Development of single-cell SPRITE: A tool for measuring heterogeneity of 3D DNA organization

Thesis by

Mary Villanueva Arrastia

In Partial Fulfillment of the Requirements for

the Degree of

Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2022

(Defended September 27, 2021)

Mary Villanueva Arrastia
ORCID: 0000-0002-0723-3574

ACKNOWLEDGEMENTS

"It takes a village to raise a child" – a well-known proverb that resonates very deeply throughout my entire graduate school journey. There have been numerous individuals and groups that have all been instrumental in my growth, both as a scientist and as an individual.

First and foremost, I definitely could not have gotten this far without the support of my advisor, Rustem Ismagilov. I was completely humbled when you accepted me into your lab and allowed me to explore my interests in biologically related research, despite knowing that my undergraduate degree was in chemistry and I had very limited knowledge in biology. You shaped me to become a better scientist by training me to think more critically about my experimental questions and design. When I first joined the lab, I was naïve and unfocused in how I approached my science, but the years under your guidance have helped me become more focused in how to approach research problems by asking thoughtful and practical questions. You presented me with multiple opportunities to grow as an independent scientist, such as attending conferences, presenting at seminars, writing grants and fellowships, preparing patents, and publishing in journals, and I gained invaluable experience through these numerous opportunities. You also presented me with the opportunities to take leadership, both through maintaining collaborations throughout my thesis project and through upholding lab's high standards of safety as one of the lab's Biosafety Coordinators. Most importantly, you served as a mentor throughout my entire graduate school journey. You were always willing to provide feedback to help me improve and constantly enlightened me with his insight and knowledge. One of the most important lessons you have taught me is that when the science doesn't work, it doesn't mean that I was a failure. Hearing that from you really helped change my perception about myself in science and how to better approach failure in research.

I would also like to thank my committee members: Mitch Guttman, Dennis Dougherty, and Long Cai. The advice, support, and feedback you have provided to me over the years have also helped me throughout the course of my graduate career towards my growth as an independent scientist. I am forever grateful for your guidance over these past years. Outside

of your role as my committee members, each of you has contributed toward my scientific growth in unique ways. Mitch, you served as the key collaborator throughout my thesis project, and I'm very grateful for the opportunities I had to learn from you and your lab and the biological insight you provided in my research. Dennis, you were one of the first people I interacted with at Caltech through the Chemical Biology recruitment dinners and the interactions I had with you and other faculty helped seal my decision to choose Caltech for my graduate studies. Long, I remember being absolutely dumbfounded the first time I interacted with you during recruitment weekend because your research was in a field I never seen or ever considered before, but it made me realize that there was a realm of possibility when it comes to research ideas and projects and that the sky is not the limit.

I also would like to thank the past and present members of the Ismagilov group for their support over these past six years. The group has served as a second home during my time here at Caltech, and I truly appreciate the opportunity to work, learn, and relax alongside an amazing group of individuals. Throughout my time in the Ismagilov group, I had the opportunity to learn and become one of the lab's Biosafety Coordinators, alongside Emily Savela and Eugenia Khorosheva. I had never imagined myself being in this role, but you were an amazing duo of Biosafety Coordinators to work with and I was fortunate to receive much guidance and knowledge from both of you to become effective in maintaining our lab's high standards for biosafety. However, you both also served as great friends to socialize with, in addition to many other folks from the lab, including Sarah Simon, Joanne Lau, Roberta Poceviciute, Tahmineh (Tami) Khazaei, Anna Romano, Aditi Narayanan (while only a rotation student at the time, you still count as a former lab member!), Asher Preska Steinberg, Joong Hwan Bahng, Matthew Curtis, and David Selck. I will never forget the crazy adventures that I have had with folks from the lab, whether they involved random lab shenanigans, exchanging crazy stories, or having fun during lab socials.

I am also truly indebted to the wonderful staff members in our lab, including Natasha (Tasha) Shelby, Sohee Lee, and Rosie Zedan. Tasha, I am incredibly grateful that you were always willing to edit my papers, grants, and fellowship applications, even taking time away from

your weekends to help in these endeavors. I also greatly appreciate that you're always willing to lend an ear whenever things just simply aren't going our way, and you constantly encourage us through every step of our journeys. I am also honored to have had the opportunity to camp outside of your home in Joshua Tree to watch and photograph the Quadrantid meteor shower. The lack of light pollution from your place really made the stars and meteors stand out! Sohee, thank you for always being available to help reserve rooms, arrange meetings, and handle day-to-day operations for me and the lab—these small things go a long way in making our lives more manageable each day! And Rosie, thank you for restocking our lab to make sure we were fully equipped for our daily lab functions. I always enjoyed talking with you whenever you were there, and I enjoyed hearing your stories of your recent and upcoming travel adventures. I felt like I was living vicariously through you! Ultimately, what I absolutely cherish the most about the lab as a whole is that there is always someone to put a smile on my face, whether I have had an amazing day or an absolute terrible day. While it is a small gesture, it goes a long way. I know that means the lab cares not only for my well-being, but also for the well-being of everyone else in the lab, and that is a trait I hope I can find in my future postdoc lab group.

Additionally, I had the fortunate opportunity to closely work with the Guttman lab during my graduate years. I particularly want to thank Joanna Jachowicz and Sofi Quinodoz, both of whom I worked closely with and learned deeply from during my graduate years. Joanna and Sofi, thank you for your patience and insight over the years in working with you throughout my graduate project. I am so grateful that I was able to collaborate with both of you over the years. Additionally, I want to say a big thank you to all of the individuals I had the opportunity to work and interact with over the years, including Noah Ollikainen, Charlotte Lai, Elizabeth Soehalim, Elizabeth Detmar, Vicky Trinh, Chris Chen, Isabel Goronzy, and Inna-Marie Strazhnik. I definitely could not have gotten to where I am today without all of your contributions along the way.

I also want to acknowledge many of the Caltech staff I've had the chance to interact with during my years at Caltech. First, I want to say thanks to Cindy Weinstein, Caltech's Chief

Diversity Officer. She was the first person at Caltech I had the opportunity to meet with, even before my official recruitment weekend. I was first introduced to her thanks to Michelle Hawley, one of the former directors of CSULA's Honors College program, after being accepted to Caltech's PhD program in Chemistry. Cindy knew the challenges that were up ahead, and she helped ease my initial nerves and answered all my questions about being a graduate student at Caltech.

Next, I want to say thank you to two wonderful Chemistry Option Representatives: Agnes Tong and Alison Ross. Before taking up her new position at the Caltech Y, Agnes helped me get settled during my initial years of graduate school. She helped become involved with Chemistry-related functions, such as leading recruitment events and organizing social events through Programming Board. It was disappointing to hear that Agnes would be leaving her position to transition over to the Caltech Y, but I was still able to drop by her office, chat with each other, and play with her dogs. When Alison took over as Option Rep, her office became a haven where I could unwind when things were overly stressful and overwhelming. We often went back and forth just ranting our stresses at each other, but we also shared fun stories and adventures with each other. During recruitment events, even though I stepped back from being directly recruitment activities, I always remember Alison being super grateful for seeing me at socials, since she always relied on my loud voice to get the attention of everyone. We'd always laughed about this, but I hope she's able to find a replacement voice to take over… or a megaphone. (Although, I recall the one year where she did hand me a megaphone so I wouldn't strain my voice. My advisor saw me with it and became instantly terrified. Perhaps one of the best laughs I had during my time in graduate school.)

In addition to the option representatives, I've enjoyed the company of the crew in our chemistry stockroom: Joe Drew, Armando Villasenor, and Greg Rolette. Though most of my interactions with them were during package pick up or delivery, they always were super friendly to talk with, and they always knew how to crack a laugh out of me. They were always curious about the designs on my graphic t-shirts (attire I'd wear almost daily), and I'd be happy to share whatever design or pun the shirts were conveying. I was also happy when

they dropped by my poster during on-campus poster sessions. Even though they didn't know much of the science, they were always interested to know what I was working on, and I'd be happy to talk with them about it.

Another Caltech staff member I would like to thank is Lauriane Queene, the Institute Biosafety Officer at Caltech. Jenia, Emily, and I would work very closely with Lauriane when addressing biosafety-related aspects in our lab. Through Lauriane, I learned a lot of what it takes to become a Biosafety Coordinator, ranging on how to assemble protocols for review by the Institutional Biosafety Committee, putting together proper Standard Operating Procedures, to even learning how to think and respond when it comes to addressing biosafety issues in our lab. She also shared her own individual experiences of handling biohazardous materials during her lab days, and she always impresses me with her knowledge and expertise. I'm incredibly fortunate to have worked alongside Lauriane, and I hope there will be someone just like her at my next workplace.

Although not directly affiliated with Caltech, Enrica Bruno serves as our lab's go-to for patent law. I got the chance to work directly with Enrica when we started filing our non-provisional patent for our scSPRITE method. She was an incredibly great person to learn patent law from, as someone who has never dealt with patent law until graduate school. She was patient to teach me the tips and tricks in laying out claims and the whole process of getting a patent approved. She was willing to let me work on-site at her firm, and I gladly took up the offer, which gave me the opportunity to directly ask her questions as they arose. Even though patent law may not be in my immediate future, I'm incredibly grateful for the opportunity to learn alongside her.

Throughout graduate school, I was fortunate to stay in Southern California, where I was born and raised. What this also means is that I remained close to my friends throughout graduate school, many of whom I've known since my high school days. First, I have to say thank you to Effie Tong-Lin, who has been my best friend since our high school days. Even though we've diverged towards different career paths, I'm happy we still continued to keep in touch.

I'm so happy to be able to engage in new food adventures, assembling puzzles, and outdoor activities with you. Thanks for sticking around for the wild ride that is graduate school. Additionally, I'm fortunate to have other close high school friends who I continue to talk and meet with on a regular basis, including William Shiraga, Jo Lee (previously Jo Bang), and Marissa Suto. Thank you all for continuing to keep in touch and for tolerating my random, crazy shenanigans. Lastly, I have to thank two wonderful high school teachers who made me interested in pursuing STEM in college and beyond: Shannon Regli and David Booze, both of whom I continue to interact and chat with outside of the classroom.

For my undergraduate studies, I attended California State University, Los Angeles (CSULA), which is where my passion for research really blossomed. I definitely would not be in this position today if it weren't for my undergraduate advisor, Frank Gomez. He invited me to his lab during my first year at CSULA, and the research I had done in his lab was my first exposure ever to scientific research. I'm incredibly grateful to have had that opportunity, as I wouldn't know where I'd be today if it weren't for the research experience in his lab. In addition, I would also like to thank Emanual Carrilho, who served as my undergraduate advisor during my summer international research program in Brazil. Through my experience in his lab, I realized my passion of wanting to do research with impact in human health, which, I didn't realize at the time, was possible unless you were a medical doctor.

Additionally, I'm so grateful to be continually surrounded by so many friends from CSULA. These friends include Valerie Phan, Justin Lee, Christina Winnoken, Samantha Cheng, Sharon Zhu, Marilyn Fabricante, Magali Cox-Espinosa (previously Magali Espinosa), Kevin Parducho, John Roferos, Jesse Garcia Castillo, and Jasmine Diep. Valerie, Justin, and Christina are my dedicated foodie/adventure crew, as we're always out and about trying the newest and most interesting restaurants and experiences. However, they also are wonderful, committed friends, who are always willing to do what they can to support me in any stage of my life. Samantha, Sharon, Marilyn, Magali, Kevin, John, Jessie, and Jasmine are friends who I first met through CSULA's Honors College. They were all fun and engaging friends through my undergraduate years, introducing me to new sights and experiences. We shared

many things with each other, many of which wouldn't be shared unless we were all very tight-knit friends. Even though we're all pursuing different paths across different areas of the US, I'm glad we're still able to come together every once in a while, whether to enjoy a good meal or to celebrate one of our big successes. I'm glad and grateful to still be friends with all of you.

Upon coming to Caltech, I was fortunate to quickly become friends with two wonderful people, Jeremy Tran and Joey Messinger. They remained my closest friends throughout my graduate years, even though we're currently located in different states throughout the US. Throughout the COVID-19 pandemic, we've remotely dined together through monthly themed Zoom lunches and played games virtually every Friday night, which introduced me to the world of speedrunning games. Even before the pandemic, we would always do things together such as food adventures, writing parties for fellowships and candidacy, planning for Chemistry social events, board game nights, pumpkin carvings, and other fun activities my mind cannot instantly recall. In addition to the fun stuff, they were always there when I needed them most, especially when I've had a bad day in lab and needed to turn to someone just for support. I'm so fortunate to have been such close friends with both of you throughout my graduate years.

Throughout my time in graduate school, I had the opportunity to participate in various Caltech groups and organizations outside of my lab work. When I first joined Caltech, one of the first groups I became involved in was CCE's 'Big Sib-Little Sib' mentoring program. Through the program, I was matched to Amy Torrens (previously Amy McCarthy), who served as my 'Big Sib' during my time at Caltech. Amy was an amazing Big Sib—we would frequently get coffee and she would answer my questions and concerns about being a graduate student at Caltech. As I learned quickly, being a graduate student at Caltech isn't easy, but Amy would help ease my concerns and provide encouragement throughout my graduate career.

Another organization I was fortunate to be directly involved in is the Diversity in Chemistry Initiative (DICI). DICI's overall mission is to provide support and celebrate the successes of underrepresented students in Chemistry at Caltech. Additionally, it provides a safe space for students to share their experiences, both good and bad. DICI provided me with that outlet where I could relate to the struggles of other folks at Caltech, and it provided me with a community that supported me throughout my academic journey. It also gave me an opportunity to engage in my passions of outreach and mentorship, and I was fortunate to have been involved in the planning and participation of various outreach events through DICI. Thanks to folks in DICI, I had an opportunity to work closely with many people who've supported me throughout my graduate years, including Trixia Buscagan, David Cagan, Javier Fajardo Jr., Krystal Vasquez, Stephanie Threatt, Joel Monroy, Jean Badroos, and Doug Rees.

In addition to DICI, I was fortunate to be involved in the Caltech Glee Club and Chamber Singers, both of which were directed by Nancy Sulahian. I had never done choral singing prior to graduate school, but Nancy was willing to take me into the Glee Club and allow me to build my choral skills among a community of both novice and experienced choral singers. I spent 4 years in Glee Club, during which I discovered I was a soprano (great for singing very high notes as a form of stress release) and learned how to sing within a choral group. Nancy also allowed me to challenge myself as a choral singer through joining Chamber Singers, and boy, was I challenged. Through Chamber Singers, I learned how to become a more confident and skilled singer, as there are much fewer people and less accompaniment support to ensure you'll sing the right notes. However, Chamber Singers provided fun opportunities to sing outside of concerts, such as performing during the Athenaeum's Holiday Gala every December. I didn't realize how much I'd enjoy choral singing until I had joined Glee Club and Chamber Singers, and I hope to continue participating in choral groups outside of Caltech.

Another group I participated in was Caltech's Meditation Mob, which was led by Lee Coleman. Many people think meditation is very easy—sit, monitor your breath, and move

on. However, it's really not that easy when done in practice, and Lee really helps in guiding you through the process. I remember the first meditation session I attended, and to me, it was difficult. Being mindful and focusing on the breath can seem easy at first, but then once you think you got a handle of it, your mind instantly takes over and thinks about some random memory that occurred either 5 weeks ago, 5 hours ago, or something happening in the near future. However, Lee really walks you through the process, always reminding us on the objective, but also acknowledging that it's okay if your mind wanders. In addition, Lee and the Meditation Mob focused on various themes aside from mindful breathing, such as loving-kindness meditation, music-based meditations, or even meditations that reflect on recent events. I'm incredibly thankful to Lee for helping me become a more mindful individual, and I always resort to meditation to help calm myself whenever situations either become overwhelming or out of my control.

One unique community I want to thank is the Pasadena Pokemon Go Community on Slack. When Pokemon Go made its worldwide debut on July 6, 2016, I instantly picked up the game, but I went on a hiatus around September 2016 because the game quickly became bland. However, I slowly picked up the game again around January 2018, and I was fortunate to learn that there was a community in Pasadena to connect players with the game. Fast forward almost 3 years later, and I'm now one of the admins who helps manage the Pasadena Pokemon Go Slack group, which consists of over 2000 members to date. I not only got to meet other fellow Caltech players, but I got to meet and become friends with so many members of the greater Pasadena community that I likely wouldn't have met otherwise. Thank you all for the wonderful community days, raid trains, special events, and go fests. They all brought joy throughout my graduate years, and I'm absolutely grateful for meeting an amazing group of people.

During my years of graduate school, I had always lived with roommates. Roommates are definitely a hit-or-miss, but I'm incredibly grateful to my most recent roommates: Nancy Mattazaro, Iryna Chatila, and Maria Chatila. We started off as strangers, but we quickly warmed to each other's company. I will always appreciate the nights where we simply

ordered pizza together and played games because why not. I also enjoyed our tradition of celebrating each other's birthdays; they made for a lovely sentiment, especially during the heart of the COVID-19 pandemic. Furthermore, I cannot forget the housecats, Jac and Carmel, who lighten up my life. While they were not my pets, they are the closest I have ever gotten to having my own pets. Despite learning that I am allergic to cats, I was willing to brave through my cat allergies just to play and interact with them.

Last but not least, I want to thank Catherine Villanueva and Francis Arrastia, my Mom and Dad, for the sacrifices, support, and encouragement not only through my graduate career but throughout my entire life. From the moment I majored in Chemistry in undergrad, you knew that it would be hard for both of you to directly support me because it was in a field that seemed relatively foreign to you. Even though you didn't always understand what I did, you continued to be there for me for every step along the way. I love you both very much and I'm excited to say that I finally completed my PhD!!

ABSTRACT

Across eukaryotic cells, DNA from each nucleus is organized in three dimensions in order to help regulate transcriptional activity. Decades of chromosome capture technologies have revealed fundamental chromatin structures, providing information about how DNA is assembled genome-wide. The majority of these methods utilize direct physical ligation of DNA molecules to generate pairwise interactions, which have provided information about short-range interactions and intra-chromosomal structures. Recent technologies have moved toward identifying multiple DNA interactions simultaneously without physical ligation of DNA molecules, revealing information about long-range interactions and inter-chromosomal structures. One of the biggest limitations of these methods is that they only study DNA organization in bulk, which misses the heterogeneity of chromosomal structures at the single-cell level. As a result, single-cell chromosome capture methods have been developed to begin probing into the cell-to-cell variability of DNA organization and answer long-standing questions regarding single-cell structure. However, single-cell methods are currently limited to identifying low-resolution, intra-chromosomal DNA interactions with few numbers of cells. This creates a need for an improved, high-throughput single-cell method that can capture high-resolution structures and simultaneous mapping of both intra- and inter-chromosomal interactions to better elucidate single-cell DNA organization. In this thesis, we describe the development of 'single-cell split-pool recognition of interactions by tag extension' (scSPRITE), a single-cell chromosome capture method that allows for mapping of high-resolution, intra- and inter-chromosomal structures across thousands of cells. Through scSPRITE, we were not only able to reveal fundamental information about single-cell DNA organizations, but we can also quantitatively measure the variability of DNA interactions from cell to cell.

PUBLISHED CONTENT AND CONTRIBUTIONS

M. V. Arrastia*, J. W. Jachowicz*, N. Ollikainen, M. S. Curtis, C. Lai, S. A. Quinodoz, D. A. Selck, R. F. Ismagilov, M. Guttman (2021). "Single-cell measurement of higher-order 3D genome organization with scSPRITE." *Nature Biotechnology*. DOI: 10.1038/s41587-021-00998-1 [* These authors contributed equally]

M.V.A. Contributed in scSPRITE method development and optimization (Figure 1); performed scSPRITE experiment to generate data used throughout the paper (Figure 1-5); performed mouse-human mixing experiment (Figure 1); generated the ensemble and single-cell heatmaps for validation of method + higher order structure (Figure 1, 2, 3, 4, S2, S3, S4); generated table and plot denoting differences in contacts between scSPRITE and scHi-C methods (Figure S2); identified regions for TAD heterogeneity (Nanog-SE; Tbx3-Lhx5; AB compartment in chr4); generated scripts and plotted heatmaps to demonstrate these differences (Figure 4, 5, S4, S5); wrote and edited the manuscript.

J.W.J. Contributed conceptually in scSPRITE method optimization (Figure 1); designed mouse-human mixing experiment and cultured cells for it (Figure 1); Analysis and plotting: generated cartoon and comparison of contacts and reads between scSPRITE and scHi-C (Figure 2); generated histograms & TAD clustering maps using the single cell TAD detection scores (Figure 4, S4); contributed conceptually to figure design, heatmaps and plots generation (Figure 1-5); contributed to discussions & biological interpretations (Figure 4, 5, S4, S5); compiled the figures (Figure 1-5); wrote and edited the manuscript.

N.O. Defined and wrote the script for the detection score analysis to quantify single cell genomic structures (Figures 2, 3, 4, S2, S3, S4); performed the virtual 4C analysis for the plot in (Figure 5;) performed cworld analysis to generate insulation scores and annotations for A and B compartments; contributed to writing the manuscript.

M.S.C. Contributed in optimizing crosslinking conditions for scSPRITE (inferred in Fig 1); contributed to optimizing in-nuclei biochemical processes (nuclei isolation, in-nuclei digestion, & dA-tailing) (Figure 1); was the major contributor in developing

the scSPRITE in-nuclei barcoding workflow and nuclei sonication conditions (Figure 1).

