Scalable Methods for *In Situ* Genomics

by

Andrew C. Payne

B.A.Sc., University of Toronto (2015)
S.M., Massachusetts Institute of Technology (2017)

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

Massachusetts Institute of Technology

September 2021

Author_____

Program in Media Arts and Sciences
August 20, 2021

Certified by_____

Edward S. Boyden
Y. Eva Tan Professor in Neurotechnology
Massachusetts Institute of Technology
Thesis Supervisor

Accepted by_____

Tod Machover
Academic Head, Program in Media Arts and Sciences

Scalable Methods for *In Situ* Genomics

by

Andrew C. Payne

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
on August 20, 2021, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

The maxim that biological structure determines function was inspired by the discovery of the DNA double helix, yet mapping the structure of genomic DNA within a cell remains challenging, and accordingly the role of genome structure and organization in determining cellular function is an open question. Mapping cellular genome organization is difficult because it requires joint measurement of linear DNA sequence and 3D spatial context, however, existing genome-scale methods lack either base-pair sequence information or direct spatial localization. To overcome these limitations, we invented *In Situ* Genome Sequencing (IGS), a set of scalable methods for simultaneously sequencing and imaging cellular genomes within intact biological samples. We first report technological developments enabling IGS, including new chemistries for *in situ* sequencing library construction, workflows for multimodal sequencing of libraries, and strategies for computational integration of spatial and genetic information. Next, we use IGS to map spatial genome organization in cultured human fibroblasts, validating and benchmarking our results against key genomic features such as chromosome positioning, chromosome folding, and repetitive sequence localization. Finally, we apply IGS to map genome organization in intact mouse early embryos, extending known features and uncovering new features of embryonic genome architecture. We characterize parent-specific changes in genome structure across embryonic stages, reveal single-cell chromatin domains in zygotes, and uncover epigenetic memory of global chromosome positioning within individual embryos. We conclude with a discussion of IGS scaling properties, by which we can anticipate many-fold future improvements in yield and resolution. We anticipate IGS and related scalable *in situ* methods will be instrumental in unifying genomics and microscopy, enabling scientists to map genome organization from single base pairs to whole organisms and ultimately to connect genome structure and function.

Thesis Supervisor: Edward S. Boyden
Title: Y. Eva Tan Professor in Neurotechnology, Massachusetts Institute of Technology

Scalable Methods for *In Situ* Genomics

by

Andrew C. Payne

This dissertation has been reviewed and approved by the following committee members

Edward S. Boyden

Y. Eva Tan Professor in Neurotechnology

Massachusetts Institute of Technology

George M. Church

Professor of Genetics

Harvard Medical School

Kevin M. Esvelt

Assistant Professor of Media Arts and Sciences

Massachusetts Institute of Technology

Contents

List of Figures

Acknowledgments

It takes a village to raise a PhD, and I cannot express strongly enough how grateful I am to the community of mentors, collaborators, colleagues, staff, friends and family that offered and continue to offer me their support.

First, I want to thank Ed Boyden, who offered me many years of extensive mentorship and advice, as well as unconditional support and freedom to chart my path as a scientist and follow my interests wherever they led. I admire his relentless focus on impact and careful meta-thinking around the art and practice of science, and I hope to carry much of that thinking into my future endeavors.

Fei Chen took me under his wing after a tumultuous start to my work, and his close mentorship and advice helped me get back on my feet, grow as a scientist, and prosper. I deeply respect how Fei combines intellectual and experimental fearlessness and playfulness, and I consider myself very lucky to have had the opportunity to learn that from him first-hand.

Adam Marblestone has been my North Star for nearly ten years -- scientifically, professionally, and philosophically -- even though, to mix my astronomical metaphors, our orbits were rarely in synchrony. I hold in high regard Adam's drive to explore the blank spaces on institutional maps as readily as scientific ones, and his rare ability to widely catalyze serendipity across people and projects. I hope our orbits will continue to interweave for many years to come.

Sam Rodriques played many roles during our years together: friend, peer, role model, collaborator, mentor, and various combinations thereof. I have great respect for Sam's capacity to inspire, bias for action, and taste for risk and reward. I'm lucky to consider him a close friend, and I look forward to seeing what roles he will play during the next act.

George Church changed the course of my life when he offered me a place in his lab after Halcyon Molecular. I am honored he chose to serve on my committee, and offer many hours of discussion and guidance, as well as thoughtful encouragement when things seemed bleak.

I'm also honored that Kevin Esvelt chose to serve on my committee, and his commitment to scientific openness and a better and safer tomorrow is an inspiration to me (as is his

impeccable taste in science fiction).

Reza Kalhor is an inspiration to me. Reza offered timely advice when our project was on the ropes, for lack of genome organization expertise, and since then he has become a valuable source of scientific and professional advice and discussion, and I am honored to call him my friend.

This dissertation could not have happened without the collaborative efforts of Zack Chiang. Zack is a computational wizard, and I will be forever impressed at how effectively his design choices allowed us to cope with the inevitable messiness of first-in-class experimental data. I admire his coolness under pressure and his talent for communication, and I can't wait to see what he'll do next.

Nor could this dissertation have happened without the collaborative efforts of Paul Reginato. Paul is the most dedicated scientist I have ever had the pleasure of working with. His initiative in applying IGS to early embryos - a high-risk and speculative effort at the time - vastly amplified the project's ultimate impact, and the gel chemistries he has pioneered since then in the context of ExGS are essential to the next generation of scalable in situ genome sequencing technologies. I wish him absolute success in his future endeavors.

I was lucky to have the opportunity to collaborate closely with Jason Buenrostro for many years. Jason and I met through serendipity -- it just so happened that he and Fei were appointed Broad Fellows and met as I was performing the first proof-of-principle experiments on what was to later become IGS. His deep experience with methods in epigenomics helped kick-start our effort at the most crucial time, and I deeply value receiving his mentorship on all aspects of genome science.

IGS was a true team effort, and I look back fondly on the countless hours we spent together overcoming seemingly impossible hurdles. Beyond the core team, I had the pleasure of working with other amazing individuals who made essential contributions to the project.

Sarah Mangiameli made key contributions to the project at multiple stages, developing models for chromosome segmentation and outlier detection, as well as scripts for beautiful 3D chromosome visualization. I'm very lucky to have had an opportunity to collaborate with her and I'm excited to see her own variations on IGS.

Evan Murray made a heroic effort to understand the failure modes of in situ ATAC-seq,

and while we were ultimately unsuccessful, I deeply appreciate his commitment and effort and am happy to see him flourishing in his current role as staff scientist at the Broad.

I'm additionally grateful to contributions from Styliani Markoulaki and Rudolf Jaenisch, who offered critical support and effort preparing early embryos for IGS, Andrew Earl, who wrote a public-facing IGS data browser, and Ajay Labade, who performed informative exploratory experiments.

I also had the pleasure to mentor two amazing UROPs. I co-mentored Shirin Shivaei with Shahar Alon on early experiments regarding in situ sequencing by synthesis, and I'm pleased to see that she is flourishing at Caltech. I also had the pleasure to co-mentor Chun-Chen (Jerry) Yao with Paul Reginato, and I remain to this day remain impressed by his diligence and brilliance attacking the hard problem of in situ sequencing by synthesis, and I wish him the best of luck in his future endeavours.

In addition to those individuals whom I collaborated with directly, there are many others whose timely advice, insights, or reagents helped me move my work forward in myriad ways.

Ting Wu also offered timely genome-science related advice on many occasions, and I'm inspired by the kindness and care with which she mentors her trainees. I'm also grateful for useful feedback on various occasions from Huy Nguyen, Shyamtanu Chattoraj, S. Dean Lee, and Nuno Martins.

Erez Lieberman-Aiden offered helpful advice at a key stage of the project, and I admire him as a founder of this field I chose to explore.

I am grateful to David Feldman for his gift of nonamer sequencing reagents, and Jonathan Strecker for his gift of Tn5 transposase.

I also acknowledge and thank the Natural Sciences of Engineering Research Council of Canada for its generous financial support of my graduate studies through an NSERC postgraduate doctoral scholarship.

Beyond the work at the core of this dissertation, I had the opportunity to collaborate with many amazing individuals on a range of projects, and I am grateful to them for many opportunities to learn and grow.

Shahar Alon helped me get started with ExSeq, and in so doing set me on the path I am on now. I deeply appreciated his diligence and patience in teaching me, as an inexperienced

graduate student, the ins-and-outs of an early incarnation of ExSeq, which was a notoriously difficult technique at the time. Shahar is one of those rare personalities that is soft-spoken, but for whom everyone will quiet down to listen, and I wish him the absolute best as he starts his new lab.

It was a true delight to work on-and-off with the entire ExSeq team. Dan Goodwin is a constant source of boundless optimism and infectious curiosity. I can't count the number of times I found myself relying on Oz Wassie's encyclopedic experimental knowledge and wisdom. I am particularly grateful for many late night conversations with Anu Sinha spanning both the intricacies of science and technology development, and the larger issues of how to ethically participate in a system known to mistreat its participants -- I have grown to consider Anu one of my moral compasses. (Our shared sense of humor didn't hurt please clap.)

I am privileged to have had the opportunity to work with Nikita Obidin on DNA elongation, and lucky to have had his friendship and domestic good company as a housemate for many years. Nikita is a rare soul with diverse interests and impeccable taste, and I expect he will go on to do great things.

I was lucky to have the opportunity to work closely with Monique Kauke on in situ screening, who in addition to being an amazing scientist, had a rare gift of kindness and extroversion, a true social butterfly. Boyden Lab socials were unusually lively when Monique organized them, and I remember those times fondly (TNG cake!). I also value the opportunity I had, with Monique, to learn the particulars of image-based directed evolution with Kiryl Piatkevich and Yong Qian.

It was also a delight to work with the "Bobae Bay" in one of my more productive forays into the world of optical connectomics. Bobae An is a brilliant scientist, diligent worker, and all-around delightful person. Her pioneering work may very well change the way we do neuroscience, and I wish her the best of luck in that continuing endeavour. I also deeply enjoyed working with Kylie Leung, whose irreverence and sass is only matched by her commitment to the team and proficiency with a label-maker; I owe her a debt of gratitude for identifying Chikorita as my spirit Pokemon. I also valued the opportunity to work with and learn from Jenny Fehring, Shubra Pandit, and Burcu Guner-Ataman during this time.

I'm especially grateful to the contingent of old-guard Boyden Lab members (circa 2015) who helped me find my feet in the early days of joining the lab. Ishan Gupta, Kata Adamala, Daniel Martin-Alarcon, and Katriona Guthrie Honea were some of the first people to help me feel like I was embedded in a community. Paul Tillburg in particular helped me dive back into experiments after a rocky first semester. I'm also happy to have enjoyed many thoughtful conversations with Noah Jakimo and Lisa Nip in our shared Media Lab space.

Charles Fracchia introduced me to the Media Lab, and continues to offer me useful advice and mentorship from beyond its walls, which I am thankful for.

I enjoyed my many conversations and nascent project ideas with Boyden Lab members and affiliates at the Wyss Institute, including Nick Barry, Kettner Griswold, and Richie Kohman.

I'm fortunate to have spent many evenings discussing science and life with Shannon Nangle, and I'm glad to consider her a close friend and fellow traveller as we boldly go where no one has gone before.

Joe Scherrer and I met and became friends during Joe's rotation in the Chen Lab, and I'm glad our friendship has continued well beyond. I'm especially grateful to Joe for introducing me to the joy of bioluminescent insect life.

I'm grateful to Jenna Aaronson for seeking out a productive period of collaboration and mentorship from myself and Sam Rodriques.

I'm very lucky to have had the opportunity to work with Desiree Dudley in her official role as Director of Neurotechnology Partnerships, as well as her unofficial role as "Ship's Councillor." Desiree is a person of rare emotional intelligence, and I'm inspired by her ability to build and cultivate community flourishing. I'm also deeply fortunate to have shared with Desiree many years of friendship and her exceptional good company as a housemate. I will be forever grateful to her for catalysing a professional and personal chain reaction which is still going strong after more than ten years.

In the waning days of my PhD I've had many positive interactions with Tony Kulesa, and I'm appreciative of his advice and support.

It almost goes without saying that the Boyden Lab attracts amazing people, all of whom I have mutually learned from or have been assisted by in one way or another, including

Orhan Celiker, Alexi Choueiri, Yi Cui, Amauche Emenari, Daniel Estandian, Rui Gao, Shannon Johnson, Changyang Linghu, Daniel Oran, Danielle Cosio, Demian Park, Sarah Sclarsic, Corban Swain, Chi Zhang, Shoh Asano, Grace Huynh, Louis Kang, Manos Karagiannis, Kristina Kitko, Nikita Pak, David Rolnick, Deblina Sarkar, Or Shemesh, Ho-Jun Suk, Christian Wentz, and Jay Yu.

Similarly, the Chen Lab is also a special place, and I am lucky to have worked with and learned from equally exceptional colleagues, including Haiqi Chen, Ehsan Habibi, Sophia Liu, Julia Morriss, Jamie Marshall, Linlin Chen, Tongtong Zhao, and Sam Padula.

It is also the case that absolutely nothing would ever get done without the amazing support staff that I have had the privilege to work with, including Macey Lavoie, Lisa Lieberson, Doug Weston, Fira Zainal, Holly Birns, and Cynthia Smith in the Boyden Lab, and Jared Spencer in the Chen Lab.

Thank you all very much for our many fruitful years together, and my apologies to anyone I might have inadvertently missed. I also could not have made it this far without the support of many friends and loved ones outside the lab, including Caitlyn Hoeflin, Christina Fong, John Brothers, Kate Murphy, and Mark Hamalainen.

Joel Dapello and Kathleen Leeper deserve particular acknowledgement as excellent housemates and friends (as well as for their impeccable taste in food, music, science, and narrative).

I want to especially thank Evgenia Nitishinskaya for her boundless love, encouragement, commitment, and enthusiasm. I couldn't, in my wildest dreams, have asked for better company on this journey, and I can't express how lucky I am.

Finally, I want to thank my brother, Eric Payne, and my parents, Rosanna and Richard Payne, for their many years of unconditional and inexhaustible love and support.

Chapter 1

Introduction

Inference of biological function from form is an essential strategy across many branches of biology. A eureka moment can arise or a hypothesis can be generated by simply looking at a picture or 3D model of a biological system (although in fairness, visual pattern recognition in humans is far from simple). Technological advances have successively extended this strategy for more than three hundred years, from the first drawings of cells in Hooke's Micrographia, to the diffraction patterns of DNA in Franklin's Photo 51, to the uncountable fluorescence micrographs of cells expressing green fluorescent protein, a graduate student's rite of passage.

Mapping biological systems is an equally essential strategy. Scientifically, maps help to generate hypotheses or constrain hypothesis building — to better navigate an uncertain theoretical or experimental landscape. In the post-genomic era, mapping DNA or RNA sequence content with high-throughput DNA sequencing technology is now commonplace, and scientists are finding increasingly clever ways to transform other variables into DNA sequence space to better fit this paradigm. The rapid and ubiquitous success of this approach has led to DNA sequencing described, with some justification, as the "new microscope" [1]. But analogy this should not be taken too literally: sequence space and three-dimensional space are distinct, and schemes to transform information from the latter into the former, despite their benefits, are still in their infancy.

Ideally one would measure both of these quantities directly, making sequence measurements with direct spatial localization. However, until recently, technology has lagged behind this need: the limited scale of traditional fluorescence microscopy in terms of colors and

throughput prohibited spatial analysis with a genome-wide view. In this work, we present a suite of methods that remove these limitations, enabling highly multiplexed measurements of genomic sequence with spatial context.

In **Chapter 2**, we describe the genome structure mapping literature. We discuss the broad need for mapping biological systems, provide historical context around genome structure mapping, offer an overview of the current state of the science, describe current methods for assaying genome structure, and outline the potential of scalable approaches based on in situ sequencing.

In **Chapter 3**, we describe technological developments along the road to scalable in situ genomics. We first detail in situ ATAC-seq, a method intended for spatially profiling the epigenomes of single cells. This project encompassed improvements in library construction which enabled amplification, detection, and sequencing of open chromatin in intact cells and tissues. The protocol ultimately proved unreliable and we did not pursue in situ ATAC-seq beyond proof-of-concept. However, the library construction methods we pioneered proved vital for the ultimate success of IGS. Next, we describe how in situ genomics may be leveraged for long-read DNA sequencing. This project was conceived after initial successes with multimodal sequencing, as described in **Chapter 4**. We speculated that a combination of DNA elongation, expansion microscopy, and in situ UMI sequencing could deliver improved long read DNA sequencing technology. We also did not pursue this potential application of scalable in situ genomics – although DNA elongation and expansion microscopy were successfully combined, sequencing library construction proved intractable. Nonetheless, such an approach may, if successfully engineered, ultimately deliver superior DNA sequencing performance.

In **Chapter 4**, we shift our focus to *in situ* genome sequencing (IGS). We outline the IGS concept and workflow, including enabling technological improvements concerning in situ genomic library construction, multimodal sequencing, and computational integration of spatial and genetic information. We next describe how IGS can recapitulate known features of genome architecture in human fibroblasts, validating the technology. Finally, we apply IGS to mouse early embryos, revealing new features of embryonic genome architecture.

Lastly, in **Chapter 5** we anticipate future improvements and describe our outlook for

both IGS in particular, and scalable in situ genomics more generally. Altogether, the work in this dissertation shows how microscopy and sequencing may be unified, opening the way to more nuanced observation of biological structure that considers both sequence and physical space.

Chapter 2

Genome structure mapping

2.1 Why genome structure mapping?

It is a truism that successful scientific endeavours depend on using the right level of conceptual abstraction at the right time. In biology, the abstract concept of the gene as the unit of heredity brought about a scientific revolution [2]. It succeeded because it did not assert a physical substrate for the gene, and indeed, it would have been premature to do so, for the tools to investigate that substrate had not yet been invented. Although successful, the abstract gene could only take investigators so far: a mechanistic understanding was eventually needed to move beyond phenomenological models of inheritance [3]. By the middle of the 20th century, the number of basic biochemical building blocks - proteins, carbohydrates, lipids, and nucleic acids - had been cataloged, and technological developments in biochemistry and physics, applied systematically to this catalog, elucidated the physical, biochemical substrate of the gene [4]. Once elucidated, that substrate turned out to be the rather innocuous substance of deoxyribonucleic acid - not protein, as had been widely suspected at the time - and the revolution in molecular biology came into its most productive period [4].

Today, certain branches of biological inquiry, such as genomics, can be considered to be between scientific revolutions of a similar kind. By analogy, genomics is somewhere after the introduction of the idea of the gene, but before the double helical model of DNA: it can provide phenomenological models of biological systems but cannot yet provide systems-level mechanistic understanding. If the analogy is to hold, scientific revolution requires that the

genomics: the full scope of genomic building blocks must be cataloged and systematically studied by appropriate technologies.

However, unlike the basic biochemical building blocks of the molecular biology revolution, there are a bewildering number of components in complex biological systems, and moreover, the cataloging process itself depends on the level of abstraction chosen for the catalog (or map, when including relationships between elements). In the early 21st century, it was widely believed that the right level of abstraction was that of *sequence*, and the human genome project, perhaps the most celebrated biomedical project in history, aimed to - and broadly speaking, succeeded at - mapping the genome in terms of sequence [5]. However, although much was gained by the completion of the project, scientific revolution has arguably not arrived [6]. In light of this relative failure, investigators are continuing the search, generating additional biomolecular maps at varying levels of abstraction. Contemporary efforts now, in addition to sequence, examine features such as abundance [7] or accessibility [8], at resolutions from tissues [9] to single cells [10] to subcellular compartments [11, 12]. These efforts are to a large extent enabled by a *technological* revolution in biomolecular measurement, underpinned by rapid improvements in high-throughput DNA sequencing [1]. Measurements tied to a sequencing assay, such as Hi-C[13, 14], typically capture a high degree of sequence-level biomolecular content in the form of sequence, but give up spatial context; when it is captured, it is by a molecular proxy rather than a direct measurement. On the other hand, there are dedicated methods to directly measure the spatial position of specific biomolecules, but they are typically low-throughput and give up nucleotide-resolution sequence information [15].

Many of the functional building blocks of genomics are thought to be lcoated with the nucleus, a subcellular compartment that has been subject to much scrutiny [16]. Investigators have found the nucleus have a high degree of spatial organization [16], and it is the focus of ongoing efforts to understand gene regulation in healthy [17] and disease states [18]; a productive level of abstraction at which to analyze the nucleus should include that spatial information, ideally in the form of a direct, sequence-level measurement. This is a means of "assumption-proofing:" much like the surprising result of the molecular biology revolution, where the substrate for heredity was found in an unexpected place, insight into function

may come from unexpected places, and therefore, measurement technologies should ideally be agnostic to which features are believed to be important [19]. In this dissertation, we describe such technologies, where sequence and spatial context are directly observed, minimizing the degree of assumption required. In the remainder of this chapter, we first outline known features of the 3D spatial architecture of the genome (henceforth "3D genome"), and then discuss the strengths and limitations of contemporary methods to study the 3D genome with simultaneous sequence and spatial context.

2.2 Genome structure: the story so far

One cannot discuss the architecture of the genome without first commenting on the architecture of the gene, or, rather, its physical substrate, DNA. When DNA was first identified by Miescher in 1869, its biological function was not yet clear, although he did determine its biochemical properties and macromolecular character [20]. The four nucleotides which make up DNA were determined by Levene, which he hypothesized to have an invariant tetranucleotide structure playing a structural biological role [21]. Thus, the idea that nucleic acid and not protein could be the substrate for the gene was not seriously challenged until rigorous experiments by Avery, with MacLeod and McCarty [22]. Although they were circumspect in their interpretation, Hershey and Chase soon corroborated their results [23]. Chargaff, spurred by Avery and colleagues, turned his attention to nucleic acids, and determined that adenine and thymine, and cytosine and guanine, were always present in molar ratios of one-to-one [24]. Franklin and Wilkins collected crucial, information rich x-ray diffraction patterns of DNA [25, 26]. Watson and Crick integrated these diverse sources of information using model building techniques pioneered by Pauling, and ultimately determined the polymeric, double helical structure of DNA: the molecular substrate of the gene [27]. The physical constraints of a structural model were of great conceptual utility in the following decades: mechanisms for e.g. replication and gene expression had to be compatible with this double helical model [28]. Thus, the model bounded the scope of possible inquiry and gave rise to what became known as the central dogma, so-called because, at the time of its conception, there was no experimental evidence, just an elegant theoretical correspondence with the model [4].

So much for the architecture of the gene, written in DNA. For the genome, we first turn our attention to the nucleus. The end-to-end length of the DNA content of a diploid human genome is hundreds of thousands of times longer than the diameter of the nucleus in which it resides, making mechanisms for packing and storage of great importance. There are many additional mechanisms (e.g. repair, transcription, replication) that must be compatible with the packaged DNA. These mechanisms are typically implemented by proteins or RNA, and the DNA-RNA-protein composite is known as chromatin [29].

The fundamental building block of chromatin is the nucleosome, a nucleoprotein particle constructed from eight histone proteins and wrapped in approximately 146 base pairs (bp) of DNA [30]. These particles are linked by short stretches of DNA up to 80 bp long, and resemble "beads on a string" in electron microscopy studies [31], this configuration is known as the 10-nm fiber . There are a large number of histone modifications which are thought to combinatorially regulate gene expression and serve other regulatory purposes such as compaction, but a detailed understanding of this so-called histone code has been elusive [32].

At a higher level of organization, groups of nucleosomes are thought to form a large structure known as the 30-nm fiber, which is typically compact and made up of heterochromatin, with various regions unfolded into the 10 nm fiber at sites of active transcription [32]. Various models have been proposed for the 30-nm fiber. The solenoid model proposes an interdigitated structure in a one-start helix configuration (i.e. each nucleosome is connected to adjacent nucleosomes on the same helix), while the zig-zag model proposes a two-start helix (ie. each nucleosome is connected to adjacent rows of nucleosomes) [33]. The nature of the fiber remains a subject of intense debate even today: recent work has suggested an ensemble of structures coexist on the fiber [34], while other work casts doubt on the 30 nm fiber altogether [35].

Ultimately, all of these spatial configurations are still defined by a single continuous DNA polymer: a chromosome. Chromosomes are the largest discrete units of nuclear organization, and were one of the earliest to be studied, being visible under light microscopy in the condensed form they adopt during metaphase [36]. In interphase, chromosomes decondense, occupying distinct territories (chromosome territory, CT) with limited intermingling [37].

These territories were first proposed by Rabl, and the term itself was coined by Boveri in the first decade of the twentieth century [37, 38]. However, nearly a century passed before their existence could be shown experimentally [39]. CTs typically maintain their positions until the next cell division; daughter cells do not exactly preserve these positions, but neither are the new positions random [40, 41]. When CTs are analyzed radially with respect to the cell center, the gene density or size of the chromosome is predictive of a CT's radial position, with smaller and more gene dense chromosomes preferring to be positioned in the nuclear interior [42, 43, 44]. Moreover, CTs positions are tissue-specific [45], and have alternate organizational patterns in disease states [46]. These spatial organizational patterns are evolutionarily preserved in primates, suggesting a functional role, although the details remain unclear [47].

The organizational patterns described so far are relatively well established phenomena and could be described as a "textbook model." However, recent work, using so-called proximity ligation methods that will be described later in this chapter, have revealed additional organization patterns at scales between that of the 30-nm fiber and the whole genome. The earliest experiments revealed a power-law scaling between contact frequency and genomic distance, with a scaling exponent suggesting that DNA is well-described by the fractal globule model discussed above [48]. Moreover, at a genomic scale, one can partition the genome into two compartments based on proximity. One of these compartments, termed A, is associated with open, accessible, and active chromatin, and the other, termed B, is associated with closed and inactive chromatin [14, 49]. The exact nature of these compartments is still being elucidated; imaging studies have begun to suggest that compartments are physical structures in single cells [50], while the compartments themselves may be phases separated [51]. At higher resolution, these methods have revealed smaller domains within these compartments; typically organized at the 10 kb to 10 Mb scale, these so-called topologically associating domains (TADs) interact frequently within their boundaries and relatively infrequently with adjacent regions [52]. Their role is still unclear, and recent work has found that although they are evolutionarily conserved [53], they do not seem to have a strong effect on gene regulation [53, 54]. At sub-TAD scales, these methods have also revealed loops, typically anchored at domain boundaries, which are also hotspots for binding sites for the transcription

factor CTCF [52, 55]. These observations have recently given rise to loop extrusion models of chromosome organization [56], although this is still an area of intense discussion [57].

2.3 Methods to study genome structure

The findings discussed above were enabled by different technologies, and as a result there are multiple methods of interrogating the spatial nature of the genome, yielding different and sometimes incommensurable observations. We will discuss the two dominant methods. The first is a family of proximity ligation based techniques, first introduced by Dekker and colleagues and termed chromosome conformation capture (3C) [58]. In the original approach, DNA in formaldehyde-crosslinked cells is first digested with a restriction enzyme, diluted, and then subject to intra-molecular ligation, creating chimeric molecules based on spatial proximity of digested fragments. The abundances of specific chimeric fragments are then determined by quantitative PCR, thereby yielding a measurement of so-called contact frequency. Given a particular polymer model, as discussed above, contact frequency is then used as a surrogate for the 3D spatial distance between the two sequences [59].

The 3C method was further developed so that multiple loci could be assayed in a single experiment: so-called 4C methods permit the contact frequency of a single locus to be measured against all other genomic loci [60], and 5C methods allow many loci within about a megabase to be assayed against another population of similar size [61]. However, the coming-of-age of high-throughput DNA sequencing permitted the remarkable development of Hi-C by Lieberman-Aiden and coworkers, which permits investigators to assay contact frequencies genome-wide [14].The development of Hi-C led to a number of findings, discussed in detail above, including evidence of fractal globule behavior, as well as compartments, domains, and loops [62]. Moreover, the method has been adapted to single cells, revealing widespread structural heterogeneity even within a single cell type [63]. However, there remain questions regarding to what extent contact frequency is interchangeable with 3D distance[64, 65], which can lead to contradictions with the second methodology, fluorescence *in situ* hybridization (FISH)[66] .

FISH is a targeted method based on light microscopy and therefore permits direct spatial

measurements of genomic architecture. It is a relatively mature technology and has been used in a variety of contexts, with early experiments identifying chromosome territories with whole-chromosome paints [67], which were then used to examine their radial and relative spatial positioning as described in the previous section. More recently, developments in large-scale oligonucleotide synthesis have enabled the creation of large libraries of targeted probes to survey specific targets [68]. The resolution of the method is typically limited by the diffraction limit of light, ~200 nm, but ingenious approaches using serial hybridization [69, 70, 71] as well as integration with super-resolution microscopy [72, 73] have permitted sub-diffraction-limited high resolution studies of chromatin domains. However, being a targeted method, FISH relies on designed synthetic oligonucleotide probes to detect known sequences, prohibiting *de novo* discovery of genomic structural features. Moreover, standard FISH is not well-suited to detection of small insertions, deletions, or polymorphisms, collectively the largest source of genetic variation [74]. Nonetheless, FISH is both powerful and scalable, with the oligopaints technology benefiting from continuing improvements in DNA synthesis [75]. Recent work in human and mouse cells has shown how oligopaints [76, 77] and related methods (e.g. SeqFISH [78]) can be used for genome scale imaging, showing how these methods are well suited for applications requiring scalable *in situ* localization of known genomic sequences.

Chapter 3

Technology Development for *In Situ* Genomics

3.1 *In situ* ATAC-seq

3.1.1 Summary

The assay for transposase-accessible chromatin with sequencing (ATAC-seq) has emerged as an efficient method to profile epigenomes. ATAC-seq measures the accessibility of chromatin to hyperactive Tn5 transposase, such that sequencing adaptors are preferentially inserted into regions of accessible ("open") chromatin, with insertion frequency quantified by next-generation sequencing. Open chromatin, defined by ATAC-seq, is a good proxy for regulatory activity, and thus ATAC-seq reveals, genome-wide, the individual regulatory elements which are active in establishing and maintaining cell type, state, and fate.

However, ATAC-seq does not capture spatial information, and open chromatin defined by ATAC-seq is not placed in spatial context, leaving out crucial structural information as reviewed in Chapter 2. Encouraged by pioneering work demonstrating how ATAC-seq adapters can be inserted into fixed nuclei to visualize open chromatin in a non-sequence-specific manner [79], we sought to use *in situ* sequencing as the core of a sequence-specific approach to spatially resolved epigenome mapping.

3.1.2 Results and Discussion

We first devised an efficient scheme for *in situ* ATAC-seq library construction. In this scheme, Tn5 tranposase inserts sequencing adapters into fixed chromatin, preferentially inserting into open chromatin. Next, hairpins are hybridized to the inserted sequencing adapters, and the genomic DNA insert is circularized by gap-fill and ligation. The circularized insert can then be amplified by rolling circle amplification, to be read out by *in situ* sequencing (**Fig. 3-1**).

We found that this scheme worked immediately in HeLa cells, yielding hundreds of amplicons per nucleus. Additionally, cells were digested and sequenced, yielding libraries with insertion lengths characteristic of ATAC-seq (**Fig. 3-2**). Further, when applied to mouse brain tissue in an early experiment, we met with significant success, generating libraries across the slice (**Fig. 3-3**).

However, these early successes were misleading. We found that results were highly variable from batch to batch, in multiple hands. In cell culture, two common failure modes were observed: jackpotting, where few cells exhibited high yields, and generally low yield in most cells (**Fig. 3-4**). The former may not have been a failure mode per se, as there are potential biological explanations. However, given the laboriousness of the protocol, the managing the risk of low yield experiments was challenging, and would undoubtably hinder adoption of the technology. Further, early results in tissue could not be reproduced, with extremely low amplicon yields in follow-on experiments.

Instead of attemping to de-bug this state of affairs, we shifted our focus to unbiased genome structure mapping, which forms the basis of Chapter 3, and which adapts many of the ideas pioneered here around sequencing library construction. In hindsight, there are a range of potential explanations for the low yield experiments, including batch variability of the Tn5 enzyme, cell culture conditions, and fixation quality to name a few. Ideally, were we or others to revisit this project, an approach based on carefully building in intermediate read-out steps in the protocol would be of great utility. For instance, this could take the form of ATAC-see based imaging and sequencing readout of paired samples at each step to assess insertion efficiency, DNA damage and/or fixation strength via insert length, etc.

3.1.3 Methods

Libraries were prepared in cells and tissue samples as follows. First, samples were rinsed with 1x PBS and fixed with with 4% formaldehyde for 10m. Next, samples were permeabilized with 0.5% triton in 1x PBS for 30 m, and washed twice with 1x PBS. 2x TD-PBS buffer (0.6x PBS, 20 mM Tris pH 8, 10 mM $MgCl_2$, 20% DMF) was prepared.

50 μL transposition mix was prepared as follows (2 μL 30 μM unloaded Tn5 transposase, 25 μL TD-PBS, 23 μL H_2O and sample was incubated with in transposition mix for for 1 hour at 37°C.

10x annealing buffer was prepared (10x TE pH 7.5, 500 mM $NaCl_2$), and 10 μM hairpins were annealed separately in 1x annealing buffer. Sample was washed 3x 15 minutes in 0.01% SDS in 1x PBS at 60°C, and then washed twice with 1x PBS at room temperature. Hairpins were hybrized at 1 μM in 2x SSC for 2 hours at 37°C, then washed twice with 1x PBS.

A 50 μL ligation mix was prepared (5 μL 10x Ampligase buffer (Lucigen), 2.5 μL T4 polymerase (NEB, 3 U/μL), 6.25 μL Ampligase (Lucigen, 5 U/μL), 1.25 μL 10 mM dNTPs (NEB), 35 μL water. Samples were inclubated in ligation mix for 30 minutes at 37 °C, and washed twice with PBS.

RCA primer was hybridized to hairpins for 3 hours at 0.5 μM in 20% formamide, 2x SSC at 37 °C. A 100 μL RCA mix was prepared (10 μL 10x Phi29 polymerase buffer (Enzymatics), 10 μL Phi29 (10 U/μL) (Enzymatics), 2.5 μL 10 mM dNTPs (Enzymatics), 0.5 μL 4 mM aminoallyl dUTP (Thermo Fisher), 77 μL water. Samples were incubated in RCA mix overnight at 30 °C, and resulting amplicons were imaged with a hybridization probe.

3.2 *In situ* genomics for long-read DNA sequencing

3.2.1 Summary

The development of DNA sequencing technologies have repeatedly revolutionized biology. Automated Sanger sequencing enabled the first draft map of the human genome [80], next-generation sequencing democratized the sequencing and assembly of many human genomes [81], and long-read sequencing recently facilitated the first complete map of a human genome,

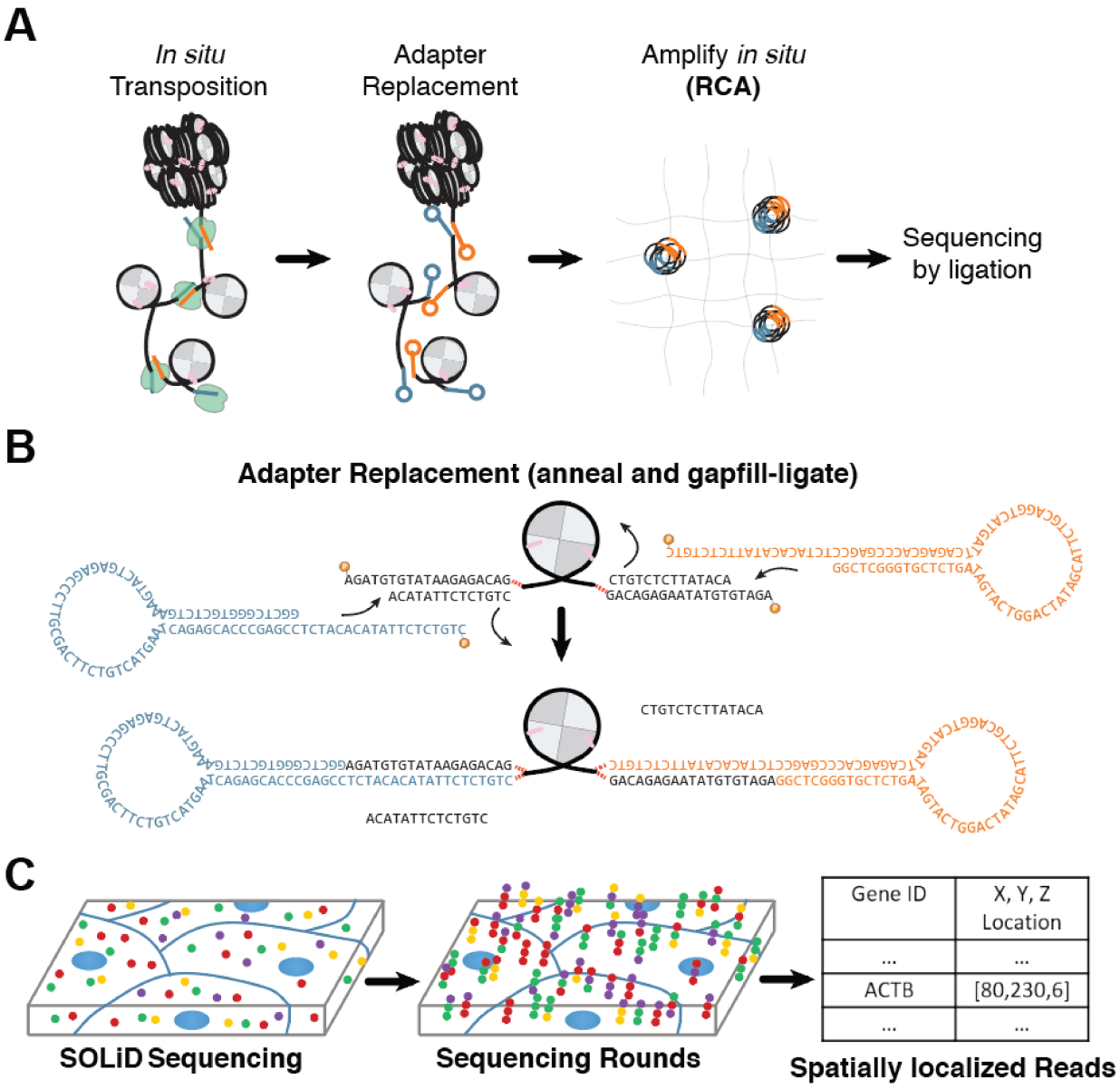

Figure 3-1. ***In situ* ATAC-seq concept**.**(a)** Schematic of sample preparation workflow for in situ ATAC-seq. In situ transposition is first performed on fixed cells or tissues, followed by adapter replacement and circularization (see **(b)**). Fragments are amplified using rolling circle amplification (RCA). **(b)** Schematic representing fragment circularization using ligation of circular adapters to transposed accessible fragments. **(c)** Schematic of sequencing spatially resolved fragments using SOLiD sequencing chemistry.

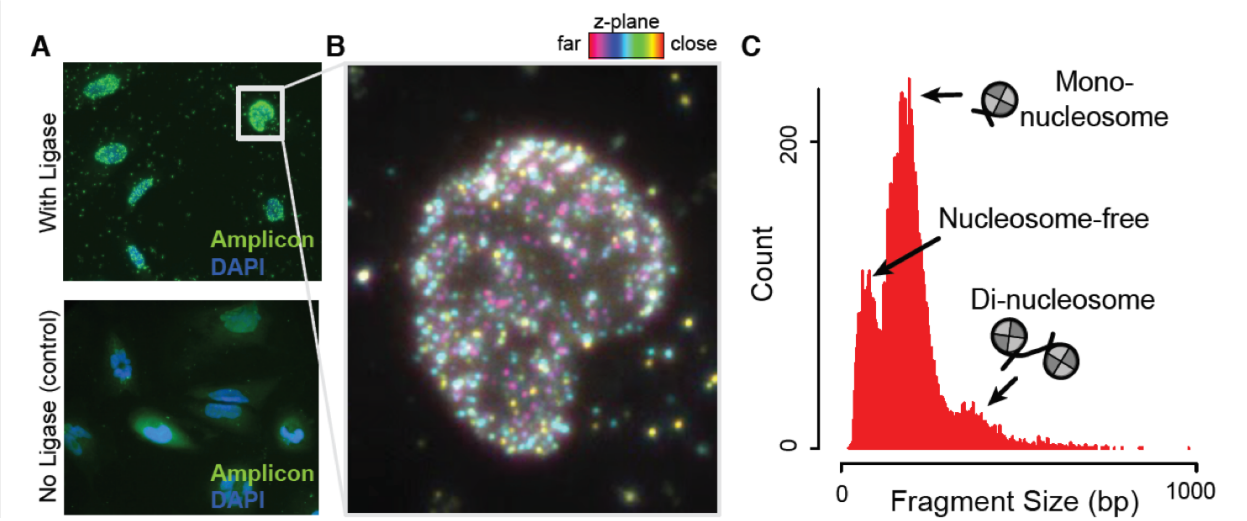

Figure 3-2. ***In situ* amplified epigenome.(a)** HeLa cells with in situ amplified ATAC-seq fragments (green) and DAPI (blue), (top) cells with ligase and (bottom) cells with no ligase control. In situ amplicons are visualized using fluorescent DNA hybridization. The image represents a compilation (max) of a z-image-stack. **(b)** A single HeLa nucleus, each pixel is colored by the maximum fluorescence at each z-image. **(c)** Paired-end sequencing of in situ amplified material on an Illumina HTS sequencer provides a fragment-size distribution similar to previous ATAC-seq studies.

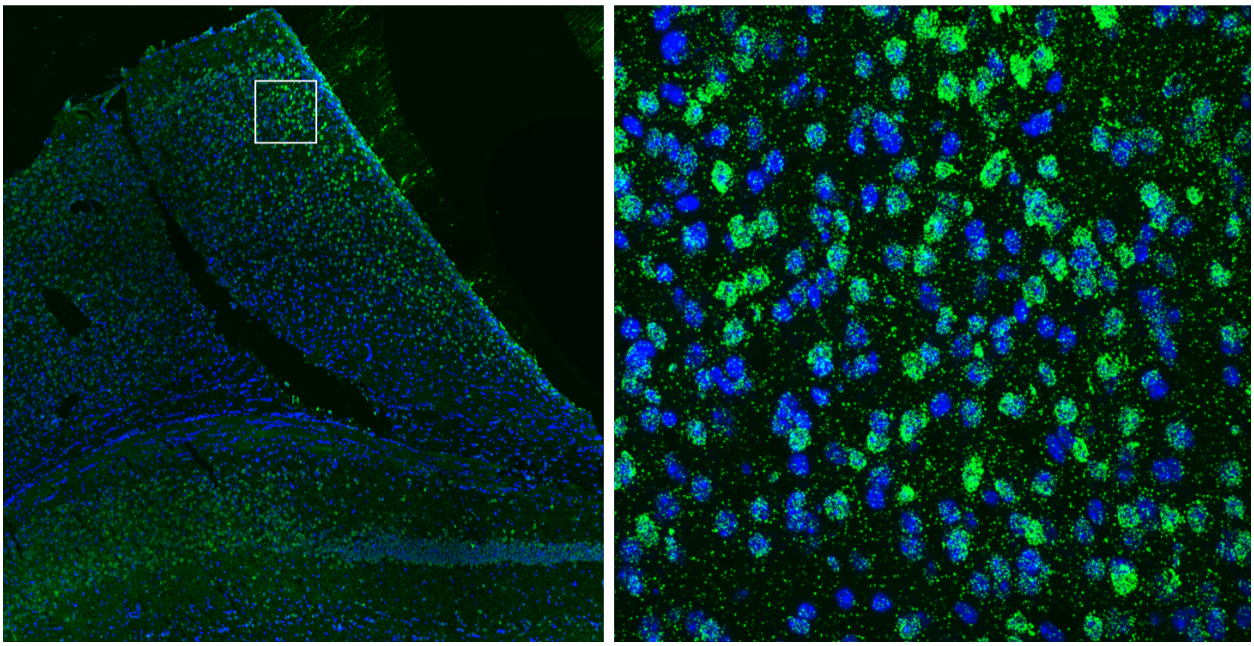

Figure 3-3. ***In situ* ATAC-seq in tissue**. Proof-of-concept experiment demonstrating high yield *in situ* ATAC-seq library construction in mouse cortex; however, these results could not be reproduced in follow-on experiments.

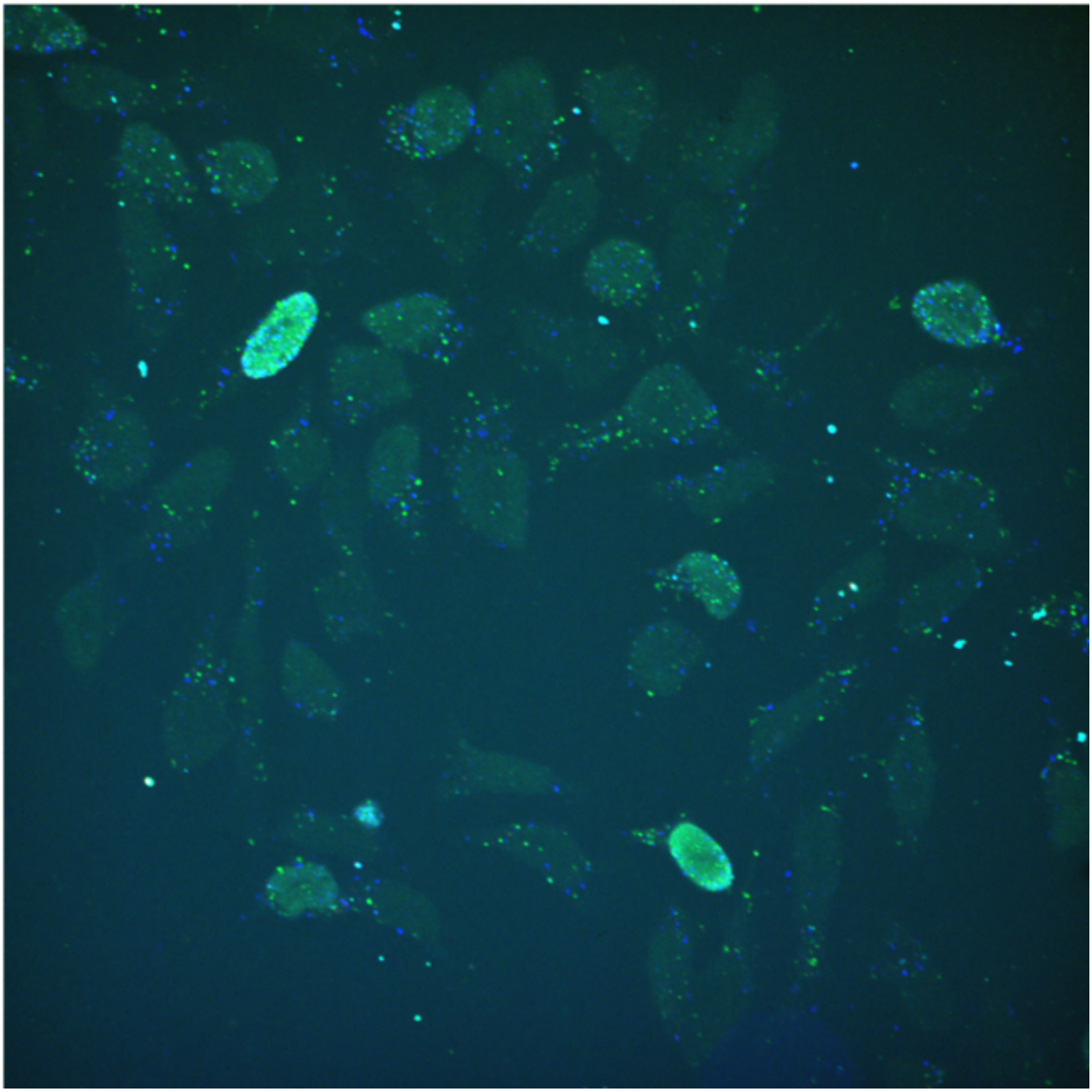

Figure 3-4. **Representative *in situ* ATAC-seq failure modes**. Representative image demonstrating both main failure modes for *in situ* ATAC-seq, jackpotting and low yield.

telomere-to-telomere [82]. Nonetheless, the appetite for sequencing continues to increase, and there remains a pressing need for DNA sequencing methods which achieve longer read lengths and reduced error rates at faster speed and lower cost. When we developed the concept of spatially localizing longer DNA molecules using spatially indexed UMIs, potentially combined with ExM, it naturally begged the question of whether such an approach could be harnessed for the purposes of improved DNA sequencing.

