

Highly Multiplexed Imaging of E. Coli  
Chromosome and Sensitive Detection of  
Single-Cell Protein

Thesis by  
Yandong Zhang

In Partial Fulfillment of the Requirements for  
the degree of  
Doctor of Philosophy

The logo for the California Institute of Technology (Caltech), featuring the word "Caltech" in a bold, orange, sans-serif font.

CALIFORNIA INSTITUTE OF TECHNOLOGY  
Pasadena, California

2019  
(Defended 05/29/2019)

© 2019

Yandong Zhang  
ORCID: 0000-0003-3291-9209

## ACKNOWLEDGEMENTS

I would like to thank my advisor Prof. Long Cai for his guidance during my PhD. I would like to thank all the Cai Lab members for their help.

## ABSTRACT

The driving force for biology research is the development of new techniques which allow high-sensitivity, high-throughput measurement in various contexts. Over the past decade, the emerging of a variety of single-cell techniques have greatly transformed our understanding of biological system. My thesis work was therefore focused on development of new single-cell techniques and use the techniques to generate new insights into biological system. Specifically, in the first part of my thesis work, we developed DNA seqFISH, a technique that allows us to image more than 100 different loci on the chromosome in single cells. We applied this technique to image *E. coli* chromosome with 50kb genomic resolution and 50nm spatial precision. Our data allows us to parse the *E. coli* chromosome structure according to their different spatial conformations and different cell-cycle stages. We identified two chromosome conformations with distinct domain structures, which is obscured from previous population-average research. We further characterized the domain structure dynamics during daughter chromosome segregation. Therefore, our data provides a high-resolution, dynamic view of *E. coli* chromosome structure.

In the second part, we developed a novel method for sensitive detection of targeted protein and its post-translational modification (PTM) isoform in single cells. Instead of depending on antibodies to distinguish targeted protein and its PTM isoform, we developed an efficient covalent barcoding strategy to barcode targeted protein inside the cells. Thereafter, targeted protein and its PTM isoform are separated by conventional gel electrophoresis, while their single-cell identity is preserved in the covalently attached oligo. By counting the attached DNA oligos using next-generation sequencing, targeted protein, and its PTM isoform can be accurately measured. We demonstrated the utility of the technology by quantification of histone protein, H2B and its mono-ubiquitination isoform, H2Bub at single-cell level. Our method revealed the single-cell heterogeneities of H2Bub/H2B ratio and its cell-cycle dynamics. Our method therefore provides an antibody-free method for sensitive detection of proteins and its isoforms in single cells.

## TABLE OF CONTENTS

Acknowledgements.....	iii
Abstract .....	iv
Table of Contents.....	v
Chapter I: Introduction .....	1
1.1 Single-cell system biology .....	1
1.2 Chromosome structure in single cells .....	2
1.2.1 Chromosome conformation capture .....	2
1.2.2 Single-cell HiC .....	4
1.2.3 Multiplexed DNA FISH.....	5
1.2.4 E. coli chromosome structure .....	6
1.3 Measure protein copy number in single cells .....	8
1.4 References .....	10
Chapter II: Highly multiplexed imaging of E. coli chromosomes reveals structural heterogeneities and dynamics during daughter chromosome segregation .....	15
2.1 Abstract.....	15
2.2 Introduction .....	15
2.3 Results and discussion.....	16
2.4 Figures .....	23
2.5 References .....	26
2.6 Supplementary information.....	28
Chapter III: Quantifying targeted protein and its post-translational modification isoforms in single cells by DNA barcoding and next-generation sequencing .....	42
3.1 Abstract.....	42
3.2 Introduction .....	42
3.3 Result and discussion .....	44
3.4 Figures .....	49
3.5 References .....	52
3.6 Supplementary information.....	54

*Chapter 1*

## INTRODUCTION

**1.1 Single-cell System Biology**

The advent of single-cell biology has greatly revolutionized our understanding of biology. For instance, by monitoring protein localization dynamics in single cells under time-lapse microscopy, researchers discovered pulsatile phosphorylation dynamics of specific transcription factors, which provides new insights into the functional regulation of transcription<sup>1,2</sup>. Another example is the development of so-called single-cell system biology, in which omics-level targets are simultaneously measured in single cells. Such high-content single-cell measurement provides unprecedented, system-level information about the single-cell states, and has been shown to be able to identify new cell types<sup>3</sup>, and reveal transcriptional dynamics of cells<sup>4</sup> and gene-regulatory networks<sup>5</sup>. In the translational research area, analyzing single cells is emerging as a new powerful tool that leads to new therapeutics. For example, isolating the neo-antigen specific single immune cells provides the first step to achieve personally precise immune therapy in cancer treatment<sup>6</sup>.

The driving force for single-cell biology is the development of new techniques which allow high-sensitivity, high-throughput measurement of single-cell components. Single-cell measurement has been challenging due to its limited amount of biological materials to be analyzed. Perhaps the most exciting development is the application of next-generation sequencing into single-cell study. For example, single-cell RNA sequencing(scRNAseq) method allows unbiased and high-throughput measurement of the transcription states of individual cells<sup>7</sup>. Besides single-cell RNA-seq, a cohort of techniques such as single-cell ATACseq (chromatin accessibility)<sup>8</sup>, single-cell HiC (3D chromosome structure)<sup>9</sup>, single-cell ChIP-seq (protein-DNA interaction)<sup>10</sup> and single-cell methylation seq (DNA methylation)<sup>11</sup> have been developed and the list has been increasing. Those new techniques provide prodigious insights into the functioning of

biological systems.

Another exciting development for single-cell biology comes from single-molecule fluorescence in situ hybridization (smFISH) developed by Raj et al.<sup>12</sup> smFISH can visualize individual RNA molecules inside the cells and thus measure the expression level of particular genes. The key advantage of smFISH is it can preserve the cell's spatial information in its native context. Recently, a great deal of progress has been made in increasing the multiplicity of smFISH so that multiple genes' expression level of single cells can be measured<sup>13-15</sup>. Our lab has devised a sequential FISH (called seqFISH) scheme so that one can increase the multiplex capacity by increasing the number of rounds of hybridization<sup>16,17</sup>. seqFISH have been demonstrated to be able to detect 10,000 genes in tissue, allowing investigation of structural organization of brain with single-cell resolution<sup>18</sup>. Alternatively, in situ sequencing methods<sup>19-21</sup> have also been developed to directly sequence transcripts inside the cells, but this method suffers from low detection efficiency.

My thesis work is driven by the need to develop new single-cell measurement techniques. In the first part, I applied the seqFISH method developed in our lab to target DNA inside the cells, and studied the chromosome organization in *E. coli* cells. The second part of my work is to develop a novel method to quantify targeted protein and its PTM isoform in single cells.

## **1.2 Chromosome Structure in Single Cells**

### **1.2.1 Chromosome Conformation Capture**

The human genome consists of over 3 billion nucleotides and needs to be compacted into a sphere smaller than a tenth of the thickness of a human hair (10 $\mu$ m). Similarly, the *E. coli* chromosome needs to be compacted 1000-fold to fit into the cells (1 $\mu$ m). Accumulating evidence has shown that the chromosome is folded into a complex, non-homogeneous structure and such spatial organization of the chromosome is tightly related

to its cellular functions, such as gene expression, replication, and segregation. For example, physical chromatin looping between the enhancer and the promoter can regulate the gene expression activity<sup>22,23</sup>. Further, recent evidence suggests that the chromosome is folded into nested

topological domains and that those domains are also involved in regulating genes by limiting enhancer-promoter interactions to only those that can occur within the domain<sup>24-26</sup>. In *E. coli*, the knockout of nucleotide-associated proteins, the main players in organizing chromosome folding, often leads to daughter chromosome segregation defects<sup>27</sup>.

To investigate the three-dimensional chromosome organization, chromosome conformation capture was developed to capture the chromatin interactions<sup>28</sup>. Coupled with high-throughput next-generation sequencing (HiC)<sup>24</sup>, one can survey all genome-wide interactions in a single experiment. In HiC, cells are first cross-linked using formaldehyde so that the chromatins that are spatially closed to each other are tethered together. The chromatin is then extracted and digested with a restriction enzyme. The restriction fragments are ligated to generate the chimeric DNA products which represent pairwise 3D interactions. A biotinylated nucleotide is also introduced at the ligation junctions, which enables specific purification of these chimeric DNAs. The crosslinks are then reversed, and DNA is purified and sequenced to identify pairwise chromatin interactions.

HiC reveals several fundamental organization principles of chromosome organization inside the mammalian cells. First, at the largest genomic scale, the chromosome is segmented into so-called A/B genomic compartments<sup>24</sup>. Essentially, the pairwise interaction matrix shows up as a “checker-board”-like pattern consisting of alternating blocks, which suggests that chromosomes are composed of two types of genomic regions that alternate along the length of the chromosomes, and the interaction frequencies between two regions of the same type tend to be higher than interaction frequencies

between regions of different types. The A/B compartments are correlated with several chromatin features, such as gene density, histone marks and DNA accessibility. Second, at smaller scale, HiC revealed the existence of sub-Mb structure that are referred to as topologically associating domains (TADs)<sup>22,29,30</sup>. Loci within TADs tend to interact more with each other than the loci outside the regions. The TADs have been thought to be involved in gene-regulatory function, such as specifying local promoter-enhancer interactions. The physical structure of TADs and how they are specified in the genome is still unclear. Finally, at even smaller genomic scale, HiC revealed point interaction, such as interactions between enhancers and promoters<sup>22,31</sup>.

Besides mammalian systems, HiC has also been applied to investigate bacteria chromosome organization. Similar to mammalian systems, the chromosome in bacteria cells is non-homogeneously organized into TAD-like structures with various sizes<sup>32-35</sup>. Transcription has been shown to play an important role in defining the domains, as long, highly transcribed genes often co-localize with domain boundaries in both *Caulobacter Crescentus* and *E. coli*<sup>33,36</sup>. Nucleoid-associated proteins (NAPs) are shown to play important roles in organizing chromosomes at various sizes in *E. coli*<sup>33</sup>. In *Bacillus subtilis*, condensin protein zips up the left arm and right arm by traveling along the chromosome<sup>35</sup>. These results have provided prodigious insights into the chromosome organization of bacteria.

### **1.2.2 Single-cell HiC**

The major limitation of usual HiC technique is that it averages across a whole population of cells. The averaging results in the loss of single-cell information and can confound the data interpretation. As illustrated by Lajoie et al.<sup>37</sup>, a “smooth” interaction matrix that shows no structures does not mean that there is no structure in the underlying genomes. Instead, it could be possible that the structures exist in every single cell and are not consistent between cells. Therefore, efforts have been made to downscale canonical HiC to single-cell HiC. In its first demonstration, single cells were isolated into single

wells and the entire HiC procedure was carried out on single cells<sup>9</sup>. With this new method, the genome conformations from single cells can be studied. Researchers found evidence for conserved topological domain organization from cell to cell, but highly variable inter-domain contacts and chromosome folding genome widely. In another study using similar method<sup>38</sup>, researchers found that the structures of individual TADs and loops vary substantially from cell to cell. By contrast, A and B compartments, lamina-associated domains, and active enhancer and promoters are organized in a consistent way on a genome-wide basis in every cell. Further scaling up the single-cell HiC technique allows researchers to profile thousands of cells from different cell-cycle stages and study the cell-cycle dynamics of chromosome structures<sup>39</sup>. Researchers found that chromosomal compartments, TADs, contact insulation, and long-range loops are governed by distinct cell-cycle dynamics. Even though in its early stages, single-cell HiC has emerged as a new powerful tool to understand chromosome structures.

### **1.2.3 Multiplexed DNA FISH**

Besides HiC, another important method for studying the 3D genome is imaging-based methods, such as live-cell fluorescent tracking and DNA fluorescent in situ hybridization (DNA FISH). Live-cell fluorescent tracking uses a modified CRISPR system<sup>40</sup> or lacO-lacI<sup>41</sup> system to label a specific DNA locus, and could track the dynamic movement of this locus inside the cells. In DNA FISH, the cells are first fixed by formaldehyde crosslinking. Oligonucleotide probes with a sequence complementary to the targeted locus are fluorescently labeled. Then the chromosome is denatured, e.g., by heating, to open the double-strand structure. The fluorescently labeled probes are allowed to hybridize to the targeted loci and the samples can be imaged under fluorescent microscopy. The merits of imaging method are that it can directly reveal the spatial structure of chromosome in situ and it is inherently single-cell measurement.