C.A.L. Wrote the pipeline to identify cell-specific barcodes and to group complexes containing the same cell-specific barcode from sequencing data.

S.A.Q. Contributed to experiments for scSPRITE method development and validation (Figure 1).

D.A.S. Conceptualized the idea for scSPRITE; contributed to scSPRITE method development (nuclei isolation, in-nuclei digestion & dA-tailing, in-nuclei barcoding workflow) (Figure 1).

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS AND/OR TABLES

Number *Page*

To my Mom and Dad

Even through the darkest moments, your unconditional love, support, and encouragement continually shines light upon me knowing you're there every step of the way.

Chapter 1

INTRODUCTION

Evolution of chromatin capture methods to illustrate chromosomal structures

Across eukaryotic organisms, every nucleus in every cell contains the same DNA sequence, which encodes for all the genes necessary to regulate everyday cell activity. In humans, this DNA sequence contains 3.2 x 10^9 base pairs (3.2 Mb), which would measure about 2 m in length if stretched out. However, the human nucleus is about 10 µm in diameter, which is 200,000 times smaller than the size of human DNA. As a result, DNA needs to be able to fold itself to not only fit inside a nucleus, but also in such a way so genes can be transcribed normally.

Over the past few decades, new technologies have been developed to elucidate how DNA is folded and organized inside nuclei to better understand how these folding events help regulate gene expression. The bulk of these technologies developed to illustrate DNA structures is through a series of assays called chromosome capture methods, or "C-methods" for short. The first of these C-methods to be developed was called 'chromosome conformation capture' (3C) in 2002[1]. Through 3C, one could detect spatially proximal DNA molecules, generally promoter and enhancer regions, through a polymerase chain reaction (PCR). However, 3C is very limited in its ability to detect multiple DNA interactions simultaneously because the method requires *a priori* knowledge of DNA regions in order to develop specific primer sequences for PCR. Later C-methods such as 'chromosome conformation capture on-chip' (4C)[2] and 'chromosome conformation capture carbon copy' (5C)[3] have evolved to better capture and analyze DNA interactions. In contrast to 3C, 4C and 5C methods are able to capture interactions between multiple loci simultaneously, allowing for the reconstruction of more complex DNA interactions. Additionally, their analyses move toward microarray or sequencing-based approaches, allowing for more high-

throughput analyses. However, these methods are greatly limited in their ability to provide unbiased, genome-wide analysis of chromosome architecture.

The biggest advancement in C-methods came in 2009 upon the development of Hi-C. In contrast to the prior C-methods, Hi-C allows for genome-wide identification of chromatin interactions[4]. Through Hi-C, new features of DNA organization were discovered, such as the division of DNA into two compartments, termed A- and B-compartments, corresponding to open and closed chromatin, respectively, and the formation of small (~100 kb–1 Mb)[4], condensed units consisting of highly enriched DNA interactions called topologically associating domains (TADs)[5]. Hi-C made it possible to study genome-wide DNA interactions to reveal fundamental features of DNA organization across prokaryotic and other eukaryotic cell types (e.g. *Drosophila*[6], plants[7, 8], yeast[9, 10]), across stages of cell development[11-14] and cell cycle[15], and between healthy and diseased cells[16-19] to reveal differences in chromatin architecture.

To this point, the way most of the C-methods capture DNA interactions is through a direct ligation event between two spatially proximal DNA molecules, usually referred to as proximity ligation. This event creates a pairwise interaction, resulting in a chimera DNA sequence; through sequencing, the chimera DNA sequence allows one to identify which two regions of DNA were spatially close together at the time of ligation. This approach captures DNA interactions over short distances, but misses many long-range DNA interactions, including inter-chromosomal DNA interactions.

Recent methods have begun to move away from proximity ligation-based approaches in order to better elucidate chromatin architecture. The first method to move away from direct ligation events between DNA molecules was a method called ‘genome architecture mapping’ (GAM)[20]. With GAM, cryosectioning is performed to generate multiple cross-sections of nuclei, followed by laser microdissection to isolate a specific nuclear region within each cross-section. DNA loci within each nuclear region get amplified, which allows for the reconstruction of 3D DNA interactions without the need for proximity ligation. Another non-proximity ligation-based method is call ‘split-pool recognition of interactions by tag

extension' (SPRITE)[21]. SPRITE utilizes split-pool barcoding to add a combinatorial barcode sequence to all DNA molecules within the same crosslinked DNA-protein complex, which not only captures DNA molecules directly next to each other, but also captures DNA molecules interacting over large genomic distances. As a result, SPRITE is able to identify hubs of inter-chromosomal interaction, and these hubs tend to cluster around nuclear bodies.

Advances in single-cell chromosome capture approaches

While the advances in previous methods have greatly expanded our knowledge about DNA architecture, a big limitation that is shared across the methods mentioned previously is that they study aspects of chromatin organization in bulk. Bulk measurements have been useful in understanding general features about DNA organization, but they generally miss rare DNA contacts or the heterogeneity of DNA interactions from cell to cell. In order to better elucidate these characteristics, single-cell chromosome capture methods were developed.

The most prominent single-cell method to date is single-cell Hi-C (scHi-C)[15, 22, 23]. The first development of scHi-C emerged in 2013[22]. scHi-C revealed the existence of chromosome territories and A/B compartments at the level of single cells, and also began to reveal that single pairwise contacts occur as stochastic events. These events were not possible to detect in bulk-based methods because the information is lost when averaging contacts over a population of cells. However, it has been generally difficult to generate broad conclusions about single-cell chromosomal structure because scHi-C methods yield sparse, intra-chromosomal datasets over low resolutions. In addition, TAD structures were unable to be captured by scHi-C, raising the question of whether these structures are actual units of DNA organization or artifacts that arise from bulk measurements. Like many of their C-method predecessors, scHi-C utilizes proximity ligation to capture DNA interactions in single cells, but only a subset of all possible DNA interactions can be identified because of its inability to capture multiple DNA interactions simultaneously. As a result, high-resolution, genome-wide measurements of single-cell chromatin interactions have been difficult to attain because of the limitations presented by proximity ligation.

Thesis outline

In this thesis, to provide an improved overview of single-cell DNA organization, I illustrate the development of a novel, single-cell method called single-cell split-pool recognition of interactions by tag extension (scSPRITE)[24]. Through scSPRITE, we are now able to measure all possible types of chromosomal structures, ranging from chromosome territories and A/B compartments to structures that were not well-studied by previous single-cell methods, such as TADs and inter-chromosomal hub interactions. In addition, scSPRITE is able to measure DNA interactions from thousands of cells simultaneously, allowing us to quantify the heterogeneity of interactions from cell to cell.

References

1. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306-1311 (2002).
2. Simonis, M. et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). *Nature Genetics* **38**, 1348-1354 (2006).
3. Dostie, J. et al. Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Research* **16**, 1299-1309 (2006).
4. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289 (2009).
5. Dixon, J.R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380 (2012).
6. Sexton, T. et al. Three-dimensional folding and functional organization principles of the drosophila genome. *Cell* **148**, 458-472 (2012).
7. Dong, P. et al. 3D chromatin architecture of large plant genomes determined by local a/b compartments. *Molecular Plant* **10**, 1497-1509 (2017).
8. Dong, Q. et al. Genome-wide Hi-C analysis reveals extensive hierarchical chromatin interactions in rice. *The Plant Journal* **94**, 1141-1156 (2018).
9. Kim, S. et al. The dynamic three-dimensional organization of the diploid yeast genome. *eLife* **6**, e23623 (2017).
10. Muller, H. et al. Characterizing meiotic chromosomes' structure and pairing using a designer sequence optimized for Hi-C. *Molecular Systems Biology* **14**, e8293 (2018).
11. Dixon, J.R. et al. Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331-336 (2015).
12. Ke, Y. et al. 3D chromatin structures of mature gametes and structural reprogramming during mammalian embryogenesis. *Cell* **170**, 367-381.e320 (2017).

13. Du, Z. et al. Allelic reprogramming of 3D chromatin architecture during early mammalian development. *Nature* **547**, 232-235 (2017).
14. Bonev, B. et al. Multiscale 3D genome rewiring during mouse neural development. *Cell* **171**, 557-572.e524 (2017).
15. Nagano, T. et al. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* **547**, 61-67 (2017).
16. Taberlay, P.C. et al. Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. *Genome Research* (2016).
17. Vilarrasa-Blasi, R. et al. Dynamics of genome architecture and chromatin function during human B cell differentiation and neoplastic transformation. *Nature Communications* **12**, 651 (2021).
18. Harewood, L. et al. Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. *Genome Biology* **18**, 125 (2017).
19. Yang, L. et al. 3D genome alterations associated with dysregulated HOXA13 expression in high-risk T-lineage acute lymphoblastic leukemia. *Nature Communications* **12**, 3708 (2021).
20. Beagrie, R.A. et al. Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* **543**, 519-524 (2017).
21. Quinodoz, S.A. et al. Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell* **174**, 744-757.e724 (2018).
22. Nagano, T. et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* **502**, 59-64 (2013).
23. Ramani, V. et al. Massively multiplex single-cell Hi-C. *Nature Methods* **14**, 263-266 (2017).
24. Arrastia, M.V. et al. Single-cell measurement of higher-order 3D genome organization with scSPRITE. *Nature Biotechnology* (2021).

Chapter 2

SINGLE-CELL MEASUREMENT OF HIGHER-ORDER 3D GENOME ORGANIZATION WITH SCSPRITE

Content published initially in Nature Biotechnology: M. V. Arrastia, J. W. Jachowicz*, N. Ollikainen, M. S. Curtis, C. Lai, S. A. Quinodoz, D. A. Selck, R. F. Ismagilov, M. Guttman (2021). DOI: 10.1038/s41587-021-00998-1*

Abstract

Although three-dimensional (3D) genome organization is central to many aspects of nuclear function, it has been difficult to measure at the single-cell level. To address this, we developed 'single-cell split-pool recognition of interactions by tag extension' (scSPRITE). scSPRITE uses split-and-pool barcoding to tag DNA fragments in the same nucleus and their 3D spatial arrangement. Because scSPRITE measures multiway DNA contacts, it generates higher-resolution maps within an individual cell than can be achieved by proximity ligation. We applied scSPRITE to thousands of mouse embryonic stem cells and detected known genome structures, including chromosome territories, active and inactive compartments, and topologically associating domains (TADs) as well as long-range inter-chromosomal structures organized around various nuclear bodies. We observe that these structures exhibit different levels of heterogeneity across the population, with TADs representing dynamic units of genome organization across cells. We expect that scSPRITE will be a critical tool for studying genome structure within heterogeneous populations.

Introduction

In eukaryotes, linear DNA is packaged in a three-dimensional (3D) arrangement in the nucleus. This includes organization of DNA regions from the same chromosome (chromosome territories)[1] which are further subdivided into megabase-sized, self-associating topologically associating domains (TADs)[2, 3] based on gene activity (active/inactive or A/B compartments)[1] and local interactions between regulatory elements (enhancer-promoter loops)[4-6]. Additionally, DNA regions from multiple chromosomes are organized around nuclear bodies that form higher-order structural units[7, 8].

Genome organization in a single nucleus affects various nuclear functions, including DNA replication[9], transcription[5, 10], and RNA processing[11, 12]. Indeed, genome structure is known to dynamically change between cell types and in individual cells across time to reflect differences in biological states[5, 13, 14]. For example, during the cell cycle, the DNA structure undergoes dramatic rearrangement from open chromatin during interphase to highly condensed metaphase chromosomes[15-17]. Similarly, gene expression levels are heterogeneous among populations of cells[18, 19], suggesting that there may be differences in enhancer-promoter contacts present in individual cells in the population.

Currently, most methods used to study nuclear organization measure ensemble structures across millions of cells and can obscure critical information about the genome organization of any given cell. For example, measuring cells across the cell cycle and averaging their DNA contacts would mask cell-cycle dependent dynamics. Additionally, several studies have shown that observation of genome structures such as TADs[13-17, 20] in single cells do not always match structures predicted from ensemble measurements[1, 3, 21]. Accordingly, genome organization observed in bulk assays may not accurately reflect specific structures that exist within biological populations.

The two main techniques for measuring genome architecture of single-cells are microscopy and single-cell HiC (scHi-C). Microscopy provides the capability to study a broad range of genomic interactions in single cells, but is generally limited to measurements of a small number of loci simultaneously[13, 14, 20] and does not provide a genome-wide view. In contrast,

scHi-C provides a genome-wide view of nuclear structure in single cells, but it requires specialized equipment (e.g. robotics), generates data for low cell numbers, and is limited to low-resolution structures (~10Mb resolution/cell)[15, 16, 22]. Additionally, because scHi-C relies on proximity-ligation to measure interactions, it has limited ability to capture long-range and higher-order interactions, such as those organized around nuclear bodies[7, 23].

To address these technological gaps, we developed single-cell split-pool recognition of interactions by tag extension (scSPRITE) to provide comprehensive, high-resolution genome-wide maps of DNA structure from thousands of single cells. scSPRITE measures both inter- and intra-chromosomal interactions and dramatically increases the number of detected DNA contacts per cell relative to existing methods. To demonstrate its utility, we measured 3D genome structures in 1,000 individual mESC nuclei and observed chromosome territories, A/B compartments, and TADs in hundreds of single nuclei. We identified higher-order structures in hundreds of single cells, including inter-chromosomal contacts around centromeres, the nucleolus, and nuclear speckles. Notably, we identified cell-to-cell heterogeneity in mESC genome structures at different levels of resolution, including at promoter-enhancer contacts of the key pluripotency gene, Nanog. Together, these observations demonstrate that scSPRITE accurately measures genome structure and provides insights into genome organization. We expect that this approach will enable future studies examining the relationship between genome organization and nuclear function in individual cells.

RESULTS

scSPRITE maps 3D structure in thousands of individual cells

To understand 3D genome organization in individual cells, we extended our previously described SPRITE protocol[7] to enable single cell measurements. Single cell SPRITE (scSPRITE) works as follows: we dissociate cells into a single cell suspension, crosslink DNA and protein complexes *in situ*, isolate and permeabilize nuclei, digest DNA using a restriction enzyme, and perform two sets of split-and-pool barcoding to (i) tag DNA fragments contained in the same nucleus and (ii) tag the 3D spatial arrangement of these fragments (Figure 1a).

To map all DNA fragments originating from one nucleus, we performed split-and-pool barcoding to generate a unique cell-specific barcode (cell-barcode) for all DNA molecules contained in a single nucleus. Briefly, we distributed permeabilized nuclei across a 96-well plate (~200,000 nuclei) where each well contained a unique DNA barcode tag, and performed ligation such that all DNA molecules in the same nucleus were labeled with the same tag. We then pooled nuclei and repeated the split-and-pool process twice more to ensure that the number of barcode combinations (96^3=884,736) exceeded the cell number (see Methods). Because single nuclei can form aggregates in suspension, we filtered nuclei to remove potential clumps before proceeding to the next step (Figure S1a).

To verify that this approach accurately tags DNA contained in a single nucleus, we tested this first set of split-and-pool barcoding in permeabilized nuclei in a mixed population of human (HEK293T) and mouse (mESC) cells. After split-pool barcoding and sequencing, we clustered reads into groups based on their cell barcodes and computed the percentage of reads that aligned exclusively to the mouse or human genome (see Methods). We found that only 3.4% of cells contained reads from both species (Figures 1b, S1b), indicating that most cell-barcodes represent single cells. Because we cannot identify collisions that lead to mixing in the same species, we extrapolate a total collision rate (~10%) from the detected collisions.

Having developed an approach to accurately tag DNA in a single nucleus, we next sought to map these DNA fragments relative to each other in 3D space. To do this, we withdrew a small fraction of the single cell-tagged nuclei (~1,500 nuclei) and sonicated them to generate spatial clusters of chromatin. We then performed three additional rounds of split-and-pool barcoding, such that all DNA fragments contained in a spatial cluster obtained the same barcode combinations, while molecules in distinct spatial clusters obtained different combinations. After sequencing, we identified DNA molecules within the same spatial complex by matching all six barcode sequences and all complexes arising from the same nucleus by matching the first three barcodes (Figures 1a, S1c, see Methods).

To validate the method, we applied scSPRITE to mESCs because their genome structure has been extensively studied[3, 15] and they display known functional heterogeneity[17, 24-26]. We sequenced ~1,500 single cells and analytically excluded cell-barcodes that were likely to represent cell aggregates using the detected collision rates measured from the previously described mixing experiment (Figures 1c, S1b). To focus on the most informative single cells, we restricted our analysis to the 1,000 cells containing the highest number of spatial clusters per cell (see Methods).

To confirm that spatial barcoding in scSPRITE accurately measures known genome structures, we merged individual cell-barcodes from scSPRITE (referred to as ensemble scSPRITE) and compared heatmaps to those previously generated by bulk SPRITE in mouse ES cells[7] (Figure 1d). We found that these maps are highly comparable across all levels of resolution (Pearson correlation r=0.92, 1 Mb genome-wide; r=0.97, 200 kb on chr2; r=0.95, 40 kb across chr6:48-54 Mb).

Together, our results demonstrate that scSPRITE tags single cells with minimal collisions and accurately measures 3D organization at different levels of resolution. While we analyzed 1,500 single cells in this experiment, the number of cells analyzed by scSPRITE can be adjusted by modifying the number of rounds of split-and-pool barcoding such that the number of barcode combinations exceeds the number of single cells (>100-fold excess, see Methods).

scSPRITE measures multiway interactions in single cells

Because each individual cell contains a single genome and contacts detected in multiple cells cannot be pooled together (as they are in bulk measurements), single cell genome structure methods need to maximize the number of contacts detected in each cell. This is the main challenge and limitation for all single cell genomic methods.

Currently, existing single cell genome structure methods (e.g. scHi-C) utilize proximity ligation and are therefore limited to measuring pairwise DNA contacts[16, 22, 27-30]. While these measurements are averaged across multiple cells, this is not possible in a single cell because a specific DNA region can only be measured once per allele. Accordingly, even with perfect efficiency, pairwise methods would be unable to measure all possible contacts present in a given structure (Figures 1e, S1e). In contrast, SPRITE captures multiway contacts among DNA molecules, which dramatically increases the structural resolution that can be obtained for an individual cell. This is because the maximal number of interactions that can be captured increases quadratically with the size of a complex[23] (Figure S1d). For example, if a crosslinked complex contains four DNA fragments, the maximum number of contacts that can be observed by pairwise methods is two, whereas the maximal number of pairwise contacts that can be identified with multiway contacts is six (Figure S1e).

Indeed, we observe an increase in the number of pairwise contacts detected for each cell using scSPRITE (average of 34,992,080/cell) compared to scHi-C[16] (average of 375,470/cell) even though the number of sequencing reads per cell is ~10-fold lower for scSPRITE (average of 83,318/cell) than for scHi-C (average of 751,172/cell) (Figure 1f, Supplemental Note 1). We observe uniform coverage across all 1 Mb bins in virtually all cells and across all 100 kb bins in more than 80% of cells (Figure 1g) with almost no bias towards any chromosome (with the exception of chromosome 8 due to a trisomy in our cell line, see Methods) (Figure 1h). Notably, we observe low variability in genomic coverage across the analyzed cells (median MAD = 14, MAD range = 0-49, median = 35, Figures S1f, S1g).

scSPRITE detects chromosome territories and compartments

To determine which DNA structures can be observed in single cells, we generated DNA contact maps from each of the 1000 individual cells. For every structure identified in the ensemble data, we computed a normalized detection score that reflects how well each single cell contact map resembles this structure compared to a randomized contact map. Briefly, for each structure we calculated an observed detection score, which defines whether each pair of genomic bins in a structure were in contact. A cell that contains all possible pairwise contacts in a given structure would have a detection score of 1, whereas a cell containing none of the expected pairwise contacts would have a detection score of -1. We normalized these observed scores to a distribution of scores generated by randomly permuting the locations of each structure (see Methods, Supplemental Note 4).

We focused on genomic structures that were previously reported to occur in single cells—chromosome territories and A/B compartments[1]. Chromosome territories are structures containing high frequencies of intra-chromosomal interactions with minimal inter-chromosomal interactions (Figure 2a). First, we looked at the contacts between chromosome 1 (chr1) and chr2 and detected clear separation of contacts into chromosome territories in both the ensemble data (Figure 2a) and in >75% of single cells (score>0, Figures 2b, S2a). Next, we quantified detection scores for every pair of chromosomes in every cell (Figure S2b, Supplemental Table 1). Although some chromosomes show stronger self-interactions than others, all chromosomes organize into territories (avg. score = 0.08, SD = 0.06, Figures 2c, S2c, see Methods for chr8). We observe that 95% of cells contain well-defined territories (Figures 2d, 2e) and only a small fraction of cells (<50 cells) do not contain observable chromosome territories (Figures 2d, S2d), and may reflect cell states containing distinct organization, such as mitotic chromosomes.

Genomes are further divided into A/B compartments, which are intra-chromosomal structures defined by open (A) or closed (B) chromatin state[1] (Figure 2f). To measure A/B compartment patterns in single cells, we first focused on a region on chr2 that has a well-defined B-A-B compartment switch observed in the ensemble scSPRITE data (Figure 2f).