Although DNA adopts a compact coiled conformation in solution, it can be elongated and immobilized on a surface, a process which applies a detectable property of spatial position to each base in the chain [83]. The idea of using this spatial information in DNA sequencing is venerable [84, 85], as in theory it promises very long reads, on the order of hundreds to thousands of kilobases [86], with single-molecule precision, mitigating issues such as amplification errors [87]. However, current DNA linearization methods have two drawbacks which limit the genomic resolution at which the linearized DNA can be analyzed. First, linearized DNA is typically analyzed using optical microscopy, and therefore the smallest adjacent genomic features that can be uniquely distinguished are limited by the classical diffraction limit of light. This is about 1 kilobase for a fully elongated molecule, which is insufficient to resolve many genomic structural variations [88]. Second, once DNA is linearized and immobilized on a surface, it cannot be subject to efficient enzymatic reactions, which are inhibited by the solid phase[1] [90] . Due to the fact that the most practical sequencing chemistries are based on sequential rounds of enzymatic processing [81], linearization methods cannot achieve base-pair level resolution of resolve loci, thus failing to measure the most common type of genetic variation.

We revisited these limitations in the context of our methods development, i.e. exploring the consequences of combining DNA elongation with ExM and UMI mapping. First, supposing a fully elongated molecule could be efficiently combined with standard ExM, a resolution of ~200 bp per diffraction-limited spot would be achievable. This length is compatible with standard read lengths on very high throughput short-read sequencers. Thus, supposing a sequencing library could be efficiently constructed such that one UMI-tagged amplicon was

[1]Some surfaces are compatible with limited enzymatic activity; however, these solid-phase reactions are still inefficient compared to the liquid phase [89].

present at each spot, *in situ* sequencing could be used to spatially index UMIs. Additionally, supposing efficient amplicon recovery, each 200 bp + UMI fragment could be spatially indexed. In the best case scenario, this would permit read lengths on the order of megabases, with per-base error rates tethered to the performance of next-gen sequencers, on the order of 0.1%. This would have been a substantial improvement over existing long-read technologies, which at the time had read lengths of ~ 100 kb and error rates of ~5% [91]. To this end, we performed proof-of-concept experiments to explore the viability of this approach.

3.2.2 Results and Discussion

We first developed a workflow to expand DNA elongated on a solid support. Although methods for expanding nucleic acids have been developed for expansion microscopy [92], they are intended for biological specimens in a cellular or tissue context: chemically fixed biomolecules within the specimen are covalently embedded in a swellable hydrogel; the ultrastructure of the specimen is then digested, and the hydrogel is expanded, physically separating the biomolecules. However, throughout this process, individual anchored biomolecules remain fixed in their native, compact, conformational state. Thus, even after expansion, individual biomolecules are localized to within a single post-expansion diffraction limited spot. As a result, although the identity and position of individual biomolecules can be recovered, their spatial structure cannot [92].

As elongated DNA is already in a fully extended state, an expansion protocol must involve 1) hydrogel anchoring, 2) DNA detachment from the solid support, and 3) DNA fragmentation to facilitate further extension. A key insight was realizing that the anchoring agent in ExFISH, a nitrogen mustard derivative, can facilitate a DNA break at a subset of bases in the presence of a strong base, such as NaOH, similar to the Maxam-Gilbert chemistry used in first-generation sequencing [93], before it was superseded by Sanger sequencing. When tested, this insight was followed by the surprising discoveries that an NaOH treatment could 1) cleave DNA from a silane support, and 2) convert a subset of acrylamide groups in an acrylamide gel to acrylate, converting an acrylamide overlay to an expansion microscopy gel in situ. Thus, we had the serendipitous good fortune to stumble on a single protocol step that would 1) fragment DNA, 2) detach the DNA from the solid support, 3) prepare an

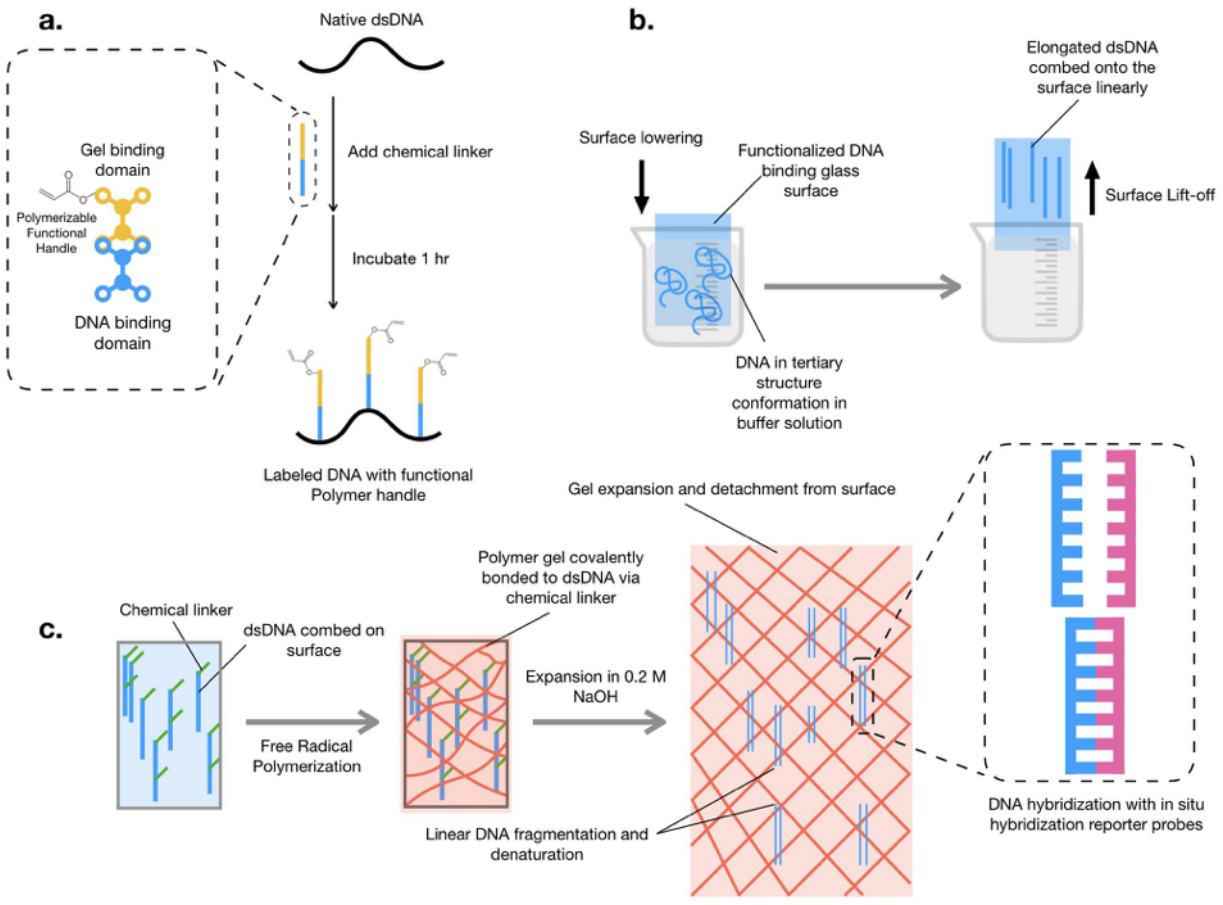

Figure 3-5. **EE-DNA concept and schematic A** DNA functionalization with a polymerizable moiety (e.g. Label-X) in solution, **B** linearization of functionalized DNA on a solid support, **C** (left, middle) formation of an acrylamide hydrogel overlay on the solid support, (middle, right) depicts simultaneous cleavage of the overlay from the support, fragmentation of hydrogel-embedded DNA, and physical expansion of the hydrogel.

expansion microscopy gel. This formed the backbone of our EE-DNA end-to-end protocol (**Fig. 3-5**) concept, which we validated on lambda phage DNA, due to its well-defined length distribution when elongated [83]. DNA was first extended on a solid vinyl silane support (**Fig. 3-6**), expanded, and detected using DNA FISH (**Fig. 3-7**). Measurement of phage lambda before and after expansion demonstrated effective elongation by expansion (**Fig. 3-8**).

We next proceeded to demonstrate that enzymatic reactions can occur on elongated DNA, which is crucial to achieve DNA sequencing. We found that elongated DNA could be detected by random hexamer extension (**Fig. 3-9**) and terminal transferase tailing (**Fig. 3-10**). However, efforts to circularize and amplify elongated DNA by adapting protocols

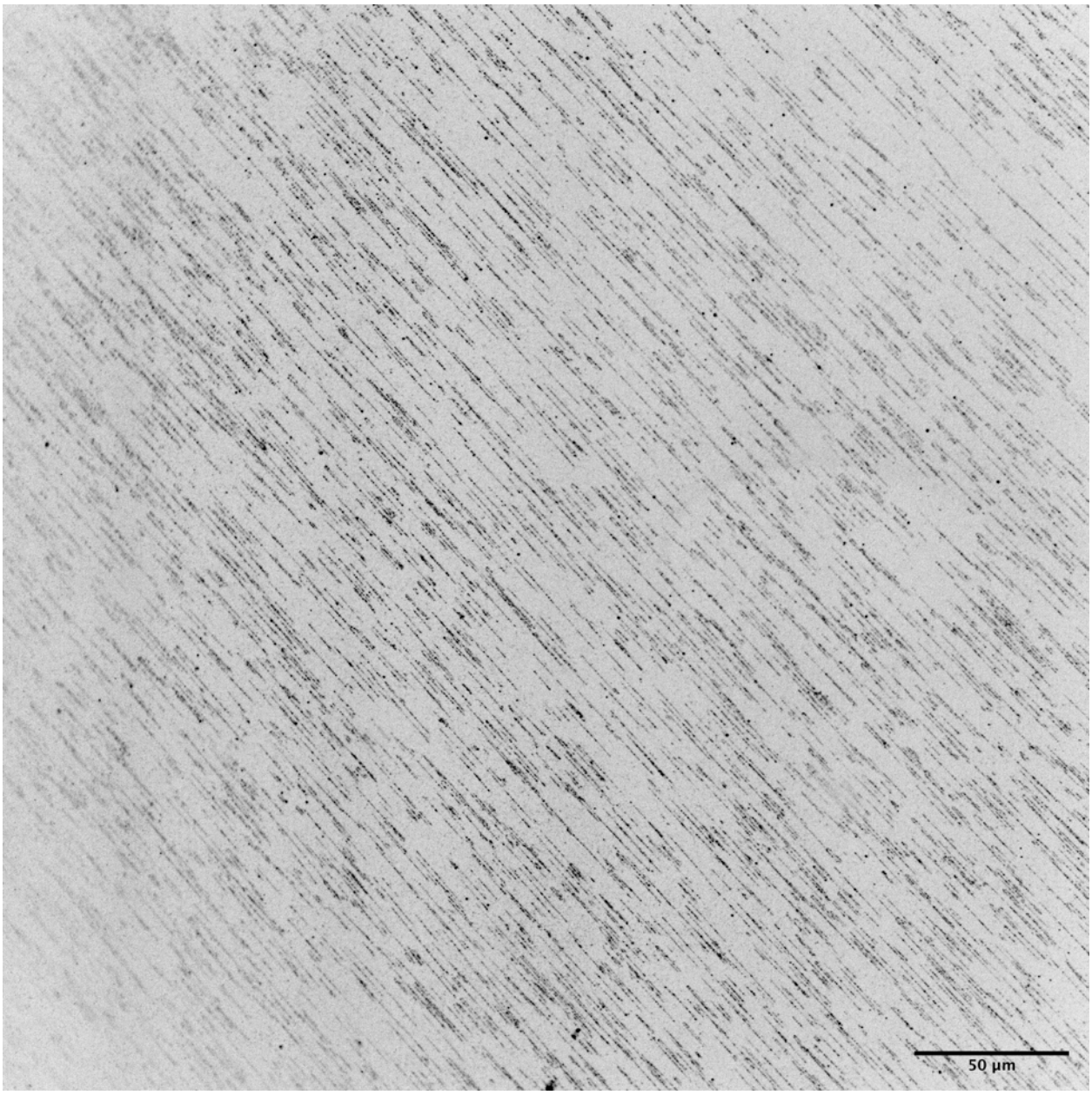

Figure 3-6. **FISH detection of hydrogel-embedded lambda phage DNA, pre-expansion**

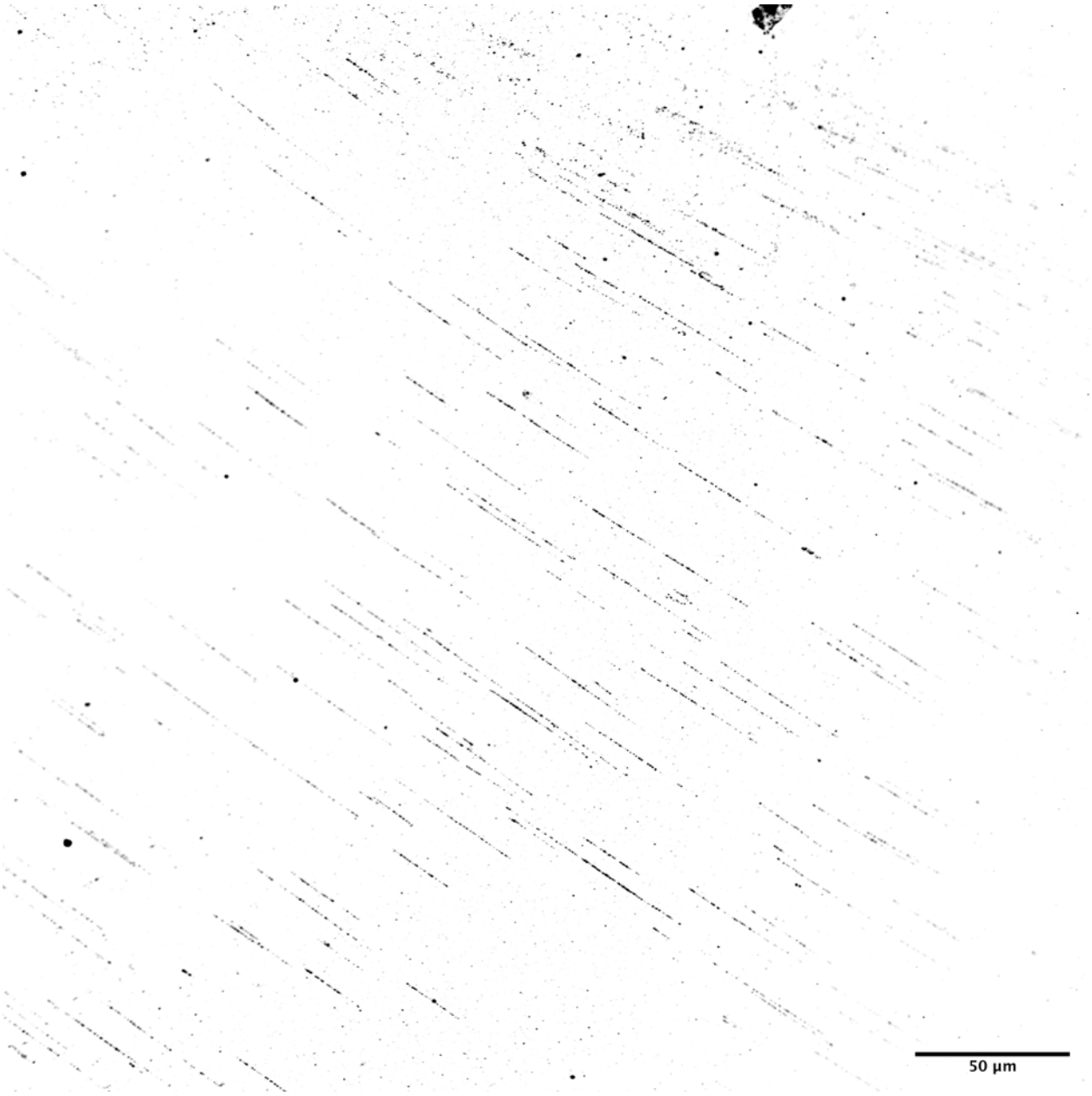

Figure 3-7. **FISH detection of hydrogel-embedded lambda phage DNA, post-expansion**

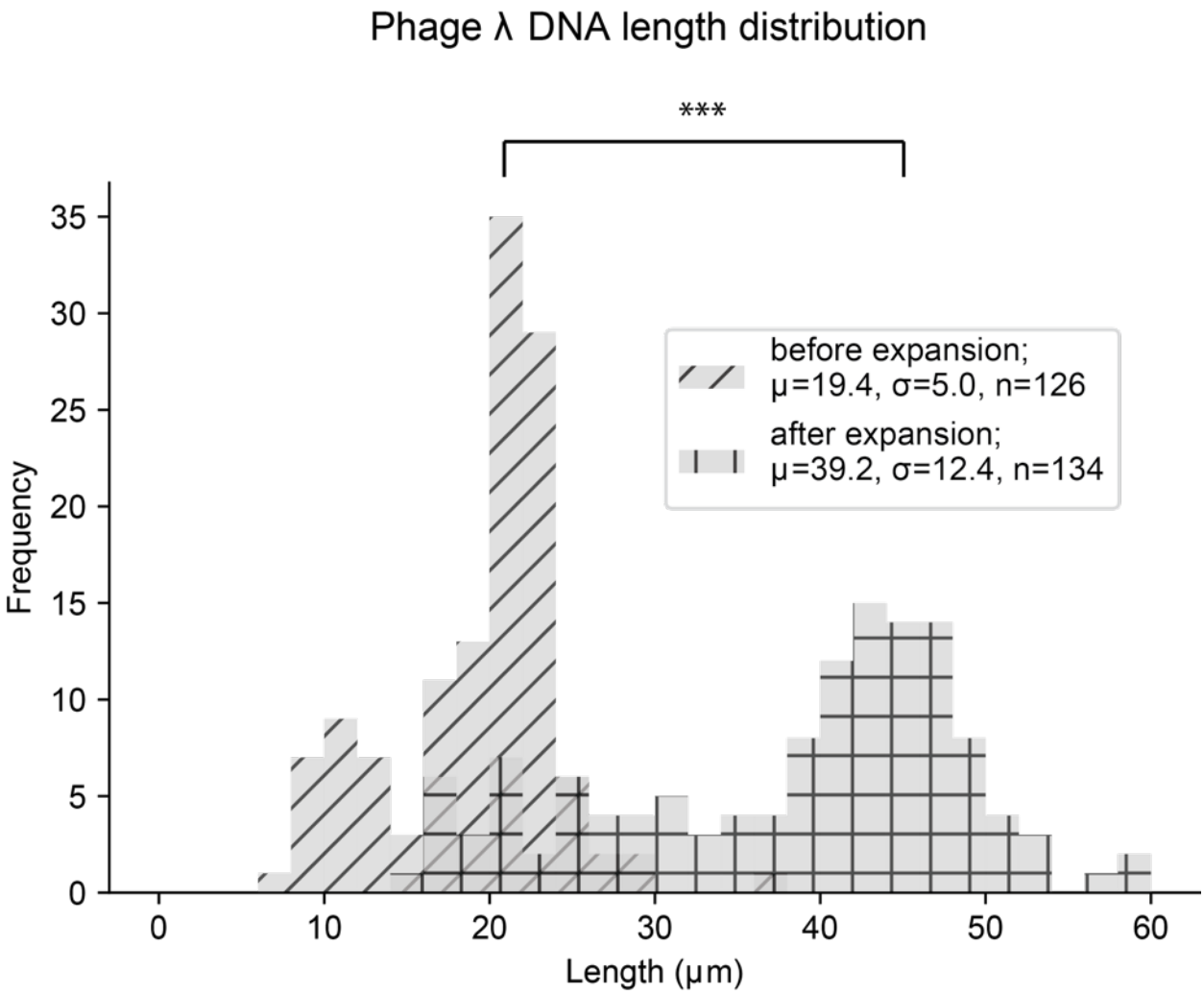

Figure 3-8. **Comparison of the length distribution of lambda phage DNA before and after expansion**. Lengths are inferred from the longest dimension of spatially proximal puncta in microscopic images. Prior to expansion, the distribution exhibits a sharp peak at 20 μm (the length of fully extended lambda phage) with a smaller peak at 10 μm (half the length, caused by DNA bound to the solid support during linearization by both of its extremities, see cite). After twofold linear expansion, lambda phage lengths significantly and are proportionally longer (*** $p < 0.001$, KS test).

used in ExSeq [94] were unsuccessful without exception.

We performed proof of concept experiments, demonstrating 1) DNA elongated on a solid substrate can be cleaved and further elongated with expansion microscopy, 2) selected enzymatic reactions can occur efficiently on expanded and elongated DNA (EE-DNA). We demonstrated (1) by comparing the length of elongated phage lambda DNA molecules to EE-lambda-DNA molecules, and we demonstrated (2) by detecting EE-lambda-DNA enzymatically by both terminal transferase tailing and random hexamer extension. However, we were unable to successfully construct an amplified DNA sequencing library, and could not identify a method to resolve this blocking issue. As the impact of this approach depends on sequencing, not just enzymatic detection, we chose not to continue this line of development. However, should this issue be resolved or circumvented, there remains the possibility of another revolution in DNA sequencing technology by using spatial information from elongated DNA, and others continue to push in this direction[95, 96].

3.2.3 Materials and Methods

3.2.3.1 Expansion and Detection of Lambda Phage DNA by FISH

Steps 1-3, DNA Functionalization, Linearization, Immobilization, and Embedding

To demonstrate the method, we have expanded, labeled, and imaged single molecules of linearized lambda phage DNA. First, to functionalize the DNA, a bifunctional crosslinker ("Label-X") bearing an alkylating moiety and an acryloyl moiety was produced from by coupling the small molecule Acryloyl-X (Thermo) to the small molecule Label-IT Amine (Mirus Bio LLC) as described in [92]. Next, this crosslinker was reacted at to full length lambda-phage DNA (NEB) (5 μg lambda-phage DNA and 1:20 Label-X in Buffer A (Mirus) at room temperature for one hour with agitation, followed by storage at -20°C).

To linearize, immobilize, and embed DNA, Label-X modified DNA was diluted to 10 pM in 150 mM MES buffer, pH 5.5. It was then elongated and immobilized on a hydrophobic vinyl silane modified coverslip (Biosurfaces, Inc.) using the molecular combing technique as described in [86]. Monomer solution (1X PBS, 4% (w/w) acrylamide, 0.2% (w/w) N,N'-Methylenebisacrylamide was mixed fresh. 0.1% (w/v) of ammonium persulfate (APS) and

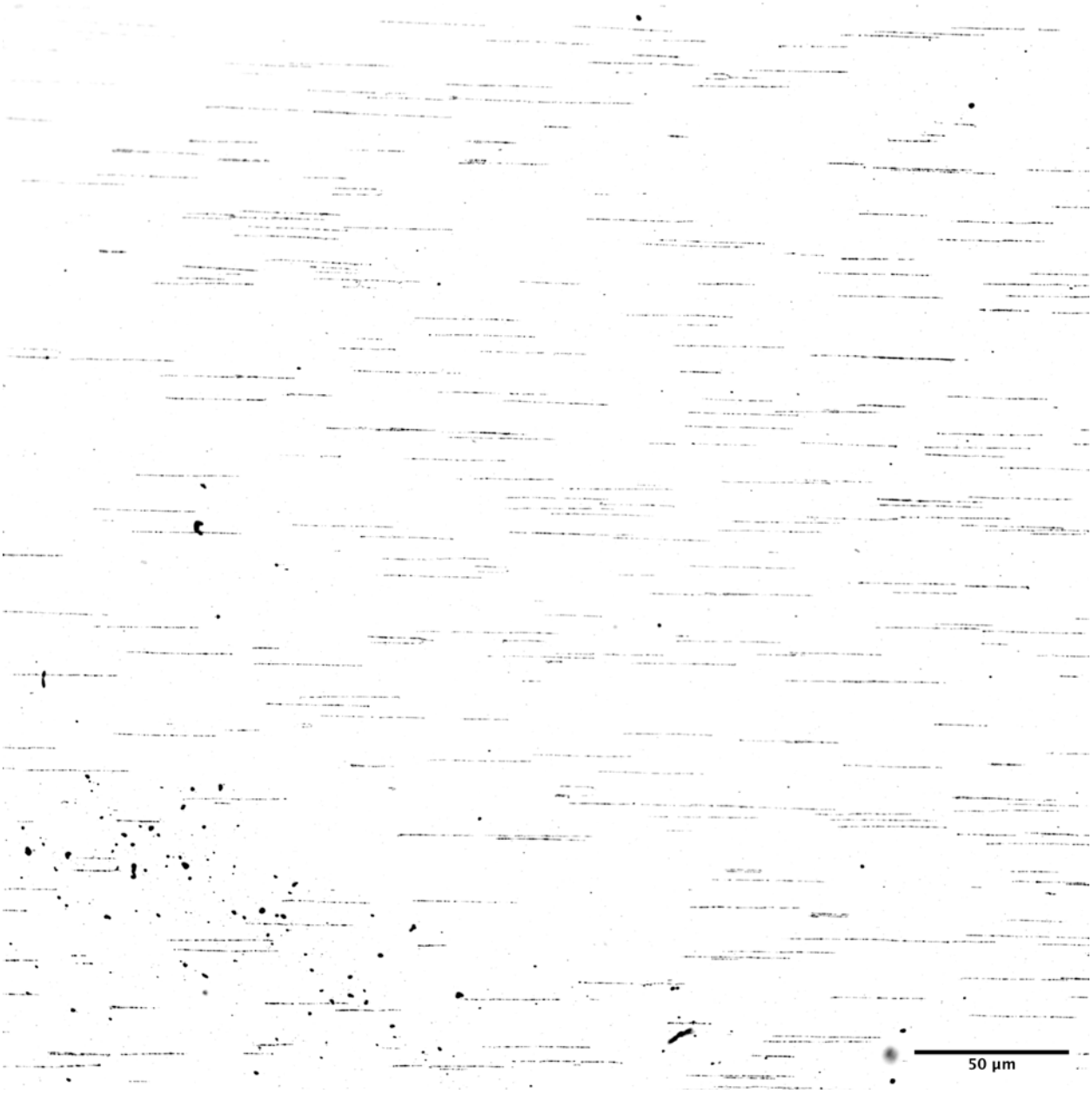

Figure 3-9. **Detection by random primer extension of expanded lambda phage DNA**

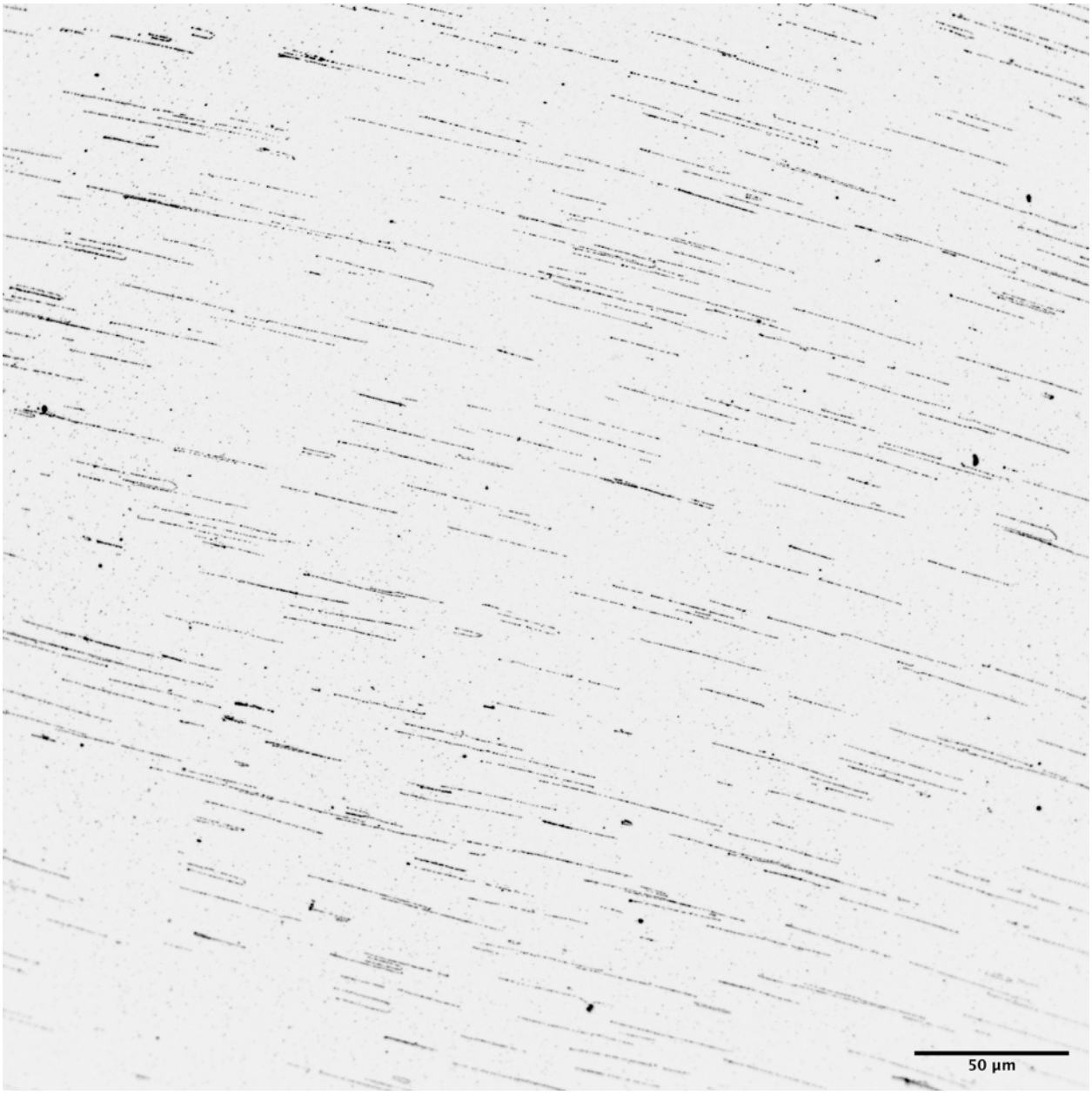

Figure 3-10. **Detection by terminal transferase tailing of expanded lambda phage DNA**

tetramethylethylenediamine (TEMED) were added to the monomer solution up to 0.2% (w/w) each and immediately brought into contact with the surface immobilized DNA. The solution was sandwiched between a second glass coverslip and incubated in a humidified chamber at 37° C for one hour, resulting in an acrylamide surface overlay, approximately 50 μm thick and adherent to the vinyl silane surface.

Steps 4-6, Surface Detachment, DNA fragmentation, Hydrogel Conversion

To cleave the overlay from the surface while retaining DNA in the gel phase, fragment the DNA, convert the gel to a swellable polymer, and denature the DNA, the overlay-coverslip sample was incubated in 0.2 M NaOH overnight. Treatment with a strong base achieves these four objectives simultaneously. First, it cleaves the surface silane bonds, reversing DNA surface immobilization, and thus detaches the DNA-hydrogel composite from the glass surface[97]. Second, it permits controlled fragmentation of the DNA. When DNA is modified by Label-X, the majority of sites are functionalized as described, and a minority of sites are damaged and rendered abasic [98], and abasic sites are efficiently cleaved by a strong base [99]. (Note that the ratio of polymerizable DNA adducts and abasic DNA sites can in principle be modulated by doping Label-X with unmodified Label-IT, thus controlling the degree of fragmentation). Third, the NaOH treatment denatures DNA [100], making it accessible for downstream hybridization or enzymatic reactions which require single stranded DNA. Fourth, it converts a portion of acrylamide hydrogel side chains into acrylate, which, as validated [101], causes the gel to expand isotropically when dialyzed with water or low salt content buffer.

Step 7: DNA labelling

To facilitate microscopic detection of DNA, an in situ hybridization was performed on the DNA-gel sample. Biotinylated lambda phage FISH probe (Enzo Life Science) was diluted to 200 ng in hybridization buffer (Molecular Probes), and the sample was immersed in this solution. The solution was briefly heated to 80°C for 3 minutes, incubated at 37°C overnight, and finally washed 3x for 30 minutes in wash buffer (Molecular Probes). The sample was then incubated with Cy5-Streptavidin (1:200 in 1x PBS) and washed 2x for 30 minutes in PBS.

Steps 7-9, Hydrogel Expansion and Microscopic Detection

The hydrogel was expanded approximately twofold from its original size by washes in 1X PBS during DNA labelling. (To expand further, the sample can be exchanged into a buffer with lower salt content, such as 0.1x PBS). The sample was imaged microscopically in 1x PBS using an Andor spinning disk (CSU-X1 Yokogawa) confocal system with a 40 $\times$ 1.15 NA water objective on a Nikon TI-E microscope body. Cy5-Streptavidin was excited with a 640 nm laser with 685/40 emission filter.

3.2.3.2 Enzymatic Detection of Lambda Phage DNA with Polymerase

Enzymatic labelling with a polymerase, rather than labelling by hybridization, was performed in order to demonstrate how hydrogel embedding, unlike the solid phase, permits facile enzymatic analysis of linearized DNA. First, steps 1-6 were performed as described in Example 1. Following overnight NaOH incubation, the sample was washed 2x 5 minutes in 1x PBS, then immediately exchanged into a primer-extension reaction (25 μM random hexamers (Thermo Fisher), 100 μM of each of dATP, dTTP, dGTP (NEB), 40 μM biotin dCTP (Thermo Fisher), 1:50 Klenow Fragment (3'$\rightarrow$5' exo-) (NEB) in 1x NEBuffer 2). The reaction was incubated at 37°C for one hour, then washed 3x 5 minutes in PBS. Newly synthesized DNA was detected by a mouse anti-biotin antibody (1:200 in 1x PBS for 30 minutes, Abcam; ab201341) amplified by an Alexa Fluor 488 Goat anti-mouse secondary (1:200 in 1x PBS, Thermo Fisher). Steps 7-9 were then performed as described above with the exception that Alexa 488 was excited with a 488 nm laser, with 525/40 emission filter.

3.2.3.3 Enzymatic Detection of Lambda Phage DNA with Terminal Transferase

An additional type of enzymatic labelling with a terminal transferase was performed to demonstrate that the enzymatic labelling in Example 2 is not a special case. First, steps 1-6 were performed as described in Example 1. Following overnight NaOH incubation, the sample was washed 2x 5 minutes in 1x PBS, then immediately exchanged into a primer-extension reaction (25 μM random hexamers (Thermo Fisher), 100 μM dNTPS (NEB), 1:50 Klenow Fragment (3'$\rightarrow$5' exo-) (NEB) in 1x NEBuffer 2). The reaction was incubated at 37°C for one hour, then washed 3x 5 minutes in PBS. The sample was then exchanged into an end-tailing reaction (100 μM biotin dCTP (Thermo Fisher), 1:20 terminal transferase

(NEB), 0.25 mM $CoCl_2$ in 1x Terminal Transferase Reaction Buffer (NEB). This reaction was incubated at 37°C for one hour, then washed 3x 5 minutes in PBS. Detection was then performed as described above.

Chapter 4

In situ genome sequencing within intact biological samples

This chapter is adapted from *In situ genome sequencing resolves DNA sequence and structure in intact biological samples*, by Andrew C Payne, Zachary D Chiang, Paul L Reginato, Sarah M Mangiameli, Evan M Murray, Chun-Chen Yao, Styliani Markoulaki, Andrew S Earl, Ajay S Labade, Rudolf Jaenisch, George M Church, Edward S Boyden, Jason D Buenrostro, and Fei Chen, published in Science, 371(6532), (2021). [102].

A.C.P. and P.L.R. developed the protocol and performed experiments. Z.D.C. developed the computational processing pipeline. A.C.P., Z.D.C., P.L.R., S.M.M., J.D.B., and F.C. performed analyses. E.M.M., C.-C.Y., and A.S.L. performed supplementary experiments. S.M. performed embryo preparation under the supervision of R.J. A.S.E. designed the interactive Shiny app. A.C.P., Z.D.C., P.L.R., E.S.B., J.D.B., and F.C. wrote the manuscript with input from all authors. G.M.C., E.S.B., J.D.B., and F.C. supervised the study.

4.1 Introduction

The genome of an organism encodes not only its genes, but also principles of spatial organization that regulate gene expression and control cellular function [103, 104]. Accordingly, mapping spatial genome organization at high resolution is important for understanding its diverse regulatory roles in health, disease, and development [105, 106]. Principles of genome architec-

ture have mostly been uncovered by methods based on DNA sequencing of chromatin contacts [13], such as Hi-C [107], and methods which probe targeted genomic loci using microscopy, such as DNA fluorescence *in situ* hybridization (FISH) [108]. Hi-C applied to populations of cells has revealed genome-wide organizing principles [109, 110, 111, 112, 113], and single-cell variations have uncovered cell-to-cell heterogeneity [114, 115, 116, 117, 118]. DNA FISH has similarly revealed genome architecture at single-cell resolution [119, 120]. More recent studies have shown how these approaches can complement each other by imaging Hi-C defined features in single cells, characterizing their heterogeneity, and validating inferred differences in chromatin conformation within and across cell types [121, 122, 123, 124, 125, 126, 127]

However, these methods cannot currently be applied jointly on the same cell, and a method to simultaneously sequence and image genomes in single cells is lacking. Efforts which combine Hi-C with microscopy [117, 128], or efforts which make FISH more like sequencing via single-nucleotide polymorphism (SNP) specific probes [129, 124], have broken important conceptual ground, but they remain limited in their imaging or sequencing throughput. Accordingly, questions requiring both genomic and spatial analysis in single cells have been difficult to address.

4.2 *In situ* genome sequencing workflow

Here we present a method for *in situ* genome sequencing (IGS). IGS enables DNA sequencing directly within intact biological samples, spatially localizing genome-wide paired-end sequences in their endogenous context and thus bridging sequencing and imaging modalities for mapping genomes. Our *in situ* sequencing workflow introduces innovations in three phases: *in situ* library construction, multimodal sequencing of libraries, and computational integration of spatial and genetic information.

In the first phase, we create an *in situ* sequencing library within fixed samples by amplifying an untargeted sampling of the genome in its native spatial context. To do this, we fix and treat samples using methods optimized for DNA FISH [129, 130]. Next, we use Tn5 transposase to randomly incorporate DNA sequencing adaptors into fixed genomic DNA by *in situ* transposition, preserving genomic fragments in their native spatial positions

[131]. We circularize these fragments *in situ* by ligation of two DNA hairpins containing a unique molecular identifier (UMI) and primer sites used for subsequent multi-modal DNA sequencing (**Fig. 4-1A ii, iii**). We then clonally amplify the resulting circular templates by rolling circle amplification, yielding *in situ* DNA sequencing libraries with up to thousands of spatially-localized amplicons per nucleus (**Fig. 4-2**). We also developed a method for modulating the effective density of sequencing libraries to optimize the number of resolvable amplicons (**Fig. 4-3**). Together, this provides an approach to clonally amplify untargeted samples of a genome, creating approximately 400-500 nm sized features for *in situ* sequencing (**Fig. 4-4)**.

In the second phase of our workflow, we sought to use reported [133] *in situ* sequencing protocols to determine the sequence and 3D positions of amplicons. However, current *in situ* sequencing methods yield short single-end reads (at most 30 bases), and are limited by imaging time [133]. This poses a challenge for genome sequencing: the human genome encodes 3 billion bases and includes highly repetitive regions, requiring long paired-end sequencing reads to resolve many regions of the genome. To address this challenge, we combined *in situ* sequencing with high-throughput paired-end DNA sequencing. To do this, we first read amplicon-specific UMIs within fixed samples using sequential rounds of *in situ* sequencing by ligation (SBL) and fluorescence imaging [133] (**Fig. 4-1B i**). Immunostaining followed by additional cycles of imaging may also be performed following *in situ* sequencing. We then dissociate the *in situ* amplicons and amplify them using PCR to produce an *in vitro* sequencing library (**Fig. 4-1B ii, iii**), which we sequence on a conventional Illumina sequencer (henceforth referred to as *ex situ* sequencing) to obtain 150 bp paired-end genomic reads tagged with *in situ* sequenced, spatially-resolved UMIs. This multimodal sequencing strategy allows us to preserve spatial information while leveraging the accuracy and read-length of paired-end sequencing on the Illumina platform, which is crucial for aligning individual reads to millions of unique genomic loci.

In the third phase, we computationally match *ex situ* paired-end sequencing reads to *in situ* amplicon positions. Briefly, we deconvolve, register, and normalize fluorescence images in order to resolve the 3D centers of amplicons across multiple rounds of imaging (**Fig. 4-5**, [130]). We then quantify the fluorescence signal of each UMI-associated amplicon across

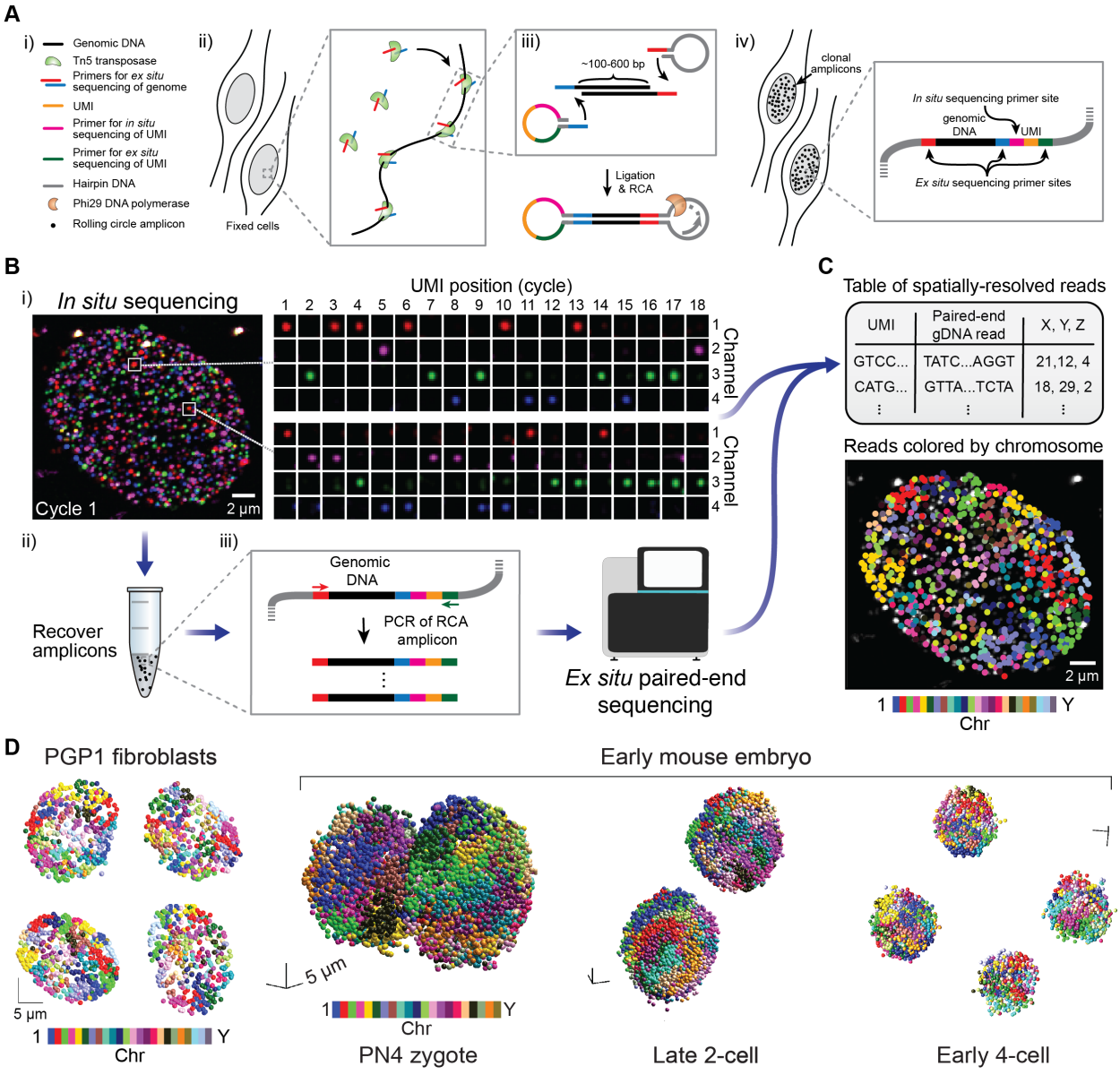

Figure 4-1. **Method for *in situ* genome sequencing**. (**A**) *In situ* genomic DNA library construction. i) Legend. ii) Adaptor insertion. iii) Insert circularization by hairpin ligation, followed by *in situ* rolling circle amplification (RCA). iv) Clonal amplicons contain primers for *in situ* and *ex situ* sequencing. (**B**) Workflow for *in situ* genome sequencing. i) *In situ* sequencing localizes unique molecular identifiers (UMIs). 4-channel imaging of two representative amplicons over 18 rounds of *in situ* sequencing. ii) Amplicon dissociation following *in situ* sequencing. iii) PCR and *ex situ* sequencing of amplicons associates genomic sequences with UMIs. (**C**) Top: paired-end sequences are spatially localized by integrating *in situ* and *ex situ* sequencing data. Bottom: matched reads, colored by chromosome, are overlaid on their imaged amplicon library (below). (**D**) *In situ* sequenced nuclei from cultured fibroblasts and intact embryos at the PN4 zygote, late 2-cell, and early 4-cell stages, with spatially-localized reads colored by chromosome.