Direct Imaging chromosomes inside the cells provides complementary insights into the chromosome organization. For example, it has been shown that each individual chromosome occupied a separate territory inside the nucleus, and highly transcribed genes

tend to be on the surface of such territory<sup>42</sup>. Live-cell tracking of individual chromosomal loci has revealed distinct dynamics during different cell-cycle stages<sup>43</sup>. Using live-cell tracking and DNA FISH in *E. coli* cells, the trajectory of the chromosome rearrangement during daughter-sister segregation is established<sup>44-47</sup>.

The major limitation of this imaging method is its capacity to measure multiple discrete locations simultaneously. Recently, our lab and other labs have greatly increased the multiplicity of the traditional FISH method by using sequential rounds of hybridization. Applying the multiplexed version of DNA FISH, one can thus image multiple genomic loci inside single cells, and generate the real physical structure of individual chromosomes inside single cells. Such data will provide new, unprecedented insights into chromosome organization and answer many important questions. Zhuang et al. firstly applied multiplexed DNA FISH to study a mammalian chromosome in which they targeted all the TADs within single chromosome<sup>48</sup>. They found that the A/B compartments defined by ensemble HiC are actually spatially separated. In another work, targeting chromosomes at higher resolution (30kb) reveals that different cells have domain structures that are distinct from each other but biased to the ensemble structure determined from HiC<sup>49</sup>. Recently, Nollmann et al. demonstrate simultaneous visualization of genome organization and transcription in intact *Drosophila* embryos. They unveil the changes in 3D chromatin organization occurring upon transcriptional activation and homologous chromosome unpairing during the awakening of the zygotic genome in intact *Drosophila* embryos<sup>50</sup>. Boettiger et al. demonstrated simultaneous detection of chromosome structure and multiple RNA species in single cells. They identified cell-type-specific physical borders between active and polycomb-repressed DNA and unexpected polycomb-independent borders<sup>51</sup>. At its early stage, multiplex DNA FISH is expected to generate more and more exciting results which will provide new insights into chromosome organization in single cells.

#### **1.2.4 *E. coli* chromosome structure**

*E. coli* has a 4.6Mb circular chromosome that needs to be folded 1,000 times to fit into

the cells. A great deal of effort has been made to understand how the *E. coli* chromosome is organized inside the cells. The DNA is supercoiled to be condensed inside the cells, as evidenced by early electron microscopy studies and polymer modeling<sup>52</sup>. Long-range interactions also play very important roles in organization principles. Early DNA-FISH studies have suggested that the Ori (the starting point of replication) region and the Ter (the ending point of replication) region form macrodomains separately<sup>53</sup>. A groundbreaking discovery about *E. coli* chromosome structure was made by Boccard et al.<sup>54</sup> In their work, they designed a genetic recombination assay to measure the contact frequency of different regions on the chromosome, conceptually similar to chromosome conformation capture. They identify six different domains inside the *E. coli* chromosomes, demonstrating the non-homogeneous organization of the *E. coli* chromosome. However, the molecular mechanism for those domain organizations remains largely unknown, except for the Ter domain where the matS/matP system has been identified<sup>55</sup>.

Unlike in eukaryotic systems, the replication and segregation of daughter chromosomes occur simultaneously. The replication starts from a single position, called OriC and two replication forks go bi-directionally to replicate the right arm and the left arm, respectively, until they converge at the terminus region. In slow growth conditions, the initiation of replication occurs after the finish of previous round of replication and cell division. In faster growth conditions, replication could initiate before the previous round of replication has finished. Using DNA-FISH and live-cell imaging, the positioning of chromosomes during replication and segregation has been studied extensively<sup>44</sup>. In slow-growing conditions, the two newly replicated Ori are segregated into two daughter cells very quickly, leading to the segregation of two daughter cells, while the Ter region remains at the center of the cells. Then at certain point which has not been fully characterized, the replicated left arm or right arm will pass the Ori, becoming the new leading part of the daughter chromosome, while Ter remains largely in the middle of the cells. This results in a “sausage” shape of the *E. coli* chromosome where the left and the right chromosome arms are on the opposite halves and the origin is in the middle of the daughter cells, and a

small genomic region between the left arm and right arm is believed to be stretched<sup>45,47,56</sup>.

Recently, HiC has been used to investigate the *E. coli* chromosome structure<sup>33</sup>. Consistent with previous results, HiC identified long-range interaction domains. Furthermore, HiC studies reveal the contribution of major evolutionarily conserved proteins in such *E. coli* chromosome organization. Specifically, the condensing complex MukBEF and the ubiquitous nucleoid-associated protein HU promote DNA contacts in the megabase range outside the Ter region. Within the Ter region, the MatP protein prevents MukBEF activity, and contacts are restricted to the ~280kb range, indicating the Ter domain has distinct properties. These results show that the *E. coli* chromosome is organized by a complex and intertwined network of contacts.

The previous imaging studies and HiC studies about *E. coli* chromosome structure have generated a major gap: while imaging studies, such as DNA FISH, have revealed highly heterogeneous, dynamic pictures of the *E. coli* chromosome during segregation, DNA FISH could only detect a few loci and cannot provide global *E. coli* pictures with high resolution. On the other hand, HiC revealed the whole genome structure with high resolution, while HiC studies only generated a static description of the *E. coli* chromosome by averaging the chromosome structures from the whole population. To fill in the gap, in the first part of my thesis, we developed the highly multiplexed DNA seqFISH method and used this method to image *E. coli* cells at different cell-cycle stages and different conformations. Our data revealed that *E. coli* cells have different conformations, which reconciles the previous imaging studies and HiC results. We further revealed the dynamic changes of chromosome conformations during chromosome segregation.

### **1.3 Measure Protein Copy Number in Single Cells**

While a great deal of progress has been made in detecting nucleic acids (DNA and RNA) in single cells, protein detection methods in single cells are still very limited. In the second part of my thesis, we aimed to develop a novel method to measure the copy number of a

target protein and its PTM isoforms in single cells.

Quantifying protein in a single cell goes back to the development of fluorescent proteins. A protein of interest can be tagged with fluorescent protein and its expression level in single cells can be measured by microscopy or flow cytometry. The limitation with fluorescent protein tagging is that it usually cannot resolve the PTMs of target protein, and is therefore not suitable for monitoring the PTM states of targeted protein. Alternatively, antibody-based immunostaining can also be used to quantify the protein inside the cells. PTM of proteins can also be targeted if a specific antibody is available. Over the past few years, highly multiplexed antibody-based single-cell protein quantification has been achieved. In one technique, named CyTOF<sup>57</sup>, the antibodies are labeled with distinct rare-earth metals and are used to label multiple protein targets inside the single cells. The single cells are analyzed in a flow-cytometry manner where different rare-earth metals are detected by mass spectrometry to quantify the proteins from single cells. The CyTOF can detect more than 100 different protein targets inside single cells. In another technique<sup>58</sup>, the antibodies are made into spatially separated arrays in a microchamber where a single cell is isolated and lysed. The proteins from single cells can be assayed in the ELISA manner by the antibody arrays. Recently, DNA-barcoded antibodies have been used to label the proteins inside the cells, and the protein quantification can be converted into the quantification of DNA oligo using next-generation sequencing<sup>59</sup>. This method holds the potential to detect both transcriptome and proteome from single cells<sup>60,61</sup>. The DNA in DNA-barcoded antibodies can also be read out using in situ hybridization, which can preserve the spatial information of the single cells<sup>62,63</sup>.

The fundamental limitation of the antibody-based detection method is that it depends on high- quality, high-specificity, high-affinity antibodies. Such antibodies are not always available, even though great efforts are being made to generate the repertoire of the whole proteome. Therefore, currently, an unbiased, discovery-driven, whole-proteome assay from a single cell is still not possible. In addition, for a certain panel of antibodies,

extensive orthogonal testing needs to be done to make sure there is no cross-reactivity. This is especially challenging to achieve when the PTM of a protein and the different splicing isoform of a protein are targeted because they are highly similar to each other. Finally, due to different binding affinities of different antibodies, the quantities of different proteins cannot be compared directly, and this might introduce artifacts during the analysis step.

Therefore, to address the issues associated with the antibody-based detection method, we developed a novel method to quantify the protein and its PTM isoform in single cells. Our method therefore provides an alternative useful tool for biologists to study protein in single cells.

#### 1.4 Reference

1. Cai, L., Dalal, C. K. & Elowitz, M. B. Frequency-modulated nuclear localization bursts coordinate gene regulation. *Nature* **455**, 485–490 (2008).
2. Regot, S., Hughey, J. J., Bajar, B. T., Carrasco, S. & Covert, M. W. High-sensitivity measurements of multiple kinase activities in live single cells. *Cell* **157**, 1724–1734 (2014).
3. Zeisel, A. et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138–1142 (2015).
4. Treutlein, B. et al. Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. *Nature* **534**, 391–395 (2016).
5. Buganim, Y. et al. Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell* **150**, 1209–1222 (2012).
6. Valdes-Mora, F. et al. Single-Cell Transcriptomics in Cancer Immunobiology: The Future of Precision Oncology. *Front. Immunol.* **9**, 2582 (2018).
7. Brennecke, P. et al. Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10**, 1093–1095 (2013).
8. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
9. Nagano, T., Wingett, S. W. & Fraser, P. Capturing Three-Dimensional Genome Organization in Individual Cells by Single-Cell Hi-C. *Methods Mol. Biol.* **1654**, 79–97 (2017).
10. Rotem, A. et al. Single-cell ChIP-seq reveals cell subpopulations defined by

- chromatin state. *Nat. Biotechnol.* **33**, 1165–1172 (2015).
11. Karemaker, I. D. & Vermeulen, M. Single-Cell DNA Methylation Profiling: Technologies and Biological Applications. *Trends Biotechnol.* **36**, 952–965 (2018).
  12. Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* **5**, 877–879 (2008).
  13. Lubeck, E., Coskun, A. F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nat. Methods* **11**, 360–361 (2014).
  14. Codeluppi, S. et al. Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat. Methods* **15**, 932–935 (2018).
  15. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
  16. Shah, S., Lubeck, E., Zhou, W. & Cai, L. In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus. *Neuron* **92**, 342–357 (2016).
  17. Shah, S. et al. Dynamics and Spatial Genomics of the Nascent Transcriptome by Intron seqFISH. *Cell* **174**, 363–376.e16 (2018).
  18. Eng, C.-H. L. et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **568**, 235–239 (2019).
  19. Lee, J. H. et al. Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nat. Protoc.* **10**, 442–458 (2015).
  20. Lee, J. H. et al. Highly multiplexed subcellular RNA sequencing in situ. *Science* **343**, 1360–1363 (2014).
  21. Wang, X. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361**, (2018).
  22. Jin, F. et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* **503**, 290–294 (2013).
  23. Deng, W. et al. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* **149**, 1233–1244 (2012).
  24. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
  25. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
  26. Gibcus, J. H. & Dekker, J. The hierarchy of the 3D genome. *Mol. Cell* **49**, 773–782 (2013).
  27. Helgesen, E., Fossum-Raunehaug, S. & Skarstad, K. Lack of the H-NS Protein

