-barcode relationships in the ground truth lineage, to compute a lower

bound on reconstruction performance (Fig. 2.3H, red line). Among the actual tree reconstruction scores (Fig. 2.3H, blue line), 25% of colonies reconstructed perfectly (RF score = 1), and the overall score distribution was significantly higher than the random control [$p < 10^{-16}$, Kolmogorov-Smirnov (K-S) test]. Further, colonies with greater normalized entropy in their edit patterns reconstructed with higher accuracy (fig. S2.8) (*41*). For instance, trees with the 40% highest normalized entropy showed performance similar to that of simulated optimal recording, statistically equivalent to the upper bound [$p > 0.2$, K-S test] (Fig. 2.3H, green lines). In applications where no ground truth is available, the entropy score thus allows one to enrich for subsets of cells likely to reconstruct with greater accuracy. Together, these results indicate that lineage recording and reconstruction can approach theoretical limits. The use of two or more arrays should therefore reconstruct colonies with greater depth and accuracy, as shown through simulations (fig. S2.9) (*24*, *44*).

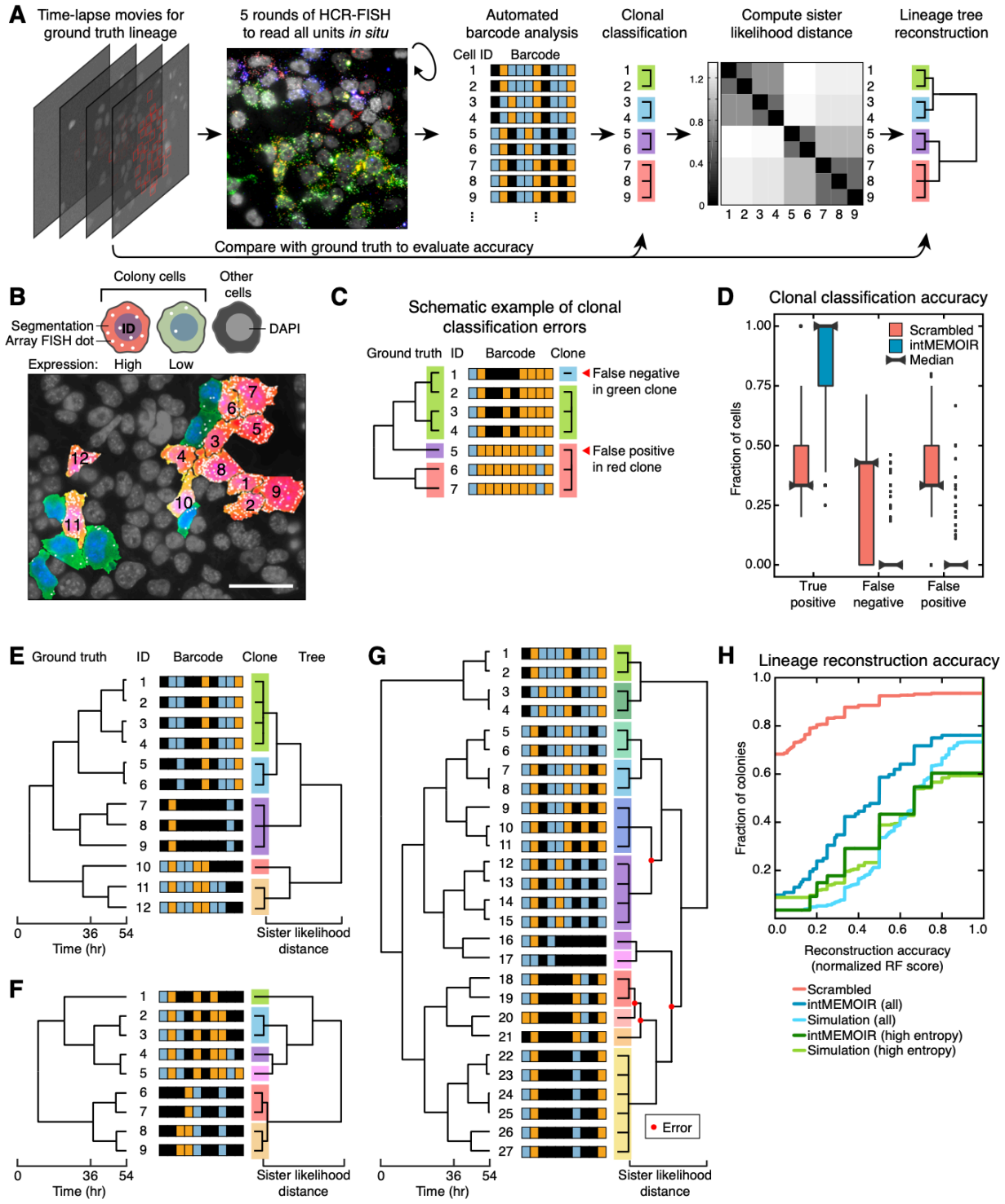

Fig. 2.3. intMEMOIR reconstructs lineage relationships.

(**A**) Quantitative analysis of the accuracy of clonal reconstruction and multigenerational lineage reconstruction. intMEM1 cells are tracked in time-lapse microscopy to establish ground truth lineage relationships (left panel). End-point HCR-FISH analysis recovers array edit states (second panel). Distinct edit patterns are used to classify cells into clones ("clonal classification," color groups), or further analyzed based on sister likelihood distance (*41*) to reconstruct lineage trees ("lineage tree

reconstruction," right). Comparison with ground truth allows quantification of reconstruction accuracy. (**B**) Cells from the same colony are segmented (highlighted in orange and green) and their array RNA HCR-FISH dots identified (white dots). Downstream analyses are performed on cells with strong array expression (orange cells) (scale bar, 50 μm). (**C**) Clonal classification errors arise either from clonal cells gaining additional edits (false negative) or convergent edits between distant relatives (false positive). (**D**) intMEMOIR demonstrates robust clonal classification accuracy, with few errors compared to scrambled control. (**E** to **G**) Lineage reconstruction examples, with ground truth lineage on the left, cell ID and their corresponding barcode states in the middle, and reconstructed lineage on the right. (E) and (F) are colonies with perfect tree reconstruction, whereas (G) shows reconstruction error in branches highlighted (red dots). (**H**) Cumulative distributions show that lineage reconstruction from intMEMOIR approaches the accuracy expected from simulations of a 10-unit trit array displaying experimentally observed edit rates (blue and cyan lines, respectively). Higher observed entropy in the edit patterns can independently identify colonies with greater reconstruction accuracy (green lines).

2.3.4 intMEMOIR reconstructs early lineage of large colony of mES cells

In many developmental contexts, it is of interest to know how distinct clones that acquired different fates were related to each other at an earlier time point (*2*, *45*). As a proof of principle for this type of analysis, we followed 36 hours of editing with an additional 70 hours of growth without editing (~6 cell divisions). We then fixed cells, analyzed array states, and classified clones (Fig. 2.4A). Imaging revealed large domains of distinct, non-redundant edits in each array element (Fig. 2.4B and fig. S2.10A). Combining these images provided a spatial map of clonal boundaries (Fig. 2.4C). Clonality broadly correlated with spatial position, as expected for colony growth. However, all clones were spatially extended and non-contiguous (Fig. 2.4C). Thus, intMEMOIR's ability to generate high digital diversity enables it to simultaneously label many intermingled clones. Further, the specific edit patterns also allow inference of clonal relatedness (Fig. 2.4D, right). Together, these results show how intMEMOIR can be used as an image-based clonal mapping system.

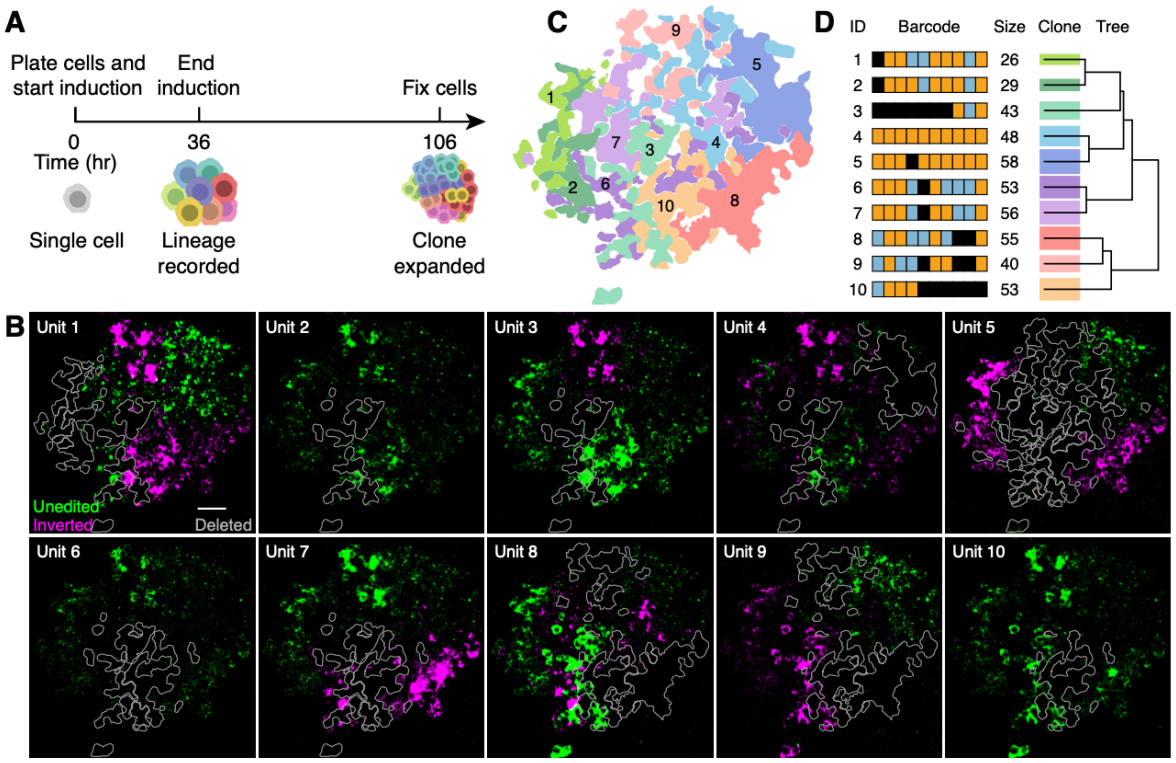

Fig. 2.4. intMEMOIR enables clonal reconstruction of large colonies.
(**A**) We induced Bxb1 in an intMEM1 colony for 36 hours of lineage recording, followed by 70 hours of clonal expansion, fixation, and imaging. (**B**) Example of HCR-FISH readout of the intMEMOIR array in a colony. Signals for the unedited and inverted states of each unit are colored green and magenta, respectively. Clones with deleted units are outlined in gray (scale bar, 50 μm). (**C**) Spatial distribution of the 10 clones in the colony. Cells are classified into clones on the basis of identical edits in their intMEMOIR arrays. (**D**) Reconstruction of the early lineage of the colony. Each clone is labeled with the number of cells it contains and its corresponding color in (C). Tree reflects reconstructed lineage relationships.

2.3.5 intMEMOIR reveals spatial organization of clones and gene expression states in *Drosophila melanogaster*

The *Drosophila melanogaster* brain provides an ideal model system to apply image-based clonal mapping *in vivo*. *Drosophila* permits rapid genetic engineering and quantitative imaging, and, while its brain development has been extensively characterized, fundamental questions about the role of lineage in fate determination remain unclear (*46*). The *Drosophila* central brain is known to develop from ~100 embryonic neuroblast progenitors per hemisphere (*46–48*), each exhibiting a distinctive lineage identity, acquired largely through spatial patterning, and controlled by lineage specific

transcription factors (*46*). A key step towards understanding central brain development is the capacity to specifically label and image all distinct clones within a single organism, along with their cellular gene expression states.

To achieve such multi-clonal labeling in a single individual, we constructed a fly line that allows controllable editing and cell-type-specific readout. We cloned the intMEMOIR array used in intMEM1 downstream of a UAS promoter also expressing mCerulean to identify array-expressing cells, and site-specifically integrated this construct using the phiC31 system (*49*). The resulting fly line, which we term "*Drosophila memoiphila*," provides a resource for general-purpose lineage analysis in flies. We then crossed these flies to an nSyb-Gal4 driver to restrict expression of the array to neurons (Fig. 2.5A) (*50*). Finally, we incorporated a Bxb1 integrase controlled by a tightly regulated heat shock inducible promoter (*51*) to record in specific time windows. Note that constitutive array expression combined with analysis of endogenous gene expression could also permit analysis of specific cell types in species that lack the Gal4 system or tissue-specific promoters.

To confirm intMEMOIR operation, we exposed flies to varying durations of heat shock during early development, collected adults, sectioned their brains, and read out the intMEMOIR array using sequential rounds of HCR-FISH (*42*). Negligible levels of editing occurred without heat shock, whereas exposure to 37ºC for 0.5 to 3 hours produced dose-dependent increases in editing, as quantified by analysis of two units (Fig. 2.5B) (*41*). These results demonstrate that Bxb1 activity can be controlled in a tight, dose-dependent manner by heat shock duration (Fig. 2.5B).

We next sought to integrate analysis of lineage, cell state, and spatial organization in a single brain. To induce editing in neuroblasts during early embryonic development, we applied a 1-hour heat shock starting 4 hours after egg laying (Fig. 2.5C). We then analyzed brain sections from adults by 15 rounds of automated smFISH, reading out not only the intMEMOIR array and mCerulean transcripts but also eight endogenous genes selected on the basis of their ability to identify diverse neuronal cell types (Fig. 2.5D and figs. S2.10B and S2.11). Of the 29 smFISH probe sets designed for this experiment (fig. S2.3B and table S4), one, targeting unedited unit 7, displayed nonspecific binding and was excluded from downstream analysis. Altogether we analyzed different sections from four brains, labeled B1 to B4 (*41*).

We first focused on a single section of brain B1 (Fig. 2.5D). We analyzed 29 clones that contained at least four cells and one inverted unit each (fig. S2.12). Most such clones consisted of tightly apposed cells (e.g. Fig. 2.5E inset, clones 1 and 2). However, a minority of clones were more dispersed, intermingling with other clones within the section (e.g. Fig. 2.5E inset, clone 3) (*52–54*) (Fig. 2.5F). Similar results were obtained in sections from three additional brains (figs. S2.13 to S2.15). These results, which are consistent with previous observations (*55*, *56*), demonstrate the ability to simultaneously map the spatial arrangements of many clones in the same brain section.

We next sought to visualize the spatial distribution of gene expression states. We used principal component analysis (PCA) (*57*), uniform manifold approximation and projection (UMAP) (*58*), and density-based clustering (DBSCAN) (*59*) to denoise, reduce dimensionality, and identify expression states (Fig. 2.5, G and H). Combining gene expression with spatial location identified known cell types including γ-aminobutyric acid (GABA)–producing neurons, dopaminergic neurons, and Kenyon cells (Fig. 2.5G) and allowed us to plot their spatial distribution within the section of brain B1 (Figure 2.5I). Cross-referencing the spatial maps of expression states and lineage also allowed simultaneous inspection of any specific region, cell state, and lineage of interest. For example, intMEMOIR captured two out of the four known lineages of Kenyon cells in this experiment (Fig. 2.5, E and I, cluster T in the right hemisphere) (*46*). Further, by labeling distinct clones in gene expression space, we were able to visualize correlations between clonal identity and cell type similarity (Fig. 2.5J, clustering of clones in gene expression space). This analysis revealed homogeneous clones containing a single cell type (Fig. 2.5J, clones 1 and 2), as well as more heterogeneous clones containing multiple cell types (Fig. 2.5J, clone 3), consistent with previous observations (*55*). Overall, cells within the same clone tended to be more similar to one another in gene expression than cells in different clones in all four brain sections (Fig. 2.5K and figs. S2.13 to S2.15).

intMEMOIR's ability to simultaneously analyze lineage, gene expression, and spatial arrangement in the same tissue could allow it to disentangle the contributions of cell-intrinsic, inherited factors and extrinsic, spatially organized cues to cell fate determination. To evaluate the contribution of neuroblast ancestry to fate determination, we compared the similarity of gene expression states between cell pairs within the same clone to cell pairs in different clones, across a range of different spatial separations. If extrinsic cues dominate, one would expect cell state similarity to strongly

correlate with spatial proximity, independent of clonal identity. By contrast, if intrinsic determinants dominate, then gene expression similarity should correlate with clonal identity, regardless of spatial separation (Fig. 2.5L). Here, among cell pairs drawn from distinct clones, transcriptional similarity showed little dependence on spatial distance (Fig. 2.5M, blue lines). However, cell pairs within the same clone showed strong cell type similarity at close distances, with a gradual relaxation of this similarity at larger distances (Fig. 2.5M, yellow-red lines). This result is robust to exclusion of the large, homogeneous population of Kenyon cells (cluster T, fig. S2.16A) and to other choices of gene expression distance metric (fig. S2.16B). Clonally restricted, spatially graded dependence of cell fate similarity was also observed in brains B2 and B4 (fig. S2.13F and S2.15F), indicating that it could be a general feature. Brain B3 did not show the effect, likely because homogenous gene expression within its clones did not provide an opportunity to detect distance dependent differences in cell fate (fig. S2.17). These relationships, although strong, would be difficult to observe without the simultaneous spatial, lineage, and gene expression analysis enabled by intMEMOIR.

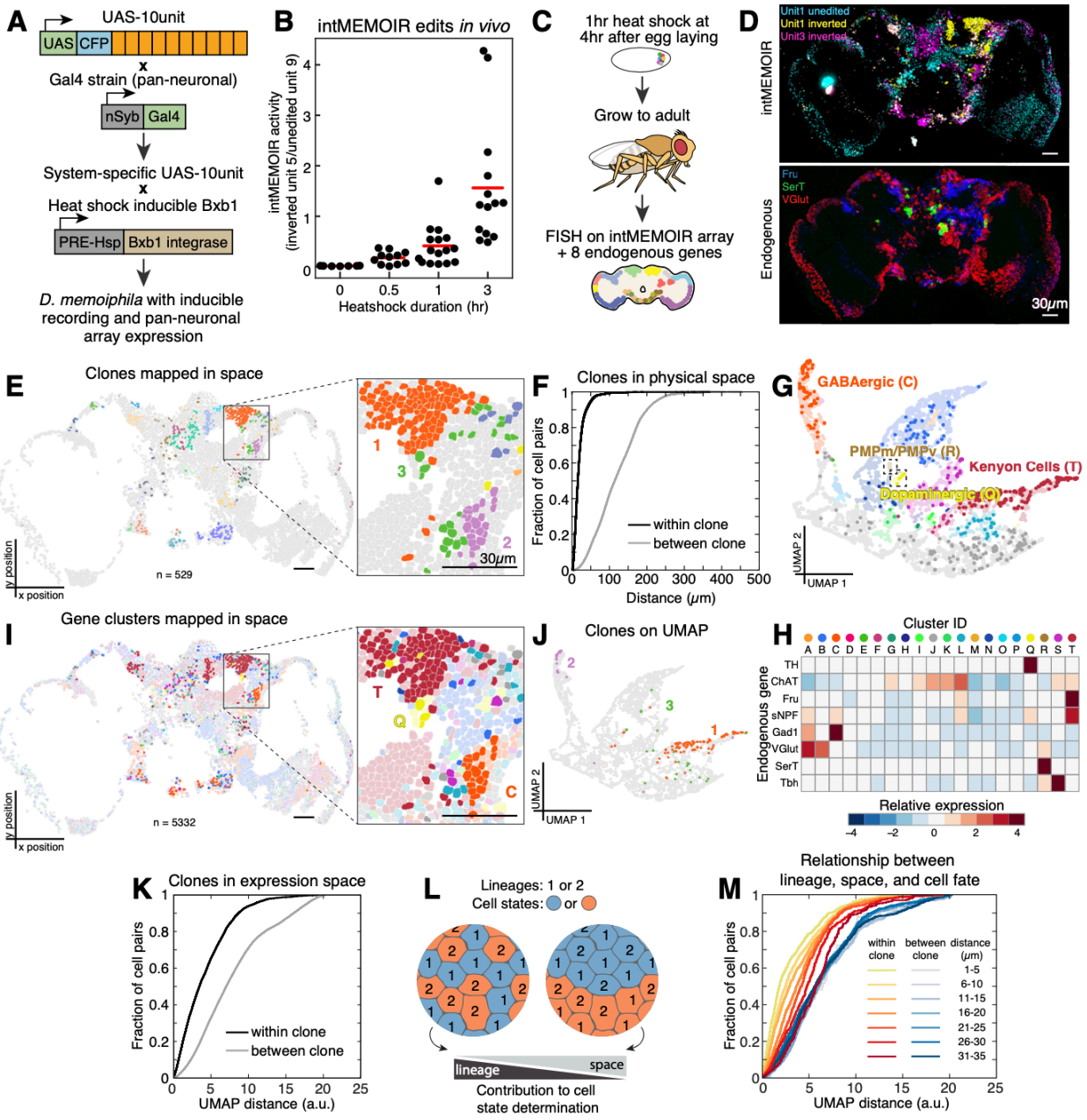

Fig. 2.5. intMEMOIR connects single-cell spatial, molecular state, and lineage information in adult *Drosophila* brain.

(**A**) A neuronal recording fly line was obtained by crossing *D. memoiphila* flies containing a site-specifically integrated UAS-10-unit array with an nSyb-Gal4 strain. Offspring were then crossed with a tight heat shock inducible Bxb1 line to produce a fly line that allows heat shock induction of editing and exhibits pan-neuronal expression of the recording array. (Note that other Gal4 drivers could be used to analyze distinct tissues or cell types.) (**B**) Editing activity, measured as the ratio of edited unit 5 over unedited unit 9, increases with heat shock duration in a dose-dependent manner. (**C**) To achieve *in situ* analysis of cell state and clonal identity, editing was induced with a 1-hour heat shock at 4 hours after egg laying. This early induction aims to label neuroblasts with distinct array states that can be inherited by all neuron progeny in the adult brain. Flies were then grown to adulthood, and their brains dissected and cryosectioned. Sequential rounds of automated smFISH

were used to read out the intMEMOIR array and eight endogenous genes: tyrosine hydroxylase (TH), choline acetyltransferase (ChAT), fruitless (Fru), short neuropeptide F precursor (sNPF), glutamic acid decarboxylase (Gad1), vesicular glutamate transporter (VGlut), serotonin transporter (SerT), and tyramine β-hydroxylase (Tbh). (**D**) Example images from brain B1 showing single cell resolution imaging of endogenous expression and array state in the same tissue sample (scale bar, 30 μm; see fig. S2.11 for additional examples). (**E**) intMEMOIR clones of brain B1 mapped in space. Segmented cells are colored by the 29 analyzed clones (n=529; scale bar, 30 μm). Inset highlights examples of clones that are clustered (clones 1 and 2) and dispersed (clone 3) in space. Cells outside the 29 clones are gray. Clone colors are consistent with (J). (**F**) Cells within the same clone (black line) are arranged closer in physical space than cells between clones (gray line). Cumulative distributions show pairwise distance between the cells. (**G**) UMAP clustering of 5,332 cells in brain B1 based on the expression of the eight endogenous genes. Four clusters are annotated by inspection based on expression patterns and *in situ* localization of the majority population. Cells among the 29 analyzed clones are highlighted with saturated colors. Cluster colors are consistent between (G), (H), and (I). (**H**) Heatmap showing the relative expression, calculated as normalized Z-score, of the eight genes in each cluster for brain B1. (**I**) Gene expression clusters mapped in space. Segmented cells are colored by cluster. Cells among the 29 analyzed clones are highlighted with saturated colors (scale bar, 30 μm). The inset highlights the same cells in (E), demonstrating clones that display similar (clones 1 and 2) and mixed (clone 3) cell states. (**J**) The three represented clones of brain B1 mapped onto UMAP space, demonstrating examples of cells whose molecular states are correlated to their neuroblast lineage to varying degrees. (**K**) Cells within the same clone (black line) are more similar in expression than cells in different clones (gray line). Cumulative distributions show pairwise UMAP expression distances. (**L**) Hypothetical observations if either neuroblast lineage or spatial distribution alone dominates cell state determination in the fly brain. (**M**) Within the same clone, larger physical distances between cells correlate with greater gene expression differences (yellow to red colors). This correlation is not observed between cells of different clones (gray to blue colors). Cumulative distributions show pairwise UMAP expression distances.