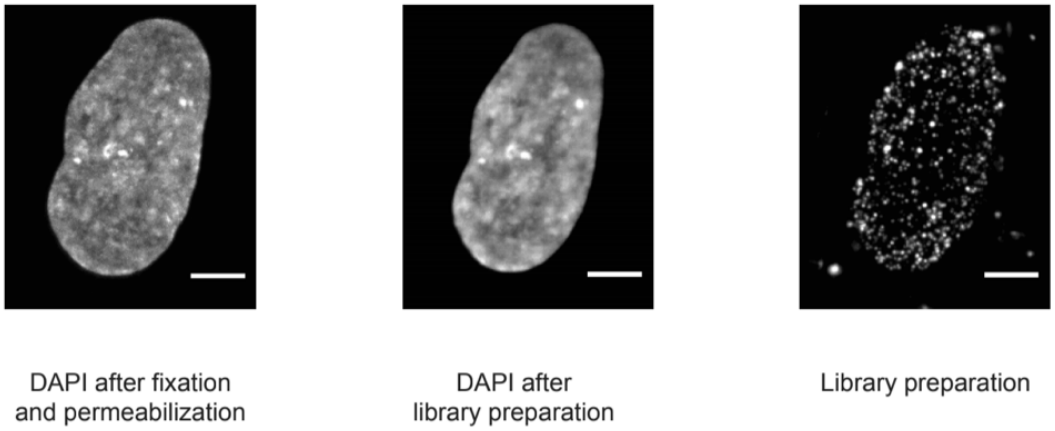

Figure 4-2. **DAPI staining before and after library preparation**. DAPI staining in a PGP1f nucleus after fixation and permeabilization (left) and after library preparation (middle) shows that morphological features are well-preserved during library preparation. Slight shrinkage of nuclei was observed over the course of library preparation, in line with what is seen during DNA FISH protocols [132]. (Right) IGS library in the same nucleus, stained with amplicon visualization oligo. Scale bars, 5 μm.

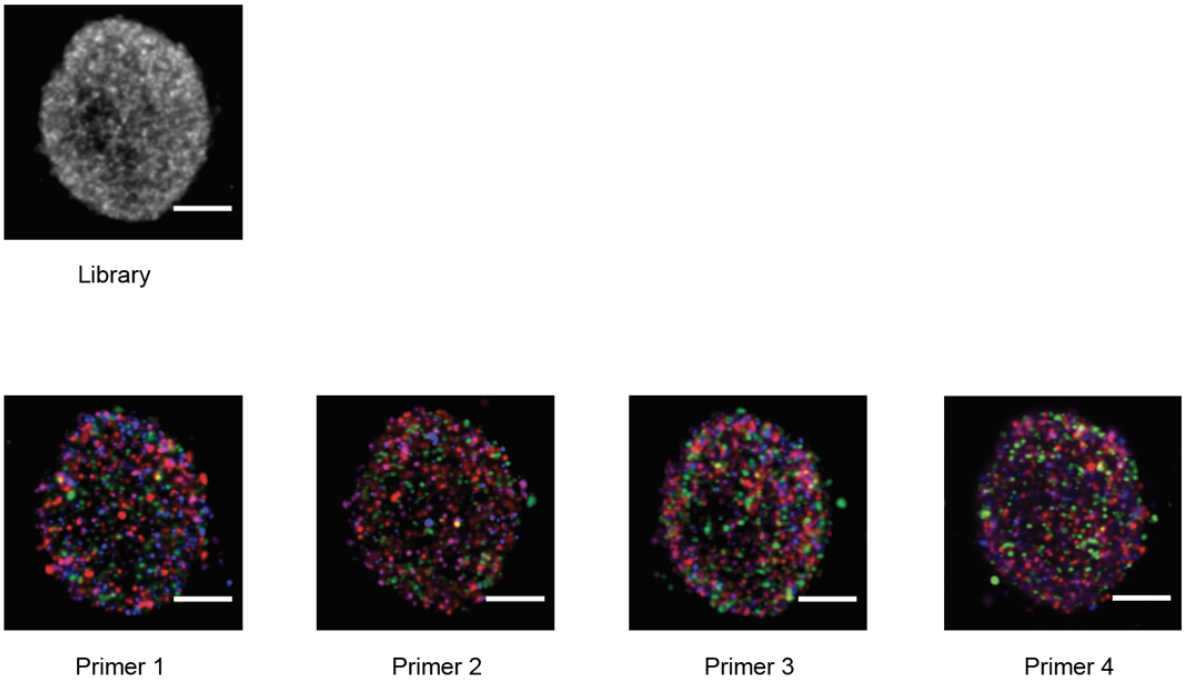

Figure 4-3. **Sub-sampled sequencing of a high-yield library**. (Above) A single z-plane through a representative PGP1f nucleus in a high-yield *in situ* genomic sequencing library stained with amplicon visualization oligo. (Below) One base of *in situ* sequencing in the same nucleus using each of four orthogonal *in situ* sequencing primers. The effective density of high-yield libraries can be modulated by using a subset of these primers. All amplicons are amplified for NGS analysis regardless of subsampling, and PCR amplicons generated from unimaged *in situ* amplicons are discarded during the UMI matching step. Scale bars, 5 μm.

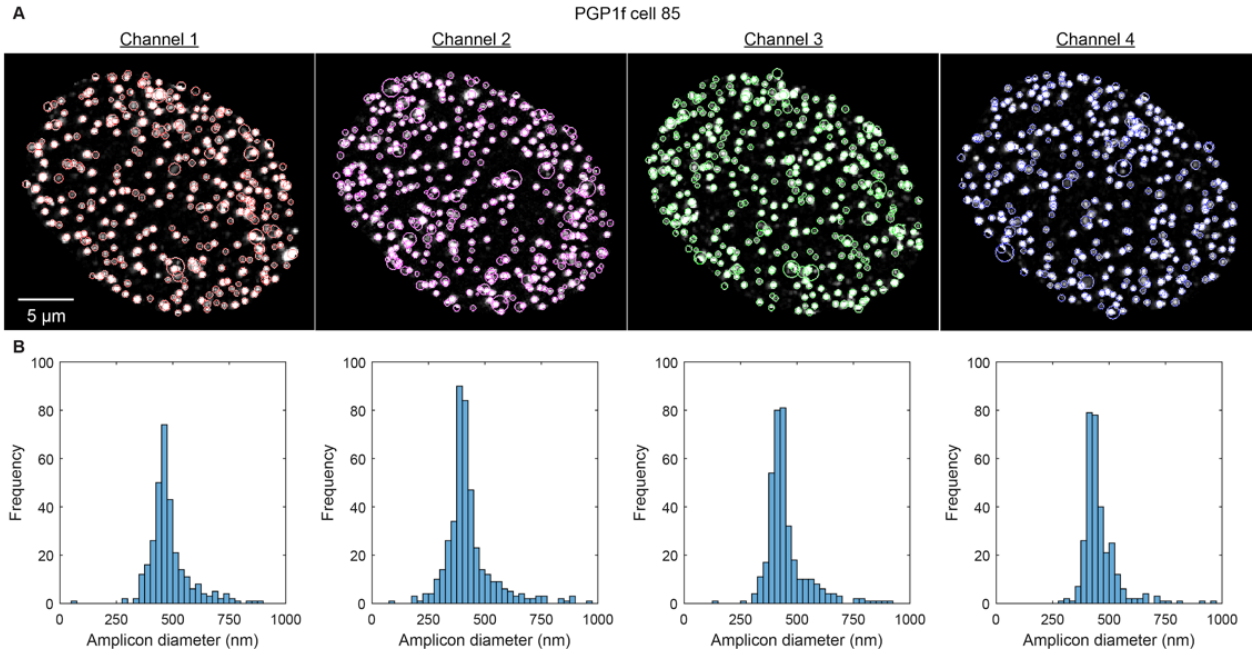

Figure 4-4. **Quantification of amplicon size**. (**A**) Representative nucleus (PGP1f cell 85) across all four sequencing channels with superimposed circles of radius 2σ (where σ is the standard deviation of the Gaussian fit for each amplicon). The images are maximum intensity projections across z-planes for the first cycle of *in situ* sequencing. (**B**) Histograms of the amplicon diameters (4σ) measured from each channel.

four color channels over all rounds of *in situ* sequencing (**Fig. 4-1B i, Fig. 4-6**. The *ex situ* sequenced reads are next associated with spatial coordinates within nuclei through error-robust matching of *in situ* and *ex situ* sequenced UMIs (**Fig. 4-1C**). To do this, we implement a probabilistic matching approach using principles from single-bit error correction [134] (**Fig. 4-6**). Collectively, the integration of these methods, which include developments across library construction, sequencing, and computational analyses, enable IGS as a general strategy for spatially mapping paired-end reads [130].

Here, we apply IGS to 106 human fibroblasts (PGP1f) and 113 cells across 57 intact early mouse embryos at the PN4 zygote (n = 24), late 2-cell (n = 20), and early 4-cell (n = 13) stages of development (**Fig. 4-1D**). Across both experiments, 66.35% of clearly resolvable amplicons (87.6% in PGP1f, 61.0% in mouse embryos) were confidently matched to an *ex situ* genomic read (**Fig. 4-7**). After cell filtering based on yield, karyotype, developmental stage, and cell cycle [130], this yielded a total of 286,335 spatially-localized genomic reads (36,602 in PGP1f, **Table 4.1**; 249,733 in mouse embryos, **Table 4.2**) with a UMI-matching false discovery rate of 0.26% (1.70% in PGP1f, 0.05% in mouse embryo, [130]). Mapped

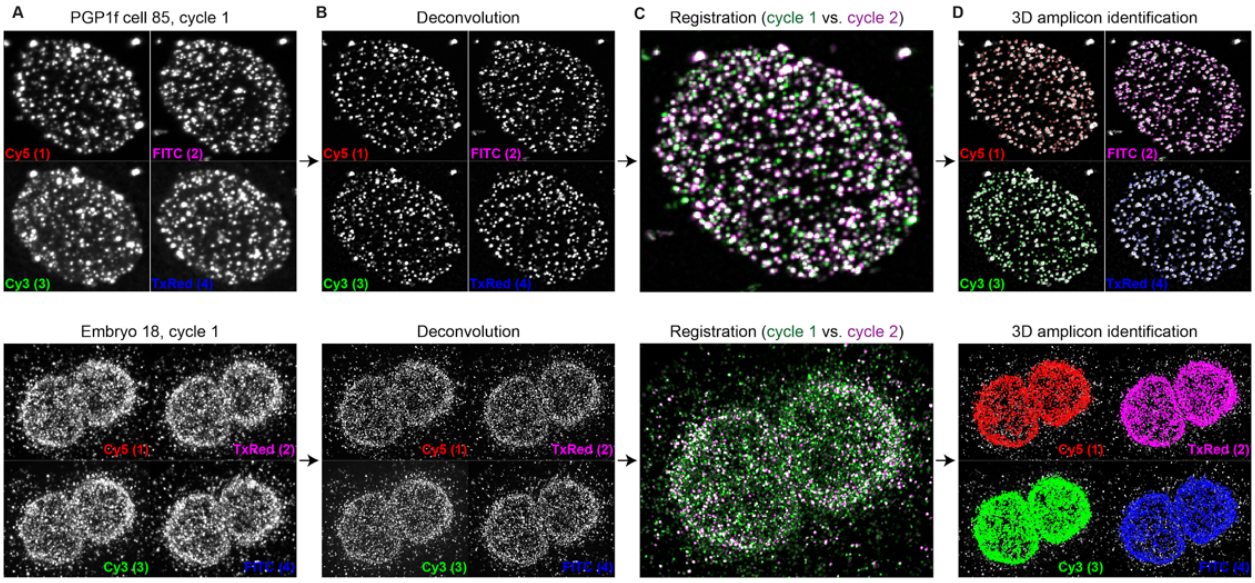

Figure 4-5. **Nuclei image processing**. (**A**) Raw *in situ* sequencing images (cycle 1) split by imaging channel for PGP1f cell 85 (top) and Embryo 18 (bottom). Preprocessing steps are performed in 3D; all displayed images are maximum intensity projections across channels and z-planes (all z planes for PGP1f, z = 50-70 = 5 μm slice for the embryo) for visualization purposes. (**B**) Deconvolution of *in situ* sequencing images (cycle 1) split by imaging channel for PGP1f cell 85 (top) and Embryo 18 (bottom). (**C**) Registration of *in situ* sequencing images (cycle 1 = green, cycle 2 = magenta) for PGP1f cell 85 (top) and Embryo 18 (bottom). (**D**) 3D amplicon identification from *in situ* sequencing images (cycle 1) split by imaging channel for PGP1f cell 85 (top) and Embryo 18 (bottom). Colored circles represent identified amplicons by imaging channel.

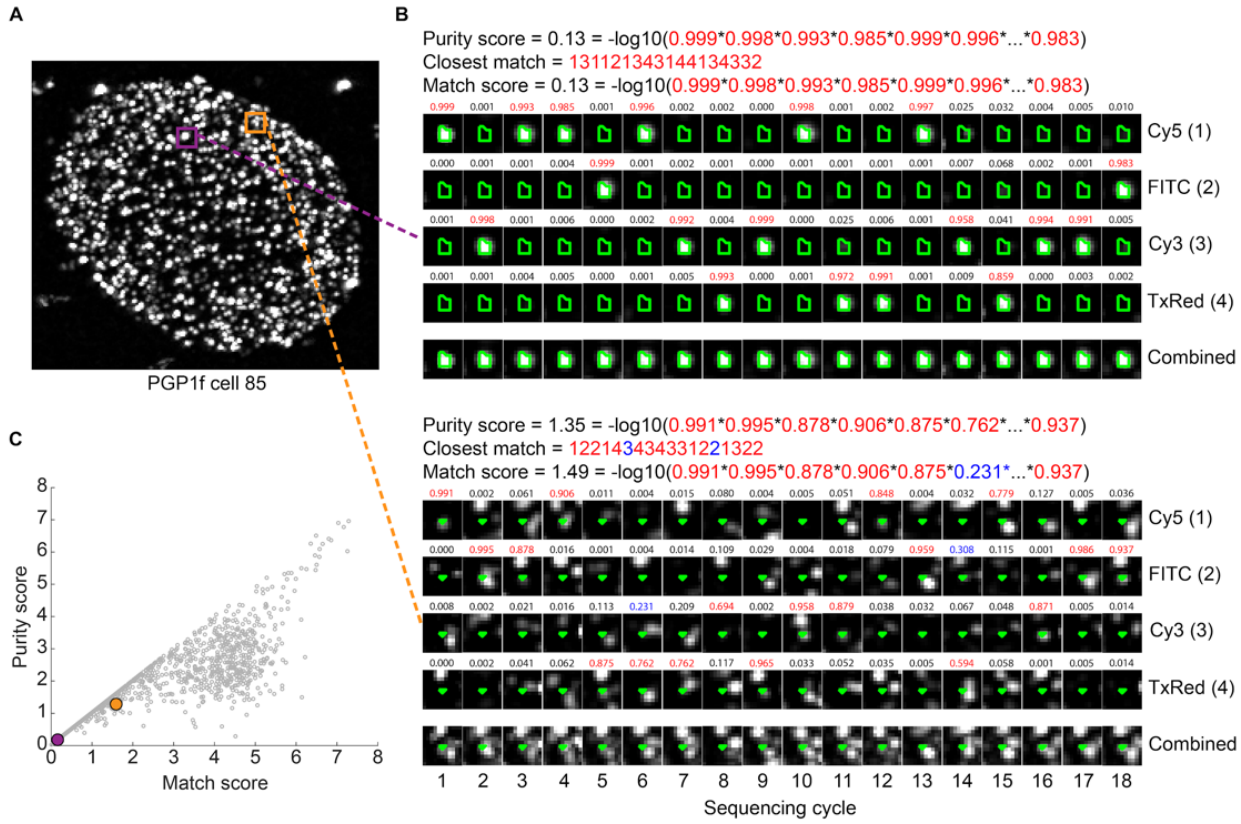

Figure 4-6. **Amplicon quantification and UMI matching**. (**A**) Representative nucleus (PGP1f cell 85) with two spatially-resolved amplicons highlighted. Image is a maximum intensity projection across channels and z-planes for the first cycle of *in situ* sequencing. (**B**) Quantification and UMI matching of the two amplicons from (A). The top set of images corresponds to the amplicon boxed in purple (in a sparsely-packed region), while the bottom set of images corresponds to the ampicon boxed in orange (in a densely-packed region). The bottom row of each set of images is a maximum intensity projection of the four channels, useful for visualizing amplicon density in the region. The green outline in each image indicates the region being quantified, while the number above each image represents the percentage of cycle fluorescence found in the corresponding channel (the sum of all channels in each cycle = 1). The purity score for each amplicon is calculated by multiplying the highest percentages for each cycle (indicated in red) and taking the negative log transformation of the product, i.e. -log10(product(max(matrix,1),2)). Each amplicon also has an associated closest UMI match. While the maximum channels of the top amplicon perfectly match its closest UMI, the bottom amplicon has two positions where the closest UMI match doesn't correspond to the maximum channels (indicated in blue). Taking the percentages from this path to calculate a match score results in a value that is greater (i.e. worse) than the lower bound of the purity score. (**C**) Comparison of purity score and match score for all amplicons in the representative nucleus shown from (A). The points corresponding to the two amplicons highlighted in (B) are enlarged and color-coded.

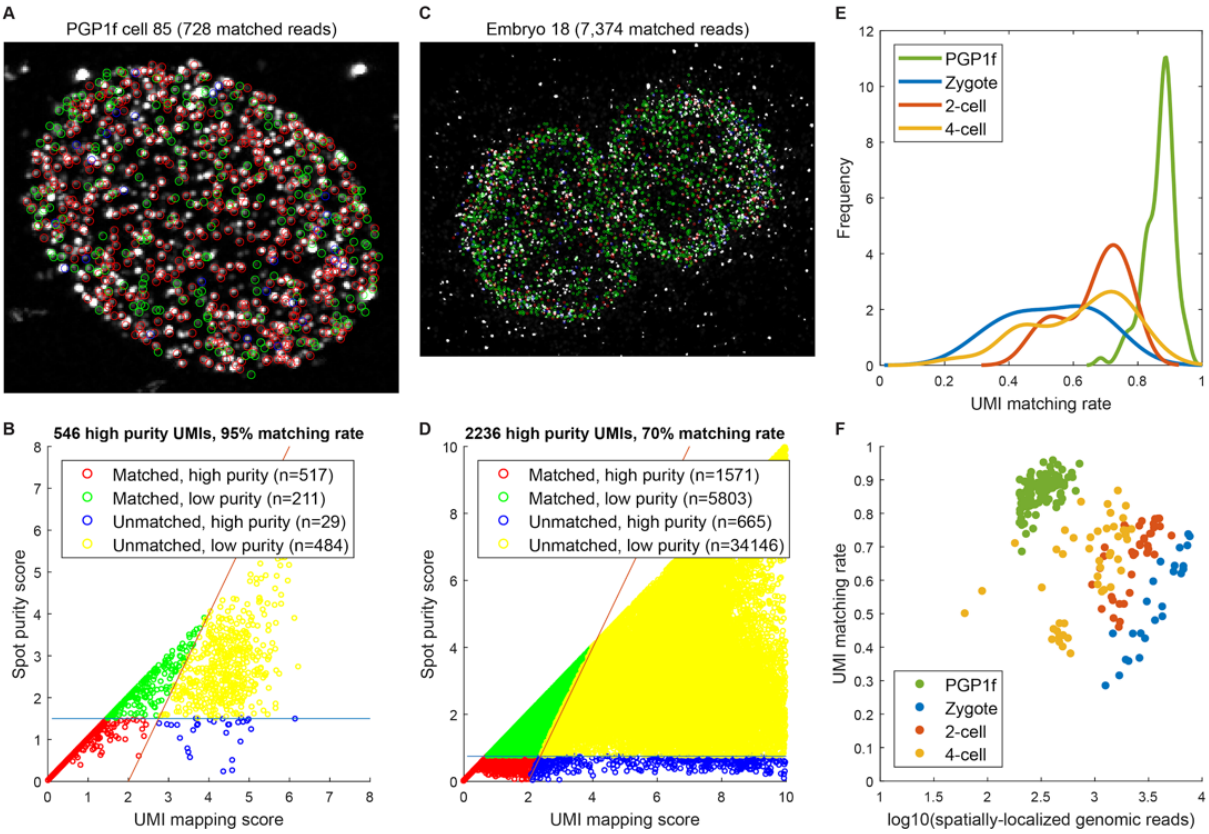

Figure 4-7. **UMI matching rate**. (**A**) Overlay of UMI matching status and a maximum intensity projection across all channels and z-planes for the first cycle of *in situ* sequencing for PGP1f cell 85. Amplicons are colored by UMI matching status, see legend in (B). Labels for unmatched, low purity amplicons are omitted for clarity. (**B**) All amplicons from PGP1f cell 85 plotted by match score and purity score. The horizontal line represents the threshold for high purity amplicons (1.5); the diagonal line represents the continuous threshold used to select valid UMI matches. (**C**) The same as (A), but for a 5 μm slice of Embryo 18 (z = 50 to 70). (**D**) The same as (D), but for Embryo 18. A higher threshold for high purity amplicons (0.75) was chosen based on the higher quality of the embryo *in situ* sequencing images. (**E**) The distribution of UMI matching rate by cell type. (**F**) Relationship between number of spatially-localized genomic reads and UMI matching rate, with points colored by cell type.

amplicons scaled with nuclear volume, spanning a median of 328 ± 114 reads per nucleus (± SD) in the PGP1f cells, to a median of 3,909 ± 2,116, 2,357 ± 1,063, and 1,074 ± 622 reads per nucleus in zygote, 2-cell, and 4-cell stage embryos, respectively (**Fig. 4-8**). Sequencing coverage across the hg38 and mm10 reference genomes was comparable to whole genome sequencing (**Fig. 4-9)**, and genomic reads did not show bias based on radial position (**Fig. 4-10**) or chromatin accessibility (**Fig. 4-11**). We also quantified the rate of detection for each genomic region across individual cells, as well as the distribution of genomic distance between sampled loci on the same chromosome (**Fig. 4-12)**.

For downstream analyses, we annotated each read based on spatial features such as

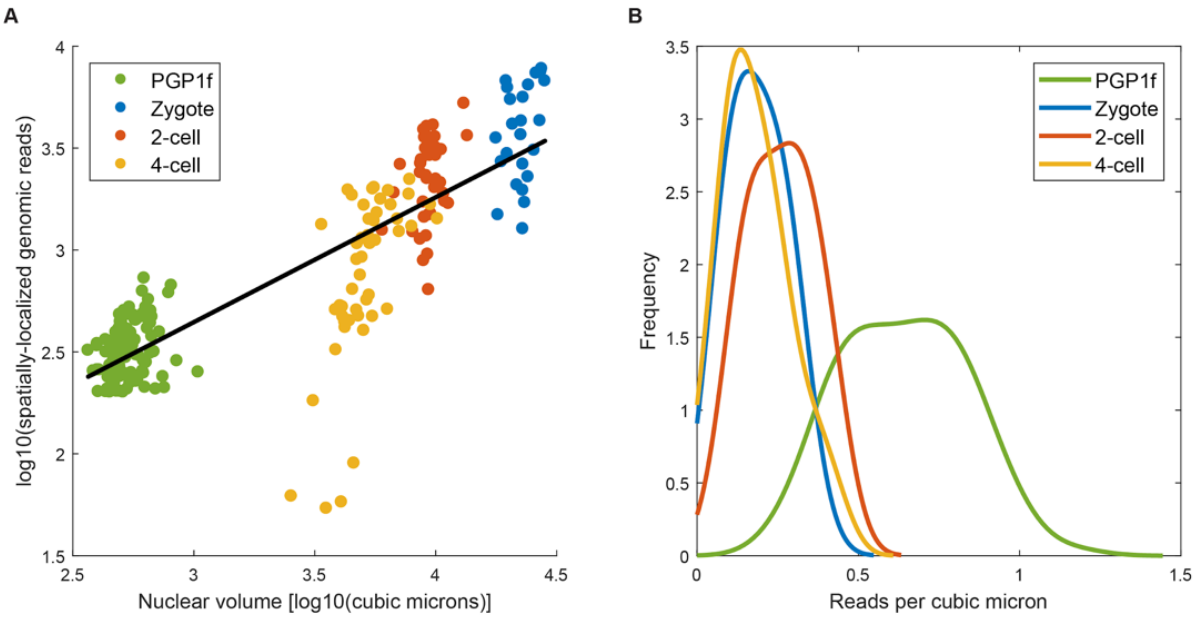

Figure 4-8. **Relationship between matched reads and nuclear volume**. (**A**) Nuclei from PGP1f and all embryonic stages plotted by number of matched reads and 3D nuclear volume, as estimated from DAPI image stacks. The line represents a linear fit on the log-transformed data. (**B**) Distribution of reads per cubic micron by cell-type and embryonic stage. The higher mean in PGP1f is likely attributable to the higher UMI matching rate (see Fig. 4-7).

inclusion in chromosome territories and distance to nuclear landmarks (nuclear lamina, centromeres, and nucleolar precursor bodies), as well as published genomic data including A/B compartments, lamina-associated domains, and GC content (**Table 4.3**). The full embryo dataset can be interactively visualized at *https://buenrostrolab.shinyapps.io/insituseq/*.

4.3 Validation of *in situ* genome sequencing in human cells

To validate that our method detects features of spatial genome organization, we first examined the locations of chromosomes in interphase human male PGP1 fibroblasts (**Fig. 4-13A, 4-13B**). We found that autosomal reads displayed a strong tendency to spatially colocalize into two distinct spatial regions, while allosomal reads were restricted to one region, confirming the known organization of chromosomes into territories [119] (**Fig. 4-14**). To systematically define these territories, we used a maximum likelihood estimation approach to assign reads to homologous chromosome clusters using both the spatial and genomic positions

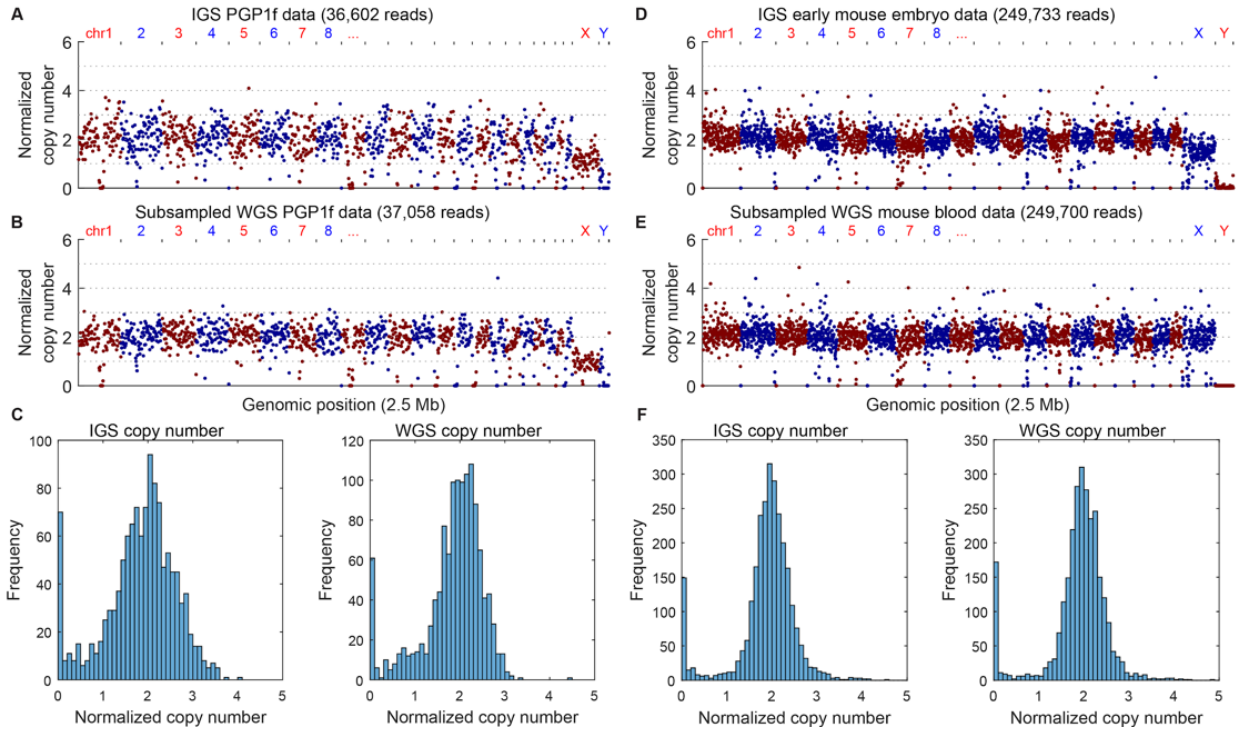

Figure 4-9. **Comparison of coverage to whole-genome sequencing**. (**A**) Normalized coverage of PGP1f IGS data across 2.5 Mb bins in hg38 (n=36,602 reads). (**B**) Normalized coverage of subsampled PGP1f whole-genome sequencing data (ENCODE accession ENCFF713HUF (*https://www.encodeproject.org/files/ENCFF713HUF/*) across 2.5 Mb bins in hg38 (n=37,058 reads). (**C**) Histograms showing the distribution of normalized coverage per autosomal bin in IGS (left) and WGS (right). (**D**) Normalized coverage of mouse early embryo IGS data across 1 Mb bins in mm10 (n=248,733 reads). (**E**) Normalized coverage of subsampled mouse blood whole-genome sequencing data (unpublished) across 1 Mb bins in mm10 (n=249,700 reads). (**F**) Histograms showing the distribution of normalized coverage per autosomal bin in IGS (left) and WGS (right).

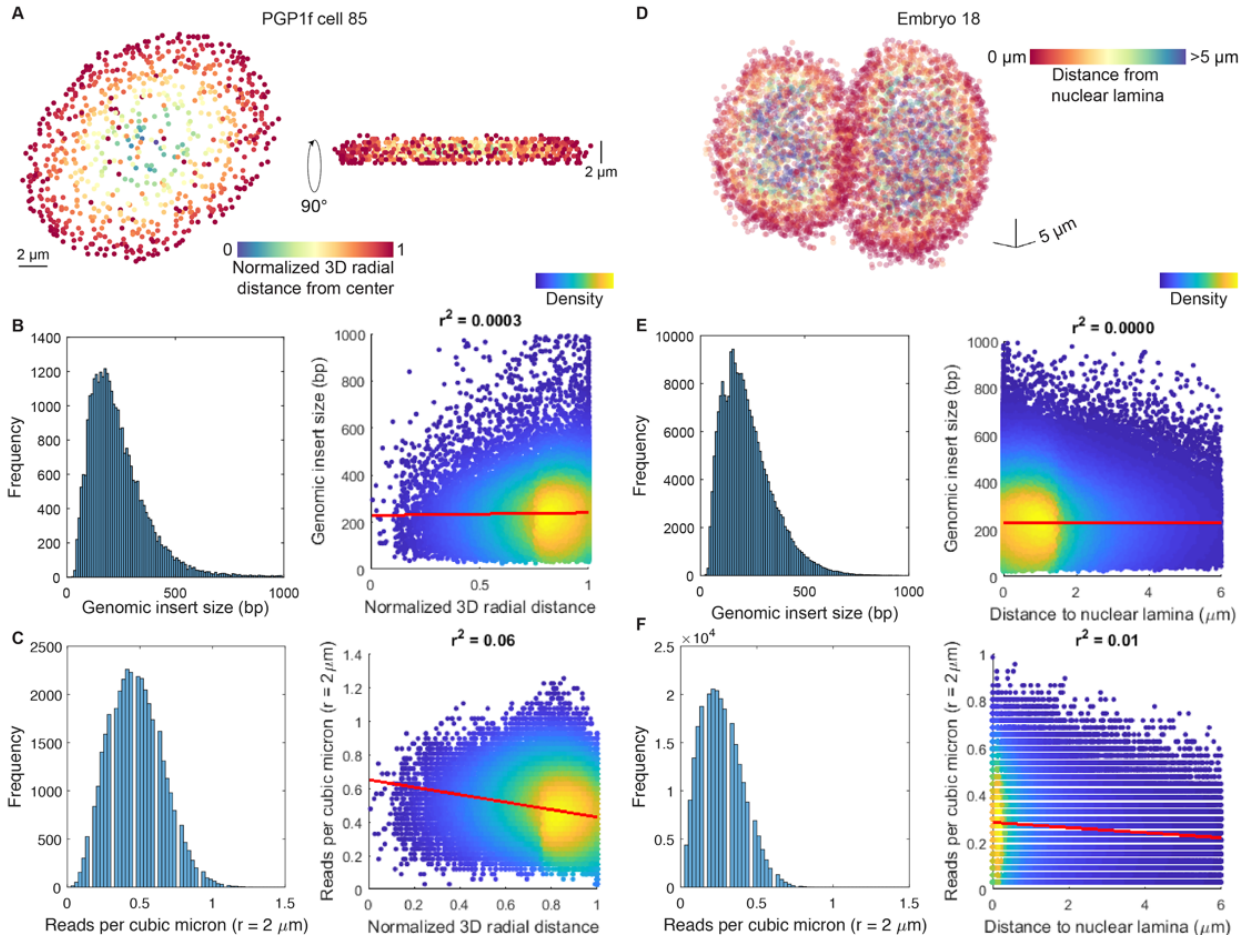

Figure 4-10. (**A**) PGP1 fibroblast (PGP1f cell 85) shown from two angles, 90 degrees apart, colored by normalized 3D radial distance. (**B**) Left: Histogram of genomic insert sizes for all PGP1f reads. Right: All reads plotted by normalized 3D radial distance and genomic insert size, colored by point density. Best fit line shown in red. (**C**) Left: Histogram of reads per cubic micron in a 2 micron radius for all PGP1f reads. Right: All reads plotted by normalized 3D radial distance and reads per cubic micron, colored by point density. Best fit line shown in red. (**D**) Zygote (Embryo 18), colored by distance to nuclear lamina. (**E**) Same as (B), but for all reads in embryos. (**F**) Same as (C), but for all reads in embryos.

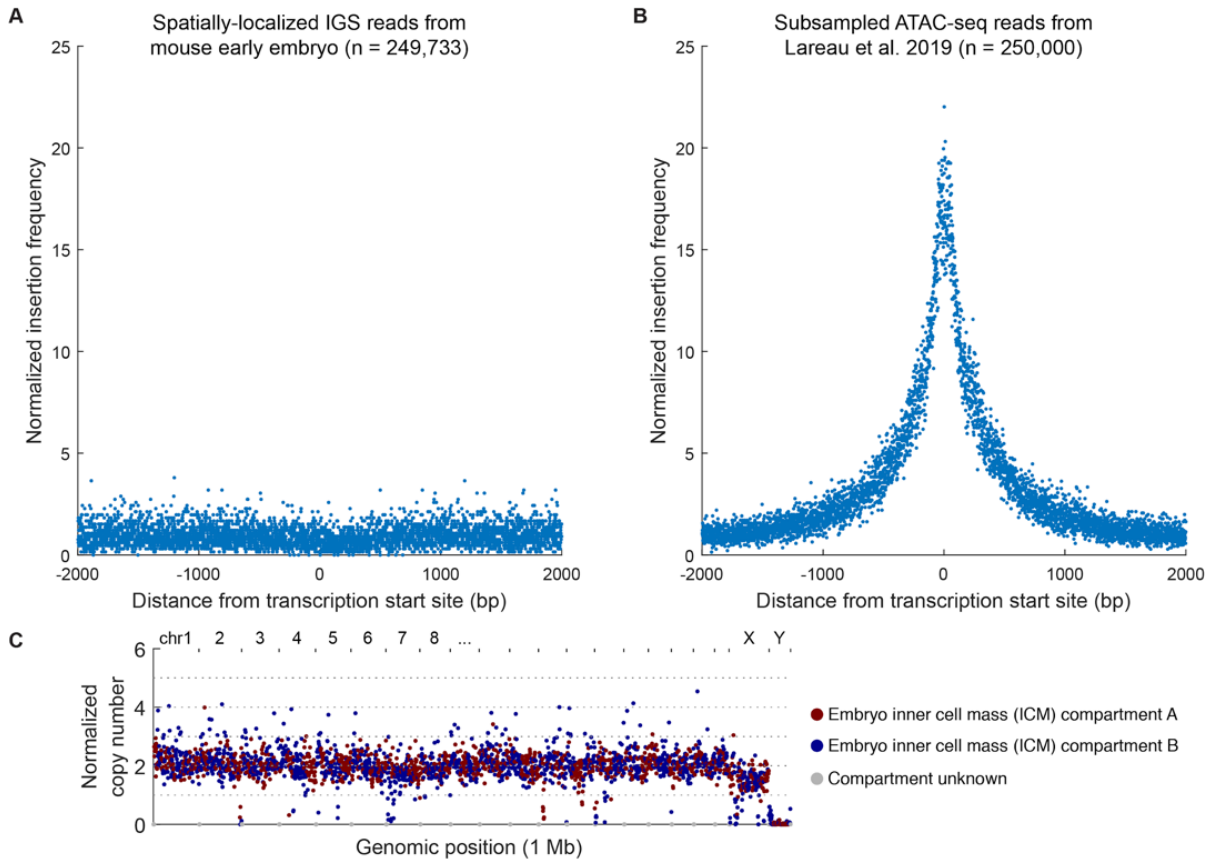

Figure 4-11. **Lack of enrichment for accessible chromatin**. (**A**) Enrichment of genomic reads at genomic positions surrounding transcription start sites in early mouse embryo IGS data (n=249,733 reads). (**B**) Enrichment of genomic reads at genomic positions surrounding transcription start sites in subsampled mouse brain single-cell ATAC-seq data from [135] (n=250,000 reads). The difference between these plots indicates that IGS does not have a bias towards accessible chromatin despite similarities in the transposase-based library preparation to ATAC-seq. (C) Normalized coverage of mouse early embryo IGS data across 1 Mb bins in mm10 (n=248,733 reads), colored by embryo inner cell mass (ICM) A/B compartment calls from Hi-C data [136].

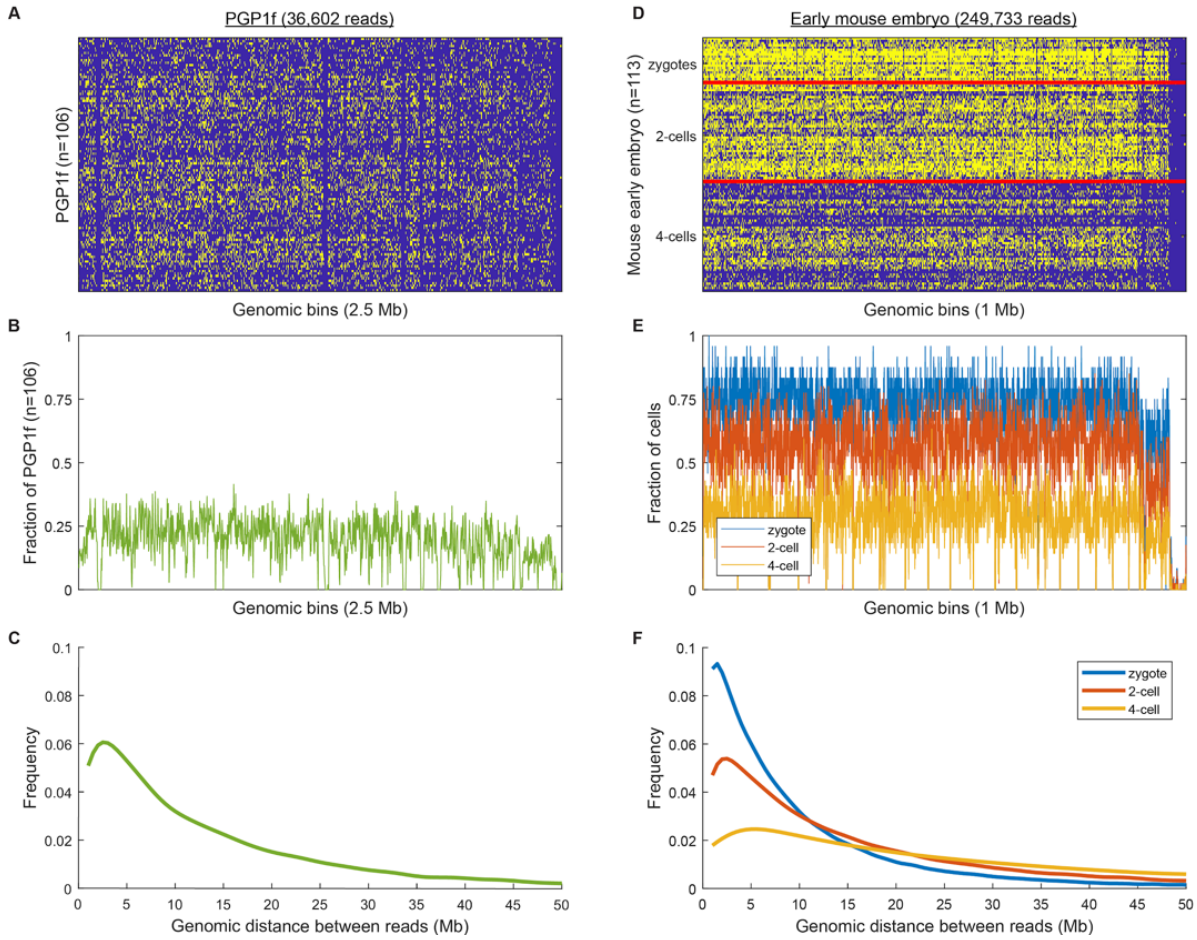

Figure 4-12. **Sampling of genomic regions in individual cells**. (**A**) A n by m matrix showing which 2.5 Mb genomic bins across hg38 (m=1,249) were sampled in individual PGP1f nuclei (n=106). Yellow indicates that the genomic region in question was sampled at least once in a given nucleus. (**B**) A line plot showing the fraction of PGP1f nuclei containing at least one read from each 2.5 Mb genomic bin. (**C**) The distribution of genomic distances between sampled loci from the same chromosome territory in PGP1f. (**D**) A n by m matrix showing which 1 Mb genomic bins across mm10 (m=2,737) were sampled in individual early mouse embryo nuclei, ordered by developmental stage (n=113). (**E**) A line plot showing the fraction of early mouse embryo nuclei containing at least one read from each 1 Mb genomic bin by developmental stage. (**F**) The distribution of genomic distances between sampled loci from the same chromosome territory in early mouse embryo nuclei by developmental stage.

of each read (**Fig. 4-15**). For chromosomes with two spatially-resolved homolog clusters, we found, via density-based thresholding, that 6.83% of reads did not spatially colocalize with either cluster (**Fig. 4-15**), a larger fraction than our estimated 1.70% UMI-matching false discovery rate in PGP1f, which may be associated with long-range chromosome looping [137]. Following spatial clustering, we visualized genome-wide conformations of individual chromosomes in single diploid cells by connecting the reads in each cluster according to genomic position (**Fig. 4-13C**).

We proceeded to characterize the positions of diploid chromosome territories across single cells by calculating the average spatial distance between pairwise genomic locations across the genome, at 10 Mb resolution. We found that blocks of short intrachromosomal distances were strongly delineated along the diagonal of this pairwise distance matrix (**Fig. 4-13D**), consistent with genomic organization into chromosome territories described above. The matrix also shows enrichment of shorter pairwise distances between smaller chromosomes. Additionally, we observed a positive association between chromosome size and radial distance from the nuclear center (**Fig. 4-13E**). These observations indicate that small chromosomes tend to be in closer proximity near the nuclear center, consistent with prior studies in human fibroblasts [140]. These results illustrate the ability of IGS to resolve diploid chromosome territories within the nuclei of single cells and to investigate the spatial positioning of chromosomes at scale.

Repetitive DNA elements, such as transposons and endogenous retroviruses, make up approximately 50% of the human genome [141, 142], and their localization is known to play a role in normal [143] and disease-associated [144] genome organization. Although FISH-based methods can measure the localization of targeted classes of repetitive sequences [140, 145], current approaches have not simultaneously mapped the localization of many classes of repetitive sequences across the nucleus. We applied IGS to simultaneously measure the localization of repetitive sequences across the genome. We focused on the 13.9% of spatially-resolved reads that do not uniquely align to the reference genome (hg38), and aggregated them into ~250 classes of repetitive elements using Repbase [146] (**Fig. 4-16A**). We found that the number of reads associated with each element was proportional to its observed frequency in hg38 (**Fig. 4-16B**), enabling an unbiased approach to studying localization of

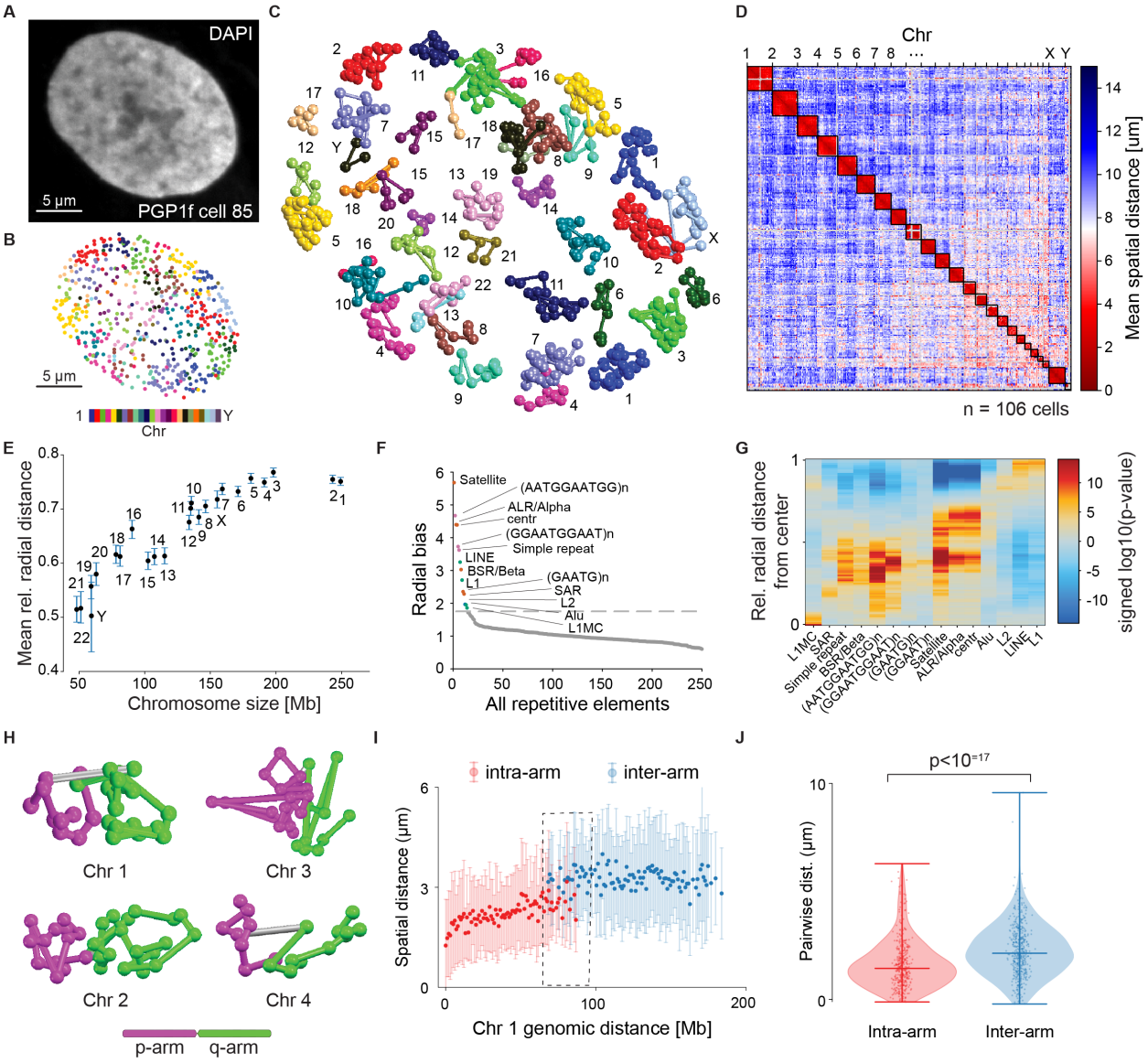

Figure 4-13. **IGS characterizes spatial features of the human genome**. (**A**) DAPI stain of a PGP1f nucleus after *in situ* library construction. (**B**) 601 spatially-localized reads in the same PGP1f nucleus, colored by chromosome. (**C**) Exploded view reveals conformations of chromosome territories, shown as *in situ* reads (balls) connected according to sequential genomic position (sticks). (**D**) Genome-wide population mean pairwise distance matrix of 106 PGP1f cells binned at 10 Mb. (**E**) Chromosome size vs. normalized mean radial distance from the nuclear center for 106 diploid-resolved PGP1f cells. Error bars denote 95% confidence interval of the mean determined by bootstrapping. (**F**) The 103 most abundant repetitive elements ordered by radial bias, defined as the variability of binned distances relative to a permuted background from the nuclear center for 106 PGP1f cells. The dashed grey line represents the threshold for elements shown in (G). (**G**) Radial enrichment/depletion by binned distance from the nuclear center for the repetitive elements with the strongest radial bias from (F). (**H**) Ball-and-stick models for Chr 1-4 in the same single-cell, demonstrating spatial polarization between the p and q arms of each chromosome. (**I**) Genomic distance vs. spatial distance for Chr 1, distinguishing intra-arm and inter-arm measurements. Error bars: standard deviation. Dashed: range in which both measurements can be compared at reasonable sampling depth (n $>$ 20 per 1 Mb bin). (**J**) Intra-arm and inter-arm distance distributions in the dashed range in (I) are distributed differently (n = 819 intra-arm, 766 inter-arm, 144 Chr 1 territories, K-S test, $p < 10^{-16}$). Violin plot indicates median and range.