- Results in Extended and Aberrantly Positioned DNA during Chromosome Replication and Segregation in *Escherichia coli*. *J. Bacteriol.* **198**, 1305–1316 (2016).
28. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
  29. Nora, E. P. et al. Spatial partitioning of the regulatory landscape of the X-inactivation center. *Nature* **485**, 381–385 (2012).
  30. Sexton, T. et al. Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* **148**, 458–472 (2012).
  31. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109–113 (2012).
  32. Le, T. B., Imakaev, M. V., Mirny, L. A. & Laub, M. T. High-resolution mapping of the spatial organization of a bacterial chromosome. *Science* **342**, 731–734 (2013).
  33. Lioy, V. S. et al. Multiscale Structuring of the *E. coli* Chromosome by Nucleoid-Associated and Condensin Proteins. *Cell* **172**, 771–783.e18 (2018).
  34. Marbouty, M. et al. Condensin- and Replication-Mediated Bacterial Chromosome Folding and Origin Condensation Revealed by Hi-C and Super-resolution Imaging. *Mol. Cell* **59**, 588–602 (2015).
  35. Wang, X., Brandão, H. B., Le, T. B. K., Laub, M. T. & Rudner, D. Z. *Bacillus subtilis* SMC complexes juxtapose chromosome arms as they travel from origin to terminus. *Science* **355**, 524–527 (2017).
  36. Le, T. B. & Laub, M. T. Transcription rate and transcript length drive formation of chromosomal interaction domain boundaries. *EMBO J.* **35**, 1582–1595 (2016).
  37. Lajoie, B. R., Dekker, J. & Kaplan, N. The Hitchhiker’s guide to Hi-C analysis: practical guidelines. *Methods* **72**, 65–75 (2015).
  38. Stevens, T. J. et al. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* **544**, 59–64 (2017).
  39. Nagano, T. et al. Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* **547**, 61–67 (2017).
  40. Chen, B. et al. Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell* **155**, 1479–1491 (2013).
  41. Milbredt, S. & Waldminghaus, T. BiFCROS: A Low-Background Fluorescence Repressor Operator System for Labeling of Genomic Loci. *G3* **7**, 1969–1977 (2017).
  42. Cremer, T. & Cremer, M. Chromosome territories. *Cold Spring Harb. Perspect. Biol.* **2**, a003889 (2010).
  43. Espeli, O., Mercier, R. & Boccard, F. DNA dynamics vary according to macrodomain topography in the *E. coli* chromosome. *Mol. Microbiol.* **68**, 1418–

- 1427 (2008).
44. Nielsen, H. J., Li, Y., Youngren, B., Hansen, F. G. & Austin, S. Progressive segregation of the Escherichia coli chromosome. *Mol. Microbiol.* **61**, 383–393 (2006).
  45. Wang, X., Liu, X., Possoz, C. & Sherratt, D. J. The two Escherichia coli chromosome arms locate to separate cell halves. *Genes Dev.* **20**, 1727–1731 (2006).
  46. Woldringh, C. L., Hansen, F. G., Vischer, N. O. E. & Atlung, T. Segregation of chromosome arms in growing and non-growing Escherichia coli cells. *Front. Microbiol.* **6**, (2015).
  47. Liu, X., Wang, X., Reyes-Lamothe, R. & Sherratt, D. Replication-directed sister chromosome alignment in Escherichia coli. *Mol. Microbiol.* **75**, 1090–1097 (2010).
  48. Wang, S. et al. Spatial organization of chromatin domains and compartments in single chromosomes. *Science* **353**, 598–602 (2016).
  49. Bintu, B. et al. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* **362**, eaau1783 (2018).
  50. Cardozo Gizzi, A. M. et al. Microscopy-Based Chromosome Conformation Capture Enables Simultaneous Visualization of Genome Organization and Transcription in Intact Organisms. *Mol. Cell* (2019). doi:10.1016/j.molcel.2019.01.011
  51. Mateo, L. J. et al. Visualizing DNA folding and RNA in embryos at single-cell resolution. *Nature* **568**, 49–54 (2019).
  52. Miller, W. G. & Simons, R. W. Chromosomal supercoiling in Escherichia coli. *Mol. Microbiol.* **10**, 675–684 (1993).
  53. Niki, H., Yamaichi, Y. & Hiraga, S. Dynamic organization of chromosomal DNA in Escherichia coli. *Genes Dev.* **14**, 212–223 (2000).
  54. Valens, M., Penaud, S., Rossignol, M., Cornet, F. & Boccard, F. Macrodomain organization of the Escherichia coli chromosome. *EMBO J.* **23**, 4330–4341 (2004).
  55. Mercier, R. et al. The MatP/matS Site-Specific System Organizes the Terminus Region of the E. coli Chromosome into a Macrodomain. *Cell* **135**, 475–485 (2008).
  56. Wang, X., Possoz, C. & Sherratt, D. J. Dancing around the divisome: asymmetric chromosome segregation in Escherichia coli. *Genes Dev.* **19**, 2367–2377 (2005).
  57. Kay, A. W., Strauss-Albee, D. M. & Blish, C. A. Application of Mass Cytometry (CyTOF) for Functional and Phenotypic Analysis of Natural Killer Cells. *Methods Mol. Biol.* **1441**, 13–26 (2016).
  58. Shi, Q. et al. Single-cell proteomic chip for profiling intracellular signaling pathways in single tumor cells. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 419–424 (2012).
  59. Shahi, P., Kim, S. C., Haliburton, J. R., Gartner, Z. J. & Abate, A. R. Abseq: Ultrahigh-throughput single cell protein profiling with droplet microfluidic barcoding. *Sci. Rep.* **7**, 44447 (2017).

60. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
61. Mimitou, E. et al. Expanding the CITE-seq tool-kit: Detection of proteins, transcriptomes, clonotypes and CRISPR perturbations with multiplexing, in a single assay. *bioRxiv* 466466 (2018). doi:10.1101/466466
62. Goltsev, Y. et al. Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell* **174**, 968–981.e15 (2018).
63. Agasti, S. S. et al. DNA-barcoded labeling probes for highly multiplexed Exchange-PAINT imaging. *Chem. Sci.* **8**, 3080–3091 (2017).

## HIGHLY MULTIPLEXED IMAGING OF E. COLI CHROMOSOMES REVEALS STRUCTURAL HETEROGENEITIES AND DYNAMICS DURING DAUGHTER CHROMOSOME SEGREGATION

### **2.1 Abstract**

In this work, we applied multiple sequential rounds of DNA FISH (DNA seqFISH) to image around 100 loci on the E. coli chromosome in single cells with 50kb genomic resolution and 50nm spatial precision. By correlating the spatial pattern from images and pairwise distance matrices, we identified two chromosome conformations with distinct domain structures, which have been obscured from previous population-average research. By parsing the E. coli chromosome structure according to different spatial conformations and different cell-cycle stages, we further characterized the domain structure dynamics occurring during the conformation transition from Ori leading to right arm/left arm leading in daughter chromosome segregation. Therefore, our data fills the gap between the imaging and chromosome conformation capture and provides a high-resolution, dynamic view of the E. coli chromosome.

### **2.2 Introduction**

The E. coli chromosome needs to be compacted 1000-fold to fit into cells, while still maintaining important cellular functions in replication and transcription<sup>1</sup>. Genetics<sup>2,3</sup>, fluorescent imaging,<sup>4,5</sup> and chromosome conformation capture<sup>6,7</sup> studies have shown that the E. coli chromosomes are organized into macrodomains/CIDs (chromosomal interacting domains)<sup>8</sup>, similar to the eukaryotic TADs<sup>9,10</sup>. The domain boundaries are often colocalized with highly transcribed genes<sup>8,11</sup>, and certain domains are organized by specialized DNA-binding proteins<sup>12,13</sup>. Recently, single-cell HiC and imaging studies showed that chromosome structure is stochastic and heterogeneous in single cells<sup>14-16</sup>. Therefore, characterizing the domain heterogeneity in a single E. coli chromosome will

provide new insights into the prokaryotic chromosome organizations. Moreover, chromosome replication and segregation are coupled in *E. coli*, and the dynamic conformational change during cell cycle is important to elucidate. While live-cell imaging experiments can visualize the dynamics of individual loci<sup>17,18</sup>, the dynamics of whole genome is difficult to track because only a few fluorescent proteins can be imaged at a time. In this work, we demonstrated a multiplexed DNA-FISH method to directly visualize around 100 chromosomal loci in *E. coli* cells. Based on these highly multiplexed single-cell chromosome structure data, we revealed the heterogeneities of chromosome organization in daughter chromosomes. By computationally reconstructing the trajectory of chromosome movements, we unveiled the dynamics of domain structures in the *E. coli* chromosome during cell-cycle progression.

### **2.3 Results and discussion**

We devised a non-barcoded sequential DNA-FISH (DNA seqFISH) scheme to image over 100 loci which cover the whole *E. coli* chromosome (**Figure 1a**). In every round of hybridization, four loci were labeled with four spectrally distinct colors. Their spatial positions were imaged, followed by stripping off the signal. The hybridization, imaging, and stripping were repeated multiple times until the whole chromosome was imaged. Specifically, the *E. coli* chromosome was first denatured and hybridized with the primary probes to the targeted loci, which are roughly evenly distributed along the genome. The probe contains 35nt “primary” sequence complementary to the genomic sequence of the chromosome, and a nongenomic 20nt overhang. A readout probe, which is complementary to the overhang sequence on the corresponding primary probe targeting a specific locus, was labeled with fluorescent dye. In every round of seqFISH, four different readout probes were flowed in to label four distinct loci. After imaging, the cells were treated with 60% (v/v) formamide solution to strip off the readout probes (**Supplementary Figure 1**). The formamide solution will destabilize the hybridization between overhang and readout sequences, but does not affect the primary probes too much, due to the melting temperature difference between the primary probe and the

readout sequence (35nt vs 20nt). We performed 27 rounds of hybridization in total, to sequentially label and image all the loci on the *E. coli* chromosome. Several example images in the DNA seqFISH experiment were shown in **Figure 1b**. We found different loci have various signal intensity and detection efficiency (**Supplementary Figure 2**). The loci with low detection efficiency were removed, and a total of 87 loci were used in further analysis. The loci were fitted to a 2D Gaussian model and the centroid was used to estimate loci position. The localization accuracy was determined to be 50nm (**Supplementary Figure 3**).

We applied DNA seqFISH to image the *E. coli* cells growing at slow rate. Under this growth condition, the *E. coli* chromosome initiates replication from the replication origin, and two replication forks proceed bidirectionally to replicate left arm and right arm, respectively. Two daughter chromosomes are segregated into two daughter cells within one cell cycle, before new replication is initiated<sup>19</sup>. For each locus, the fraction of cells in which two dots could be detected is shown in **Supplementary Figure 4a**. Replication origin (Ori) has the highest fraction, and end of replication (Ter) has the lowest, reflecting the fact that Ori is the first replicated and Ter is the last replicated. The number of dots detected in each cell increases with the cell size, an indicator commonly used to mark out cell-cycle progression. The detection efficiency is estimated to be around 20–40% (**Supplementary Figure 4b**).

**Figure 1c** shows the reconstructed chromosome images in several example cells from different cell-cycle stages. The loci are color barcoded according to the genomic positions (Red: Ori(O); Green: right arm(R); Purple: left arm(L); Blue: Ter(T)). We considered the relative positioning of these four regions on the chromosomes and identified the conformations that would recapitulate previous results from DNA-FISH or FROS experiments<sup>20–23</sup>. At the earliest cell-cycle stage, only one Ori region (Red) was observed in the cells, and the *E. coli* chromosome was organized into a “sausage”-shape, with the left arm and the right arm on the opposite cell halves, and the replication origin (Ori) close to the middle cells. The positioning of Ter region is somewhat flexible

(**Supplementary Figure 5a**). Thereafter (stage 2), the Ori regions from the two new replicated daughter chromosomes are at the two polar positions of the cells, flanked by the left arm and right arm, while the Ter region is at the center of the cells (OL/R T L/RO). Then at later stages, the right arm or left arm passes over the Ori in one of the daughter cells, while in the other daughter cell Ori is still leading the chromosome (LOR T L/RO or ROL T L/RO). Finally, the right arm and left arm passes over the Ori in both daughter cells, while the Ter remains in the center of the cells, forming one translational symmetry (ROL T ROL) conformation and two mirror-symmetry (LOR T ROL and ROL T LOR) conformations. More examples are shown in **Supplementary Figure 5a**. The average cell size gradually increases over stage 2 and after (**Supplementary Figure 5c**), confirming that these snapshots reflect conformational changes during cell-cycle progression. In summary, our DNA seqFISH data captures the chromosome structures from different cell-cycle stages, which are consistent with previous results but have much higher resolution.