2.4 Discussion

How the lineage history of a cell affects its future potential is central to development but challenging to systematically address in most systems. Clonal imaging methods such as MARCM revolutionized lineage analysis but can discriminate a limited number of clones per animal (*15*). Conversely, sequencing-based recording approaches such as GESTALT can provide higher throughput lineage information but do not preserve spatial information (*3*, *11*, *16*, *17*, *19*, *20*, *22*). intMEMOIR allows high-density, cell-autonomous, digital editing with imaging-based readout for lineage reconstruction and is compatible with FISH gene expression measurements, providing the means to simultaneously analyze single cell lineage, spatial organization, and gene expression data in the same tissue.

We implemented intMEMOIR in two biological contexts: mouse embryonic stem cells and fly neural development. Analysis of the three-way lineage-space-expression relationship in the brains revealed a spatially graded, lineage-dependent correlation in cell state, and illustrated the potential of our system to reveal new insights. In the future, labeling at additional time points and analysis of larger numbers of genes should allow more detailed dynamic analysis of development over a broader range of cell types and time scales.

The current intMEMOIR implementation can produce a theoretical maximum of 59,049 possible outcomes from a single array. This diversity is sufficient for accurate tree reconstruction across several cell divisions. Additional recording arrays would exponentially increase the number of array states, enabling reconstruction of deeper trees over longer developmental time scales (Fig. 2.3 and fig. S2.9). Further, the system could be extended to allow additional, independent "channels" consisting of distinct, orthogonal integrases and their corresponding sets of arrays (*60*). By validating independent editing by distinct integrases and making their activities conditional, one could create systems that allow reconstruction of the dynamic activity histories of signaling pathways and transcription factors (fig. S2.2) (*23*). Thus, intMEMOIR provides a versatile, extensible basis for developing new recording applications.

Several aspects of intMEMOIR should allow its use in other species and biological contexts. First, its genetically compact array design facilitates engineering of transgenic animal lines, as exemplified by *D. memoiphila*, which can be crossed with other fly lines to enable recording in desired cell types at appropriate developmental times and in specific genetic backgrounds (Fig. 2.5A). Second, a minimal intMEMOIR implementation only requires regulated integrase activity and array expression prior to imaging. Editor activity can be regulated using diverse system-appropriate methods, including Gal4-UAS, heat shock promoters, cell type specific promoters, or Cre driver lines (*61*). Similarly, array expression can be either selectively induced in cell types of interest (e.g. pan-neuronal expression), constitutively expressed in live tissue, or potentially even expressed in fixed tissue using *in situ* T7 transcription (*62*). Even without cell type specific array expression, array states in cell types of interest can be identified using FISH on relevant marker genes within the same workflow. Thus, we anticipate that intMEMOIR should be readily adaptable to other model organisms and developmental contexts. More generally, as cell atlas projects develop, it should be possible to incorporate spatial and morphological data and lineage relationships alongside molecular

profiles. The outcome would provide a richer view of cellular states and histories, allowing exploration of interactions among clonal lineages and analysis of developmental variation between individual organisms.

2.5 Methods Summary

We performed all tissue culture experiments with E14 mES cells (ATCC). For flow cytometry experiments shown in Figure 2.2, mES cells were cotransfected with mTagBFP2 (as cotransfection marker), integrase, and the sample's corresponding prototype trit reporter. We performed flow cytometry 2 days after transfection. We constructed intMEM1 by sequentially introducing a TIGRE locus landing pad [modified from (*39*)] and Tet3G via PiggyBac (System Biosciences). We then integrated TRE-Bxb1-ecDHFR into the TIGRE landing pad. We also integrated the 10-unit intMEMOIR array into the Rosa26 locus using CRISPR-Cas9. Finally, we integrated mTurquoise2 via PiggyBac. The cell lines underwent multiple rounds of clonal selection during this process, and the final intMEM1 line is monoclonal.

For time-lapse intMEM1 imaging experiments, cells were plated on glass bottom 24-well plates (Eppendorf) coated with Laminin-511 (BioLamina) overnight. intMEM1 cells were co-cultured with parental E14 cells to increase total cell density to support growth and survival. For the lineage reconstruction experiments shown in Figures 2.2 and 2.3, we induced recording for 36 hours by the addition of the inducers trimethoprim and doxycycline to the culture media. We then halted the induction by washing off the induction media and replacing it with regular culture media, followed by another 18 hours of expansion. We then fixed cells and performed five sequential rounds of HCR-FISH (*41*, *42*) to read out the intMEMOIR array state, followed by immunostaining and imaging of E-cadherin and β-catenin to facilitate cell segmentation.

To reconstruct the ground truth lineage, we manually tracked the cells in the time-lapse images using a modified version of the EasyTrack software [available at (*63*)]. We then identified individual array edit states using Ilastik (*64*), manual curation, and a custom analysis pipeline in Matlab [available at (*63*)]. Lineage trees were then reconstructed using a maximum likelihood approach (*41*). Similarly, for the large colony lineage reconstruction experiment in Figure 2.4, cells were induced for 36 hours, followed by 70 hours of growth with no induction, fixation, and HCR-FISH. Clone boundaries and

barcodes for this colony were analyzed manually.

D. memoiphila fly lines were generated by site-specifically integrating the UAS-Ceru-10unit and PRExpress-Bxb1 constructs into the attp2 and VK27 sites, respectively (Bestgene Inc.). UAS-Ceru-10-unit flies were first crossed with an nSyb-Gal4 line (Bloomington). Offspring were crossed with the PRExpress-Bxb1 line to generate an autonomous recording line with pan-neuronal expression of the intMEMOIR array. For the heat shock inducibility experiment, the flies were maintained at 25°C and heat shocked in 37°C water baths for the specified durations at the embryo stage. Adult fly brains were dissected, fixed, and cryosectioned onto coverslips pre-treated with 3-aminopropyltriethoxysilane (Sigma). We installed SecureSeal Hybridization Chambers (Grace Bio-Labs) onto the coverslips to perform HCR-FISH (*41*) for intMEMOIR array readouts.

For the early neuroblast labeling experiment, fly embryos obtained 4 hours after egg laying were heat shocked at 37°C for 1 hour. The resulting adult flies were incubated at 29°C overnight to enhance the activity of Gal4 prior to brain collection, fixation, and cryosection onto coverslips pretreated with 1% bind-silane (GE) and poly-D-lysine (Sigma). To read out the intMEMOIR arrays and eight endogenous genes, we used an automated imaging and fluidics delivery system to perform multiple rounds of smFISH (*41*, *65*). Both HCR-FISH and automated smFISH are viable readout methods, but the latter offers higher throughput. For downstream analysis, we segmented the fly cells manually and used a custom Matlab program [available at (*63*)] to call the array edit states. To analyze the gene expression data for each brain, we applied PCA (*57*), UMAP (*58*), and DBSCAN (*59*) to denoise, reduce the dimensionality, and cluster the dataset. UMAP distance between cell pairs was then calculated as the Euclidean distance between their two-dimensional UMAP coordinates. We also calculated the physical Euclidean distance between cell pairs. On the basis of their intMEMOIR array state, cell pairs were then subsequently divided into “within clone” and “between clones” for analysis. To disentangle the contribution of space and lineage to gene expression, we also further binned the data by the physical Euclidean distance between cell pairs.

All data, code, analysis, and sequence files are freely available on (*63*), and the full materials and methods are available in (*41*).

2.6 Acknowledgements

We thank L. Sanchez-Guardado, H. Choi, C. Calvert, G. Shin, C. Tischbirek, Y. Takei, S. Shah, and N. Pierson for technical assistance and advice; A. Askary, X. Gao, F. Horns, D. Chadly, C. Su, and other members of the Elowitz lab for critical feedback on the manuscript; and A. Shur, P. Meyer, R. Lu, and J. Linton for scientific input and advice. **Funding:** This research was supported by the Allen Discovery Center program, a Paul G. Allen Frontiers Group advised program of the Paul G. Allen Family Foundation (grant UWSC10142 to M.B.E., C.L., and L.C.), the National Institutes of Health (NIH) (grant R01 MH116508 to M.B.E., C.L., and L.C.), and Burroughs Wellcome Fund CASI (K.L.F.), and M.B.E. is a Howard Hughes Medical Institute investigator. M.B.E. acknowledges Fritz Thyssen Stiftung for support for a visiting fellowship to Berlin. **Authors contributions:** K.K.C., M.W.B., A.A.G., K.L.F., L.C., C.L., and M.B.E. designed research. K.K.C., M.W.B., M.C., S.Y., S.C., and N.K. performed experiments. A.A.G., M.W.B., K.K.C., T.H., C.L., and M.B.E. analyzed data, K.K.C., M.W.B., A.A.G., C.L., and M.B.E. wrote the manuscript. **Competing interests:** K.L.F., K.K.C., L.C., and M.B.E. are inventors on a patent application for recording technologies. **Data and materials availability:** Plasmids to implement intMEMOIR in mES cells and *Drosophila melanogaster* are available from the Addgene repository (ID: 158387, 158389, 158390, and 158391), *Drosophila melanogaster* lines are available from the Bloomington repository (RRID: BDSC_90853 and BDSC_90854), and the intMEM1 cell line will be made available from C.L. and M.B.E. under the terms of the Uniform Biological Material Transfer Agreement (UBMTA). The data, code, and analysis to generate the results in the manuscript and the sequence information of relevant constructs are freely available on (*63*).

2.7 Materials and Methods

Plasmids preparation

Constructs were cloned using standard methods. Due to the repetitive sequence, inverted *attP*s were difficult to amplify *in vitro*, therefore PCR-based cloning methods were avoided for these regions. Mammalian constructs involving serine integrases Bxb1, phiC31, R4, and TP901 were cloned from, or used directly as, plasmid gifts from Mitsuo Oshimura (*66*). All constructs reported in this manuscript are listed in table S1, and sequence maps for constructs generated for the intMEMOIR system are available at (*63*). The Bxb1 and intMEMOIR array constructs are available on Addgene:

TIGRE-TRE-[poor kozak]Bxb1-ecDHFR-BGHpA: https://www.addgene.org/158390/
R26-pCAG-Ceru-10unit-BGHpA: https://www.addgene.org/158387/
PRExpress-Bxb1-hsp70pA: https://www.addgene.org/158391/
UAS-Ceru-10unit: https://www.addgene.org/158389/

Tissue culture

All tissue culture experiments were done with E14 mouse embryonic stem (mES) cell line (ATCC catalog number CRL-1821). Cells were cultured in humidified chambers at 37°C and 5% CO_2, with filtered media composed of GMEM (Sigma), 15% FBS, PSG (100 units/mL penicillin, 100 µM/mL streptomycin, 2 mM L-glutamine) (ThermoFisher), 1mM sodium pyruvate (ThermoFisher), 1X Minimum Essential Medium Non-Essential Amino Acids (MEM NEAA, ThermoFisher), and 100µM 2-Mercaptoethanol (ThermoFisher), with 1,000 units/mL Leukemia Inhibitory Factor (LIF, Millipore) added after filtering. Cells were maintained on polystyrene plates coated with 0.1% gelatin.

Flow cytometry

For flow cytometry experiments shown in Figures 2.2, C to E and S2.2, D and E, mES cells were plated on 24 well plates at approximately 70% confluency. Cells in each well were then cotransfected with 200 ng mTagBFP2, 400 ng integrase, and 400 ng of the sample's corresponding prototype trit reporter. The transfections were performed with Lipofectamine LTX and PLUS reagent overnight (ThermoFisher).

Flow cytometry was performed two days after transfection on CytoFlex (Beckman Coulter). Cells were lifted from the plate with StemPro Accutase (ThermoFisher) and resuspended in buffer made of Hank's Balanced Salt Solution (HBSS), 2.5mg/mL Bovine Serum Albumin (BSA), and 1mM EDTA. They were then filtered through a 40 µm cell strainer prior to flow cytometry. These experiments, including their respective transfections, were conducted in triplicate (Fig. 2.2C to E).

Flow cytometry data were analyzed using the EasyFlow Matlab program developed by Yaron Antebi, the version used for this manuscript available at (*63*), and the latest version available at

https://antebilab.github.io/easyflow/. We gated for single cells using forward and side scatter (FSC and SSC), then gated for cells expressing high levels of the cotransfection marker mTagBFP2 to enrich for the transfected population in downstream analysis. For figure S2.2, D and E, we plotted the resulting distributions of Citrine and mCherry fluorescence for the relevant triplicates. For Figure 2.2C, we determined the median Citrine and mCherry fluorescence, background subtracted the fluorescence detected in the no integrase negative control, and calculated the mCherry/Citrine ratio for each replicate. We then plotted the average ratio of the experimental triplicates, with error bars representing the standard error of the mean (SEM). For Figure 2.2, D and E, we determined the median Citrine and mCherry fluorescence for each sample, background subtracted the fluorescence detected in the no integrase negative control, and averaged the values over the experimental triplicates. The resulting values were then normalized to the values for matching GT *att* sites and plotted for comparison, with the error bars representing normalized SEM.

Characterization of additional members of the serine integrase family

To characterize the ability of additional, non-Bxb1 serine integrases to function in mES cells, we constructed stable reporter cell lines containing either an integrase-specific reporter construct (fig. S2.2, A and B), or a 4 unit array with palindromic *att* sites (fig. S2.2C). These cell lines were then transiently transfected with their respective integrases, and the results evaluated through hybridization chain reaction FISH (described in section below).

To construct the stable reporter mES cell lines, the reporter constructs were site specifically inserted into the Rosa26 locus through Cas9-mediated homologous recombination by cotransfection of 600 ng of the reporter plasmid with 200 ng of pX330 Cas9 (gRNA sequence: CAGGACAACGCCCACACACC), followed by 500 µg/mL geneticin selection.

Polyclonal stable cell lines were used to generate figure S2.2, A and B. Monoclonal cell lines were selected for figure S2.2C. The cells were then transiently co-transfected with 600 ng of their corresponding integrase and 200 ng of puromycin resistance plasmid. After one day, 1 µg/mL puromycin was added for two days to enrich for transfected cells in downstream experiments. After ending selection, cells recovered in regular media for one day, and were then plated on glass bottom 96-well plates (Cellvis) coated with 20 µg/mL of Laminin-511 (BioLamina).

For figure S2.2, A and B, cells were fixed one day after plating for HCR-FISH. Integrase activities were calculated as the percentage of manually counted cells with inverted reporters, out of all cells with the invariable barcode (fig. S2.2A). For figure S2.2C, cells were fixed approximately 4 hours after plating for HCR-FISH. The results were analyzed by a custom Matlab pipeline available at (*63*). Briefly, the program automatically segments the cells and counts HCR-FISH dots above manually determined thresholds. To account for different array expression levels, dot counts were normalized to nuclear CFP intensities before using them to call a unit state (i.e. unedited, inverted, and deleted), while the unnormalized dot counts were used to resolve rare conflicts between unedited and inverted calls. The 4 unit array constructs contain a barcode without *att* sites located in the middle of the array, and only cells with unedited middle barcodes detected were used for analysis.

All transfections were performed with Lipofectamine LTX and PLUS reagent (ThermoFisher) overnight, on mES cells plated at approximately 70% confluency on a 24-well plate. All monoclonal selections involving site-specific integrations were screened with PCR.

intMEM1 cell line construction

To construct intMEM1, we began by integrating a landing pad containing FRT sites into the TIGRE locus using Cas9-mediated homologous recombination (construct modified from (*39*)). This was achieved through cotransfection of 600 ng of TIGRE-LandingPad-FRT-partialHygro-SV40pA and 200 ng of pX330 Cas9 (gRNA sequence: CTGCCATAACACCTAACTTT), followed by selection with 10 μg/mL blasticidin. After selecting a clone with correct integration, we introduced constitutive pEF1α-Tet3G through PiggyBac transposition (System Biosciences), transfecting with 600 ng of the Tet3G plasmid and 200 ng of the transposase, followed by 1 μg/mL puromycin treatment, and again selected for a single clone. We then inserted TRE-Bxb1-ecDHFR into the TIGRE landing pad with FlpE recombinase through cotransfection of 600 ng of TIGRE-TRE-(poorKozak)Bxb1-ecDHFR-BGHpA and 200 ng of FlpE, followed by 100 μg/mL hygromycin selection. The resulting polyclonal line was transfected with the 10-unit intMEMOIR array targeted to the Rosa26 locus through Cas9-mediated homologous recombination, by cotransfection of 600 ng of R26-pCAG-Ceru-10unit-BGHpA and 200 ng of pX330 Cas9 (gRNA sequence: CAGGACAACGCCCACACACC), 500 μg/mL geneticin selection, and monoclonal selection.

Finally, we increased the fluorescence of these cells for time-lapse movie tracking by integrating PGK-mTurquoise2-Blast by PiggyBac, using 600 ng of the marker plasmid and 200 ng of transposase, followed by blasticidin and a second round of hygromycin selection. A final round of monoclonal selection resulted in the intMEM1 cell line. All transfections were performed with Lipofectamine LTX and PLUS reagent (ThermoFisher) overnight, on mES cells plated at approximately 70% confluency on a 24-well plate. All monoclonal selections involving site-specific integrations were screened with PCR.

Time-lapse imaging for ground truth lineage

Cells were plated on glass bottom 24-well plates (Eppendorf) coated with 20 µg/mL of Laminin-511 (BioLamina) overnight. Approximately 6,000 intMEM1 cells were seeded onto the coated wells, along with 18,000 parental E14 cells to increase cell density to support growth and survival. Media was changed prior to the start of the movie to remove any unattached cells. Imaging was done with an Olympus IX81 inverted epi-fluorescence microscope with Photometrics Prime 95b sCMOS camera, 20x air objective (0.75 numerical aperture), and equipped with an environmental chamber. intMEMOIR recording was initiated by adding 10 µM TMP (to block the DHFR degron) and 100 ng/mL doxycycline (to activate the TRE3G promoter). Inducers were omitted in negative control samples. For each position, images were acquired every 15 minutes in both the visible light (DIC) and fluorescent (CFP) channels. 36 hours after the start of the movie we halted induction by washing off the induction media and replacing it with regular culture media. 54 hours after the start, we terminated time-lapse imaging and promptly fixed the sample at room temperature with 4% formaldehyde in PBS for 5 minutes, followed by HCR-FISH protocol (below).

Constructing ground truth lineage

Ground truth lineage trees were constructed by manually tracking the cells in the time-lapse images using a modified version of the EasyTrack software developed by Yaron Antebi (freely available at (*63*) and https://github.com/AntebiLab/EasyTrack/tree/Memoir). Cells were primarily tracked by their CFP fluorescence. Ground truth trees could begin at either the one or two cell stage depending on the colony's cell cycle at the start of image acquisition, and were rooted at the two cell stage if

the parent cells in question were likely sisters based on proximity, cell morphology, CFP intensity, as well as their cell movements and cycles in the subsequent frames. Ground truth trees for all colonies were outputted as Newick strings (table S2 for colonies used in lineage reconstruction).

Hybridization Chain Reaction (HCR) FISH and imaging

Overview of imaging workflow:

The imaging protocol consists of fixation and permeabilization steps, followed by multiple rounds of primary probe binding and signal amplification by HCR (fig. S2.3A). Below, we describe each of these steps in more detail.

Fixation and permeabilization:

HCR-FISH in tissue culture began after fixing the samples at room temperature with 4% formaldehyde in PBS for 5 minutes (as described above). Fixed cells were washed with PBS, followed by permeabilization in 70% RNase-free ethanol at -20^{o}C overnight, and stored in 70% ethanol for up to 3 days at -20^{o}C. Permeabilized cells were washed with 20% formamide wash buffer in 2X SSCT at room temperature for 5 minutes and pre-hybridized in 30% probe hybridization buffer at 37^{o}C for 30 minutes.

Primary probe hybridization:

Primary probe hybridizations and hairpin amplifications were then carried out as previously described for HCR v3.0 (Molecular Instruments) (*42*). Primary probes for each round of hybridization were prepared in probe hybridization buffer (warmed to 37^{o}C) at 4 nM per probe. The pre-hybridization solution was then replaced with the probe solution with an overnight incubation at 37^{o}C. The samples were then washed 4 times with warm 30% probe wash buffer at 37^{o}C, with 15 minutes incubation accompanying each wash. Finally, samples were washed once with 5X SSCT at room temperature for 5 minutes.

HCR amplification:

Samples were incubated in amplification buffer at room temperature for 30 minutes. Hairpins for amplification were prepared by snap cooling each at the stock concentration of 3 μM. This was done by heating the individual hairpins to 95^{o}C for 90 seconds, then cooling them to room temperature in

the dark for 30 minutes. The cooled hairpins were then mixed and prepared in amplification buffer at 60 nM final concentration for each hairpin. The pre-amplification solution on the sample was then replaced with the hairpin mix and incubated at room temperature from 4 hours to overnight. During incubation and for all subsequent steps, the sample plate was protected from light by covering it with aluminum foil except during pipetting and/or imaging.

Amplification was ended with two 5 minute washes, two 30 minute washes, and one 5 minute wash of 5X SSCT at room temperature. Finally, the cells were imaged in 5X SSCT.