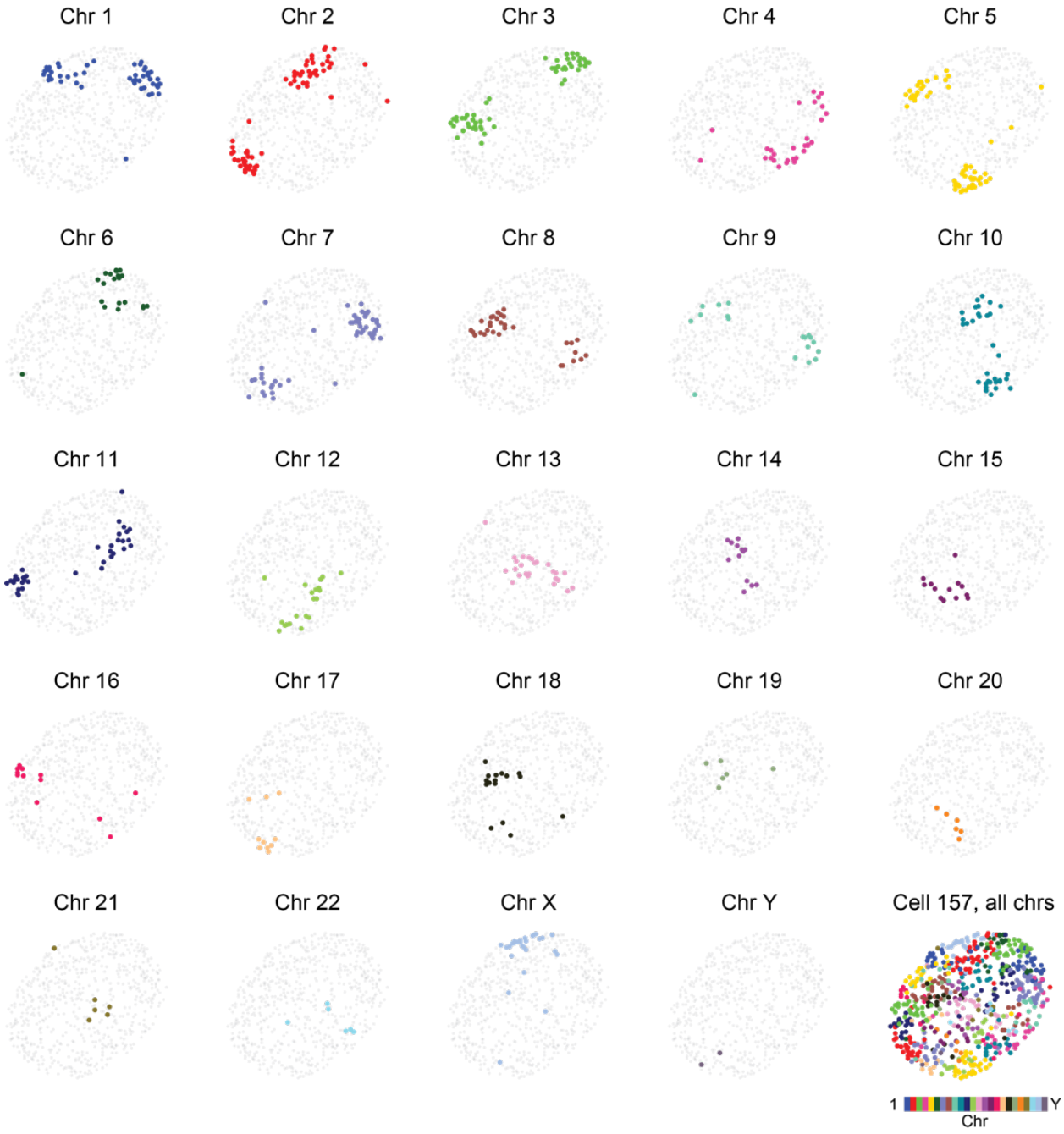

Figure 4-14. **"2n" colocalization pattern (PGP1f)**. "2n" co-localization pattern of spatially-localized reads colored by chromosome, before maximum likelihood clustering into territories. Here the measurement is broken down into individual chromosomes in a single cell for visual clarity. Most reads mapping to a given autosome typically cluster into two territories, while reads mapping to sex chromsomes typically cluster into one. ~7% of reads do not colocalize with one of these large territories, possibly due to a combination of UMI FDR and chromosome looping [138, 139].

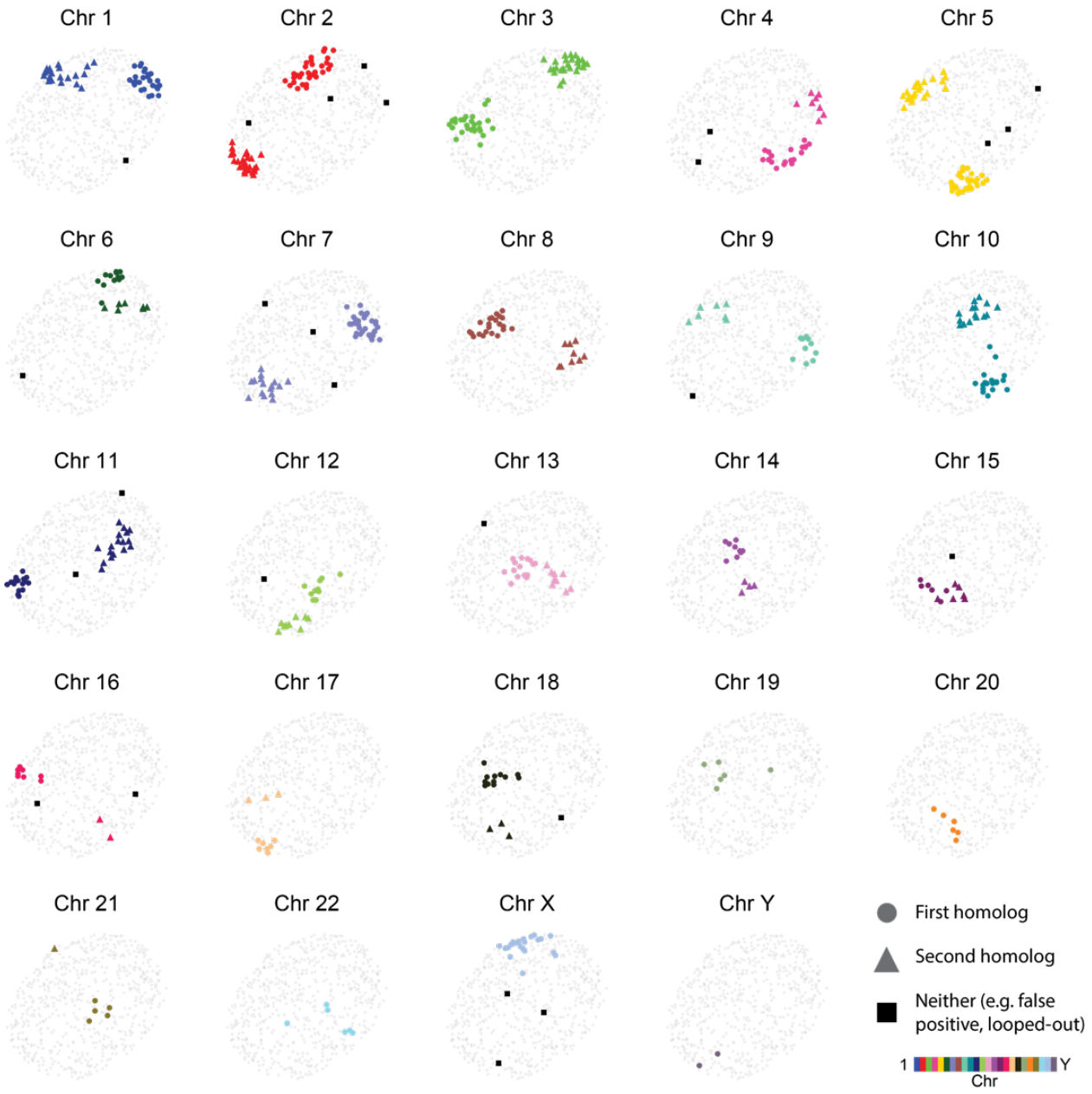

Figure 4-15. **Territory classification after maximum likelihood clustering (PGP1f).** Individual reads in a single cell after maximum likelihood-based clustering, into territories. The discrete spatial nature of individual chromosomes in interphase allows most reads to be assigned to a specific chromosome territory (denoted as a circle or triangle for each homolog), resolving the diploid nature of the genome for these reads. The genomic positions of these reads were used to resolve ambiguities when chromosomes are close to each other (preventing purely spatial classification); nonetheless, small chromosomes are more challenging to resolve than larger ones when they are in close proximity. Reads which do not spatially co-localize with a territory can also be identified by this clustering method, permitting them to be accounted for in downstream analyses.

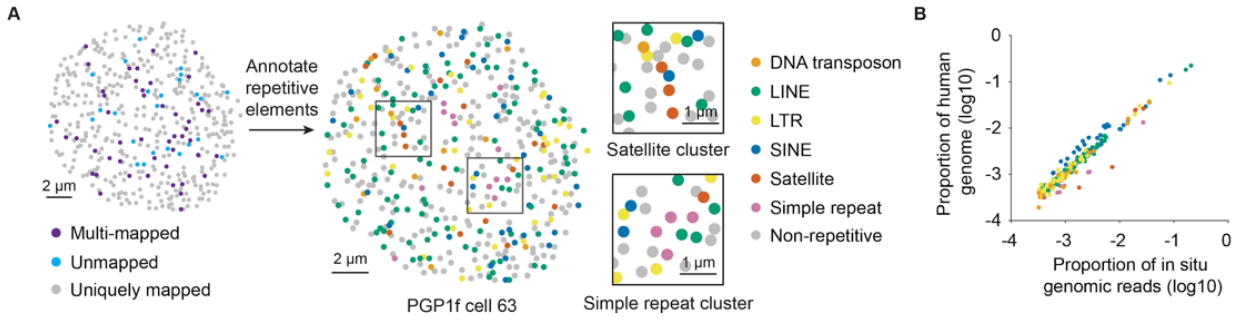

Figure 4-16. **Repetitive element frequency (PGP1f)**. (**A**) Spatially-localized reads in a representative nucleus (PGP1f cell 63) colored by alignment status (left) and repetitive element class (right). The boxes highlight two regions that appear to contain spatial clusters of Satellite and Simple Repeat elements respectively. (**B**) The observed number of reads associated with each repeat annotation is proportional to their expected frequency in hg38.

repetitive elements.

Given our observations of radial patterns of chromosome positioning and the known radial organization of heterochromatin [147], we investigated whether repetitive elements displayed radial patterns. To do this, we compared the radial distribution of reads containing repetitive elements to permuted distributions to identify classes of repetitive elements with the strongest radial bias (**Fig. 4-13F,** [130]). We confirmed reports that Alu elements are depleted ~1 µm from the nuclear edge [140], further validating our approach (**Fig. 4-13G**). We also found that certain types of repetitive heterochromatin, such as satellite DNA, show enrichment towards the nuclear center, while others, including AT-rich L1 elements, are overrepresented at the nuclear edge. These findings demonstrate the ability of IGS to simultaneously identify spatial localization patterns of many different repetitive sequences *de novo* in an untargeted and genome-wide manner.

Having shown that IGS confirms known features of global genome organization, we next asked whether we could characterize the structures of individual chromosomes. DNA FISH and Hi-C studies have found that chromosome arms can be individually compartmentalized in a fashion similar to chromosome territories [113, 148]. Independent localization of chromosome arms was apparent as stripes in the genome-wide distance matrix (**Fig. 4-13D**), and could be visualized in single chromosomes colored by their p and q arms (**Fig. 4-13H**). We computed pairwise distances for reads within each chromosome territory (Chrs 1-11 and Chr X, **Fig. 4-17)** and fit a power law to this relationship as described [121]. Separate treatment

of p and q arms resulted in an improved fit compared to fitting of all of Chr 1 (**Fig. 4-18**), in line with the expectation of differential scaling across the centromere. Indeed, intra-arm and inter-arm pairs of loci exhibited two scaling regimes when treated separately (**Fig. 4-13I, Fig. 4-19**). In chromosomes where we had high coverage (Chr 1-11 and Chr X), inter-arm distances were significantly larger than intra-arm distance for the range of genomic distances present in both distributions (**Fig. 4-13J**, 56-87 Mb for Chr 1; boxed region, **Fig. 4-13I**; K-S test, $p < 10^{-16}$; **Fig. 4-20**). These results extend observations of spatially polarized chromosome arms [149], and demonstrate the ability of IGS to characterize subchromosomal spatial structure. Taken together, these findings highlight the unique ability of IGS to simultaneously interrogate broad features of genome organization, including chromosome positioning, chromosome folding, and the localization patterns of repetitive sequences.

4.4 In situ genome sequencing in intact early mouse embryos

The spatial organization of the genome is extensively remodeled in early embryogenesis, as the initially separate parental genomes undergo major reorganization after fertilization to prime the organism for zygotic genome activation (ZGA) [150] and, subsequently, lineage-specific cell fates [151]. Studies have linked chromatin and epigenetic remodeling to various phenomena including sequence-specific localization of chromatin to nuclear landmarks [152, 153], parent specific-chromatin domain organization in single cells [154, 155, 156, 157], and developmental specification of clonal lineages [158, 159, 160]. Given the importance of spatial features, sequence-specificity, and intercellular relationships in these phenomena, we sought to apply IGS in intact early embryos to characterize genome organization in early embryogenesis across length scales.

We applied IGS to intact early mouse embryos (B6C3F1 females × B6D2F1 males) spanning the PN4 zygote (3,909 ± 2,116 reads/cell, median ± SD), late 2-cell (2,357 ± 1,063 reads/cell), and early 4-cell (1,074 ± 622 reads/cell) stages of development (**Fig. 4-21A**).

Collectively, imaging and sequencing methods have shown that some of the structural

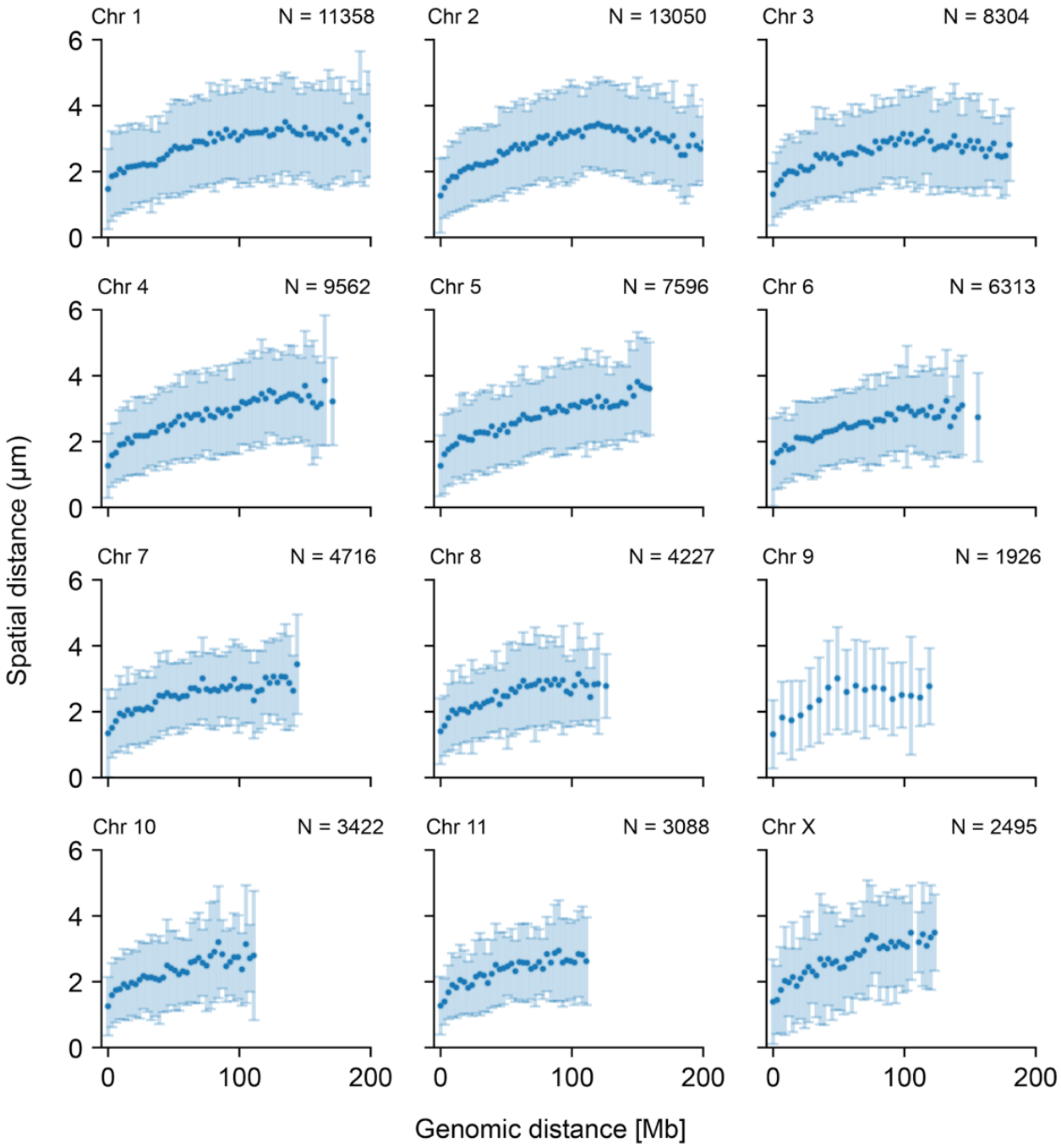

Figure 4-17. **Relationship between genomic vs spatial distance for Chr 1-11 + X (PGP1f)**. Distances were computed pairwise for each territory, and binned at 3 Mb (rather than 1 Mb as in Fig. 4-13I) to ensure sufficient coverage in smaller chromosomes. Chr 9 is an exception and was instead binned at 7 Mb because of a dearth of inter-arm measurements due to its large centromere and small p arm. N, number of pairwise distance measurements. The mean of each bin ± 1 SD is plotted. Bins with fewer than 20 measurements were excluded.

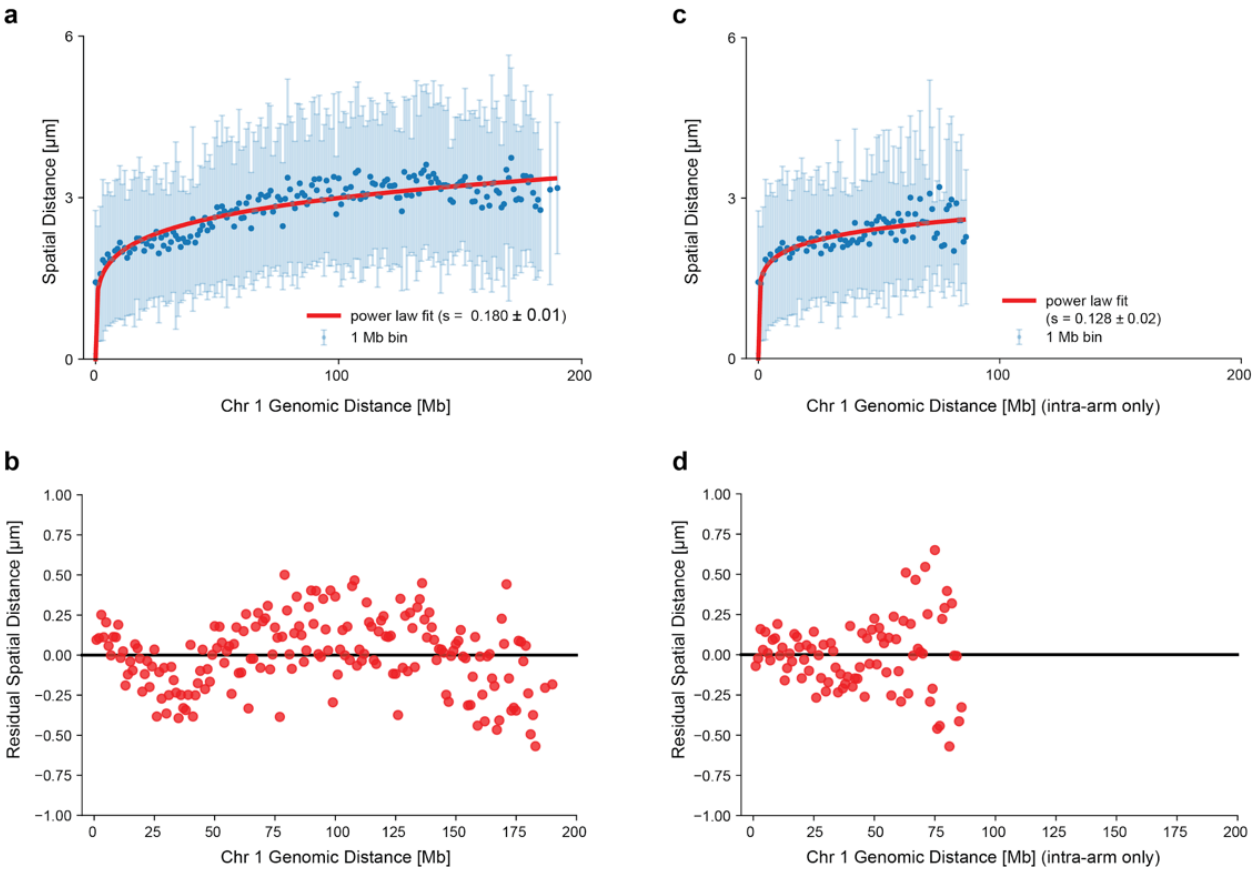

Figure 4-18. **Chr 1 genomic vs mean spatial distance, power law function fit (PGP1f)**. (**A**) Ensemble Chr 1 genomic vs mean spatial distance. Distances were computed pairwise for each Chr 1 territory (n = 144 territories, 11358 pairwise distances), binned at 1 Mb, and plotted (mean ± SD) alongside a power law function fit to these measurements (scaling exponent s = 0.180 ± 0.01, 95% CI). (**B**) Residuals from the power-law fit to all of Chr 1 in a). The residuals are nonuniform around 0 and are autocorrelated at lag 1 (Ljung-Box test, p < 10-5). (**C**) Pairwise distances restricted to intra-arm measurements (n = 5788 pairwise distances), yielding a smaller scaling exponent (s = 0.128 ± 0.02). (**D**) Residuals from the intra-arm power law fit in (C). The residuals are not significantly autocorrelated at lag 1 (Ljung-Box test, p = 0.46). For both (A) and (C), bins with fewer than 20 measurements were excluded.

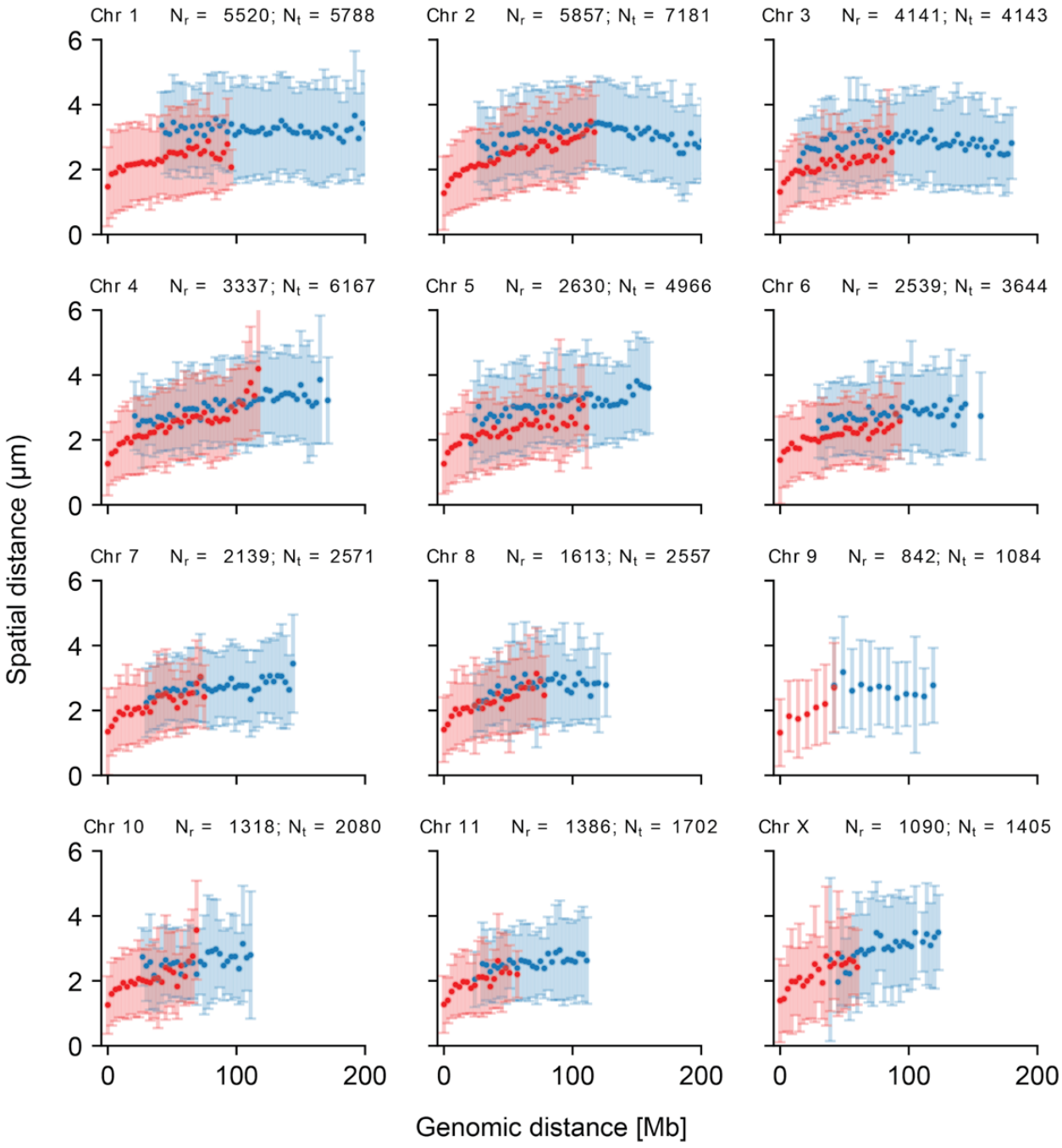

Figure 4-19. **Genomic vs spatial distance within and between chromosome arms, Chr 1-11 + X (PGP1f)**. Genomic distance versus spatial distance for Chr 1-11 & X, distinguishing between intra-arm and inter-arm measurements. Measurements were then computed, binned, and plotted as described in **Fig. 4-17**. N_r, intra-arm measurements; N_t, inter-arm measurements. The mean of each bin ± 1 SD is plotted.

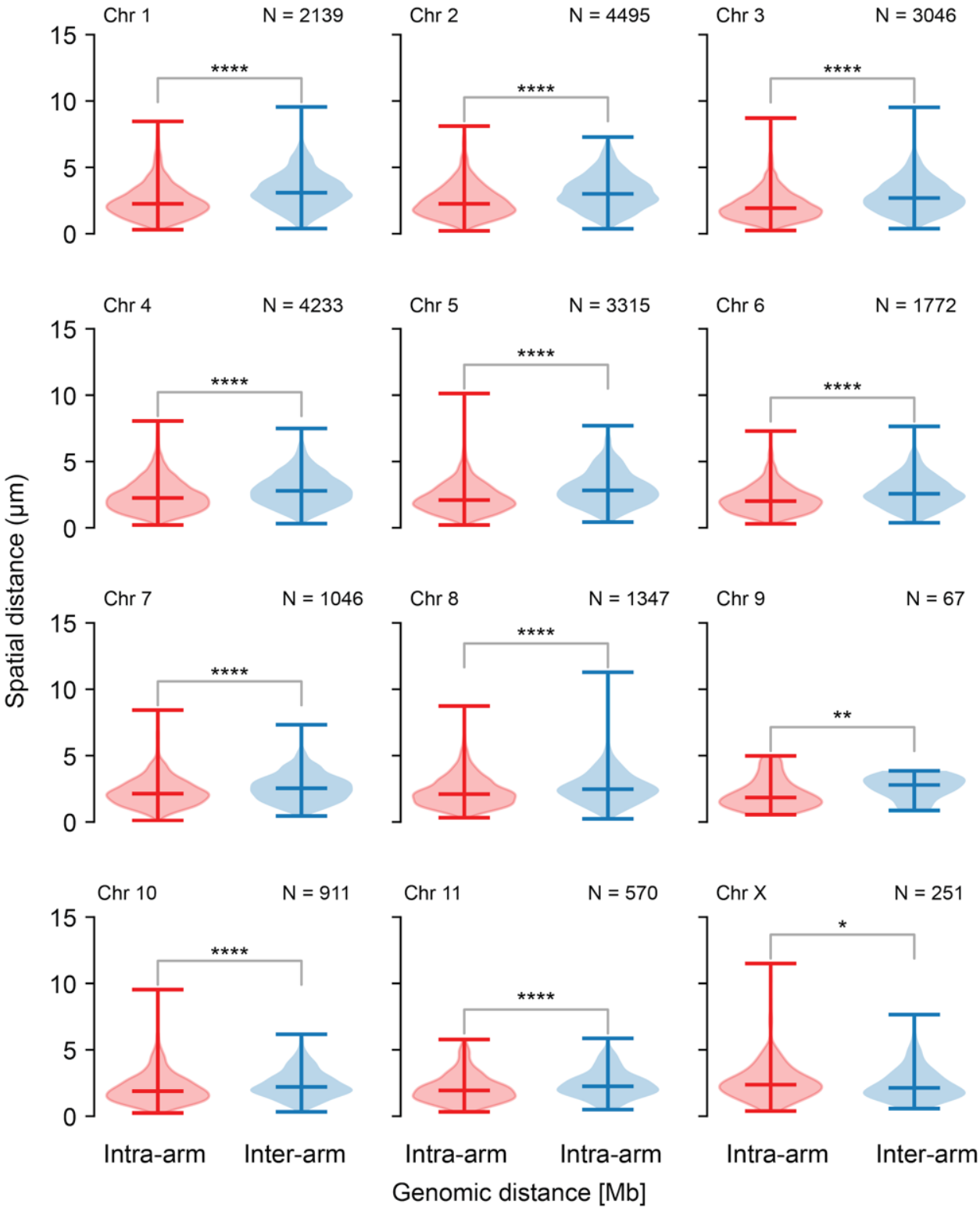

Figure 4-20. **Distribution of intra-arm and inter-arm pairwise distances, Chr 1-11 + X (PGP1f).** Distribution of intra-arm and inter-arm pairwise distances in genomic distance ranges shared by both types of measurement. Bins in **Fig. 4-19** with at least 20 instances of each type of measurement were considered. There is a significant difference between the two distributions for each chromosome, with higher mean inter-arm distances in general. Chr 9 is sparse due to the idiosyncrasies described above. N, number of pairwise distance measurements in the shared genomic distance range. ****, $p < 0.0001$, ***, $p < 0.001$, **, $p < 0.01$, *, $p < 0.05$; all significance comparisons by K-S test.

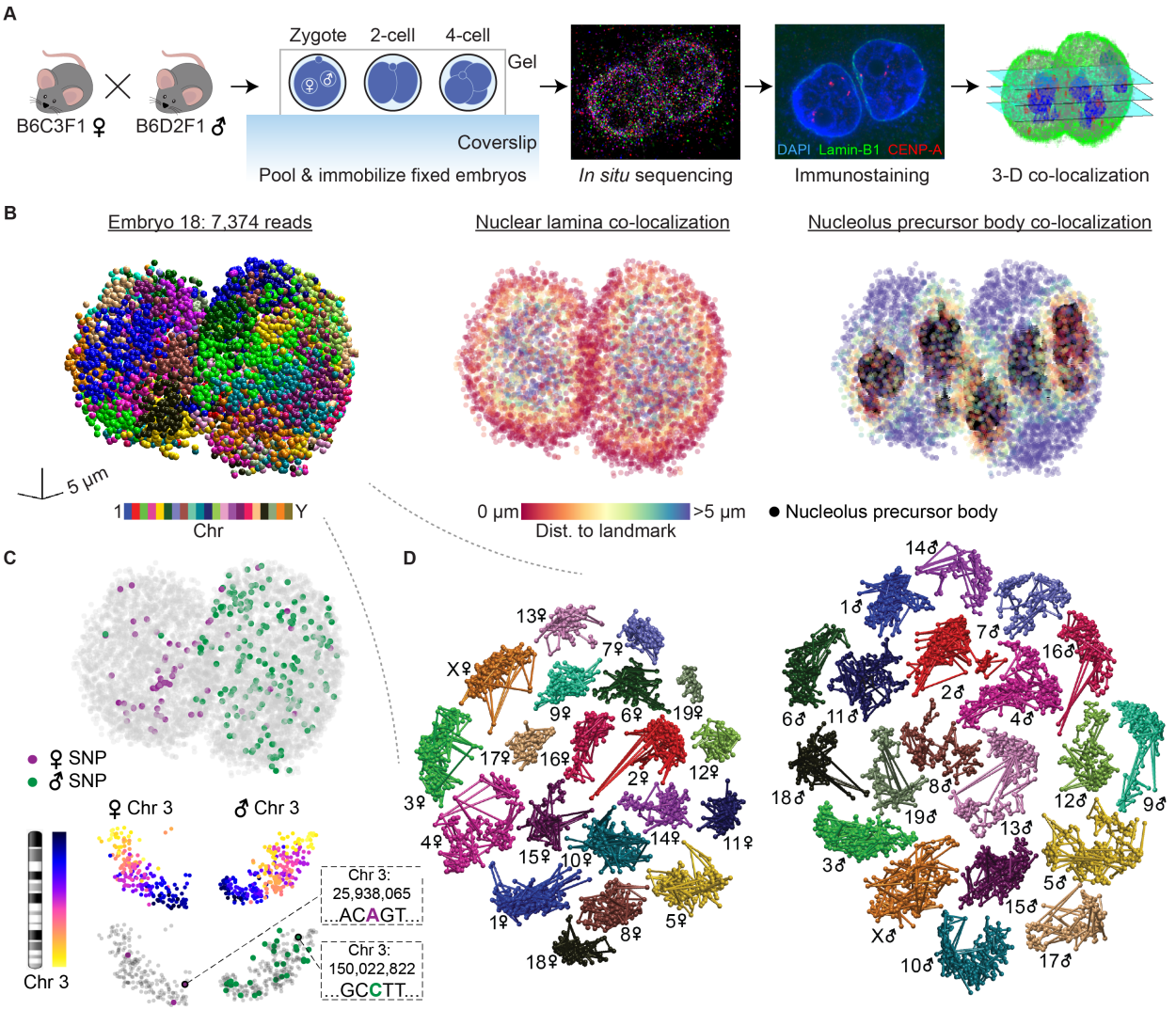

Figure 4-21. **IGS enables high-resolution genomic and spatial profiling of intact early mouse embryos**. (**A**) Workflow. B6C3F1 x B6D2F1 embryos at the zygote, 2-cell, and 4-cell stages are pooled, fixed and immobilized in a polyacrylamide gel. Following *in situ* sequencing, DAPI and immunofluorescence staining of CENP-A and Lamin-B1 are performed. (**B**) Representative zygote with 7,374 spatially-localized reads colored by chromosome (left), distance to the nuclear lamina (middle), and distance to nearest nucleolus precursor body (right). (**C**) Amplicons from (B), with reads colored by parental haplotype assignment for the intact embryo (top), reads colored by genomic position for Chr 3 homologs (middle), and reads colored by parental haplotype assignment for Chr 3 homologs (bottom). Boxes show two haplotype-informative Chr 3 SNPs. (**D**) An exploded view of chromosome territories from (B) for the maternal (left) and paternal pronuclei (right).

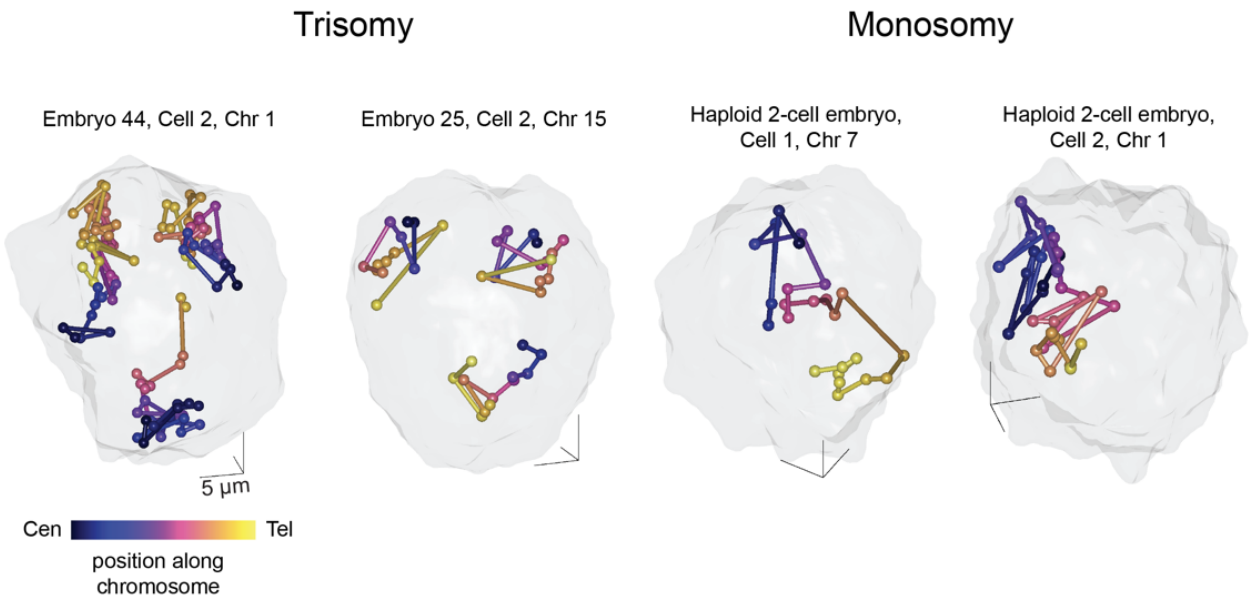

Figure 4-22. **Aneuploidy in embryonic cells**. Examples are shown of trisomy and monosomy in four different embryonic cells, detected by IGS.

changes in early development are associated with nuclear landmarks, such as the centromeres [161], nuclear lamina [153] and nucleolus precursor bodies (NPBs) [162]. To demonstrate that IGS is compatible with other imaging modalities and to investigate the organizational roles of these landmarks, we performed co-immunostaining for CENP-A (centromere) and Lamin-B1 (nuclear lamina), in addition to staining with DAPI (used to locate NPBs). The resulting images were segmented and registered to the *in situ* sequencing data in 3D, enabling us to directly localize genomic reads relative to these landmarks (**Fig. 4-21B**, [130]).

In order to resolve the maternal and paternal genomes within single cells, we first confirmed the presence of chromosome territories in all stages. As with the PGP1 fibroblasts, we found that reads originating from a particular autosome could generally be separated into two distinct spatial clusters per nucleus (or one cluster per allosome in male embryos). We then filtered cells based on yield, karyotype, developmental stage, and cell cycle (**Fig. 4-22** [130]. After filtering, we found a nearly equivalent rate of reads that did not colocalize with chromosome territories as in PGP1f (6.95%). Relative to reads within territories, these non-colocalizing reads were significantly depleted from regions proximal to the nuclear lamina and NPBs (**Fig. 4-23** K-S test, $p < 10^{-51}$ and 10^{-109} respectively).

To assign parent-of origin to each territory, we identified spatially-localized reads that

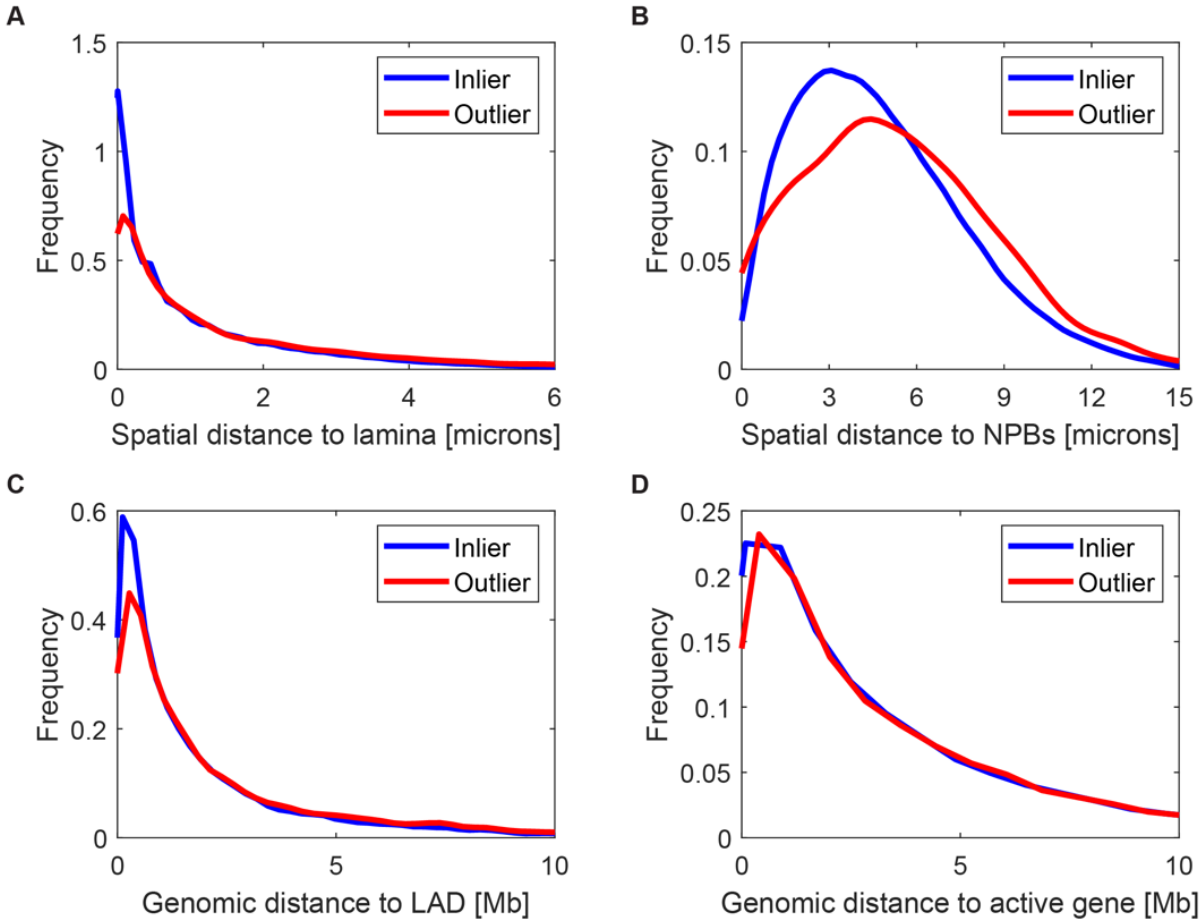

Figure 4-23. **Association of embryo non-colocalizing loci with nuclear landmarks and genomic annotations**. (**A**) Density plot showing how inlier points (reads that colocalize with a chromosome territory) and outlier points (reads that do not colocalize) differ in their spatial proximity to the nuclear lamina. Distributions are significantly different (K-S test, $p < 10^{-51}$). (**B**) Same as (A) but for spatial distance to the nearest nucleolus precursor body (NPB). Distributions are significantly different (K-S test, $p < 10^{-109}$). (**C**) Same as (A) but for genomic distance to the nearest lamina-associated domain, identified via DamID. Distributions are significantly different (K-S test, $p < 10^{-12}$). (**D**) Same as (A) but for genomic distance to the nearest highly-expressed gene in the corresponding developmental stage. Distributions are not significantly different (K-S test, $p > 0.01$).

overlapped a genomic position with a heterozygous SNP in either of the parental strains. 1.40% and 1.64% of genomic reads were uniquely assigned to the maternal (B6C3F1) and paternal (B6D2F1) genomes respectively, resulting in an average of 67 haplotype-informative reads per cell. To validate these assignments, we visualized the positions of haplotype-informative reads in the PN4 zygote (**Fig. 4-21C**). At this stage of development, the parental genomes remain segregated in the larger paternal and smaller maternal pronuclei. Based on this known feature, we assigned each chromosome territory to either the maternal or paternal genome in a semi-supervised manner (**Fig. 4-21D**). We found that 97.1% of haplotype-informative reads were concordant with this assignment, where non-concordant reads may be attributable to genomic sequencing errors, UMI matching errors, or strain impurities. We then used our haplotype-informative reads to assign entire chromosome territories to parent-of-origin across the 2-cell and 4-cell stage embryos [130]. This approach enables a strong majority of reads (82.26%), even those not overlapping a SNP, to be assigned to parent-of-origin through co-localization with haplotype-resolved reads in the same territory.