We next calculated the mean distance between each pair of loci and constructed the pairwise distance matrix. Cells from stage 2 and later stages are selected. Since in those cells the Ter are localized in the center, we divided the cells into two halves from the center, isolating single daughter chromosomes from each half cell. The pairwise distance within daughter chromosomes was calculated and mean pairwise distance was used to construct the matrix.

We firstly compared our pairwise distance matrix with published HiC data of *E. coli* from the same growth condition at the designed probe positions. The pairwise distance matrix from seqFISH shows a high similarity to the HiC contact frequency matrix (**Figure 2a**). Accordingly, the mean distance and contact frequency measured from HiC have high correlation (Pearson correlation 0.87, **Supplementary Figure 6**). Such high correlation between mean distance and contact frequency is also observed in mammalian cells<sup>16,24</sup>, as well as in *Drosophila* embryo cells<sup>25</sup> across different resolutions, indicating that this is probably a universal property. However, the scaling coefficient in *E. coli* ( $3.65 \pm 0.04$ ) is

lower than that in mammalian cells ( $4.99 \pm 0.05$ )<sup>16</sup>, which reflects distinct chromatin folding in *E. coli*.

We defined the insulation score (**Supplementary Figure 7**) to identify the domain boundaries and measure their insulation strengths, as in previous HiC work. We identified 11 domain boundaries in total (**Figure 2b**), which divide the chromosome into domains that agree well with previous imaging, genomic, and HiC studies<sup>2,5,6</sup>. Similarly with HiC studies, we found ribosomal operons colocalized with domain boundaries, indicating that transcription plays an important role in shaping the chromosomes.

Previous imaging experiments and our data have shown that in the late cell-cycle stages, the daughter chromosomes have Ter-Right arm-Ori-Left arm (TLOR) or Ter-Left arm-Ori-Right arm (TROL) conformations (**Figure 1c**). We asked whether there was a domain structure difference between the two conformations. We selected daughter chromosomes that had either TLOR or TROL patterns based on their images and calculated their mean pairwise distance matrices. Strikingly, we observed that two pairwise distance matrices have very distinct patterns (**Figure 2c and d**). The insulation score profiles calculated from each pairwise distance matrix show that while most domain boundaries could still be identified in both configurations, their insulation strengths are different. The most significant differences occur at the Ter domain boundaries. In the TROL configuration, locus 38 (39' on a 100min genomic map) forms a clear domain boundary (high insulation score) that separates the left arm from the Ter domain. Such domain boundary is very weak in ROLT configurations. In contrast, the domain boundary around locus 22 (26') clearly separates the right arm from the Ter domain in the ROLT configuration, but this is not obvious in the LORT configuration. Along with this, the distance between the Ter region and the left arm is large in the LORT conformation and vice versa in the ROLT configuration, which agrees with imaging results. These results revealed the domain structure heterogeneity in daughter chromosomes that is obscured by population-average measurement, and suggest that the Ter domain boundaries might play an important role in the global deposition pattern of daughter chromosomes.

During the chromosome segregation, the right arm or left arm passed the Ori to establish the ROLT or LORT configurations in daughter chromosomes. Previous live-cell imaging can only visualize a few loci during this conformation transition, and the global chromosome conformations are missing in the contexts of domain structures. Our data allows us to investigate the dynamics of domain structural change at the whole-chromosome level during this process. We first parsed the chromosomes according to their different stages and different spatial conformations. Daughter chromosomes were aligned so that the Ter is at the rightmost position. The vectors from the left arm to Ori and from the right arm to Ori along the long axis in each single cell were calculated and plotted in **Figure 3a**. Two major axial regions were identified in the plot: In the vertical axis (indicated by the dashed arrow), the right-arm-to-Ori vector is positive, meaning the right arm is not passing the Ori; in contrast, the left-arm-to-Ori vector flipped direction from positive to negative values, which corresponds to the process in which the left arm passes over the Ori region. In the horizontal axis (indicated by the dashed arrow), the left-arm-to-Ori direction does not change, but the right-arm-to-Ori vector flipped direction from positive to negative values, which corresponds to the process in which the right arm passes over the Ori regions. We took the cells along these two axes and divided them into overlapping bins and calculated the mean pairwise distance matrices (**Figure 3a**).

Based on the pairwise distance matrices from different stages and different conformations, we built a pseudo-time trajectory that reflects the conformation changes during the transition from Ori-leading to right-arm-/left-arm-leading conformation. First, the pairwise distance matrices of different stages and different conformations were visualized in their PC space (**Figure 3b, Supplementary Figure 8**). The first two principal components, which account for 50% of the total variance, show a clear bifurcation trajectory. The two trajectories are in alignment with the two axes we defined previously, demonstrating the emergence of two different conformations. We then plotted the pairwise distance matrices along two trajectories (**Figure 3c**). The pairwise distance matrices were rotated so that the Ori region is in the center for easy visualization. We found that in the very first stage, the off-diagonal center is at the Ori region, indicating a

conformation that Ori is leading the chromosome, and left arm/right arm are flanking along the long axis of the cells. Then, along the pseudo-time trajectory, the off-diagonal center moves either to the right arm or left arm. The loci along the right arm and left arm become the leading part of the chromosome and are flanked by their neighboring regions along the cellular long axis. The anti-diagonal axis becomes more and more asymmetrical along the trajectory. This asymmetry indicates that the two regions flanking the leading part are compacted differentially, corresponding to the right arm and left arm being stretched in their respective configurations during this process.

We investigated the dynamics of domain structures (**Figure 3c, Supplementary Figure 9**) along the pseudo-time trajectory. We found that in the ROLT trajectory, the domain boundaries on the right-arm side (locus 80 (93'), locus 4 (3'), locus 14 (16'), locus 22(26')) show distinct dynamics (**Figure 3d**). For domain boundary locus 80, its insulation score decreases in the early stages and then increases in later stages. The insulation score for locus 4 keeps decreasing over the trajectory. The insulation score for locus 14 does not change much over this process and the locus 22's insulation score keeps increasing during this process. We believe that such dynamics might reflect that the domains are dynamically merged and re-structured during the chromosome segregations.

In addition, the insulation score for domain boundary locus 22 gradually increases along the ROLT trajectory, which agrees with our previous result that locus 22 forms a strong domain boundary in the final ROLT configuration. The domain boundary of locus 22 segmented this chromosomal region into right-arm domain and Ter domain (**Figure 3e**). We found that the loci within the right-arm domain maintain an almost constant distance from each other along the pseudo-time trajectory. Similarly, the distance between loci within the Ter domain does not change much. In contrast, the inter-domain distance between them increases during this process (**Figure 3e**). This suggests that, when the right arm passes over the Ori and moves far away from the Ter, the loci within the domain moves coordinately as a whole entity and has distinct dynamics with its neighboring domains.

In summary, by correlating the spatial pattern from imaging with pairwise distance matrices, we revealed that the daughter chromosomes have distinct chromosome conformations, which are obscured by previous population-average measurements. We further revealed the cell-cycle-related dynamics of these two states at the domain level, demonstrating the dynamic domain organization in genetically identical cells<sup>16,26</sup>. These findings fill the gap between ensemble genome-wide HiC and fluorescence-imaging research, and provide new insights into the *E. coli* chromosome organization dynamics during cell cycle. In the future, it would be interesting to investigate the effect of nuclear associate proteins on the overall chromosome structure<sup>6</sup>. It would be also interesting to study how the transcription activity affects chromosome conformations by simultaneously detecting RNA and DNA inside the cells<sup>25,27</sup>.

## 2.4 Figures

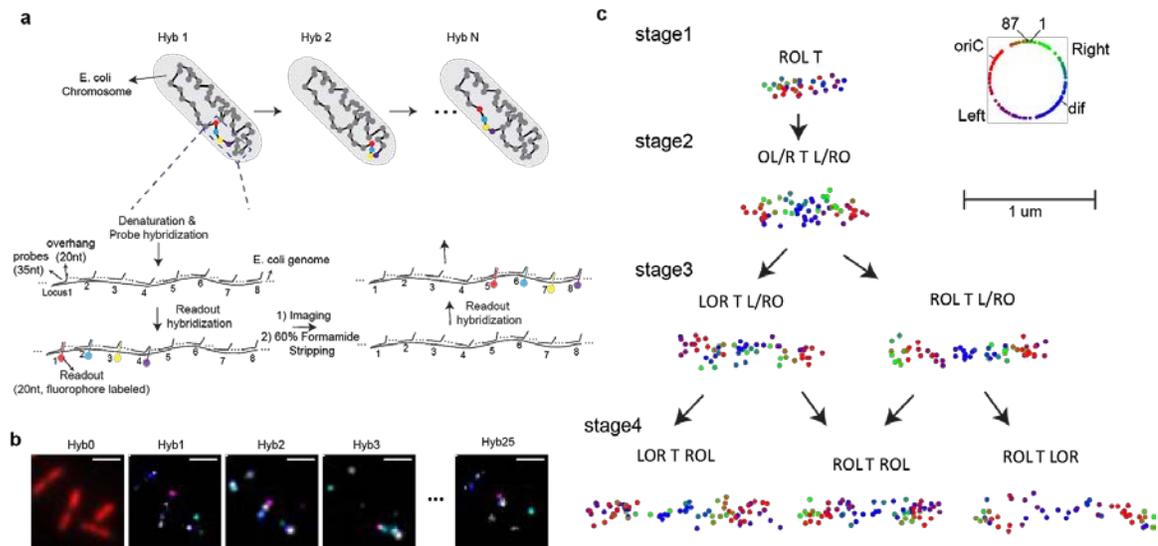


Figure 1. a) Scheme for DNA seqFISH in *E. coli*. In every round of FISH, four chromosomal loci were labeled with four spectrally distinct fluorescent dyes, imaged, and then the signal was stripped. This process was repeated sequentially until the entire chromosome was imaged. The zoom-in shows the hybridization scheme. The *E. coli* chromosome was first denatured and primary probes were hybridized to their targeted chromosomal regions. Fluorescently labeled readouts were hybridized to the overhang part of probes. After imaging, the readouts were stripped off with 60% formamide treatment, generating a clean background ready for the next round of seqFISH. b) Example images captured during seqFISH (Scale bar: 1  $\mu$ m). The DNA-FISH dot was fitted with a 2D Gaussian model and the centroid was used to estimate locus position. c) Examples of single-cell chromosome structures by seqFISH. The loci were labeled in a color palette according to their genomic positions (Ori: red; Right arm: Green; Left arm: Purple; Ter: blue). Example cells from different cell-cycle stages with different conformations are shown.