We calculated the detection score for that region in individual cells and observed segregation of DNA into A/B compartments in >65% of single cells (score>0, Figures 2g, S2e). Next, using our ensemble data, we defined all regions that correspond to a compartment switch (B-A-B or A-B-A) genome-wide (224 regions, Supplemental Table 2) and quantified their detection scores for each cell (Figure S2f). We observed that individual regions are more variable in single cells than chromosome territories (avg. score = 0.03, SD = 0.06, Figures 2h, S2f), but are still present in ~95% of cells (Figure 2i). We looked more closely into three regions with different average detection scores (Region 1, score = 0.12, SD = 0.16; Region 2, score = -0.01, SD = 0.14; Region 3, score = -0.10, SD = 0.11) and observed that the variability in the A/B compartment structure in single cells is indeed higher for the regions with lower detection scores (Figures 2j, S2g) (i.e. Region 3>Region 2>Region 1). This suggests that the detection score metric that we developed is useful to identify cells and regions of variable structures. We observe a small detection bias towards active regions (A compartments, 45% of observed reads versus 39% expected reads) (Figure S2h).

Together, our results demonstrate that scSPRITE can detect known genomic interactions such as chromosome territories and A/B compartments in single cells and can be used to measure structural variability between individual cells.

Inter-chromosomal hubs are organized around nuclear bodies

The nucleus is further organized around various nuclear bodies that form higher-order inter-chromosomal contacts[7, 8]. These contacts have not been previously explored in single cells at the genome-wide scale because existing single-cell proximity-ligation methods are limited in their ability to detect inter-chromosomal contacts[16, 22, 23, 29]. scSPRITE measures, on average, an almost ten-fold increase in the proportion of inter-chromosomal contacts per cell than scHi-C (54% and 6%, respectively) (Figure 3a), which makes it a suitable method to study higher-order organization in the nucleus. We focused on three types of known inter-chromosomal structures: inactive regions associated with nucleoli, active chromatin around nuclear speckles, and centromeric and peri-centric regions organized into chromocenters.

Inactive DNA hubs are known to organize around the nucleolus[7], a nuclear body that is formed around transcription of ribosomal DNA (rDNA) regions[12]. In mESCs, regions on chr12, chr15, chr16, chr18, chr19 contain rDNA clusters that form Nucleolar Organizing Regions (NORs). We first explored contacts between two NOR-containing regions on two pairs of chromosomes (chr18/chr19 and chr12/chr19) that were previously reported to form strong interactions in mESCs[7]. We observed similar interaction patterns between these regions in the ensemble SPRITE data and in individual cells (score >0 in 54% and 61% of cells, respectively) (Figures 3b, 3c, S3a, S3b). We compared the frequencies of contacts detected by scSPRITE (specifically how often these two regions are in the same cluster) to the frequencies of their co-occurrence at the same nucleolus measured by microscopy (where the nucleolus is visualized by nucleolin immunostaining and DNA regions are visualized by DNA-FISH)[7]. We focused our analysis specifically on 1 Mb regions targeted by DNA-FISH probes (three NOR-containing chromosome pairs and two control chromosome pairs) and observed a strong correlation between these datasets ($R^2 = 0.88$, Figure S3c), indicating that single cell measurements generated by scSPRITE are comparable to those observed by microscopy. Similarly, we observed a strong correlation between scSPRITE and SPRITE data for these NOR regions ($R^2 = 0.88$, Figure S3d). To look at genome-wide interactions of NORs, we quantified the percent of single cells that contain each nucleolar contact (Figure S3e) and observed that on average, 38% of cells contained each nucleolar pair (Figure S3e). The most frequent contacts are formed between NORs on chr18 (3-10 Mb) and chr19 (29-37 Mb or 25-28 Mb) which are both observed in >50% of cells and the least frequent contacts are observed between chr15 (67-71 Mb) and chr18 (57-60 Mb) which are observed in <20% of cells (Figure S3e). In all cases, we observed that NORs interact more frequently than random non-NOR containing regions (Figure 3d).

Next, we looked at active nuclear hubs organized around nuclear speckles—structures enriched in pre-mRNA splicing factors[11, 31]. First, we focused on the previously reported inter-chromosomal interactions formed by precise regions of mouse chr2/chr4 and chr2/chr5[7] and observed these contacts in 53% and 38% of cells (score>0, Figures 3e, 3f, S3f, S3g). Next, we quantified the percent of single cells that contain each pair of interacting speckle

regions (Figure S3h). We detected speckle interactions in an average of 34% of cells (Figure S3h), with interactions between regions on chr4 (128-142 Mb) and chr5 (112-126 Mb) observed in >50% of cells and between chromosome 2 (117-181 Mb) and chromosome 13 (55-58 Mb) observed in <10% of cells (Figure S3h). When we calculated the frequency of contacts per 1 Mb bin of every interacting speckle region, we observed that most speckle regions interact more frequently than random regions but less frequently that NORs (Figure 3d).

Finally, we explored centromeric and peri-centromeric heterochromatin regions (PCH). Centromeres and peri-centromeres are long stretches of repetitive DNA essential for chromosome stability and segregation[32] , and have been shown to come into close proximity to form inter-chromosomal structures called chromocenters[32] (Figure 3g). Because PCH regions are not mapped in the genome, we focused our analysis on the first 10 Mb of each chromosome. First, we made single cell contact maps and calculated detection scores for two pairs of PCH regions (chr1/chr11 and chr4/chr11) (Figures 3g, 3h, S3i, S3j); we detected formation of these inter-chromosomal interactions in 54% and 80% of cells, respectively (score >0, Figures 3h, S3i, S3j). Next, we looked at genome-wide interactions of PCH regions and quantified the percent of single cells that contain each PCH contact (Figure S3k). We observed that on average, 49% of cells contained two different PCH-containing regions in close proximity. Notably, the PCH region of chr11 forms pairs with other PCH regions most frequently (80% of cells) and PCH region of chr14 interacts least frequently with other PCH regions (30% of cells) (Figure S3k). More generally, when we calculated the frequency of PCH interactions per each 1 Mb region of PCH, we observed that these regions form pairs more frequently than random regions of the same size (Figure 3d). We note that after size normalizations (see Methods), chromosomes that contained NORs displayed a higher contact frequency between their centromeric regions (Figure 3i), consistent with previous observations by microscopy[33, 34] .

The results of these analyses demonstrate that scSPRITE can capture various higher-order contacts reflecting inter-chromosomal interactions across multiple cells and involving structures of different sizes and transcriptional output (active versus inactive hubs). We note

that centromere-proximal and nucleolar contacts were not detectable even in the ensemble scHi-C data[16] (Figure S3l). Although the ensemble scHi-C[16] was able to identify speckle interactions, the single-cell interaction maps lacked information on these structures (Figure S3l).

TADs are heterogeneous across individual cells

Topologically associating domains (TADs) are intra-chromosomal structures in which contiguous regions of the genome have been shown to interact more with themselves than with surrounding regions[2, 3, 35]. However, these observations are mainly based on bulk measurements, and whether TADs exist in single cells has been debated[13, 15, 16, 20]. Specifically, it is unclear whether the inability to observe TADs in single cells reflects the technical limitation of current single cell methods (e.g. low-resolution structures) or if these DNA structures are not present in individual genomes. Because scSPRITE generates higher resolution structures in individual cells, we asked whether it can detect TADs in single cells.

We first defined all TADs present in mESC using the ensemble scSPRITE data (Supplemental Table 3), which are comparable to TADs defined from HiC data[3] (Pearson correlation r = 0.70, Figure S4a, S4b). We used these genomic coordinates to score each of these TADs in every single cell. First, we focused our analysis on a region of chromosome 4 (124.8-126.7 Mb) where we observed strong evidence for TADs in the ensemble scSPRITE dataset (Figure 4a). Using the genomic locations defined from the ensemble data, we detected TAD-like structures in >75% of single cells (score>0, Figures 4b, S4c), suggesting that most individual cells contain this specific TAD structure with the same boundaries.

To explore the heterogeneity of TAD structures in single cells, we performed two analyses. First, we looked at the average representation of all TADs in each cell by averaging the TAD detection score for each region (identified in the ensemble dataset) in each individual cell; we found that the majority of cells contain TADs (95% of cells with score>0, Figure 4c). Second, we explored whether individual TADs are more or less variable across individual

cells by averaging the TAD detection score for individual TADs across cells. We found that most TADs are highly variable between cells (65% of cells with score<0, Figure 4d) and noticed that highly variable TADs are not randomly distributed, but cluster in shared genomic regions (variable TAD regions; Figures 4e, S4d)

To explore these variable TAD regions, we focused on a specific example which showed a low detection score suggesting its structural variability (chr4: 38.5 – 43.6 Mb, average score across the three TADs identified in this region = 0.00, SD = 0.06, Figure S4f). We identified two groups of cells containing differences in genome organization at that region (Figure 4f). Specifically, we detected a population of cells that contain an alternative TAD that spans the boundary of the ensemble-defined A/B compartment (Figures 4f, S4f). When focusing exclusively on cells that contain this alternative TAD, we found that the A/B compartments defined in those cells are distinct from the ensemble population (Figures 4f, 4g). We confirmed that these distinct structural states are not explained by differences in cell cycle (Figure S4g) or by other major structural changes between these two groups of cells (Figure S4h). This suggests that this region is present in at least two distinct—and mutually exclusive—structural states in different cells in the population.

Together, our results demonstrate that the scSPRITE method can detect TAD-like genome organization in individual cells and identifies structural differences at the level of TADs in single cells. More studies are required to define if these cell-to-cell variabilities and region-to-region differences are functionally relevant and if they are characterized by other features like transcription, specific chromatin marks, or weak insulation boundaries (Supplemental Note 2).

scSPRITE detects heterogeneity across long-range contacts

We next asked if scSPRITE could detect structural changes that reflect biologically significant long-range DNA contacts, such as the interactions between promoters and super enhancers (SE) or between regions enriched in polycomb group proteins (PcGs)[36] . SE are

large domains enriched in H3K27 acetylation that are thought to modulate gene expression by forming loops with promoters[6]. Bulk genome-wide studies have shown that SE can form long- and short-range interactions with the same promoter[37-39], but it remains unclear whether these interactions occur simultaneously in the same cell. Similarly, DNA regions bound by PcGs have been shown to interact across long distances to regulate gene expression[36]; however it remains unclear how heterogenous these long-range interactions are in a population of cells.

We focused on two examples of long-range interactions in mESCs: (i) the Nanog locus, a key pluripotency factor in ES cells whose promoter interacts with multiple enhancers over a broad range of distances (up to 300 kb)[37, 40] (Figure 5a); and (ii) the Tbx3 locus, a transcription factor involved in the maintenance of pluripotency[41] whose locus interacts with another PcG enriched gene, Lhx5 (760 kb downstream) (Figure 5b).

We selected cells with coverage over the Nanog and Tbx3 regions of interest and split them into two groups based on whether we observed a contact between the target locus and the long-range enhancer (300kb upstream for Nanog, 760kb downstream for Tbx3) (Figure S5a). We computed the frequency of contacts between the target locus and all 40 kb bins for each group of cells (Figure 5c, 5d). We noticed that in the group with long-range interactions detected, short-range interactions were significantly weaker ($p<0.001$) and on average, three times less frequent (Figure 5c, 5d). Additionally, we observed that detected long-range interactions span across a TAD border identified in the ensemble dataset for both Nanog and Tbx3 examples (Figures S5a, S5d). We confirmed that the observed structural differences were not caused by technical differences (e.g. number of reads in each group of cells) (Figures S5b, S5e) or different cell cycle phases (Figures S5c, S5f).

Our results demonstrate that in cells where either the Nanog or Tbx3 locus contacts the long-range region, the locus is less likely to form a contact with the short-range region (and vice versa) (Figure 5e). Surprisingly, we detect long-range and short-range interactions in a similar number of cells, suggesting that both of these states are present at comparable frequencies in mESCs. Whether such heterogeneity is a more global occurrence or restricted

to specific loci (e.g. transcription factors regulating pluripotency), and what (if any) functional role these distinct structures might play remains to be determined (Supplemental Note 3).

DISCUSSION

Here we described scSPRITE, a method to generate high-resolution genome-wide maps of 3D DNA organization in thousands of single cells. scSPRITE expands the toolkit of genome-wide, single cell sequencing-based methods with an approach that enables high-resolution structural views across a broad spectrum of DNA interactions into high-throughput contact maps of the entire genome. In contrast to existing methods, scSPRITE does not require specialized equipment, techniques, or training, and provides increased resolution from a lower number of sequencing reads across a larger number of cells. Because of this, we expect that it will expand the availability of single cell genome structure measurements to any molecular biology laboratory. Additionally, we expect that scSPRITE can be scaled to work with as few as hundreds or as many as several thousands of cells simultaneously.

Our results reveal several novel insights about the heterogeneity of genome organization in mouse ES cells. Specifically, we detected long-range higher-order interactions of both active (nuclear speckle) and inactive (centromeres and nucleolar contacts) chromatin regions as well as heterogenous organization of TADs and enhancer-promoter contacts between individual single cells. We note that our experiments were performed in a population of mESCs cultured using a “2 inhibitor” (2i) cocktail that is thought to promote ground-state pluripotency and display more homogenous expression profiles across single cells[24, 26, 42] (Supplemental Note 3). Nonetheless, our results suggest that even in these conditions, nuclear organization can be heterogeneous. Whether these cell-to-cell differences in 3D structure impact gene expression or have other functional significance remains to be determined.

Although our initial study focused on mESCs, scSPRITE can be applied to different cell types or homogenized tissues that are composed of mixed cell populations. One of the current challenges with studying complex tissues (e.g. brain) or disease states (e.g. tumors) is the heterogeneity of their cellular composition. The application of scSPRITE to such cell populations will enable studies of intrinsically heterogeneous systems and provide an accurate global view of their 3D genome organization. Accordingly, we expect that scSPRITE will provide the field with a path toward understanding the relationship between 3D genome organization and genome function in single cells.

METHODS

Cell types and culture conditions

We developed scSPRITE using mouse and human cells, focusing primarily on mouse embryonic stem cells (mESCs) because their genome structure has been extensively studied[3, 15].

We used a male ES cell line (bsps derived from V6.5 ES cell line, provided by K. Plath) and cultured them in serum-free 2i/LIF medium as previously described[43]. We suspect that this mES cell line displays trisomy in Chromosome 8 because the average number of reads aligning to chr8 is about 33% greater than the average number of reads across the other chromosomes (Figure 1h).

HEK293T, a female human embryonic kidney cell line transformed with the SV40 large T antigen was obtained from ATCC (#CRL-1573) and cultured in complete media consisting of DMEM (#11965092, GIBCO, Life Technologies; Carlsbad, CA) supplemented with 10% FBS (Seradigm Premium Grade HI FBS, VWR), 1X penicillin-streptomycin (GIBCO, Life Technologies), 1X MEM non-essential amino acids (GIBCO, Life Technologies), and 1 mM sodium pyruvate (GIBCO, Life Technologies), and maintained at 37°C under 5% CO2. For maintenance, 800,000 cells were seeded into 10 mL of complete media every 3-4 days in 10 cm plates.

Single Cell SPRITE protocol

Cell crosslinking

Media from mESCs was removed and washed once with 1X PBS. Cells on the 10 cm plates were then trypsinized using 2 mL of 0.025% Trypsin-EDTA (prewarmed to 37°C). Plates were incubated at 37°C for 5 min, and the trypsinized cells were mixed by pipetting to break up any clumps. We added 8 mL of pre-heated wash solution (DMEM/F12 + BSA, prewarmed to 37°C) to the plate to inactivate trypsin before transferring the cells to a conical tube. Cells were centrifuged at 330 g for 3 min, and the supernatant was discarded. Cells were washed once with 1X PBS at a ratio of 4 mL of PBS per 1 x 10^7 cells and centrifuged again at 330 g for 3 min. After the wash, 4 mL of 2 mM disuccinimidyl glutarate (DSG,Life Technologies, #20593, Carlsbad, CA) prepared in 1X PBS was added per 1 x 10^7 cells to the conical tube, and the solution was mixed thoroughly by pipetting to remove clumps. The cells in DSG solution were gently shaken for 45 min at room temperature. Following incubation with DSG, 200 µL of 2.5 M glycine were added per 1 mL of DSG solution previously added to quench the reaction, and the tube was gently shaking for 5 min at room temperature. Cells were then centrifuged at 1000 g for 4 min and the supernatant was discarded. Cells were washed with 1X PBS at a ratio of 4 mL of PBS per 1 x 10^7 cells and centrifuged again at 1000 g for 4 min. After the wash, 4 mL of 1% formaldehyde (16% (w/v) ampules, Life Technologies, #28908, Carlsbad, CA) prepared in pre-warmed (37°C) 1X PBS) was added per 1 x 10^7 cells to the conical tube and the solution was mixed thoroughly. The cells in formaldehyde solution were then gently shaking for 10 min at room temperature. Following incubation with formaldehyde, 200 µL of 2.5 M glycine was added per 1 mL of formaldehyde solution previously added to quench the reaction, and the tube was gently shaking for 5 min at room temperature. Cells were then centrifuged at 1000 g for 4 min and the supernatant was removed. Cells were twice washed with cold 1X PBS + 0.5% BSA (w/v) solution, and centrifugation was done at 4°C at 1000 g for 4 min. Following the washes, enough cold 1X PBS + 0.5% BSA solution was added to get a cell concentration of 5 x 10^6 cells/mL. Crosslinked cells were then aliquoted in new 1.5 mL low-bind Eppendorf tubes,

centrifuged (2000 g for 5 min) to remove the supernatant, and flash-frozen in liquid nitrogen. Cells were kept at -80°C until used for analyses.

Cell lysis and nuclei preparation

Crosslinked cells were thawed from -80°C and were kept on ice during the cell lysis procedures. Initially, 1.4 mL of lysis buffer #1 (50 mM HEPES pH 7.4, 1 mM EDTA pH 8.0, 1 mM EgTA pH 8.0, 140 mM NaCl, 0.25% TritonX-100, 0.5% IGEPAL CA-630, 10% glycerol, 1X proteinase inhibitor cocktail (PIC)) was added per 1 x 10^7 cells. The cell solution was mixed thoroughly before incubating on ice for 10 min. Cells were pelleted afterwards at 900 g for 8 min at 4°C, and the supernatant was removed. Following, 1.4 mL of lysis buffer #2 (10 mM Tris-HCl pH 8, 1.5 mM EDTA, 1.5 mM EgTA, 200mM NaCl, 1X PIC) was added per 1 x 10^7 cells. Again, the cell solution was mixed thoroughly before incubating on ice for 10 min. Cells were pelleted afterwards at 900 g for 9 min at 4°C, and the supernatant was removed. Afterwards, the cells were washed in 800 µL of 1.2X CutSmart solution (from 10X CutSmart stock (NEB, #B7204S, Ispwich, MA)) and pelleted at 900 g for 2 min. Supernatant was removed and a fresh 400 µL of 1.2X CutSmart solution was added carefully to not resuspend the pellet. Then, 6 µL of 20% SDS was added to the tube, and the cells were thoroughly resuspended. The cell solution was mixed on an Eppendorf ThermoMixer C at 1200 rpm for 60 min at 37°C to isolate nuclei. Next, 40 µL of 20% Triton X-100 was added to the same tube to quench the reaction, and the solution was left mixing on the same instrument at 1200 rpm for 60 min at 37°C. Lastly, 30 µL of 5000 U/mL HpyCH4V (NEB, #R0620L, Ispwich, MA) was added to the same tube to allow for DNA to be digested in-nuclei. In-nuclei digestion was performed for 4 h at 37°C while shaking at 1200 rpm. HpyCH4V is a 4-base pair restriction enzyme that performs blunt-end cutting at TGCA sequences. This particular enzyme was chosen since it was able to perform in-nuclei enzymatic restriction digestion and eliminated the need to perform any additional DNA strand repair steps after restriction digest. After 4 hours of restriction digest, the average DNA fragment size was 823 bp.

Following digestion, nuclei were pelleted at 900 g for 2 min, the supernatant was removed, and the nuclei washed three times with 1X PBS, 1 mM EDTA, 1 mM EgTA, and 0.1% Triton X-100 solution at 900 g for 2 min. Following the washes, the nuclei concentration was assessed by loading 6 µL of the solution into a disposable hemocytometer (4-Chip Disposable Hemocytometer, Bulldog Bio, #DHC-N420, Portsmouth, NH). After determining nuclei concentration, 5 x 10^5 nuclei were transferred by pipetting into a new 1.5 mL low-bind Eppendorf tube. In this new tube, 25 µL of dA-tail reaction buffer and 10 µL of Klenow Fragment were added to the nuclei (both reagents were part of NEBNext dA-Tailing Module (NEB, #E6053L, Ispwich, MA)). The tube was filled to 250 µL using nuclease-free H_2O and dA-tailing was performed in-nuclei at 37°C for 90 min while shaking at 1200 rpm. The reaction was then stopped with the addition of 200 µL of 1X PBS, 50 mM EDTA, 50 mM EgTA, and 0.1% Triton X-100. The nuclei pellet was spun down at 900 g for 2 min and washed twice using 400 µL of 1X PBS, 1 mM EDTA, 1 mM EgTA, and 0.1% Triton X-100 solution. Following the washes, the nuclei were resuspended in fresh 1X PBS, 1 mM EDTA, 1 mM EgTA, and 0.1% Triton X-100 solution, and nuclei concentration was determined again using the hemocytometer as described previously.

In-nuclei combinatorial barcoding

To uniquely identify DNA sequences originating from the same cell, combinatorial barcoding was performed in-nuclei (Figure 1a). In our specific experiments, we utilized three rounds of combinatorial barcoding in the following order: “DNA phosphate modified” (DPM), “odd” tagging, and “even” tagging (these tags are described in the original SPRITE paper[7]). The resulting tags were pre-loaded onto a 96-well plate, with each well containing 2.4 µL of a uniquely barcoded tag at a concentration of 45 µM. Nuclei previously dA-tailed were washed twice in a solution of 1X PBS, 0.1% Triton X-100, and 0.3% BSA (w/v), and nuclei concentration was reassessed using a hemocytometer, as described previously.