Materials:

30% probe hybridization buffer, 30% probe wash buffer, amplification buffer, and HCR amplification hairpins were purchased from Molecular Instruments. Hairpins used for the intMEM1 experiments were (in the format of HCR initiator-fluorophore): B1-Alexa594, B2-Alexa647, B3-Alexa546, and B4-Alexa488. The probe binding regions for each intMEMOIR unit, along with their corresponding initiators, are listed in table S3, and the probes can be purchased from Molecular Instruments with order IDs 3049 and 3092.

Rehybridization:

Between hybridization rounds, probes were removed via DNase I treatment (Roche). Briefly, cells were washed with 1X DNase buffer, followed by incubation with 1 Kunitz unit/μL DNase I in 1X buffer for 2 to 4 hours at 37^{o}C. Digestion was ended by washing the cells 3 times with 30% probe wash buffer, incubating the final wash for 15 minutes at 37^{o}C. Finally, cells were washed once with 5X SSCT before the pre-hybridization step for the next round of HCR-FISH.

Imaging:

Cells were imaged using a Nikon Eclipse Ti inverted fluorescence microscope, with an Andor Zyla 4.2 sCMOS camera and a 60x oil objective (1.4 numerical aperture). For all HCR-FISH channels, each field of view was acquired with 0.5 μm z-steps for 20 z-slices. Maximum intensity projections from the in-focus slices were then used for downstream analysis.

<u>Antibody staining</u>

Upon completion of all rounds of HCR-FISH readout and a final round of DNase I treatment to remove any HCR-FISH signals, cells underwent antibody staining for membrane markers E-cadherin and β-catenin to facilitate segmentation. Immunostaining was performed following standard protocols, with most incubation and washing steps carried out on a gentle rocker. Briefly, samples were blocked with blocking buffer made in PBS (5% BSA, 1% DMSO, and 0.2% Triton X-100) for 1 hour at room temperature. They were then incubated with primary antibodies E-cadherin (R&D Systems, AF648, 1:20) and β-catenin (Abcam ab6301clone15B8, 1:750) overnight at 4^{o}C. The following day, they were washed 5 times with PBST for 5 minutes each, then incubated with secondary antibodies (donkey anti-goat IgG 647 A21447, and donkey anti-mouse IgG 488 A21202, respectively) diluted 1:1000 in blocking buffer for 3 hours at room temperature. Finally, samples were washed 5 times with PBS for 5 minutes each, followed by imaging in fresh PBS.

Analysis of HCR-FISH readout in mES cells

To segment individual cells and identify their array edit states, we used a custom analysis pipeline in Matlab (fig. S2.4, available at (*63*)). For cell segmentation, immunofluorescence images of E-cadherin were first preprocessed with Ilastik (*64*) to generate a membrane probability map. The positions of cells in the final frame of the ground truth lineage analysis were used as watershed seeds overlaid on the membrane probability maps. The resulting segmented images were visually examined and manually curated. The four channels used for HCR-FISH analysis did not show significant fluorescent crosstalk (fig. S2.5).

For array state determination, the centers of the mRNA dots were determined using a Laplacian of Gaussian filter in Matlab. Barcode mRNA locations were called when multiple units localized to the same spot. Each barcode mRNA state was determined by looking at the binary state of each of the twenty unit HCR-FISH images. All of the barcodes located in each cell were used to generate a consensus barcode state. Cells with fewer than 50 detected units were excluded from analysis.

Calculation of mutual information between recording units

We pooled the observed states from all 1,453 cells and built a frequency matrix $\Gamma_{3\times10}$ representing each of the 10 recording units and the observed frequency of each one of the three possible intMEMOIR states which define the distribution $P(x)$ per site. For each pair of sites, we then computed the joint distribution $P(x,y)$ from the observed frequencies of pairs of states e.g. $[(1,1),(1,0),(1,2),]$. We then combined the probabilities in Γ with the joint distribution to build a matrix of pairwise Mutual Information using Shannon's formula, using log_3 to normalize the maximum entropy of a single unit to 1 trit.

Lineage analysis of large mES cell colony

Cells were plated on glass bottom 24-well plates (Eppendorf) coated with 20 µg/mL of Laminin-511 overnight. Approximately 1,000 intMEM1 cells were seeded onto the coated wells and induced with 10 µM TMP and 100 ng/mL doxycycline, along with 9,000 parental E14 cells to increase cell density to support growth and survival. Induction lasted 36 hours, followed by approximately 70 hours of growth with no induction. Media was changed daily, and the cells were fixed at the end of the experiment with 4% formaldehyde in PBS for 5 minutes, followed by HCR-FISH protocol described above. Clone boundaries and barcode analysis for this colony were analyzed by hand.

D. memoiphila fly line generation

Fly lines containing UAS-Ceru-10unit and PRExpress-Bxb1-hsp70pA were site-specifically integrated into the attp2 and VK27 sites, respectively, using phiC31 (Bestgene Inc.). Flies with the 10-unit array were first crossed with an nSyb-Gal4 line (R57C10-Gal4, attp40, Bloomington Drosophila stock center) for pan-neuronal expression of the intMEMOIR array. The offspring were then crossed with the PRExpress-Bxb1-hsp70pA to generate the line capable of autonomous recording for downstream experiments. The generated fly lines are available from Bloomington *Drosophila* stock center with the following RRID:

PRExpress-Bxb1-hsp70pA: BDSC_90853
nSyb-Gal4; UAS-Ceru-10unit: BDSC_90854

D. memoiphila characterization

To determine if we could tune the edits in *D. memoiphila* embryo and read out the results in adult fly brains, we placed parents of the PRExpress-Bxb1-hsp70pA x nSyb-Gal4; UAS-Ceru-10unit cross in fresh vials overnight at 25ºC to collect eggs. 3 to 4 hours after removing the parents, the embryos were heat shocked in 37ºC water bath for 30 minutes, 1 hour, and 3 hours for the respective samples. Negative control samples were always kept at 25ºC and not heat shocked. The resulting adult flies were sacrificed, and their brains were dissected in PBS and fixed in 4% paraformaldehyde (PFA) in PBS for 20 minutes. Samples were then washed 3 times with PBS for 10 minutes, transferred to Optimal Cutting Temperature (OCT) compound, and frozen on dry ice. Samples were cut into 20 µm-thick sections on a cryostat and transferred onto coverslips that had been pre-treated with 3-aminopropyltriethoxysilane (Sigma A3648, diluted to 2% v/v in Acetone), followed by post-fix with 4% PFA in PBS for approximately 25 minutes, 3 rinses with PBS, 1 rinse with 70% ethanol, and permeabilized in 70% ethanol overnight at 4ºC. Tissues were then cleared with 8% SDS for 5 minutes at room temperature, rinsed once with PBS, then rinsed 3 times with 70% ethanol. After air drying the samples, we installed SecureSeal Hybridization Chambers (Grace Bio-Labs) onto the coverslips. intMEMOIR array states were then read out with HCR v3.0 as described above (Molecular Instruments) (*42*). Imaging for these samples were done on a Nikon Eclipse Ti inverted microscope, with a spinning disc unit Yokogawa CSU-W1, an electron-multiplying charge-coupled device camera Andor iXon Ultra, and a 40x oil objective (1.3 numerical aperture).

intMEMOIR activity was evaluated by calculating the ratio of inverted-unit-5 to unedited-unit-9. The mES cell HCR-FISH data demonstrated that these two units are efficiently edited (Fig. 2.2I), and the two states were both probed in the same round of HCR-FISH, enabling comparison of the exact same fields of view and eliminating errors or bias that might result from image alignment. Each data point in Fig. 2.5B corresponds to one imaging position.

intMEMOIR labeling of early neuroblast lineages

To label neuroblast lineages at early embryonic stages, parents of the PRExpress-Bxb1-hsp70pA x nSyb-Gal4; UAS-Ceru-10unit cross were placed in fresh vials at 25ºC to collect freshly laid eggs. The vials were inspected every hour and, upon observing egg laying, the parents were removed, and

4 hours later the embryos were heat shocked at 37ºC for 1 hour. The resulting adult flies, up to 1 week old, were incubated at 29ºC overnight to enhance activity of the Gal4 transcription factor on the UAS promoter prior to brain collection and cryosection, followed by smFISH readout (fig. S2.3B and below).

smFISH readout in *D. melanogaster* brain section

To simultaneously examine spatial organization, cell state, and lineage information in the same tissue (Fig. 2.5D to M), the dissected fly brains were cryosectioned and attached onto coverslips treated with 1% bind-silane (GE 17-1330-0) and poly-D-lysine (Sigma P6407), and the resulting samples prepared and analyzed with sequential, automated rounds of smFISH in a manner similar to previously described (*65*). Briefly, sections were post-fixed with 4% PFA at room temperature for 15 minutes, followed by three PBS washes. They were then permeabilized in 70% ethanol (either at 4˚C overnight or at room temperature for 2 hours), cleared with 8% SDS in 1X PBS for 20 minutes at room temperature, then washed with 70% ethanol prior to two rounds of overnight primary probe hybridizations at 37˚C (in order to separate the hybridizations for unedited and inverted units). After each hybridization, samples were washed with 2X SSC for 3 times, incubated in 40% formamide in 2X SSC for 30 minutes at 37˚C, followed by 3 additional rounds of 2X SSC wash. They were then stained with 100 µg/mL Concanavalin A-488 (ThermoFisher) in PBS with 0.1% BSA and 0.1% Triton X-100 for more than 5 hours at room temperature to facilitate segmentation in downstream analysis. After staining, the sample was washed three times with PBS plus 0.5% Triton X-100 (with an extended 5 minute incubation for the final wash), and stained with 10 µg/ml DAPI in 4X SSC for 15 seconds. Next, an anti-bleaching buffer solution made of 10% (w/v) glucose, 1:100 diluted catalase, 0.5 mg/ml glucose oxidase and 50 mM pH 8 Tris-HCL in 4X SSC was flowed through the samples. We used an automated imaging and fluidics delivery system described in (*65*) to image mCerulean transcripts, the intMEMOIR units, and 8 endogenous genes. Two fluorophores, 647 and Cy3B, were used to read out the 29 targets in 15 rounds of hybridizations. The probe sequences and their corresponding readout channels are listed in table S4. The microscope used in this system was a Leica DMi8, with confocal scanner unit Yokogawa CSU-W1, Andor Zyla 4.2 Plus sCMOS camera, 63x oil objective Leica 1.40 NA, and an ASI MS2000 stage. Each field of view was acquired with 1 µm z-steps for 14 or 15 z-slices, across 647 nm, 561 nm, 488 nm, and 405 nm fluorescent

channels. Maximum intensity projections of the slices were used for example images in Figs. 2.5D and S2.11. For each brain, 1 z-slice with in-focus FISH signals (determined with mCerulean FISH signal to avoid bias), was then selected for downstream analysis.

Analysis of smFISH readout in *D. melanogaster* brain section

Fly cells were segmented manually. A custom Matlab program (available at (*63*)) was used to determine the barcode state, as with the mES cells. Clones which had at least one barcode inversion and at least 4 cells were chosen for downstream analysis.

Gene expression analysis in brain section (Brain B1 as the example for the description below)

The brain data set comprises gene expression, location, and intMEMOIR state for 5,332 individual cells. For each cell, we recorded the average pixel value as the expression level for the 8 endogenous genes. To investigate the structure of the gene expression space, we constructed a gene expression matrix $M_{m\times n}$, where n = 8 genes, and m = 5,332. Based on this matrix, the analysis pipeline we built to delineate gene expression clusters consists of several steps: 1. Scaling gene expression using a z-transform, such that all genes have mean=0 and standard deviation=1. 2. Denoising data by applying PCA (*57*) to the scaled data and retaining principal components accounting for 80% of the total variance (6 PCs). 3. The 6-dimensional data were then transformed into a UMAP embedding of 4 dimensions followed by the DBSCAN clustering algorithm (*59*) (sklearn DBSCAN with eps = 0.3) which resulted in a total of 20 distinct clusters. 4. For visualization, the 6-dimensional data set was transformed and projected into 2 dimensions using UMAP (*58*) (default parameters, Python UMAP v0.3.10). 5. Finally, we mapped the gene expression clusters (as color labels) into either the UMAP space (Fig. 2.5G) or physical space (Fig. 2.5I).

Determining the relationship between clonality, physical distance, and gene expression distance

The physical distance was calculated as the Euclidean distance between all pairs of barcoded cells chosen for analysis. This data set was then divided into two groups: pairs within the same clone or pairs from two different clones, and plotted as a cumulative histogram (Fig. 2.5F, 'within clone' and

'between clone', respectively). For gene expression space, the Euclidean distance was calculated between cell pairs using the UMAP coordinates (e.g. Fig. 2.5K and S2.16A). To disentangle the relative contribution of physical distance and lineage to gene expression, the aforementioned data was further binned by the physical Euclidean distance between cell pairs (Fig. 2.5M). Pearson correlation was also used as an alternative metric for gene expression distance (figs. S2.16B and S2.17).

Lineage and statistical analysis

Here we describe procedures for two types of lineage analysis (Fig. 2.1A). First, we discuss assignment of individual cells to clones, i.e. groups of cells that share a common ancestor at the time of editing (clonal classification). Second, we discuss the hierarchical assignment of cells or clones into multi-generational lineage trees (lineage tree reconstruction). In both cases, we describe an analytical framework and experimental validation using the data in Figure 2.3. These data were obtained from experiments in which Bxb1 was expressed for ~3 generations, followed by an additional ~1-2 generations of clonal expansion without Bxb1 induction (Fig. 2.2G).

Clonal classification

intMEMOIR can classify cells into clones based on shared array state inherited from a common ancestor that was uniquely labeled at a specific point in the past. Clonal analysis can be used both to address specific biological questions, and to provide the 'leaves' of more detailed lineage tree reconstruction (below).

Ideally, clonal classification should group cells in such a way that each cell is more closely related to other cells in its own group than to any cell in other groups. In general, a given lineage tree can generate multiple, distinct clonal classifications depending on which edits occurred at what point in the tree. For example, the tree in figure S2.6 could show multiple distinct sets of unique edit patterns (colors), all consistent with the true lineage.

An experimental clonal classification is made in a straightforward way by grouping cells with identical edit patterns into putative clones. To assess the accuracy of such a clonal classification, we

must first determine if it is consistent with the ground truth lineage tree observed by direct time-lapse imaging (Fig. 2.3A), and, second, quantify the number and types of classification errors, if any (Fig. 2.3C).

The following algorithm assigns an accuracy score to a given putative clone, labeled R. To do so, it considers all subtrees (partitions) of the ground truth lineage tree and asks whether any subtree exactly matches the inferred clone. If such a subtree exists, then the clonal classification is considered accurate. If not, we identify the subtree that most closely matches the clone, which we label as S, and quantify its deviation from the putative clone. This deviation is computed by first classifying each cell in S as either a true positive (appears in both S and R), a false positive (appears in R but not S), or a false negative (appears in S, but not R). We then count the number of cells in each of these three categories and compute a clone score:

$$\text{score} = \frac{TP}{TP + FP + FN}$$

Here, TP, FP, and FN denote the number of cells that are true positive, false positive, or false negative, respectively. A higher score indicates a higher fraction of true positive cells (greater accuracy). Results from this analysis are plotted in Fig. 2.3D.

Lineage reconstruction

To reconstruct a multi-generation lineage tree from observed edit patterns, we first develop a relatedness metric for pairs of cells (or clones) based on their edit patterns. The metric is based on the likelihood of a sister relationship. We then use this metric to reconstruct lineage trees in such a way that cells that score higher on this sister likelihood metric are grouped more closely together on the reconstructed tree. Finally, we validate this procedure and quantify its accuracy.

To develop the metric, we start by modeling the molecular events that generate the final edit patterns. We assume each unedited memory element can be edited stochastically at a constant, empirically determined rate per cell generation per memory element, denoted μ_k, where $k = 1..10$ indexes the memory element within the array (see Fig. 2.2I). We also incorporate the empirical transition probabilities for each of the two possible outcomes of each state (Fig. 2.2I). We then define the

probability distribution P_g for observing each of the 10-unit array states that occur in the colony of interest starting from an unedited array, after g generations. P_g is computed for each cell in the colony, and represents the probability of observing that cell's specific array state, independent of the states of other cells in the colony.

Next, we define the pairwise distance metric for a single memory element. We denote the conditional probability of observing any two specific memory states in a pair of sister cells as $P_g^{sis}(i,j)$, where i and j index two specific, different cells $(i \neq j)$. To convert this probability into a distance metric, we need to normalize it by comparing the likelihood of observing these two memory states in a pair of sister cells to the likelihood of them occurring independently in two unrelated cells. That is, we define the distance metric as

$$d_{i,j} \equiv \frac{P_g(i)P_g(j)}{P_g^{sis}(i,j)} .$$

Memory units edit independently (Fig. 2.2J). Therefore, it is possible to extend this distance metric for a single memory element to the level of a complete array in a straightforward manner, by replacing the single unit probabilities with products over all the units:

From this we obtain the $K = 10$ unit array distance metric:

$$d_{i,j}^{array} \equiv \frac{\prod_{k=1}^{10} P_g(i)_k P_g(j)_k}{\prod_{k=1}^{10} P_g^{sis}(i,j)_k}$$

Deriving probability distributions for the intMEMOIR system

This distance metric is independent of many details of the recording system. To apply it to intMEMOIR data, we first need to derive expressions for the distributions P_g and P_g^{sis}. The recording units have an initial state, denoted *1*, that can be edited irreversibly into either of two states, denoted *0* and *2*. The probability that a given unit is edited during a cell division is μ_k, and the probability that no edit happens during a cell division is $(1-\mu_k)$. For simplicity, we first derive P_g assuming only two possible states: unedited (1) and edited (0). For a given unit, the probability that no editing happens for *g* generations (cell divisions) is then

$$P_g(1) = (1-\mu)^g \quad (1)$$

The probability that an edit occurred at some point in the past is defined by the geometric distribution:

$$P_g(0) = \sum_{g=1}^{G} (1-\mu)^{g-1}\mu \quad (2)$$

This expression considers all possible times at which the edit could have happened, e.g. in the first generation, the second generation, or even in the last generation. Once the edit occurs, the unit can no longer be edited. (Below, we will extend this analysis to the case of multiple edit outcomes.)

By applying the geometric series, we can show that P_g is well defined as a probability distribution for all values of *g* such that:

$$P_g(0) + P_g(1) = 1 \quad (3)$$

We derive Eq. 3 by first expanding Eq. 2:

$$P_g(0) = \mu + (1-\mu)\mu + (1-\mu)^2\mu + (1-\mu)^3\mu + ... + (1-\mu)^{g-1}\mu$$

Combining eqs. 1 and 2 we obtain the total probability:

$$P_g(0) + P_g(1) = \mu[1 + (1-\mu) + (1-\mu)^2 + ... + (1-\mu)^{g-1}] + (1-\mu)^g$$

We then use the following identity for the geometric series:

$$\sum_{k=0}^{n-1} qw^k = q\frac{1-w^k}{1-w}$$

By setting $q = 1$, $w = 1-\mu$ and $k = g$, we obtain:

$$P_g(0) + P_g(1) = \mu\frac{1-(1-\mu)^g}{\mu} + (1-\mu)^g = 1$$

Which shows that P_g is well-defined for all values of g.

Three-state model

We now extend the model by considering three possible edit outcomes: $\{1,0,2\}$. The probabilities of observing the recording unit in each of three possible states at generation g become:

$$P_g(1) = (1-\mu)^g$$

$$P_g(0) = \sum_{g=1}^{G}(1-\mu)^g \mu\alpha$$

$$P_g(2) = \sum_{g=1}^{G}(1-\mu)^g \mu(1-\alpha) \quad (4)$$

Here, α denotes the probability of an edited unit going to state *0* and $1-\alpha$ is the probability of it reaching state 2. Note that, in a similar way, this framework could also be generalized to larger numbers of editing outcomes. The transition probability distribution $P(z_{g-1} \to i)$ represents all the ways in which an individual unit can change state during a single cell division cycle:

$$P(1 \to 1) = 1-\mu$$

$$P(1 \to 0) = \mu\alpha$$

$$P(1 \to 2) = \mu(1-\alpha) \quad (5)$$

$$P(0 \to 0) = 1$$

$$P(2 \to 2) = 1$$

The probabilities of all other transitions are zero, consistent with the irreversibility of intMEMOIR editing. Further, these transition rates are assumed to be time independent.

Sister likelihood

We now have all elements necessary to compute the sister likelihood scores that are the basis of our distance metric. We calculate the conditional probability that an unobserved parental state z in the

previous generation transitioned into the observed states i, j. For simplicity, first consider just a single cell state, i:

$$P_g(i|z_{g-1}) = P(z_{g-1} \rightarrow i|z_{g-1})P_{g-1}(z) \tag{6}$$

Where $P_{g-1}(z)$ is the probability of observing the parental state z in the previous generation and can be calculated using Eq. 4. Considering now two cells i, j, this transition probability becomes the joint distribution:

$$P_g(i, j|z_{g-1}) = P(z_{g-1} \rightarrow i|z_{g-1})P(z_{g-1} \rightarrow j|z_{g-1})P_{g-1}(z) \tag{7}$$

Eq. 7 provides the probability that the observed states i, j came from the unobserved parental state z. Since we don't actually observe z, we need to account for all possible parental states to obtain the total sister probability P_g^{sis}:

$$P_g^{sis}(i, j) = \sum_{z=0}^{2} P_g(i, j|z_{g-1}) \tag{8}$$

Finally, considering an array of k independent units we extend this and calculate the product:

$$P_g^{sis}(C_i, C_j) = \prod_{k=1}^{10} P_g^{sis}(i, j)_k \tag{9}$$

Where C_i, C_j represent a specific pair of array states, e.g. in two different cells or clones.

Hierarchical lineage tree

The calculations above enable us to compute the pairwise distance matrix $d_{i,j}^{array}$, defined above, for any actual data set. As the final step in reconstruction, we built a dendrogram from $d_{i,j}^{array}$ by applying divisive clustering, a top-down approach in which all observations start in one cluster, and two-way splits are performed recursively as one moves down the lineage tree, terminating at the leaves (individual cells or clones). Divisive clustering was implemented with the *DIANA* function from the **R** *cluster* package. Examples of the resulting trees are shown in Figures 2.3E to 2.3G. A complete list of reconstructed trees is provided in table S2 in the Newick tree format.

Assessing tree accuracy

To quantitatively assess reconstruction accuracy compared to ground truth (Fig. 2.3H), we used the Robinson-Foulds distance metric, as implemented in the **R** *phytools* package, to compute the difference between the reconstructed and ground truth trees. For this analysis, all cells sharing the same array state are collapsed into a single clonal tree "leaf." Occasionally, analysis of ground truth trees revealed convergent edits producing identical array states in distantly related cells (false positive events; see discussion of clonal analysis, above). These events prevent one from unambiguously collapsing identical array states in the ground truth tree. In these cases, we randomly retain one of the array states.