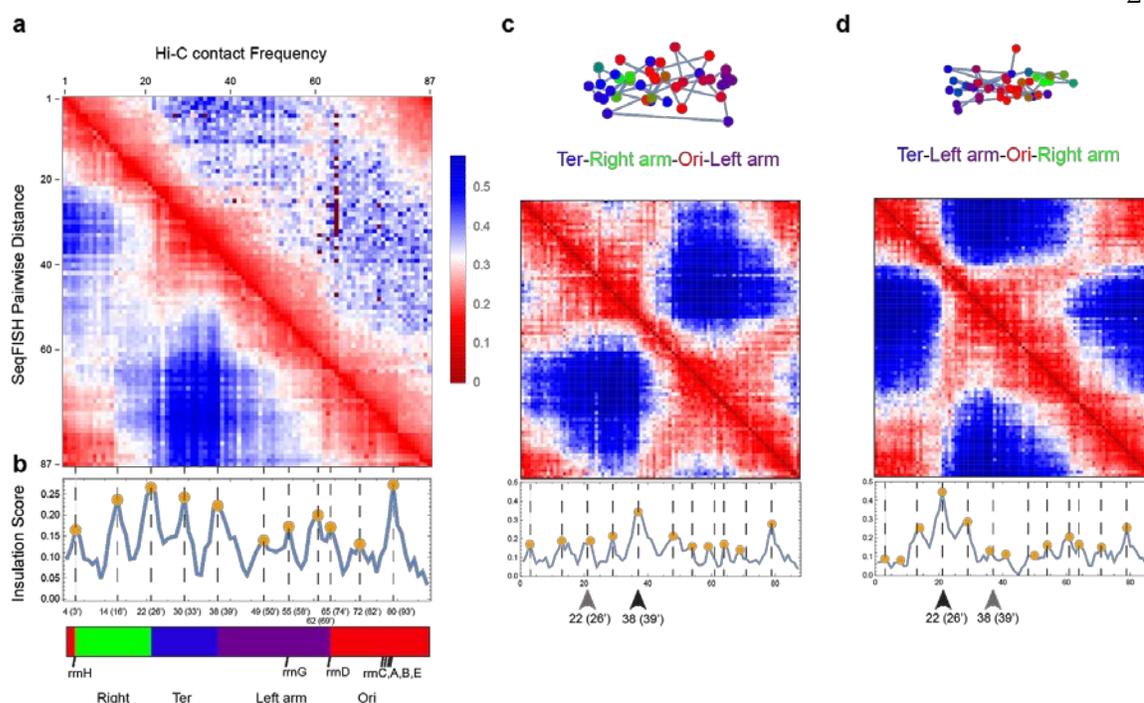


Figure 2. a) Pairwise distance matrix versus contact frequency matrix. Bottom: mean pairwise distance measured from seqFISH. Top: Contact frequency measured from Hi-C at the same probe positions. b) Insulation score of pairwise distance matrix. The domain boundaries are identified as peaks (yellow dots) in insulation score profiles. Bottom: The domain boundaries segment the *E. coli* chromosome into macrodomains. Ribosomal operons are highlighted and colocalized with domain boundaries. c) and d) daughter chromosomes with TROL and TLOR conformations and their pairwise distance matrices and insulation score profiles. The peaks identified from insulation score profiles are labeled with yellow dots. The dashed lines are the domain boundaries identified from the ensemble pairwise distance matrix as in b). The arrows indicate the domain boundaries that define the Ter domain, which show different insulation level in two conformations.

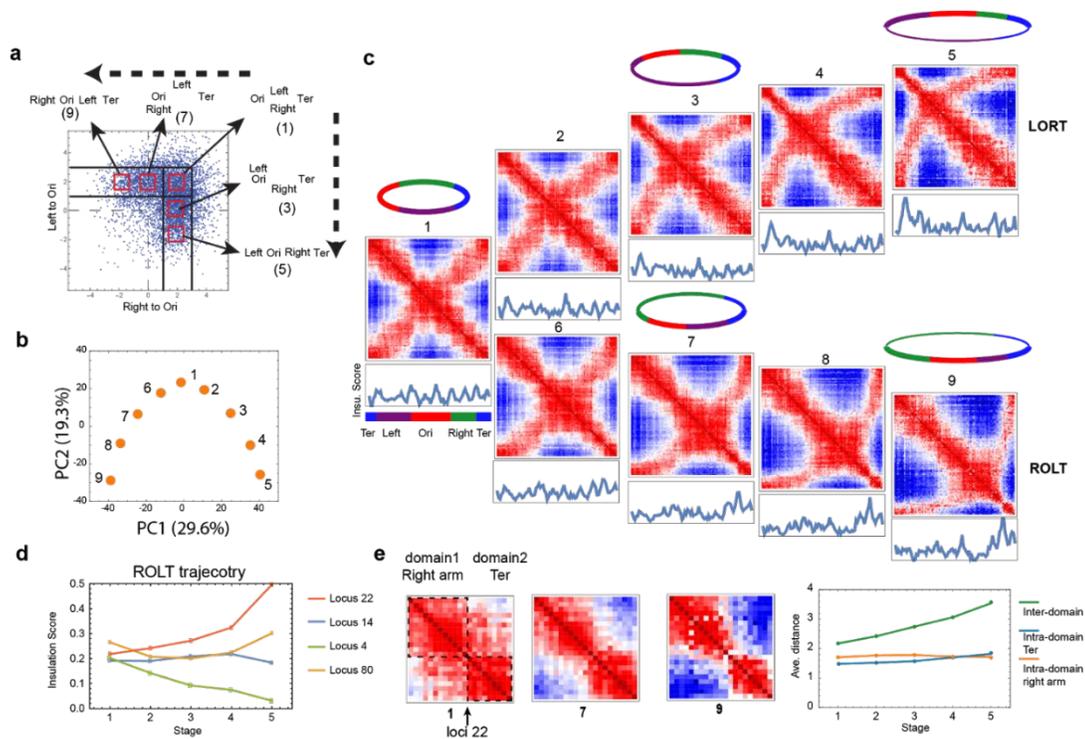


Figure 3. a) The relative positioning of right arm to Ori and left arm to Ori in each single cell. The positive values mean Ori is leading and negative values mean right arm/left arm is leading. Each dot represents a single cell. Different chromosome positioning patterns can be identified as shown. Two major axes (vertical and horizontal), representing two conformation trajectories, were identified, as indicated by the solid lines and dashed arrows. The cells along these two trajectories are divided into several bins (indicated by red boxes), and their pairwise distance matrices are calculated and shown in c). b) The pairwise distance matrices are visualized in their first two PCs. The bifurcation indicates the emergence of two trajectories. c) The pairwise distance matrices and their insulation scores. Cartoons of *E. coli* chromosomes were drawn for a few stages to show the chromosomal movement. The thinner lines illustrate the “stretched regions” during this process. d) The dynamic change of insulation score for some domain boundaries along the ROLT trajectory. e) The pairwise distance matrices of the chromosomal region around domain boundary locus 22 in the ROLT trajectory. The values within the dashed boxes represent intra-domain distance. The inter-domain distance was calculated using the value outside the dashed boxes.

## 2.5 Reference

1. Le, T. B. K. & Laub, M. T. New approaches to understanding the spatial organization of bacterial genomes. *Curr. Opin. Microbiol.* 22, 15–21 (2014).
2. Valens, M., Penaud, S., Rossignol, M., Cornet, F. & Boccard, F. Macrodomain organization of the *Escherichia coli* chromosome. *EMBO J.* 23, 4330–4341 (2004).
3. Duigou, S. & Boccard, F. Long range chromosome organization in *Escherichia coli*: The position of the replication origin defines the non-structured regions and the Right and Left macrodomains. *PLoS Genet.* 13, e1006758 (2017).
4. Espeli, O., Mercier, R. & Boccard, F. DNA dynamics vary according to macrodomain topography in the *E. coli* chromosome. *Mol. Microbiol.* 68, 1418–1427 (2008).
5. Niki, H., Yamaichi, Y. & Hiraga, S. Dynamic organization of chromosomal DNA in *Escherichia coli*. *Genes Dev.* 14, 212–223 (2000).
6. Lioy, V. S. et al. Multiscale Structuring of the *E. coli* Chromosome by Nucleoid-Associated and Condensin Proteins. *Cell* 172, 771–783.e18 (2018).
7. Cagliero, C., Grand, R. S., Jones, M. B., Jin, D. J. & O’Sullivan, J. M. Genome conformation capture reveals that the *Escherichia coli* chromosome is organized by replication and transcription. *Nucleic Acids Res.* 41, 6058–6071 (2013).
8. Le, T. B., Imakaev, M. V., Mirny, L. A. & Laub, M. T. High-resolution mapping of the spatial organization of a bacterial chromosome. *Science* 342 VN -, 731–734 (2013).
9. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485, 376–380 (2012).
10. Sexton, T. et al. Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell* 148, 458–472 (2012).
11. Le, T. B. & Laub, M. T. Transcription rate and transcript length drive formation of chromosomal interaction domain boundaries. *EMBO J.* 35, 1582–1595 (2016).
12. Mercier, R. et al. The MatP/matS Site-Specific System Organizes the Terminus Region of the *E. coli* Chromosome into a Macrodomain. *Cell* 135, 475–485 (2008).
13. Marbouty, M. et al. Condensin- and Replication-Mediated Bacterial Chromosome Folding and Origin Condensation Revealed by Hi-C and Super-resolution Imaging. *Mol. Cell* 59, 588–602 (2015).
14. Nagano, T. et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* 502, 59–64 (2013).
15. Stevens, T. J. et al. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* 544, 59–64 (2017).

16. Bintu, B. et al. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* 362, eaau1783 (2018).
17. Fisher, J. K. et al. Four-dimensional imaging of *E. coli* nucleoid organization and dynamics in living cells. *Cell* 153, 882–895 (2013).
18. Wang, X., Possoz, C. & Sherratt, D. J. Dancing around the divisome: asymmetric chromosome segregation in *Escherichia coli*. *Genes Dev.* 19, 2367–2377 (2005).
19. Michelsen, O., Teixeira de Mattos, M. J., Jensen, P. R. & Hansen, F. G. Precise determinations of C and D periods by flow cytometry in *Escherichia coli* K-12 and B/r. *Microbiology* 149, 1001–1010 (2003).
20. Nielsen, H. J., Li, Y., Youngren, B., Hansen, F. G. & Austin, S. Progressive segregation of the *Escherichia coli* chromosome. *Mol. Microbiol.* 61, 383–393 (2006).
21. Liu, X., Wang, X., Reyes-Lamothe, R. & Sherratt, D. Replication-directed sister chromosome alignment in *Escherichia coli*. *Mol. Microbiol.* 75, 1090–1097 (2010).
22. Wang, X., Liu, X., Possoz, C. & Sherratt, D. J. The two *Escherichia coli* chromosome arms locate to separate cell halves. *Genes Dev.* 20, 1727–1731 (2006).
23. Woldringh, C. L., Hansen, F. G., Vischer, N. O. E. & Atlung, T. Segregation of chromosome arms in growing and non-growing *Escherichia coli* cells. *Front. Microbiol.* 6, (2015).
24. Wang, S. et al. Spatial organization of chromatin domains and compartments in single chromosomes. *Science* 353, 598–602 (2016).
25. Cardozo Gizzi, A. M. et al. Microscopy-Based Chromosome Conformation Capture Enables Simultaneous Visualization of Genome Organization and Transcription in Intact Organisms. *Mol. Cell* (2019). doi:10.1016/j.molcel.2019.01.011
26. Fudenberg, G. et al. Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep.* 15, 2038–2049 (2016).
27. Mateo, L. J. et al. Visualizing DNA folding and RNA in embryos at single-cell resolution. *Nature* 568, 49–54 (2019).

## 2.6 Supplementary information

### 2.6.1 Experiments

#### Cell culture

*E. coli* MG1655 is cultured at 37°C in minimal M9 media supplemented with 0.2% glucose. At this slow-growing condition (doubling time ~80mins), the cell will finish division before starting another round of replication<sup>1</sup>. Therefore, the cell will contain 1–2 copies of the chromosomes.

#### DNA seqFISH probe design

Each primary probe contained three parts: a 35nt target sequence complementary to the genomic region of interest; a 4-nt spacer; and a 20nt readout sequence. The 35nt target sequence was generated according to the following criteria: the GC content was between 40% and 70%, the probe had no more than 18nt homology to any other part of the genome, and the probe was complementary to the coding strand of the genes so that it would not hybridize with RNA inside the cells. For most loci we targeted, around 100 probes were generated. For a few loci, we generated 48 probes and 24 probes for the purpose of comparison. A 20nt readout sequence was attached to the 35nt sequence with a 4nt spacer. The readout sequence was the same within all the probes of a specific locus, and distinct between different loci. The above designed oligo pool was synthesized using the enzymatic application protocol<sup>2</sup>.

The readout probes, which were complementary to the readout sequence, were conjugated with fluorescent dyes (Alexa 594, Alexa 647 (Life Technologies), cy3b, cy7 (GE Health)) using methods described previously<sup>3</sup>.