To perform in-nuclei barcoding, 2 x 10^5 nuclei were withdrawn and transferred into a new 1.5 mL low-bind Eppendorf tube, and filled to 1125 µL using a solution of 1X PBS, 0.1% Triton X-100, and 0.3% BSA (w/v). The nuclei solution was well-mixed before loading 11.2

µL of nuclei solution into each well of a 96-well plate. Each well was then supplemented with 6.4 µL of ligation mix (220 µL of 2X Instant Sticky Master Mix (NEB, #M0370, Ispwich, MA), 352 µL 5X Quick Ligase Buffer (NEB, #B6058S, Ispwich, MA), and 132 µL 1,2-Propanediol (Sigma, #398039, St. Louis, MO)). The 96-well plate was sealed after loading a ligation mix and was mixed on an Eppendorf ThermoMixer C at 20°C. The reaction was performed for 3 h while mixing at 1600 rpm for 30 s every 5 min.

After performing in-nuclei DNA ligation, 20 µL of 1X PBS, 50 mM EDTA, 50 mM EgTA, and 0.1% Triton X-100 solution was added to each well and incubated for 10 min at 20°C to stop the ligation reaction. Next, a solution of 80 µL of 1X PBS, 50 mM EDTA, 50 mM EgTA, and 0.1% Triton X-100 (w/v) was added to each well, and all the contents of the well plate were pooled together into a new 15 mL conical tube. The 96-well plate was washed once with a solution of 100 µL of 1X PBS, 50 mM EDTA, 50 mM EgTA, and 0.1% Triton X-100, and pooled together into the same conical tube. Nuclei were pelleted at 800 g for 10 min, and all but 1 mL of supernatant was removed from the tube. The nuclei were resuspended before transferring to a new 1.5 mL non-low bind Eppendorf tube. In the new Eppendorf tube, nuclei were washed twice with a solution of 500 µL of 1X PBS, 0.1% Triton X-100, and 0.3% BSA (w/v) at 900 g for 2 min. This in-nuclei ligation process was repeated two more times resulting in a total of three tags (the "DPM," "odd," and "even" tags) being ligated to DNA fragments.

Once the three rounds of in-nuclei barcoding process was completed, the nuclei were filtered through a 10 µm mesh filter (PluriStrainer, #43-10010-50, Spring Valley, CA) into a new 1.5 mL non-low bind Eppendorf tube to ensure we only isolated single cells (Figure S1a). Filtered nuclei were then pelleted at 900g for 2 min, and the supernatant was removed. Then nuclei were resuspended and washed twice in lysis buffer #3 (1.5 mM EDTA, 1.5 mM EgTA, 100 mM NaCl, 0.1% sodium deoxycholate, 0.5% sodium lauroyl sarcosinate) at 900g for 2 min. Next, the concentration of nuclei was determined and 1,500 nuclei were withdrawn and used in the following steps of the protocol.

Scaling the number of cells to analyze and determining the number of barcoding rounds

In our experiments, we utilized a final concentration of 1,500 individual nuclei and performed three rounds of barcoding to generate 96^3 (884,736) barcode combinations. This results in 590-fold excess barcode combinations to the number of cells analyzed and results in <1 expected cell "collisions" (where cells obtain the same complete barcode string). To provide some intuition on these numbers, we note that the probability that any two cells will have a "collision" is defined by a Poisson distribution with a mean (λ) defined by the number of cells divided by the number of barcode combinations. The probability of observing two or more cells with the same barcode in this distribution is defined as the p(x>1). The expected number of collisions is the number of measured cells multiplied by this probability of collision. Accordingly, analyzing 10,000 cells with 100-fold barcode excess (100,000 barcode combinations) would yield <1 expected cell collisions. Thus, the number of cells analyzed can be adjusted to enable the analysis of larger numbers (or smaller numbers) based on the needs of the application. Adjusting cell numbers may require adjusting the numbers of rounds of barcoding to enable accurate separation of individual cells. We recommend between 10-100-fold excess barcode combinations to the number of cells analyzed. The exact excess utilized depends on how many potential collisions would be tolerated in the final output.

Sonication

1500 nuclei were placed into a Covaris microtube-15 and filled to 15 µL using lysis buffer #3. The Covaris tube was placed in the Covaris M220 Focused-ultrasonicator (Covaris, Woburn, MA), and sonication was performed for 2 min under specific settings (water temperature 6°C, incident power 30W, duty cycle 3.3) to release DNA complexes from nuclei. The tube was then removed from the instrument and set on ice.

At this step of the protocol, it is important to proceed with all the sonicated nuclei as sampling them further will lead to a loss of nuclei fragments and will prevent the analysis of DNA

structure in single cells. We also recommend adjusting the sonicated number of cells to sequencing abilities in order to achieve satisfactory coverage per cell.

NHS (N-hydroxysuccinimide) beads coupling

After sonication, sample containing crosslinked DNA complexes was coupled to NHS-beads as previously described[7]. Briefly, NHS-Activated Magnetic Beads (Life Technologies, #88826, Carlsbad, CA) were activated for coupling. First, 600 µL of NHS-beads were withdrawn and placed in a 1.5 mL low-bind Eppendorf tube. The tube was placed on a DynaMag-2 magnet, and the supernatant was removed. The beads were washed once with 600 µL of ice-cold 1M HCl, and the supernatant was removed again and replaced with 600 µL of ice-cold 1X PBS. After removing 1X PBS, the beads were resuspended in 500 µL of 1X PBS + 0.1% SDS. Additionally, 85 µL of 1X PBS + 0.1% SDS was added to the previously sonicated nuclei solution, mixed, and added to the bead solution. The complexes were then coupled to NHS-beads on an Eppendorf ThermoCycler C overnight at 4°C while shaking at 1200 rpm. After coupling, the flowthrough was removed and 600 µL of 1M Tris-HCl pH 7.5, 0.5 mM EDTA, 0.5 mM EgTA, and 0.1% Triton X-100 was added to the beads to quench the remaining NHS groups; this was done at 4°C at 1200 rpm for 60 min. Once the beads were quenched, the flowthrough was removed, and the beads were washed twice in cold RLT2+ buffer (0.2% Sodium lauryl sarcosinate, 1 mM EDTA, 1 mM EgTA, 10 mM Tris-HCl pH 7.5, 0.1% Triton X-100, 0.1% NP-40, filled to the final volume with RLT (Qiagen, #79216, Valencia, CA)). This was followed by three washes in M2 buffer (50 mM NaCl, 20 mM Tris-HCl pH 7.5, 0.2% Triton X-100, 0.2% NP-40, 0.2% sodium deoxycholate). The beads were then resuspended in a mix of M2 buffer and H_2O (58% M2, 42% H_2O) to attain a total volume of 1125 µL of M2 buffer, H_2O, and beads.

Spatial barcoding/complex-specific barcoding

Next, spatial barcoding of the DNA complexes on beads was performed as described previously[7]. First, the bead solution was well-mixed and loaded into each well of a 96-well plate (11.2 µL of bead solution per well). Each well of the plate contained 2.4 µL of uniquely

barcoded tag at a concentration of 4.5 µM. Next, each well was supplemented with 6.4 µL of ligation mix (220 µL of 2X Instant Sticky Master Mix (NEB, #M0370, Ispwich, MA), 352 µL 5X Quick Ligase Buffer (NEB, #B6058S, Ispwich, MA), and 132 µL 1,2-Propanediol (Sigma, #398039, St. Louis, MO)). The 96-well plate was sealed after loading a ligation mix and was mixed on an Eppendorf ThermoMixer C at 20°C. The reaction was performed for 60 min with mixing at 1600 rpm for 30s every 5 min. Afterwards, the reaction was stopped by adding 60 µL of RLT2+ buffer to each well before pooling the solutions of each well into a 25 mL reservoir. Each well was then rinsed once with 100 µL of RLT2+ buffer to remove residual beads and pooled into the same 25 mL reservoir. The solution was then transferred to a 15 mL conical tube, which was then placed on a magnet to remove most of the RLT2+ buffer from the beads. With about 2 mL of RLT2+ buffer remaining, the beads were resuspended and transferred to a low-bind 1.5 mL Eppendorf tube, which was placed on a DynaMag-2 magnet to remove the remaining RLT2+ buffer. The beads were washed three times with 600 µL of M2 buffer. This process of split-pool barcoding on beads was repeated until the three additional tags were added. After the last round of split-pool barcoding was completed, the beads were resuspended in 600 µL of MyK buffer (20 mM Tris-HCl pH 8.0, 0.2% SDS, 100 mM NaCl, 10 mM EDTA, 10 mM EgTA, 0.5% Triton X-100) following the washes.

We performed three rounds on spatial tagging because it provided sufficient barcode combinations to uniquely label DNA complexes coming from each individual cell. Briefly, the mouse genome contains 2.5×10^9 nucleotides, which when divided per the average fragment size of DNA post digestion (823 bp) results in 3.04×10^6 DNA fragments per cell. If we do three rounds of barcoding, we provide 884,736 number of combinations which exceeds number of DNA molecules 3.4 times. Importantly, during scSPRITE barcoding we distribute clusters of DNA molecules not single molecules so the actual number of barcode combinations will exceed number of spatial clusters much more than our calculation.

Library Preparation

To ensure we capture all information coming from single cells, we need to sequence all DNA molecules that were bound to the beads. The bead solution was split equally into 10 low-bind Eppendorf 1.5 mL tubes, with each tube containing 60 µL of beads in MyK buffer. Next, an additional 32 µL of MyK buffer and 8 µL of Proteinase K (NEB, #P8107S, Ispwich, MA) were added to each tube. All 10 tubes were placed on an Eppendorf ThermoCycler C, and reverse crosslinking proceeded overnight at 60°C while shaking at 1200 rpm. Next, the tubes were placed on a DynaMag-2 magnet, and the MyK and Proteinase K solution were transferred to 10 new low-bind Eppendorf tubes. The beads from each of the tubes were washed once with 20 µL of H_2O, and then transferred to the same tube containing each respective MyK and Proteinase K solution. DNA from each of the tubes were purified using the Clean-and-Concentrator-5 columns (Zymo, #D4004, Irvine, CA) using 5X binding buffer to increase yield. Purified DNA from each column was eluted in 10 new Eppendorf 1.5mL tubes using 12 µL of water. Each of the tubes were filled to 30 µL using 15 µL Q5 Hot Start High-Fidelity 2X Master Mix (NEB, #M0493S, Ispwich, MA), 1.5 µL of 20X Evagreen (Biotium, #31000-T, Fremont, CA), 1.2 µL of 25 µM indexed Illumina primers, and 0.3 µL of H_2O. Real-time PCR amplification proceeded for 14 cycles, which was when the libraries entered exponential amplification, but had not plateaued. Following amplification, each of the libraries was diluted 4-fold prior to running on a 1% Agarose E-gel (Life Technologies, #G402001, Carlsbad, CA) with a E-Gel 1-Kb Plus DNA Ladder (Life Technologies, #10488090, Carlsbad, CA) as a reference. After the run, the gel was cut between 300 and 1000 bp marks to remove primer dimers, small non-specific amplicons, and long DNA amplicons. Libraries from the gel were purified using a Gel Purification Kit (Zymo, #D4002, Irvine, CA) as described by the manufacturer, and 20 µL of H_2O was used to elute libraries off the column.

To estimate the number of unique molecules in our libraries, the molarity of our libraries was determined using the concentration of our library from Qubit 3.0 Fluorometer (using the Qubit dsDNA high-sensitivity assay kit) and the average library size (bp) using an Agilent

Tapestation 2200 (using the Agilent high-sensitivity D1000 ScreenTape and reagents). This in addition to estimated losses during library cleanup allowed us to estimate the number of unique molecules in our libraries. The libraries were sequenced with a read depth of 2.4X to ensure that we are able to map the DNA contained in each cluster.

scSPRITE Data Generation

scSPRITE data was generated using Illumina paired-end sequencing on the Novoseq through Novogene Corporation. Reads were sequenced with at least 120 bp in Read 1 for genomic DNA information and the DPM tag and 95 bp in Read 2 to read the other five remaining tags (odd—even—odd—even—Y-even) (Figure S1c). We generated 1,269,693,929 reads from the scSPRITE library made from ~1500 cells. From the FastQC report, we observe a normal distribution of GC content per sequence (Read 1: normal distribution between ~15-71%; Read 2: normal distribution between ~27-59%).

Sequencing analysis pipeline

The full barcode sequence was identified by combining the DNA tag sequence from the beginning of Read 1 and the remaining five barcode tags from Read 2 (Figure S1c). The tags were identified from a table of known tag sequences as previously described[7], with Odd and Even tags allowing up to two mismatches and DPM and Y-even tags allowing zero mismatches. Out of 1,269,693,929 reads sequenced, we identified: 26,546,674 (2.1%) of reads with 0 barcodes, 62,183,357 (4.9%) of reads with 1 barcodes, 116,086,266 (9.1%) of reads with 2 barcodes, 291,410,130 (22.9%) of reads with 3 barcodes, 33,689,683 (2.7%) of reads with 4 barcodes, 107,535,755 (8.5%) of reads with 5 barcodes, and 632,242,064 reads (49.8%) that contained the full six barcode sequence. Any reads that lacked the full six barcode sequence (DPM—odd—even—odd—even—Y-even) in the expected order were discarded from further analysis and considered not-usable for identifying cell-of-origin. The remaining 632,242,064 reads are therefore considered usable and were kept for downstream alignment and filtering. Before alignment, Read 1 was trimmed to a length of 100 bp (Figure S1c).

Alignment and filtering of reads

The trimmed reads containing the full six barcode sequence were mapped to pre-indexed mm9 reference genome using STAR 2.6.1 using the following parameters: --

outFilterMultimapNmax 50 --outFilterScoreMinOverLread 0.30 --outFilterMatchNminOverLread 0.30 --outFilterIntronMotifs None --alignIntronMax 50000 --alignMatesGapMax 1000 --genomeLoad NoSharedMemory --outReadsUnmapped Fastx --alignIntronMin 80 --alignSJDBoverhangMin 5 --sjdbOverhang 100 --limitOutSJcollapsed 10000000 --limitIObufferSize=300000000. SAMtools 1.9 was applied to filter mapped reads, and only uniquely mapped reads (-q 255) were kept. Alignments that had overlapped a masked region as denoted by Repeatmasker (UCSC, milliDiv < 140) were removed using bedtools (version 2.25.0). Finally, reads that were aligned to a mm9 non-unique region of the genome were removed by excluding alignments that mapped to regions by the ComputeGenomeMask program (read length = 35 nt). After these filtration steps, all BAM files that corresponded to the same sample but contained different Illumina primers at sequencing were pooled together before cluster identification (Figure S1c).

Cluster barcode and cell barcode identification

To identify SPRITE clusters, all reads that contained the same six barcode sequences were grouped together into a single cluster. All reads containing the same six barcode sequences that started at the same genomic position were excluded to remove possible PCR duplicates. This led to 161,989,473 remaining reads. Once identified, a SPRITE cluster file was generated where each line contained the cluster barcode name and corresponding genomic alignments. Once the cluster barcodes were identified, the cell barcodes were identified by grouping clusters together that contained the same DPM, first Odd, and first Even barcode sequences. This grouping can create on the order of hundreds of thousands cell barcode files, but the majority of these files contain fewer than 10 clusters. As a result, only the largest 4000 cell barcode files based on file size were selected for downstream filtration, and the remaining cell barcode files were removed from the directory (Figure S1c).

Selecting single cells for analysis

Once the largest 4000 cell barcode files were identified, these files underwent additional in silico filtration to select files the most informative cells for analysis (Figure S1c). The files

were rank-ordered based on the number of clusters. The 1500 cell barcode files with the largest number of clusters from the initial 4000 files were selected, consistent with the initial number of cells used for the scSPRITE experiment. To ensure we selected only single cells for downstream analysis and not cell doublets, we removed the top 3.4% percent of cells as determined from the detected collision rate calculated from the results of human-mouse mixing experiment (Figures 1b, S1b). To ensure that we focus on the cells with most information per cell (number of reads/cDNA cluster/DNA contacts), we selected the top 1000 cell barcode files containing the most number of clusters per cell for downstream single-cell analysis. This led to 107,181,084 usable reads from the top 1000 cell barcode files.

Next, in the 1000 cells, we calculate the size distribution of DNA clusters per each cell and remove large clusters (>10,000 reads/cluster) from further analysis. We have previously reported[7] that clusters larger than >10,000 reads/cluster contain less information about higher-resolution structures (i.e. TADs) and most likely contain big chunks of nuclei that are composed of several chromosomes. We consider them less informative for the type of DNA interactions/structures (background) and therefore remove them from the further analysis. Excluding all reads in the >10,000-read clusters led to 83,318,292 remaining reads that were utilized for all downstream analyses.

Human-mouse mixing experiment

To determine the percent of single cells that are mixed together during scSPRITE (from crosslinking until the end of in-nuclei barcoding), we performed an in-nuclei part of the scSPRITE experiment using cell types from different species—mouse and human. We perform only the in-nuclei barcoding step because we have previously shown[7] that the spatial barcoding step used in bulk SPRITE leads to minimal collisions if total number of NHS beads is in an excess to total number of clusters in a sample (their mixing human/mouse experiment detects more than 99% of reads aligning to one species).

Mouse embryonic stem cells (bsps) and human cells (HEK293T) were harvested and resuspended into a single cell solution, then 30 x 10^6 cells per each cell type were mixed together in equal quantities and crosslinked, digested, dA-tailed, and barcoded in-nuclei as described above. Additionally, for the experiments described in Figure S1b, we mixed equal numbers of mouse and human cells post-crosslinking but i) pre-digestion, ii) post digestion, or iii) proceeded to the next step without mixing. Four rounds of in-nuclei barcoding were done (DPM, odd, even, y-even) (8 x 10^7 barcode combinations and 2x10^5cells). Next, nuclei were filtered through a 10-µm filter (PluriStrainer), 300 nuclei were removed as a new sample, reverse-crosslinked, and we proceeded as described above. 10% of the total purified libraries were sequenced using MiSeq; reads were then aligned to combined human/mouse genome using STAR alignment (hg19 and mm9 reference genomes). The best alignment was taken into consideration and if reads align equally well, they were considered as multi mappers and removed from the further analysis. Reads were sorted into individual cells based on cell-specific barcodes, and we focused only on cell-barcodes that had more than 1000 reads per cluster.

Next, we calculated the percentage of reads that aligned to each genome for each identified cell-barcode. We categorized cell-barcodes as mouse- or human-derived when they contained >95% single-species reads, and as mixed when they contained <95% single species reads. We then calculated a fraction of human only, mouse only, and mixed cell-barcodes (Figure 1b) and reported the percent of mixed cell-barcodes as detected collision

rate. Detected collision rate was further used to estimate thresholds used for cell filtering (Figures S1c, see Methods), and to calculate total collision rate. Total collision rate represents an estimation of all possible collisions and relies on the assumption that cells from the same species show similar collision rates as cells from mixed population, but we cannot detect them in our mixing experiment. It is calculated as follows: detected collision rate (mixed cells) + detected collision rate (human cells) + collision rate (mouse cells).

We note that despite starting with equal numbers of human and mouse cells, we observe bias in the final libraries with a higher number of mouse cells than human cells which results in better coverage per single human cell than mouse cell (Figure 1b). This is likely caused by the fact that we observed that during the full scSPRITE procedure (nuclei isolation, DNA digestion, in-cell barcoding), human HEK293 fibroblast cells are more susceptible to fragmentation and as a consequence lead to higher cell loss. We believe that this results in an unequal read distribution observed in our experiment and is consistent with other mouse/human mixing experiments when genomic methods like scHi-C are used[29].

Data analysis

Contact maps

Generation of ensemble heatmaps from scSPRITE

The generation of pairwise contact frequency matrices for ensemble scSPRITE was done similarly as was done for SPRITE[7]. For each cluster in the ensemble scSPRITE dataset, we gathered all possible pairs of reads. The pairwise contact frequency for each genomic bin i and j was then determined by counting the pairs of reads from each cluster, where both reads in a pair overlap with both i and j bins. These are unweighted clusters. To minimize the effect larger clusters contribute toward the number of pairwise contacts between any two bins, we also generated downweighted pairwise contact frequency matrices. The pairwise contact frequency was downweighted by a factor of 2/n, where n represents the number of reads in each cluster. The unweighted and downweighted contact frequency matrices were then normalized using Hi-Corrector[44]. In addition, low coverage bins and contacts in the same bin are masked in heatmaps.

To assess how well ensemble scSPRITE mapped known genomic structures, we compared the mouse ES cluster file from ensemble scSPRITE with the original mouse ES cluster file from SPRITE[7]. We used unweighted pairwise contact frequency matrices for genome-wide (1 Mb res) and AB compartment (200 kb res) for both ensemble scSPRITE and SPRITE, but using clusters containing fewer than 1,000 reads/cluster. Downweighted pairwise contact frequency matrices were used for TAD comparison (40 kb res) for both ensemble scSPRITE and SPRITE, but using all clusters.