4.4.1 Developmental transitions in embryonic genome organization

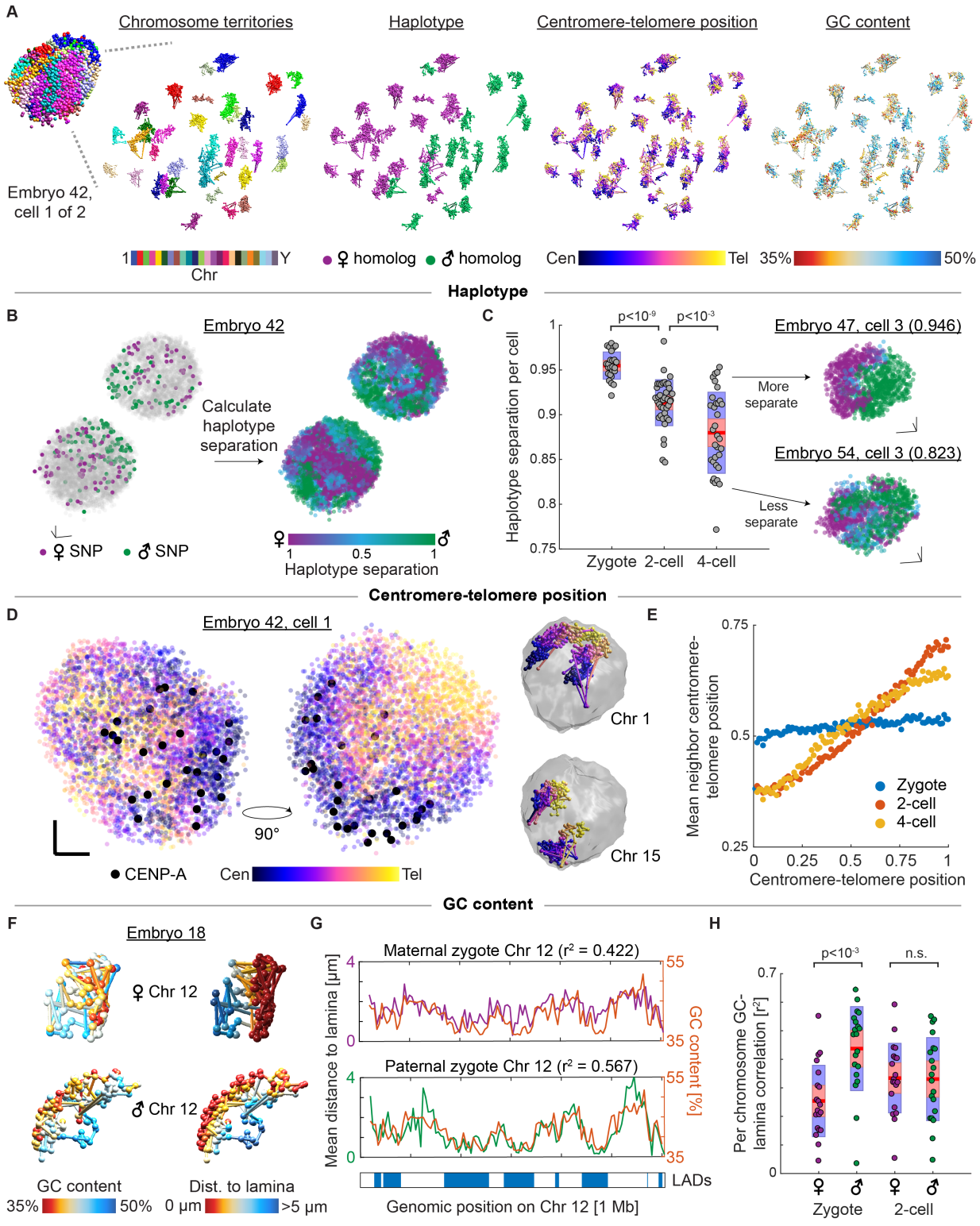

Figure 4-24

Figure 4-24. **IGS characterizes developmental transitions in embryonic genome organization**. (**A**) Exploded view of a single nucleus from a 2-cell embryo colored by chromosome territories, haplotype, centromere-telomere position, and GC content. (**B**) 2-cell embryo with spatially-localized reads colored by parental haplotype assignment (left) and haplotype separation score (right). (**C**) Boxplots showing mean haplotype separation score per cell across developmental stages (left; K-S test, $p < 10^{-8}$ and $p < 10^{-3}$). Grey dots represent mean scores of single cells. Distribution mean (red line), 95% confidence interval, (red box), and 1 standard deviation (blue box) are indicated. Two cells representing extreme scores (> 1 SD) are shown (right). (**D**) Nucleus from (A) with spatially-localized reads colored by centromere-telomere position, shown from two angles 90 degrees apart (left). Black dots indicate the position of CENP-A as identified from immunostains. Chr 1 and Chr 15 homologs from this cell are shown (right) to illustrate the Rabl-like configuration. (**E**) Mean centromere-telomere position of spatial neighbors as a function of centromere-telomere position for each stage. (**F**) Chr 12 homologs from a representative zygote with spatially-localized reads colored by GC content (left) and distance to lamina (right). (**G**) Plots showing the relationship between GC content and average distance to the nuclear lamina for 1 Mb bins in Chr 12 of the maternal and paternal zygotic pronuclei. Zygotic lamina-associated domains (LADs) defined by DamID are displayed below. (**H**) Boxplots showing Spearman's ρ between GC content and distance to lamina for 1 Mb bins, partitioned by haplotype and developmental stage (K-S test, $p < 10^{-5}$ and n.s.). Dots represent single chromosomes. Distribution mean (red line), 95% confidence interval, (red box), and 1 standard deviation (blue box) are indicated. n = 24 zygotes, 40 2-cell, 49 4-cell nuclei for all panels. Scale bars: 5 μm in all directions.

We then sought to examine previously-described principles of global genome organization, focusing on parental haplotype [163, 164], centromere-telomere position, [161, 117], and GC content [107, 117] (**Fig. 4-24A**).

We began by examining the spatial separation of parental genomes, as imaging studies have shown that maternal and paternal chromatin are spatially polarized in early embryos [163, 164]. To quantify the spatial separation of parental genomes across developmental stages, we analyzed the spatial inter-chromosomal neighbors of each read, and calculated a haplotype spatial separation score (**Fig. 4-24B**). We then averaged the separation scores for all reads in each cell. We found that the mean separation score significantly decreased between the zygote and 2-cell stages and between the 2-cell and 4-cell stages (K-S test, $p < 10^{-8}$ and $p < 10^{-3}$, **Fig. 4-24C**), consistent with earlier studies [163, 164]. The standard deviation of mean separation scores increased with each developmental stage (SD = 0.015 for zygotes, 0.026 for 2-cell, 0.045 for 4-cell), indicating that the degree of parental genome intermixing is heterogeneous within the embryo. Further, we observed no evidence that particular chromosomes were more likely to break this separation than others (**Fig. 4-25**). These results are concordant with the hypothesis that gradual mixing is a consequence of global chromosome repositioning following mitosis.

We next examined global spatial organization of the genome along the centromere-

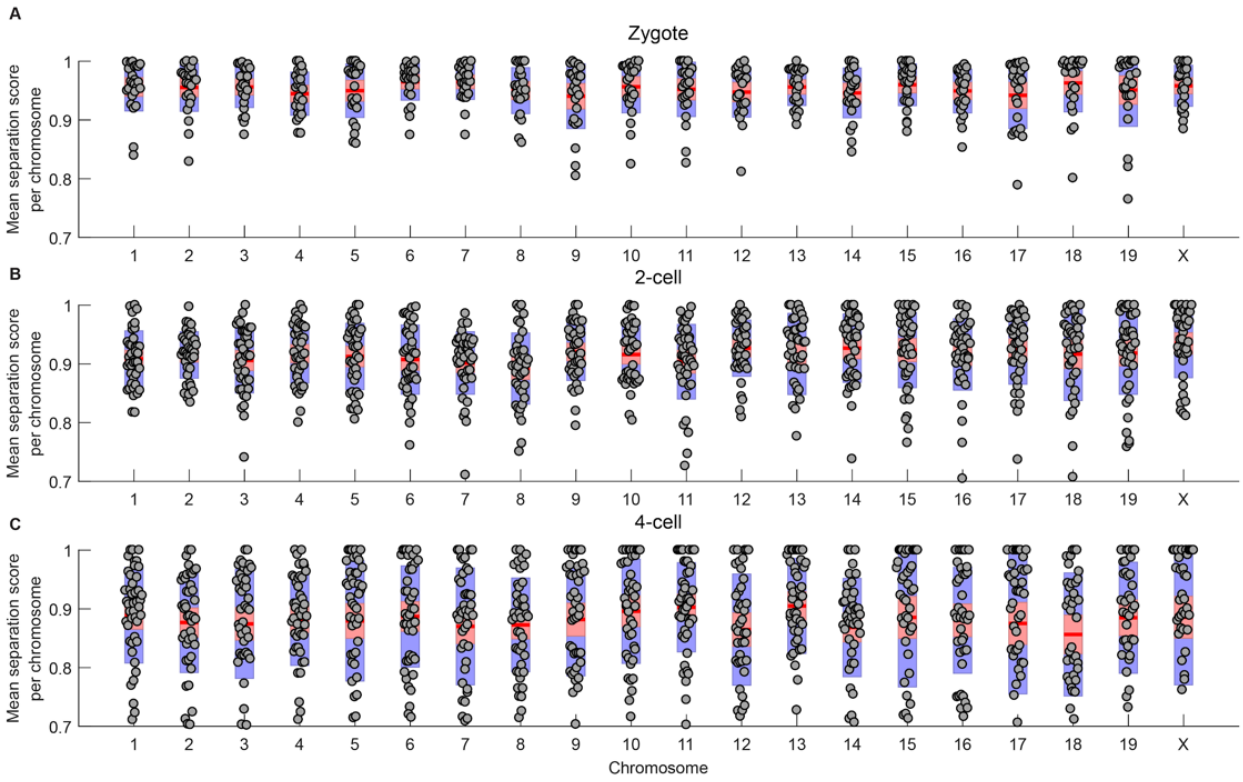

Figure 4-25. **Haplotype separation scores across chromosomes and developmental stages**. (**A**) Boxplots showing mean (red line), 95% confidence interval (red box), and 1 SD (blue box) to represent the distribution of separation scores (**Methods**) for each chromosome in zygotes. Each overlaid grey dot represents the separation score averaged over all spatially-localized reads mapped to the associated chromosome in a single cell. (**B**) Same as (A) but for the 2-cell stage. (**C**) Same as (A) but for the 4-cell stage.

telomere axes of chromosomes. Mouse chromosomes are acrocentric and are known to be arranged in a Rabl-like configuration in early embryos, in which centromeres cluster toward one side of the nucleus and distal telomeres cluster toward the other [163, 161]. To confirm this configuration in our data, we first measured the polarity of the CENP-A stain and found that centromeres in the 2 and 4-cell stages were significantly clustered toward one side of the nucleus (K-S test, $p < 10^{-4}$ and $p < 10^{-8}$). To analyze this configuration for all chromosome positions, we assigned each read a centromere-telomere score based on its genomic position along its chromosome. When we visualized these scores in a nucleus from a 2-cell embryo, we observed that the centromere-telomere scores were highly polarized, which was supported by co-localization of the CENP-A immunostain (**Fig. 4-24D**). To quantify this polarization across all stages, we calculated a spatial neighborhood centromere-telomere score for each read [130]. We then examined the relationship between centromere-telomere scores and neighborhood scores across all reads, and observed much stronger correlation in the 2-cell and 4-cell stages (Pearson's $r = 0.519$ and 0.502) than in the zygote ($r = 0.074$, **Fig. 4-24E**). The functional consequences of this transition to a Rabl-like configuration in 2- and 4-cell embryos remains unclear. In other contexts, this configuration is thought to be an extension of anaphase chromosome positioning into interphase, perhaps without cellular function [165]. On the other hand, it has also been hypothesized to restrict chromatin entanglement [166]; thus, it may be involved in constraining genome structure to enable the short cell cycles of the early embryo.

Finally, we examined the role of GC content in genome structure, which is strongly associated with A/B compartmentalization [107]. To study this effect, we first visualized individual homologs of Chr 12 from zygotic pronuclei (**Fig. 4-24F**). We observed that genomic reads from GC-poor regions tended to localize to the periphery of the nucleus, in line with reports describing the localization of the inactive B compartment [117]. To quantify this effect, we measured the correlation between GC content and distance to nuclear lamina across Chr 12 in all zygotes and observed that these two factors were correlated in both the paternal and maternal homologs (Spearman's $\rho = 0.794$, 0.649 respectively, **Fig. 4-24G**). We applied this approach to all chromosomes and found that the paternal homologs were significantly more correlated than their maternal counterparts in the zygote, but not in the

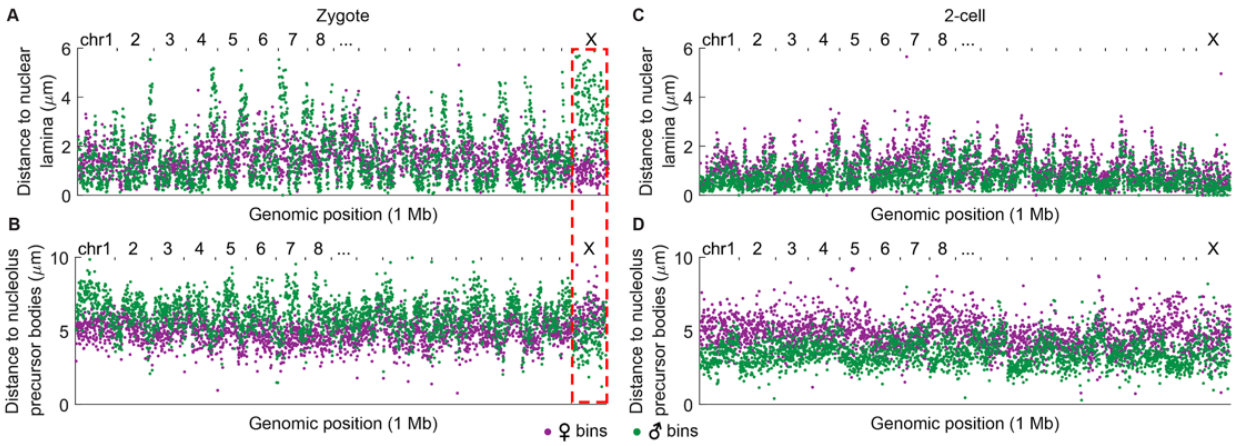

Figure 4-26. **Spatial distance to nuclear landmarks in the zygote and 2-cell embryos**. (**A**) 1 Mb haplotype-resolved genomic bins in the zygote, plotted by position in the mm10 genome and the average spatial distance of reads in the bin to the nuclear lamina, colored by parent-of-origin. The dotted red box highlights the unique spatial localization of the paternal X chromosome away from the nuclear lamina and close to a nucleolus precursor body. (**B**) Same as (AB), but for the average distance of reads in the bin to the nearest nucleolus precursor body. (**C**) Same as (A), but for the 2-cell stage. (**D**) Same as (A)**,** but for the average spatial distance of reads in the bin to the nearest nucleolus precursor body in the 2-cell stage.

2-cell stage (**Fig 4-24H;** K-S test, $p < 10^{-5}$, n.s.). This suggests that the degree of GC compartmentalization may be influenced by the differing biological histories of the pronuclei [154].

Intriguingly, when we examined the relationship between genomic position and distance to nuclear landmarks, we observed that Chr X seemed to localize especially far from the nuclear lamina and toward the NPBs in paternal pronuclei (**Fig. 4-26**). This finding extends models of the role of NPBs in establishing epigenetic asymmetry between parental X chromosomes rapidly after fertilization and in advance of imprinted X inactivation [152]. Taken together, these results demonstrate the ability of IGS to characterize 3D genome organization across diverse developmental stages with parent-specific resolution, and with respect to nuclear landmarks.

4.4.2 Detection of single-cell domains chromatin domains in zygotes

Next, we used IGS to examine subchromosomal spatial organization. We focused on our data with the highest genomic resolution, the zygotic pronuclei, where large-scale parent-specific reorganization of chromatin is thought to play an important role in ZGA [167]. First, we characterized the scaling relationship between mean spatial and genomic pairwise distance

in the zygotic parental genomes. Consistent with previous reports [154], we found that each parental genome had distinct scaling properties (**Fig. 4-30A, 4-30B**).

Reports analyzing genome structure in zygotes have suggested that paternal chromatin exhibits unusually weak higher order structure ($>$ 2 Mb) [155, 156, 154]. In accordance with these reports, we found that spatial distance matrices generated from the population ensemble indeed exhibited little off-diagonal structure (**Fig. 4-30C, 4-27 all chromosomes**. Furthermore, evidence suggests that paternal zygotic chromatin exhibits unusually weak A/B compartmentalization [155, 156, 154], and unusually large lamina associated domains [153]. Our ensemble data corroborated these reports and correlated well with Hi-C (mean Pearson's r = 0.84) when analyzed in terms of lamin proximity (**Fig. 4-30D, 4-28**, all chromosomes). However, when we examined single-cell distance matrices, we found that, unlike the ensemble, single paternal pronuclei generally exhibited large blocks of spatially associated chromatin (**Fig. 4-30E,** left). To distinguish these blocks from population-defined topological domains identified in Hi-C studies, we term them single-cell domains (SCDs).

To better understand the nature of SCDs in paternal zygotes, we systematically identified individual domains in single cells [130]. The SCDs we identified corresponded well to spatially distinct clusters identified by visual inspection (**Fig. 4-30E**, right). When we examined SCDs across cells and chromosomes, we observed that they were large (median size 17.5 Mb, 10 Mb inter-quartile range) relative to canonical features defined by Hi-C, and had heterogeneous sizes and boundary positions (**Fig. 4-29**). We proceeded to assess the strength of all SCD boundaries in single cells, and found they were significantly stronger and more variable than boundaries identified by the same method in the ensemble matrices (**Fig. 4-30F;** K-S test, $p < 10^{-17}$; 95% CI for Cohen's d, (0.56, 0.77), determined by bootstrapping [130])). Together, these observations suggest that the weak ensemble structure may be explained by the variability of single-cell structures. Finally, we investigated the association of SCDs with nuclear landmarks, which may suggest organizing principles, and found that their boundaries and interiors were, respectively, significantly more lamin distal and more lamin proximal than expected (**Fig. 4-30G** [130]**)**. We found this striking in light of previous electron microscopy studies showing discrete micron-scale lamina-associated chromatin domains following ZGA [168].

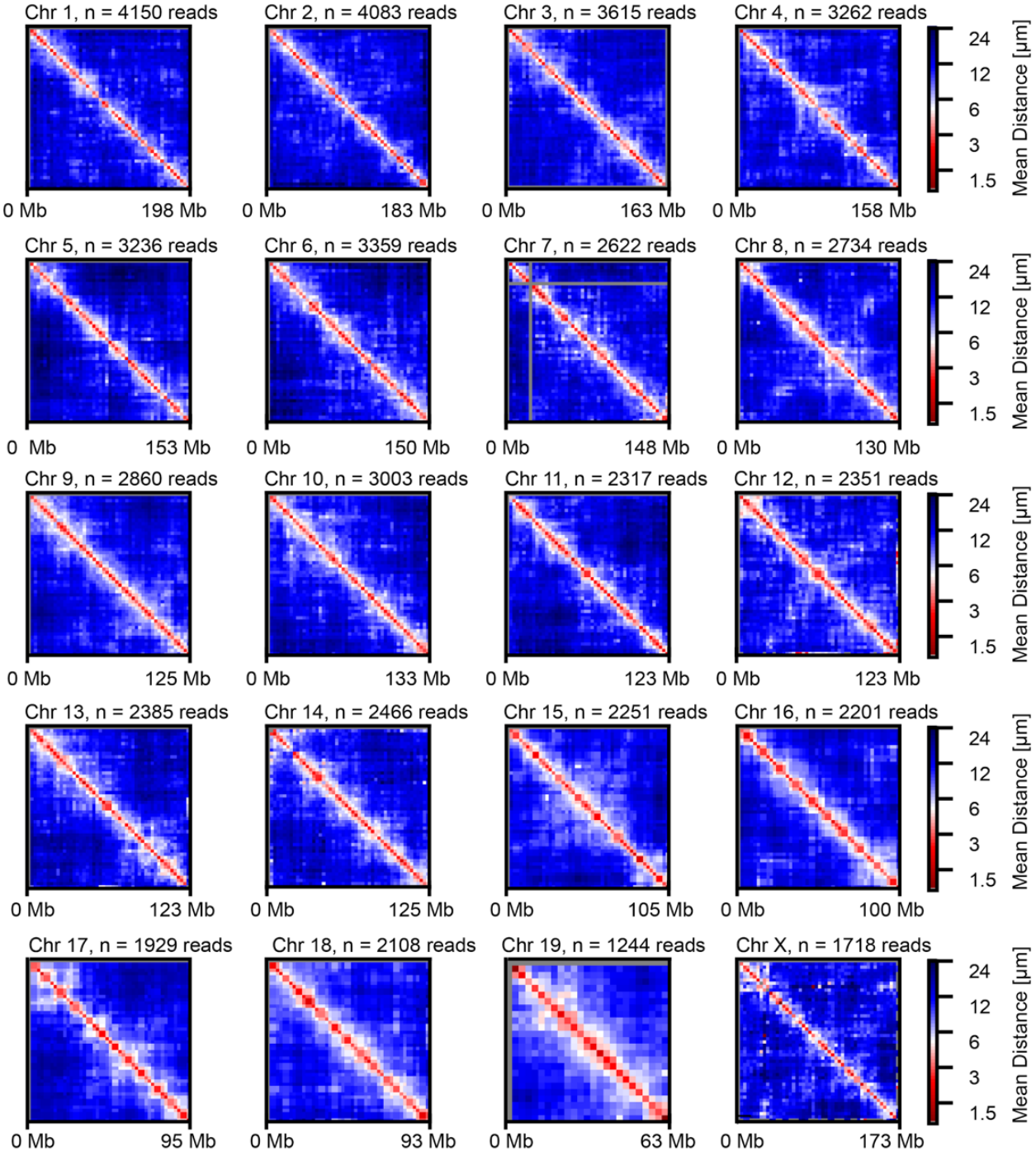

Figure 4-27. **Paternal zygotic ensemble distance matrices across all chromosomes**. Mean spatial distance matrices for the ensemble of paternal zygotic chromosomes. In comparison to single-cell matrices (e.g. Fig. 4-30E), the ensemble matrices show little in the way of off-diagonal structure. Number of reads analyzed for each chromosome indicated. All autosomes describe a population ensemble of 24 pronuclei, except for ChrX, which describes an ensemble of 14 pronuclei.

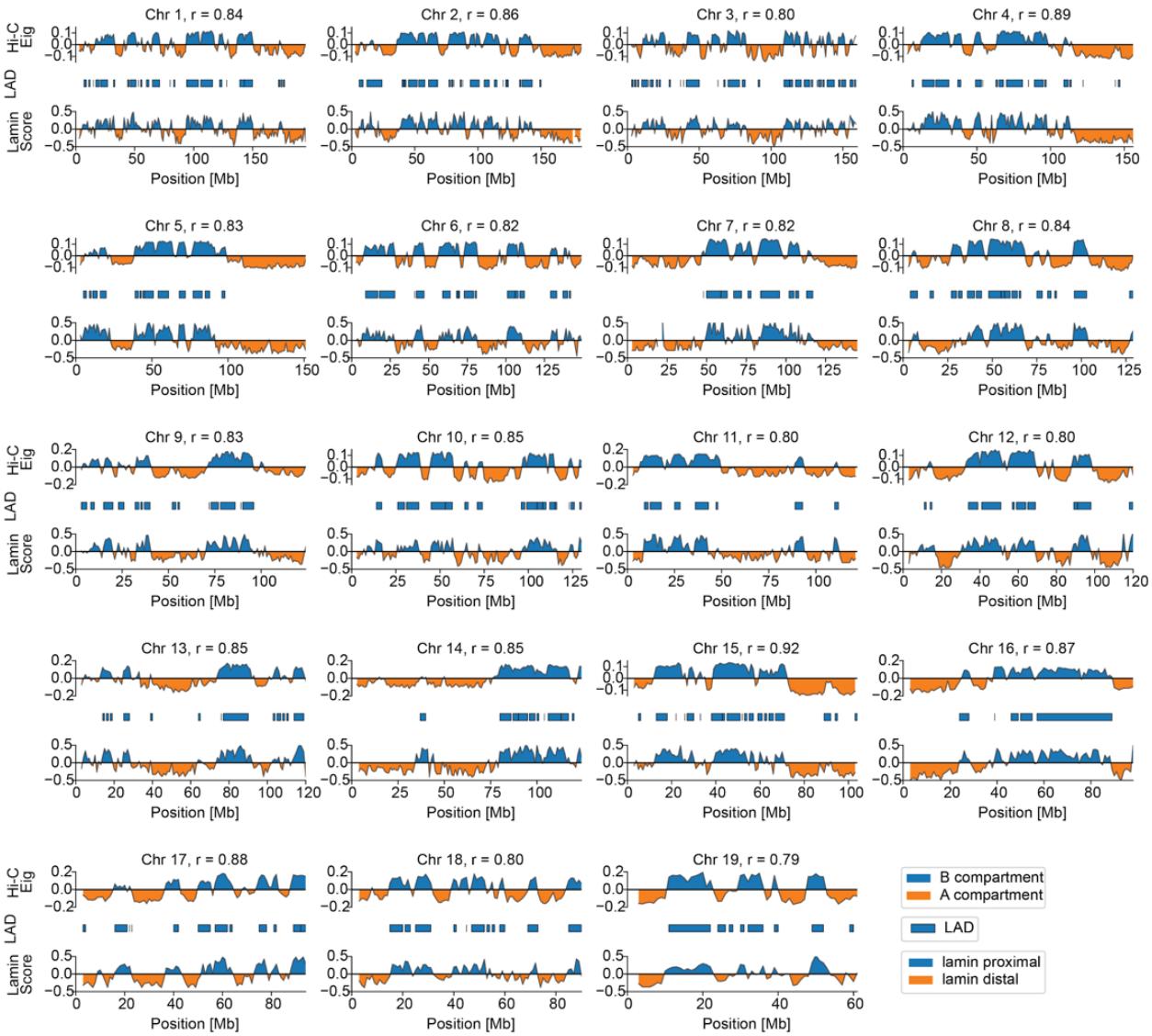

Figure 4-28. **Comparison of IGS with Hi-C and DamID in the paternal zygote**. Each subplot shows, for a specific chromosome in paternal zygotes, Hi-C-defined compartmental status (top), DamID-defined LAD status (mid), and IGS-defined lamin proximity (bottom). Pearson correlation coefficient between lamin proximity score and Hi-C eigenvalue is indicated. Number of reads analyzed for each chromosome indicated. All analyses describe a population ensemble of 24 pronuclei. Chr X not indicated due to lack of DamID data.

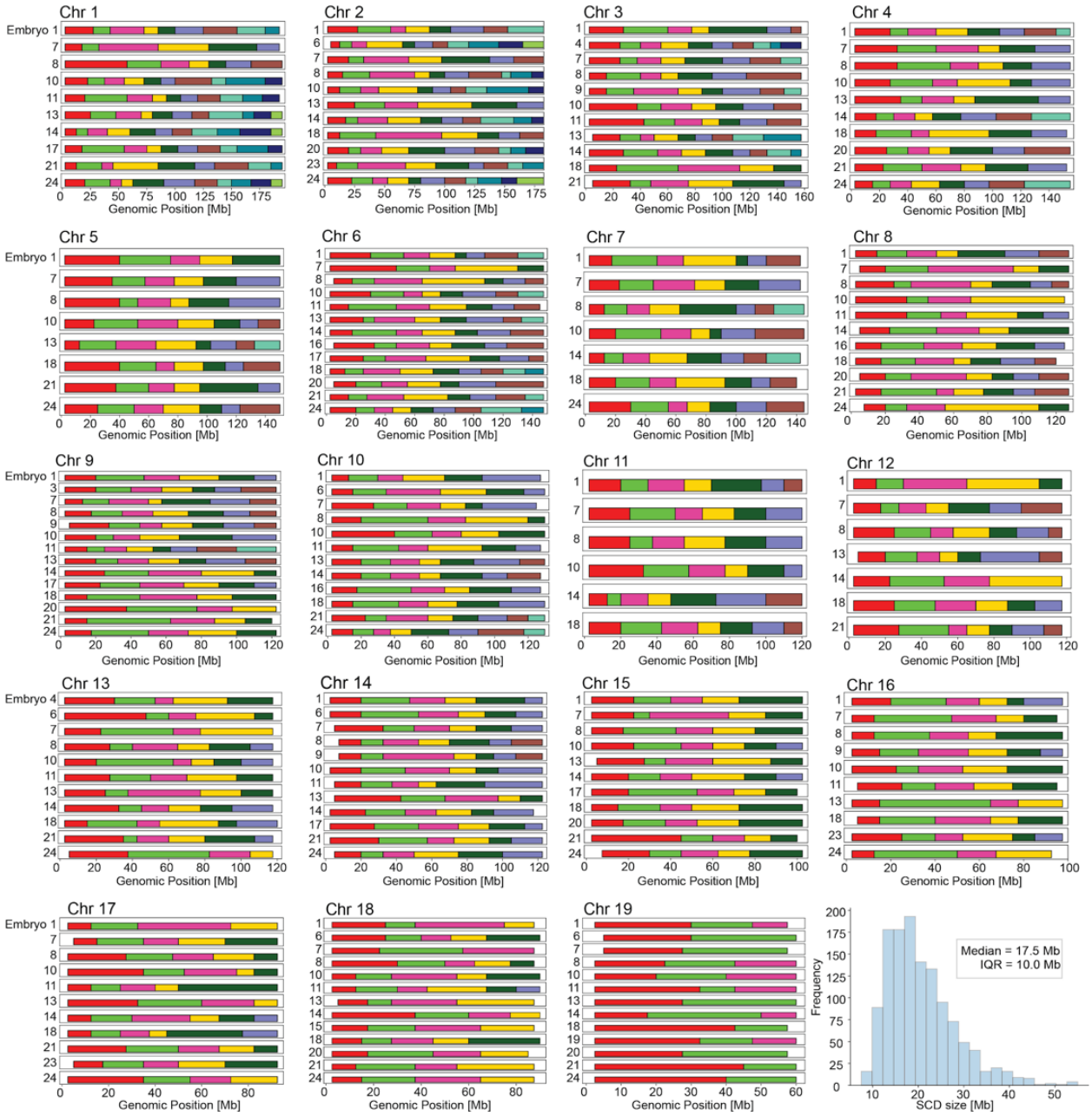

Figure 4-29. **Single-cell domains have heterogeneous sizes and boundaries**. For all high-coverage (>90%) paternal zygotic autosomes analyzed at a matrix resolution of 2.5 Mb, SCDs were detected at that resolution (**Methods**) and are indicated. Lower right, histogram of 1262 detected SCDs spanning two or more bins. Median SCD size and IQR are indicated.

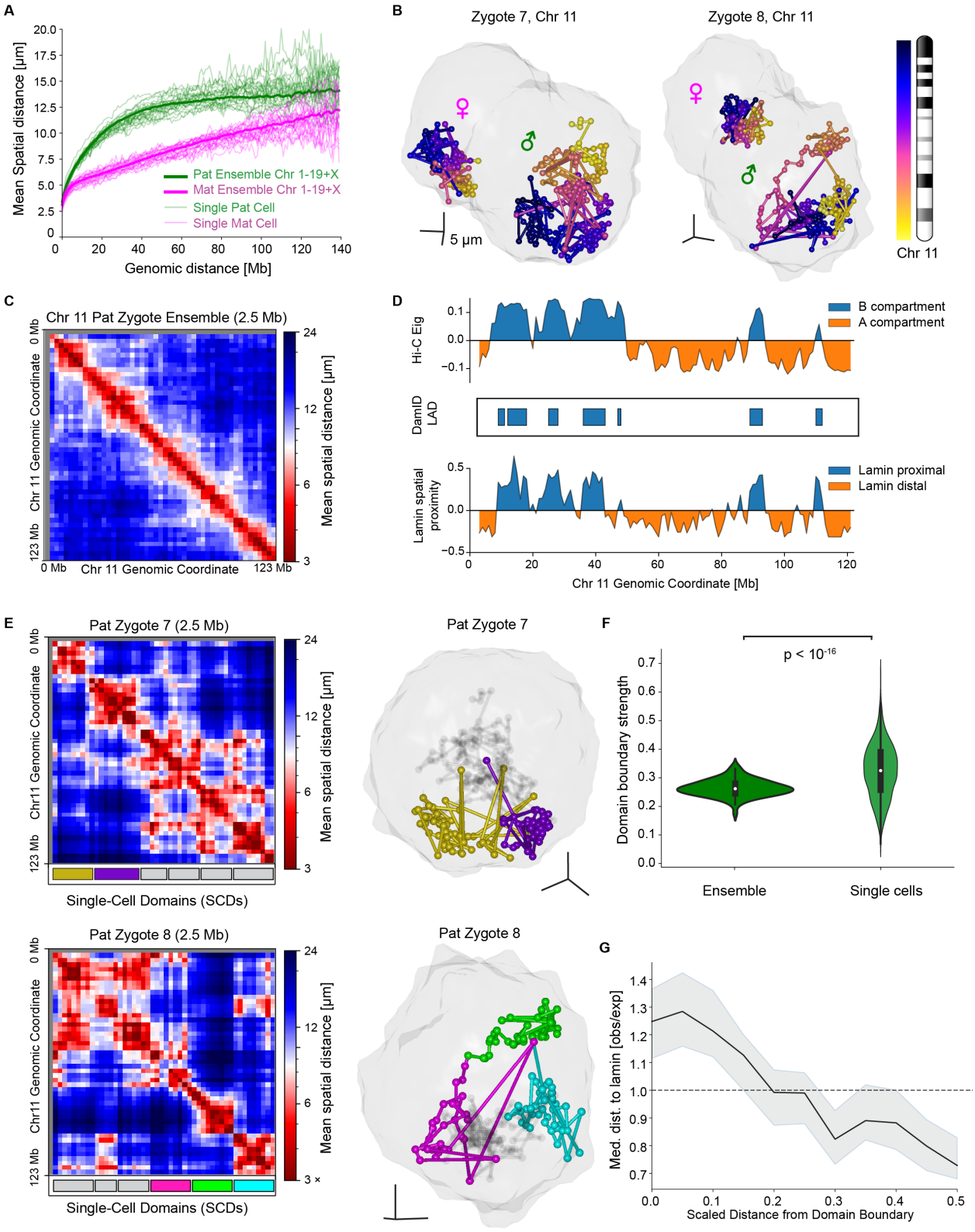
A
Mean Spatial distance [µm]
Genomic distance [Mb]
Pat Ensemble Chr 1-19+X
Mat Ensemble Chr 1-19+X
Single Pat Cell
Single Mat Cell
B
Zygote 7, Chr 11
Zygote 8, Chr 11
5 µm
Chr 11
C
Chr 11 Pat Zygote Ensemble (2.5 Mb)
Chr 11 Genomic Coordinate
Mean spatial distance [µm]
D
Hi-C Eig
B compartment
A compartment
DamID LAD
Lamin spatial proximity
Lamin proximal
Lamin distal
Chr 11 Genomic Coordinate [Mb]
E
Pat Zygote 7 (2.5 Mb)
Chr11 Genomic Coordinate
Single-Cell Domains (SCDs)
Pat Zygote 8 (2.5 Mb)
Pat Zygote 7
Pat Zygote 8
F
p < 10-16
Domain boundary strength
Ensemble
Single cells
G
Med. dist. to lamin [obs/exp]
Scaled Distance from Domain Boundary

Figure 4-30

Figure 4-30. **IGS characterizes developmental transitions in embryonic genome organization**.(**A**) Global relationship between genomic and spatial distance in zygotes for all chromosomes, distinguishing the parental genomes. (**B**) Visualization of Chr 11 homologs in two zygotes according to parent-of-origin. (**C**) Population ensemble mean spatial distance matrix for paternal Chr 11, constructed at 2.5 Mb resolution (24 zygotic pronuclei, 2317 reads). (**D**) Comparison across measurement modalities for the population of paternal zygotic Chr 11. Top row: Hi-C defined eigenvalues and compartment calls. Middle row: DamID-defined population lamina associated domains. Bottom row: lamin-proximal and lamin-distal regions defined with IGS (24 zygotic pronuclei, 2317 reads). (**E**) Top left: single-cell mean distance matrix for paternal Chr 11 in a representative zygote, with single-cell domain boundaries (SCDs) marked below (263 reads). Top right: visualization of individual paternal SCDs in the same zygote. To assist visualization, two SCDs are shown in color (purple, gold), while the remaining SCDs are shown in grey. Bottom left: single-cell mean distance matrix for paternal Chr 11 in a second representative zygote, with single-cell domain boundaries marked below (213 reads). Bottom right: visualization of three paternal SCDs in the second zygote. To assist visualization, three SCDs are shown in color (magenta, lime, cyan) while the remaining SCDs are shown in grey. (**F**) Comparison of single-cell and ensemble domain boundary strengths spanning all detectable boundaries in Chr 1-19+X (74 ensemble boundaries, 1057 single-cell boundaries, K-S test, $p < 10^{-17}$). (**G**) Scaled distance from SCD boundary versus observed/expected median distance to nuclear lamina, measured genome-wide (Chr 1-19+X, N = 1262 SCDs). Envelope indicates 95% confidence interval determined by bootstrapping.

Recently, both polymer simulations [154] and direct observation of chromatin structure [123] have shown how variable domain-like structures can exist in single cells when higher order ensemble structure is lacking. We speculate that, given the weak ensemble structure, the SCDs observed here may involve a similar phenomenon. The SCDs described here are larger than canonical Hi-C defined features, and were not detected in earlier single-cell Hi-C studies in zygotes [154, 157], perhaps because they are organized on length scales which are relatively less accessible to Hi-C measurements. It may be interesting to investigate the extent to which SCDs are governed by mechanisms related to the nuclear lamina, which perhaps modulates underlying epigenetic [147] or polymer-intrinsic [154] domain-forming behaviors of chromatin.

4.4.3 Epigenetic memory of global chromosome positioning

Embryonic development is thought to involve epigenetic transmission of structural and regulatory features of chromatin organization through clonal cell lineages within individual embryos [160]. These mechanisms play an important role in breaking initial symmetry [158], engaging clonal lineage-specific gene expression programs [169], and cell fate commitment [151]. Intercellular asymmetries influencing the developmental fate of clonal lineages have been reported as early as the 4-cell stage within individual embryos [170, 169, 158, 159].

In order to study clonal lineage-specific features at the single-cell level, it is necessary to resolve and compare cells within the same embryo. Chromosome territories have been found to form early in interphase and subsequently maintain their relative positions until prophase [171, 172, 119], so we reasoned that comparison of chromosome positioning would be a robust way to quantify the similarity of global genome organization between interphase cells. Live-cell studies using non-specific photopatterning of the nucleus have demonstrated similarity in global genome organization between sister cells in culture [172], and indeed, visual inspection of pairs of chromosomes suggested that cells within an embryo share similar chromosomal positions (**Fig. 4-31A**). We quantified similarity by comparing single-cell autosome distance matrices of pairs of cells within and between individual embryos (**Fig. 4-31A, 4-31B,** [130]). In 2-cell embryos, we found that global chromosome positioning in sister cells was significantly more correlated than in pairs of cells from different embryos (K-S test, $p < 10^{-15}$; **Fig. 4-31C, Fig. 4-32**). These results suggest that cells within 2-cell embryos may share memory of their common initial chromosome positioning during zygotic metaphase, if not earlier.

Next, we asked whether the similarity shared by sister cells in 2-cell embryos might be epigenetically transmitted across the second cell division, i.e. to cousin cells. While earlier work has not found heritability in the radial positioning of individual loci [173, 174], widely varying degrees of similarity in global genome organization between mother and daughter cells have been reported [172, 171]. We constructed putative clonal lineage trees within each 4-cell embryo by using the ranked correlations of autosome distance matrices for each pair of cells to classify putative sister and cousin cells. The most correlated pair of cells in each embryo was designated as one set of putative sister cells, thus implying the remainder of the tree (**Fig. 4-31D, Fig 4-33**, [130]). As expected by the definition of the tree, we found that global chromosome positioning was significantly more correlated between putative sister cells than between pairs of cells from different 4-cell embryos. However, we also found that the same held true for putative cousin cells, which was not expected (**Fig. 4-31E**, **Fig. 4-32**; K-S test, $p < 10^{-14}$ and $p < 10^{-3}$; 95% CI for Cohen's d, (0.55, 1.29), determined by bootstrapping, [130]). Together, these results demonstrate clonal lineage-specific similarity in global chromosome positioning in early embryos, and imply that epigenetic memory of

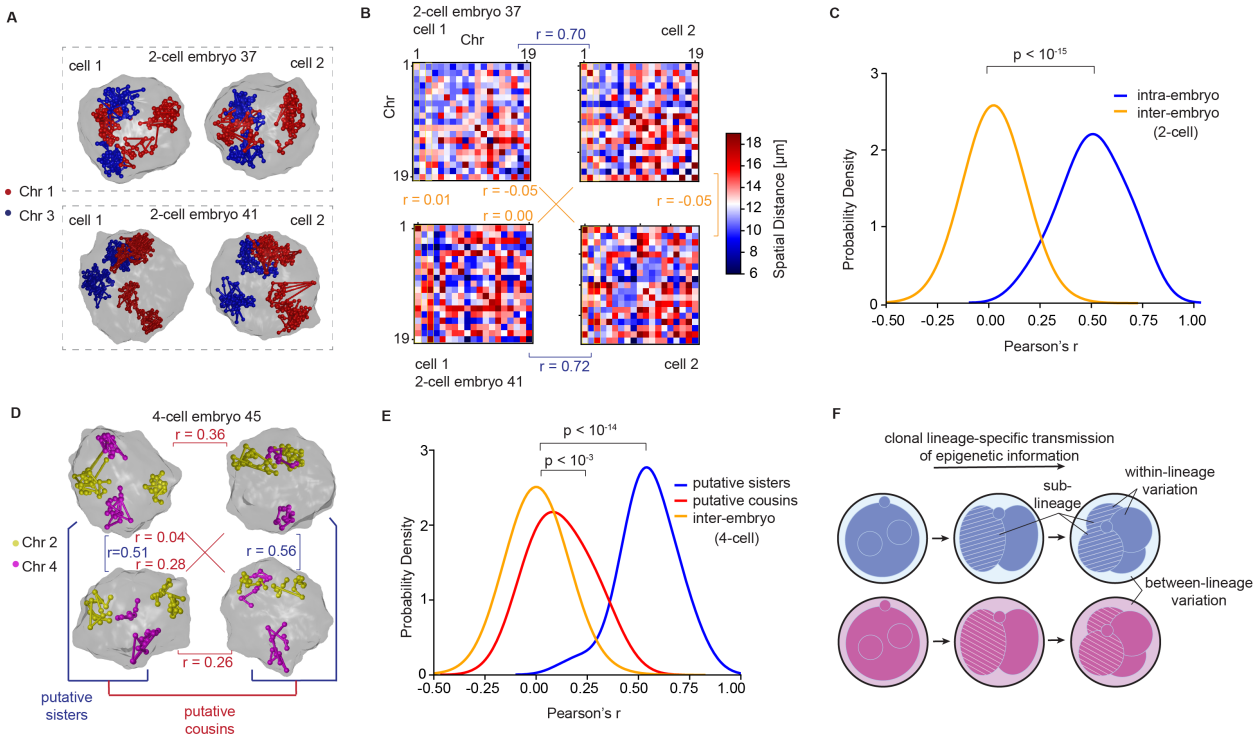

Figure 4-31. **IGS enables high-resolution genomic and spatial profiling of intact early mouse embryos**. (**A**) Positioning of Chr 1 and 3 in the cells of 2-cell embryos 37 (top) and 41 (bottom). (**B**) Pairwise correlations between autosome distance matrices for the cells in (A). Intra-embryo and inter-embryo correlations are shown in blue and orange, respectively. (**C**) Probability distributions of correlations between autosome distance matrices for intra-embryo and inter-embryo pairs of cells among 2-cell embryos. K-S test, $p < 10^{-15}$; n = 20 intra-embryo pairs and n = 760 inter-embryo pairs, among 20 2-cell embryos. (**D**) Positioning of Chr 2 and 4 in the cells of 4-cell embryo 45. Pairs of cells are putatively classified as sister and cousin cells based on correlation of global chromosome positioning, with the most correlated pair classified as sisters. Correlations between sister and cousin cells are shown in blue and red, respectively. (**E**) Probability distributions of correlations between autosome distance matrices for pairs of putative sister cells, cousin cells, and inter-embryo pairs of cells among 4-cell embryos. K-S test, $p < 10^{-14}$ for sisters vs. inter-embryo and $p < 10^{-3}$ for cousins vs. inter-embryo; n = 18 sister pairs, n = 36 cousin pairs, and n = 933 inter-embryo pairs, among 13 4-cell embryos. (**F**) Model of epigenetic memory transmission within clonal lineages.

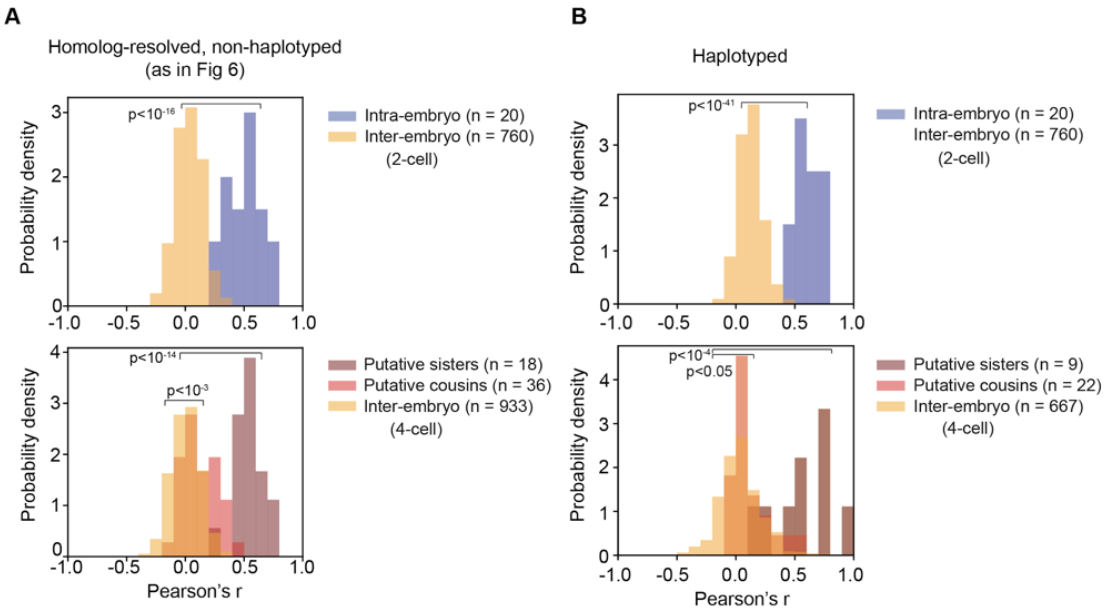

Figure 4-32. **Correlation of global autosome positioning for pairs of blastomeres**. (**A**) Histogram representation of the data in Fig. 6C (top) and 6F (bottom). Correlations of global autosome positioning for intra- and inter-embryonic pairs of cells for 2-cell embryos (top), and for putative sister, putative cousin, and inter-embryonic pairs of cells for 4-cell embryos (bottom). Correlations were calculated in a homolog-resolved but non-haplotyped manner (**Methods**). (**B**) As in (A), but with correlations calculated in a haplotype-resolved manner (**Methods**). The number of pairs compared in each category is indicated in the legend. K-S test was used to test for significance in all cases.