#### Sample preparation and primary probe hybridization

After the cell culture reaches O.D. 0.2, formaldehyde (Thermo Scientific 28908) was

added directly into cell culture to final concentration of 1%. The cells were fixed at their normal growing condition (37°C, constant agitation) for 20mins and final concentration of 0.25M glycine was added to quench the formaldehyde for 5mins. Fixed cells were collected by centrifuge, washed with 2xSSC (Invitrogen 15557-044 diluted in Ultrapure water (Invitrogen 10977-015)), and treated with lysozyme (1ug/ml in GTE buffer (25mM Tris pH8.0, 50mM Glucose, 10mM EDTA)) for 10mins. The cells were washed with 2xSSC after lysozyme treatment. Then the cells were dispersed on to home-made amine-modified coverslips by centrifuge. We found that the cells adhered to the amine-modified coverslip very well and did not move during our experiment. A custom-made flowcell was then attached to the coverslip. To denature the E. coli chromosome, denaturing buffer (70% formamide (v/v) (Invitrogen AM9344), 2xSSC, 10% Dextran Sulfate (Sigma D8906) in Ultrapure water) was added into the flow cell. The flowcell was sealed and heated to 85°C for 10mins on a heat block. The primary probe pool was dissolved in primary hybridization buffer (40% formamide (v/v), 2xSSC, 1mg/ml BSA, 10% Dextran Sulfate) to a final concentration of 10uM and added into the flowcell. The flowcell was sealed and incubated at 37°C for at least 24hrs in order for the primary probes to hybridize.

#### Sequential rounds of hybridization

After primary probe hybridization, the flowcell was washed at room temperature with washing buffer (30% formamide (V/V), 2xSSC, 0.1% Triton (Sigma 93443)) for 2hrs and washed with 2xSSC several times. The flowcell was then mounted onto the microscope and connected to custom-built fluidics. The cells were first stained with DAPI (Sigma D8417) (1uM in 2xSSC) for 5mins and imaged to select regions of interest. Then blue fluorescent (365/415) beads (0.1um, Thermo Scientific F8805) were flowed in. The beads attach to the surface of the coverslips and serve as the fiducial markers during sequential rounds of hybridization experiments.

For each round of hybridization, the following protocol was used: Hybridization buffer (10% formamide (v/v), 2xSSC, 1mg/ml BSA, 10% Dextran Sulfate) with four different

colors of adapters complementary to the four different readout sequences, at 100nM each, were flowed into the flowcell and incubated for 30mins at room temperature. Then the flowcell was washed three times with washing buffer. Afterwards, anti-bleaching buffer (50mM Tris pH 8.0, 300mM NaCl, 2xSSC, 3mM Trolox (Sigma 238812), 0.8% D-glucose (Sigma G7528), 100-fold diluted Catalase (Sigma C3155), 0.5mg/ml glucose oxidase (Sigma G2133)) was flowed into the flowcell and samples were imaged. After imaging, stripping buffer (60% formamide (v/v), 2xSSC) was flowed into the flowcell, incubated for 5mins, and then washed with 2xSSC three times. After stripping, the samples were imaged again to check the completeness of the signal quenching and establish the background for the next round of hybridization. For imaging, a Nikon Ti Eclipse with PFS autofocus microscope was used. The microscope, the motorized stage (ASI MS2000), and the fluidic system were controlled through a custom script written in Micromanager software. In the experiments, snapshots of multiple regions of interest were taken with a fixed z position autofocus lock. The experiments were done using a home-made automation system which integrates fluidics handling and microscope imaging acquisition via Micromanager.

#### Localization precision measurement

In a separate set of experiments, we performed sequential FISH in which the same locus was targeted with four different colors in four sequential rounds of hybridization. The FISH signal was then processed following the same procedure (see Data Analysis, below). The above experiment is similar to the hybridization, washing, and imaging procedure in the real experiments, and accounts for all the errors introduced into our experimental procedure. The localization accuracy of our system was then estimated by calculating the standard deviation of localization distribution. The standard deviation is estimated to be 50nm (**Supplementary Figure 3**).

### 2.6.2 Data Analysis

#### Image registration

The fiducial markers, blue fluorescent (365/415) beads, were imaged together with FISH images in each round of hybridization. The bead images were fit to 2D Gaussian functions to determine their center positions in x and y. The drifts between different rounds of hybridization were determined from bead positions and used to correct FISH dot positions.

#### Determination of loci's spatial positions

Only the central quarter of the image was used for processing, in order to reduce uneven illumination and chromatic aberration. FISH signals were thresholded to reject weak, non-specific signals and, in some cases, low leftover signals from previous hybridizations. The FISH dots were fit to 2D Gaussian functions to determine their center positions in x and y. Dots with low fitting score were discarded. The positions were further corrected, respectively, with the drifts measured before.

#### Cell segmentation

The DAPI image was taken before sequential rounds of hybridization and used for cell segmentation. The image was converted into morphological components using a predetermined threshold in Mathematica 10.0. Each morphological component corresponds to one cell. The segmentation results were further manually curated and corrected.

#### Single-cell chromosome structure reconstruction

The pixelated cell positions obtained from cell segmentation were uniformly expanded by two pixels. The loci in the expanded cell were attributed to this cell. In further analysis, the cells were aligned with their longitudinal axis, and the centers of the cells were aligned to the coordinate origins.

We calculated the number of dots and cell size in a single cell (**Supplementary Figure 4**). The cell size was measured as the area within the bounding box of the reconstructed

single-cell chromosome images. We observed that, as cell size increased, the number of detected dots also increased, in agreement with the assumption that the chromosomes were being progressively replicated and segregated during the cell cycle<sup>4</sup>, and confirming that cell size could be used as a rough cell-cycle marker. Example cells with different conformations were manually selected and presented in **Figure 1c** and **Supplementary Figure 5a**.

In further analysis, cells with a size bigger than 30a.u. were selected and used. In the majority of these cells, two Ori regions could be observed that were well separated into two half cells, and the Ter region was localized in the center of the cells (data not shown). This allowed us to isolate single daughter chromosomes by dividing the cells into two halves at the center. Using this simple method, we were able to capture single daughter chromosome structures with decent accuracy and coverage, even though we possibly losing some loci, especially near the Ter region.

Different daughter chromosome conformations were determined based on the relative positioning of Ori region, right arm region, and left arm region. The Ori region was defined as a chromosomal region centered by replication origin (OriC). The Ter region was a chromosomal region centered by dif site (the end of replication). The right-arm and left-arm regions were chromosomal regions which were centered by the midpoint between OriC and dif on the two arms, respectively (**Supplementary Figure 5b**). Specifically, the mean position of dots from the chromosomal region between locus 68 and locus 80 was calculated and used to represent the Ori region. Similarly, the mean position of dots from the chromosomal regions between locus 4 and locus 15 was calculated and used to represent the right-arm region. The mean position of dots from chromosomal regions between locus 50 and locus 62 was calculated and used to represent the left-arm region. Using these mean positions, the cells with different conformations shown in **Figure 1c** were identified, and their size distribution was plotted in **Supplementary Figure 5c**.

Intra-chromosomal pairwise distance matrix

As stated before, cells with a size bigger than 30a.u. were selected, and single daughter chromosomes were isolated. Within a single sister chromosome, we calculated the pairwise Euclidean distance and the mean distance across all chromosomes in order to generate the ensemble intra-chromosome pairwise distance matrix.

#### Pairwise distance versus Hi-C contact frequency

The E. coli Hi-C data was obtained from Lioy et al.'s work<sup>5</sup>. For each pair of loci in our probe sets, their contact frequency was determined by averaging the contact frequency of the genomic regions covered by the loci's probes. The Hi-C contact frequency and mean spatial distance were log transformed and fitted by a linear function (**Supplementary Figure 6**). The high correlation indicates that the Hi-C and FISH are complementary techniques to measure the chromosome interaction and provide cross validation to each other, which is in alignment with other research<sup>6</sup>.

#### Insulation Score and domain boundary identification

The insulation score for a specific locus was defined as  $-\log((a1/b+a2/b)/2)$ : where  $a1$  is the average distance between its upstream five loci;  $a2$  is the average distance between its downstream five loci; and  $b$  is the average distance between upstream and downstream (**Supplementary Figure 7**). This definition of insulation score follows previous Hi-C work<sup>7</sup> and measures the level this locus insulates its upstream and downstream regions. The insulation scores for all the loci along the chromosome were calculated and the peaks in the insulation score profiles were identified as the domain boundaries. A standard zero- derivative method with Gaussian blurring up to scale 1 was used for peak detection. A five-loci window was chosen, as it generates domain boundaries that agree most closely with a visual inspection.

#### Pseudo-time trajectory reconstruction

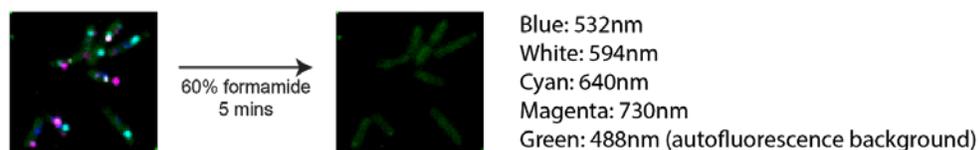
To infer fine-grained pseudo-time trajectories, we parsed the chromosomes according to their different stages and different spatial conformations. Daughter chromosomes

were aligned so that the Ter was at the rightmost position, then the mean positions representing Ori region, right-arm region, and left-arm region in each single daughter chromosome were calculated using the methods described previously. The vector from left arm to Ori and from right arm to Ori along the long axis in each single cell was then calculated and plotted in **Figure 3a**. In the plot, most cells have both positive left-to-Ori and positive right-to-Ori (the first quadrant), which correspond to the Ori-leading conformation. Some cells have positive left-to-Ori and negative right-to-Ori (the second quadrant), which correspond to the right-arm-leading conformations. Some cells have positive right-to-Ori and negative left-to-Ori (the fourth quadrant), which correspond to the left-arm-leading conformations. The cells in the third quadrant account for a very small fraction of the whole population, and are probably from low quality cells with wrong positioning patterns. These observations lead to the identification of two axial regions (**Figure 3a**). In the vertical axis, the left-arm-to-Ori vector is positive and the right arm-to-Ori vector flipped from positive to negative values, which corresponds to the process in which the right arm passes over the Ori region. In the other horizontal axis, the right-arm-to-Ori direction does not change, but the left-arm-to-Ori vector flipped from positive to negative values, which corresponds to the process in which the left arm passes over the Ori regions. We took the cells along these two axes and divided them into overlapping bins based on their left-to-Ori and right- to-Ori vectors: Vertical axis: right-to-Ori: [3,1], left-to-Ori: [3,1], [2, 0], [1, -1], [0, -2], [-1, -3]; Horizontal axis: left-to-Ori: [3,1], right-to-Ori: [3, 1], [2, 0], [1, -1], [0, -2], [-1, -3]. The mean pairwise distance matrices were calculated from each bin.

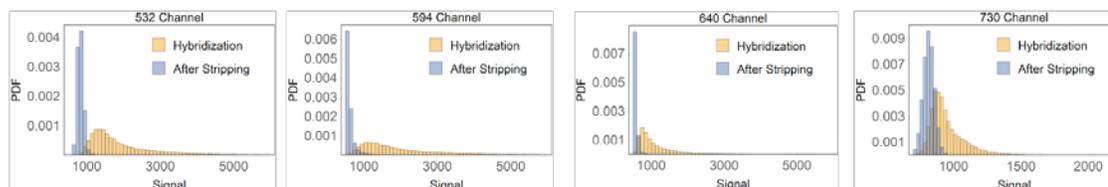
To infer trajectory, we analyzed the pairwise distance matrices using PCA (**Supplementary Figure 8**) and plotted them using the first two PCs. We observed a trajectory structure with two branches. The branches match along with the two axes we defined previously and correspond to the formation of ROLT and LORT conformations.