Generation of single cell heatmaps from scSPRITE

Similarly to ensemble scSPRITE, single cell contact frequency matrices were generated at 1 Mb and 40 kb resolutions for all 1000 filtered cells. Contact frequency matrices were made similarly as described previously for ensemble scSPRITE, where each value in the matrix

reflects the number of clusters containing a read pair at genomic bin i and j. Single cell maps remained unweighted unless otherwise stated.

Comparison of ensemble and single cell SPRITE chromosome territory heatmaps

Genome-wide 1 Mb resolution contact maps were generated for the ensemble data set by pooling clusters containing fewer than 10,000 reads/cluster from the filtered 1000 single cells dataset. The resulting contact matrix for the ensemble dataset represents the non-downeighted, contact frequency for each pair of 1 Mb bins throughout the genome. The ensemble contact matrix was normalized by performing HiCorrector before plotting.

For the single cell maps, genome-wide 1 Mb resolution contact maps were generated by using clusters fewer than 10,000 reads/cluster for each single cell. The resulting contact matrix for each single cell represents the number of clusters that contained each pair of 1 Mb bins throughout the genome. Each single cell contact matrix was normalized by dividing every value in the contact matrix by the largest value in the matrix, resulting in a value between 0 to 1.

Insulation scores and A / B compartment annotation

Insulation scores and annotations for A and B compartments were calculated from the ensemble scSPRITE dataset using cworld (https://github.com/dekkerlab/cworld-dekker). Insulation scores were calculated using contact maps binned at 40 kb resolution, and A and B compartment annotations were calculated using contact maps binned at 200 kb and 1 Mb resolution. Insulation scores were calculated using the script matrix2insulation.pl with the parameters "--ss 80000 --im iqrMean --is 480000 --ids 320000" and compartment annotations were calculated using the script matrix2compartment.pl with default parameters. We used the output file ending in "insulation.boundaries.bed." These TAD regions correspond to the interval between two insulation boundaries. To quantitatively compare TADs between ensemble scSPRITE and HiC, we computed the correlation coefficient

between the insulation scores for each 40 kb genomic bin (using the ".insulation" file output by the matrix2insulation.pl script).

Detection scores for 3D genome structures

Detection scores were calculated to identify various 3D genome structures in single cells. These structures included chromosome territories, A/B compartments, TADs, centromere interactions, nuclear speckle interactions and nucleolar interactions. Each score reflects how clearly defined a given structure is in a single cell. The scores were calculated using a binary contact matrix for each cell, which defined whether or not each pair of genomic bins were in contact in that cell. For example, a clearly defined chromosome territory in a single cell consists of chromosomes interacting more with themselves than with each other (illustration Figure 2b).

To normalize detection scores, an expected detection score was calculated for each 3D genome structure in each cell. The expected detection score was calculated as the mean detection score for 1000 randomized structures, which were generated by randomly shuffling the genomic coordinates of known structures. The normalized detection score for each structure in each cell was calculated as the observed detection score minus the expected detection score (Supplemental Note 4).

Detection scores were calculated for each structure in each cell as follows:

●Chromosome territories: (observed intra-chromosomal contacts) / (total possible intra-chromosomal contacts) – (observed inter-chromosomal contacts) / (total possible inter-chromosomal contacts). Genome-wide scores were calculated for every possible pair of chromosomes between chr1-19, excluding combinations between the same chromosomes (e.g. chr1-chr1), amounting to 171 combinations (Supplemental Table 1) from binary matrices at 1 Mb resolution (171 combinations because chrA-chrB = chrB-chrA).

●Compartments: (observed intra-compartment contacts) / (total possible intra-compartment contacts) – (observed inter-compartment contacts) / (total possible inter-compartment

contacts). Genome-wide scores were calculated for all 224 regions across all chromosomes in which we detected a compartment switch in our ensemble dataset (e.g. chr1 has 21 regions and chr3 has 0 regions) (Supplemental Table 2). A compartment switch is defined as a transition between "A to B to A" or "B to A to B" compartments. Scores were calculated from binary matrices at 1Mb resolution.

●TADs: (observed intra-TAD contacts) / (total possible intra-TAD contacts) – (observed inter-TAD contacts) / (total possible inter-TAD contacts). Genome-wide TAD scores were calculated +/- 1 Mb from TAD boundary region, and these were calculated for all 2,602 TAD boundary regions that we detected in our ensemble dataset (Supplemental Table 3). Scores were calculated from binary matrices at 40 kb resolution.

●Centromere interactions: (observed centromere-centromere contacts) / (total possible centromere-centromere contacts) – (observed centromere-non-centromere contacts) / (total possible centromere-non-centromere contacts). Centromere interactions were defined as interactions between positions 3 Mb and 13 Mb of each chromosome.

●Nuclear speckle interactions: (observed speckle-speckle contacts) / (total possible speckle-speckle contacts) – (observed speckle-non-centromere contacts) / (total possible speckle-non-centromere contacts). Nuclear speckle interactions were defined as interactions between the following nuclear speckles regions[7]: Chr2 (164-174, 177-181 Mb), Chr4 (128-142 Mb, 147-155 Mb), Chr5 (112-126 Mb), Chr8 (123-127 Mb), Chr11 (95-103, 115-121 Mb), Chr13 (55-58 Mb), Chr15 (76-79 Mb), Chr17 (25-30 Mb).

●Nucleolar interactions: (observed nucleolar-nucleolar contacts) / (total possible nucleolar-nucleolar contacts) – (observed nucleolar-non-centromere contacts + observed non-nucleolar-non-nucleolar contacts) / (total possible nucleolar-non-centromere contacts + total possible non-nucleolar-non-nucleolar contacts). Nucleolar interactions were defined as interactions between the following nucleolar regions[7]: Chr12 (5-17 Mb, 25-32 Mb), Chr15 (3-6, 67-71 Mb), Chr16 (5-8 Mb), Chr18 (3-10, 13-24 Mb, 25-33 Mb, 39-42 Mb, 57-60 Mb), Chr19 (11-24 Mb, 25-28 Mb, 29-37 Mb, 48-53 Mb, 58-61 Mb).

Calculation of median absolute deviation (MAD) scores for scSPRITE

For each single cell in scSPRITE, we calculated the number of reads in each 1 Mb bin for every chromosome genome-wide (chr1-19). Once these reads were counted, we calculated the median absolute deviation (MAD) value for each cell based on the number of reads in each 1 Mb bin genome-wide to determine the variability of coverage.

Analysis of higher-order structures

Comparison of intra-chromosomal versus inter-chromosomal contacts

The percent of intra-chromosomal and inter-chromosomal contacts for each cell was calculated from the 1000 cells in scHi-C16 and from the filtered 1000 cells from scSPRITE (cluster size threshold of <10,000 reads/cluster). For scHi-C, because every cluster is a pairwise contact, we counted the number of pairwise contacts that were intra-chromosomal contacts (two contacts in the same chromosome) and inter-chromosomal contacts (two contacts coming from different chromosomes). For scSPRITE, we counted all pairwise contacts per cluster, where the number of pairwise contacts can be expressed as a binomial coefficient of "n choose 2," where n is the number of reads per cluster. From this, we then counted the number of intra-chromosomal and inter-chromosomal contacts. This was repeated for all clusters in each cell. The percent of inter-chromosomal contacts was determined by dividing the number of inter-chromosomal contacts by the sum of the number of intra- and inter-chromosomal contacts.

Frequencies of higher-order inter-chromosomal interactions

To determine the frequency of centromeric, speckle, or nucleolar interactions in single cells, we used the following metrics: i) percentage of cells that contain a given interaction in each 1 Mb bin and ii) normalized mean interaction value.

The percentage of cells containing centromere-proximal, speckle, or nucleolar interactions is determined by looking through the filtered 1000 single cells, focusing on clusters below 10,000 reads/cluster, and counting the number of cells containing at least one interaction between all 1 Mb bins (i and j) in the given centromere, speckle, and nucleolar regions, respectively. The genomic regions of these higher-order structures were defined previously in the section titled "Detection scores for 3D genome structures." To determine the expected frequency of cells that would contain these interactions by chance, we generated random genomic regions that were size matched to each feature. We generated 1,000 random permutations of each feature. For each permutation, we computed the percentage of single cells showing a contact between these random bins.

We get the normalized mean interaction value by first calculating an interaction matrix between all 1 Mb genomic bins, where the values in the interaction matrix were the percent of cells containing an interaction between each pair of 1 Mb bins, and then calculating the mean value for pairs of regions in this interaction matrix representing centromere-proximal, speckle, or nucleolar regions. For example, to determine the mean interaction value for cells containing an interaction between the centromere-proximal regions on chromosome 1 and chromosome 2, we calculated the mean value in this interaction matrix for chromosome 1 positions 3,000,000 to 13,000,000 with chromosome 2 positions 3,000,000 to 13,000,000.

Higher order structures in scHi-C data

Ensemble and single-cell contact maps from scHi-C16 were plotted to visualize centromere, speckle, and nucleolar interactions. The single-cell barcode from scHi-C that was referenced was "hyb_2i-1CDES-1CDES_p10.H9-adj."

DNA-FISH comparison with ensemble scSPRITE analysis

For the FISH analysis, we focused on the same chromosomal loci pairs that were originally analyzed in SPRITE. These pairs include two control chromosomal pairs and four NOR chromosomal pairs. The chromosomal loci pairs are listed below:

- Control 1: Chr3 (15-16 Mb) and Chr15 (4-5 Mb)
- Control 2: Chr3 (15-16 Mb) and Chr19 (18-19 Mb)
- NOR 1: Chr12 (6-7 Mb) and Chr15 (4-5 Mb)
- NOR 2: Chr15 (4-5 Mb) and Chr18 (3-4 Mb)
- NOR 3: Chr18 (3-4 Mb) and Chr19 (18-19 Mb)

We first compared contact frequency values from the loci pairs listed above between ensemble scSPRITE and SPRITE to determine how well the two methods correlated with each other. For both ensemble scSPRITE and SPRITE, we generated 1 Mb resolution, genome-wide pairwise contact frequency maps using clusters containing fewer than 10,000 reads/cluster. These contact maps were normalized using HiCorrector. Using both the ensemble scSPRITE and SPRITE normalized contact frequency maps, we then pulled out the contact frequency value from each of six loci pairs, plotted their values using a scatter plot, and calculated the coefficient of determination (R^2).

We generated 1 Mb resolution genome-wide contact frequency matrices for each single cell using clusters containing fewer than 10,000 reads/cluster. For each chromosomal loci pair, a cell contained that loci pair interaction if there was at least one read in the bin containing both loci. To calculate the percent of cells for each loci pair, we divided the number of cells containing the loci pair interaction by the total number of cells used in the analysis.

Calculation of % of reads coming from A/B compartments

To get the expected percentage of reads that fell into either the A or B compartments in our ensemble dataset, we used the data from the ensemble, genome-wide compartment switch analysis. We counted the number of 1 Mb bins that were classified as being in either A or B compartments genome-wide (except for chr3 and chrX). To get the expected percentage of

A or B reads in our ensemble dataset, we then divided the number 1 Mb bins in A or B compartments, respectively, by the total 1 Mb bins counted.

To get the percentage of reads in A or B compartments in single cells, we looked into the genome-wide reads (with the exception of chr3 and chrX) in each single cell file. From there, we sorted reads into A or B compartments depending on the data from the ensemble, genome-wide compartment switch analysis. Once sorted, we then divided the number of reads that fell into A or B compartments by the total number of reads counted for that cell to determine the percentage of reads in A or B compartments, respectively.

Contact maps of regions with heterogeneous structures

To identify regions of heterogeneity, we manually looked through a genome-wide heatmap using the ensemble scSPRITE dataset to look for emerging TAD-like structures in between designated A/B compartments and TAD regions based on the previously identified A/B regions and TAD boundary regions, respectively. Once a region was identified in a given chromosome, 40 kb weighted, single-cell contact maps were made for that specific chromosome, where the contact frequency values in each 40 kb bin are weighted by cluster size. In the 40 kb single-cell maps, the two 40-kb bins that made up the outermost interaction of the pseudo-TAD structure in the ensemble dataset were used to look for this same interaction in the single-cell dataset (further referred to as “bin A” and “bin B”).

Long-range interactions

Detection of heterogeneity in long-range interactions

For the interactions studied in this paper (Phc1 at the Nanog locus and Lhx5 at the Tbx3 locus), we used the 40-kb bins containing the locations of the Phc1 enhancer and Nanog

promoter and the locations of the Lhx5 gene and the Tbx3 gene as identified previously[38]. In every single cell, we first identified cells containing a contact anywhere along bin A and bin B in that chromosome to ensure that coverage was accounted for. For Phc1 and Nanog, bins A and B are Chr6 122,280,000-122,320,000 bp and Chr6 122,640,000-122,680,000 bp, respectively. For Tbx3 and Lhx5, bins A and B are Chr5 120,120,000-120,160,000 bp & Chr5 120,880,000-120,920,000 bp, respectively. On average we detect a contact in ⅓ of total cell number which we believe is technical and is due to non-sufficient coverage of every region per cell. Once the cells with coverage were identified, we identified and grouped cells in this set that contained or lacked the interaction at the intersection of bin A and bin B. For the SE-promoter interaction at the Nanog locus, we identified 308 cells with read coverage, of which 159 cells contained the Nanog-Phc1 contact and 149 cells lacked the Nanog-Phc1 contact. For the SE-promoter interaction at the Tbx3 locus, we identified 301 cells with read coverage, of which 152 cells contained the Tbx3-Lhx5 contact and 149 cells lacked the Tbx3-Lhx5 contact.

For the cells with and without an interaction over the AB compartment boundary (Figure 4f), a similar approach was done as described above for grouping. We first identified cells based on coverage anywhere along bin A and bin B, which was Chr4 40,120,000-40,160,000 bp and Chr4 40,920,000-40,960,000 bp. This region was chosen since the ensemble scSPRITE map displayed a high contact frequency at this point, which happened to be over an AB compartment transition. Once the cells with coverage were identified, we identified and grouped cells that had contained or lacked an interaction within 120 kb of bin A and B (i.e. Chr4 40,080,000-40,200,000 and Chr4 40,880,000-41,000,000 bp). Unlike the promoter-enhancer examples where a known bin contains the promoter and enhancer loci, this information does not exist for contacts over a compartment boundary. Therefore, we provided a wider base pair range to sort the cells into those two groups. Of the 379 cells identified with read coverage, 199 cells contained an interaction over the AB compartment, and 170 lacked this interaction.

Virtual 4C analysis

To identify contacts with a specific locus, such as in the Nanog and Tbx3 examples, we first calculated a contact frequency matrix for all pairs of genomic bins at 40 kb resolution. Using the cells that were grouped into sets either containing or lacking interactions with a specific locus, we combined each cell's individual contact frequency matrix to create an ensemble contact frequency matrix for each set. Each ensemble contact frequency matrix was normalized by Hi-Corrector[44]. To convert this contact matrix to a 1-dimensional profile of contacts, we simply used the values in the row of the contact matrix corresponding to the locus of interest.

Significance and variance estimation

To determine the variance and significance of the observed contacts between these two groups, we performed a bootstrap method. Specifically, we generated random groups of cells by sampling with replacement from the initially defined groups. This approach allows us to estimate how much of the observed signal is dependent on individual cells in the population and how stable these estimates are across cells in the group. We generated 1000 random bootstrap groups for each of the two groups and computed the average and standard deviation across these permutations. To define the significance of differences between these two groups, we computed a p-value using the unpaired two-sided t-test with Welch's correction between the bootstrap values in group A versus group B.

Comparison of cells with & without SE-promoter contact

We compared the number of reads and contacts from the cells containing and lacking the SE-promoter contact from the Nanog and Tbx3 examples to determine if there was any bias that contributed to differences in their respective virtual 4C plots. For the Nanog-Phc1 example, we focused our analysis on the cells that contained or lacked the Nanog-Phc1 contact, as described previously. For each cell, we went through every cluster and calculated the number of genome-wide reads and contacts from each cluster. We then summed the number of reads

and contacts from all the clusters in each cell, and then repeated this process for all the cells in the two groups. We then used the Kolmogorov–Smirnov test to calculate the statistical significance between the two groups. This same analysis was repeated for the Tbx3-Lhx5 example.

ChIPseq data

We downloaded the call sets from the ENCODE portal (https://www.encodeproject.org/) with the following identifiers: H3K27ac ENCSR000CDE, H3K4me3 ENCSR000CBG, H3K27me3 ENCSR000CFN.

Cell-cycle analysis

We computationally sorted the cells into M, G1, G2, or S phases of cell cycle based on the parameters described previously[16]. After categorizing the cells by phase, we calculated the percentage of cells in each corresponding cell cycle phase in the sets that contained or lacked a particular interaction.

For the SE-promoter interaction at the Nanog locus, 152 of the 159 cells (95.6%) containing the Nanog-Phc1 contact and 145 of the 149 cells (97.3%) lacking the Nanog-Phc1 contact were sorted into cell cycle phases. For the SE-promoter interaction at the Tbx3 locus, 146 of the 152 cells (96.1%) containing the Tbx3-Lhx5 contact and 148 of the 149 cells (99.3%) lacking the Tbx3-Lhx5 contact were sorted into cell cycle phases. For the Chr4 AB heterogeneity example, 195 of the 199 cells (98.0%) containing the interaction of the AB compartment boundary and 166 of the 170 cells (97.6%) lacking this interaction were sorted into cell cycle phases. The other cells were identified as “Unknown” and were not included in the cell cycle plot.

ACKNOWLEDGEMENTS

We would like to thank the assistance from Fan Gao from Caltech's Bioinformatics Resource Center and Igor Antoshechkin from Caltech's Millard and Muriel Jacobs Genetics and Genomics Laboratory. We would like to thank Chris Chen, Vicky Trinh, Elizabeth Detmar, Elizabeth Soehalim, Aditi Narayanan, and Isabel Goronzy for their contributions in helping develop scSPRITE and the analysis. We would like to thank Matt Thompson's laboratory for allowing us to use their MiSeq instrument and the ENCODE Consortium and the ENCODE production laboratory of Bing Ren, UCSD for making their data publicly available. We also thank Natasha Shelby and Shawna Hiley for contributions to writing and editing this manuscript, and Inna-Marie Strazhnik for helping with illustrations.

FUNDING

This work was funded by the NIH 4DN Program (U01 DA040612 and U01 HL130007), NHGRI GGR Program (U01 HG007910), New York Stem Cell Foundation (NYSCF-R-I13), Sontag Foundation, and funds from Caltech. M.V.A. and S.A.Q. were funded by a NSF GRFP Fellowship. M.V.A. was additionally funded by the Earle C. Anthony Fellowship (Caltech). M.G. is a NYSCF-Robertson Investigator.

DATA AVAILABILITY STATEMENT

The datasets (Figures 1-5; supplemental figures S1-S5) generated during and analyzed during the current study are available in the GEO repository under accession number GSE154353 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE154353). scSPRITE software is available at https://github.com/caltech-bioinformatics-resource-center/Guttman_Ismagilov_Labs.

CONFLICTS OF INTEREST

This paper is the subject of a patent application filed by Caltech. R.F.I. has a financial interest in Talis Biomedical Corp. S.A.Q. and M.G. are inventors on a patent owned by Caltech on SPRITE. The remaining authors declare no competing interests.

REFERENCES

1. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289 (2009).
2. Nora, E.P. et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* 485, 381-385 (2012).
3. Dixon, J.R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485, 376-380 (2012).
4. Dekker, J. & Mirny, L. The 3D genome as moderator of chromosomal communication. *Cell* 164, 1110-1121 (2016).
5. Freire-Pritchett, P. et al. Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem cells. *eLife* 6, e21926 (2017).
6. Whyte, Warren A. et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* 153, 307-319 (2013).
7. Quinodoz, S.A. et al. Higher-order inter-chromosomal hubs shape 3d genome organization in the nucleus. *Cell* 174, 744-757.e724 (2018).
8. Mao, Y.S., Zhang, B. & Spector, D.L. Biogenesis and function of nuclear bodies. *Trends in Genetics* 27, 295-306 (2011).
9. Miura, H. et al. Single-cell DNA replication profiling identifies spatiotemporal developmental dynamics of chromosome organization. *Nature Genetics* 51, 1356-1368 (2019).
10. Kagey, M.H. et al. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* 467, 430-435 (2010).
11. Chen, Y. et al. Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler. *Journal of Cell Biology* 217, 4025-4048 (2018).
12. Pederson, T. The Nucleolus. *Cold Spring Harbor Perspectives in Biology* 3 (2011).
13. Finn, E.H. et al. Extensive heterogeneity and intrinsic variation in spatial genome organization. *Cell* 176, 1502-1515.e1510 (2019).

14. Wang, S. et al. Spatial organization of chromatin domains and compartments in single chromosomes. *Science* 353, 598 (2016).
15. Stevens, T.J. et al. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* 544, 59-64 (2017).
16. Nagano, T. et al. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* 547, 61-67 (2017).
17. Ma, X., Ezer, D., Adryan, B. & Stevens, T.J. Canonical and single-cell Hi-C reveal distinct chromatin interaction sub-networks of mammalian transcription factors. *Genome Biology* 19, 174 (2018).
18. Mohammed, H. et al. Single-cell landscape of transcriptional heterogeneity and cell fate decisions during mouse early gastrulation. *Cell Reports* 20, 1215-1228 (2017).
19. Buettner, F. et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology* 33, 155-160 (2015).
20. Bintu, B. et al. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* 362, eaau1783 (2018).
21. Giorgetti, L. et al. Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell* 157, 950-963 (2014).
22. Nagano, T. et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* 502, 59-64 (2013).
23. O'Sullivan, J.M., Hendy, M.D., Pichugina, T., Wake, G.C. & Langowski, J. The statistical-mechanics of chromosome conformation capture. *Nucleus* 4, 390-398 (2013).
24. Kolodziejczyk, Aleksandra A. et al. Single cell RNA-sequencing of pluripotent states unlocks modular transcriptional variation. *Cell Stem Cell* 17, 471-485 (2015).
25. Guo, F. et al. Single-cell multi-omics sequencing of mouse early embryos and embryonic stem cells. *Cell Research* 27, 967-988 (2017).