Calculation of entropy in colonies

In applications where no ground truth is available, we would like to develop a predictive metric that could be used to enrich for colonies that are likely to reconstruct with greater accuracy. Multiple variables, including spatial arrangements of cells or morphological similarity could in principle be informative. However, we reasoned that the most useful and generalizable metric would be one based only on the observed edit patterns, since this information should be available in all applications independent of systems-specific biological features.

Shannon's entropy provides an ideal metric to quantify information content in a discrete data set such as a list of edit states. To apply it to colonies, we pooled the edit states from all cells within each colony. We then constructed a 10×3 matrix, Γ representing the frequency of observing each of the 10 recording units in each of the 3 edit states. We can then apply Shannon's formula to each unit to obtain its individual entropy, e.g. $H_k = -\sum_{i=0}^{2} \Gamma_{i,k} \log \Gamma_{i,k}$ is the entropy of the k^{th} unit. The total entropy of the colony is then obtained by adding the entropies of the individual sites: $H_{colony} = \sum_{k=1}^{10} H_k$. Finally, we scaled the entropy by the fraction of edited sites in the colony, π, to obtain an informative score of expected reconstruction quality. We confirmed that selecting colonies with higher normalized entropy enriched for better reconstruction (Figs. 2.3H and S2.8).

2.8 Supplementary Figures

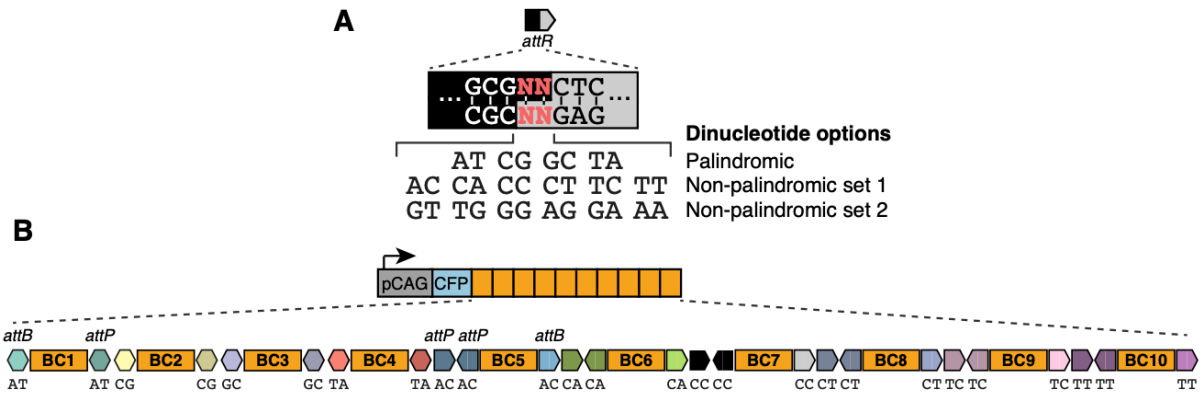

Fig. S2.1. 10 central dinucleotide variants in the *att* site enable the arrangement of 10 independent memory units in one array.
(**A**) The central dinucleotide (red) of *attP/B* form base pairs during recombination, dictating specificity and orthogonality of the sites. In principle, out of the 16 possible dinucleotide combinations, 10 could confer orthogonality in an array: four are palindromic, and the two sets of six non-palindromic dinucleotides are reverse complements of one another, so only one set could be used. (**B**) Schematic of the 10-unit array as designed with annotated *att* sites and barcodes (BC).

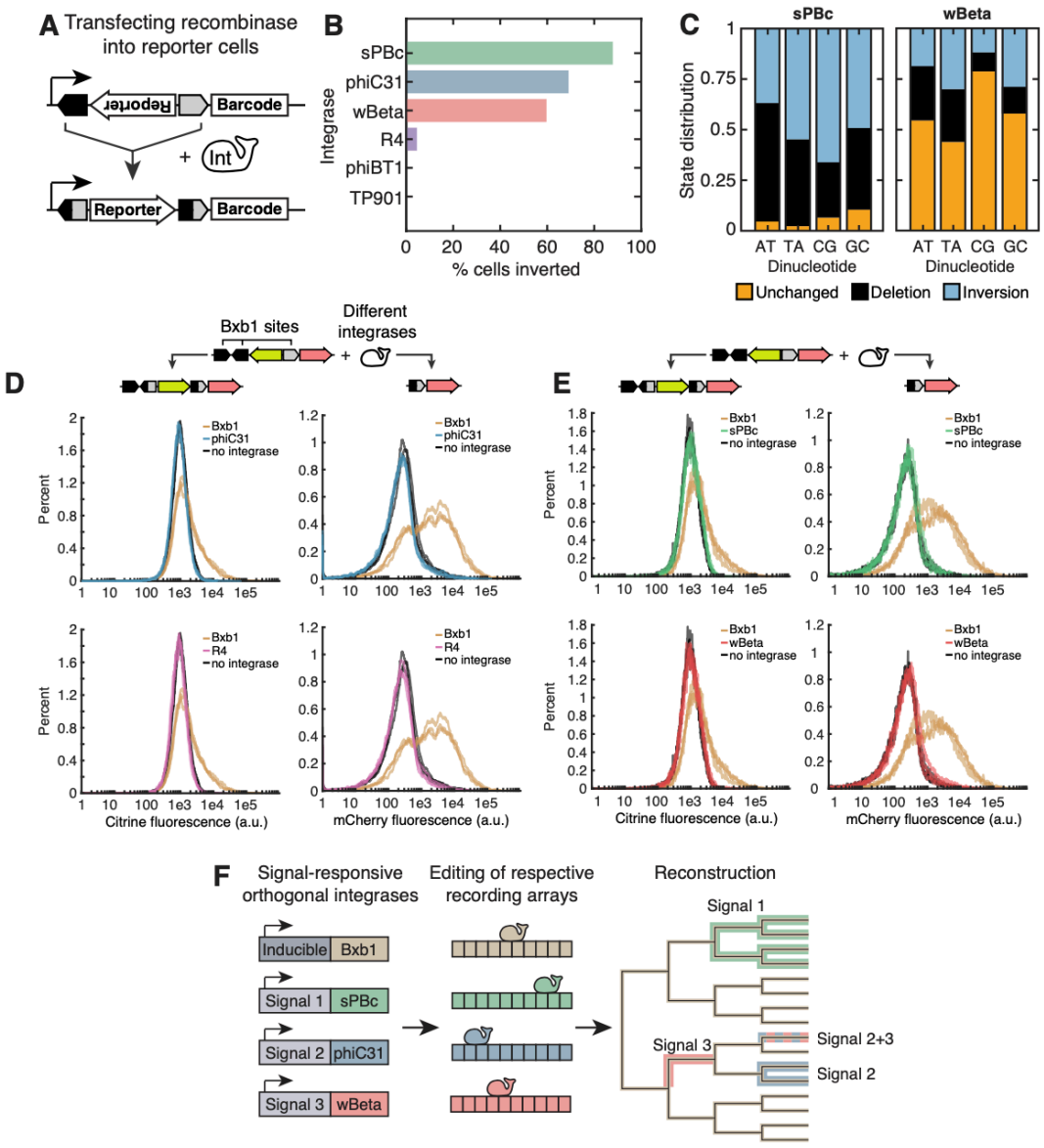

Fig. S2.2. Additional members of the serine integrase family function in mES cells.
(**A**) To assess the activity of different serine integrases in mES cells (*30*), we integrated reporter constructs with an *attP/B* flanked unit followed by a barcode. Active integrase inverts the unit upon transfection, which can be detected via HCR-FISH. (**B**) Percentage of cells with inverted reporter after transfection. sPBc, phiC31, and wBeta are active in mES cells (n=91, 165, and 139 cells, respectively); R4 show weak activity (n=88 cells), and no activity was detected for phiBT1 and TP901 (n=172 and 171 cells, respectively). (**C**) Additional serine integrases can mediate inversion and deletion between *att* sites with palindromic dinucleotides. sPBc and wBeta reporter cell lines with a 4 unit array were transfected with their corresponding integrases, and their relative edit frequencies analyzed via HCR-FISH (n=307 and 264 cells, respectively). (**D** and **E**) Fluorescent reporter assay demonstrates that other serine integrases operate orthogonally to Bxb1 *att* sites, opening the possibility for orthogonal recording in the same cell. This is true for phiC31 and R4, as shown in (D), and wBeta and sPBc, as shown in (E), with n=3 for each sample. (**F**) Schematic illustrating simultaneous lineage tracking and signal recording using orthogonal serine integrases.

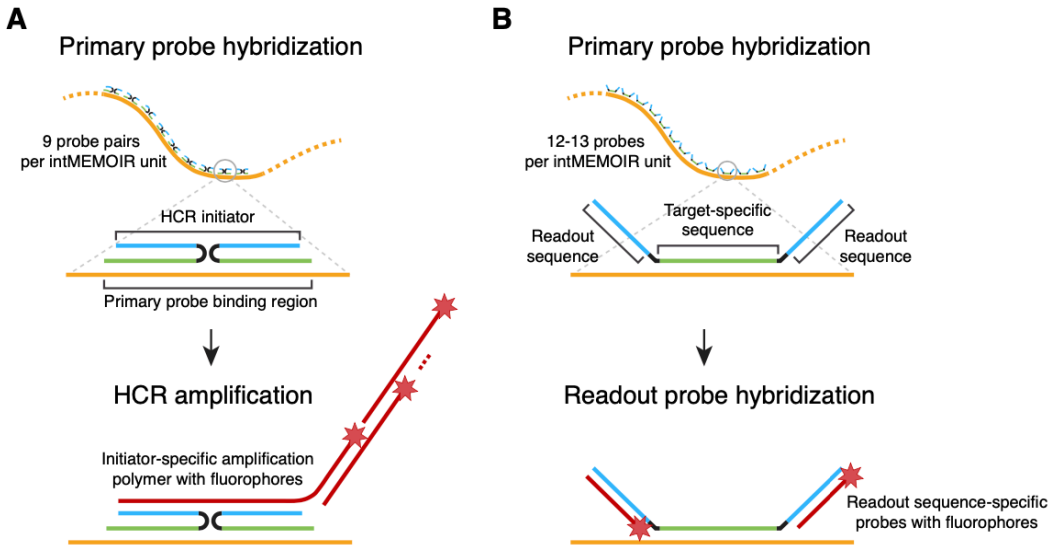

Fig. S2.3. Two fluorescent in situ hybridization methods were used to read the intMEMOIR array.
(**A**) We used HCR v3.0 (*42*) to read the array state in the intMEM1 experiments (Figs. 2.2, G to J, 3, 4, S4, and S5) and to determine the effect of heat shock duration on *D. memoiphila* editing (Fig. 2.5B). 9 probe pairs were used for each unit, and the signals were subsequently amplified through HCR. The primary probe binding regions of each unit, the probe pairs' HCR initiator ID, and the fluorophore used in the corresponding amplification hairpins are listed in table S3. We also used the HCR-FISH method, with different barcodes and corresponding probes, to test additional members of the serine integrase family (fig. S2.2, A to C) (**B**) We used automated smFISH (*65*) to read the array state and endogenous genes in *D. memoiphila* experiments (Fig. 2.5C onward, and S2.11 onward). 12-13 primary probes were used for each intMEMOIR unit, 15 probes were used for mCerulean, and 24 probes were used for each endogenous gene. The target-specific sequences, readout sequences and ID, and the fluorophore used with the corresponding readout probes are listed in table S4.

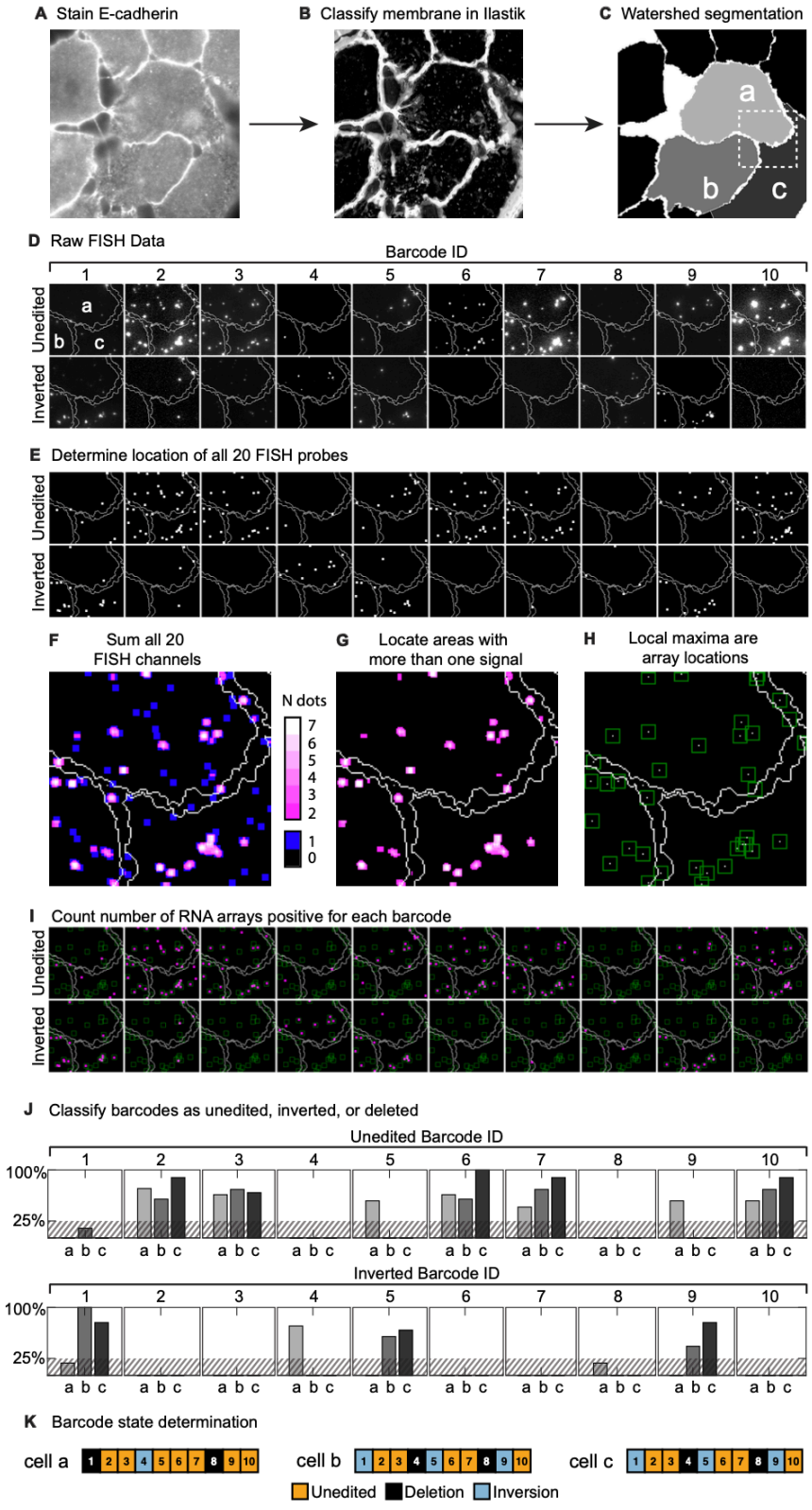

Fig. S2.4. Cell segmentation and barcode determination.
(*Figure legend continued on next page*)

Here we show the workflow for classifying array states in individual cells. (**A**) In order to segment individual cells, we acquired image stacks of cells stained with an E-cadherin antibody that localizes to cell membranes. (**B**) We trained the image analysis program Ilastik to classify pixels in these images as membrane or non-membrane. (**C**) A 3-dimensional watershed algorithm (Matlab) was used to segment cells from the pixel classification stack, using the final annotated cell positions from the movie as watershed seeds. Obvious segmentation errors were fixed by adding additional seeds and cutting joined cells. After the watershed, stacks were converted to 2-dimensional images by maximum projection. Here, cells 'a', 'b' and 'c' were shaded and labeled for subsequent panels. (**D**) Maximum intensity projections of the images for the 20 array channels. (**E**) RNA molecule locations were detected for each HCR-FISH probe set by finding the local maximum after applying a Laplacian filter. The points were dilated to account for small errors in localization. (**F**) Dilated points from all 20 channels were summed to identify locations with two or more barcodes as barcode arrays. Blue locations contain only a single detected dot and were discarded. (**G**) Locations with multiple detected unit reads are shown in magenta and white, and considered validated array signals. (**H**) Each local maximum was designated as an array location. For visualization, green boxes were drawn around the array locations. (**I**) The presence or absence of HCR-FISH signal (magenta dots) was determined for each barcode array location (green boxes) and tallied for each cell. Cells with greater than 50 validated unit reads were retained for downstream analysis. (**J**) Cells were considered positive for a given unit when a signal was present in at least 25% of the array locations in the cell. (**K**) From these results, the final array states were determined for each cell.

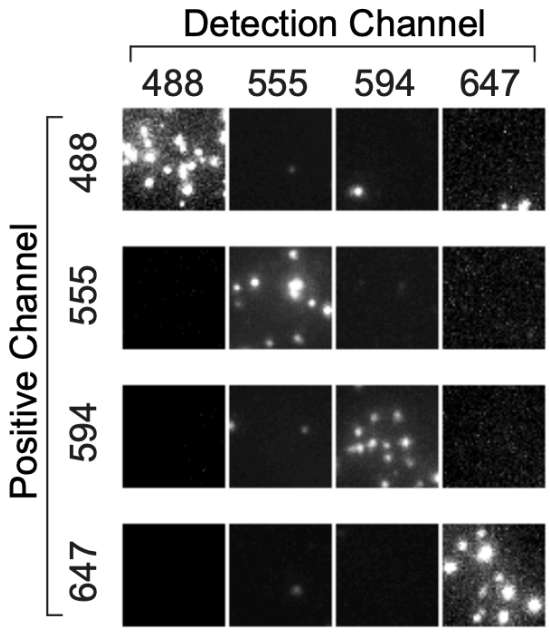

Fig. S2.5. The four HCR-FISH fluorescent channels show minimal crosstalk.
Each row shows a portion of an intMEM1 cell positive for a single channel. Only the true positive channel shows significant fluorescent signal above background. Fluorescent crosstalk would have appeared as signal in the negative channel, at the same location as the true positive channel. Each column is shown with the same exposure, brightness, and contrast.

Fig. S2.6. A given lineage tree can generate multiple, distinct clonal classifications.
Seven possible subtree assignments can accurately describe this simple five-cell tree, where all cells within a clone are more closely related to each other than to any cell outside of the clone. In one extreme, each individual cell can be classified as its own clone. At the other extreme, in the absence of editing, all cells are grouped into a single clone.

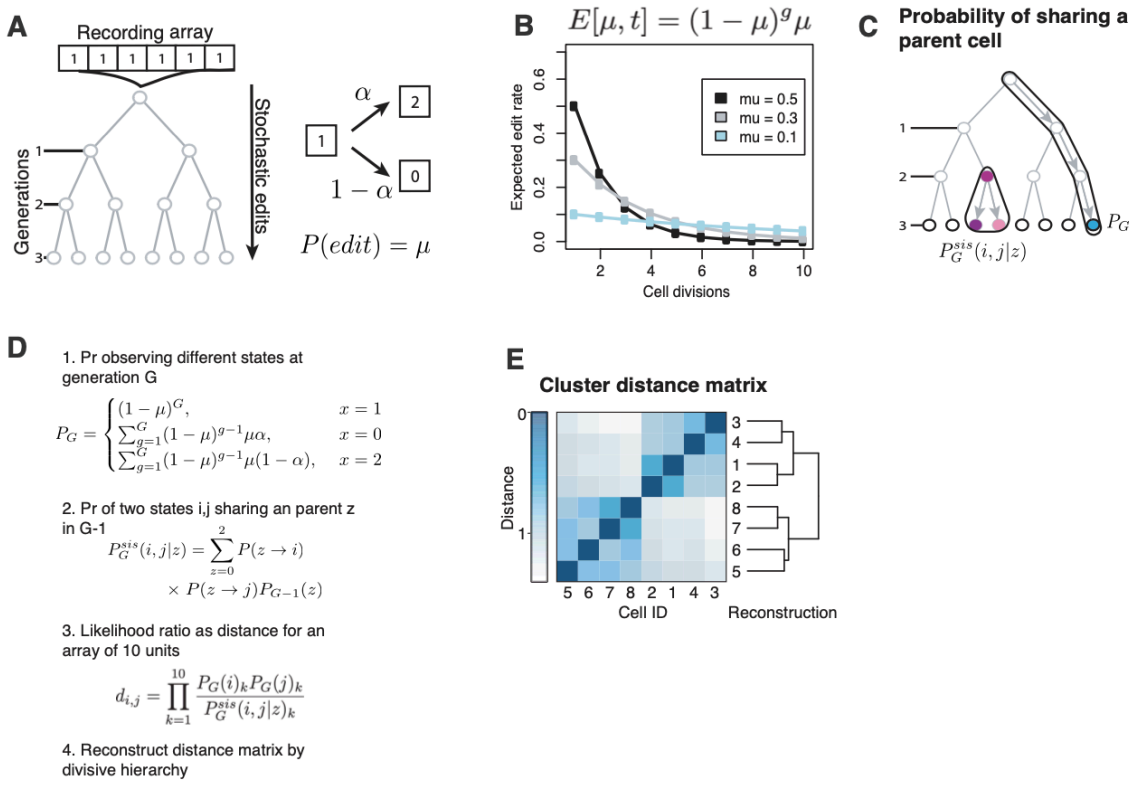

Fig. S2.7. Lineage simulation and reconstruction based on maximum likelihood of sister relationships.

(**A**) Schematic of the lineage simulation method. We define a two-parameter stochastic model where μ equals the edit rate in units of edits per site per generation, and α denotes the probability that the edit goes to state 2**.** The model assumes that cells divide synchronously and at a constant rate such that at generation G the lineage comprises 2^G cells. (**B**) For a constant μ, the number of new edits appearing in each generation decays exponentially as dictated by the equation $E[\mu, t] = (1-\mu)^g \mu$ where g is the number of generations (cell divisions) and $E[\mu, t]$ is the expected fraction of edited sites. For intMEMOIR, the experimental value of μ is ~0.1-0.3 (**C**) Schematic of the reconstruction approach. We first compute the probability that a trit is in either of the three possible states at generation G, combining the transition probabilities shown in A and the equation from B and called this distribution P_G (blue cell). For sister likelihood, we compute the probability that two cells (i, j) share a parent z in the previous generation. (**D**) Equations for reconstructing lineages based on sister likelihood. 1. The probability that a recording unit is in either of the three possible states at generation G, independently of the other cells. 2. The probability that a parent cell z at $G-1$ transitions into the states i, j. This equation assumes that the daughter cells (i, j) inherit the state of z and then edit with probability μ. Since the recording is irreversible, the only valid transitions are $1 \to 0$ and $1 \to 2$; once a cell reaches either state 2 or 0, all its daughters will inherit that state with Pr =1. We finally sum over all possible states of the parent cell z. 3. We can then calculate the joint probability for the 10 units as the product of the probabilities of each unit. And compare this number to the probabilities of observing the states (i, j) assuming no sister relationship, which are just the product of their P_G probabilities in the numerator. This ratio quantifies the likelihood of observing a given pair of array states for two sister cells compared to two unrelated cells. 4. This likelihood provides a pairwise distance metric that we then use to reconstruct the lineage tree. (**E**) Once we computed the likelihood

ratio for all pairs of cells, we can cluster the matrix using divisive hierarchical clustering, which starts by partitioning the data set into the most distinct groups, then it proceeds to partition each subgroup into two groups iteratively until each group contains only one cell. Ideally, each partition of the algorithm would correspond to a cell division event.