### 2.6.3 Supplementary Figures

a

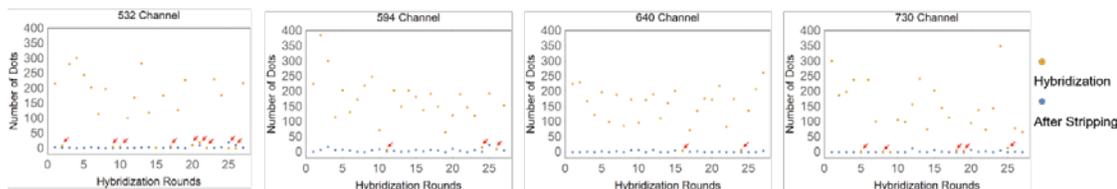


b

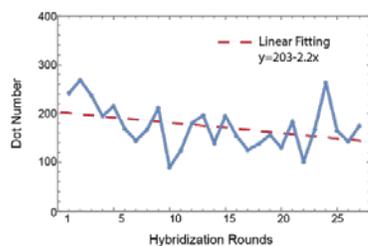


**Supplementary Figure 1.** a) Example images before and after signal stripping. b) Image intensity profiles before and after stripping for four different channels. The signals decrease substantially after stripping.

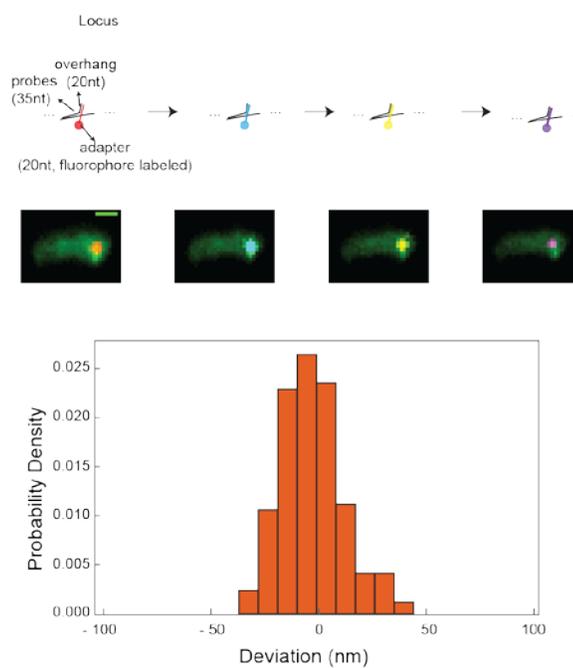
a



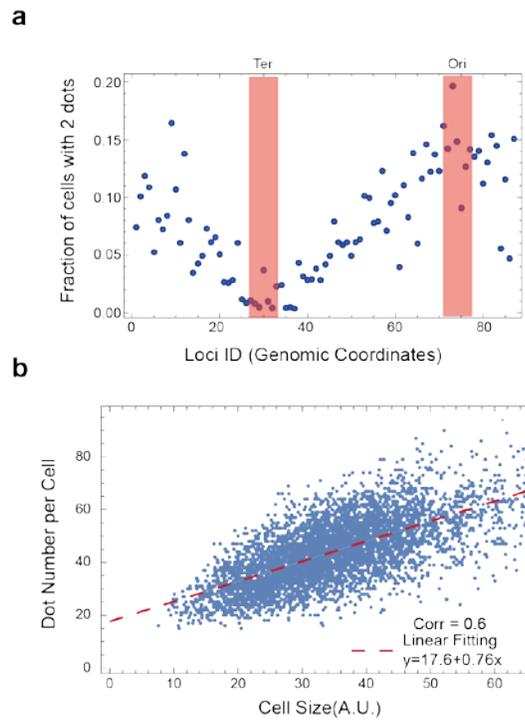
b



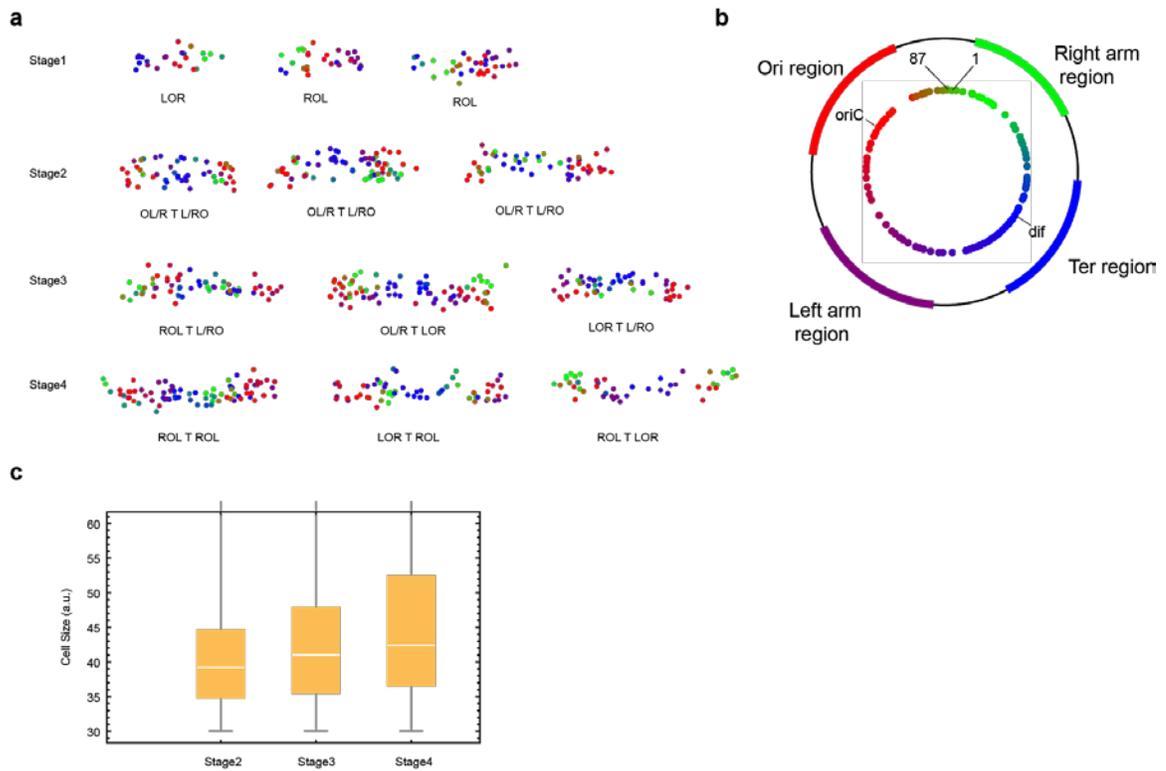
**Supplementary Figure 2.** a) Number of dots detected over 27-round hybridization for four different channels. Yellow dots represent the number of dots in hybridization images and blue ones represent the number of dots in after-stripping images. The red arrows indicate the hybridizations that have low efficiency and are discarded in further analysis. b) The number of dots for all four channels across 27-round hybridization experiment. From the linear fitting, we estimated a 20% loss of signal during the seqFISH experiments.



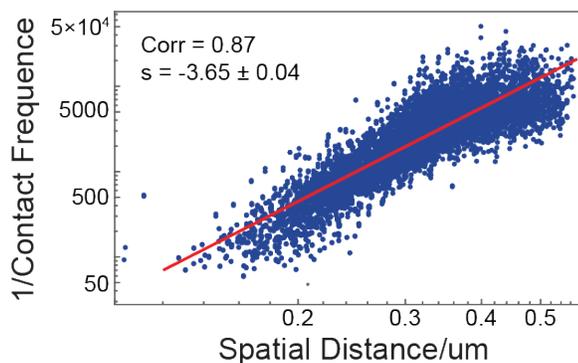
**Supplementary Figure 3.** The localization accuracy of DNA seqFISH. The same locus was imaged in four sequential rounds of hybridization using four different fluorescent channels. The localization error was determined to be 50nm (Full width at half maximum (FWHM) of the localization error distribution).



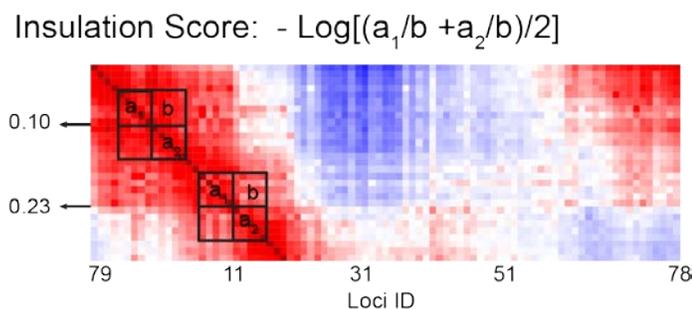
**Supplementary Figure 4.** a) The fraction of cells that have two dots (indicating the locus has been replicated in this cell) for every locus. The Ori has the highest fraction and the Ter has the lowest fraction, which agrees with the fact that the Ori is the first to be replicated and the Ter is the last to be replicated. b) Number of dots versus cell size. Each dot represents single cell. From the linear fitting (dashed red line), we estimated the detection efficiency is about 20-40%.



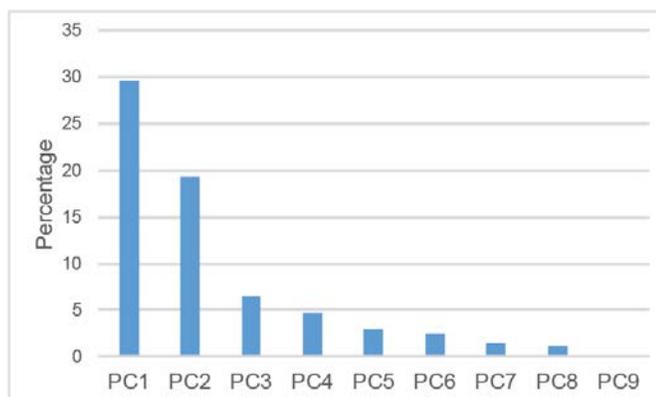
**Supplementary Figure 5.** a) Example cells with different chromosomal conformations. b) The chromosomal regions that represent Ori, right arm, left arm, and Ter. The mean positions of these loci are calculated in single daughter chromosomes and used to represent the chromosomal regions. c) The cell size distribution for cells at stage 2, 3, and 4.



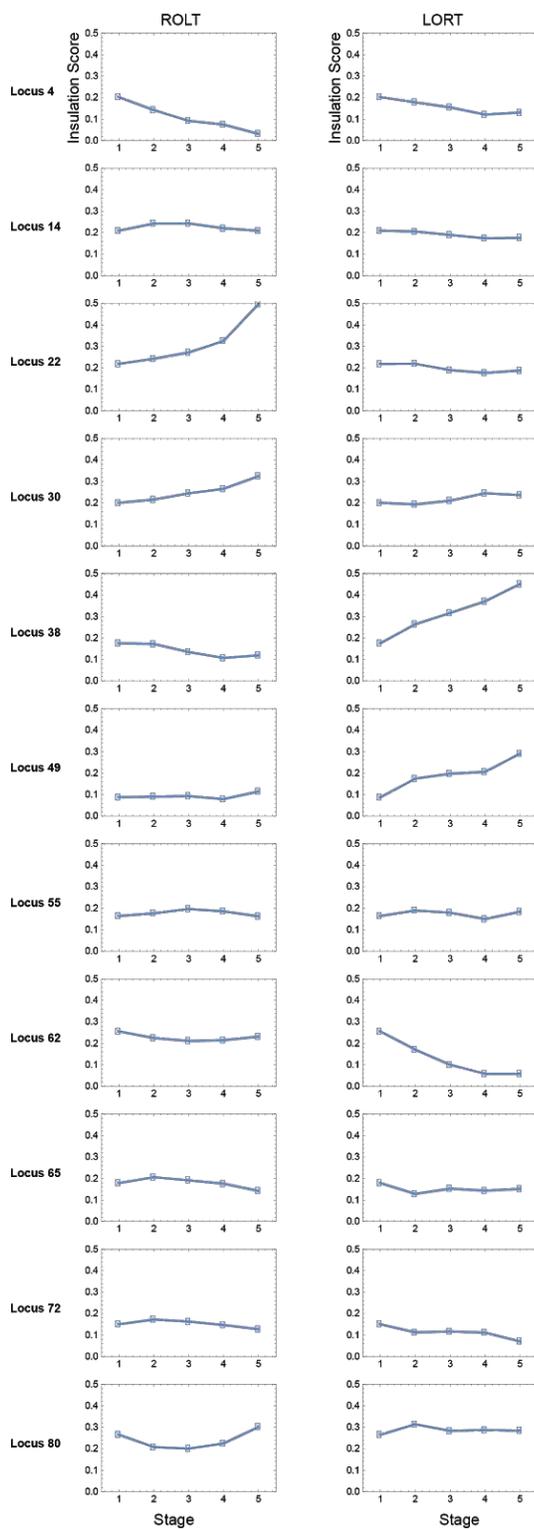
**Supplementary Figure 6.** The correlation between ensemble-averaged spatial distance obtained by seqFISH and contact frequency obtained from HiC after normalization.