26. Ghimire, S. et al. Comparative analysis of naive, primed and ground state pluripotency in mouse embryonic stem cells originating from the same genetic background. *Scientific Reports* 8, 5884 (2018).
27. Lee, D.-S. et al. Simultaneous profiling of 3D genome structure and DNA methylation in single human cells. *Nature Methods* 16, 999-1006 (2019).
28. Zhou, S., Jiang, W., Zhao, Y. & Zhou, D.-X. Single-cell three-dimensional genome structures of rice gametes and unicellular zygotes. *Nature Plants* 5, 795-800 (2019).
29. Ramani, V. et al. Massively multiplex single-cell Hi-C. *Nature Methods* 14, 263-266 (2017).
30. Ramani, V. et al. Sci-Hi-C: A single-cell Hi-C method for mapping 3D genome organization in large number of single cells. *Methods* 170, 61-68 (2020).
31. Spector, D.L. & Lamond, A.I. Nuclear Speckles. *Cold Spring Harbor Perspectives in Biology* 3 (2011).
32. Guenatri, M., Bailly, D., Maison, C.l. & Almouzni, G.v. Mouse centric and pericentric satellite repeats form distinct functional heterochromatin. *Journal of Cell Biology* 166, 493-505 (2004).
33. Almouzni, G. & Probst, A.V. Heterochromatin maintenance and establishment: Lessons from the mouse pericentromere. *Nucleus* 2, 332-338 (2011).
34. Strongin, D.E., Groudine, M. & Politz, J.C.R. Nucleolar tethering mediates pairing between the IgH and Myc loci. *Nucleus* 5, 474-481 (2014).
35. Dowen, Jill M. et al. Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* 159, 374-387 (2014).
36. Pachano, T., Crispatzu, G. & Rada-Iglesias, A. Polycomb proteins as organizers of 3D genome architecture in embryonic stem cells. *Briefings in Functional Genomics* 18, 358-366 (2019).
37. Blinka, S., Reimer, Michael H., Jr., Pulakanti, K. & Rao, S. Super-enhancers at the *Nanog* locus differentially regulate neighboring pluripotency-associated genes. *Cell Reports* 17, 19-28 (2016).

38. Novo, C.L. et al. Long-range enhancer interactions are prevalent in mouse embryonic stem cells and are reorganized upon pluripotent state transition. *Cell Reports* 22, 2615-2627 (2018).

39. Schoenfelder, S. et al. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Research* 25, 582-597 (2015).

40. Apostolou, E. et al. Genome-wide chromatin interactions of the nanog locus in pluripotency, differentiation, and reprogramming. *Cell Stem Cell* 12, 699-712 (2013).

41. Russell, R. et al. A dynamic role of TBX3 in the pluripotency circuitry. *Stem Cell Reports* 5, 1155-1170 (2015).

42. Kalmar, T. et al. Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. *PLOS Biology* 7, e1000149 (2009).

43. Engreitz, Jesse M. et al. RNA-RNA interactions enable specific targeting of noncoding rnas to nascent pre-mRNAs and chromatin sites. *Cell* **159**, 188-199 (2014).

44. Li, W., Gong, K., Li, Q., Alber, F. & Zhou, X.J. Hi-Corrector: A fast, scalable and memory-efficient package for normalizing large-scale Hi-C data. *Bioinformatics* **31**, 960-962 (2015).

MAIN FIGURES AND CAPTIONS

Figure 1

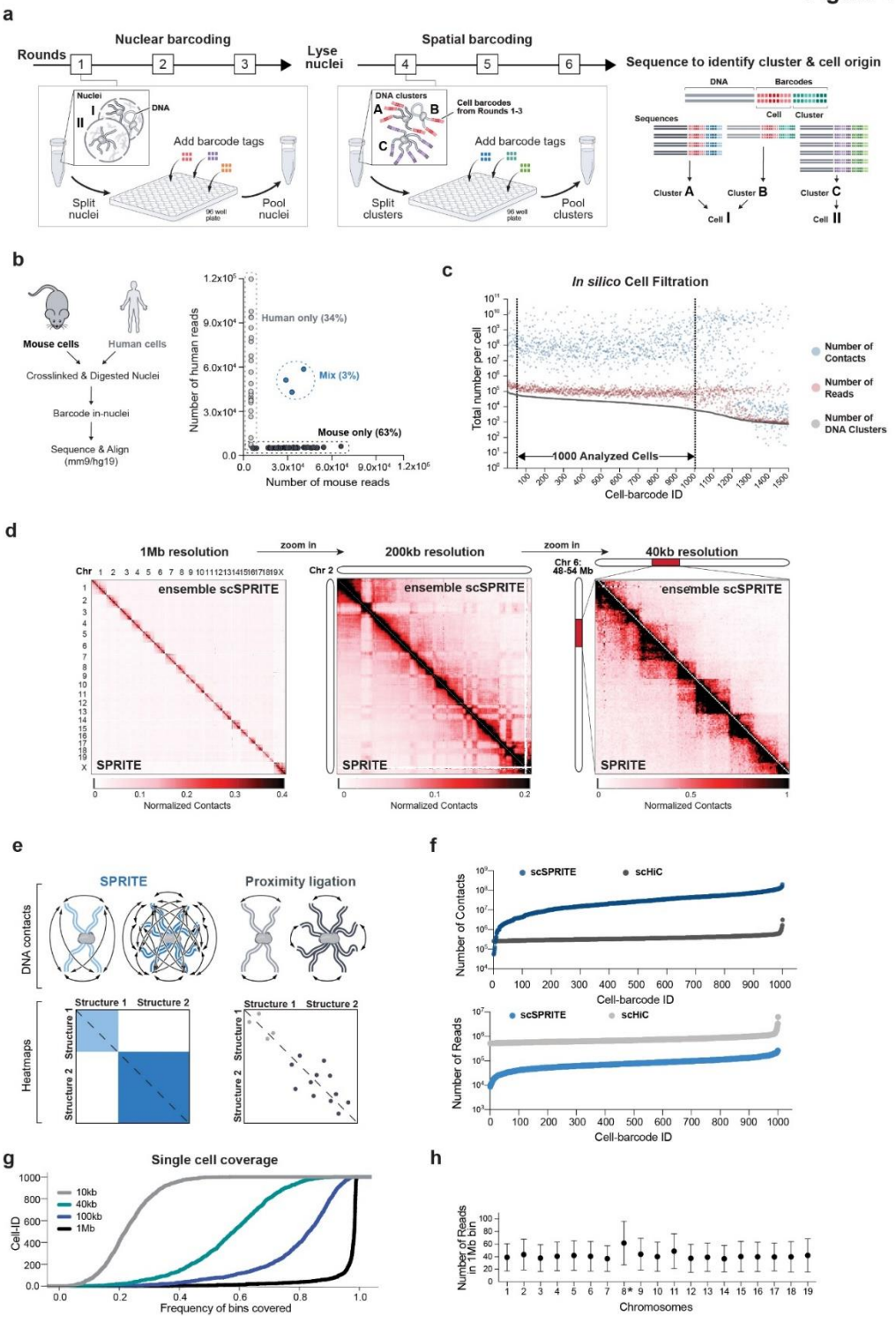


Figure 1: Single cell SPRITE—a single cell method to map DNA structure genome-wide.

a. Schematic of scSPRITE protocol. **b.** Validation of in-nuclei barcoding step on mixed cell population (human-mouse cells); number of reads for each identified cell barcode ID is plotted. Threshold of >95% single species reads was applied to identify mouse or human only cells; cell-barcodes > 1,000 reads are plotted. **c.** Number of contacts (blue), reads (red), and DNA clusters (grey) plotted for the 1,500 cells. Dashed lines represent filtration steps: left of the dashed lines—cell aggregates estimated based on detected collision rate from Figure 1b, right of the dashed lines—cells with low number of reads/contacts **d.** Comparison of merged scSPRITE (upper diagonal, "ensemble scSPRITE") and bulk SPRITE[7] (lower diagonal). Chromosome territories across all chromosomes at 1 Mb resolution (left); A/B compartments on chromosome 2 at 200 kb resolution (middle); TADs within a 18-Mb region of chromosome 6 at 40 kb resolution (right). **e.** Schematic illustration of multiway interactions (SPRITE-derived methods) and pairwise interactions (proximity ligation methods) and examples of heatmaps**. f.** Number of contacts (top) and number of reads (bottom) obtained from scSPRITE (blue) and scHi-C[16] (grey). **g.** Genomic coverage per 1 Mb, 100 kb, 40 kb, 10 kb bins in individual 1,000 cells. **h**. Average number of reads per single cell in 1 Mb bins of each chromosome (n = 1,000 cells). Average (dots) and SD (bars) are shown; asterisk marks chromosome with detected trisomy.

Figure 2

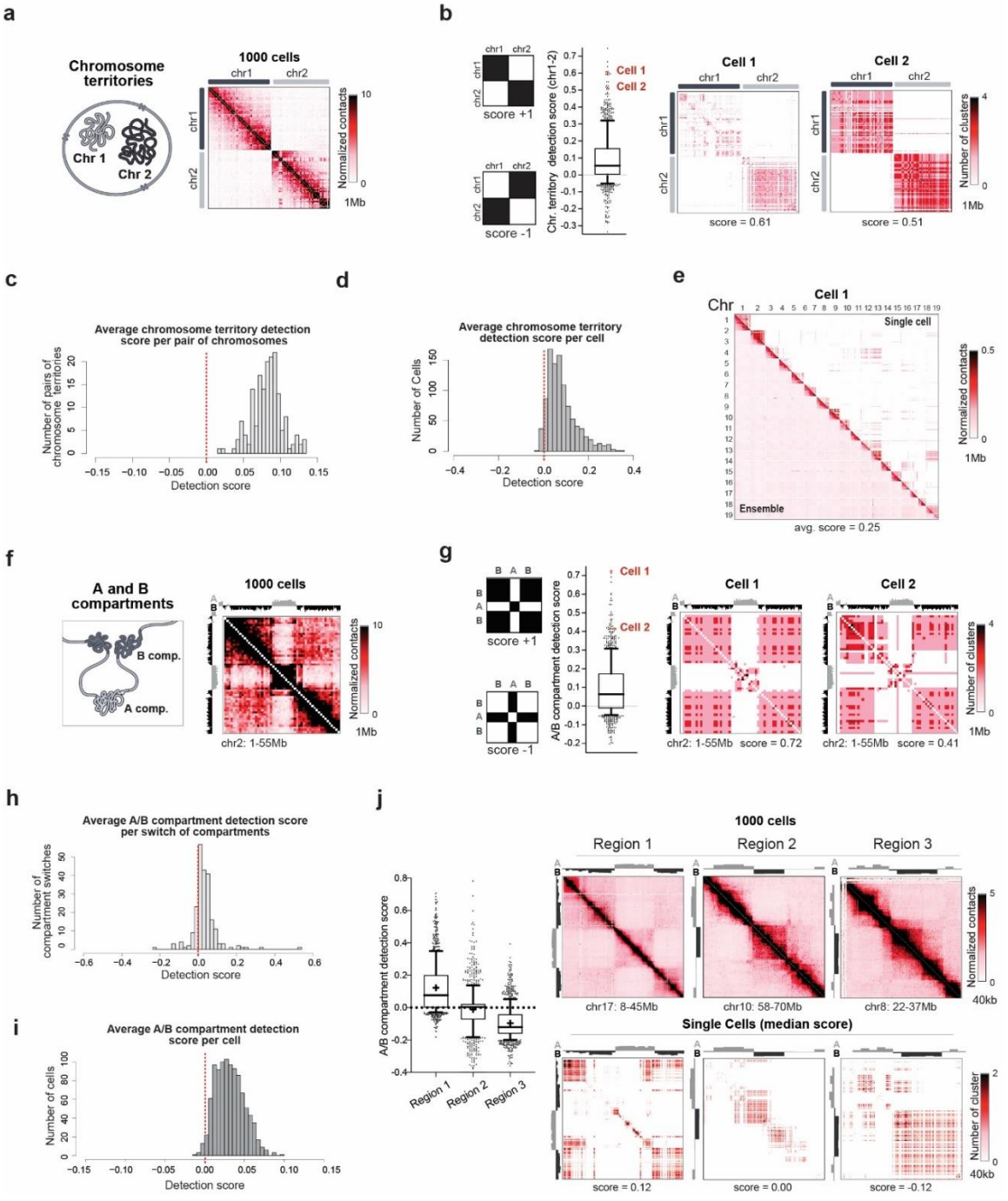


Figure 2: scSPRITE accurately measures single cell DNA interactions at different resolutions by capturing multiway interactions.

a. Illustration of chromosome territories for chr1 and chr 2 (left) and ensemble scSPRITE heatmap (right) of the same structures; downweighted contact map at 1 Mb resolution. **b.** Chromosome territory normalized detection scores for 1,000 individual cells between chr1 and chr2. Left: representation of structures with max. score (+1) and min. score (-1) Center: Box plot where whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line represents the median, red dots represent single cell examples (n = 1,000 cells). Right: single cell examples of chr1 and chr2 territories, plotted as number of DNA clusters at 1 Mb resolution. **c.** Normalized detection scores across all 1,000 cells per each pair of chromosome territories detected in ensemble scSPRITE data; score = 0 (red line). **d.** Normalized detection scores across all pairs of chromosome territories detected in ensemble scSPRITE data per single cell; score = 0 (red line). **e.** Chromosome territories (chr1-19) in ensemble scSPRITE (left) and in a single cell (right, detection score = 0.25). **f.** Illustration of A/B compartment in chr2:0-55 Mb (left) and ensemble scSPRITE heatmap (right); downweighted contact map at 1 Mb resolution. **g.** A/B compartments detection scores for 1,000 individual cells. Left: representation of structures with max score (+1) and min. score (-1). Center: Box plot where whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line represents the median, red dots represent single cell examples (n = 1,000 cells). Right: single cell examples of A/B compartments in chr2:0-55 Mb, plotted as number of DNA clusters at 1 Mb resolution. **h.** Normalized detection scores across all 1,000 cells per each compartment switch; score = 0 (red line). **i.** Compartment detection scores across all compartments per single cell; score = 0 (red line). **j.** Examples of three different regions containing a high (Region 1), medium (Region 2), and low (Region 3) median compartment switch score. For each region's box plot: whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line represents the median, red dots represent single cell examples (n = 1000 cells). Heatmaps for each region are shown in both ensemble scSPRITE (above) and single cell (below).

Figure 3

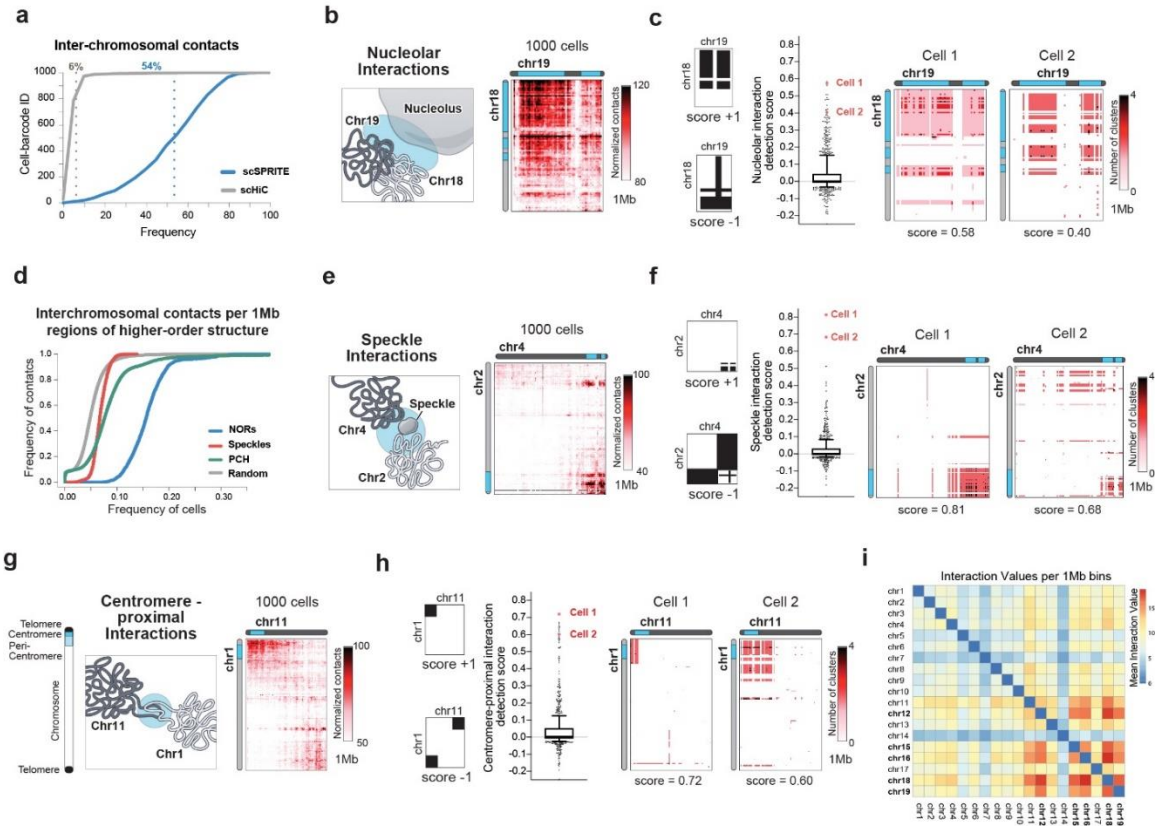


Figure 3: scSPRITE identifies inter-chromosomal structures genome-wide in hundreds of single mESC

a. Quantification of inter-chromosomal contacts from the top 1,000 cells by scHi-C[16] (grey) and scSPRITE (blue). The dashed lines represent the mean percentage of inter-chromosomal contacts. **b.** Nucleolar interaction between chr18 and chr19: illustration (left) and ensemble scSPRITE heatmap (right); contact map at 1 Mb resolution. **c.** Nucleolar interactions detection scores for 1,000 cells (middle). Box plot where whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line represents the median, red dots represent single cell examples (n = 1000 cells). Representation of structures with max score (+1) and min. score (-1) (left). Single cell examples (right); plotted as number of DNA clusters at 1 Mb resolution. **d.** Frequency of NOR (blue), speckle (red), and PCH (green) higher-order interactions in comparison to randomly shuffled regions of the same size (grey) in 1000 individual cells. **e.** Speckle interaction between chr2 and chr4: illustration

(left) and ensemble scSPRITE heatmap (right); contact map at 1 Mb resolution. **f.** Speckle interaction detection scores for 1,000 individual cells (middle). Box plot where whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line represents the median, red dots represent single cell examples (n = 1000 cells). Representation of structures with max score (+1) and min. score (-1) (left). Single cell examples (right); plotted number of DNA clusters at 1 Mb resolution. **g.** PCH interactions between chr1 and chr11: illustrations (left) and ensemble scSPRITE heatmap (right); contact map at 1 Mb resolution. **h.** PCH region detection scores for 1000 individual cells (middle). Box plot where whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line represents the median, red dots represent single cell examples (n = 1,000 cells). Representation of structures with max score +1 and min. score -1 (left). Single cell examples (right); plotted as number of DNA clusters at 1 Mb resolution. **i.** Mean interaction value of inter-chromosomal PCH contacts (normalized to number of reads per region) for each pair of chromosomes. NOR-containing chromosomes are shown in bold.