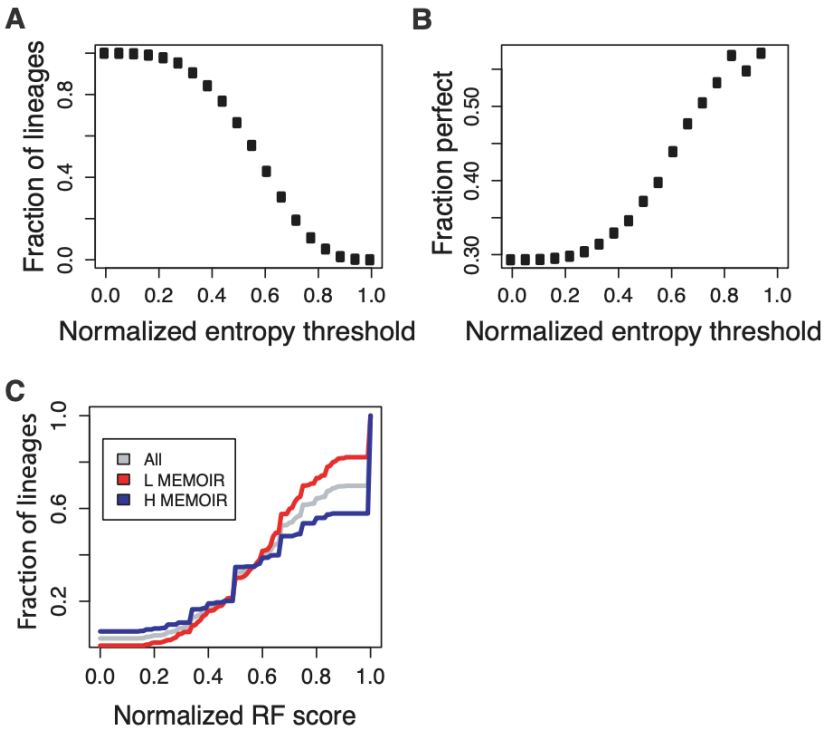

Fig. S2.8. Barcode entropy enriches for colonies that reconstruct with greater accuracy. We computed the entropy for a lineage as the sum of the individual entropies for each trit, using Shannon's formula. The normalized entropy is then computed as the lineage's entropy times the fraction of edited sites for that lineage, scaled by the maximum such that the metric has a range from [0,1]. This simulated dataset comprises 3000 lineages. (**A**) The fraction of lineages with normalized entropy larger than the threshold. (**B**) The fraction of perfectly reconstructed lineages for increasing thresholds of normalized entropy. For a given threshold value, we split the dataset and calculated the fraction of perfect trees in the high-entropy set. Note that the number of lineages analyzed decreases with increasing entropy thresholds, as shown in (A). (**C**) As an example, using a threshold of 0.6, we obtain a high-accuracy set of colonies that exhibit a fraction of perfect trees > 0.4 (compare high entropy colonies, 'H MEMOIR', with low entropy colonies, 'L MEMOIR'). Note that the threshold is arbitrary and can be tuned to maximize the numbers of colonies and minimize the false discovery rate.

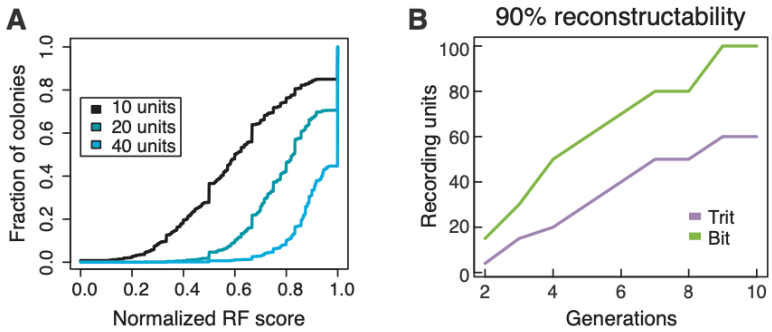

Fig. S2.9. Additional intMEMOIR arrays increase reconstruction accuracy and depth. (**A**) Accuracy in the reconstruction of simulated lineages using increasing numbers of recording units. Parameters were estimated from experimental data. The structure of the lineage trees used in the simulation are those observed experimentally. Using 40 units arranged as 4 intMEMOIR arrays, more than 50% of lineages can be reconstructed perfectly. (**B**) For a given accuracy (90%), the number of recording units necessary for reconstruction scales with the depth of the lineage. The calculation assumes binary lineages with no cell death.

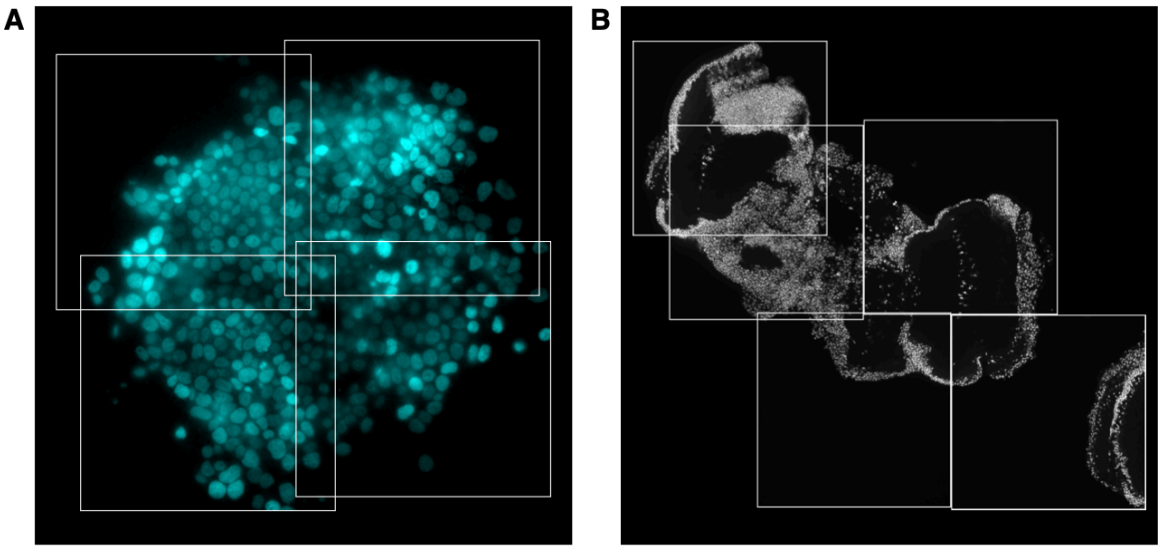

Fig. S2.10. Example of stitched microscope images.
(**A**) Data shown in Figure 2.4 were derived from four overlapping microscope positions (white squares) that were digitally combined. (**B**) Microscope images shown in Figure 2.5 and S2.11 were similarly derived from five positions (white squares).

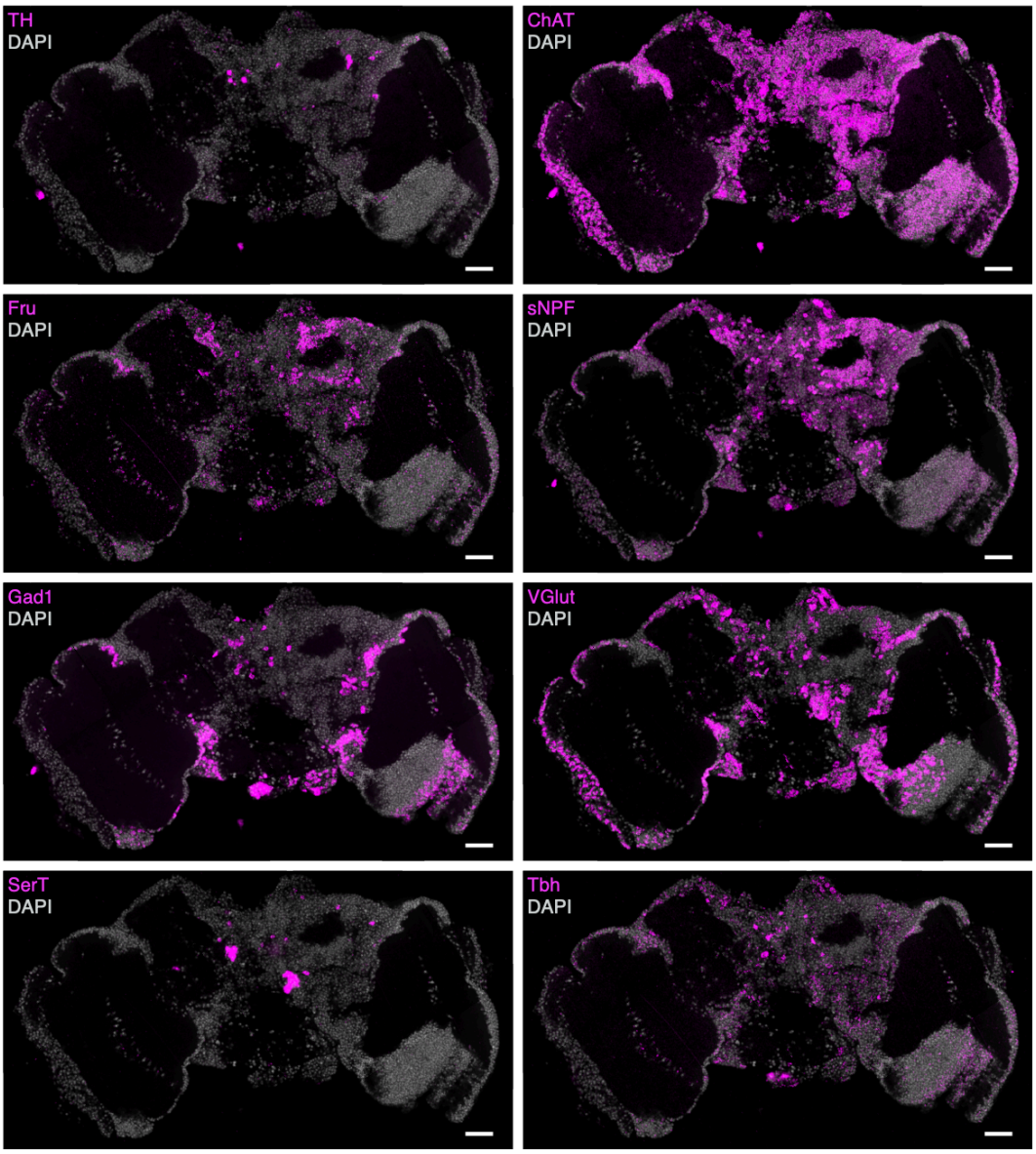

Fig. S2.11. Probing the expression of 8 endogenous genes in an adult *Drosophila* brain section with smFISH.
In addition to the intMEMOIR array, we probed for the expression of 8 endogenous genes in the same brain section: tyrosine hydroxylase (TH), choline acetyltransferase (ChAT), fruitless (Fru), short neuropeptide F precursor (sNPF), glutamic acid decarboxylase (Gad1), vesicular glutamate transporter (VGlut), serotonin transporter (SerT), and tyramine ß-hydroxylase (Tbh). Endogenous genes and DAPI signals are shown in magenta and gray, respectively (scale bar, 30 μm).

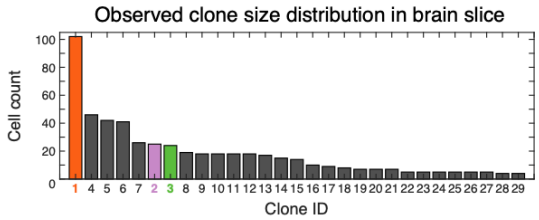

Fig. S2.12. intMEMOIR recovers clone sizes *in vivo*.
Cells in the adult *Drosophila* brain section were segmented, their array states determined (see Materials and Methods), and clones with ≥four cells and at least one unit inverted were chosen for downstream analysis. Clones 1, 2, and 3, as shown in Fig. 2.5E, are highlighted in their corresponding colors.

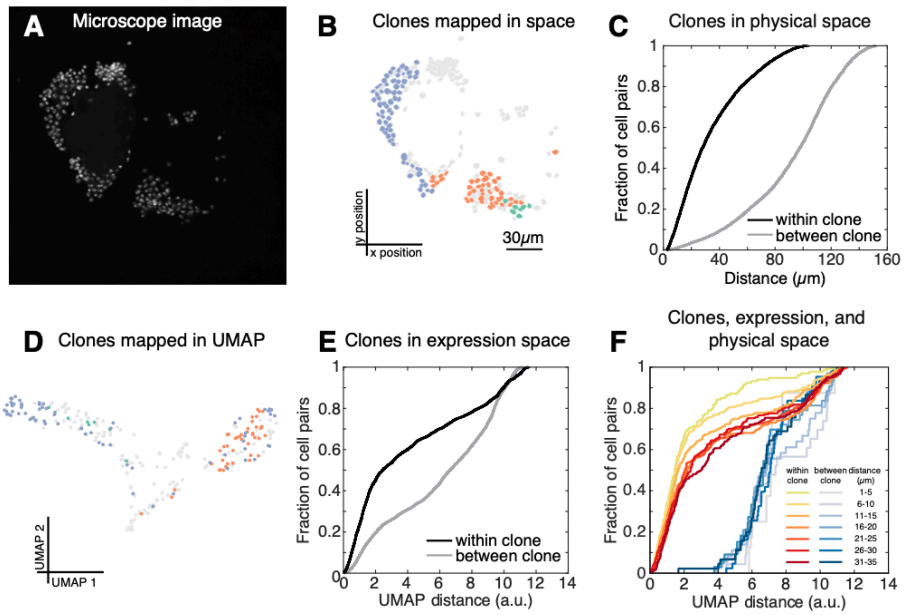

Fig. S2.13. Brain B2, a section of D. melanogaster antenna lobe.
(**A**) A section of brain B2 stained with DAPI. (**B** and **C**) Cells in the same clone were closer in physical space than cells in different clones, as seen on the spatial map (B) and in the cumulative distributions (C). In (B), segmented cells are colored by the analyzed clones. Grey cells were excluded from analysis (see Materials and Methods) (scale bar, 30 μm). (**D** and **E**) Cells within a clone were more similar in gene expression space than cells in different clones, as seen on the UMAP (D) and in the cumulative distributions (E). (**F**) Within a clone, but not between different clones, cell pairs exhibited spatially graded cell type similarity (cf. Figure 2.5).

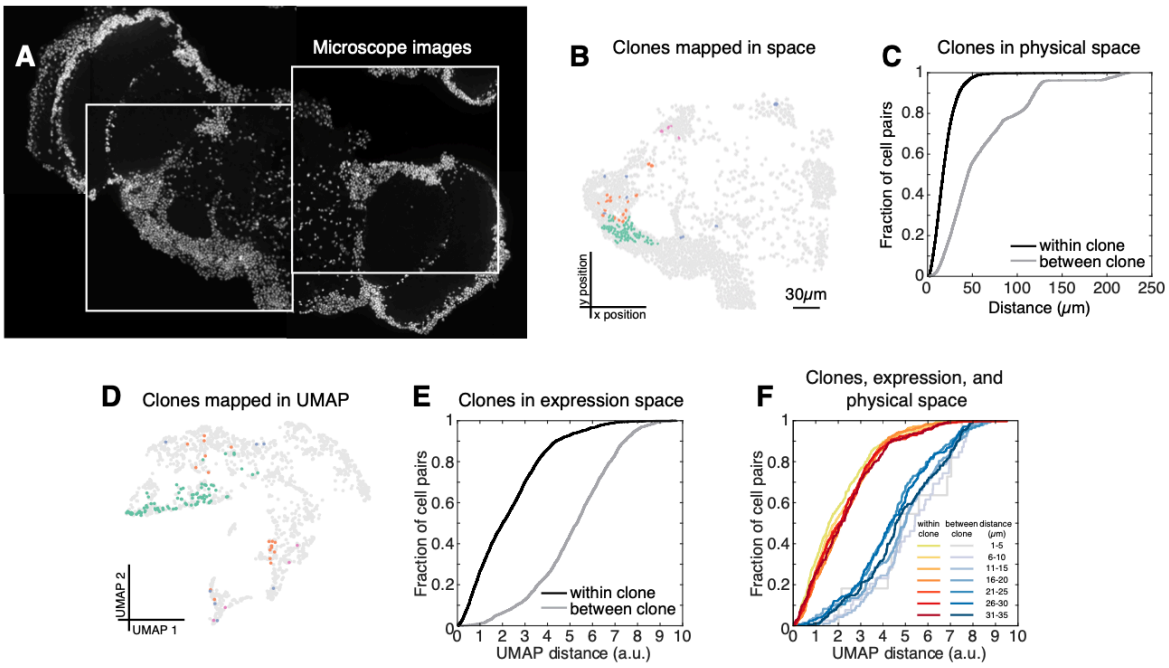

Fig. S2.14. Brain B3, a section of a D. melanogaster brain.
(**A**) A section of brain B3 stained with DAPI. Downstream analyses were done on stitched images from two microscope positions (white squares). (**B** and **C**) Cells in the same clone were closer in physical space than cells in different clones, as seen on the spatial map (B) and in the cumulative distributions (C). In (B), segmented cells are colored by the analyzed clones. Grey cells were excluded from analysis (see Materials and Methods) (scale bar, 30 μm). (**D** and **E**) Cells within a clone were more similar in gene expression space than cells in different clones, as seen on the UMAP (D) and in the cumulative distributions (E). (**F**) This brain does not show a spatially graded dependence on cell fate similarity within clones, likely due to its reduced diversity of captured cell types (fig. S2.17). Cf. Figure 2.5.

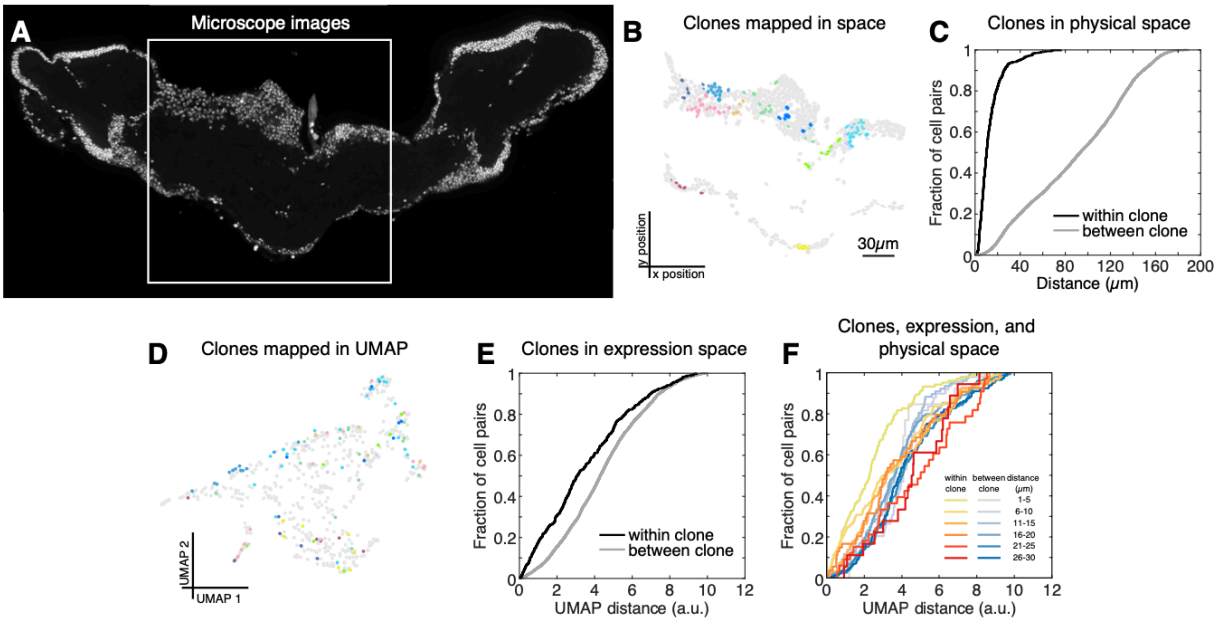

Fig. S2.15. Brain B4, a section of a D. melanogaster brain.
(**A**) A section of brain B4 stained with DAPI. Downstream analyses were done on a single microscope position capturing the central brain region (white square). (**B** and **C**) Cells in the same clone were closer in physical space than cells in different clones, as seen on the spatial map (B) and in the cumulative distributions (C). In (B), segmented cells are colored by the analyzed clones. Grey cells were excluded from analysis (see Materials and Methods) (scale bar, 30 μm). (**D** and **E**) Cells within a clone were more similar in gene expression space than cells in different clones, as seen on the UMAP (D) and in the cumulative distributions (E). (**F**) Within a clone, but not between different clones, cell pairs exhibited spatially graded cell type similarity (cf. Figure 2.5).

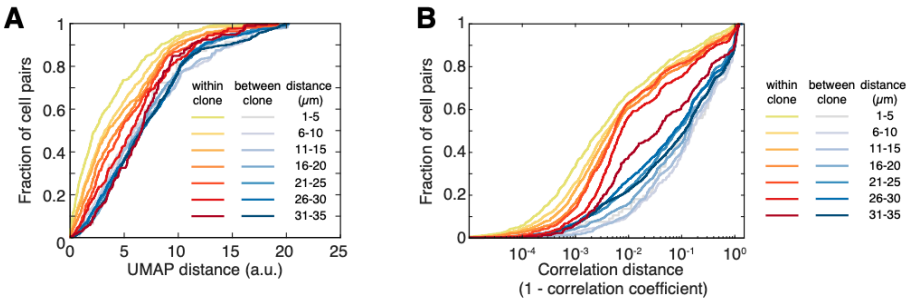

Fig. S2.16. Intra-lineage spatial distribution of cells predicts fate similarity in subsamples or with alternative distance metric.

(**A**) To determine if the relationship observed in Fig. 2.5M was skewed by the large Kenyon cell clone, we repeated the analysis omitting those cells. The results still showed a strong role for lineage in cell fate determination at close distances. (**B**) Clonal dependence of cell type similarity at short distances is observed when we use Pearson correlation to measure gene expression similarity, demonstrating that it is robust to the choice of distance metric.