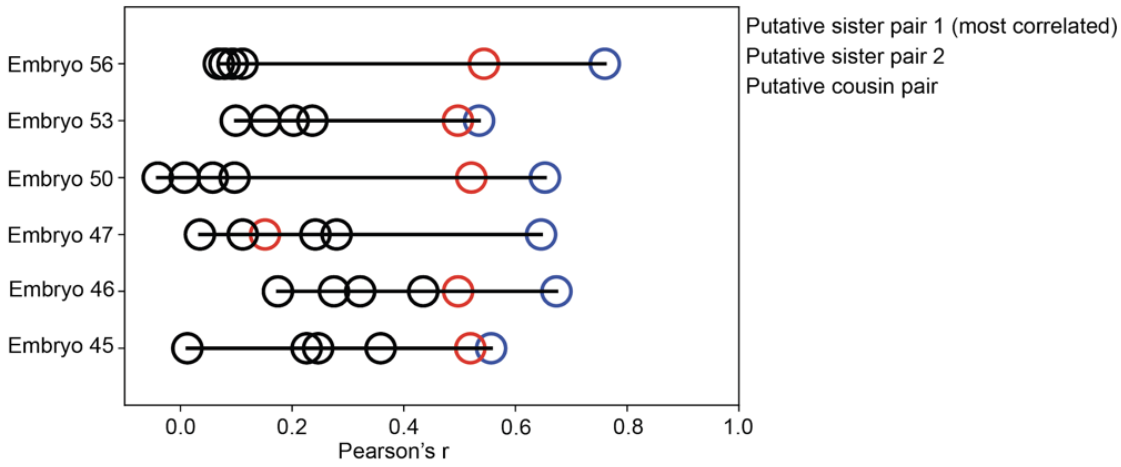

Figure 4-33. **Correlations of putative sister and cousin pairs for each complete four-cell embryo**. Correlations for each pair of cells in each four-cell embryo for which the genomes of all four cells had >150 reads . The putative sister pair 1 is the most correlated pair in the embryo (blue), and putative sister pair 2 is the other non-overlapping pair in the embryo (red).

chromosome positioning is transmitted from mother to daughter cells during the second cleavage (**Fig. 4-31F**). While the mechanisms are not fully clear, this memory may reflect minimal relative repositioning of chromosomes during congression to and departure from the metaphase plate, perhaps due to the Rabl configuration and short cell cycles of the early embryo. This mitotic heritability of global chromosome positioning may influence processes that affect the viability and phenotype of the developing organism, such as rates of homologous recombination (HR)-mediated double-strand break (DSB) repair [175, 176] and the distribution of translocations [177, 178, 179] in the early embryo.

4.5 Conclusion

Using microscopy to look at biological structure has been an essential strategy for more than three hundred years. However, in the post-genomic era, the limited scale of traditional microscopy in terms of colors and throughput prohibits looking at things with a genome-wide view. In this thesis, we present a set of methods that cut through these limitations, enabling highly multiplexed molecular measurements with spatial context, i.e. scalable *in situ* genomics.

In this work we present *in situ* genome sequencing, unifying sequencing and imaging of genomes in intact samples. This unified approach enables *de novo* discovery of spatial organization of genomes across length scales, from single-cell subchromosomal domains to intercellular relationships. Because IGS is both sequencing and imaging-based, it can be extended in either modality based on the needs of specific experiments. We demonstrate this in early mouse embryos through integration of genotype information to spatially resolve the maternal and paternal genomes, through integration of immunofluorescence to localize genomic loci relative to nuclear landmarks, and by using whole-embryo spatial information to infer clonal lineages. This contextual information enabled us to uncover single-cell chromatin domains in zygotes with lamin-distal boundary positions and lamin-proximal interiors, as well as heritable correlations in global chromosome positioning within single early embryos.

While extant methods such as multiplexed DNA FISH and single-cell Hi-C are well-suited to measuring aspects of spatial genome organization, they cannot currently be combined in the same cell. With IGS, we spatially localize hundreds to thousands of genomic loci in single cells, achieving genomic resolutions comparable to recent genome-wide approaches based on targeted DNA FISH in fibroblasts [76, 77]. Unlike these targeted methods, we further show how IGS can perform untargeted sequence localization, resolve genome structure of maternal and paternal alleles, and be applied in 3D nuclei and thick intact samples. However, due to its genome-wide sampling frequency (at most ~1 Mb in this report), IGS is currently limited in its ability to systematically examine specific genetic loci in specific cells. Targeted DNA FISH or single-cell Hi-C are thus currently more appropriate for applications requiring high-resolution interrogation of genomic features such as TAD boundaries [123, 180, 154] or enhancer-promoter loops [127, 117], and chromosome painting methods [181] may be preferred when high-throughput visualization of chromosomes territories is required. IGS therefore joins an ecosystem of conceptually new approaches [182, 183, 184] complementary to these more well-established methods.

We expect that improvements to IGS will further enable the study of genome sequence, structure, and function. In addition to the cultured cells and early embryos presented here, we anticipate extension of IGS to a broader range of cell types and intact tissues. Outside of development, IGS may be well-suited to study cancers, in particular how copy number

instability and translocations contribute to tumor heterogeneity and alter nuclear morphology [185]. The transposase-based library construction used in IGS may also be extended to measure the spatial localization of the accessible genome [186, 131]. Further, because nuclear volume is the primary constraint on the amplicon yield of IGS, we anticipate many-fold improvements in yield and resolution, either through smaller amplicons [187], or preferably through integration of IGS with Expansion Microscopy [188, 189], which simultaneously increases nuclear volume by 50-100-fold and enables super-resolved imaging by physical expansion of samples. Finally, following our proof-of-concept integration with immunostaining, we expect increasingly multiplexed multi-omic [190] variations of IGS will be possible. We anticipate *in situ* genome sequencing will be instrumental in unifying genomics and microscopy, and therefore sequence and structure.

4.6 Materials and Methods

4.6.1 Brief Methods

Library construction

Cells were grown and fixed on a glass coverslip, and embryos were fixed in solution and immobilized in polyacrylamide gel in a 6-well plate. Phosphorylated DNA adapters were inserted into fixed genomic DNA *in situ* by incubating samples 1 hr (cells) or overnight (embryos) with transposase. Hairpins were hybridized to the adapters on either side of the insert, and the complex was circularized by gap-fill ligation. Hairpins contained either a UMI and primers for *in situ* and *ex situ* sequencing, or an RCA primer hybridization site. RCA primers were hybridized, and RCA was performed overnight, with aminoallyl-dUTP spiked into the reaction. Amplicons were crosslinked by reacting with BSPEG(9).

***In situ* sequencing**

For cells, the coverslip was mounted in a flow cell and *in situ* sequencing reactions were performed using automated fluidics. For embryos, reactions were performed manually in a 6-well plate. Samples were treated with calf intestinal phosphatase before the first primer hybridization and before each cleavage reaction. *In situ* sequencing was performed using

sequencing-by-ligation chemistry. Samples were exchanged into an imaging buffer following each round of sequencing. Images were acquired using confocal microscopy. Immunostaining and immunofluorescence imaging were performed after *in situ* sequencing.

***Ex situ* sequencing**

Samples were transferred into solution and used as input to a PCR reaction. The resulting library was sequenced using high-throughput paired-end sequencing and then then aligned to hg38 for PGP1f and mm10 for the early mouse embryos. Aligned reads that overlapped alleles that differ between the B6C3F1 and B6D2F1 strains were annotated as haplotype-informative.

Computational integration of *in situ* and *ex situ* sequencing data

In situ and *ex situ* sequencing data were computationally integrated using a probabilistic UMI matching approach that borrows principles from single-bit error correction to account for signal intermingling between densely-packed amplicons, as well as the decay of sequencing quality over successive rounds of *in situ* sequencing.

Detailed methods for sample preparation, library construction, multimodal sequencing, imaging, image analysis, and all analyses of data are described in **Section 4.6.2**. A description of methods for kit-free synthesis of *in situ* sequencing reagents is given in **Section 4.6.3** and **Tables 4.1** and **4.2**. A description of cost, complexity and throughput of the method is found in **Section 4.6.4** and **Table 4.3**.

4.6.2 Detailed Materials and Methods

4.6.2.1 Sample processing and library preparation

Experiments performed

Two separate IGS experiments were performed, with one experiment examining cultured fibroblasts and the other examining embryos.

Cell culture, fixation and permeabilization

PGP1f (Coriell GM23248) human primary skin cells were cultured to 70% confluence at 37°C with 5% CO_2 in 100 mm TC-Treated Culture Dishes (Corning) in Dulbecco's Modified Eagle Medium (DMEM), high glucose, GlutaMAX Supplement (Gibco 10566016), supple-

mented with 15% (v/v) fetal bovine serum (FBS) (Life Technologies, Inc 16140071), 1% (v/v) MEM Non-Essential Amino Acids Solution (Gibco 11140050), 100 units/mL penicillin and 100 μg/mL streptomycin. Cells were then treated with 0.25% Trypsin-EDTA (Gibco 25200056) for ~3 min at 37°C, mixed with culture media, centrifuged at 200xg for 5 min, and resuspended in 10 mL of fresh culture media. 125 μL of resuspended cells were then seeded into ethanol-sterilized, Matrigel-treated wells formed by 9 mm CultureWell silicone gaskets (Grace BioLabs 103240) attached to 40 mm circular coverslips (Bioptechs 40-1313-0319), except for the cells shown in **Fig. 4-3**, for which 90 μL of resuspended cells were seeded into Matrigel-treated CultureWell chambered coverglass wells (Grace BioLabs 112358), which were pre-sterilized. Ethanol sterilization was performed by incubating the well with 70% ethanol for 15 min followed by aspiration and air-drying for 30 min. Matrigel treatment was performed by incubating the well with Matrigel Matrix, LDEV-free (Corning) diluted 1:50 in culture medium for 30 min. Cells were incubated overnight to allow them to attach to the coverslip. Cells were then washed once with 1x phosphate-buffered saline (PBS) (Gibco 10010023) and fixed with methanol-free paraformaldehyde (Electron Microscopy Sciences 15710) diluted to 4% in PBS for 10 min, washed with PBS, permeabilized with 0.5% Triton-X 100 (Sigma 93443) in PBS for 10 min, and then washed twice with PBS. Fixed, permeabilized cells were stored at 4°C in PBS up to three weeks.

Embryo collection

Cryopreserved viable embryos resulting from a cross of B6D2F1/Hsd males (Envigo) and superovulated B6C3F1/Hsd females (Envigo) were purchased from Embryotech Laboratories, Inc. Zygotes, 2-cell, and 4-cell embryos were collected at 22 hours, 43 hours, and 49 hours post human chorionic gonadotropin injection, respectively, by Embryotech Laboratories, Inc.

Embryo thawing, removal of zona pellucida, and fixation

Embryos were thawed as per instructions from Embryotech Laboratories, Inc.: straws containing embryos cryopreserved in freezing medium were thawed at room temperature for 2 minutes and then incubated in a 37°C water-bath for 1 minute, after which they were gently pushed out of the straw into a droplet of EmbryoMax Advanced KSOM Medium (Sigma MR-101). Embryos were then transferred to another droplet of the same medium

under oil and incubated for 1 hour at 37°C with 5% CO_2 to recover from cryopreservation. The embryos were then briefly incubated in EmbryoMax Acidic Tyrode's solution (Sigma MR-004-D) to remove the zona pellucida and then transferred to a droplet of M2 medium (Sigma M7167). The zona-free embryos were then fixed in methanol-free paraformaldehyde 15710) diluted to 4% in PBS containing 1% polyvinylpyrrolidone, MW 360 kD (K90) (VWR AAJ61381-30) for 10 minutes at room temperature, followed by two rinses and final storage in PBS containing 1% polyvinylpyrrolidone, MW 360 kD (K90). Embryos were embedded in 4% polyacrylamide in a glass-bottom plate within 1 hr of fixation, to keep them immobilized during library preparation and *in situ* sequencing.

Embedding of embryos in polyacrylamide gel

The inner surface of a glass-bottom plate was incubated for 1 minute with 3 - (Trimethoxysilyl) propyl methacrylate (Sigma 440159) diluted 1:1000 in 80% ethanol, 2% acetic acid, 18% H_2O, and then washed 3 times in ethanol to functionalize the glass surface with methacrylate moieties. A cut piece of a glass microscope slide, which would be used in casting the polyacrylamide gels, was incubated with Sigmacote (Sigma SL2) for 1 minute and then washed 3 times in H_2O to make the surface hydrophobic. Near each edge of the hydrophobic glass, two layers of Invisible Tape (Universal 83412) were attached to the glass to form spacers roughly 100 μm thick. Fixed embryos were quickly transferred through 3 droplets of polyacrylamide gel monomer solution (4% acrylamide (VWR 97064-870), 0.5% N,N'-diallyltartramide (Sigma 156868), 1X PBS, 0.1% Tween-20 (Sigma 11332465001), 0.1% ammonium persulfate (Sigma A3678), 0.1% N,N,N',N'-tetramethylethylenediamine (Sigma T7024), 1:100 dilution of Fluorescent Tetraspeck 0.2 μm beads (Life Technologies, T7280)) and then transferred to the 6-well plate along with a small droplet of monomer solution taken from the final droplet. Tween was included in the monomer solution to prevent the embryos from sticking to surfaces. The hydrophobic glass slide with tape spacers was placed over the droplet to confine the monomer solution and embryos to a thin layer. The glass-bottom plate was placed in an air-tight container with a few ml of water in the bottom, which was then purged of oxygen by placing the needle of a N_2 gas line into one of two holes poked in the lid of the container for two minutes. Water is included in the container to prevent gels from drying. The holes in the lid of the container were sealed, and the container was incubated at 37°C for 2 hours.

The hydrophobic glass was removed from the gel using tweezers, and the gel was washed 3 times in PBS.

Permeabilization of embryos

Following embedding in polyacrylamide gel, the embryos were incubated in 0.5% Triton-X 100 in PBS for 15 minutes, rinsed once and then washed 3 times for 5 minutes in PBS.

Tn5 purification

Tn5 was purified as previously described [191]. E. coli cells (NEB C3013) harboring pTBX1-Tn5 were grown in terrific broth to an OD of 0.65 before addition of IPTG at 0.25 mM. Tn5 expression was induced at 23°C for 16-20 hours before harvesting by centrifugation and storage at -80°C until purification. 20 g of thawed E. coli pellet was lysed in 200 mL HEGX buffer (20 mM HEPES-KOH ph 7.2, 800 mM NaCl, 1 mM EDTA, 0.2% Triton, 10% glycerol) with cOmplete protease inhibitor (Roche) and 10 uL of benzonase (Thermo-Fisher Scientific). Cells were lysed using a LM20 microfluidizer device (Microfluidics) and cleared by centrifugation at 9000 x g for 30 min. 5.25 mL of 10% PEI (pH 7) was added dropwise to a stirring lysate solution to remove E. coli DNA and the resulting precipitation was removed by centrifugation for 10 min. Cleared supernatant was added to 30 mL of washed and equilibrated chitin resin (NEB), mixed by end- over-end at 4°C for 30 min. Resin was washed by gravity flow with 1L HEGX buffer. To elute Tn5, 75 mL HEGX buffer with 100 mM DTT was added to column, 30 mL drawn through the resin before sealing the column and storing at 4°C for 48h to allow for intein cleavage and elution of free Tn5. Eluted Tn5 was dialyzed into 2x Tn5 dialysis buffer (100 HEPES, 200 NaCl, 2 EDTA, 0.2 Triton, 20% glycerol), with two exchanges of 1L of buffer. The final solution was concentrated to 50 mg/mL as determined by A280 absorbance (A280 = 1 =0.616 mg/mL =11.56 mM) and flash frozen in liquid nitrogen before storage at -80°C.

Loading Tn5 transposase

Adaptor sequences (**Table S4)** were annealed by resuspending the strands at 50 uM each in 10 mM Tris-HCl, pH 8.0, 50 mM NaCl, followed by a thermal ramp from 65°C to 25°C over 1 hr. Annealed adaptors were mixed 1:1 with glycerol and stored at -20°C until Tn5 loading. Tn5 transposase was loaded by combining the two adaptors, Tn5 dilution buffer (50 mM Tris-HCl pH 7.5, 0.1 mM EDTA, 100 mM NaCl, 0.1% NP-40, 1 mM DTT, 50%

glycerol) and 14.8 μM Tn5 transposase at a ratio of 1:1:1:1 and incubating for 30 min at RT. Loaded Tn5 was stored at -20°C until library preparation. We note that while unloaded Tn5 transposase was provided to us as a gift from a colleague who produced and purified it in-house, it is commercially available from Lucigen as Cat # TNP92110.

Hairpin snap-cooling

Hairpins (**Table S4**) were resuspended at 1 μM total hairpin concentration in 4X saline sodium citrate buffer (SSC), and then snap-cooled by heating at 95°C for 90 sec and then incubating on ice for 30 min.

Library preparation

Cells and gel-embedded embryos were treated with 0.1N HCl for 5 min at room temperature (RT) and washed 3x with 1X PBS. For adaptor transposition, cells were incubated for 1 hr (or overnight, the cells shown in **Fig. 4-3**) and embryos were incubated overnight at 37°C with loaded Tn5 diluted 1:20 in 0.3X PBS, 10 mM Tris pH 8.5, 5 mM $MgCl_2$. We estimate that our in-house purified Tn5 has 2- to 4-fold the activity of Tn5 available through Lucigen as Cat # TNP92110; titration or adjustment of incubation time for transposition may be needed to determine the optimal Tn5 concentration. Samples were then rinsed once and washed 3 times for 10 min at 37°C with 50 mM EDTA, 0.01% SDS in 1X PBS, and then washed twice with 1X PBS. Hairpins were hybridized to the adaptors by incubating samples with 250 nM annealed R2 hairpin, 62.5 nM each of annealed R1 hairpins A-D (cells) or 15.625 nM each of annealed R1 hairpins 1-16 (embryos) in 4X SSC for 1.5 hr (cells) or 2 hr (embryos) at 37°C. Samples were then washed 3 times for 5 min in PBS. Gap-fill and ligation was performed by incubating samples with 0.5 U/μL (cells) or 0.1 U/μL (embryos, cells shown in **Fig. 4-3**) Ampligase Thermostable DNA Ligase (Lucigen A32750), 0.2 U/μL (cells) or 0.04 U/μL (embryos, shown in **Fig. 4-3**) Phusion High-Fidelity DNA Polymerase (New England Biolabs M0530), 50 μM dNTPs, 50 mM KCl, 20% formamide in 1x Ampligase Reaction Buffer (Lucigen) for 30 min at 37°C followed by 15 min at 45°C. (Following our first experiment, we found that the lower enzyme concentrations performed equally well as the higher concentrations.) Samples were then washed 3 times for 5 minutes in PBS. Rolling circle amplification (RCA) primers (**Table S4**) were hybridized by incubating the cells with 0.5 μM RCA primer in 2x SSC, 30% formamide for 2.5 hr (cells) or 3 hr (embryos) at 37°C,

followed by washing once for 30 minutes (cells) or twice for 15 minutes (embryos) at 37°C with 2x SSC, 30% formamide. Samples were then washed 3 times with PBS. RCA was performed by incubating cells overnight with 1 U/μL Phi29 DNA Polymerase in 1X Phi29 DNA Polymerase Reaction Buffer (New England Biolabs M0269), and incubating embryos and the cells shown in **Fig. 4-3** overnight with EquiPhi29 DNA Polymerase (Thermo Scientific) in 1X EquiPhi29 DNA Polymerase Reaction Buffer with 1 mM DTT. RCA reactions for all samples included 250 μM dNTPs and 50 μM aminoallyl dUTP. (Partway through the study, we found that amplicon yields were higher when using EquiPhi29 compared to standard Phi29.) Samples were then rinsed and washed 3 times for 5 minutes with PBS. The samples were then treated for 1 hr at RT with 18 mg/mL BS(PEG)9 (Thermo Scientific 21582) in 10% DMSO, 1X PBS, to cross-link the amplicons. For visualization, an oligonucleotide (**Table 4.2**) modified with Cy3 was hybridized to the amplicons by incubating the cells with 100 nM oligo in 10% formamide, 4X SSC. Fluorescent Tetraspeck 0.2 μm beads were diluted 1:100 in PBS and applied to cells for 2 min before aspirating and washing with PBS.

4.6.2.2 *In situ* sequencing and immunostaining

Automated *in situ* sequencing in PGP1f cells

For cultured cells, we performed automated *in situ* sequencing by integrating automated fluidics, controlled by MATLAB, with a spinning disk confocal microscope (see *Imaging*), controlled by NIS-Elements AR software. The fluidics system was composed of a modular valve positioner with HVXM 8-5 valve (Hamilton 36766), PTFE laboratory tubing (Finemech S1810-12), an FCS2 flow cell (Bioptechs 060319-2), and a peristaltic pump (Rainin RP-1). A National Instruments Data Acquisition (NI-DAQ, USB-6008) card was used to connect the pump and the valve positioner to a computer. Sequencing reagents that required cooling (ligation mix, CIP solution, cleave 2, imaging buffer) were stored in a Mini Dry Bath (Fisher Scientific) set to 4°C.

To prepare the sample for sequencing, the silicone gasket was removed from the 40 mm round coverslip containing the cells and library, and the coverslip was placed inside the flow cell and connected to the fluidics. The visualization oligo was removed from the amplicons by continuously flowing strip solution (80% formamide in H_2O) over the sample at 500 μL/min

for 10 min, and the sample was washed by continuously flowing instrument buffer (SOLiD Buffer F, 1:10 diluted) over the sample at 1000 μL/min for 5 min. All automated washing in the protocol going forward was performed in this manner.

In situ sequencing was performed using SOLiD chemistry as described previously [192], with modification. (While we purchased the SOLiD reagents commercially, we note that they are no longer being sold as a kit. However, equivalent reagents can be synthesized by Integrated DNA Technologies and equivalent buffers can be prepared as simple formulations [193] (**Note S1**). Four rounds of *in situ* sequencing were performed using each of five primers (**Table S4**), for a total of 20 rounds. All steps were performed using the automated fluidics and imaging setup described above, at RT. First, the sample was treated for 10 min with Quick Calf Intestinal Alkaline Phosphatase (Quick CIP) (New England Biolabs, M0508) diluted 1:20 in 1x CutSmart Buffer (New England Biolabs, B7204), and then 70% of the dead volume of CIP solution in the flow cell and fluidic lines was pumped back into the stock tube, to save reagents. Next, the sample was incubated for 10 min with a mixture of N-0 sequencing primers A-D (**Table 4.1)** at 625 nM each in 5x SASC (0.75 M sodium acetate, 75 mM tri-sodium citrate, pH 7.5). Four rounds of sequencing were subsequently performed. Each round of sequencing was performed as follows, except on the fourth round, where steps 3 and 4 were excluded: 1) A ligation reaction was performed by incubating with ligation mix (12 U/μL Rapid T4 DNA Ligase (Enzymatics, L6030-HC-L), 1:40 dilution of SOLiD sequencing oligos (SOLiD, 4475669), 1x T4 DNA Ligase Buffer (Enzymatics, B603)) for 10 min. After the ligation reaction, 70% of the dead volume of ligation mix in the flow cell and fluidic lines was pumped back into the stock tube, to save reagents. The sample was washed. 2) Imaging buffer (SOLiD Buffer A) (SOLiD, 4463024) was pumped into the flow cell, and then the sample was imaged (see *Imaging*) and washed. Photobleaching was performed after imaging by illuminating with the 594 nm laser line for 25 seconds. 3) A dephosphorylation reaction was performed by treating the sample with Quick CIP, as described above. The sample was washed with buffer F. 4) A cleave reaction was performed by treating the sample first with SOLiD buffer C (SOLiD, 4458932) for 6 min and then with SOLiD buffer B (SOLiD, 4463021) for 6 min. The sample was washed with instrument buffer.

The sequencing primers were then stripped by continuously flowing strip solution (80% formamide in H_2O) over the sample at 500 μL/min for 10 min, and a mixture of N-1 primers A-D (recessed by one base at the 5' end, as compared to the N-0 primers A-D) were hybridized to the sample. Sequencing, stripping, and primer hybridization were repeated using N-2, N-3 and N-4 primers A-D. For the N-3 and N-4 primers, the first sequenced base was not imaged, because it is determined by the sequence of the primer binding site and provides no additional information on top of the first sequenced base of primer N-2, which distinguishes between primer binding sites A-D.

The sample was then stained with 100 nM Cy3-modified visualization oligo in 10% formamide, 4X SSC, for 45 minutes at RT, and with 1 μg/mL DAPI in 1X PBS for 3 minutes at RT, and then imaged for use in downstream image registration and segmentation.

For stratified sequencing in **Fig. 4-3**, only one sequencing primer was used at a time, and the sequencing protocol was performed by hand.

***In situ* sequencing in embryos**

For gel-embedded embryos, sequencing reactions were performed manually at RT. Before sequencing, the embryos were stained with Cy3-modified visualization oligo and DAPI as described above for cultured cells and imaged for use in downstream image registration and segmentation. The visualization oligo was stripped by rinsing and washing 2 times for 5 minutes in strip solution, and the embryos were rinsed and washed 4 times for 5 minutes in instrument buffer. The embryos were then treated with CIP solution for 45 min, rinsed and washed 5 times for 5 minutes in instrument buffer.

In situ sequencing was performed on the embryos as described for cultured cells with the following modifications: i) between each step, washes consisted of a rinse followed by 3 washes for 5 minutes each in instrument buffer, except following incubation with ligation mix, for which 3 3 washes for 10 minutes each were performed, and except for CIP solution, for which 5 washes for 5 minutes each were performed to thoroughly remove chloride ions that can cause precipitation with the silver ions in the SOLiD buffer C solution; ii) For the five primer hybridizations, a pool of primers 1-16 [N-0, N-1, N-2, N-3, or N-4] was used, at a concentration of 500 nM each and incubated for 45 minutes. iii) ligation mix used a 1:400 dilution of SOLiD sequencing oligos and included 1 mg/mL bovine serum albumin

(Thermo Scientific 15561020); iv) the ligation reaction was incubated for 90 minutes; v) the incubation with SOLiD buffer C was incubated twice for 5 minutes followed by incubation with SOLiD buffer B twice for 5 minutes; vi) the first base of sequencing was collected for primer N-3, which provided the information to distinguish primer binding sites 1-16; vii) GLOX imaging buffer with Trolox was used (see below); viii) imaging was performed using different software and a different microscope (see *Imaging*, below); ix) primer stripping was performed by incubating 2 times for 5 minutes in strip solution.

In order to image thick samples (i.e early embryos), we used a different imaging buffer with improved antifade properties. The imaging buffer used for *in situ* sequencing of embryos was prepared as follows: i) a GLOX stock solution was prepared by dissolving 70 mg of glucose oxidase (Sigma G2133) in 1 ml of 200 mM Tris, pH 8.0, 50 mM NaCl, gently mixing with 250 μL of 10-60 mg/mL catalase solution (Sigma C100), centrifuging at maximum speed for 1 minute, taking the supernatant, and storing at 4°C for up to 2 days; ii) immediately before imaging, GLOX stock solution was diluted 1:50 in 100 mM Tris pH 8.0, 25 mM NaCl, 2 mM Trolox (Sigma 238813), 10% glucose. The well of the glass bottom plate to be imaged was completely filled with GLOX imaging buffer and sealed with an adhesive polypropylene film (VWR 60941-070) to prevent contact with oxygenated air.

Immunostaining of embryos

Following *in situ* sequencing, embryos were immunostained. Lamin-B1 was stained with mouse Lamin B-1 Antibody (B10) (Santa Cruz; cat #sc-374015). CENP-A was stained with rabbit CENP-A (C51A7) mAb (Cell Signaling Technology cat #2048). anti-Lamin-B1 and anti-CENP-A were jointly diluted 1:1000 and 1:100 respectively in 3% BSA / 1x PBS. Embryos were stained overnight at 4C. After staining, samples were washed 3x in 1x PBS and detected with goat anti-mouse Alexa Fluor 488 (Thermo Fisher A32723) and goat anti-rabbit Alexa Fluor 647 (Thermo Fisher A32733). Secondaries were jointly diluted 1:200 in 1x PBS and the sample was stained for 1 hour. After staining, samples were washed 3x in 1x PBS and imaged.

Imaging

Imaging for cultured cells was performed using a Yokogawa CSU-W1 confocal spinning disk with Borealis modification coupled to a Nikon Ti-E inverted microscope with a Zyla

4.2 PLUS sCMOS camera, controlled by NIS-Elements AR software. The lasers, power, and emission filters used to image the fluorophores were: 100 mW solid state 405 nm smart diode laser at 50% power with 450/50 filter for DAPI, 150 mW solid state OPSL 488 nm laser at 70% power with 525/50 filter for FITC, 100 mW solid state OPSL 560 nm laser at 100% power with 582/15 filter for Cy3 and Alexa 546, 100 mW solid state DPSS 594 nm laser at 100% power with 624/40 filter for Texas Red, and 110 mW solid state OPSL 642 nm laser at 70% power with 685/40 filter for Cy5. A 1.40 NA 60x Plan Apochromat Lambda oil immersion objective lens (Nikon) with 0.3 μm step size and 200 ms exposure time were used for all images.

Imaging for embryos was performed using an Andor CR-DFLY-201-40 confocal spinning disk coupled to a Nikon Ti-E inverted microscope with a Zyla 4.2 PLUS sCMOS camera, controlled by Andor Fusion 2.0 software. The lasers and emission filters used to image the fluorophores were: 100 mW solid state 405 nm laser at with 450/50 filter for DAPI, 150 mW solid state 488 nm with 525/50 filter for FITC/AlexaFluor 488, 150 mW OBIS LS solid state OPSL 561 nm laser with 582/15 filter for Cy3, 100 mW OBIS LS 594 nm OPSS laser with 631/36 filter for Texas Red, 150 mW OBIS LX solid state 637 nm laser with 676/37 filter for Cy5/AlexaFluor 647. The dichroic mirror Andor CR-DFLY-DMQD-01 (405/488/561/640) was used for imaging all fluorophores except Texas Red, for which Andor CR-DFLY-DMQD-04 (405-445/514/594/730) was used. Images were collected using a 1.15 NA CFI Apo Long Working Distance Lambda S 40XC water immersion objective lens (Nikon) with 0.4 μm step size. For imaging the hybridization probe in the Cy3 channel, the power was 100% and exposure was 200 ms. For *in situ* sequencing, the power levels were 50% for FITC, 100% for Cy3, 100% for Texas Red, and 100% for Cy5. The exposure times were adjusted for each ligation number to account for reduction in brightness across successive ligations on the same primer. For FITC, Cy3, Texas Red, and Cy5, exposure times were: 300 ms, 300 ms, 100 ms, and 100 ms for the first ligation; 450 ms, 450 ms, 150 ms, and 150 ms for the second ligation; 675 ms, 675 ms, 225 ms, and 225 ms for the third ligation; 1000 ms, 1000 ms, 340 ms, and 340 ms for the fourth ligation. For DAPI imaging power was 50% and exposure time was 100 ms. For antibody imaging, power and exposure times were 100% and 150 ms for AlexaFluor 488 (lamin B1), 100% and 100 ms for AlexaFluor 647 (CENP-A).

4.6.2.3 *Ex situ* sequencing

Dissociation and PCR amplification of amplicons

Following *in situ* sequencing of cultured cells, the coverslip was removed from the flow cell. We carefully cut the coverslip down to the size of the imaged region (~250 cells) using a diamond scribe, and placed the remaining fragment in a PCR tube. The coverslip was cut to fit the sample in a PCR tube. A conventional Illumina sequencing library was prepared from the sample as follows: we first fully immersed the trimmed coverslip fragment by adding 100 μL PCR mix (50 μL Phusion U Hot Start PCR Master Mix (Thermo F533), 0.5 μL of 100 uM P5 primer, 0.5 μL of 100 μM v2_Ad2.41 primer, 49 μL ultrapure water), and performed a first round of amplification by incubation at 8°C for 10 min, 98°C for 30 s, 10 cycles of [98°C for 10 s, 69°C for 15 s, 72°C for 30 s], 72°C for 5 min, 4°C hold. To avoid overamplification, we then eluted 5 μL for quantification by qPCR, as described in [194], maintaining the rest of the sample at 4 C. We then performed an additional 15 cycles of amplification on the remaining 95 μL of sample as described above. Following PCR, the reaction was eluted from the PCR tube, column purified (DNA Clean and Concentrator-5, Zymo) using 475 μL DNA binding buffer (i.e. 1:5) and eluted into 20 μL water.

Following *in situ* sequencing of polyacrylamide-embedded embryos, the gels were washed with ultra pure water, scored with a razor, scraped off of the glass-bottom plate in pieces using a disposable polypropylene spatula, and transferred using a paintbrush (Blick Art Supplies, 06170-7030) to PCR tubes containing 34 ul of PCR mix each. PCR mix was prepared as follows: 20 μL Phusion U Hot Start PCR Master Mix, 1 μL 20 μM P5 primer, 2 μL 10 μM i7 indexed Nextera primer (Illumina FC-131-1002), 11 μL ultra pure H_2O. The gel fragments were assumed to contain 6 μL of ultra pure water. The tubes were flicked several times, then incubated on ice 3 times for 10 minutes with flicking between each incubation. The tubes were then pre-amplified by thermocycling according to the following program: 98°C for 30 s, 5 cycles of [98°C for 45 s, 72°C for 80 s], 4°C hold. The tubes were then flicked and incubated at 4°C overnight to allow amplicons to diffuse out of the gel. The tubes were flicked again. 1.5 μL of the pre-amplified libraries was diluted 1:10 in PCR mix with 1x SYBR Green followed by quantification by qPCR. Half the remaining volume of the

pre-amplified libraries was then diluted 1:10 in PCR mix and thermocycled according to the following program: for cycles for each diluted library (ranging from 18-24 cycles): 98°C for 30 s, [# qPCR cycles required to reach $\frac{1}{2}$-maximum + 1] cycles of [98°C for 10 s, 72°C for 40 s], 72°C for 5 min, 4°C hold. The number of cycles used in the latter PCR reaction ranged from 18-24. Dilution of the pre-amplified libraries was to dilute putative carry-over from the in situ sequencing, which dramatically improved yield.

***Ex situ* Illumina sequencing**

PCR-amplified libraries from cultured cells were sequenced using one lane of a HiSeq 2500. 84 bp paired-end reads were collected with a 25 bp i5 index read. Data was analyzed using the Broad Picard Pipeline, which includes de-multiplexing and data aggregation. PCR-amplified libraries from embryos were sequenced on NovaSeq 6000. 143 bp paired-end reads were collected with a 21 bp i5 index read. Sequencing data was demultiplexed using bcl2fastq2 (v2.19.1). E*x situ* Illumina sequencing accounted for the majority of the cost for IGS.

***Ex situ* sequence alignment and processing**

Ex situ sequenced UMIs were appended to the headers of associated paired-end sequencing reads. Reads were then trimmed for sequencing adapters using a custom python script and aligned to hg38 and mm10 using bowtie2 [195] with --very-sensitive and -k 5 parameters. The resulting BAM files were sorted by genomic coordinates, UMIs were moved from the header to a new read group tag, and optical duplicates were removed using Picard MarkDuplicates (*http://broadinstitute.github.io/picard/*)). 76,879,813 reads were sequenced from PGP1f cells with a 96.11% alignment rate to hg38 and a 90.5% duplication rate. 986,237,536 reads were sequenced from embryos with a 97.94% alignment rate to mm10 and a 99.7% duplication rate.

For the PGP1f data, the filtered BAM file was split into three files -- one for uniquely-aligning reads, a second for multi-mapped reads, and a third for unmapped reads. UMIs for the uniquely-aligned and multi-mapped reads were then used to group PCR duplicates together using UMI-tools group with parameter --edit-distance-threshold 2 to facilitate UMI error correction [196]. For groups in which the inferred true UMI was ambiguous, the highest quality base was chosen based on quality scores from original UMI FASTQ file instead of randomly, as implemented in UMI-tools. Multi-mapped and unmapped reads were collapsed

into a single entry per UMI to be added to a comprehensive list of observed UMIs. This list was then filtered for occurrences of index swapping based on the frequency of PCR duplicates for each unique UMI-genomic location combination. Lastly, the UMIs were converted to colorspace sequences using a di-base encoding table [197].

4.6.2.4 Image processing and UMI matching

Field of view image processing

All image processing and UMI matching steps were implemented in custom MATLAB scripts available from the GitHub repository. For all fields of view, 3D image stacks for each sequencing cycle and DAPI stains were registered to cycle 1 using normalized cross-correlation to correct for shifts that may have occurred between imaging rounds. For the PGP1f data, the 3D image stacks for each channel were also corrected for spectral and physical drift by calculating a 3D rigid transformation based on the positions of fluorescent Tetraspeck beads, which were detected using a 3D peak finder. Bounds for each nucleus were defined by performing threshold-based segmentation on the DAPI image stack, and nuclei located at the edge of the field of view that were not fully imaged in every sequencing cycle were excluded from further analysis. In some ($<$10%) of embryos, bounds were manually added for nuclei that were not automatically detected.

Nucleus image processing

A five-dimensional (x by y by z by channel by cycle) image stack was created for each nucleus by cropping fields of view based on nuclei segmentation bounds. Next, stacks were deconvolved using a high pass Gaussian filter to improve the resolution of densely-packed amplicons in the nucleus. The images in each stack were registered to each other along the cycle dimension using an iterative approach. For embryos, each channel from sequencing cycle 1 was independently registered to the visualization oligo image via a 3D affine transformation. For all data, images from a single cycle were collapsed across the channel dimension using a maximum intensity projection, and registered to the collapsed cycle 1. The resulting transformation was then applied to the uncollapsed images from all four channels separately, which were then re-registered independently to the collapsed cycle 1. This process was repeated for all cycles. In the PGP1f data, the rate of signal phasing between cycles was

calculated by identifying pixels that lose most, but not all of their signal in subsequent cycles. The phasing rate was then calculated from these pixels and subtracted stack-wide to correct any residual fluorescence resulting from sequencing inefficiencies. Lastly for all data, images from each cycle were normalized by applying quantile normalization such that the total fluorescence values from each channel were equal (**Fig. 4-5**).

Amplicon identification and size quantification

Amplicon centers were identified by applying a 3D peak finder to the normalized image stacks, with peaks under a percentile-based threshold being removed from downstream analysis. Peaks from separate images with identical 3D coordinates were collapsed into a single entry. To quantify the distribution of amplicon sizes in PGP1f, a Gaussian fit was simultaneously performed on every putative amplicon within a single nucleus (**Fig. 4-4**). The majority of amplicons had diameters between 400-500 nm, and in almost all cases, the identified peak was confirmed to be at the center of the amplicon. Each amplicon was thus localized to a single 0.108 by 0.108 by 0.3 or 0.4 μm voxel (0.3 in PGP1, 0.4 in embryos).

***In situ* UMI processing**

For each amplicon, a region of interest was defined by selecting nearby pixels with high correlation to the peak over all images. Each region was then quantified over all channels and cycles by summing the fluorescence values of all pixels. The resulting two-dimensional matrix (channels by cycles) was normalized such that the sum of squares for each column = 1. Since a large fraction of amplicons fall in densely-packed subcellular volumes, regions were iteratively refined to maximize the purity score, which is calculated for a set of pixels by taking the negative log transform of the maximum fluorescence values across channels and multiplying the value from each cycle together i.e. -log10(product(max(matrix,1),2)). Following this iterative refinement, each amplicon was associated with a final colorspace probability matrix representing the region with the highest signal purity (**Fig. 4-6**).

Generating spatially-resolved reads via UMI matching

UMI matching was facilitated by a set of colorspace probability matrices derived from the spatially-resolved amplicons in the *in situ* images, and a list of observed UMIs and associated genomic reads from the *ex situ* sequencing. A consensus colorspace sequence was generated from each probability matrix by taking the channel with the maximum probability from

each cycle. Each consensus sequence was then compared against all observed sequences using Hamming distance. All sequences with Hamming distances less than 4 were saved as potential matches. If there were no observed sequences with Hamming distance less than 4, the threshold was incremented by 1 until one or more sequences were found. For each potential matching sequence, a match score was calculated by tracing the path of the observed sequence though the colorspace probability matrix. Each value along the path was multiplied together and the final product was negative log transformed. Lower match scores indicate a more probable match, due to the negative log transformation. The potential matching sequence with the lowest match score was associated with each amplicon, generating spatially-resolved reads (**Fig. 4-6**). For the embryos, each nucleus was only allowed to contain matches from a single sequencing well. A consensus well was called for each nucleus, and any matched reads from a different well were excluded.

UMI match filtering

Spatially-resolved reads were filtered based on both purity score, a measure of how clearly an amplicon could be resolved in the images, and match score, a measure of how well the observed *ex situ* UMIs match the colorspace probability matrix. Reads with lower purity scores were allowed to have higher match scores, analogous to a continuous Hamming distance threshold. In order to calculate the rate of spatially-resolved amplicons that successfully match an *ex situ* genomic read, a set of high quality reads was defined by taking all reads with a purity score less than 1.5 in PGP1f and 0.75 in embryos. Matching rate was then calculated by dividing the number of high quality reads passing filter over the number of toal high quality reads (**Fig. 4-7**).

UMI matching false discovery rate

False discovery rate (FDR) was calculated independently for each putative amplicon. For each corresponding colorspace probability matrix, a $4^{18/19}$ tree representing all possible 18/19-base UMIs was generated (18 in PGP1f, 19 in embryos), with branches over the match threshold being pruned to save computational resources. The total number of possible UMIs that pass the match filter for each amplicon was then divided by $4^{18/19}$ and multiplied by the number of observed *ex situ* UMIs per sequencing well to yield a per-amplicon FDR.

Registration of immunofluorescence images

For the embryo data, immunofluorescence (IF) images of lamin B1 and CENP-A were collected along with an additional DAPI stain after sequencing. To register these images to the *in situ* sequencing images, this additional DAPI stain was registered to the original DAPI stain imaged at the time of sequencing to calculate a 3D affine transformation for each nucleus. This transformation was then applied to the IF images.

Segmentation of nuclear landmarks

For the embryo data, segmentation of nuclear landmarks was performed in one of two ways. The spatial location of the nuclear lamina was segmented from the lamin B1 immunofluorescence images using a percentile-based threshold. Both lamin B1 in the nuclear interior and nuclear lamina invaginations were included in the final segmentation. The spatial positions of nucleolus precursor bodies and centromeres were segmented from the DAPI stain and CENP-A immunofluorescence images respectively by training an object classification workflow for each developmental stage in Ilastik [198]. Distance to these landmarks for each spatially-localized read was calculated by generating pairwise distances to all segmented voxels and finding the minimum spatial distance.

4.6.2.5 Data filtering and quality control

Cell filtering

For PGP1f, 88 cells with fewer than 200 spatially-localized reads and 27 cells with detected aneuploidy were excluded. This resulted in a dataset containing 106 cells with 328 ± 114 reads per cell (mean ± SD; **Table 4.1**), equivalent to an average of one genomic locus per ~18 Mb across each diploid human genome.

For embryos, 4 embryos with fewer than 600 spatially-localized reads were excluded. Embryos were visually inspected after DAPI staining, and 1 mitotic zygote and 1 PN5 zygote were detected and excluded. In one 2-cell embryo that was excluded for having fewer than 600 localized reads, both cells were also found to be mitotic. 1 haploid 2-cell embryo was identified based on visual inspection of data and excluded. Polar bodies were detected by visual inspection and removed. This resulted in a dataset containing 24 zygotes, 40 nuclei from 2-cell embryos, and 49 nuclei from 4-cell embryos, with 3,909 ± 2,116, 2,357 ± 1,063, and 1,074 ± 622 reads per nucleus (mean ± SD) for zygotes, 2-cell, and 4-cell

embryos, equivalent to an average of one locus per 1.3 Mb, 2.1 Mb, and 4.7 Mb across each diploid mouse genome, respectively (**Table S2**). We note that for 3 four-cell embryos, *in situ* sequencing data could not be collected from one of the cells due to z-dependent optical inhomogeneities. We expect that this effect will be mitigated in future experiments using optical clearing techniques [199].

Amplicon density inside and outside of nuclei

In cultured cells, amplicons were detected using peak detection in ImageJ for maximum projection images of *in situ* sequencing libraries stained with a visualization probe. A nuclear mask was produced in ImageJ using DAPI staining of nuclei, with a manually adjusted brightness threshold. Amplicons coinciding with the DAPI mask, or within 8 pixels of the mask, were counted as being associated with nuclei. The 8-pixel padding was used because amplicons at the periphery of the nucleus sometimes fell <1 μm outside the boundary of the DAPI mask. This analysis was performed for every second field of view in the dataset. Mean $\pm$ standard deviation amplicon density was $0.94 \pm 0.06/\mu m^2$ inside nuclei, and $0.006 \pm 0.001/\mu m^2$ outside nuclei.

Estimation of nuclear volume and density of spatially-localized reads

The volume of each nucleus was estimated by taking the number of voxels segmented in the DAPI stain and multiplying by the volume of a single voxel, 0.108 μm * 0.108 μm * 0.3 or 0.4 μm (0.3 in PGP1, 0.4 in early mouse embryos). The read density of each nucleus was calculated by dividing the number of spatially-localized reads by the estimated volume of the nucleus.

Comparison of genomic coverage to whole-genome sequencing

Spatially-localized reads were partitioned into 2.5 Mb (for PGP1f) and 1 Mb (for embryos) genomic bins. The number of reads in each bin was considered to be the raw coverage of each bin. Bins were normalized by calculating the GC content of each bin, and then dividing each by the mean of the 500 bins with the most similar GC content. The resulting values were then divided by the median normalized coverage of autosomal bins and multiplied by 2 to scale the coverage values to copy number. PGP1f whole-genome sequencing data for comparison was downloaded from ENCODE accession ENCFF713HUF (*https://www.encodeproject.org/files/ENCFF713HUF/*) and subsampled to match the size

of the PGP1f IGS data set. Whole-genome sequencing data from mouse blood is currently unpublished, but analogous to whole-genome sequencing from any other mouse cell type. It was subsampled to match the size of the embryo IGS data set. Whole-genome sequencing data was normalized for GC content and scaled to copy number in the same fashion as the IGS data.

Quantification of chromatin accessibility bias

A list of transcription start sites (TSSs) in mm10 was obtained from the UCSC Genome Browser (*https://genome.ucsc.edu/cgi-bin/hgTables*). All spatially-localized reads within 2000 bp of each TSS were identified. The starting positions of each identified read relative to the TSS were calculated and aggregated into bins. Enrichment was calculated by dividing the number of reads in each bin by the average of reads 1800-2000 bp upstream of the mean TSS. The ATAC-seq mouse brain data set for comparison to IGS was obtained from a previous study [135]. The data was subsampled to match the size of the embryo IGS data set, and TSS enrichment was calculated in the same fashion.

4.6.2.6 Data annotation

Spatial clustering of reads in cultured cells

For cultured cells, amplicons were clustered and resolved into homologous chromosome pairs using an approach based on maximum likelihood estimation. Prior to clustering, we first identify and outlier points using DBSCAN, where the search radius parameter is set to 3.5 μm and a minimum of two neighbors is required to identify a core point. Cluster assignments were initially generated for all paired chromosomes using k-medoids (k=2). In many cases this was adequate to resolve the homologs, as they localized to different regions of the nucleus and could be clustered based on spatial coordinates alone; these cases were identified manually by visual inspection and used to infer the homolog assignments for the entire data set as follows. Using the subset of well-separated homologous chromosomes, we constructed the joint probability distribution of genomic and physical pairwise distances. The distribution was smoothed using a gaussian kernel and normalized over the length of the chromosome and the range of observed physical distances. To perform maximum likelihood estimation, we first initialized the cluster assignments using k-medoids. Starting

with the amplicon having the smallest genomic position, we assign each segment along the chromosome (in the preliminary cluster) a probability based on the empirically-determined genomic and physical distance distribution. The likelihood is taken as the product of these probabilities across both clusters. We systematically transfer each amplicon to the other clusters, and accept the new grouping if the likelihood increases. The cluster assignments are finalized when no increase in likelihood results from transferring any amplicon to another cluster. This approach identified clusters corresponding to territories, as well as smaller clusters corresponding to amplicons that did not colocalize with a territory. Amplicons that did not colocalize with a territory were excluded from downstream analyses.