Figure 4

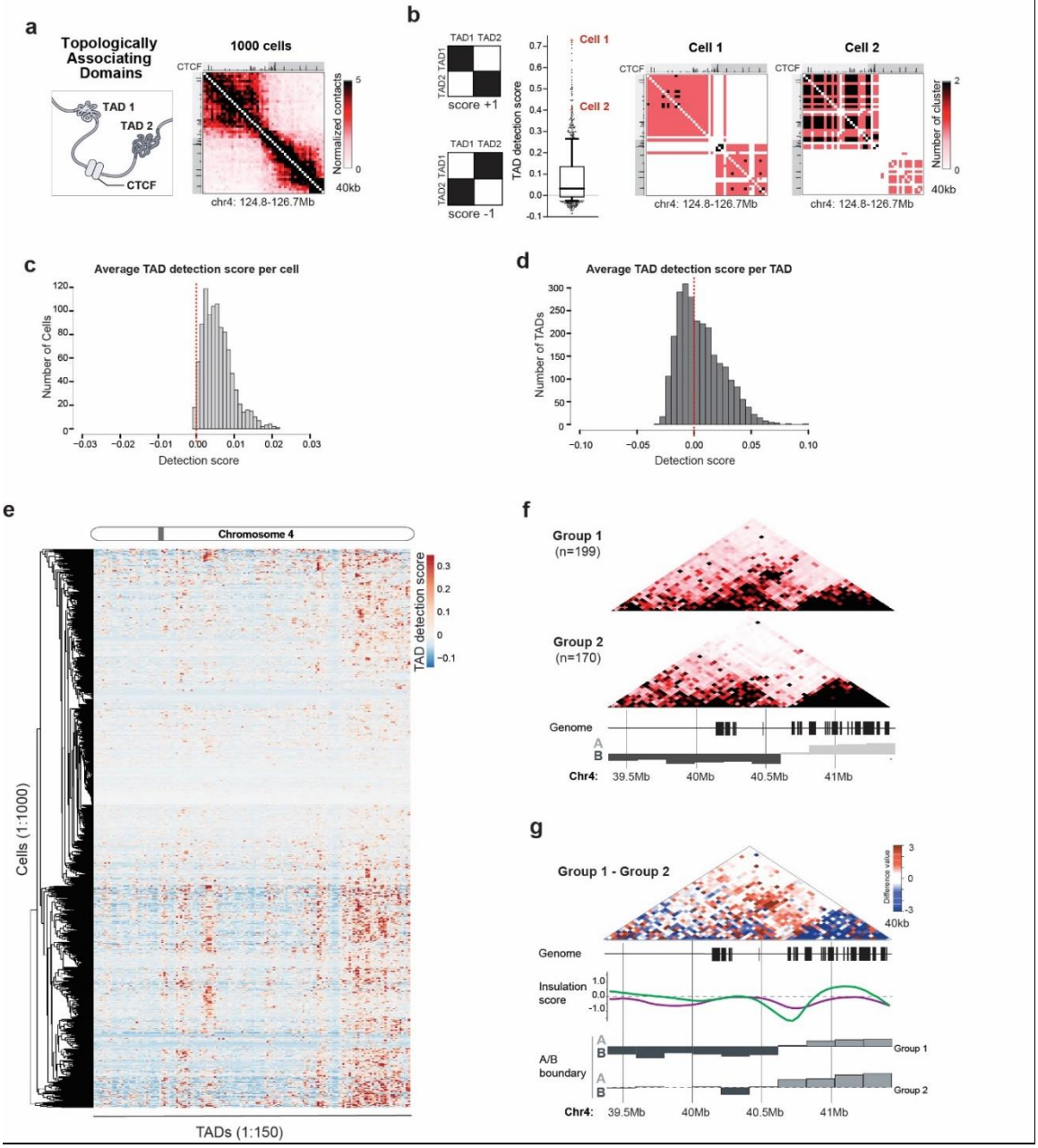


Figure 4: TADs are heterogeneous units present in the genomes of individual mESCs

a. TAD structure between 124.8-126.7 Mb of chr4: illustration (left) and scSPRITE heatmap (right); pairwise contact map at 40 kb resolution. **b.** TAD detection scores for 1,000 cells (middle). Box plot where whiskers represent the 10th and 90th percentiles, box limits

represent the 25th and 75th percentiles, black line represents the median, red dots represent single cell examples (n = 1000 cells). Representation of structures with max score (+1) and min. score (-1). Single cell examples (right), plotted as number of DNA clusters at 40 kb resolution. **c.** Normalized detection scores across all 1,000 cells per each TAD detected in ensemble scSPRITE data; red line marks score = 0. **d.** TAD detection scores across all TADs detected in ensemble scSPRITE data per single cell; red line marks score = 0. **e.** TAD detection scores across 1000 cells (clustered based on score similarity pattern): columns represent the strength of TAD detection scores for all TADs detected across chr4 in ensemble scSPRITE; grey bar indicates the variable region described in Figure S4c. **f.** Ensemble heatmaps across 39.4-41.4 Mb region of chr4 representing cells containing (Group 1, top) or lacking (Group 2, bottom) the contact emerging over the boundary of A/B compartment. **g.** Difference contact map across 39.4-41.4Mb of chr4 made by subtracting the normalized contacts in Group 2 from Group 1 (Figure 4f). Insulation scores for cells in Group 1 (purple) and Group 2 (green) are plotted.

Figure 5

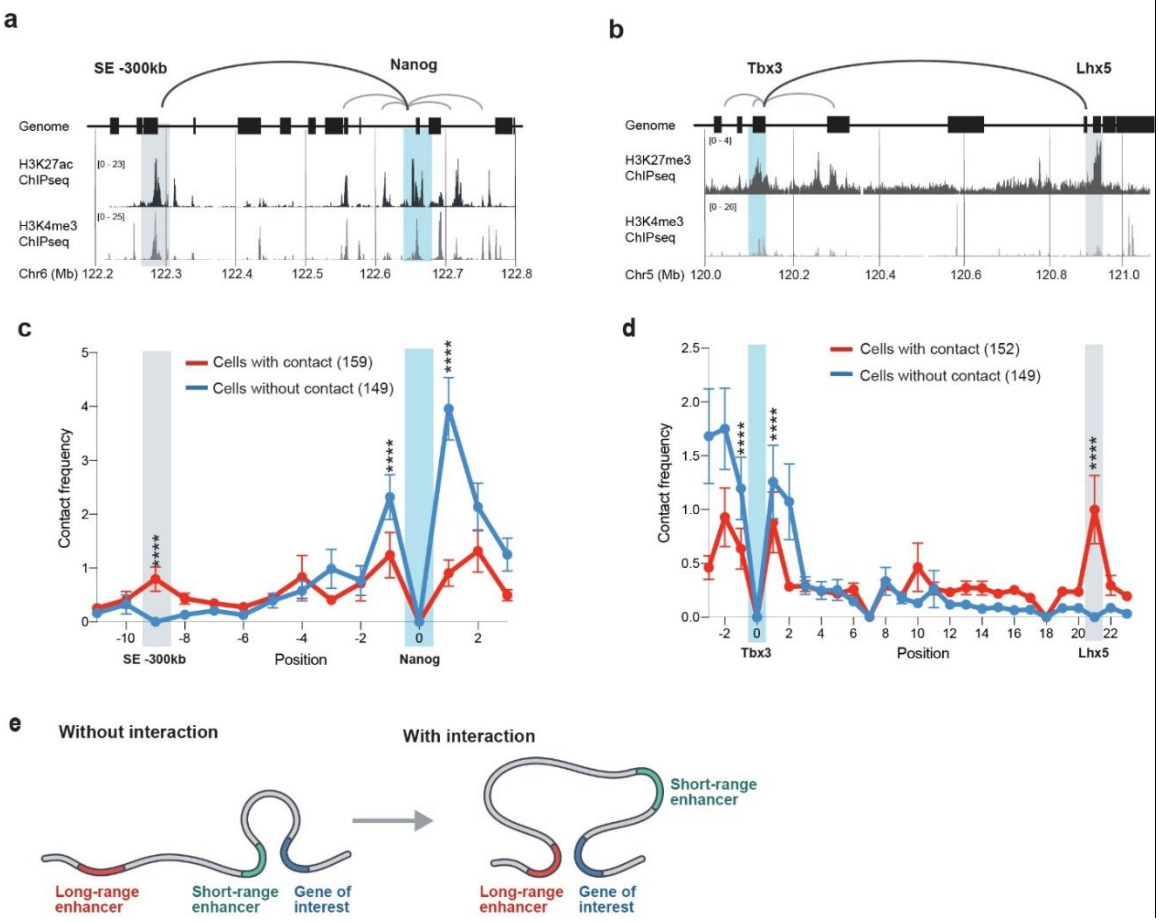


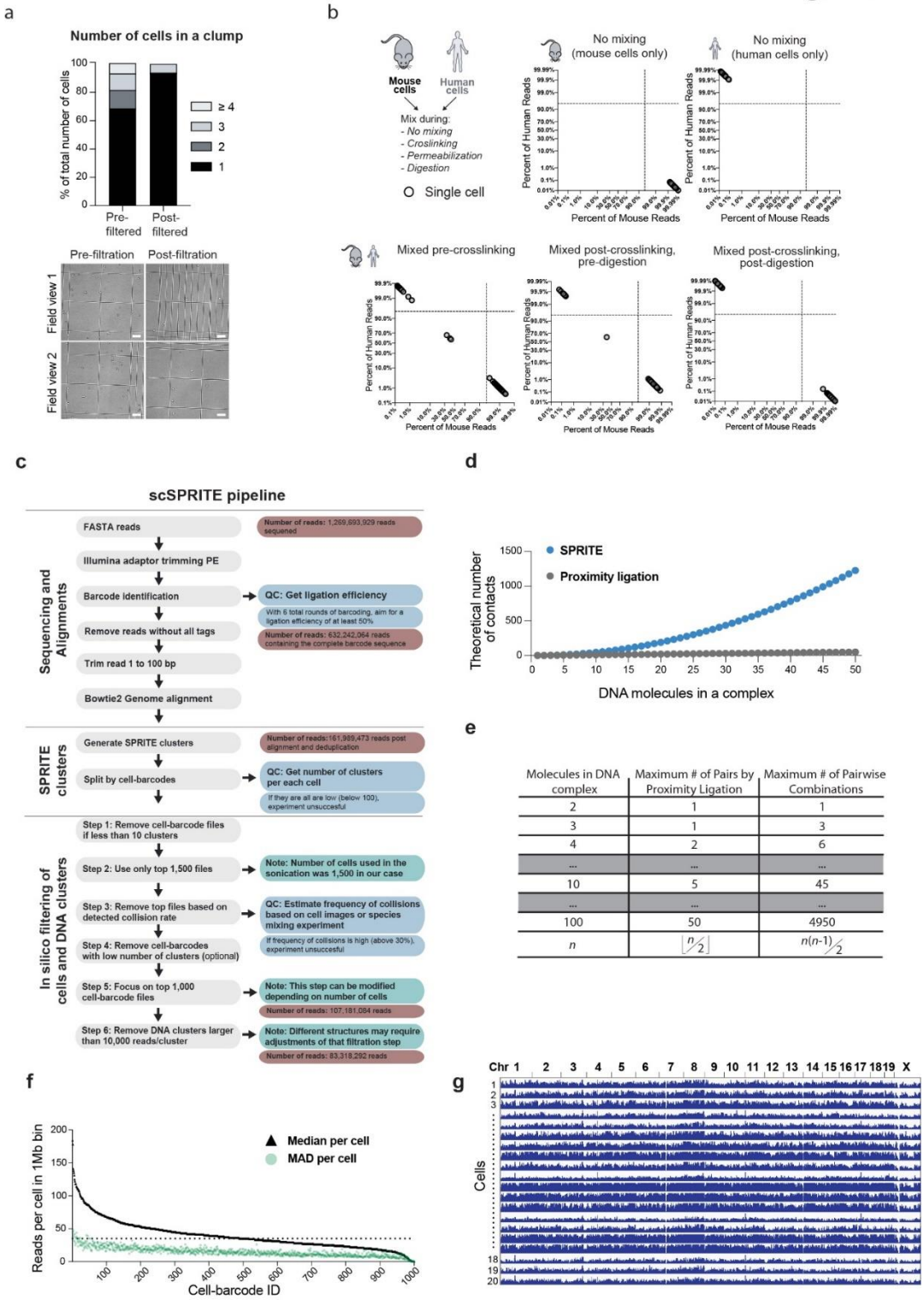
Figure 5: Heterogeneous structural states formed by Nanog and Tbx3 loci in individual mESC.

a. Representation of the Nanog locus and its DNA interactions with SE: 122.2-122.8 Mb region in chr6 with corresponding ChIPseq tracks for H3K27ac and H3K4me3; Nanog-SE interaction (black lines). **b.** Representation of Tbx3 locus and its DNA interactions with Lhx5: 120.0-121.0 Mb region in chr5 with the corresponding ChIPseq tracks for H3K27me3 and H3K4me3; Tbx3-Lhx5 interaction (black line). **c.** Normalized contact frequency plot between Nanog locus and 122.2-122.8 Mb surrounding region in chr6; shown cells containing (red) or lacking (blue) the contact between the Nanog locus and SE -300 kb. Each position refers to a 40 kb bin. Asterisks denote statistical significance ($p<0.0001$,unpaired

two-sided t-test with Welch's correction) between the two groups at the specified positions (n = 1000 random bootstrap groups for each of the two groups). Error bars represent one standard deviation. **d.** Normalized contact frequency plot between Tbx3 locus and 120.0-121.0 Mb surrounding region in chr5; shown cells containing (red) or lacking (blue) the contact between the Tbx3 locus and Lhx5. Each position refers to a 40 kb bin. Asterisks denote statistical significance ($p<0.0001$, unpaired two-sided t-test with Welch's correction) between the two groups at the specified positions (n = 1000 random bootstrap groups for each of the two groups). Error bars represent one standard deviation. **e.** Schematic illustrating differences in structure when a gene of interest lacks (left) and contains (right) the long-range enhancer interaction.

SUPPLEMENTARY FIGURES

Figure S1

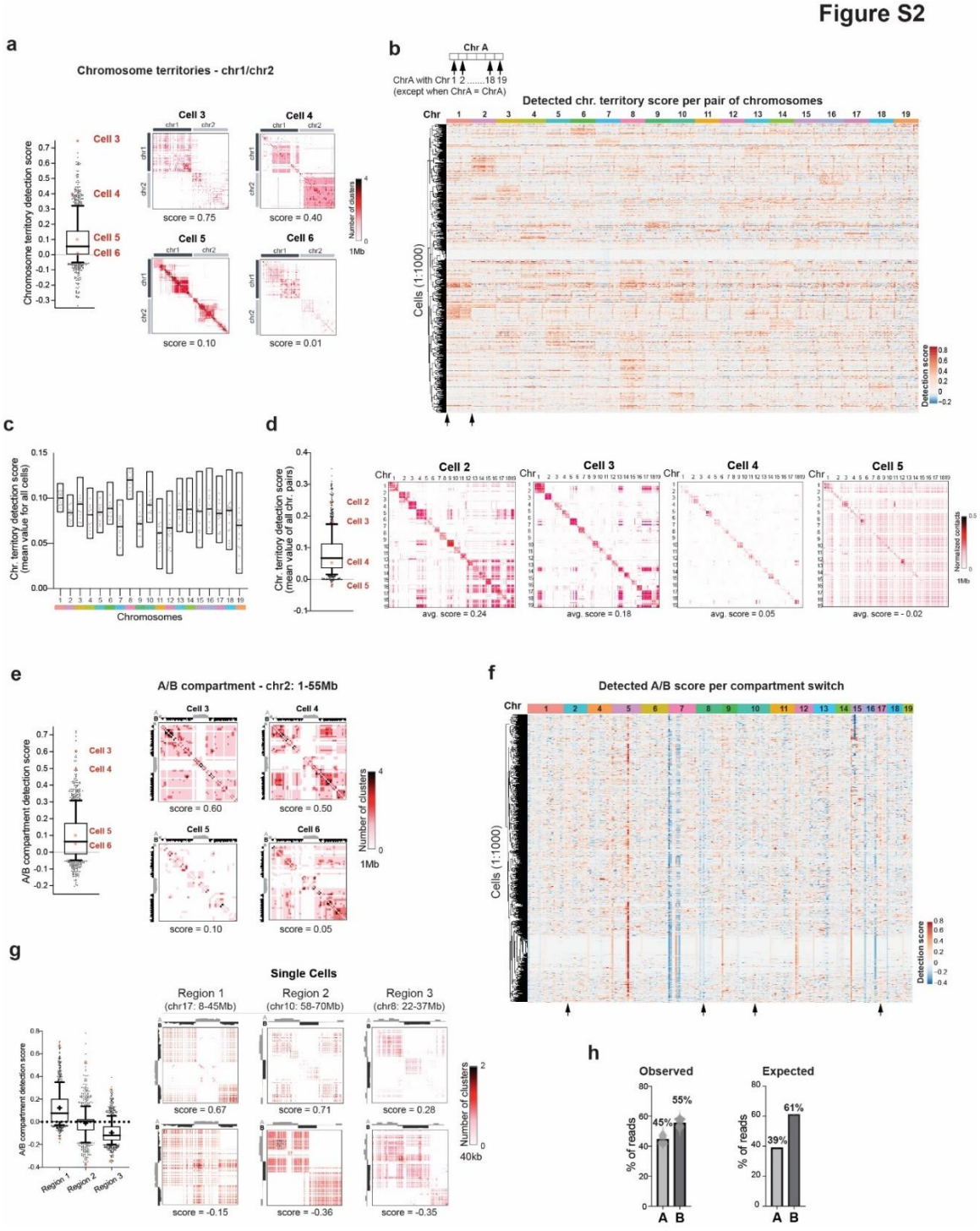


Molecules in DNA complex	Maximum # of Pairs by Proximity Ligation	Maximum # of Pairwise Combinations
2	1	1
3	1	3
4	2	6
...	...	...
10	5	45
...	...	...
100	50	4950
n	$\lfloor n/2 \rfloor$	$n(n-1)/2$

Supplementary Figure 1: scSPRITE generate single cell maps with high genomic coverage.

a. Quantification of cell aggregation. Top: number of cells in clumps pre- and post-filtration (singlets, doublets, triplets, etc). Bottom: microscope images (10x) of cells pre- and post-filtration step, scale bar 100 μm. **b.** Validation of in-nuclei barcoding step of the protocol on mixed cell population (human-mouse cells): no mixing (top middle and top right), mixing before crosslinking (bottom left), mixing after crosslinking (bottom middle), and mixing after in-nuclei restriction digest (bottom right). **c.** Schematic of the computational analysis pipeline for processing scSPRITE data. **d.** Theoretical number of contacts measured by SPRITE-derived methods and HiC-derived methods over increasing numbers of DNA molecules per complex. **e.** Maximum number of pairwise interactions that can be obtained from proximity ligation (HiC-derived methods) and complex barcoding (SPRITE-derived methods). **f.** Genome-wide coverage for the filtered 1,000 cells: the median (black triangular points) and median absolute deviation (MAD) (green circular points) values were calculated per cell using the number of reads per 1 Mb bin genome-wide (chr1-19). **g.** Genomic coverage of 20 random cell barcodes; 1 Mb bin per chromosome.

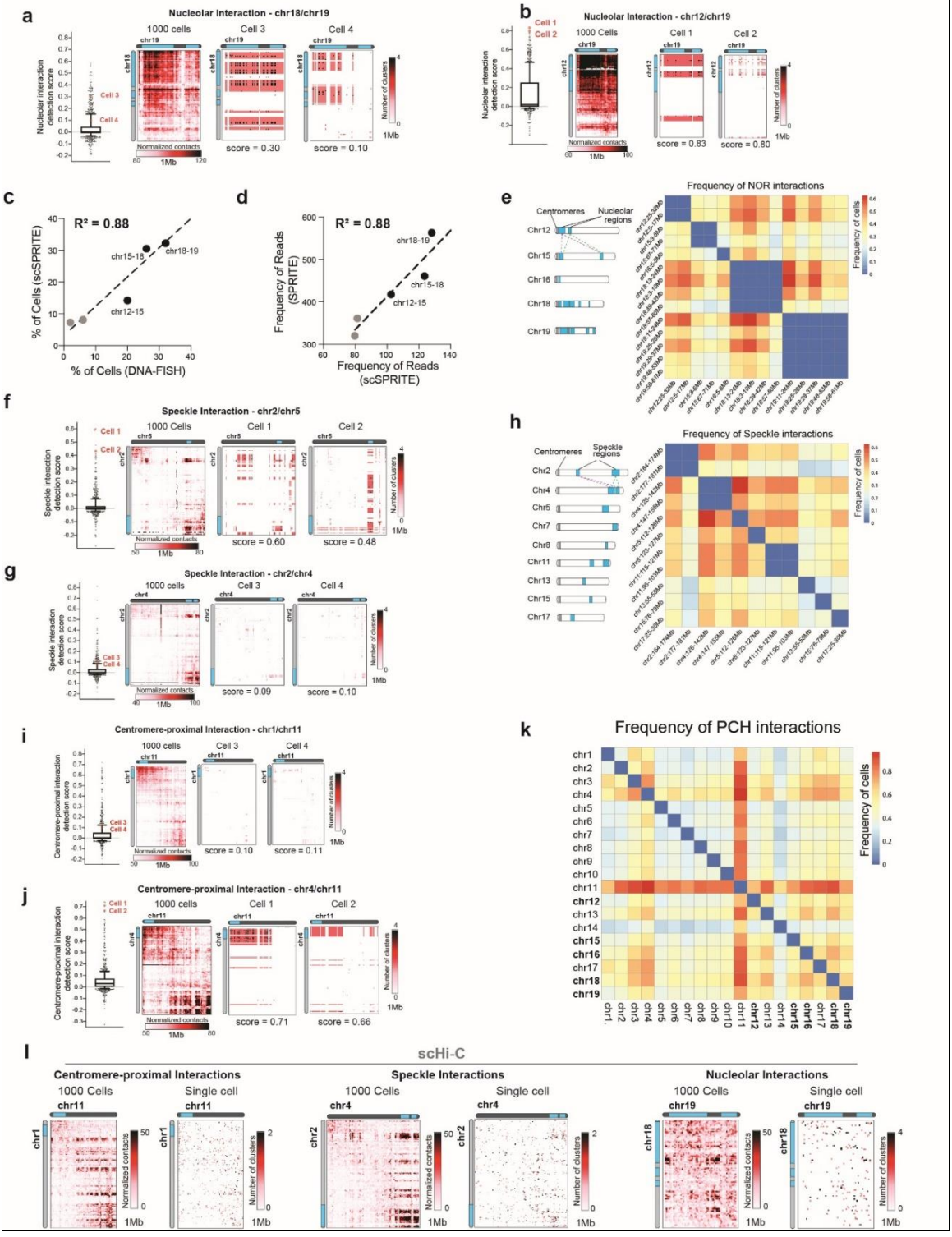
Figure S2



Supplementary Figure 2: Known chromosomal structures can be measured genome-wide in hundreds of single mESCs by scSPRITE.

a. Additional single cell examples of chromosome territory structure between chr1 and chr2; plotted as number of DNA clusters at 1 Mb resolution. Box plot represents normalized detection scores between chr1 and chr2, where whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line represents the median, red dots represent single cell examples (n = 1000 cells). **b.** Chromosome territory scores across 1000 cells (clustered based on similarity pattern). Columns represent chromosome territory detection scores for all pairs of chromosomes with the reference chromosome. Arrows represent chromosome territory scores between chr1 and chr2, which were analyzed in this paper. **c.** Quantification of chromosome territory scores with respect to each chromosome. Boxplots show the range of chromosome territory scores, the average score (black line), and individual pairs of chromosome territory scores (grey dots). **d.** Box plot represents average chromosome territory detection scores from all genome-wide (chr1-19) chromosome pairs., where whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line represents the median, red dots represent single cell examples (n = 1000 cells) (left). Additional single cell examples of genome-wide (chr1-19) chromosome territories (right). **e.** Additional single cell examples of A/B compartments detected within 0-55Mb in chr2; plotted number of DNA clusters at 1 Mb resolution (right). Box plot represents normalized detection scores between 0-55Mb in chr2, where whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line represents the median, red dots represent single cell examples (n = 1000 cells). **f.** Representation of compartment switching scores across 1,000 cells (clustered based on score similarity pattern). Columns represent the strength of compartment switching detection scores for compartments that switched from “B-to-A-to-B” or “A-to-B-to-A” genome-wide (chr1-19). Arrows represent compartment switching scores for chr2 1-55 Mb, chr8 22-37 Mb, chr10 58-70 Mb, and chr17 8-45 Mb, all of which were analyzed in this paper. **g.** Additional single cell examples of compartment switching from Region 1, Region 2, and Region 3 (right). For each region’s box plot: whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line represents the median, red dots represent single cell examples (n = 1000 cells). **h.** Expected (right) and observed (left) coverage of reads in the A and B compartment.