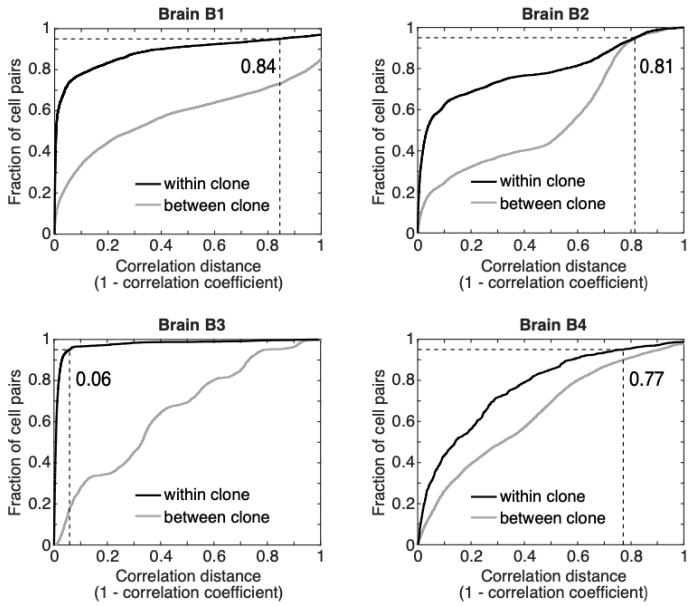

Fig. S2.17. Brain B3 displays the least diversity within clones.
The Pearson correlation was used to determine the amount of diversity within and between clones in each brain and is shown as a cumulative histogram. Here, the correlation distance is defined as 1 - correlation coefficient. For Brain B3, 95% of cell pairs within the captured clones have a correlation distance of less than 0.06, showing very little diversity within clones. 95% of cell pairs within clones in Brains B1, B2, and B4 have correlation distances of less than 0.84, 0.81, and 0.77, respectively.

2.9 Online Supplementary Materials

Supplementary materials below are available online as separate files at:

K.-H. K. Chow, M. W. Budde, A. A. Granados, M. Cabrera, S. Yoon, S. Cho, T.-H. Huang, N. Koulena, K. L. Frieda, L. Cai, C. Lois, M. B. Elowitz, Imaging cell lineage with a synthetic digital recording system. Science. 372 (2021), doi:10.1126/science.abb3099.

Table S1. List of constructs used in the manuscript.

Table S2. Ground truth and reconstructed lineage trees.

Table S3. HCR probe binding regions and information.

Table S4. Automation smFISH probe sequences and information.

Movies S1 to S3. Time-lapse imaging and tracking of intMEM1 cells used for lineage reconstruction in Figure 3, E, F, and G. Cells were imaged as shown in Fig. 3A. Ground truth lineage trees were constructed by manually tracking the cells in the time-lapse images using a modified version of the EasyTrack software developed by Yaron Antebi (freely available at (*63*) and https://github.com/AntebiLab/EasyTrack/tree/Memoir). Tracked cells and division events are indicated with open and filled squares, respectively. Cells were primarily tracked by their CFP fluorescence. Arbitrary cell numbers in the final frame were assigned for subsequent analysis.

2.10 References

1. C. Blanpain, B. D. Simons, Unravelling stem cell dynamics by lineage tracing. *Nat. Rev. Mol. Cell Biol.* 14, 489–502 (2013).
2. K. Kretzschmar, F. M. Watt, Lineage tracing. *Cell*. 148, 33–45 (2012).
3. M. B. Woodworth, K. M. Girskis, C. A. Walsh, Building a lineage from single cells: genetic techniques for cell lineage tracking. *Nat. Rev. Genet.* 18, 230–244 (2017).
4. S.-H. S. Wu, J.-H. Lee, B.-K. Koo, Lineage Tracing: Computational Reconstruction Goes Beyond the Limit of Imaging. *Mol. Cells*. 42, 104–112 (2019).
5. C. S. Baron, A. van Oudenaarden, Unravelling cellular relationships during development and regeneration using genetic lineage tracing. *Nat. Rev. Mol. Cell Biol.* 20, 753–765 (2019).
6. D. Frumkin, A. Wasserstrom, S. Kaplan, U. Feige, E. Shapiro, Genomic variability within an organism exposes its cell lineage tree. *PLoS Comput. Biol.* 1, e50 (2005).

7. S. J. Salipante, M. S. Horwitz, Phylogenetic fate mapping. *Proc. Natl. Acad. Sci. U. S. A.* 103, 5448–5453 (2006).

8. N. Navin, J. Kendall, J. Troge, P. Andrews, L. Rodgers, J. McIndoo, K. Cook, A. Stepansky, D. Levy, D. Esposito, L. Muthuswamy, A. Krasnitz, W. R. McCombie, J. Hicks, M. Wigler, Tumour evolution inferred by single-cell sequencing. *Nature*. 472, 90–94 (2011).

9. S. Behjati, M. Huch, R. van Boxtel, W. Karthaus, D. C. Wedge, A. U. Tamuri, I. Martincorena, M. Petljak, L. B. Alexandrov, G. Gundem, P. S. Tarpey, S. Roerink, J. Blokker, M. Maddison, L. Mudie, B. Robinson, S. Nik-Zainal, P. Campbell, N. Goldman, M. van de Wetering, E. Cuppen, H. Clevers, M. R. Stratton, Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature*. 513, 422–425 (2014).

10. M. A. Lodato, M. B. Woodworth, S. Lee, G. D. Evrony, B. K. Mehta, A. Karger, S. Lee, T. W. Chittenden, A. M. D'Gama, X. Cai, L. J. Luquette, E. Lee, P. J. Park, C. A. Walsh, Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science*. 350, 94–98 (2015).

11. R. U. Sheth, H. H. Wang, DNA-based memory devices for recording cellular events. *Nat. Rev. Genet.* 19, 718–732 (2018).

12. L. S. Ludwig, C. A. Lareau, J. C. Ulirsch, E. Christian, C. Muus, L. H. Li, K. Pelka, W. Ge, Y. Oren, A. Brack, T. Law, C. Rodman, J. H. Chen, G. M. Boland, N. Hacohen, O. Rozenblatt-Rosen, M. J. Aryee, J. D. Buenrostro, A. Regev, V. G. Sankaran, Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. *Cell*. 176, 1325–1339.e22 (2019).

13. H. Zong, J. S. Espinosa, H. H. Su, M. D. Muzumdar, L. Luo, Mosaic analysis with double markers in mice. *Cell*. 121, 479–492 (2005).

14. J. Livet, T. A. Weissman, H. Kang, R. W. Draft, J. Lu, R. A. Bennis, J. R. Sanes, J. W. Lichtman, Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature*. 450, 56–62 (2007).

15. H.-H. Yu, C.-H. Chen, L. Shi, Y. Huang, T. Lee, Twin-spot MARCM to reveal the developmental origin and identity of neurons. *Nat. Neurosci.* 12, 947–953 (2009).

16. S. D. Perli, C. H. Cui, T. K. Lu, Continuous genetic recording with self-targeting CRISPR-Cas in human cells. *Science*. 353 (2016), doi:10.1126/science.aag0511.

17. A. McKenna, G. M. Findlay, J. A. Gagnon, M. S. Horwitz, A. F. Schier, J. Shendure, Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*. 353, aaf7907 (2016).

18. W. Pei, T. B. Feyerabend, J. Rössler, X. Wang, D. Postrach, K. Busch, I. Rode, K. Klapproth, N. Dietlein, C. Quedenau, W. Chen, S. Sauer, S. Wolf, T. Höfer, H.-R. Rodewald, Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. *Nature*. 548, 456–460 (2017).

19. R. Kalhor, K. Kalhor, L. Mejia, K. Leeper, A. Graveline, P. Mali, G. M. Church, Developmental

barcoding of whole mouse via homing CRISPR. *Science*. 361 (2018), doi:10.1126/science.aat9804.

20. W. Tang, D. R. Liu, Rewritable multi-event analog recording in bacterial and mammalian cells. *Science*. 360 (2018), doi:10.1126/science.aap8992.

21. A. Alemany, M. Florescu, C. S. Baron, J. Peterson-Maduro, A. van Oudenaarden, Whole-organism clone tracing using single-cell sequencing. *Nature*. 556, 108–112 (2018).

22. M. M. Chan, Z. D. Smith, S. Grosswendt, H. Kretzmer, T. M. Norman, B. Adamson, M. Jost, J. J. Quinn, D. Yang, M. G. Jones, A. Khodaverdian, N. Yosef, A. Meissner, J. S. Weissman, Molecular recording of mammalian embryogenesis. *Nature*. 570, 77–82 (2019).

23. K. L. Frieda, J. M. Linton, S. Hormoz, J. Choi, K.-H. K. Chow, Z. S. Singer, M. W. Budde, M. B. Elowitz, L. Cai, Synthetic recording and in situ readout of lineage information in single cells. *Nature*. 541, 107–111 (2017).

24. I. Salvador-Martínez, M. Grillo, M. Averof, M. J. Telford, Is it possible to reconstruct an accurate cell lineage using CRISPR recorders? *Elife*. 8 (2019), doi:10.7554/eLife.40292.

25. W. M. Stark, M. R. Boocock, D. J. Sherratt, Catalysis by site-specific recombinases. *Trends Genet.* 8, 432–439 (1992).

26. M. C. A. Smith, R. Till, M. C. M. Smith, Switching the polarity of a bacteriophage integration system. *Mol. Microbiol.* 51, 1719–1728 (2004).

27. J. Bonnet, P. Subsoontorn, D. Endy, Rewritable digital data storage in live cells via engineered control of recombination directionality. *Proc. Natl. Acad. Sci. U. S. A.* 109, 8884–8889 (2012).

28. K. Rutherford, P. Yuan, K. Perry, R. Sharp, G. D. Van Duyne, Attachment site recognition and regulation of directionality by the serine integrases. *Nucleic Acids Res.* 41, 8341–8356 (2013).

29. P. C. M. Fogg, S. Colloms, S. Rosser, M. Stark, M. C. M. Smith, New applications for phage integrases. *J. Mol. Biol.* 426, 2703–2716 (2014).

30. Z. Xu, L. Thomas, B. Davies, R. Chalmers, M. Smith, W. Brown, Accuracy and efficiency define Bxb1 integrase as the best of fifteen candidate serine recombinases for the integration of DNA into the human genome. *BMC Biotechnol.* 13, 87 (2013).

31. C. A. Merrick, J. Zhao, S. J. Rosser, Serine Integrases: Advancing Synthetic Biology. *ACS Synth. Biol.* 7, 299–310 (2018).

32. B. P. Zambrowicz, A. Imamoto, S. Fiering, L. A. Herzenberg, W. G. Kerr, P. Soriano, Disruption of overlapping transcripts in the ROSA geo 26 gene trap strain leads to widespread expression of -galactosidase in mouse embryos and hematopoietic cells. *Proceedings of the National Academy of Sciences*. 94 (1997), pp. 3789–3794.

33. P. Soriano, Generalized lacZ expression with the ROSA26 Cre reporter strain. *Nat. Genet.* 21, 70–71 (1999).

34. M. Sadelain, E. P. Papapetrou, F. D. Bushman, Safe harbours for the integration of new DNA in the human genome. *Nat. Rev. Cancer*. 12, 51–58 (2011).

35. H. A. Grunwald, V. M. Gantz, G. Poplawski, X.-R. S. Xu, E. Bier, K. L. Cooper, Super-Mendelian inheritance mediated by CRISPR-Cas9 in the female mouse germline. *Nature*. 566, 105–109 (2019).

36. I. Espinosa-Medina, J. Garcia-Marques, C. Cepko, T. Lee, High-throughput dense reconstruction of cell lineages. *Open Biol.* 9, 190229 (2019).

37. P. Ghosh, A. I. Kim, G. F. Hatfull, The orientation of mycobacteriophage Bxb1 integration is solely dependent on the central dinucleotide of attP and attB. *Mol. Cell*. 12, 1101–1111 (2003).

38. S. D. Colloms, C. A. Merrick, F. J. Olorunniji, W. M. Stark, M. C. M. Smith, A. Osbourn, J. D. Keasling, S. J. Rosser, Rapid metabolic pathway assembly and modification using serine integrase site-specific recombination. *Nucleic Acids Res.* 42, e23 (2014).

39. L. Madisen, A. R. Garner, D. Shimaoka, A. S. Chuong, N. C. Klapoetke, L. Li, A. van der Bourg, Y. Niino, L. Egolf, C. Monetti, H. Gu, M. Mills, A. Cheng, B. Tasic, T. N. Nguyen, S. M. Sunkin, A. Benucci, A. Nagy, A. Miyawaki, F. Helmchen, R. M. Empson, T. Knöpfel, E. S. Boyden, R. C. Reid, M. Carandini, H. Zeng, Transgenic mice for intersectional targeting of neural sensors and effectors with high specificity and performance. *Neuron*. 85, 942–958 (2015).

40. M. Iwamoto, T. Björklund, C. Lundberg, D. Kirik, T. J. Wandless, A general chemical method to regulate protein stability in the mammalian central nervous system. *Chem. Biol.* 17, 981–988 (2010).

41. Detailed materials and methods are available as supplementary materials.

42. H. M. T. Choi, M. Schwarzkopf, M. E. Fornace, A. Acharya, G. Artavanis, J. Stegmaier, A. Cunha, N. A. Pierce, Third-generation in situ hybridization chain reaction: multiplexed, quantitative, sensitive, versatile, robust. *Development*. 145 (2018), doi:10.1242/dev.165753.

43. D. F. Robinson, L. R. Foulds, Comparison of phylogenetic trees. *Math. Biosci.* 53, 131–147 (1981).

44. K. Sugino, J. Garcia-Marques, I. Espinosa-Medina, T. Lee, Theoretical modeling on CRISPR-coded cell lineages: efficient encoding and optimal reconstruction. *bioRxiv* (2019), p. 538488.

45. J. M. Kebschull, A. M. Zador, Cellular barcoding: lineage tracing, screening and beyond. *Nat. Methods*. 15, 871–879 (2018).

46. T. Lee, Wiring the Drosophila Brain with Individually Tailored Neural Lineages. *Curr. Biol.* 27, R77–R82 (2017).

47. R. Urbach, G. M. Technau, Neuroblast formation and patterning during early brain development in Drosophila. *Bioessays*. 26, 739–751 (2004).

48. S. R. Spindler, V. Hartenstein, The Drosophila neural lineages: a model system to study brain development and circuitry. *Dev. Genes Evol.* 220, 1–10 (2010).

49. A. C. Groth, M. Fish, R. Nusse, M. P. Calos, Construction of transgenic Drosophila by using the site-specific integrase from phage phiC31. *Genetics*. 166, 1775–1782 (2004).

50. O. Riabinina, D. Luginbuhl, E. Marr, S. Liu, M. N. Wu, L. Luo, C. J. Potter, Improved and expanded Q-system reagents for genetic manipulations. *Nat. Methods*. 12, 219–222 (2015).

51. A. Akmammedov, M. Geigges, R. Paro, Single vector non-leaky gene expression system for Drosophila melanogaster. *Sci. Rep.* 7, 6899 (2017).

52. K. Dumstrei, F. Wang, C. Nassif, V. Hartenstein, Early development of the Drosophila brain: V. Pattern of postembryonic neuronal lineages expressing DE-cadherin. *J. Comp. Neurol.* 455, 451–462 (2003).

53. K. Ito, W. Awano, K. Suzuki, Y. Hiromi, D. Yamamoto, The Drosophila mushroom body is a quadruple structure of clonal units each of which contains a virtually identical set of neurones and glial cells. *Development*. 124, 761–771 (1997).

54. S.-L. Lai, T. Awasaki, K. Ito, T. Lee, Clonal analysis of Drosophila antennal lobe neurons: diverse neuronal architectures in the lateral neuroblast lineage. *Development*. 135, 2883–2893 (2008).

55. M. Ito, N. Masuda, K. Shinomiya, K. Endo, K. Ito, Systematic analysis of neural projections reveals clonal composition of the Drosophila brain. *Curr. Biol.* 23, 644–655 (2013).

56. H.-H. Yu, T. Awasaki, M. D. Schroeder, F. Long, J. S. Yang, Y. He, P. Ding, J.-C. Kao, G. Y.-Y. Wu, H. Peng, G. Myers, T. Lee, Clonal development and organization of the adult Drosophila central brain. *Curr. Biol.* 23, 633–643 (2013).

57. I. T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments. *Philos. Trans. A Math. Phys. Eng. Sci.* 374, 20150202 (2016).

58. E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. H. Kwok, L. G. Ng, F. Ginhoux, E. W. Newell, Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* (2018), doi:10.1038/nbt.4314.

59. M. Ester, H.-P. Kriegel, J. Sander, X. Xu, Others, in *Kdd* (1996), vol. 96, pp. 226–231.

60. F. Zhu, M. Gamboa, A. P. Farruggio, S. Hippenmeyer, B. Tasic, B. Schüle, Y. Chen-Tsai, M. P. Calos, DICE, an efficient system for iterative genomic editing in human pluripotent stem cells. *Nucleic Acids Res.* 42, e34 (2014).

61. S. Ausländer, M. Fussenegger, From gene switches to mammalian designer cells: present and future prospects. *Trends Biotechnol.* 31, 155–168 (2013).

62. A. Askary, L. Sanchez-Guardado, J. M. Linton, D. M. Chadly, M. W. Budde, L. Cai, C. Lois, M. B. Elowitz, In situ readout of DNA barcodes and single base edits facilitated by in vitro

transcription. *Nat. Biotechnol.* 38, 66–75 (2020).

63. K.-H. Chow, M. Budde, A. Granados, M. Cabrera, S. Yoon, S. Cho, T.-H. Huang, N. Koulena, K. Frieda, L. Cai, C. Lois, M. Elowitz, Data for "Imaging cell lineage with a synthetic digital recording system" (2020) (Version 1.0) [Data set]. CaltechDATA. https://doi.org/10.22002/D1.1444.

64. S. Berg, D. Kutra, T. Kroeger, C. N. Straehle, B. X. Kausler, C. Haubold, M. Schiegg, J. Ales, T. Beier, M. Rudy, K. Eren, J. I. Cervantes, B. Xu, F. Beuttenmueller, A. Wolny, C. Zhang, U. Koethe, F. A. Hamprecht, A. Kreshuk, ilastik: interactive machine learning for (bio)image analysis. *Nat. Methods*. 16, 1226–1232 (2019).

65. C.-H. L. Eng, M. Lawson, Q. Zhu, R. Dries, N. Koulena, Y. Takei, J. Yun, C. Cronin, C. Karp, G.-C. Yuan, L. Cai, Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature*. 568, 235–239 (2019).

66. S. Yamaguchi, Y. Kazuki, Y. Nakayama, E. Nanba, M. Oshimura, T. Ohbayashi, A method for producing transgenic cells using a multi-integrase system on a human artificial chromosome vector. *PLoS One*. 6, e17267 (2011).

Chapter 3

BUILDING SERINE INTEGRASE-BASED, IMAGE-READABLE RECORDING SYSTEMS

3.1 There are several design principles for a serine integrase based, image-readable recording system

We have built the foundation of intMEMOIR, and it is now poised at an exciting junction where future work can increase its lineage reconstruction potential, introduce more recording channels, or employ it to answer biological questions. The goal of this chapter is to provide essential information for these developments. I will discuss unpublished design principles that we learned while developing intMEMOIR and end the chapter with suggestions for future directions.

Most of the topics I touch on are outlined in Figure 3.1A and many of them are centered around the goal of achieving robust expression of the array in mammalian cells. Because our system relies on single molecule FISH readout, we require larger recording units than most sequencing-based technologies to ensure that sufficient numbers of probes can bind to each target and generate a strong signal. At the same time, strong expression (i.e. many transcripts) reduces the chance of miscalls from false positives and negatives FISH dots (Fig. 3.1B). However, we quickly discovered that it was nontrivial to transcribe a large, non-coding array in mammalian cells. Thus, we embarked on a journey to improve the array expression and, in the process, discovered a number of principles that I hope would be useful not only for intMEMOIR, but also for the design of future synthetic recording constructs (Fig. 3.1C).

It is worth noting that many experiments described in this chapter were done during the exploratory phase of the intMEMOIR project and some were, in retrospect, not optimally controlled. Nevertheless, I will describe our interpretation of the results and, when relevant, point out their caveats.

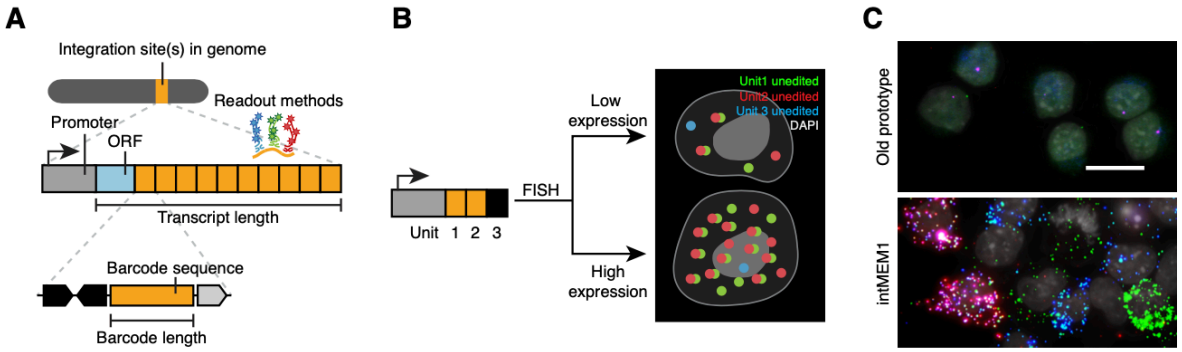

Fig 3.1. Design details significantly impact the recording system's performance.
(**A**) A number of factors can affect the expression level and overall performance of serine integrase-based recording arrays. (**B**) Increasing the array transcript count reduces the frequency of miscalls that result from FISH dots with false positives or false negatives signals. (**C**) Improving the array design can significantly improve its expression levels in mammalian cells. (scale bar, 25 μm). The top panel is an old prototype probed by traditional smFISH and the bottom panel is the final intMEM1 array probed by HCR-FISH.

3.2 The promoter affects the array's expression level and inter-unit deletion frequency

The promoter is the first and foremost consideration when designing a synthetic array. To summarize our experience: a single strong and inducible promoter would be ideal for the array expression.

The 10-unit array is strongly expressed by the CAG and UAS promoters in mammalian cells and *Drosophila*, respectively. TRE, a great choice due to its inducibility, and EFα1 promoters also robustly expressed shorter arrays in mammalian cells but were not tested for the full length 10-unit construct. Other candidates that showed robust expression in the literature, such as UbC, may be worth investigating if additional promoters are needed (*1*). Further, while developing the original MEMOIR system, we learned that CMV promoters tend to produce heterogeneous and bursty expressions. As such, they should be avoided for future designs.