Aneuploid cells were identified by finding the chromosomes that had more than two clusters for autosomes, or more than one territory for allosomes, and visually confirming that these extra clusters corresponded to territories, and not amplicons unassociated with a territory. These cells were excluded from the final data set (see **Cell Filtering**).

Spatial clustering of reads in embryos

For each chromosome, spatial clusters were identified using k-means clustering (k=2) with semi-supervised correction based on visual inspection of neighboring genomic positions. When territories for a given chromosome overlapped such that reads could not be confidently assigned to a cluster, the reads were not given a cluster annotation and were specially annotated as overlapping (**Table 4.3**) (0/475 diploid chromosome pairs in zygotes, 41/779 in 2-cell embryos, 95/872 in 4-cell embryos). For cells that had fewer than 150 total reads (4 cells total, all in the highest z-position of a 4-cell embryo), clusters were not assigned. Clusters were not assigned to Chr Y. During visual inspection, three instances of triploidy were observed for individual chromosomes in 2-cell embryos (embryo 25, cell 1 Chr 10; embryo 25, cell 2, Chr 15; and embryo 44 cell 2 Chr 1).

Following initial clustering, for each chromosome in the embryos, outlier points were detected using a DBSCAN algorithm with a variable search radius of 3.5 multiplied by the mean distance between data points in the nucleus, and the requirement that two neighbor points define a core point. Points that were identified as outliers by the algorithm were excluded from our further analyses.

Calculation of percentage of reads falling outside a chromosome territory

To find the percentage of spatially-localized reads that fell outside of a chromosome territory in PGP1f, the number of reads that were not part of the two largest clusters identified by MLE spatial clustering (or, for allosomes, the single largest cluster) for a chromosome in a cell was divided by the number of reads that were part of the two largest clusters (or, for allosomes, single largest cluster). The resulting percentage was 6.83%.

To calculate an equivalent percentage for embryos , the number of amplicons identified by DBSCAN as outliers in our annotated clusters of chromosomal reads (see **Spatial clustering of reads in embryos**) was divided by the number of amplicons not identified as outliers. The resulting percentage was 6.98%.

Assignment of parent-of-origin to spatially-localized reads

Genotype information (.vcf files) for the B6C3F1 and B6D2F1 mouse strains was obtained from *https://www.sanger.ac.uk/data/mouse-genomes-project/*. All spatially-localized reads overlapping a SNP unique to one of the two strains were identified. If a read contained a non-reference base at a heterozygous SNP position, it was assigned to the corresponding parent-of-origin. However, if the non-reference base was present in less than 90% of a read's PCR duplicates, it was excluded as a putative sequencing error. Any reads containing SNPs from both the maternal and paternal genomes were marked as conflicting.

Assignment of parent-of-origin to territories

For embryos, parent-of-origin was assigned to chromosome territories when possible. For zygotes, parent-of-origin was assigned manually, by visually identifying which pronucleus each territory belonged to and assigning paternal origin to the territories in the larger pronucleus. For 2- and 4-cell embryos, the following procedure was used: 1) Assign each territory to a parent-of-origin based on the majority of parent-of-origin-specific SNPs in the territory, leaving the territory unassigned if it did not contain any SNPs or a majority could not be determined. For cells containing only one territory for the X-chromosome, assign it to maternal origin, and assign all Y chromosomes to paternal origin; 2) For territories left unassigned, if their homolog was assigned, then assign them to the opposite assignment of their homolog; 3) If the parent-of-origin assignments for two homologs are the same, but one of the assignments was based on a 1-0 majority in parent-of-origin SNPs and the other was not, resolve the contradiction by changing the homolog with the 1-0 majority; 4) If a contra-

diction in parent-of-origin assignment between homologs could not be resolved, drop both assignments. SNPs in outlier reads were not used in the assignments. Chromosomes that were aneuploid or whose homologs spatially overlapped in a given cell were excluded from parent-of-origin assignments. This procedure led to parent-of-origin assignments for 75.6% of territories in 2-cell embryos and 43.7% of territories in 4-cell embryos. To validate this method of assigning parent-of-origin, we predicted parent-of-origin for zygotic chromosome territories and compared the assignments with ground-truth manual assignments. We found that 79.6% of zygotic chromosome territories could be assigned by this method, of which 97.1% agreed with manual assignments.

Relative radial distance from nuclear center

For PGP1f, to compare the radial positions of chromosomes, we first constructed a 2D convex hull for each cell using the 2D projection of each read within the cell. We then found the relative radial position of each read in the cell. To do this, we defined $\mathbf{r_0}$ as the vector connecting the centroid of the hull and a given read, and then, defined a positive constant c as the minimum scaling factor required for $\mathbf{r_0}$ to intersect a facet of the convex hull. We minimized this constant by solving the equation of the plane for the line coincident with $\mathbf{r_0}$ and each facet of the convex hull. $1/c$ is thus the relative radial distance of the read to the nuclear center. The radial position for each chromosome was defined as the mean radial position of all reads mapping to that chromosome, aggregated across all cells. 95% CI was determined by bootstrapping.

Repetitive DNA element annotation

For PGP1f, both uniquely-mapping and multi-mapping spatially-resolved reads were annotated for overlap with repetitive DNA elements. For uniquely-mapping reads, the genomic positions of each paired-end read were looked up from the BAM file and examined for overlap with the RepeatMasker database [146] using bedtools [200]. For multi-mapping reads, this process was performed for all potential genomic alignments, and the resulting annotations were collapsed into a single entry.

4.6.2.7 PGP1f analysis

Genome-wide mean distance map:

To construct a genome-wide distance map, we first constructed a square matrix with a size equal to the linear hg38 human genome binned at 10 Mb. Then, for each diploid-resolved chromosome territory, we computed the pairwise distance between each read in the territory and every other read in its home cell *except* for the reads associated with its homologous pair in that cell; we then assigned each pairwise distance to its corresponding bin in the matrix. Finally, we flattened the matrix by computing the mean of each bin.

Comparison of repetitive element frequency

All repetitive element (rep_name, rep_family, and rep_class) annotations from the RepeatMasker database [146] were used for this analysis. To calculate the frequency of each repetitive element in IGS data, the number of reads corresponding to each element was divided by the total number of spatially-resolved reads. To calculate the approximate frequency of each repetitive element in the reference genome, the number of genomic bases corresponding to each element in the RepeatMasker database was summed and divided by 3 billion.

Radial distribution of repetitive elements

All repetitive element (rep_name, rep_family, and rep_class) annotations from the RepeatMasker database ([146] were used for this analysis. A distribution for each repetitive element was created by aggregating the relative radial distance measurements for that element across all cells. The repetitive element annotations were then shuffled for each cell 500 times, and the same aggregation was performed on each permutation to create a null distribution for each element. The observed and null distributions for each element were reorganized into 100 bins, each representing a ring 0.05-0.1 microns wide from the nuclear center. A z-score was calculated for each bin and each element, with negative values indicating depletion, and positive values indicating enrichment. The z-scores were then converted to log10(p-values) for visualization. The repetitive elements displaying the most radial bias were identified by selecting those with the greatest standard deviation in bin values. The identified elements with the strongest radial bias were ordered by enrichment profile by sorting by the position of bin with the highest value.

Relationship between genomic and spatial distance:

To characterize the relationship between genomic and spatial distances for ensembles

of chromosomes in PGP1, we first computed the pairwise genomic and spatial distances between each read localized within each individual diploid-resolved chromosome territory, separately keeping track of whether each measurement was within an arm or between arms. For visualization, these measurements were binned (typically at 1 or 3 Mb) either together (**Fig. 4-17, 4-18A-B**), or separately (e.g. **Fig. 4-13, 4-18B-C, 4-19, 4-20**), and were plotted as bin means ± SD. Power laws were fit to unbinned pairwise distance measurements by nonlinear least squares, and 95% confidence intervals were estimated from the sample variance. Residuals were calculated from the bin means and the power law fit and tested for autocorrelation using the Ljung-Box test at lag 1. To compare the distributions of intra- and inter-arm distances, bins were filtered such that only genomic distances containing at least 20 of both types of measurement were retained, denoting a range of shared genomic distances. The empirical spatial distance distributions of intra-arm and inter-arm measurements in this range were then compared by KS test.

In embryos, we took a similar approach, with a few differences. First, reads were distinguished by parent-of-origin rather than homolog. Second, ensemble measurements were binned at 100 kb. Third, curves for individual pronuclei were also computed.

4.6.2.8 Developmental transitions analysis

Separation score

Separation scores were calculated for each spatially-localized read by taking its 100 nearest neighbors in 3D space (excluding reads belonging to the same chromosome territory) and dividing the number of neighbors assigned to either the maternal or paternal genome by the total number of neighbors with an assignment, and selecting the higher fraction. For the boxplots shown in **Fig. 4-24C**, a mean separation score for each nucleus was calculated by averaging the separation score of all its spatially-localized reads.

Centromere polarization

Nuclei from the embryo data set were partitioned by developmental stage. To quantify the degree of centromere polarization at each stage, the spatially-localized reads closest to each segmented CENP-A loci were selected. The spatial positions of these reads were averaged to create a weighted center, and then the distance from the weighted center to the center of

the entire nucleus was measured. This process was repeated for each nucleus. To determine if the distribution of distances between the weighted centromere centers and nuclei centers was significant, the distances to the centromere in each nucleus were randomly permuted. The same weighted center analysis was performed, and the significance for each stage was calculated by performing a two-sample K-S test on the real and permuted distributions.

Rabl-like configuration

Each spatially-localized read was assigned a centromere-telomere score between 0 and 1 by dividing its genomic position by the total length of the corresponding chromosome. Rabl scores were calculated for each read by taking its 100 nearest neighbors in 3D space (excluding reads belonging to the same chromosome territory) and averaging the centromere-telomere score of its neighbors. For the plot shown in **Fig. 4-24E**, reads were first partitioned by developmental stage and then split into 100 bins according to their centromere-telomere score (from 0-0.01 to 0.99-1). The "mean neighbor telomere-centromere position" shown on the y-axis was calculated by taking the median Rabl score for all reads falling within a particular bin. The r values reported represent the Pearson correlation between all spatially-localized reads' centromere-telomere scores and Rabl scores.

Relationship between GC content and distance to nuclear landmarks

Spatially-localized reads in the zygote and 2-cell stage were partitioned into non-overlapping haplotype-resolved 1 Mb bins spanning mm10. The GC content of each bin was calculated using the mm10.gc5Base.bw file from the UCSC Genome Browser. The average distance of each bin to the nuclear lamina and to the nucleolus precursor bodies (NPBs) was calculated by iteratively averaging all of the reads in the bin. The distances from reads originating from the same cell were averaged together first, and then all remaining distances were averaged yielding a final value. Correlations between GC content and average distance to nuclear landmarks were performed for each homolog separately and were reported as r^2 so they can be interpreted in the context of variance explained.

4.6.2.9 Single-cell domain analysis

IGS comparison with Hi-C

Allele-resolved Hi-C data for zygotes and 2-cell embryos were obtained from a published

study [136]. allValidPairs.txt.gz files were temporarily split, lifted over from mm9 to mm10 using the UCSC liftOver utility, and re-combined. The re-combined files were then converted to .hic format [201] and loaded into Juicer [202] to calculate compartment eigenvalues for each 1 Mb bin in mm10.

Next, using our IGS data, a lamin proximity score was calculated for each 1 Mb bin in the mouse genome, defined per-bin as the mean-centered probability that a read was closer than 500 nm to a lamin immunostain. The Pearson correlation coefficients of the Hi-C eigenvalues and the lamin proximity scores were determined per chromosome.

Single-cell domain boundary detection

To detect chromatin domains in single cells, first, a mean pairwise distance matrix was constructed at 2.5 Mb resolution for each single chromosome in each single paternal pronucleus. Matrices were then filtered using a 90% coverage threshold. For the remaining high-coverage matrices, gaps were resolved by linear interpolation. Within a sliding window along the diagonal, we calculated all pairwise distances between bins on each side of the window center ('intra-domain') and between bins on opposite sides of the center ('inter-domain'). The mean intra-domain and inter-domain distances were then calculated and a boundary score was assigned at the window center, defined as the difference of the means divided by the sum of the means. SCD boundaries were then defined as local maxima in the resulting boundary score vector, with peak-finding parameters set by visual inspection of boundary calls in representative cells. To avoid edge noise, peaks found within three bins of the start or end of the matrix were not considered. Boundaries and boundary scores were called and calculated for ensemble matrices using the approach described above without further modification.

Scaled domain distance from lamin

To examine the relationship between SCDs and the nuclear lamina, first, domain boundaries were found in high-coverage matrices as described above. Next, the 'observed distance-to-lamin profile' for scaled domains was constructed: for each sequential pair of boundaries, we considered all reads from that single chromosome with a genomic position falling between the boundaries. Using a fixed number of bins (N) for all pairs of boundaries in all chromosomes, each read was binned based on its relative genomic position between the two

boundaries. The distance to the nuclear lamina for that read was then recorded in the corresponding bin. After processing all reads, the rightmost N/2 bins of distances were reflected and concatenated to the leftmost N/2 bins of distances, resulting in a final observed distance-to-lamin distribution spanning a scaled distance of 0 to 0.5 (i.e. 0 to N/2) from SCD boundaries. To calculate the expected distance-to-lamin vector, for each pair of observed boundaries, the position of the boundary-pair was randomly shifted within the chromosome, while keeping the genomic distance between the two boundaries constant (to control for edge effects, shifted boundaries were required to remain within the minimum and maximum matrix boundaries). The read positions and lamin-distances were then evaluated and binned for the shifted boundaries as described above. This approach was used to generate a number of random samples for each chromosome proportional to the genomic size of the chromosome, and the final 'expected distance-to-lamin profile' was constructed as described above. A 95% confidence interval for the median observed-over-expected distance from lamin was then determined per bin by bootstrapping.

4.6.2.10 Global chromosome positioning analysis

Construction of single-cell autosome distance matrices

For each autosome territory, the mean spatial position of all reads in the territory was calculated to find approximate centers for the territories. For homolog-resolved but non-haplotyped analysis (**Fig. 4-31**, **Fig. 4-29A**), a 19 x 19 single-cell autosome distance matrix was constructed by 1) for each pair of chromosomes, finding the distances between the centers of all inter-pair homologs (ie between Chr A homolog 1 and Chr B homolog 1, Chr A homolog 1 and Chr B homolog 2, etc, without regard for parent-of-origin); and 2) taking the mean inter-pair distance. This procedure allowed construction of a distance matrix that accounted for the separate positioning of each chromosome territory while remaining agnostic to haplotype, allowing the use of all chromosome territories regardless of whether haplotype could be assigned. For chromosomes in which the two territories were overlapping, and thus could not be confidently broken into two clusters (see **Spatial clustering of reads in embryos**), reads were was assigned randomly into two clusters, since their pairwise distances with other chromosomes would be very similar due to their spatial co-localization. For the

three cells with instances of triploid chromosomes (see **Spatial clustering of reads in embryos**) (**Fig. 4-22**), those chromosomes were excluded from analysis, resulting in blank entries (NaN) in columns and rows corresponding to the triploid chromosome in the distance matrix for that cell.

For haplotype-resolved analysis (**Fig. 4-29B**), the single-cell autosome distance matrix was constructed by finding the pairwise distance between all autosome territories for which parent-of-origin was assigned, leaving blank entries (NaN) in columns and rows corresponding to autosome territories for which parent-of-origin could not be assigned. For chromosomes in which the two territories were overlapping, and could thus not be confidently assigned to a haplotype based on reads containing informative SNPs, the haplotype was assigned randomly, since their pairwise distances with other chromosomes would be very similar due to their spatial co-localization.

Correlation of single-cell autosome distance matrices

To find the correlation between two single-cell autosome distance matrices, we first unraveled the upper diagonal of each distance matrix, including the diagonal, and then calculated the Pearson correlation coefficient between the two vectors. For haplotype-resolved analysis and cells which contained triploid chromosomes, the single-cell autosome distance matrices had missing entries (see above), so the vectors were modified to contain only entries that were present in the matrices for both cells being compared. Pairs of cells that shared fewer than 3 sets of haplotyped autosomes (6 territories) were discarded from the haplotype-resolved analysis.

Construction of putative clonal lineage trees

For each 4-cell embryo, the most correlated pair of cells was taken to be a pair of putative sister cells. When all four cells in the embryo passed the threshold of 150 reads per cell (see **Spatial clustering of reads in embryos**), the other two cells were taken to be the second pair of putative sister cells, which was also the second-most correlated pair in 5/6 such embryos (**Fig. 4-32**). All other pairs were taken to be putative cousin cells. When only three cells in a 4-cell embryo were available, the pairs of cells that were not the most correlated were taken to be putative cousin cells.

4.6.3 Kit-free synthesis of *in situ* sequencing reagents

We note that while we purchased the SOLiD sequencing reagents commercially, they are no longer being sold as a kit. However, while we were preparing our manuscript, Nguyen et al. [193] reported methods for producing reagents functionally equivalent to SOLiD sequencing reagents independently of Applied Biosystems. We provide this note as guidance to those who wish to produce their own sequencing reagents.

4.6.3.1 SOLiD Oligos

Nguyen et al. recently reported that oligos containing the cleavable backbone linker used in SOLiD sequencing oligos can be obtained from Integrated DNA Technologies (IDT) using internal 3'-thio deoxyinosine, an off-catalog internal oligo modification that can be requested as a custom order, followed by a phosphorothioate modification. The IDT key for the modification is N/i3Thio-dI/N, where N is any nucleotide, and /i3Thio-dI/ is internal 3'-thio deoxyInosine.

Based on this report in Nguyen et al., we recommend the following oligo sequences that are functionally equivalent to SOLiD oligos:

Sequence (modifications using the code of IDT)	**Dye**	**Dibase encoded**
/56-FAM//ideoxyI//ideoxyI//i3Thio-dI/NNNAA	FAM	AA
/56-FAM//ideoxyI//ideoxyI//i3Thio-dI/NNNCC	FAM	CC
/56-FAM//ideoxyI//ideoxyI//i3Thio-dI/NNNGG	FAM	GG
/56-FAM//ideoxyI//ideoxyI//i3Thio-dI/NNNTT	FAM	TT
/5Cy3/ /ideoxyI//ideoxyI//i3Thio-dI/NNNAC	Cy3	AC
/5Cy3/ /ideoxyI//ideoxyI//i3Thio-dI/NNNCA	Cy3	CA
/5Cy3/ /ideoxyI//ideoxyI//i3Thio-dI/NNNGT	Cy3	GT
/5Cy3/ /ideoxyI//ideoxyI//i3Thio-dI/NNNTG	Cy3	TG
/5TexRd-XN//ideoxyI//ideoxyI//i3Thio-dI/NNNAG	TXR	AG
/5TexRd-XN//ideoxyI//ideoxyI//i3Thio-dI/NNNGA	TXR	GA
/5TexRd-XN//ideoxyI//ideoxyI//i3Thio-dI/NNNCT	TXR	CT

Sequence (modifications using the code of IDT)	Dye	Dibase encoded
/5TexRd-XN//ideoxyI//ideoxyI//i3Thio-dI/NNNTC	TXR	TC
/5Cy5/ /ideoxyI//ideoxyI//i3Thio-dI/NNNAT	Cy5	AT
/5Cy5/ /ideoxyI//ideoxyI//i3Thio-dI/NNNTA	Cy5	TA
/5Cy5/ /ideoxyI//ideoxyI//i3Thio-dI/NNNCG	Cy5	CG
/5Cy5/ /ideoxyI//ideoxyI//i3Thio-dI/NNNGC	Cy5	GC

Table 4.1. Oligonucleotide sequences for SOLiD oligos. Sequences are written using the oligonucleotide modification key from Integrated DNA Technologies.

4.6.3.2 Oligo Cleavage Buffers

Nguyen et al reported that 50 mM $AgNO_3$ in H_2O can be used as a replacement for SOLiD Buffer C. Additionally, we have found that the following buffer can be used as a replacement for SOLiD Buffer B:

Replacement for SOLiD Buffer B	
Reagent	**Concentration**
MESNA (2-mercaptoethanolsulfate)	50 mM
tri-sodium citrate	30 mM
sodium acetate	300 mM
pH to 7.5	

Table 4.2. Replacement for SOLiD Buffer B. *This buffer excludes chloride ions, which will precipitate with the silver ions left over from the first cleavage buffer. As such, HCl should not be used for adjusting pH.

4.6.4 Cost, complexity, and throughput

4.6.4.1 Cost

Table 4.3 below provides a cost breakdown estimate for the most costly reagents required to perform IGS at a scale equivalent to our experiment with early mouse embryos. Volumes correspond to library preparation and *in situ* sequencing of a pooled experiment performed using 300 µm total volume. We note that reagent cost is dominated by *ex situ* Illumina sequencing, which could be reduced by sequencing fewer reads. In our sequencing data generated by NovaSeq, we obtained a duplication rate of 99.72%, indicating library saturation and the potential for a reduction in *ex situ* sequencing.

Reagent	**Cost per iteration**	**Iterations**	**Cost**
Tn5 transposase, (Lucigen Cat # TNP92110)*	519.60	1	519.60
Ampligase DNA ligase	3.41	1	3.41
Phusion DNA polymerase	8.56	1	8.56
EquiPhi29 DNA polymerase	65.64	1	65.64
BSPEG(9)	16.09	1	16.09
SOLiD sequencing oligo	1.31	20	26.25
T4 DNA ligase	7.86	20	157.20
Quick CIP	2.58	16	41.28
NovaSeq S2 sequencing, 2x 150 bp	5300.00	1	5300.00
Total cost			**6138.03**

Table 4.3. **IGS cost breakdown.** Cost breakdown for the most costly reagents required to perform IGS using 300 μl total volume for each incubation. *Our Tn5 transposase was a gift from stocks produced in-house by a colleague. The estimated cost given here is an estimate for using Tn5 transposase purchased from Lucigen.

4.6.4.2 Complexity

In addition to the reagents described in **Table ??**, IGS requires a dedicated confocal fluorescence microscope for overnight automated imaging in order to collect many bases of *in situ* sequencing data. Imaging requires setup of automated imaging protocols, which is accessible within imaging software packages (we used NIS-Elements AR and Andor Fusion 2.0 imaging software to automate imaging of PGP1f and early mouse embryos, respectively, as described in **Section 4.6.2**).

For automated sequencing, which we used in the collection of data from PGP1f, an automated fluidics setup is required. As described in **Section 4.6.2**, we performed automated fluidics by using a custom MATLAB script to control a modular valve positioner and peristaltic pump, which were connected to a computer via a National Instruments Data Acquisition card.

4.6.4.3 Throughput

The throughput-limiting step for IGS is imaging time for *in situ* sequencing, which scales linearly with the number of bases sequenced, the number of fields of view, the thickness of the sample, and the imaging exposure time. Using the exposure times, optimized imaging buffer, and z-step size we used for imaging in embryos, we estimate that 49.4 cells, 1.2 embryos, or simply 123.5 z-steps could be collected per hour of total imaging time over 18 bases of imaging. To estimate the times per embryo and per cell, we have assumed the average z-height required to capture IGS data from embryos in our dataset (41.7 μm), cells with 10 μm nuclei, and 10 cells per FOV.

We note that speed could be improved by using brighter, more photostable dyes for the sequencing reagents. This is an immediate possibility using custom synthesis of the sequencing oligos, as described in **Section 4.6.3.**

Chapter 5

Outlook and Conclusion

In this dissertation, we have detailed a set of methods for in situ genomics, with the bulk of our focus directed towards IGS. However, as noted in the previous chapter, IGS is subject to several limitations, e.g. amplicon density is bounded by the nuclear volume, and spatial resolution is diffraction limited. This poses challenges for scaling IGS towards truly genome-scale measurements in single cells. In this concluding chapter, we briefly describe how we will overcome these limitations by combining IGS with expansion microscopy [188], in which samples are physically expanded with isotropic and nanoscale precision. These improvements will enable both super-resolution imaging and amplicon yields many-fold higher than previously demonstrated. We term this variant Expansion *In Situ* Genome Sequencing (ExGS).

How might expansion microscopy be combined with IGS? We recently demonstrated how this might be possible by combining ExM with *in situ* RNA sequencing, which we term Expansion Sequencing or ExSEQ [94]. In ExSEQ, RNA molecules are embedded in a swellable hydrogel. Any proteins present in the sample are then digested to render it structurally homogeneous, permitting isotropic hydrogel expansion with nanoscale precision. An RNA sequencing library can then be constructed within the expanded hydrogel and sequenced in a similar fashion to IGS, including an ex situ sequencing step (**Fig. 5-1**). Thus, if a protocol to expand genomic DNA can be effectively developed, we anticipate progress to be rapid.

At the time of writing of this thesis, we have developed a preliminary ExGS protocol.

A key insight is to perform all library construction steps prior to amplification within an unexpanded sample, and use the RCA primer as a handle for hydrogel embedding. This is crucial, as we found genomic DNA directly anchored with Label-X [92] cannot be amplified by phi29. We speculate the polymerase cannot read through the lesion and/or is sterically hindered by the polymer. We have applied this protocol to cells and early embryos, mirroring our earlier experiments (**Fig. 5-2**), which will permit us to validate higher resolution / yield results against our earlier data.

Overall, the ExGS process is expected to yield several key advantages. First, because amplification is competitive in space, larger volumes enable higher amplicon densities per cell. This scales favorably, as the cube of the linear expansion factor. For genome structure studies, this means the genome can be analyzed at higher genomic resolution (**Fig. 5-3**), allowing crucial biological structures such as TADs [109, 110, 112] and enhancer-promoter interactions [203] to be detectable de novo at modest levels of expansion. Second, because amplification occurs after expansion, super-resolution is achieved during imaging, and spatial resolutions of 100 nm or less are practical. This permits close biological contacts e.g. TAD interactions to be more readily detected [123]. Finally, because the hydrogel is a quasi-in vitro environment that is mostly water, there is substantially less scattering than in conventional biological specimens, permitting deeper imaging depths in intact samples. However, careful validation experiments and/or cross-validation efforts will need to be conducted, to show that nanoscale structure is preserved [204].

Beyond the scalability of ExGS, we anticipate other areas of improvement. An appealing property of optical methods is the ability to simultaneously image multiple different types of biomolecules in the same sample, which we show at proof-of-concept scale with immunostaning in IGS. However, true multi-omic analysis of RNA and protein alongside genome structure will enable new frontiers of analysis, permitting direct structure-function relationships to be observed between genome architecture and gene expression [71]. Additionally, variations on the core IGS/ExGS protocol may include fusions of the Tn5 transposase to other proteins, including transcription factors, for comprehensive in situ mapping of their binding sites in the context of spatial genome architecture [205]. Altogether, we anticipate a bright future for *in situ* genomics.

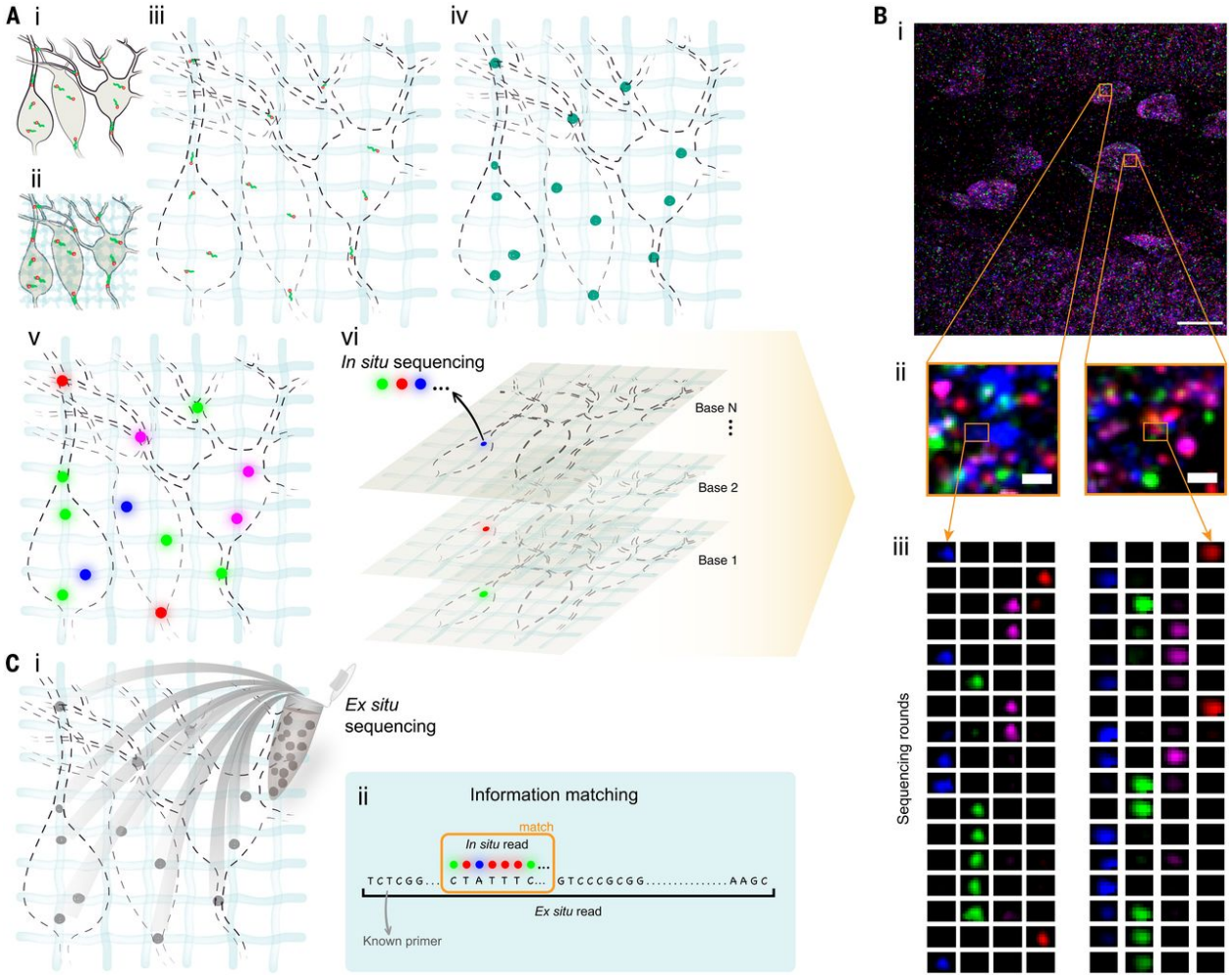

Figure 5-1. **ExSeq concept and workflow**.(A) ExSeq schematic. (i) A specimen is fixed, and RNA molecules (green) are bound by an anchor (orange). (ii) The specimen is embedded in a swellable gel material (light blue, not to scale), mechanically softened, and then expanded with water (iii). RNA molecules are anchored to the gel. (iv) RNA molecules are reverse transcribed and amplified using FISSEQ. (v) In situ sequencing. Colored dots indicate the colors used in the sequencing chemistry. (vi) In each sequencing round, colors (blue, magenta, green, and red) reveal the current base of the cDNA. (B) Example of ExSeq from a 50-μm-thick slice of mouse dentate gyrus. (i) One sequencing round, with two zoomed-in regions (ii) and puncta histories obtained over the course of 17 rounds of in situ sequencing (iii). (C) Ex situ sequencing. (i) After in situ sequencing, cDNA amplicons are eluted from the sample and resequenced ex situ with next-generation sequencing. (ii) In situ reads are matched to their longer ex situ counterparts, focusing on unique matches, augmenting the effective in situ read length. Scale bars in (B) are 17 μm in (i) (in biological, i.e., pre-expansion units used throughout, unless otherwise indicated) and 700 nm in (ii).

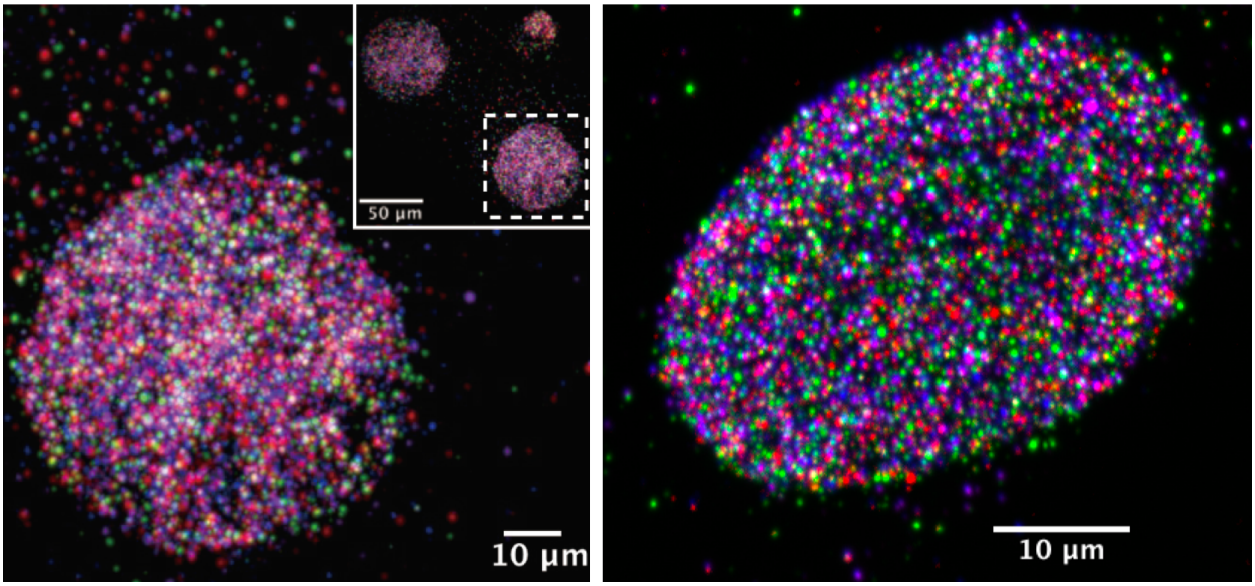

Figure 5-2. **Preliminary ExGS data**. Left: One base of ExGS sequencing in an expanded 2-cell embryo. Right: One base of ExGS sequencing in an expanded PGP1f fibroblast. Both samples expanded approximately 3.5x in linear dimension.

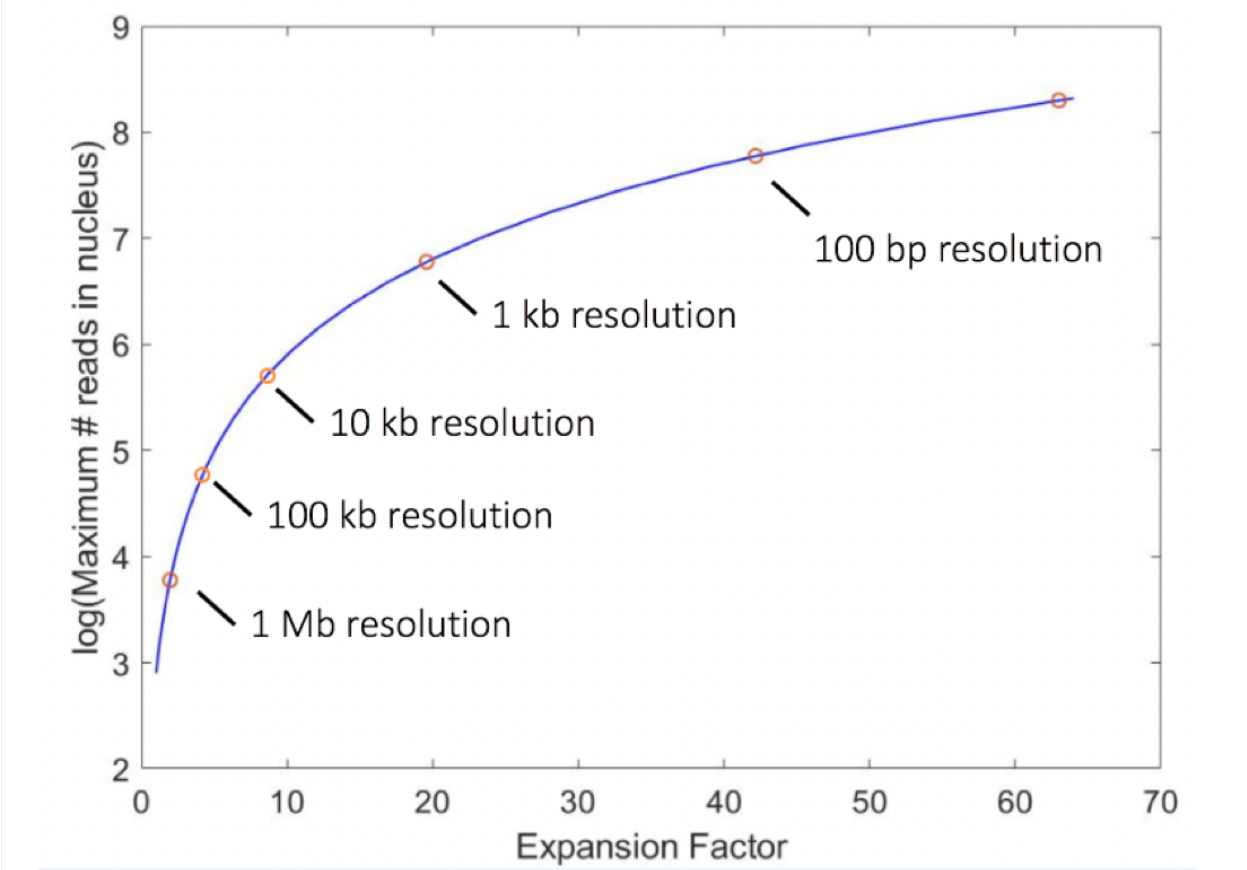

Figure 5-3. **ExGS scaling properties**. Theoretical amplicon yield curve as a function of linear expansion, given a 100 μm^3 nucleus and 0.125 μm^3 amplicons.

Bibliography

[1] Jay Shendure, Shankar Balasubramanian, George M Church, Walter Gilbert, Jane Rogers, Jeffery A Schloss, and Robert H Waterston. DNA sequencing at 40: past, present and future. *Nature*, 550(7676):345–353, October 2017.

[2] Gregor Johann Mendel and British Museum. *Experiments in Plant Hybridisation: Mendel's Original Paper in English Translation*. 1965.

[3] Thomas Hunt Morgan. The physical basis of heredity /, 1919.

[4] Horace Freeland Judson. *The Eighth Day of Creation: Makers of the Revolution in Biology*. January 2004.

[5] Eric S Lander. Initial impact of the sequencing of the human genome, 2011.

[6] Stephen S Hall. Revolution postponed. *Sci. Am.*, 303(4):60–67, October 2010.

[7] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63, January 2009.

[8] Jason D Buenrostro, Beijing Wu, Howard Y Chang, and William J Greenleaf. ATAC-seq: A method for assaying chromatin accessibility Genome-Wide. *Curr. Protoc. Mol. Biol.*, 109:21.29.1–9, January 2015.

[9] GTEx Consortium. Human genomics. the Genotype-Tissue expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, May 2015.

[10] Mathias Uhlén and Emma Lundberg. Webinar | a high-resolution look at the human cell: Introducing the human cell atlas, 2017.

[11] Peter J Thul, Lovisa Åkesson, Mikaela Wiking, Diana Mahdessian, Aikaterini Geladaki, Hammou Ait Blal, Tove Alm, Anna Asplund, Lars Björk, Lisa M Breckels, Anna Bäckström, Frida Danielsson, Linn Fagerberg, Jenny Fall, Laurent Gatto, Christian Gnann, Sophia Hober, Martin Hjelmare, Fredric Johansson, Sunjae Lee, Cecilia Lindskog, Jan Mulder, Claire M Mulvey, Peter Nilsson, Per Oksvold, Johan Rockberg, Rutger Schutten, Jochen M Schwenk, Åsa Sivertsson, Evelina Sjöstedt, Marie Skogs, Charlotte Stadler, Devin P Sullivan, Hanna Tegel, Casper Winsnes, Cheng Zhang, Martin Zwahlen, Adil Mardinoglu, Fredrik Pontén, Kalle von Feilitzen, Kathryn S Lilley, Mathias Uhlén, and Emma Lundberg. A subcellular map of the human proteome. *Science*, 356(6340), May 2017.

[12] Furqan M Fazal, Shuo Han, Pornchai Kaewsapsak, Kevin R Parker, Jin Xu, Alistair N Boettiger, Howard Y Chang, and Alice Y Ting. Atlas of subcellular RNA localization revealed by APEX-seq.

[13] J Dekker. Capturing chromosome conformation, 2002.

[14] Erez Lieberman-Aiden, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A Mirny, Eric S Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, October 2009.

[15] Jeffrey M Levsky and Robert H Singer. Fluorescence in situ hybridization: past, present and future. *J. Cell Sci.*, 116(Pt 14):2833–2838, July 2003.

[16] Wendy A Bickmore. The spatial organization of the human genome. *Annu. Rev. Genomics Hum. Genet.*, 14:67–84, July 2013.

[17] Job Dekker and Leonid Mirny. The 3D genome as moderator of chromosomal communication. *Cell*, 164(6):1110–1121, March 2016.

[18] Peter Hugo Lodewijk Krijger and Wouter de Laat. Regulation of disease-associated gene expression in the 3D genome, 2016.

[19] Adam H Marblestone and Edward S Boyden. Designing tools for assumption-proof brain mapping. *Neuron*, 83(6):1239–1241, September 2014.

[20] Ralf Dahm. Friedrich miescher and the discovery of DNA, 2005.

[21] P A Levene and W A Jacobs. Yeast nucleic acid.

[22] O T Avery, C M MacLeod, and M McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. 1944. *Mol. Med.*, 1(4):344–365, May 1995.

[23] A D Hershey and M Chase. GENETIC RECOMBINATION AND HETEROZYGOSIS IN BACTERIOPHAGE, 1951.

[24] B Magasanik and E Chargaff. Studies on the structure of ribonucleic acids. 1951. *Biochim. Biophys. Acta*, 1000:17–33, 1989.

[25] Rosalind E Franklin and R G Gosling. Molecular configuration in sodium thymonucleate. 1953. *Nature*, 421(6921):400–1; discussion 396, January 2003.

[26] M H F Wilkins, A R Stokes, and H R Wilson. Molecular structure of deoxypentose nucleic acids. *Nature*, 171(4356):738–740, April 1953.

[27] J D Watson and F H C Crick. A structure for deoxyribose nucleic acid, 2017.

[28] G Gamow. Possible relation between deoxyribonucleic acid and protein structures, 1954.

[29] Tony Kouzarides. Chromatin modifications and their function, 2007.

[30] J T Finch, L C Lutter, D Rhodes, R S Brown, B Rushton, M Levitt, and A Klug. Structure of nucleosome core particles of chromatin, 1977.

[31] J Hozier, M Renz, and P Nehls. The chromosome fiber: evidence for an ordered superstructure of nucleosomes. *Chromosoma*, 62(4):301–317, July 1977.

[32] T Jenuwein and C D Allis. Translating the histone code. *Science*, 293(5532):1074–1080, August 2001.

[33] David J Tremethick. Higher-Order structures of chromatin: The elusive 30 nm fiber, 2007.

[34] Viviana I Risca, Sarah K Denny, Aaron F Straight, and William J Greenleaf. Variable chromatin structure revealed by in situ spatially correlated DNA cleavage mapping. *Nature*, 541(7636):237–241, January 2017.

[35] Horng D Ou, Sébastien Phan, Thomas J Deerinck, Andrea Thor, Mark H Ellisman, and Clodagh C O'Shea. ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science*, 357(6349), July 2017.

[36] Theodor Schwann, M J Schleiden, and Henry Smith. Microscopical researches into the accordance in the structure and growth of animals and plants. tr. from the german of dr. th. schwann ... by henry smith ., 1847.

[37] T Cremer and C Cremer. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat. Rev. Genet.*, 2(4):292–301, April 2001.

[38] Thomas Cremer and C Cremer. Rise, fall and resurrection of chromosome territories: a historical perspective. part i. the rise of chromosome territories. *Eur. J. Histochem.*, 50(3):161–176, July 2006.

[39] T Cremer, C Cremer, H Baumann, E K Luedtke, K Sperling, V Teuber, and C Zorn. Rabl's model of the interphase chromosome arrangement tested in chinise hamster cells by premature chromosome condensation and laser-UV-microbeam experiments, 1982.

[40] Joachim Walter, Lothar Schermelleh, Marion Cremer, Satoshi Tashiro, and Thomas Cremer. Chromosome order in HeLa cells changes during mitosis and early g1, but is stably maintained during subsequent interphase stages. *J. Cell Biol.*, 160(5):685–697, March 2003.

[41] Daniel Gerlich, Joël Beaudouin, Bernd Kalbfuss, Nathalie Daigle, Roland Eils, and Jan Ellenberg. Global chromosome positions are transmitted through mitosis in mammalian cells, 2003.

[42] S Boyle. The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells, 2001.

[43] Jenny A Croft, Joanna M Bridger, Shelagh Boyle, Paul Perry, Peter Teague, and Wendy A Bickmore. Differences in the localization and morphology of chromosomes in the human nucleus, 1999.

[44] Andreas Bolzer, Gregor Kreth, Irina Solovei, Daniela Koehler, Kaan Saracoglu, Christine Fauth, Stefan Müller, Roland Eils, Christoph Cremer, Michael R Speicher, and Thomas Cremer. Three-Dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes, 2005.

[45] Luis A Parada, Philip G McQueen, and Tom Misteli. Tissue-specific spatial organization of genomes. *Genome Biol.*, 5(7):R44, June 2004.

[46] Karen J Meaburn, Prabhakar R Gudla, Sameena Khan, Stephen J Lockett, and Tom Misteli. Disease-specific gene repositioning in breast cancer. *J. Cell Biol.*, 187(6):801–812, December 2009.

[47] Hideyuki Tanabe, Stefan Müller, Michaela Neusser, Johann von Hase, Enzo Calcagno, Marion Cremer, Irina Solovei, Christoph Cremer, and Thomas Cremer. Evolutionary conservation of chromosome territory arrangements in cell nuclei from higher primates. *Proc. Natl. Acad. Sci. U. S. A.*, 99(7):4424–4429, April 2002.

[48] Leonid A Mirny. The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res.*, 19(1):37–51, January 2011.

[49] Reza Kalhor, Harianto Tjong, Nimanthi Jayathilaka, Frank Alber, and Lin Chen. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.*, 30(1):90–98, December 2011.

[50] Siyuan Wang, Jun-Han Su, Brian J Beliveau, Bogdan Bintu, Jeffrey R Moffitt, Chao-Ting Wu, and Xiaowei Zhuang. Spatial organization of chromatin domains and compartments in single chromosomes. *Science*, 353(6299):598–602, August 2016.