Figure S3

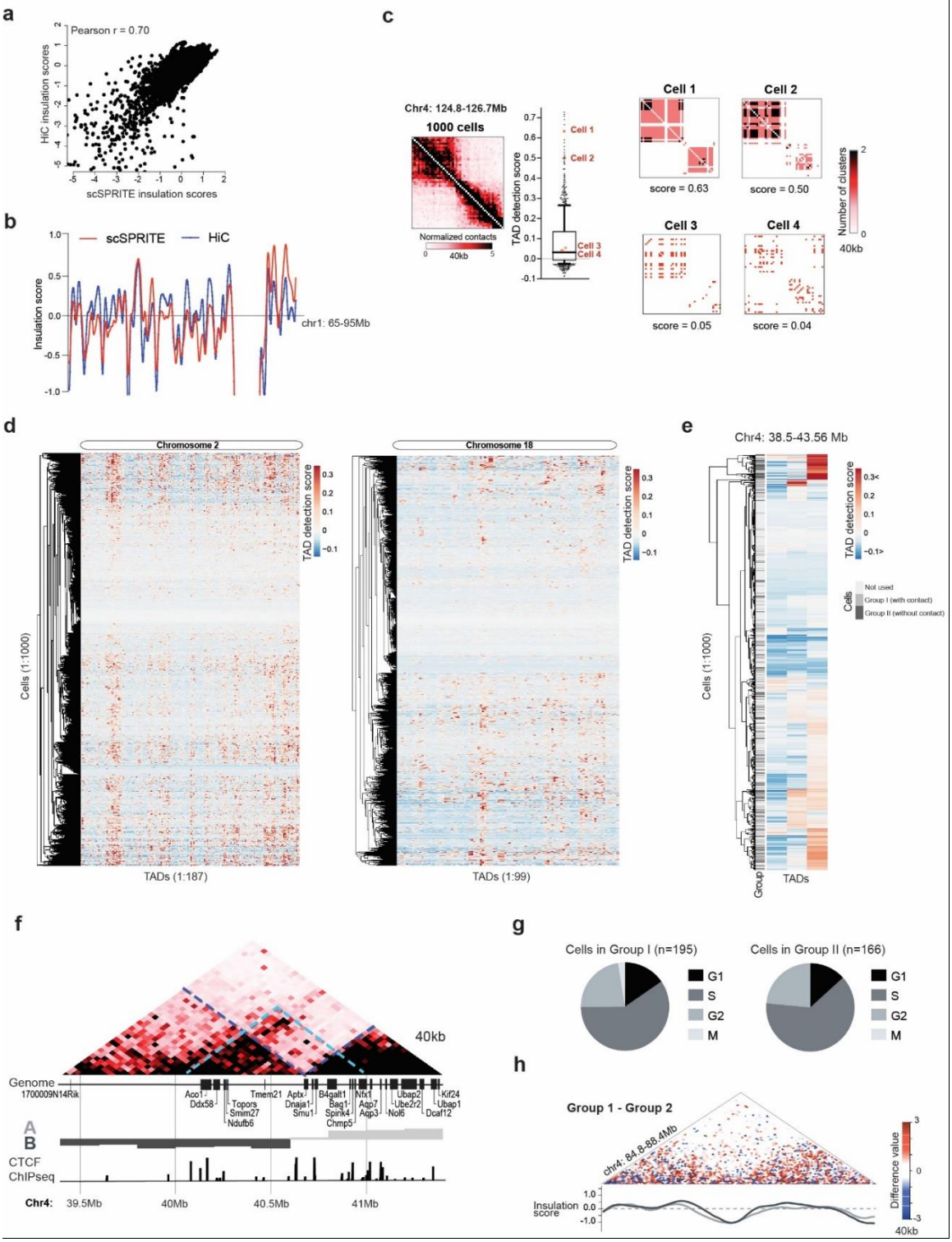

a
Nucleolar Interaction - chr18/chr19
1000 cells
Cell 3
Cell 4
Nucleolar interaction detection score
Normalized contacts
Number of clusters
score = 0.30
score = 0.10
b
Nucleolar Interaction - chr12/chr19
Cell 1
Cell 2
1000 Cells
score = 0.83
score = 0.80
c
R² = 0.88
chr15-18
chr18-19
chr12-15
% of Cells (scSPRITE)
% of Cells (DNA-FISH)
d
R² = 0.88
Frequency of Reads (SPRITE)
Frequency of Reads (scSPRITE)
e
Frequency of NOR interactions
Centromeres
Nucleolar regions
Chr12
Chr15
Chr16
Chr18
Chr19
Frequency of cells
f
Speckle Interaction - chr2/chr5
score = 0.60
score = 0.48
g
Speckle Interaction - chr2/chr4
score = 0.09
score = 0.10
h
Frequency of Speckle interactions
Speckle regions
Chr2
Chr4
Chr5
Chr7
Chr8
Chr11
Chr13
Chr15
Chr17
i
Centromere-proximal Interaction - chr1/chr11
score = 0.10
score = 0.11
j
Centromere-proximal Interaction - chr4/chr11
score = 0.71
score = 0.66
k
Frequency of PCH interactions
l
scHi-C
Centromere-proximal Interactions
Speckle Interactions
Nucleolar Interactions
1000 Cells
Single cell

Supplementary Figure 3: Higher-order structures are identified genome-wide in hundreds of single mESC by scSPRITE method.

a. Additional single cell examples of nucleolar interactions detected between chr18 and chr19; plotted number of DNA clusters at 1 Mb resolution; detection scores below contact map (right). Box plot represents normalized detection scores between chr18 and chr19, where whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line represents the median, red dots represent single cell examples (n = 1000 cells). **b.** Nucleolar interaction between chr12 and chr19: detection scores for 1000 cells (middle). Box plot where whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line represents the median, red dots represent single cell examples (n = 1000 cells). Representation of structures with max score (+1) and min. score (-1) (left) and ensemble scSPRITE heatmap (middle); contact map at 1 Mb resolution. Single cell examples (right); plotted number of DNA clusters at 1 Mb resolution. **c.** Relative correlation of the percent of cells from scSPRITE vs DNA-FISH containing inter-chromosomal interactions at specified 1 Mb regions targeted by DNA-FISH probes. Control chromosomes (grey points) and nucleolar associating chromosomes (black dots) are plotted. **d.** Relative correlation of the contact frequency from scSPRITE vs the contact frequency from SPRITE containing inter-chromosomal interactions targeted by DNA-FISH probes. Control chromosomes (grey points) and nucleolar associating chromosomes (black dots) are plotted. **e.** Frequency of cells containing inter-chromosomal nucleolar contacts (normalized to number of reads per region) for each pair of nucleolar associating chromosomes. . **f.** Single cell examples of speckle interaction detected between chr2 and chr5; plotted number of DNA clusters at 1 Mb resolution. Box plot represents normalized detection scores between chr2 and chr5, where whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line represents the median, red dots represent single cell examples (n = 1000 cells). **g.** Additional single cell examples of speckle interactions detected between chr2 and chr4; plotted number of DNA clusters at 1 Mb resolution. Box plot represents normalized detection scores between chr2 and chr4, where whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line

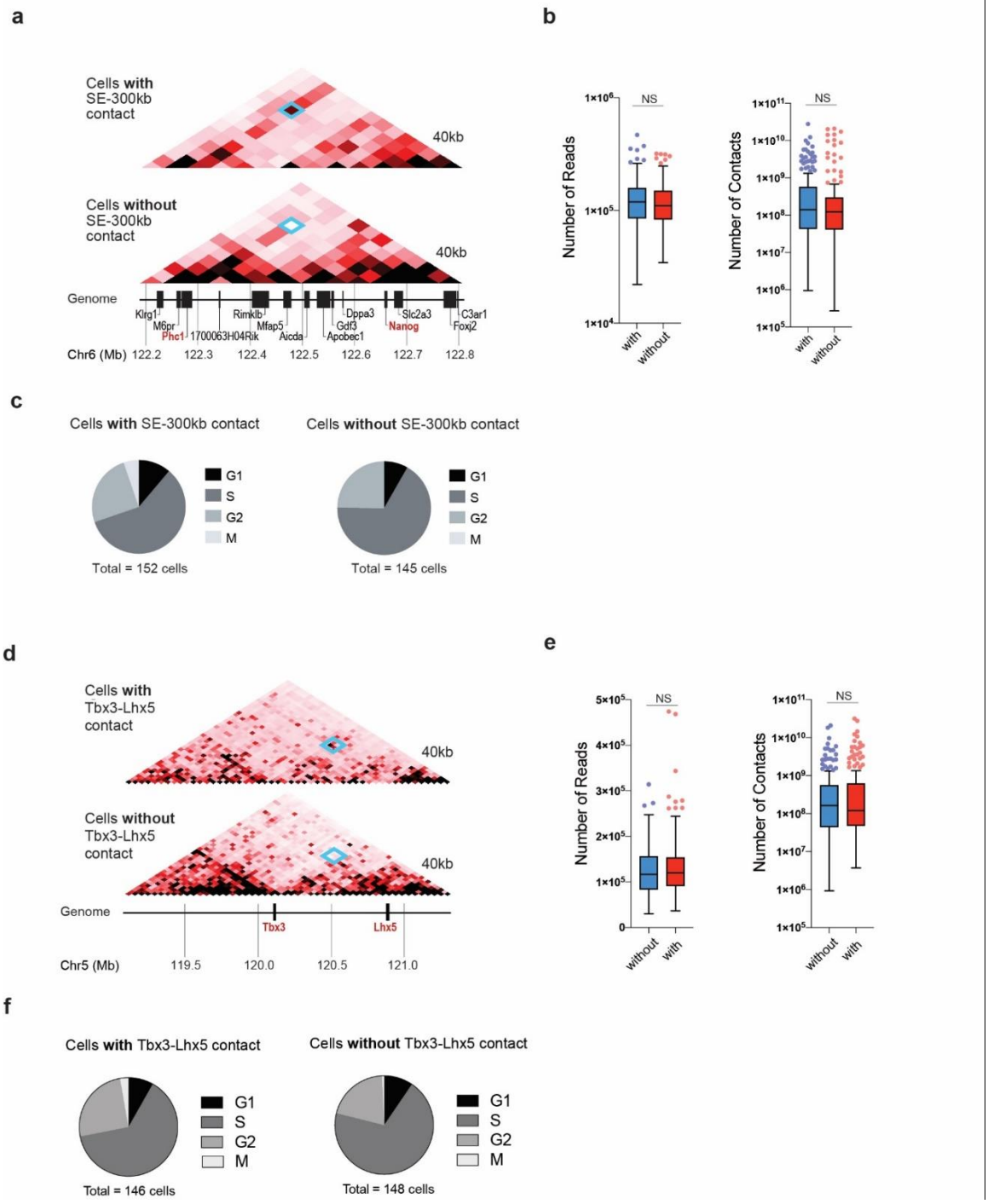
represents the median, red dots represent single cell examples (n = 1000 cells). **h.** Frequency of cells containing inter-chromosomal speckle contacts (normalized to number of reads per region) for each pair of speckle associating chromosomes. **i.** Additional single cell examples of centromere-proximal interactions detected between chr1 and chr11; plotted number of DNA clusters at 1 Mb resolution. Box plot represents normalized detection scores between chr1 and chr11, where whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line represents the median, red dots represent single cell examples (n = 1000 cells). **j.** Single cell examples of chr4 and chr11 centromere-proximal regions.

Figure S4


a
Pearson r = 0.70
HiC insulation scores
scSPRITE insulation scores
b
scSPRITE
HiC
Insulation score
chr1: 65-95Mb
c
Chr4: 124.8-126.7Mb
1000 cells
Normalized contacts
40kb
TAD detection score
Cell 1
Cell 2
Cell 3
Cell 4
score = 0.63
score = 0.50
score = 0.05
score = 0.04
Number of clusters
d
Chromosome 2
Chromosome 18
TAD detection score
Cells (1:1000)
TADs (1:187)
TADs (1:99)
e
Chr4: 38.5-43.56 Mb
TAD detection score
Cells
Not used
Group I (with contact)
Group II (without contact)
Cells (1:1000)
Group
TADs
f
40kb
Genome
1700009N14Rik
Aco1
Ddx58
Topors
Smim27
Ndufb6
Tmem21
Aptx
Dnaja1
Smu1
B4galt1
Bag1
Spink4
Chmp5
Nfx1
Aqp7
Aqp3
Ubap2
Ube2r2
Nol6
Kif24
Ubap1
Dcaf12
A
B
CTCF
ChIPseq
Chr4:
39.5Mb
40Mb
40.5Mb
41Mb
g
Cells in Group I (n=195)
Cells in Group II (n=166)
G1
S
G2
M
h
Group 1 - Group 2
chr4: 84.8-88.4Mb
Insulation score
Difference value
40kb

Supplementary Figure 4: TADs are heterogeneous units present in the genomes of individual mESCs.

a. Genome-wide correlation of insulation scores between ensemble scSPRITE and HiC[3] from mouse ES cells at 40 kb resolution. **b.** Insulation score profile of ensemble scSPRITE (red) and HiC[3] (blue) at 40 kb resolution at chr1 65-95 Mb. **c.** Additional single cell examples of TAD-like structures between 124.8-126.7Mb of chr4; plotted number of DNA clusters at 40 kb resolution; detection scores below contact map. Box plot represents normalized detection scores between 124.8-126.7Mb of chr4, where whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line represents the median, red dots represent single cell examples (n = 1000 cells). **d.** TAD detection scores across 1,000 cells (clustered based on score similarity pattern) in chr2 (left) and chr18 (right). Columns represent the strength of TAD detection scores for all TADs detected across chr2 or chr18, respectively, in ensemble scSPRITE. **e.** TAD detection scores across 1,000 cells between 38.5-48.56 Mb of chr4. Each line represents the strength of TAD detection scores in this given region from a single cell. Cells are either in Group 1 or 2 in Figure 4f or not used. **f.** Ensemble heatmap from all 1000 cells between 39.4-41.4Mb of chr4 representing strong TADs detected in bulk (blue lines), and weak emerging TADs (green line) over the A/B boundary. **g.** Fraction of cells in each cell cycle phase from the set of single cells containing (left) or lacking (right) the contact between the boundary region (Figure 4f). **h.** Difference contact map across a control region 84.8-88.4 Mb of chr4 made by subtracting the normalized contacts from cells in Group II from Group I (Figure 4f). Insulation scores for cells in Group I (dark grey) and Group II (light grey) are plotted.



Supplementary Figure 5: Structural heterogeneity in long-range interactions is revealed by scSPRITE.

a. Ensemble heatmaps across 122.2-122.8 Mb region in chr6 representing cells containing (top) or lacking (bottom) the contact between the Nanog locus and the -300 kb SE. Blue square shows the contact. **b.** Number of genome-wide reads (left) and number of genome-wide contacts (right) for groups of cells with and without the Nanog-SE interaction. For each box plot, whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line represents the median (with = 159 cells, without = 149 cells). No statistical significance between the two groups were seen based on the Kolmogorov–Smirnov two-sided test. **c.** Fraction of cells in each cell cycle phase from the set of single cells containing (left) or lacking (right) the contact between the Nanog locus and the SE 300kb upstream of Nanog. **d.** Heatmaps between 119.24-121.28Mb in chr5 of pooled cells either containing (top) or lacking (bottom) the contact between the Tbx3 locus and Lhx5. Blue square shows the contact. **e**. Number of genome-wide reads (left) and number of genome-wide contacts (right) for groups of cells with and without the Tbx3-Lhx5 interaction. For each box plot, whiskers represent the 10th and 90th percentiles, box limits represent the 25th and 75th percentiles, black line represents the median (with = 152 cells, without = 149 cells). No statistical significance between the two groups were seen based on the Kolmogorov–Smirnov two-sided test. **f.** Fraction of cells in each cell cycle phase from the set of single cells containing (left) or lacking (right) the contact between the Tbx3 locus and the Lhx5.

SUPPLEMENTARY NOTES

Supplemental Note 1: Pairwise contacts measured by SPRITE and Hi-C capture distinct structural information.

While scSPRITE captures a larger number of pairwise contacts than scHi-C (**Figure 1f, Figure S1e**), a direct comparison of the number of pairwise contacts and the specific nuclear structures identified are not exactly equivalent. Specifically, SPRITE captures contacts that are often missed or underrepresented in Hi-C data. For example, we previously showed that the types of contacts that we observe by SPRITE are dependent on the cluster sizes. Whereas small clusters (2-10 reads/cluster) capture features that are virtually identical to Hi-C, larger SPRITE clusters (>10 reads/cluster) capture longer range interactions that are under-represented in Hi-C data. Accordingly, scSPRITE contacts will be distributed across additional nuclear structures than those measured by HiC. For example, we observe 54% inter-chromosomal contacts compared with just 6% for HiC. As such, if we measured an equal number of SPRITE and Hi-C contacts, we would expect fewer of the SPRITE contacts to be present in TADs and other structures that are also observed by HiC because the total contacts would be distributed across these structures as well as the additional nuclear structures measured.

Supplemental Note 2: Impact of structural heterogeneity on function

We detected structural heterogeneity in several regions forming long-range interactions across borders of A/B compartment (**Figures 4f, 4g, S4d**) and TADs (**Figures 5a, 5b, S5a, S5d**). Although we are not aware of any reports describing functional heterogeneity among these regions (in terms of gene expression), we are hesitant to claim that these structural changes do not have an impact on gene expression levels because we cannot measure the relationship between 3D structure and gene expression in the same single cell. We therefore cannot exclude the possibility that only a small fraction of cells display change in these genes.

Supplemental Note 3: Structural heterogeneity of mESC cultured in 2i/LIF conditions

It is known that gene expression is more heterogeneous when mESCs are cultured in LIF/serum conditions and more homogeneous when mESC are cultured in 2i/LIF which promotes ground state pluripotency[1, 2]. Specifically, it is well documented that the heterogeneity of Nanog expression is primarily observed in the population of mESC grown in serum/LIF (not 2i/LIF)[3], which suggests that this limited functional heterogeneity is likely to minimize the amount of heterogeneity observed in nuclear structure. However, we can detect the differential structural state of the Nanog locus in 2i/LIF culture conditions (**Figures 5, S5**). Because we cannot measure the structural change and the expression profiles in the same single cell, we cannot exclude the possibility that only a small fraction of cells display change in Nanog expression, and therefore we are hesitant to claim that the structural changes do not have an impact on Nanog expression levels. One possible scenario is that nuclear organization of individual mES cells in 2i/LIF may be more heterogeneous or dynamic as previously thought based on the expression profiles. This hypothesis requires more direct studies.

Supplemental Note 4: Quantification of the normalized detection score

In the supplemental tables, we provide an example of normalized detection scores for chromosome territories, A/B compartments, and TADs (**Supplemental Table 1, 2, 3**). However, these scores will vary slightly each time the scores are generated due to its normalization to the expected detection score, which is generated from randomized structures. The randomized structures vary each time score calculations are run and, as a consequence, will provide a slightly different expected detection score from run-to-run.

References

1. Blinka, S. & Rao, S. Nanog expression in embryonic stem cells – an ideal model system to dissect enhancer function. *BioEssays* **39**, 1700086 (2017).

2. Blinka, S., Reimer, Michael H., Jr., Pulakanti, K. & Rao, S. Super-enhancers at the nanog locus differentially regulate neighboring pluripotency-associated genes. *Cell Reports* **17**, 19-28 (2016).

3. Kalmar, T. et al. Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. *PLOS Biology* **7**, e1000149 (2009).