3.2.1 Transcriptional interference may preclude the use of tandem or bidirectional arrays

When using PolII promoters, there should only be one promoter per integration site. For any synthetic system, reducing the number of necessary integrations increases germline heritability and system portability. Thus, we considered the possibility of increasing intMEMOIR memory by

integrating multiple tandem arrays into the same locus of the cell (multiple arrays are discussed in Chapter 3.6.1). However, in that configuration, transcriptional interference is an important factor to consider (*2*). Typically, small arrays (=< four units) are very well expressed in mammalian cells. However, when two 2-unit arrays (with one static barcode each) are expressed in tandem, each array's expression becomes poor, and some cells appear to exclusively express only one of the two arrays (Fig. 3.2A). On rare occasions, we also observed cells with transcripts that showed all three FISH signals colocalized in the same dot, suggesting that they existed in one long, run-on transcript. Note that this experiment was performed on a polyclonal line, so we cannot rule out the possibility of incorrect integration. Nonetheless, the decrease in expression levels observed in this preliminary experiment was quite striking, and, combined with existing literature on transcriptional interference and numerous anecdotal encounters by we and others in the lab, we decided to avoid the use of tandem PolII arrays.

We also encountered transcriptional interference with bidirectional TRE (biTRE) [sequence available at (*3*)]. We first tested this promoter as a possible alternative to the tandem arrangement. To do so, we created a polyclonal stable line with a biTRE fluorescent test construct in the TIGRE locus, expressing citrine and mCherry from the two directions (Fig. 3.2B, top). Upon overnight induction with 100ng/mL of doxycycline, we observed strong expression from both directions within the same cells by flow cytometry (Fig. 3.2B, bottom).

Encouraged by this result, we next tested intMEMOIR arrays. To our surprise, when we performed this test with a biTRE 4x2-unit array using HCR-FISH readout, the expression profiles were poor. Most of the transcripts were retained in the nucleus and, similar to the tandem arrays, some cells appear to exclusively express only one direction of the construct (Fig. 3.2C). We suspected the lack of cytoplasmic RNA may be due to poor RNA stability (potentially caused by the lack of a 5' H2B-mCerulean we typically place before our array; see Chapter 3.4). Thus, the biTRE promoter may be alternating its transcription between the two directions and, due to the poor stability of the RNA, we rarely capture cells with both transcripts at the same time.

If our suspicions are correct, it would explain why the stable protein reporters did not show exclusive expression in one direction (Fig. 3.2B), and highlight the fact that fluorescent proteins may not be the ideal reporters to test the stability and expression levels of different array designs. Also note,

however, that that this experiment was performed on a polyclonal line for Rosa26 integration, which has a lower success rate than the Flp-recombinase-mediated insertion into the TIGRE locus used in Figure 3.2B, so our observations could be due to artifacts of incorrect integration. Nevertheless, based on this preliminary result, we placed the biTRE design on hold. Overall, the tandem array and biTRE results are also the reasons why, despite the difficulty of transcribing a long synthetic array, we did not split them into smaller, independent expression arrays.

3.2.2 Co-transcriptional editing increases the rate of crosstalk

Lastly, co-transcriptional editing increases the rate of inter-unit deletions, so a promoter that is off during editing and on before imaging would be ideal for future systems. As discussed in Chapter 2.3.1, the intMEMOIR units operate largely orthogonally due to the unique central dinucleotides of their *att* sites. However, occasional instances of cross-talk still occur and mostly manifest as inter-unit deletions instead of inversions. Puzzled by this observation, we hypothesized that the RNA polymerase may prevent the integrase from successfully completing the recombination reaction after its initial DNA cleavage. For example, the polymerase could "knock off" the DNA-integrase complex mid-recombination before the *attL/R* sites are successfully ligated; or, in the case of failed recombination between orthogonal sites, before the *attP/B* sites have a chance to reform (*4*).

To begin testing this hypothesis, we integrated a 4-unit array driven by the TRE promoter into the TIGRE locus. This array contains a static barcode labelled UnitS without its own *attP/B* pairs , which should always remain intact in the absence of crosstalk (Fig. 3.2D). Using the polyclonal stable line, we transfected the cells with either Tet3G alone, Tet3G + Bxb1, or Tet3G + Bxb1 + dox, which correspond to no editing, editing without transcription, and co-transcriptional editing, respectively. The constructs are designed such that Tet3G will be stably integrated via piggyBac, while Bxb1 should only be expressed transiently after transfection.

Approximately five days later, we induced each condition with 100ng/mL of doxycycline overnight to turn on the array transcription, prior to fixing and analysis with HCR-FISH targeting unedited Units 1, S, 3, and 4 (Fig. 3.2D, left). Interestingly, when we counted the percentage of CFP positive cells without unedited UnitS, we observed that co-transcriptional editing (condition 3) resulted in the highest rate of crosstalk (Fig. 3.2D, middle). To reduce the chance of counting cells with no array

expression in the polyclonal line, we also performed the analysis after gating for cells that are expressing at least one of the unedited units we FISHed for. The observation becomes even more pronounced under this criteria (Fig. 3.2D, right). Thus, based on these results, we concluded that co-transcriptional editing likely increases the rate of crosstalk, and that future designs of intMEMOIR should aim to use promoters that are off during recording, and only turned on immediately prior to or even after sample fixation (*5*).

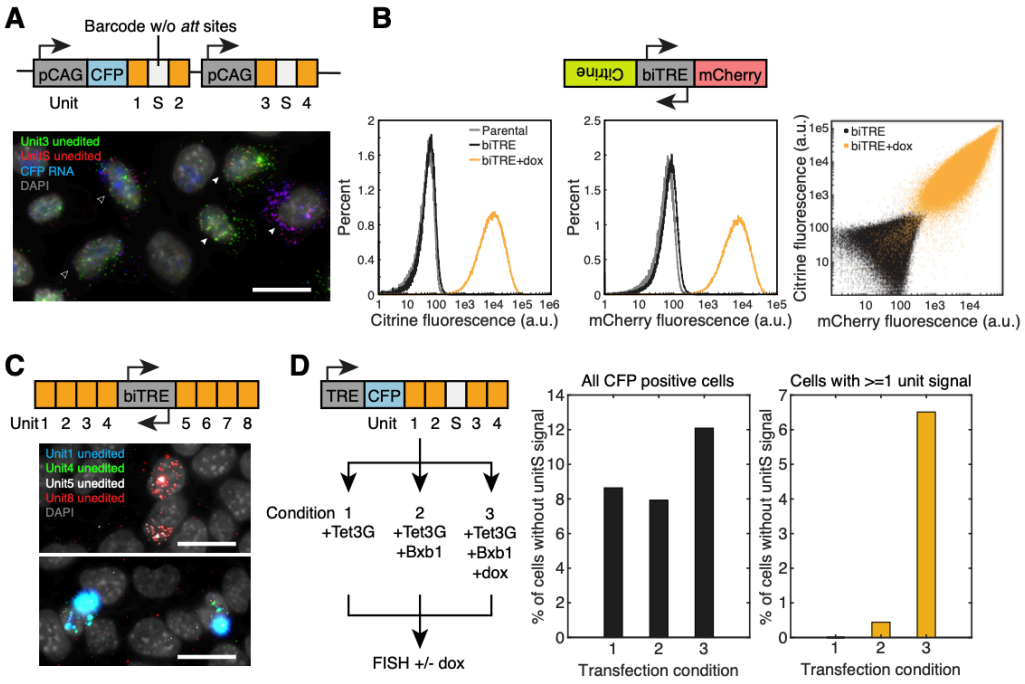

Fig. 3.2. Promoter choice influences array performance.
(**A**) Tandem arrays are poorly expressed, likely due to transcriptional interference. Here, the construct is integrated into a polyclonal stable cell line into the Rosa26 locus. The indicated units are then read out through traditional smFISH (scale bar, 25 μm). Filled arrowheads indicate cells that are expressing from only one of the two arrays, and empty arrowheads indicate cells that are expressing both arrays poorly. (**B**) Fluorescent reporter suggests that bidirectional TRE promoter is tightly off in the absence of doxycycline, and can be induced to express the proteins from both directions. The experiment was done on a polyclonal stable line with the construct integrated into the TIGRE locus. (**C**) Bidirectional TRE promoter appears to suffer from transcriptional interference, as suggested by FISH results on a polyclonal cell line with a 4x2-unit cell line integrated into the Rosa26 locus. Further, the 4-unit arrays are poorly expressed in these cells, potentially due to the lack of a 5' H2B-mCerulean. (**D**) Co-transcriptional editing increases the rate of crosstalk. The experiment was done by transfecting a polyclonal stable cell line with TRE-4-unit integrated into the TIGRE locus.

3.3 The barcodes affect the array's integration efficiency, readout, edit rate, and expression level

The overall length of the array affects its integration efficiency and expression level, while the length of the individual barcodes determine the FISH methods we could use. Furthermore, the sequence of the barcode can also significantly impact the expression of the array. Given these considerations, I recommend using the shortest possible barcode that could still produce a robust FISH signal and efficient editing. Finally, the barcode sequence should be pre-screened *in silico* to avoid sequence similarity with endogenous genes, premature polyadenylation signals, and potential splice sites.

3.3.1 The length of the array affects its integration efficiency and expression level

The length of the array is inversely correlated with its integration efficiency. Over the course of developing intMEMOIR, we have integrated a range of constructs into the Rosa26 locus, with insert sizes that ranged from approximately 4.5kb to 11kb. These integrations were done with CRISPR-mediated homology directed repair. Although no quantitative data is available, screening for monoclonal cell lines with the correct integration has consistently been more difficult with the larger constructs. Even when the clone's successful integration is confirmed on the 5' and 3' ends by PCR, subsequent FISH experiments can show that only a part of the construct is transcribed, suggesting a truncated insert. Further, many conventional *in vivo* transgenic methods such as viral delivery and pronuclear injections have either strict size limitations or decreased success rate with larger constructs (*6*). Thus, for the ease of implementing intMEMOIR in new systems, future versions should strive to reduce the size of the insert and/or consider the use of alternative integration strategies (*6*, *7*).

Beyond integration efficiency, longer arrays also have lower expression levels. We tested this hypothesis by cloning non-coding regions of various lengths after an H2B-mCerulean, mimicking the structure of our recording arrays (Fig. 3.3A, top). We then transiently transfected these constructs into mES cells with mCherry as the cotransfection marker. The resulting CFP fluorescence, gated on high mCherry cells, are then used as a measure of the transcript expression level and stability. From the results, we see a clear decrease in median CFP fluorescence with increasing 3' UTR length (Fig. 3.3A, bottom). Note, however, that the sequences chosen for the 3' UTR were not the same nor

controlled for each tested length. As such, the sequence itself could also contribute to variations in expression levels. Nonetheless, this result is consistent with the increasing difficulty to achieve robust expression of longer arrays. Thus, this again argues for a smaller array design for future systems.

3.3.2 The length of the barcodes determines the system's compatible readout methods and potentially the unit edit efficiency

While decreasing the size of the barcode could resolve many of the above concerns, we also need to ensure that it retains a sufficient number of probe binding sites to produce a robust FISH signal for downstream analysis. The exact number of probe binding sites required will depend on the barcode sequence and the readout methods. Of the three methods that we have used with intMEMOIR, traditional smFISH, seqFISH, and HCR-FISH produced the weakest to strongest signals, respectively (Table 3.1). With seqFISH and HCR-FISH, we read out 500 bp barcodes, while smFISH were used for barcodes around 900 bp. The original MEMOIR system also used smFISH for barcodes as short as 400bp, but the signal to noise ratio was suboptimal (*8*). Overall, even shorter lengths than 500 bp may be achievable with more optimized target sequences and signal amplification offered by seqFISH and HCR-FISH.

Another concern for shortening the barcode is the possibility that the DNA length between *attP/B* pairs may affect the integrase's edit efficiency. To invert or delete DNA, serine integrase dimers bind to *attP* and *attB* sites, which in turn bind to each other to form a tetrameric complex that mediates recombination (*9*). Thus, theoretically, sufficient length of DNA must exist between the *attP/B* pairs to avoid steric hindrance to the complex formation. We have never tested for this lower limit, and we observed efficient editing and near 1 to 1 inversion to deletion ratio in our 500 bp units (Fig. 2.2I). However, if future systems seek to lower the length of the barcodes, it would be interesting and important to observe its effect on edit efficiencies.

FISH method	Signal	Speed	# of rounds	Automation
Traditional smFISH	+	++	9	No
HCR-FISH	+++	+	5	No
seqFISH	++	+++	15	Yes

Table 3.1. The three different FISH methods used in MEMOIR and intMEMOIR development have different strengths and weaknesses.
Of the three methods, automated seqFISH offered the best throughput, while HCR-FISH may be preferable when significant signal amplifications are needed (e.g. small barcode, tissue with high background). Note that the "# of rounds" column indicates the number of sequential rounds of FISH that we have used for our analyses and not the upper limits of the technique. Similarly, the "Automation" column also only indicates whether we have the automation setup for that method.

3.3.3 Barcode sequence affects array expression

Lastly, the sequence of the barcode itself may affect the array expression. We encountered this problem in early versions of the 4-unit intMEMOIR array, where the detection efficiency would drop sharply in the middle of the array (Fig. 3.3B). Puzzled by this observation, we hypothesized that a problematic barcode may be causing the transcript to terminate prematurely. To test this hypothesis, we took the last barcode we efficiently detected (labeled in red in Figure 3.3B) and rearranged it to the back of the array (Fig. 3.3C, top). Consistent with our suspicion, the detection efficiency of the array is restored (Fig. 3.3C). In addition, we also tested another 4-unit array with one static barcode that does not contain the red barcode. This array, too, appears efficiently transcribed. Thus, although these results were performed on polyclonal stable lines, they convincingly demonstrated that the red barcode is causing premature transcription termination. Upon further investigation, we discovered that the barcode sequence contained four copies of the polyadenylation sequence "AATAAA", the likely cause of our observations. Taken together, we began to perform *in silico* screens to eliminate all barcodes with polyadenylation sequences and splice sites for future array designs [code developed by Mark W. Budde, available at (*3*)].

Curiously, during the troubleshooting period, we discovered that flanking the 4-unit array between a pair of 5' and 3' splice sites also improved expression in the array with the red barcode (Fig. 3.3E,

left). Searching the literature revealed that the presence of a 5' splice site may prevent the recognition of nearby poly(A) signals (*10*). To test if this was the cause of the improved expression, we transfected two additional constructs that contained only one of the two splice sites (Fig. 3.3E, middle and right). Consistent with our expectation, the 5' splice site is necessary and sufficient to improve the expression of our array.

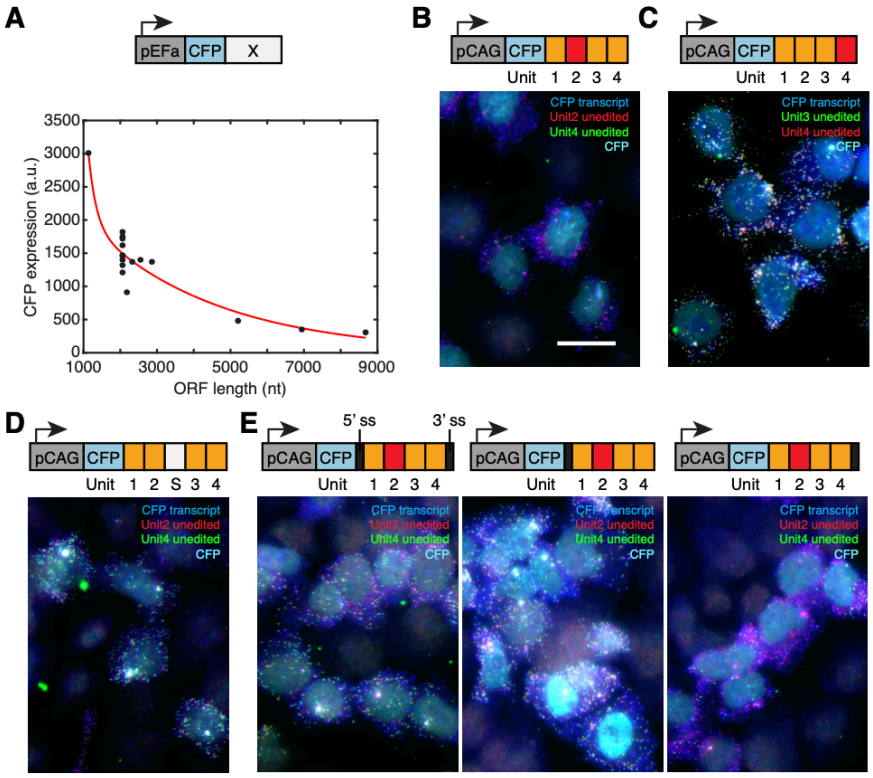

Fig. 3.3. Array and barcode length affect expression.
(**A**) Increasing the length of the 3' UTR following an H2B-mCerulean decreases the fluorescence of the transiently transfected cells. This result suggests that longer recording arrays (i.e. 3' UTR) decrease array expression level. (**B** to **E**) Barcode sequence can affect array expression. Experiments were done on polyclonal stable lines integrated in Rosa26 locus and read out through traditional smFISH (scale bar, 25 µm). (B) shows that transcription appears to drop off after one barcode, represented in red in the array schematic. The problem is resolved when the red barcode is moved to the end of the array or removed altogether, as shown in (C) and (D). (E) demonstrates that the presence of a 5' splice site is able to counteract what appears to be premature transcription termination.

3.4 Additional factors likely affect array performance

Overall, we have discussed a number of factors that could affect intMEMOIR performance, including the array's promoter choice, array length, barcode length, and barcode sequence. In addition to these factors, I also have several suggestions that are extrapolated from our experience, but lack explicit results to support the claims. As such, they should be treated as discussion points that require further investigation.

First and foremost, as I briefly suggested in Chapter 3.2.1's discussion on biTRE, the presence of H2B-mCerulean in the 5' end of our array appears to improve expression. Historically, this open reading frame (ORF) was introduced into our construct to facilitate mES cell colony selection for correct integration. Therefore, it was not included in all arrays. Interestingly, the arrays without this ORF tended to show either poor expression or retained transcripts in the nuclei. Thus, although we never explicitly tested whether the 5' H2B-mCerulean improved expression, we have since included it in all our PolII expressing arrays in mammalian cells. Further, if the ORF does improve expression, it would be interesting to investigate whether the same effect could be achieved with any ORF, H2B alone, or mCerulean alone. Note that the *Drosophila* UAS-10-unit construct, which does have robust expression, only contains mCerulean (without H2B).

Second, an ideal recording system minimizes its perturbation to the cell. This ensures that we are recording the natural process we hope to study, and not artifacts generated by the recording system itself. This was one of our original motivators to switch the editor from Cas9 to serine integrases (MEMOIR vs. intMEMOIR). Unlike Cas9-generated double-strand breaks (DSBs), serine integrase edits do not engage the endogenous DNA repair mechanism of the cell (*9*). As such, the recording likely creates less overall burden to the host cells. Furthermore, DSBs can be repaired through multiple pathways, and different cells may have different pathway preferences (*11*). With that in mind, Cas9 edits may also result in unnecessary variations in edit rate and outcomes between different cell types, tissues, or even organisms. Thus, serine integrases offer an overall more bio-orthogonal choice.

On a related note, compared to random integrations, site-specific integration of the recording array into a safe harbor locus also reduces our chances of perturbing the normal functions of the cell. This is an important consideration if we wish to employ other integration methods for future versions of intMEMOIR (discussed in Chapter 3.6.1).

Third, the memory unit design could also be improved in three ways. First, although it is not strictly necessary when many units are expressed on the same transcript, having a positive signal for deletion (instead of relying on the absence of unedited and inverted signals) could reduce miscalls during FISH readout. Second, future designs should strive to be compatible with sequencing-based readouts. Although it is not a primary goal of intMEMOIR, this compatibility would greatly improve the versatility of the system, allowing researchers to choose their preferred method of readout without needing to generate additional cell or model organism lines. Third, for applications where editing needs to be run to completion or near-completion, it may be preferable to introduce a third *attP* into the unit such that the "unedited" state becomes a third, terminal outcome. This ensures that we retain a maximum diversity of 3^{10}, and not 2^{10} (i.e. deletion and inversion only). However, it is also important to note that a third *attP* will increase the sequence repetitiveness and overall array length, which may lead to other complications.

Lastly, creating landing pad cell lines in safe harbor loci could significantly improve the turnover rate for testing future recording array designs. To generate a germline heritable system with robust expression and minimum perturbation to the endogenous genome, it is often preferred to site-specifically integrate our system into a safe harbor locus like Rosa26 in intMEMOIR. However, integrating through CRISPR-mediated homology directed repair, while significantly more efficient than integrating by homology alone, is still relatively inefficient compared to those mediated by recombinases, especially when the donor construct is large (*6*). We observed this difference when comparing our polyclonal stable lines for constructs integrated in the TIGRE locus, which were mediated by the Flp-recombinase, against those integrated in the Rosa26 locus, which were done through CRISPR. Due to this low rate of correct integration, quantitative analyses of array performance often require us to generate monoclonal cell lines with confirmed integration, which is a slow and labor intensive process. Thus, it would greatly streamline future tests if we obtained or generated mES cell lines with recombinase landing pads in common safe harbor sites. Some of these may already be available from existing publications (*6*, *7*, *12*, *13*). Related, mES cells containing mouse artificial chromosomes (MAC) with multiple integration sites may also be a useful resource for rapid tests or even mouse line generation (discussed in Chapter 3.6.2) (*14*).

I have summarized the main suggestions for designing future Pol II arrays in a checklist below. The list is based on our experiences described in this chapter, and should only be used as a reference.

Promoter

☐ Strong, consistent expression

☐ Inducible

☐ No transcriptional interference

Barcode

☐ Sufficient length to produce robust FISH signal

☐ Efficiently edited with approximately 1:1 ratio for inversion and deletion outcomes

☐ No splice sites

☐ No poly(A) signals

☐ No sequence homology with endogenous genes

☐ Compatible with sequencing readout

☐ As short as possible while satisfying the other criteria

Single integration

☐ Use safe-harbor sites that minimize perturbation to the cell

☐ Site is not silenced in cell type of interest

☐ Site is compatible with inducible promoters (e.g. low leakiness)

☐ Site has high integration efficiency to enable fast testing (e.g. using recombinase landing pads)

Multiple integrations

☐ There is minimal inter-array crosstalk

☐ Each array can be distinguished by FISH (e.g. via unique static barcodes)

☐ Variations in expression levels due to position effects is acceptable

Other considerations

☐ Array has a 5' ORF (e.g. H2B-mCerulean)

☐ Array should be portable (i.e. easy to implement in new cell lines and organisms)

3.5 The Zombie expression system may offer an improved design that eliminates co-transcriptional editing.

In 2020, Askary et al. from our lab published the Zombie system, which uses bacteriophage promoters to drive the barcode transcription in fixed cells (*5*). Since then, we have designed a Zombie version of the intMEMOIR that may offer several improvements over the original (Fig. 3.4). In this design, each unit is expressed by its own T7 and T3 promoter for a Zombie-based readout. This could potentially give us better signal than a single PolII promoter transcribing the entire 10-unit array, especially if they are not integrated in safe harbor loci. In addition, because Zombie transcription happens after the cells have been fixed, this design also avoids co-transcriptional editing, which increases the rate of cross-talk in our array (see Chapter 3.2.2).