[51] Martin Falk, Yana Feodorova, Natalia Naumova, Maxim Imakaev, Bryan R Lajoie, Heinrich Leonhardt, Boris Joffe, Job Dekker, Geoffrey Fudenberg, Irina Solovei, and Leonid A Mirny. Heterochromatin drives compartmentalization of inverted and conventional nuclei. *Nature*, 570(7761):395–399, June 2019.

[52] Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, April 2012.

[53] Jan Krefting, Miguel A Andrade-Navarro, and Jonas Ibn-Salem. Evolutionary stability of topologically associating domains is associated with conserved gene regulation. *BMC Biol.*, 16(1):87, August 2018.

[54] Iain Williamson, Lauren Kane, Paul S Devenney, Eve Anderson, Fiona Kilanowski, Robert E Hill, Wendy A Bickmore, and Laura A Lettice. Developmentally regulated shh expression is robust to TAD perturbations.

[55] Suhas S P Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, and Erez Lieberman Aiden. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, December 2014.

[56] Geoffrey Fudenberg, Maxim Imakaev, Carolyn Lu, Anton Goloborodko, Nezar Abdennur, and Leonid A Mirny. Formation of chromosomal domains by loop extrusion. *Cell Rep.*, 15(9):2038–2049, May 2016.

[57] Suhas S P Rao, Su-Chen Huang, Brian Glenn St Hilaire, Jesse M Engreitz, Elizabeth M Perez, Kyong-Rim Kieffer-Kwon, Adrian L Sanborn, Sarah E Johnstone, Gavin D Bascom, Ivan D Bochkov, Xingfan Huang, Muhammad S Shamim, Jaeweon Shin, Douglass Turner, Ziyi Ye, Arina D Omer, James T Robinson, Tamar Schlick, Bradley E Bernstein, Rafael Casellas, Eric S Lander, and Erez Lieberman Aiden. Cohesin loss eliminates all loop domains. *Cell*, 171(2):305–320.e24, October 2017.

[58] J Dekker. Capturing chromosome conformation, 2002.

[59] Job Dekker, Marc A Marti-Renom, and Leonid A Mirny. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.*, 14(6):390–403, June 2013.

[60] Marieke Simonis, Petra Klous, Erik Splinter, Yuri Moshkin, Rob Willemsen, Elzo de Wit, Bas van Steensel, and Wouter de Laat. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4c). *Nat. Genet.*, 38(11):1348–1354, November 2006.

[61] Josée Dostie, Todd A Richmond, Ramy A Arnaout, Rebecca R Selzer, William L Lee, Tracey A Honan, Eric D Rubio, Anton Krumm, Justin Lamb, Chad Nusbaum, Roland D Green, and Job Dekker. Chromosome conformation capture carbon copy (5c): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.*, 16(10):1299–1309, October 2006.

[62] James Fraser, Iain Williamson, Wendy A Bickmore, and Josée Dostie. An overview of genome organization and how we got there: from FISH to Hi-C, 2015.

[63] Takashi Nagano, Yaniv Lubling, Tim J Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D Laue, Amos Tanay, and Peter Fraser. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, October 2013.

[64] Luca Giorgetti and Edith Heard. Closing the loop: 3C versus DNA FISH. *Genome Biol.*, 17(1):215, October 2016.

[65] Geoffrey Fudenberg and Maxim Imakaev. FISH-ing for captured contacts: towards reconciling FISH and 3C. *Nat. Methods*, 14(7):673–678, July 2017.

[66] Iain Williamson, Soizik Berlivet, Ragnhild Eskeland, Shelagh Boyle, Robert S Illingworth, Denis Paquette, Josée Dostie, and Wendy A Bickmore. Spatial genome organization: contrasting views from chromosome conformation capture and fluorescence in situ hybridization. *Genes Dev.*, 28(24):2778–2791, December 2014.

[67] T Cremer, P Lichter, J Borden, D C Ward, and L Manuelidis. Detection of chromosome aberrations in metaphase and interphase tumor cells by in situ hybridization using chromosome-specific library probes. *Hum. Genet.*, 80(3):235–246, November 1988.

[68] Brian J Beliveau, Eric F Joyce, Nicholas Apostolopoulos, Feyza Yilmaz, Chamith Y Fonseka, Ruth B McCole, Yiming Chang, Jin Billy Li, Tharanga Niroshini Senaratne, Benjamin R Williams, Jean-Marie Rouillard, and Chao-Ting Wu. Versatile design and synthesis platform for visualizing genomes with oligopaint FISH probes. *Proc. Natl. Acad. Sci. U. S. A.*, 109(52):21301–21306, December 2012.

[69] Bogdan Bintu, Leslie J Mateo, Jun-Han Su, Nicholas A Sinnott-Armstrong, Mirae Parker, Seon Kinrot, Kei Yamaya, Alistair N Boettiger, and Xiaowei Zhuang. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science*, 362(6413), October 2018.

[70] Guy Nir, Irene Farabella, Cynthia Pérez Estrada, Carl G Ebeling, Brian J Beliveau, Hiroshi M Sasaki, S Dean Lee, Son C Nguyen, Ruth B McCole, Shyamtanu Chattoraj, Jelena Erceg, Jumana AlHaj Abed, Nuno M C Martins, Huy Q Nguyen, Mohammed A Hannan, Sheikh Russell, Neva C Durand, Suhas S P Rao, Jocelyn Y Kishi, Paula Soler-Vila, Michele Di Pierro, José N Onuchic, Steven P Callahan, John M Schreiner, Jeff A Stuckey, Peng Yin, Erez Lieberman Aiden, Marc A Marti-Renom, and C-Ting Wu. Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling. *PLoS Genet.*, 14(12):e1007872, December 2018.

[71] Leslie J Mateo, Sedona E Murphy, Antonina Hafner, Isaac S Cinquini, Carly A Walker, and Alistair N Boettiger. Visualizing DNA folding and RNA in embryos at single-cell resolution. *Nature*, 568(7750):49–54, April 2019.

[72] Alistair N Boettiger, Bogdan Bintu, Jeffrey R Moffitt, Siyuan Wang, Brian J Beliveau, Geoffrey Fudenberg, Maxim Imakaev, Leonid A Mirny, Chao-Ting Wu, and Xiaowei Zhuang. Super-resolution imaging reveals distinct chromatin folding for different epigenetic states. *Nature*, 529(7586):418–422, January 2016.

[73] Brian J Beliveau, Alistair N Boettiger, Maier S Avendaño, Ralf Jungmann, Ruth B McCole, Eric F Joyce, Caroline Kim-Kiselak, Frédéric Bantignies, Chamith Y Fonseka, Jelena Erceg, Mohammed A Hannan, Hien G Hoang, David Colognori, Jeannie T Lee, William M Shih, Peng Yin, Xiaowei Zhuang, and Chao-Ting Wu. Single-molecule super-resolution imaging of chromosomes and in situ haplotype visualization using oligopaint FISH probes. *Nat. Commun.*, 6:7147, May 2015.

[74] Yong-Shu He, H E Yong-Shu, Wen Zhang, and Zhao-Qing Yang. Structural variation in the human genome, 2009.

[75] Brian J Beliveau, Eric F Joyce, Nicholas Apostolopoulos, Feyza Yilmaz, Chamith Y Fonseka, Ruth B McCole, Yiming Chang, Jin Billy Li, Tharanga Niroshini Senaratne, Benjamin R Williams, et al. Versatile design and synthesis platform for visualizing genomes with oligopaint fish probes. *Proceedings of the National Academy of Sciences*, 109(52):21301–21306, 2012.

[76] Huy Q Nguyen, Shyamtanu Chattoraj, David Castillo, Son C Nguyen, Guy Nir, Antonios Lioutas, Elliot A Hershberg, Nuno M C Martins, Paul L Reginato, Mohammed Hannan, Brian J Beliveau, George M Church, Evan R Daugharthy, Marc A Marti-Renom, and C-Ting Wu. 3D mapping and accelerated super-resolution imaging of the human genome using in situ sequencing. *Nat. Methods*, 17(8):822–832, August 2020.

[77] Jun-Han Su, Pu Zheng, Seon S Kinrot, Bogdan Bintu, and Xiaowei Zhuang. Genome-Scale imaging of the 3D organization and transcriptional activity of chromatin. *Cell*, 182(6):1641–1659.e26, September 2020.

[78] Yodai Takei, Jina Yun, Shiwei Zheng, Noah Ollikainen, Nico Pierson, Jonathan White, Sheel Shah, Julian Thomassie, Shengbao Suo, Chee-Huat Linus Eng, et al. Integrated spatial genomics reveals global architecture of single nuclei. *Nature*, 590(7845):344–350, 2021.

[79] Xingqi Chen, Ying Shen, Will Draper, Jason D Buenrostro, Ulrike Litzenburger, Seung Woo Cho, Ansuman T Satpathy, Ava C Carter, Rajarshi P Ghosh, Alexandra East-Seletsky, et al. Atac-see reveals the accessible genome by transposase-mediated imaging and sequencing. *Nature methods*, 13(12):1013–1020, 2016.

[80] Eric S Lander. Initial impact of the sequencing of the human genome. *Nature*, 470(7333):187–197, February 2011.

[81] Jay Shendure, Shankar Balasubramanian, George M Church, Walter Gilbert, Jane Rogers, Jeffery A Schloss, and Robert H Waterston. DNA sequencing at 40: past, present and future. *Nature*, 550(7676):345–353, October 2017.

[82] Karen H Miga, Sergey Koren, Arang Rhie, Mitchell R Vollger, Ariel Gershman, Andrey Bzikadze, Shelise Brooks, Edmund Howe, David Porubsky, Glennis A Logsdon, Valerie A Schneider, Tamara Potapova, Jonathan Wood, William Chow, Joel Armstrong, Jeanne Fredrickson, Evgenia Pak, Kristof Tigyi, Milinn Kremitzki, Christopher Markovic, Valerie Maduro, Amalia Dutra, Gerard G Bouffard, Alexander M Chang, Nancy F Hansen, Amy B Wilfert, Françoise Thibaud-Nissen, Anthony D Schmitt, Jon-Matthew Belton, Siddarth Selvaraj, Megan Y Dennis, Daniela C Soto, Ruta Sahasrabudhe, Gulhan Kaya, Josh Quick, Nicholas J Loman, Nadine Holmes, Matthew Loose, Urvashi Surti, Rosa Ana Risques, Tina A Graves Lindsay, Robert Fulton, Ira Hall, Benedict Paten, Kerstin Howe, Winston Timp, Alice Young, James C Mullikin, Pavel A Pevzner, Jennifer L Gerton, Beth A Sullivan, Evan E Eichler, and Adam M

Phillippy. Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, 585(7823):79–84, July 2020.

[83] A Bensimon, A Simon, A Chiffaudel, V Croquette, F Heslot, and D Bensimon. Alignment and sensitive detection of DNA by a moving interface. *Science*, 265(5181):2096–2098, September 1994.

[84] Andrew C Payne, Michael Andregg, Kent Kemmish, Mark Hamalainen, Charlotte Bowell, Andrew Bleloch, Nathan Klejwa, Wolfgang Lehrach, Ken Schatz, Heather Stark, Adam Marblestone, George Church, Christopher S Own, and William Andregg. Molecular threading: Mechanical extraction, stretching and placement of DNA molecules from a Liquid-Air interface. *PLoS One*, 8(7):e69058, July 2013.

[85] Xavier Michalet, Rosemary Ekong, Françoise Fougerousse, Sophie Rousseaux, Catherine Schurra, Nick Hornigold, Marjon van Slegtenhorst, Jonathan Wolfe, Sue Povey, Jacques S Beckmann, and Aaron Bensimon. Dynamic molecular combing: Stretching the whole human genome for High-Resolution studies. *Science*, 277(5331):1518–1523, September 1997.

[86] Atanas Kaykov, Thibaud Taillefumier, Aaron Bensimon, and Paul Nurse. Molecular combing of single DNA molecules on the 10 megabase scale. *Sci. Rep.*, 6:19636, January 2016.

[87] Melanie Schirmer, Umer Z Ijaz, Rosalinda D'Amore, Neil Hall, William T Sloan, and Christopher Quince. Insight into biases and sequencing errors for amplicon sequencing with the illumina MiSeq platform. *Nucleic Acids Res.*, 43(6):e37, March 2015.

[88] Yong-Shu He, H E Yong-Shu, Wen Zhang, and Zhao-Qing Yang. Structural variation in the human genome, 2009.

[89] Aditya Gupta, Kristy L Kounovsky-Shafer, Prabu Ravindran, and David C Schwartz. Optical mapping and nanocoding approaches to whole-genome analysis, 2016.

[90] Rodolphe Marie, Jonas N Pedersen, Kalim U Mir, Brian Bilenberg, and Anders Kristensen. Concentrating and labeling genomic DNA in a nanofluidic array, 2018.

[91] Miten Jain, Hugh E Olsen, Daniel J Turner, David Stoddart, Kira V Bulazel, Benedict Paten, David Haussler, Huntington F Willard, Mark Akeson, and Karen H Miga. Linear assembly of a human centromere on the Y chromosome. *Nat. Biotechnol.*, 36(4):321–323, March 2018.

[92] Fei Chen, Asmamaw T Wassie, Allison J Cote, Anubhav Sinha, Shahar Alon, Shoh Asano, Evan R Daugharthy, Jae-Byum Chang, Adam Marblestone, George M Church, Arjun Raj, and Edward S Boyden. Nanoscale imaging of RNA with expansion microscopy. *Nat. Methods*, 13(8):679–684, August 2016.

[93] A M Maxam and W Gilbert. Sequencing end-labeled DNA with base-specific chemical cleavages. *Methods Enzymol.*, 65(1):499–560, 1980.

[94] Shahar Alon, Daniel R Goodwin, Anubhav Sinha, Asmamaw T Wassie, Fei Chen, Evan R Daugharthy, Yosuke Bando, Atsushi Kajita, Andrew G Xue, Karl Marrett, Robert Prior, Yi Cui, Andrew C Payne, Chun-Chen Yao, Ho-Jun Suk, Ru Wang, Chih-Chieh Jay Yu, Paul Tillberg, Paul Reginato, Nikita Pak, Songlei Liu, Sukanya Punthambaker, Eswar P R Iyer, Richie E Kohman, Jeremy A Miller, Ed S Lein, Ana Lako, Nicole Cullen, Scott Rodig, Karla Helvie, Daniel L Abravanel, Nikhil Wagle, Bruce E Johnson, Johanna Klughammer, Michal Slyper, Julia Waldman, Judit Jané-Valbuena, Orit Rozenblatt-Rosen, Aviv Regev, IMAXT Consortium, George M Church, Adam H Marblestone, and Edward S Boyden. Expansion sequencing: Spatially precise in situ transcriptomics in intact biological systems. *Science*, 371(6528), January 2021.

[95] Rodolphe Marie, Jonas N Pedersen, Loic Bærlocher, Kamila Koprowska, Marie Pødenphant, Céline Sabatel, Maksim Zalkovskij, Andrej Mironov, Brian Bilenberg, Neil Ashley, Henrik Flyvbjerg, Walter F Bodmer, Anders Kristensen, and Kalim U Mir. Single-molecule DNA-mapping and whole-genome sequencing of individual cells. *Proc. Natl. Acad. Sci. U. S. A.*, 115(44):11192–11197, October 2018.

[96] Nicholas Boyd and Kalim Mir. Sequencing by emergence: Modeling and estimation. March 2021.

[97] Lutz Rösch, Peter John, and Rudolf Reitmeier. Silicon compounds, organic, 2000.

[98] Natsuko Kondo, Akihisa Takahashi, Koji Ono, and Takeo Ohnishi. DNA damage induced by alkylating agents and repair pathways, 2010.

[99] A M Maxam and W Gilbert. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U. S. A.*, 74(2):560–564, February 1977.

[100] Xiaofang Wang, Hyun Jeong Lim, and Ahjeong Son. Characterization of denaturation and renaturation of DNA for DNA hybridization. *Environ. Health Toxicol.*, 29:e2014007, September 2014.

[101] Jae-Byum Chang, Fei Chen, Young-Gyu Yoon, Erica E Jung, Hazen Babcock, Jeong Seuk Kang, Shoh Asano, Ho-Jun Suk, Nikita Pak, Paul W Tillberg, Asmamaw T Wassie, Dawen Cai, and Edward S Boyden. Iterative expansion microscopy. *Nat. Methods*, 14(6):593–599, June 2017.

[102] Andrew C Payne, Zachary D Chiang, Paul L Reginato, Sarah M Mangiameli, Evan M Murray, Chun-Chen Yao, Styliani Markoulaki, Andrew S Earl, Ajay S Labade, Rudolf Jaenisch, George M Church, Edward S Boyden, Jason D Buenrostro, and Fei Chen. In situ genome sequencing resolves DNA sequence and structure in intact biological samples. *Science*, 371(6532), February 2021.

[103] Boyan Bonev and Giacomo Cavalli. Organization and function of the 3D genome, 2016.

[104] M Jordan Rowley and Victor G Corces. Organizational principles of 3D genome architecture. *Nat. Rev. Genet.*, 19(12):789–800, December 2018.

[105] Hui Zheng and Wei Xie. The role of 3D genome organization in development and cell differentiation. *Nat. Rev. Mol. Cell Biol.*, 20(9):535–550, September 2019.

[106] Malte Spielmann, Darío G Lupiáñez, and Stefan Mundlos. Structural variation in the 3D genome, 2018.

[107] Erez Lieberman-Aiden, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A Mirny, Eric S Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, October 2009.

[108] Jeffrey M Levsky and Robert H Singer. Fluorescence in situ hybridization: past, present and future. *J. Cell Sci.*, 116(Pt 14):2833–2838, July 2003.

[109] Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, April 2012.

[110] Tom Sexton, Eitan Yaffe, Ephraim Kenigsberg, Frédéric Bantignies, Benjamin Leblanc, Michael Hoichman, Hugues Parrinello, Amos Tanay, and Giacomo Cavalli. Three-dimensional folding and functional organization principles of the drosophila genome. *Cell*, 148(3):458–472, February 2012.

[111] Suhas S P Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, and Erez Lieberman Aiden. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, December 2014.

[112] Elphège P Nora, Bryan R Lajoie, Edda G Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, Nynke L van Berkum, Johannes Meisig, John Sedat, Joost Gribnau, Emmanuel Barillot, Nils Blüthgen, Job Dekker, and Edith Heard. Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature*, 485(7398):381–385, April 2012.

[113] Reza Kalhor, Harianto Tjong, Nimanthi Jayathilaka, Frank Alber, and Lin Chen. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.*, 30(1):90–98, December 2011.

[114] Takashi Nagano, Yaniv Lubling, Tim J Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D Laue, Amos Tanay, and Peter Fraser. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, October 2013.

[115] Takashi Nagano, Yaniv Lubling, Csilla Várnai, Carmel Dudley, Wing Leung, Yael Baran, Netta Mendelson Cohen, Steven Wingett, Peter Fraser, and Amos Tanay. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature*, 547(7661):61–67, July 2017.

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

[117] Tim J Stevens, David Lando, Srinjan Basu, Liam P Atkinson, Yang Cao, Steven F Lee, Martin Leeb, Kai J Wohlfahrt, Wayne Boucher, Aoife O'Shaughnessy-Kirwan, Julie Cramard, Andre J Faure, Meryem Ralser, Enrique Blanco, Lluis Morey, Miriam Sansó, Matthieu G S Palayret, Ben Lehner, Luciano Di Croce, Anton Wutz, Brian Hendrich, Dave Klenerman, and Ernest D Laue. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*, 544(7648):59–64, April 2017.

[118] Longzhi Tan, Dong Xing, Chi-Han Chang, Heng Li, and X Sunney Xie. Three-dimensional genome structures of single diploid human cells. *Science*, 361(6405):924–928, August 2018.

[119] Thomas Cremer and Marion Cremer. Chromosome territories. *Cold Spring Harb. Perspect. Biol.*, 2(3):a003889, March 2010.

[120] James Fraser, Iain Williamson, Wendy A Bickmore, and Josée Dostie. An overview of genome organization and how we got there: from FISH to Hi-C, 2015.

[121] Siyuan Wang, Jun-Han Su, Brian J Beliveau, Bogdan Bintu, Jeffrey R Moffitt, Chao-Ting Wu, and Xiaowei Zhuang. Spatial organization of chromatin domains and compartments in single chromosomes. *Science*, 353(6299):598–602, August 2016.

[122] Diego I Cattoni, Andrés M Cardozo Gizzi, Mariya Georgieva, Marco Di Stefano, Alessandro Valeri, Delphine Chamousset, Christophe Houbron, Stephanie Déjardin, Jean-Bernard Fiche, Inma González, Jia-Ming Chang, Thomas Sexton, Marc A Marti-Renom, Frédéric Bantignies, Giacomo Cavalli, and Marcelo Nollmann. Single-cell absolute contact probability detection reveals chromosomes are organized by multiple low-frequency yet specific interactions. *Nat. Commun.*, 8(1):1–10, November 2017.

[123] Bogdan Bintu, Leslie J Mateo, Jun-Han Su, Nicholas A Sinnott-Armstrong, Mirae Parker, Seon Kinrot, Kei Yamaya, Alistair N Boettiger, and Xiaowei Zhuang. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science*, 362(6413), October 2018.

[124] Guy Nir, Irene Farabella, Cynthia Pérez Estrada, Carl G Ebeling, Brian J Beliveau, Hiroshi M Sasaki, S Dean Lee, Son C Nguyen, Ruth B McCole, Shyamtanu Chattoraj, Jelena Erceg, Jumana AlHaj Abed, Nuno M C Martins, Huy Q Nguyen, Mohammed A Hannan, Sheikh Russell, Neva C Durand, Suhas S P Rao, Jocelyn Y Kishi, Paula Soler-Vila, Michele Di Pierro, José N Onuchic, Steven P Callahan, John M Schreiner, Jeff A Stuckey, Peng Yin, Erez Lieberman Aiden, Marc A Marti-Renom, and C-Ting

Wu. Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling. *PLoS Genet.*, 14(12):e1007872, December 2018.

[125] Quentin Szabo, Daniel Jost, Jia-Ming Chang, Diego I Cattoni, Giorgio L Papadopoulos, Boyan Bonev, Tom Sexton, Julian Gurgo, Caroline Jacquier, Marcelo Nollmann, Frédéric Bantignies, and Giacomo Cavalli. TADs are 3D structural units of higher-order chromosome organization in drosophila. *Science Advances*, 4(2), February 2018.

[126] Elizabeth H Finn, Gianluca Pegoraro, Hugo B Brandão, Anne-Laure Valton, Marlies E Oomen, Job Dekker, Leonid Mirny, and Tom Misteli. Extensive heterogeneity and intrinsic variation in spatial genome organization. *Cell*, 176(6):1502–1515.e10, March 2019.

[127] Leslie J Mateo, Sedona E Murphy, Antonina Hafner, Isaac S Cinquini, Carly A Walker, and Alistair N Boettiger. Visualizing DNA folding and RNA in embryos at single-cell resolution. *Nature*, 568(7750):49–54, March 2019.

[128] Andrés M Cardozo Gizzi, Diego I Cattoni, Jean-Bernard Fiche, Sergio M Espinola, Julian Gurgo, Olivier Messina, Christophe Houbron, Yuki Ogiyama, Giorgio L Papadopoulos, Giacomo Cavalli, Mounia Lagha, and Marcelo Nollmann. Microscopy-Based chromosome conformation capture enables simultaneous visualization of genome organization and transcription in intact organisms. *Mol. Cell*, 74(1):212–222.e5, April 2019.

[129] Brian J Beliveau, Alistair N Boettiger, Maier S Avendaño, Ralf Jungmann, Ruth B McCole, Eric F Joyce, Caroline Kim-Kiselak, Frédéric Bantignies, Chamith Y Fonseka, Jelena Erceg, Mohammed A Hannan, Hien G Hoang, David Colognori, Jeannie T Lee, William M Shih, Peng Yin, Xiaowei Zhuang, and Chao-Ting Wu. Single-molecule super-resolution imaging of chromosomes and in situ haplotype visualization using oligopaint FISH probes. *Nat. Commun.*, 6:7147, May 2015.

[130] Sai Ma, Bing Zhang, Lindsay M LaFave, Andrew S Earl, Zachary Chiang, Yan Hu, Jiarui Ding, Alison Brack, Vinay K Kartha, Tristan Tay, Travis Law, Caleb Lareau, Ya-Chieh Hsu, Aviv Regev, and Jason D Buenrostro. Chromatin potential identified by shared Single-Cell profiling of RNA and chromatin. *Cell*, 183(4):1103–1116.e20, November 2020.

[131] Xingqi Chen, Ying Shen, Will Draper, Jason D Buenrostro, Ulrike Litzenburger, Seung Woo Cho, Ansuman T Satpathy, Ava C Carter, Rajarshi P Ghosh, Alexandra East-Seletsky, Jennifer A Doudna, William J Greenleaf, Jan T Liphardt, and Howard Y Chang. ATAC-see reveals the accessible genome by transposase-mediated imaging and sequencing. *Nat. Methods*, 13(12):1013–1020, December 2016.

[132] Irina Solovei, Antonio Cavallo, Lothar Schermelleh, Françoise Jaunin, Catia Scasselati, Dusan Cmarko, Christoph Cremer, Stanislav Fakan, and Thomas Cremer. Spatial preservation of nuclear chromatin architecture during three-dimensional fluorescence in situ hybridization (3D-FISH). *Exp. Cell Res.*, 276(1):10–23, May 2002.

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

[134] Jeffrey R Moffitt, Junjie Hao, Guiping Wang, Kok Hao Chen, Hazen P Babcock, and Xiaowei Zhuang. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proc. Natl. Acad. Sci. U. S. A.*, 113(39):11046–11051, September 2016.

[135] Caleb A Lareau, Fabiana M Duarte, Jennifer G Chew, Vinay K Kartha, Zach D Burkett, Andrew S Kohlway, Dmitry Pokholok, Martin J Aryee, Frank J Steemers, Ronald Lebofsky, and Jason D Buenrostro. Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat. Biotechnol.*, 37(8):916–924, August 2019.

[136] Zhenhai Du, Hui Zheng, Bo Huang, Rui Ma, Jingyi Wu, Xianglin Zhang, Jing He, Yunlong Xiang, Qiujun Wang, Yuanyuan Li, Jing Ma, Xu Zhang, Ke Zhang, Yang Wang, Michael Q Zhang, Juntao Gao, Jesse R Dixon, Xiaowo Wang, Jianyang Zeng, and Wei Xie. Allelic reprogramming of 3D chromatin architecture during early mammalian development. *Nature*, 547(7662):232–235, July 2017.

[137] E V Volpi, E Chevret, T Jones, R Vatcheva, J Williamson, S Beck, R D Campbell, M Goldsworthy, S H Powis, J Ragoussis, J Trowsdale, and D Sheer. Large-scale chromatin organization of the major histocompatibility complex and other regions of human chromosome 6 and its response to interferon in interphase nuclei. *J. Cell Sci.*, 113 (Pt 9):1565–1576, May 2000.

[138] Sheel Shah, Yodai Takei, Wen Zhou, Eric Lubeck, Jina Yun, Chee-Huat Linus Eng, Noushin Koulena, Christopher Cronin, Christoph Karp, Eric J Liaw, Mina Amin, and Long Cai. Dynamics and spatial genomics of the nascent transcriptome by intron seqFISH. *Cell*, 174(2):363–376.e16, July 2018.

[139] E V Volpi, E Chevret, T Jones, R Vatcheva, J Williamson, S Beck, R D Campbell, M Goldsworthy, S H Powis, J Ragoussis, J Trowsdale, and D Sheer. Large-scale chromatin organization of the major histocompatibility complex and other regions of human chromosome 6 and its response to interferon in interphase nuclei. *J. Cell Sci.*, 113 (Pt 9):1565–1576, May 2000.

[140] Andreas Bolzer, Gregor Kreth, Irina Solovei, Daniela Koehler, Kaan Saracoglu, Christine Fauth, Stefan Müller, Roland Eils, Christoph Cremer, Michael R Speicher, and Thomas Cremer. Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol.*, 3(5):e157, May 2005.

[141] A P Jason de Koning, Wanjun Gu, Todd A Castoe, Mark A Batzer, and David D Pollock. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.*, 7(12):e1002384, December 2011.

[142] Karen H Miga, Sergey Koren, Arang Rhie, Mitchell R Vollger, Ariel Gershman, Andrey Bzikadze, Shelise Brooks, Edmund Howe, David Porubsky, Glennis A Logsdon, Valerie A Schneider, Tamara Potapova, Jonathan Wood, William Chow, Joel Armstrong, Jeanne Fredrickson, Evgenia Pak, Kristof Tigyi, Milinn Kremitzki, Christopher Markovic, Valerie Maduro, Amalia Dutra, Gerard G Bouffard, Alexander M Chang, Nancy F Hansen, Amy B Wilfert, Françoise Thibaud-Nissen, Anthony D Schmitt, Jon-Matthew Belton, Siddarth Selvaraj, Megan Y Dennis, Daniela C Soto, Ruta Sahasrabudhe, Gulhan Kaya, Josh Quick, Nicholas J Loman, Nadine Holmes, Matthew Loose, Urvashi Surti, Rosa Ana Risques, Tina A Graves Lindsay, Robert Fulton, Ira Hall, Benedict Paten, Kerstin Howe, Winston Timp, Alice Young, James C Mullikin, Pavel A Pevzner, Jennifer L Gerton, Beth A Sullivan, Evan E Eichler, and Adam M Phillippy. Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, July 2020.

[143] Axel Cournac, Romain Koszul, and Julien Mozziconacci. The 3D folding of metazoan genomes correlates with the association of similar repetitive elements. *Nucleic Acids Res.*, 44(1):245–255, January 2016.

[144] Valentina Casa and Davide Gabellini. A repetitive elements perspective in polycomb epigenetics. *Front. Genet.*, 3:199, October 2012.

[145] Michael Hausmann, Jin-Ho Lee, Aaron Sievers, Matthias Krufczik, and Georg Hildenbrand. COMBinatorial oligonucleotide FISH (COMBO-FISH) with uniquely binding repetitive DNA probes. *Methods Mol. Biol.*, 2175:65–77, 2020.

[146] Weidong Bao, Kenji K Kojima, and Oleksiy Kohany. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA*, 6:11, June 2015.

[147] Martin Falk, Yana Feodorova, Natalia Naumova, Maxim Imakaev, Bryan R Lajoie, Heinrich Leonhardt, Boris Joffe, Job Dekker, Geoffrey Fudenberg, Irina Solovei, and Leonid A Mirny. Heterochromatin drives compartmentalization of inverted and conventional nuclei. *Nature*, June 2019.

[148] Steffen Dietzel, Anna Jauch, Dirk Kienle, Guoquiong Qu, Heidi Holtgreve-Grez, Roland Eils, Christian Munkel, Michael Bittner, Paul S Meltzer, Jeffrey M Trent, et al. Separate and variably shaped chromosome arm domains are disclosed by chromosome arm painting in human cell nuclei. *Chromosome Research*, 6(1):25–33, 1998.

[149] Héloïse Muller, José Gil, Jr, and Ines Anna Drinnenberg. The impact of centromeres on spatial genome architecture. *Trends Genet.*, 35(8):565–578, August 2019.

[150] Clemens B Hug and Juan M Vaquerizas. The birth of the 3D genome during early embryonic development. *Trends Genet.*, 34(12):903–914, December 2018.

[151] Claire Chazaud and Yojiro Yamanaka. Lineage specification in the mouse preimplantation embryo. *Development*, 143(7):1063–1074, April 2016.

[152] Satoshi H Namekawa, Bernhard Payer, Khanh D Huynh, Rudolf Jaenisch, and Jeannie T Lee. Two-step imprinted X inactivation: repeat versus genic silencing in the mouse. *Mol. Cell. Biol.*, 30(13):3187–3205, July 2010.

[153] Máté Borsos, Sara M Perricone, Tamás Schauer, Julien Pontabry, Kim L de Luca, Sandra S de Vries, Elias R Ruiz-Morales, Maria-Elena Torres-Padilla, and Jop Kind. Genome–lamina interactions are established de novo in the early mouse embryo. *Nature*, 569(7758):729–733, May 2019.

[154] Ilya M Flyamer, Johanna Gassler, Maxim Imakaev, Hugo B Brandão, Sergey V Ulianov, Nezar Abdennur, Sergey V Razin, Leonid A Mirny, and Kikuë Tachibana-Konwalski. Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature*, 544(7648):110–114, April 2017.

[155] Zhenhai Du, Hui Zheng, Bo Huang, Rui Ma, Jingyi Wu, Xianglin Zhang, Jing He, Yunlong Xiang, Qiujun Wang, Yuanyuan Li, Jing Ma, Xu Zhang, Ke Zhang, Yang Wang, Michael Q Zhang, Juntao Gao, Jesse R Dixon, Xiaowo Wang, Jianyang Zeng, and Wei Xie. Allelic reprogramming of 3D chromatin architecture during early mammalian development. *Nature*, 547(7662):232–235, July 2017.

[156] Yuwen Ke, Yanan Xu, Xuepeng Chen, Songjie Feng, Zhenbo Liu, Yaoyu Sun, Xuelong Yao, Fangzhen Li, Wei Zhu, Lei Gao, Haojie Chen, Zhenhai Du, Wei Xie, Xiaocui Xu, Xingxu Huang, and Jiang Liu. 3D chromatin structures of mature gametes and structural reprogramming during mammalian embryogenesis. *Cell*, 170(2):367–381.e20, July 2017.

[157] Samuel Collombet, Noémie Ranisavljevic, Takashi Nagano, Csilla Varnai, Tarak Shisode, Wing Leung, Tristan Piolot, Rafael Galupa, Maud Borensztein, Nicolas Servant, Peter Fraser, Katia Ancelin, and Edith Heard. Parental-to-embryo switch of chromosome organization in early embryogenesis. *Nature*, 580(7801):142–146, April 2020.

[158] Maria-Elena Torres-Padilla, David-Emlyn Parfitt, Tony Kouzarides, and Magdalena Zernicka-Goetz. Histone arginine methylation regulates pluripotency in the early mouse embryo. *Nature*, 445(7124):214–218, January 2007.

[159] Adam Burton, Julius Muller, Shengjiang Tu, Pablo Padilla-Longoria, Ernesto Guccione, and Maria-Elena Torres-Padilla. Single-cell profiling of epigenetic modifiers identifies PRDM14 as an inducer of cell fate in the mammalian embryo. *Cell Rep.*, 5(3):687–701, November 2013.

[160] Adam Burton and Maria-Elena Torres-Padilla. Chromatin dynamics in the regulation of cell fate allocation during early embryogenesis. *Nat. Rev. Mol. Cell Biol.*, 15(11):723–734, November 2014.

[161] Tiphaine Aguirre-Lavin, Pierre Adenot, Amélie Bonnet-Garnier, Gaétan Lehmann, Renaud Fleurot, Claire Boulesteix, Pascale Debey, and Nathalie Beaujean. 3D-FISH analysis of embryonic nuclei in mouse highlights several abrupt changes of nuclear

organization during preimplantation development. *BMC Dev. Biol.*, 12:30, October 2012.

[162] Helena Fulka and Fugaku Aoki. Nucleolus precursor bodies and ribosome biogenesis in early mammalian embryos: Old theories and new discoveries 1. *Biol. Reprod.*, 94(6), June 2016.

[163] W Mayer, A Smith, R Fundele, and T Haaf. Spatial separation of parental genomes in preimplantation mouse embryos. *J. Cell Biol.*, 148(4):629–634, February 2000.

[164] Judith Reichmann, Bianca Nijmeijer, M Julius Hossain, Manuel Eguren, Isabell Schneider, Antonio Z Politi, M Julia Roberti, Lars Hufnagel, Takashi Hiiragi, and Jan Ellenberg. Dual-spindle formation in zygotes keeps parental genomes apart in early mammalian embryos. *Science*, 361(6398):189–193, July 2018.

[165] C R Cowan, P M Carlton, and W Z Cande. The polar arrangement of telomeres in interphase and meiosis. rabl organization and the bouquet. *Plant Physiol.*, 125(2):532–538, February 2001.

[166] Maxime Pouokam, Brian Cruz, Sean Burgess, Mark R Segal, Mariel Vazquez, and Javier Arsuaga. The rabl configuration limits topological entanglement of chromosomes in budding yeast. *Sci. Rep.*, 9(1):6795, May 2019.

[167] Miler T Lee, Ashley R Bonneau, and Antonio J Giraldez. Zygotic genome activation during the Maternal-to-Zygotic transition, 2014.

[168] Kashif Ahmed, Hesam Dehghani, Peter Rugg-Gunn, Eden Fussner, Janet Rossant, and David P Bazett-Jones. Global chromatin architecture reflects pluripotency and lineage commitment in the early mouse embryo. *PLoS One*, 5(5):e10531, May 2010.

[169] Nicolas Plachta, Tobias Bollenbach, Shirley Pease, Scott E Fraser, and Periklis Pantazis. Oct4 kinetics predict cell lineage patterning in the early mammalian embryo. *Nat. Cell Biol.*, 13(2):117–123, February 2011.

[170] Karolina Piotrowska-Nitsche, Aitana Perea-Gomez, Seiki Haraguchi, and Magdalena Zernicka-Goetz. Four-cell stage mouse blastomeres have different developmental properties. *Development*, 132(3):479–490, February 2005.

[171] Joachim Walter, Lothar Schermelleh, Marion Cremer, Satoshi Tashiro, and Thomas Cremer. Chromosome order in HeLa cells changes during mitosis and early g1, but is stably maintained during subsequent interphase stages. *J. Cell Biol.*, 160(5):685–697, March 2003.

[172] Daniel Gerlich, Joël Beaudouin, Bernd Kalbfuss, Nathalie Daigle, Roland Eils, and Jan Ellenberg. Global chromosome positions are transmitted through mitosis in mammalian cells. *Cell*, 112(6):751–764, March 2003.

[173] Jop Kind, Ludo Pagie, Havva Ortabozkoyun, Shelagh Boyle, Sandra S de Vries, Hans Janssen, Mario Amendola, Leisha D Nolen, Wendy A Bickmore, and Bas van Steensel. Single-cell dynamics of genome-nuclear lamina interactions. *Cell*, 153(1):178–192, March 2013.

[174] Inga Thomson, Susan Gilchrist, Wendy A Bickmore, and Jonathan R Chubb. The radial positioning of chromatin is not inherited through mitosis but is established de novo in early G1. *Curr. Biol.*, 14(2):166–172, January 2004.

[175] Cheng-Sheng Lee, Ruoxi W Wang, Hsiao-Han Chang, Daniel Capurso, Mark R Segal, and James E Haber. Chromosome position determines the success of double-strand break repair. *Proc. Natl. Acad. Sci. U. S. A.*, 113(2):E146–54, January 2016.

[176] E R Phillips and P J McKinnon. DNA double-strand break repair and development. *Oncogene*, 26(56):7799–7808, December 2007.

[177] Evelyne Vanneste, Thierry Voet, Cédric Le Caignec, Michèle Ampe, Peter Konings, Cindy Melotte, Sophie Debrock, Mustapha Amyere, Miikka Vikkula, Frans Schuit, Jean-Pierre Fryns, Geert Verbeke, Thomas D'Hooghe, Yves Moreau, and Joris R Vermeesch. Chromosome instability is common in human cleavage-stage embryos. *Nat. Med.*, 15(5):577–583, May 2009.

[178] Yu Zhang, Rachel Patton McCord, Yu-Jui Ho, Bryan R Lajoie, Dominic G Hildebrand, Aline C Simon, Michael S Becker, Frederick W Alt, and Job Dekker. Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell*, 148(5):908–921, March 2012.

[179] Jesse M Engreitz, Vineeta Agarwala, and Leonid A Mirny. Three-dimensional genome architecture influences partner selection for chromosomal translocations in human disease. *PLoS One*, 7(9):e44196, September 2012.

[180] Jennifer M Luppino, Daniel S Park, Son C Nguyen, Yemin Lan, Zhuxuan Xu, Rebecca Yunker, and Eric F Joyce. Cohesin promotes stochastic domain intermingling to ensure proper regulation of boundary-proximal genes. *Nat. Genet.*, 52(8):840–848, June 2020.

[181] Brandon D Fields, Son C Nguyen, Guy Nir, and Scott Kennedy. A multiplexed DNA FISH strategy for assessing genome architecture in caenorhabditis elegans. *Elife*, 8, May 2019.

[182] Robert A Beagrie, Antonio Scialdone, Markus Schueler, Dorothee C A Kraemer, Mita Chotalia, Sheila Q Xie, Mariano Barbieri, Inês de Santiago, Liron-Mark Lavitas, Miguel R Branco, James Fraser, Josée Dostie, Laurence Game, Niall Dillon, Paul A W Edwards, Mario Nicodemi, and Ana Pombo. Complex multi-enhancer contacts captured by genome architecture mapping. *Nature*, 543(7646):519–524, March 2017.

[183] S A Quinodoz, N Ollikainen, B Tabak, A Palla, J M Schmidt, E Detmar, M M Lai, A A Shishkin, P Bhat, Y Takei, V Trinh, E Aznauryan, P Russell, C Cheng, M Jovanovic, A Chow, L Cai, P McDonel, M Garber, and M Guttman. Higher-Order

inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell*, 174(3), July 2018.

[184] Gabriele Girelli, Joaquin Custodio, Tomasz Kallas, Federico Agostini, Erik Wernersson, Bastiaan Spanjaard, Ana Mota, Solrun Kolbeinsdottir, Eleni Gelali, Nicola Crosetto, and Magda Bienko. GPSeq reveals the radial organization of chromatin in the cell nucleus. *Nat. Biotechnol.*, pages 1–10, May 2020.

[185] Daniele Zink, Andrew H Fischer, and Jeffrey A Nickerson. Nuclear structure in cancer cells. *Nat. Rev. Cancer*, 4(9):677–687, September 2004.

[186] Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang, and William J Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, 10(12):1213–1218, December 2013.

[187] Carl-Magnus Clausson, Linda Arngården, Omer Ishaq, Axel Klaesson, Malte Kühnemund, Karin Grannas, Björn Koos, Xiaoyan Qian, Petter Ranefall, Tomasz Krzywkowski, Hjalmar Brismar, Mats Nilsson, Carolina Wählby, and Ola Söderberg. Compaction of rolling circle amplification products increases signal integrity and signal-to-noise ratio. *Sci. Rep.*, 5:12317, July 2015.

[188] Fei Chen, Paul W Tillberg, and Edward S Boyden. Optical imaging. expansion microscopy. *Science*, 347(6221):543–548, January 2015.

[189] Shahar Alon, Daniel R Goodwin, Anubhav Sinha, Asmamaw T Wassie, Fei Chen, Evan R Daugharthy, Yosuke Bando, Atsushi Kajita, Andrew G Xue, Karl Marrett, Robert Prior, Yi Cui, Andrew C Payne, Chun-Chen Yao, Ho-Jun Suk, Ru Wang, Chih-Chieh (jay) Yu, Paul Tillberg, Paul Reginato, Nikita Pak, Songlei Liu, Sukanya Punthambaker, Eswar P R Iyer, Richie E Kohman, Jeremy A Miller, Ed S Lein, Ana Lako, Nicole Cullen, Scott Rodig, Karla Helvie, Daniel L Abravanel, Nikhil Wagle, Bruce E Johnson, Johanna Klughammer, Michal Slyper, Julia Waldman, Judit Jané-Valbuena, Orit Rozenblatt-Rosen, Aviv Regev, IMAXT Consortium, George M Church, Adam H Marblestone, and Edward S Boyden. Expansion sequencing: Spatially precise in situ transcriptomics in intact biological systems. May 2020.

[190] Chenxu Zhu, Sebastian Preissl, and Bing Ren. Single-cell multimodal omics: the power of many. *Nat. Methods*, 17(1):11–14, January 2020.

[191] Simone Picelli, Åsa K Björklund, Björn Reinius, Sven Sagasser, Gösta Winberg, and Rickard Sandberg. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.*, 24(12):2033–2040, December 2014.

[192] Je Hyuk Lee, Evan R Daugharthy, Jonathan Scheiman, Reza Kalhor, Thomas C Ferrante, Richard Terry, Brian M Turczyk, Joyce L Yang, Ho Suk Lee, John Aach, Kun Zhang, and George M Church. Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat. Protoc.*, 10(3):442–458, March 2015.

[193] Huy Q Nguyen, Shyamtanu Chattoraj, David Castillo, Son C Nguyen, Guy Nir, Antonios Lioutas, Elliot A Hershberg, Nuno M C Martins, Paul L Reginato, Mohammed Hannan, Brian J Beliveau, George M Church, Evan R Daugharthy, Marc A Marti-Renom, and C-Ting Wu. 3D mapping and accelerated super-resolution imaging of the human genome using in situ sequencing. *Nat. Methods*, 17(8):822–832, August 2020.

[194] Jason D Buenrostro, Beijing Wu, Howard Y Chang, and William J Greenleaf. ATAC-seq: A method for assaying chromatin accessibility Genome-Wide. *Curr. Protoc. Mol. Biol.*, 109:21.29.1–9, January 2015.

[195] B Langmead and S L Salzberg. Langmead. 2013. bowtie2. *Nat. Methods*, 9:357–359, 2013.

[196] Tom Smith, Andreas Heger, and Ian Sudbery. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res.*, 27(3):491–499, March 2017.

[197] Tim Massingham and Nick Goldman. Error-correcting properties of the SOLiD exact call chemistry. *BMC Bioinformatics*, 13:145, June 2012.

[198] Stuart Berg, Dominik Kutra, Thorben Kroeger, Christoph N Straehle, Bernhard X Kausler, Carsten Haubold, Martin Schiegg, Janez Ales, Thorsten Beier, Markus Rudy, and Others. Ilastik: interactive machine learning for (bio) image analysis. *Nat. Methods*, pages 1–7, 2019.

[199] Hiroki R Ueda, Ali Ertürk, Kwanghun Chung, Viviana Gradinaru, Alain Chédotal, Pavel Tomancak, and Philipp J Keller. Tissue clearing and its applications in neuroscience. *Nat. Rev. Neurosci.*, 21(2):61–79, February 2020.

[200] A R Quinlan and I M Hall. BEDTools: a flexible suite of utilities for comparing genomic features. bioinformatics [internet] 26: 841–842, 2010.

[201] Nicolas Servant, Nelle Varoquaux, Bryan R Lajoie, Eric Viara, Chong-Jian Chen, Jean-Philippe Vert, Edith Heard, Job Dekker, and Emmanuel Barillot. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.*, 16:259, December 2015.

[202] Neva C Durand, Muhammad S Shamim, Ido Machol, Suhas S P Rao, Miriam H Huntley, Eric S Lander, and Erez Lieberman Aiden. Juicer provides a One-Click system for analyzing Loop-Resolution Hi-C experiments. *Cell Syst*, 3(1):95–98, July 2016.

[203] Stefan Schoenfelder and Peter Fraser. Long-range enhancer–promoter contacts in gene expression control. *Nature Reviews Genetics*, 20(8):437–455, 2019.

[204] Ivona Kubalova, Marketa Schmidt Cernohorska, Martina Huranova, Klaus Weisshart, Andreas Houben, and Veit Schubert. Prospects and limitations of expansion microscopy in chromatin ultrastructure determination. *Chromosome Research*, 28(3):355–368, 2020.

[205] Kenneth Pham, Alexandria Nikish, and Jennifer E Phillips-Cremins. See (quence) and ye shall find: higher-order genome folding in intact single cells. *Molecular Cell*, 81(6):1130–1132, 2021.