Apart from adapting the T7/T3 promoters, we also made several modifications to the units themselves. First of all, they are designed to produce a positive signal for deletion to facilitate state-calling (see Chapter 3.4). Second, the barcodes themselves now have common primer binding sites, which increases their compatibility with sequencing based readouts. Third, the pair of *attP*s in these constructs are direct repeats, which could theoretically decrease the chance of truncation mutations during cloning. Instead of using inverted repeats, the *attP*s behave as if they are facing opposite directions through the use of reverse complement dinucleotides (fig. S2.1). The prototype 10-unit Zombie construct has already been cloned, and it is available for implementation and future characterization in both mammalian and zebrafish cells (Table 3.2).

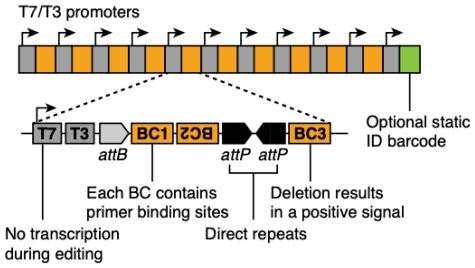

Fig. 3.4. The Zombie intMEMOIR design offers several potential improvements.
The new design provides a positive signal for deleted units, eliminates co-transcriptional editing, and is compatible with sequencing readout.

3.6 intMEMOIR can be the foundation for future recording systems

By expanding on the existing intMEMOIR system, we can develop new versions that feature greater recording capacity and capabilities. As demonstrated in Chapter 2, the existing intMEMOIR system can accurately reconstruct 3-4 generations of mES cell lineage. Increasing the memory of the system beyond 10-units would allow us to reconstruct more accurate lineage over longer timescales. Alternatively, we could also design the system to record molecular event histories in addition to lineage, resulting in a "decorated" lineage tree (Fig. 1.1). Both of these improvements would be very valuable to the study of normal and disease development.

3.6.1 Increasing memory increases reconstruction depth and accuracy

Not just for intMEMOIR, but for almost all synthetic recording systems, the depth and accuracy of reconstruction depend on the available memory (fig. S2.9). The most straightforward way to increase the memory of intMEMOIR would be to introduce more recording arrays. One array gives 10 recording units, or 3^{10} possible outcomes, while n arrays give 10n units, or 3^{10n} outcomes. To achieve this, we could randomly integrate a library of barcoded arrays into the cell through methods such as piggyBac or lentivirus transduction (Fig. 3.5A). However, there are several concerns with this approach, such as varied expression and integrase accessibility due to chromosome position effects and decreased germline transmissibility for *in vivo* applications. Another important concern is the possibility of inter-array recombination causing aneuploidy (*15*): because our editor is a recombinase, cross-talk between arrays could result in large scale genomic deletion or even chromosome translocation. However, unlike tyrosine recombinases, serine integrases irreversibly convert their *attP/B* into *attL/R*. Because of this irreversible "destruction" of their target sites, the number of target sites will progressively decrease with editing, and the resulting rate of aneuploidy may be lower than those observed with Cre and Flp. Nevertheless, it is important to empirically confirm that editing scattered arrays will not significantly perturb the systems we wish to study.

3.6.2 Orthogonal integrases can increase lineage reconstruction depth and accuracy

Additional orthogonal integrases can also expand our system. The current intMEMOIR only uses Bxb1, one member of a large family of serine integrases that could, theoretically, be used in parallel to enable both deeper lineage reconstruction and molecular event recording (*16–18*). The idea of using orthogonal integrases for event recording has already been discussed in Chapter 3 (fig. S3.2), so I will primarily focus on its potential to increase reconstruction depth and accuracy. However, it is worth noting that, unlike CRISPR-Cas9 based systems, additional intMEMOIR recording channels are enabled by orthogonal proteins (instead of orthogonal gRNAs). Thus, our system is able to record events on both the transcriptional and protein levels, enabling protein-based designs that incorporate synNotch, CHOMP, inducible dimerization domains, and others (*19–21*).

In terms of increasing memory, orthogonal integrases' arrays can, theoretically, be inserted into the same genomic locus with little to no crosstalk (Fig. 3.5A). These designs have the advantage of keeping the system confined within safe harbor locus and reducing the number of necessary constructs, simplifying germline heritability. Due to the transcriptional interference discussed in Chapter 3.2.1, however, it would likely be necessary to transcribe these arrays from the same promoter or through the Zombie system described in Chapter 3.5. Alternatively, they are also compatible with scattered integration approaches through transposons and viral transductions.

Apart from directly increasing the number of memory units, additional integrases may also enable us to reconstruct deeper lineages with fewer memory units. This could be achieved through the use of integrase cascades. In this design, an integrase would edit for a couple of generations before excising itself and activating the next integrase in line. When tuned correctly, each integrase would only record small and highly accurate trees that can then be stitched together to form the complete lineage record (Fig. 3.5B). Cascades like this could be simply implemented with self-excising serine integrases (Fig. 3.5C), or with more complex schemes such as two separate drug inducible cascades that, by manually altering the inducers every couple of generations, would better synchronize the transition from one integrase to the next across the colony (Fig. 3.5D). Simulations are needed to quantify exactly how much benefit can be gained from having these cascades.

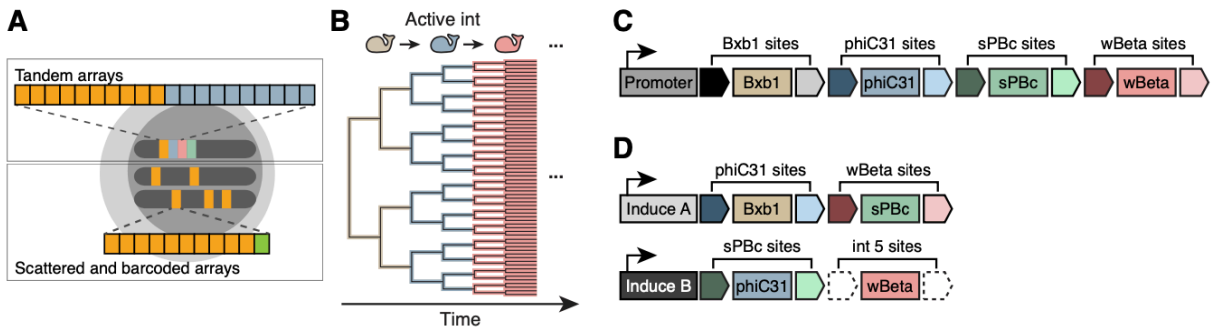

Fig. 3.5. Future versions of intMEMOIR could enable greater reconstruction depth and accuracy.
(**A**) Additional memory units can be introduced as tandem arrays (top) scattered integration (bottom). In the latter case, transcripts from each array are distinguished from one another through a unique identifying barcode. (**B**) Integrase cascades can, theoretically, increase the reconstruction depth without significantly increasing the amount of required memory. (**C**) A simple cascade relies on the self-excision of each integrase. (**D**) A more elaborate design uses two orthogonally inducible cascades to potentially enable synchronized integrase activation across all cells in the colony.

3.6.3 Generating the intMEMOIR mouse line and combining the system with other recording technologies could broaden its biological applications

Briefly, I would also like to touch on the topic of mouse line generation. Although *Drosophila* is a powerful model organism, many biological questions are specific to mammalian systems. As such, generating intMEMOIR mouse lines is an important future direction. Unfortunately, previous attempts with blastocyst injection of the intMEM1 line have failed. We suspect this may be due to the cell line's high passage number, which likely exceeded 200. Future work could tackle this challenge by recreating the intMEMOIR line in low passage mES cells, or by exploring pronuclear injections into zygotes (*22*). In particular, a recent publication that describes an *in vivo*, Bxb1-mediated integration of transgenes into the Rosa26 locus in mice could be a useful resource for generating mouse lines with non-Bxb1 intMEMOIR arrays (*6*).

Finally, I would like to end on the reminder that intMEMOIR is not mutually exclusive with other recording systems, and can be combined with other technologies to form complementary and more powerful recording technologies. For example, one could leverage the fast editing rates of integrases such as Bxb1, and use intMEMOIR to record developmental stages with rapid cell division, and use other slower editing systems or even somatic mutations to reconstruct periods of slower proliferation.

3.7 Summary

We have discussed several design principles for a serine integrase-based, image-readable recording system and suggested a few future directions. Much of the information presented here is learned from the challenges and failures we encountered while creating the original intMEMOIR system, and I hope they would facilitate the design and construction of future synthetic recording systems.

In summary, Figure 3.6 lists many of the general qualities we should look for in future versions of the intMEMOIR system. In addition, Table 3.2 contains a selected list of materials that may be useful for the testing and development of future systems. Most data, analyses, code, and sequence files described in this chapter are available at (*3*).

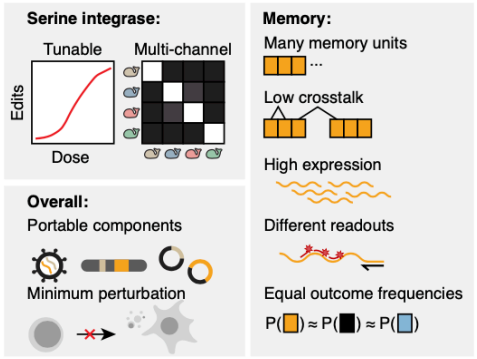

Fig. 3.6. Future versions of intMEMOIR should aim to have several qualities.
Although not an exhaustive list, future intMEMOIR systems should aim to have several important qualities for its editor, recording array, and overall design.

intMEMOIR

Type	Name	Purpose	Internal source	External source	Notes
mES cell line	intMEM1	As used in the published manuscript. Contains the complete intMEMOIR system.	nGC0244		Monoclonal, nGC0222 background
mES cell line	Inducible Bxb1	Preceeds intMEM1. Contains dox and TMP inducible Bxb1.	nGC0222		Polyclonal, nGC0170 background
mES cell line	TIGRE LP + Tet3G	Preceeds the inducible Bxb1 line. Contains the TIGRE landing pad and Tet3G.	nGC0170		Monoclonal, nGC0143 background
mES cell line	TIGRE LP	Preceeds the TIGRE LP + Tet3G line. Contains the TIGRE landing pad.	nGC0143		Monoclonal
D. melanogaster	PRExpress-Bxb1-hsp70pA	Drosophila line with heat shock inducible Bxb1.	Lois lab	Bloomington: BDSC_90853	
D. melanogaster	nSyb-Gal4; UAS-Ceru-10unit	Drosophila line with pan-neuronal expression of the 10 unit array.	Lois lab	Bloomington: BDSC_90854	
Plasmid	TIGRE-TRE-[poor kozak]Bxb1-ecDHFR-BGHpA	Dox and TMP inducible Bxb1 in the TIGRE donor vector. For mammalian cells.	pGC0100	Addgene: 158390	
Plasmid	R26-pCAG-Ceru-10unit-BGHpA	10 unit intMEMOIR array in the Rosa26 donor vector. For mammalian cells.	pGC0101	Addgene: 158387	
Plasmid	PRExpress-Bxb1-hsp70pA	Heat shock inducible Bxb1 with phiC31 attB integration site. For Drosophila.	pGC0102	Addgene: 158391	
Plasmid	UAS-Ceru-10unit	UAS driven 10 unit intMEMOIR array with phiC31 attB integration site. For Drosophila.	pGC0103	Addgene: 158389	

Other integrases

Type	Name	Purpose	Internal source	External source	Notes
mES cell line	New int 4mer xtalk 20-10	Rosa26 integrated 4 unit array of sPBc with a static barcode, used to test integrase edit rates.	nGC0180		Monoclonal, nGC0143 background
mES cell line	New int 4mer xtalk 20-11	Rosa26 integrated 4 unit array of sPBc with a static barcode, used to test integrase edit rates.	nGC0181		Monoclonal, nGC0143 background
mES cell line	New int 4mer xtalk 21-5	Rosa26 integrated 4 unit array of wBeta with a static barcode, used to test integrase edit rates.	nGC0182		Monoclonal, nGC0143 background
mES cell line	New int 4mer xtalk 21-7	Rosa26 integrated 4 unit array of wBeta with a static barcode, used to test integrase edit rates.	nGC0183		Monoclonal, nGC0143 background
mES cell line	R26 integrase test construct phiC31 polyclonal	Rosa26 integrated 1 invertable unit of phiC31 with a static barcode, used to test integrase activity.	nGC0102		Polyclonal
mES cell line	R26 integrase test construct R4 polyclonal	Rosa26 integrated 1 invertable unit of R4 with a static barcode, used to test integrase activity.	nGC0103		Polyclonal
mES cell line	R26 integrase test construct wBeta polyclonal	Rosa26 integrated 1 invertable unit of wBeta with a static barcode, used to test integrase activity.	nGC0104		Polyclonal
mES cell line	R26 integrase test construct sPBc polyclonal	Rosa26 integrated 1 invertable unit of sPBc with a static barcode, used to test integrase activity.	nGC0105		Polyclonal
mES cell line	R26 integrase test construct fBT1 polyclonal	Rosa26 integrated 1 invertable unit of phiBT1 with a static barcode, used to test integrase activity.	nGC0106		Polyclonal
mES cell line	R26 integrase test construct TP901 polyclonal	Rosa26 integrated 1 invertable unit of TP901 with a static barcode, used to test integrase activity.	nGC0107		Polyclonal

Zombie

Type	Name	Purpose	Internal source	External source	Notes
mES cell line	piggybac ZOMBIE 10mer Bxb1 "old"	Prototype line, contains Zombie 10 unit arrays (pGC0104) integrated through piggyBac.	nGC0249		Polyclonal, from Maria
Plasmid	PB-Bxb1-old Zombie intMEMOIR array	Prototype Zombie intMEMOIR array in piggyBac backbone.	pGC0104		From Maria
Plasmid	Tol2-Bxb1-old Zombie intMEMOIR array	Prototype Zombie intMEMOIR array in Tol2 backbone. For Zebrafish.	pGC0105		From Maria
Plasmid	PB-Bxb1-new Zombie intMEMOIR array	Prototype Zombie intMEMOIR array in piggyBac backbone with a new set of barcodes.	pGC0106		From Maria
Plasmid	Tol2-sPBc Zombie intMEMOIR array	Prototype Zombie intMEMOIR array for sPBc integrase in Tol2 backbone. For Zebrafish.	pGC0107		From Maria. Unit 2 *attP* may be mutated

Table 3.2. A list of useful materials for future intMEMOIR development.

3.8 References

1. C.-M. Chen, J. Krohn, S. Bhattacharya, B. Davies, A Comparison of Exogenous Promoter Activity at the ROSA26 Locus Using a PhiC31 Integrase Mediated Cassette Exchange Approach in Mouse ES Cells. *PLoS ONE*. **6** (2011), p. e23376.

2. K. Shearwin, B. Callen, J. Egan, Transcriptional interference – a crash course. *Trends in Genetics*. **21** (2005), pp. 339–345.

3. K. Edmonds, M. Budde, S. Yoon, M. Cabrera, M. Elowitz, Data for Ph.D. thesis, "Chapter 3: Building serine-integrase based, image-readable recording systems" (2021) (Version 1.0) [Data set]. CaltechDATA. https://doi.org/10.22002/D1.2195.

4. C. A. Merrick, J. Zhao, S. J. Rosser, Serine Integrases: Advancing Synthetic Biology. *ACS Synth. Biol.* **7**, 299–310 (2018).

5. A. Askary, L. Sanchez-Guardado, J. M. Linton, D. M. Chadly, M. W. Budde, L. Cai, C. Lois, M. B. Elowitz, In situ readout of DNA barcodes and single base edits facilitated by in vitro transcription. *Nat. Biotechnol.* **38**, 66–75 (2020).

6. B. E. Low, V. Hosur, S. Lesbirel, M. V. Wiles, Efficient targeted transgenesis of large donor DNA into multiple mouse genetic backgrounds using bacteriophage Bxb1 integrase. *bioRxiv* (2021), p. 2021.09.20.461117, , doi:10.1101/2021.09.20.461117.

7. E. I. Ioannidi, M. T. N. Yarnall, C. Schmitt-Ulms, R. N. Krajeski, J. Lim, L. Villiger, W. Zhou, K. Jiang, N. Roberts, L. Zhang, C. A. Vakulskas, J. A. Walker II, A. P. Kadina, A. E. Zepeda, K. Holden, J. S. Gootenberg, O. O. Abudayyeh, Drag-and-drop genome insertion without DNA cleavage with CRISPR-directed integrases. *bioRxiv* (2021), p. 2021.11.01.466786, doi:10.1101/2021.11.01.466786.

8. K. L. Frieda, J. M. Linton, S. Hormoz, J. Choi, K.-H. K. Chow, Z. S. Singer, M. W. Budde, M. B. Elowitz, L. Cai, Synthetic recording and in situ readout of lineage information in single cells. *Nature*. **541**, 107–111 (2017).

9. G. D. V. D. Karen Rutherford, The ins and outs of serine integrase site-specific recombination. *Curr. Opin. Struct. Biol.* **24**, 125 (2014).

10. N. J. Proudfoot, Ending the message: poly(A) signals then and now. *Genes Dev.* **25**, 1770–1782 (2011).

11. R. Scully, A. Panday, R. Elango, N. A. Willis, DNA double-strand break repair-pathway choice in somatic mammalian cells. *Nat. Rev. Mol. Cell Biol.* **20**, 698–714 (2019).

12. X. Chi, Q. Zheng, R. Jiang, R. Y. Chen-Tsai, L.-J. Kong, A system for site-specific integration of transgenes in mammalian cells. *PLoS One*. **14**, e0219842 (2019).

13. B. Tasic, S. Hippenmeyer, C. Wang, M. Gamboa, H. Zong, Y. Chen-Tsai, L. Luo, Site-specific integrase-mediated transgenesis in mice via pronuclear injection. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 7902–7907 (2011).

14. Y. Yoshimura, K. Nakamura, T. Endo, N. Kajitani, K. Kazuki, Y. Kazuki, H. Kugoh, M. Oshimura, T. Ohbayashi, Mouse embryonic stem cells with a multi-integrase mouse artificial chromosome for transchromosomic mouse generation. *Transgenic Res.* **24**, 717–727 (2015).

15. M. Lewandoski, G. R. Martin, Cre–mediated chromosome loss in mice. *Nature Genetics*. **17** (1997), pp. 223–225.

16. Z. Xu, L. Thomas, B. Davies, R. Chalmers, M. Smith, W. Brown, Accuracy and efficiency define Bxb1 integrase as the best of fifteen candidate serine recombinases for the integration of DNA into the human genome. *BMC Biotechnol.* **13**, 1–17 (2013).

17. M. S. Gomide, T. T. Sales, L. R. C. Barros, C. G. Limia, M. A. de Oliveira, L. H. Florentino, L. M. G. Barros, M. L. Robledo, G. P. C. José, M. S. M. Almeida, R. N. Lima, S. K. Rehen, C. Lacorte, E. O. Melo, A. M. Murad, M. H. Bonamino, C. M. Coelho, E. Rech, Genetic switches designed for eukaryotic cells and controlled by serine integrases. *Commun Biol*. **3**, 255 (2020).

18. M. G. Durrant, A. Fanton, J. Tycko, M. Hinks, S. S. Chandrasekaran, N. T. Perry, J. Schaepe, P. P. Du, P. Lotfy, M. C. Bassik, L. Bintu, A. S. Bhatt, P. D. Hsu, Large-scale discovery of recombinases for integrating DNA into the human genome. *bioRxiv* (2021), doi:10.1101/2021.11.05.467528.

19. L. Morsut, K. T. Roybal, X. Xiong, R. M. Gordley, S. M. Coyle, M. Thomson, W. A. Lim, Engineering Customized Cell Sensing and Response Behaviors Using Synthetic Notch Receptors. *Cell*. **164**, 780–791 (2016).

20. X. J. Gao, L. S. Chong, M. S. Kim, M. B. Elowitz, Programmable protein circuits in living cells. *Science*. **361**, 1252–1258 (2018).

21. B. H. Weinberg, J. H. Cho, Y. Agarwal, N. T. H. Pham, L. D. Caraballo, M. Walkosz, C. Ortega, M. Trexler, N. Tague, B. Law, W. K. J. Benman, J. Letendre, J. Beal, W. W. Wong, High-performance chemical- and light-inducible recombinases in mammalian cells and mice. *Nat. Commun.* **10**, 4845 (2019).

22. L. M. Ittner, J. Götz, Pronuclear injection for the production of transgenic mice. *Nature Protocols*. **2** (2007), pp. 1206–1215.

CONCLUSION

Our recording projects began in 2014, around when I first joined Michael's lab. Driven by the need in the field, we developed a synthetic lineage tracking system with imaging-based readout, which was demonstrated in mES cells. The system was a novel proof-of-principle that demonstrated the possibility to image single cell lineage history, but its reliance on genomically distributed recording units with only one edit outcome limited its reconstruction capabilities and germline transmissibility. Since then, we returned to the drawing board and focused on developing a new recorder that retains the unique strengths of the original while addressing its flaws. The efforts finally culminated into intMEMOIR, a system that allows us to simultaneously analyze single-cell lineage, cell state, and spatial organization *in vitro* and *in vivo*.

Unlike most other technologies in the lineage tracking field, intMEMOIR uses serine integrases to edit a transcribed DNA array of 10 independently editable memory units. Each of the units can exist in one of three states: unedited, deleted, or inverted. Upon transcription, this produces three different RNA species (intact, absent, and reverse complement, respectively) that could be unambiguously distinguished from one another underneath the microscope using FISH. The entire array collectively allows up to 59,049 possible edit outcomes: a significant improvement over existing image-readable recording systems. We implemented the system in an mES cell line, where we evaluated its clonal classification and lineage reconstruction accuracy, and in a *Drosophila melanogaster* line, where we demonstrated its ability to disentangle the relative contributions of lineage and spatial organization to cell fate determination in adult brains.

Overall, we have built the foundation for a serine-integrase based, image-readable recording system. In the future, increasing the memory of the system would increase the depth and accuracy of its reconstruction. Furthermore, intMEMOIR could also serve as the foundation for systems with parallel recording channels to capture information such as the timing, duration, and magnitude of molecular events experienced by the individual cells and their ancestors. These improvements and new insights would enable the reconstruction of "decorated" lineage trees that would contribute to the construction of a developmental atlas. Ultimately, through developing, applying, and

disseminating these tools, we hope to bring us one step closer to deconvoluting the underlying mechanism of cell fate determination in development, homeostasis, and diseases.