**Supplementary Figure 7.** Definition of insulation score. For each locus, the average distance from its upstream five-loci region ( $a_1$ ), the average distance from its downstream five-loci region ( $a_2$ ), and the average distance between its upstream and downstream regions ( $b$ ) are used for insulation score calculation.



**Supplementary Figure 8.** The percentage of variance explained by each principal component.



**Supplementary Figure 9.** The insulation score of domain boundaries changes along the ROLT and LORT trajectories. Different domain boundaries have different dynamics.

#### 2.6.4 Supplementary Reference

1. Reeve, B., Hargest, T., Gilbert, C. & Ellis, T. Predicting translation initiation rates for designing synthetic biology. *Front Bioeng Biotechnol* 2, 1 (2014).
2. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* 348, aaa6090 (2015).
3. Lubeck, E. & Cai, L. Single-cell systems biology by super-resolution imaging and combinatorial labeling. *Nat. Methods* 9, 743–748 (2012).
4. Nielsen, H. J., Li, Y., Youngren, B., Hansen, F. G. & Austin, S. Progressive segregation of the *Escherichia coli* chromosome. *Mol. Microbiol.* 61, 383–393 (2006).
5. Lioy, V. S. et al. Multiscale Structuring of the *E. coli* Chromosome by Nucleoid-Associated and Condensin Proteins. *Cell* 172, 771–783.e18 (2018).
6. Wang, S. et al. Spatial organization of chromatin domains and compartments in single chromosomes. *Science* 353, 598–602 (2016).
7. Lajoie, B. R., Dekker, J. & Kaplan, N. The Hitchhiker’s guide to Hi-C analysis: practical guidelines. *Methods* 72, 65–75 (2015).

## QUANTIFYING TARGETED PROTEIN AND ITS POST-TRANSLATIONAL MODIFICATION ISOFORMS IN SINGLE CELLS BY DNA BARCODING AND NEXT-GENERATION SEQUENCING

### 3.1 Abstract

Currently, using specific antibodies to distinguish proteins and their PTM isoforms is the only way to quantify a protein and its PTMs in single cells. Here, we report a novel, antibody-free strategy to quantify targeted protein and its PTM isoform in single cells. In this method, we developed an efficient in situ barcoding strategy to barcode proteins in single cells, through a spycatcher/spytag system and combinatorial indexing. Thereafter, the tagged protein and its PTM isoform are separated by conventional gel electrophoresis, while their single-cell identity is preserved in the covalently attached oligo. By counting the attached DNA oligos using next-generation sequencing, single-cell proteins and protein isoforms can be accurately measured. We demonstrated the utility of the technology by quantification of histone protein H2B and its monoubiquitination isoform, H2Bub, at the single-cell level. Our method revealed the single-cell heterogeneities of the H2Bub/H2B ratio and its cell-cycle dynamics, which are obscured by previous methods of ensemble measurement.

### 3.2 Introduction

Protein PTMs modify proteins and regulate their function after the protein has been translated. Protein expression level and its PTM states show substantial single-cell heterogeneity. For example, histone PTMs have been associated with a variety of processes inside the cells<sup>1</sup>. Many transcription factors have also been shown to undergo pulsatile phosphorylation/dephosphorylation cycle to regulate multiple gene expressions in a coordinated fashion<sup>2-4</sup>. Therefore, accurate quantification of the protein

and its PTM states in single cells is greatly needed to understand such complex behaviors of cells. However, conventional biochemical methods that can differentiate and quantify different protein isoforms, such as Western blot, are generally challenging to downscale to the single-cell level<sup>5</sup>. Alternatively, tagging a fluorescent protein to a target protein has been widely used to quantify the protein expression level in single cells. But this method is usually unable to track the PTM state of the targeted protein<sup>6</sup>. The antibody-based immunocytochemistry method can quantify different proteins and different PTMs in single cells. Fluorescence<sup>7</sup>, mass spectrometry<sup>8</sup>, and next-generation sequencing<sup>9,10</sup> are used as readout signals towards high multiplicity proteome detection in single cells. Yet existing methods depend on high-quality antibodies with high specificity and high binding affinities, which are not always available<sup>11</sup>. In addition, each antibody has to be tested individually and validated for each experimental method in order to ensure successful identification of targets<sup>12,13</sup>. More importantly, there are restrictions on the ability to directly compare the readouts from different antibodies, due to various binding affinities and target specificities. All these issues are even more serious when targeting a protein of interest and its PTM isoforms, due to their chemical similarities.

Therefore, to accurately quantify targeted protein and its PTM isoforms in complex cell populations, we reported a novel method based on covalent DNA barcoding and next-generation sequencing (**Figure 1a**). We developed a chemical biology approach to covalently attach a DNA oligo to the targeted protein with high efficiency and specificity. We then implemented a combinatorial barcoding method to incorporate single-cell barcodes and unique molecular identifiers (UMIs) into oligos, which can then be read out in next-generation sequencing. This combinatorial barcoding scheme does not require special instruments (the whole process involves pipetting steps), and bypasses the need to manipulate single cells. Finally, instead of depending on antibodies, we separated targeted proteins and their PTM isoforms (H2B and H2Bub, demonstrated in this work) by conventional gel electrophoresis. Protein-oligo conjugate

was extracted from the gel, and the oligo part was PCR amplified to generate sequencing library. Since the protein oligo is covalently ligated, the single-cell identity of the proteins will be preserved in the oligos during the process and can be quantified by a next-generation sequencer. As a proof-of-concept example, we used our method to quantify histone protein H2B and its monoubiquitination isoform H2Bub in single yeast cells. Our result revealed the cell-cycle dynamics of the H2Bub/H2B ratio in single cells.

### 3.3 Results and Discussion

First we aimed to develop an approach to in situ label DNA oligo to targeted proteins inside the cells. Among various approaches we tested, we found that the spyttag/spycatcher reaction system is highly efficient for in situ labeling of a DNA oligo to a target protein. Spyttag is a 13-amino-acid peptide that forms an isopeptide with spycatcher, a 20kd protein. This reaction does not require any cofactors and can happen rapidly at room temperature with very high specificity<sup>14</sup>. We constructed *S. cerevisiae* yeast strains containing spyttag at the C-terminal of proteins of interest. The 3xFLAG tag was included with spyttag for antibody detection in order to check ligation efficiency by Western blot. We then synthesized the spycatcher-DNA oligo conjugate (**Supplementary Figure 1**) in vitro. Briefly, a cysteine is introduced at the C-terminal of spycatcher. After purifying the protein in the reductive condition, we labeled the thiol group in the protein with maleimide-PEG4-tetrazine to obtain spycatcher-tetrazine. Then 5' amine-modified oligo(20nt) was reacted with NHS ester-TCO to generate TCO-oligo. Finally, spycatcher-tetrazine and TCO- oligo react with each other via the orthogonal click chemistry to form a spycatcher-oligo conjugate.

We tested the efficiency of the spycatcher/spyttag system for in situ DNA tagging. Histone protein H2B was targeted as an example. Yeast cells with H2B bearing a spyttag were fixed by formaldehyde and permeabilized by spheroplasting. Additional fixation was done after spheroplasting to further keep the proteins from diffusing away from the fixed-cell matrix. The resulting cell pellet was reacted with spycatcher oligo. **Figure 1c**

shows the Western blot result of the whole-cell lysate, before and after reaction using the anti-FLAG antibody. The band corresponding to the target protein shifts up after reacting with spycatcher oligo, indicating successful conjugation of oligos to the target protein. We estimated the reaction efficiency to be over 90%, as seen from the intensity of Western blot. We tested several other target proteins with different copy numbers, with different cellular localizations (High copy: H2B (nucleus) and PRE1(cytoplasm); low copy: SNF1 and GLC7). The in situ tagging efficiency is high (above 90%) for the targeted proteins (**Supplementary Figure 2**), showing the general applicability of the spytag/spycatcher system method.

We next devised a combinatorial cellular barcoding scheme via sequential rounds of “pool- split” T7 oligo ligation to uniquely barcode a target protein in single cells (**Figure 1a**). Essentially, the cells, after in situ spytag/spycatcher DNA tagging, are distributed into a 96-well plate. Each well contains oligos with a well-specific cell barcode sequence, as well as a common adapter sequence for T7 ligation (**Figure 1b**). The barcode oligo is ligated to the protein by T7 ligase with presence of the adapter oligo. After ligation, the cells from different wells are pooled together and redistributed into a new 96-well plate. The second cell barcode is T7 ligated to the protein-oligo conjugate again. The combinatorial barcoding scheme avoids the synthesis of hundreds of different spycatcher-oligo conjugates, and the whole process involves only pipetting, avoiding physical manipulation and isolation of single cells. After these two rounds of barcoding, the cells are pooled again and roughly 900 cells are aliquoted for uniquely sampling single cells, which effectively avoids barcode collisions<sup>15</sup> (**Supplementary Figure 6**). In addition, a 12nt random-base UMI sequence is included in the first round of DNA barcode oligo (**Figure 1b**), which will encode each individual protein from single cells with a UMI. The use of UMIs effectively corrects the biases generated in next-generation sequencing<sup>16</sup>, as the protein copy numbers in single cells can be then determined through counting the number of UMIs. A 12nt random-based sequence is used in this case, with a barcoding capacity much higher than the protein copy number, so that each individual protein from single cells are most likely to be labeled uniquely,

as shown in later sequencing result analysis (**Supplementary Figure 8**).

We tested the T7 ligation efficiency of barcode oligos. Yeast cells after spycatcher-oligo conjugation were subject to two rounds of pool-split T7 ligation reactions and analyzed by Western blot. **Figure 1c** shows that bands corresponding to the target protein shift up after two rounds of ligation, indicating that the cell barcode is successfully ligated. Under optimized conditions, the T7 ligation efficiency is estimated to be more than 90%. We also tested the T7 ligation for a variety of other proteins (**Supplementary Figure 2**). In addition, the cell morphology was checked under microscopy and was well preserved after the spycatcher-oligo reaction and two rounds of ligation, validating that the cells can be used as compartments during the pool-split ligation process (**Supplementary Figure 3**).

Furthermore, we designed and synthesized “dummy” cells, in which the targeted proteins were labeled with the same-length oligo as barcoded cells so that they would co-migrate with barcoded cells, but would not be amplified during the PCR (**Supplementary Figure 4**). As such, the small number of barcoded cells (900 uniquely barcoded cells after second-round barcoding) piggy-backed on the large number of dummy cells (~ 1 million cells) to avoid severe sample loss when handling a small number of samples. The dummy oligo also had a TAMRA dye at 3' end for band visualization on gel (**Supplementary Figure 4**). We separated histone protein H2B and its monoubiquitination isoform H2Bub by SDS-PAGE after barcoding oligo ligation. The mono-ubiquitinated H2B is 7kD heavier and shows as an upper band in Western blot (**Figure 1c**). Bands corresponding to different protein isoforms were then cut from the gel, and the protein-oligo conjugate was extracted from the gel piece (**Supplementary Figure 4**). Dissolvable polyacrylamide gel was used to achieve high-efficiency gel recovery<sup>17</sup> (**Supplementary Figure 5**). The oligo from the extracted protein-oligo complex was made into sequencing libraries by PCR amplification.

We applied a bioinformatic pipeline to identify cell barcode and count UMIs. The total number of reads per cell barcode was plotted in descending order (**Figure 2a**). A group

of barcodes with high number of reads are clearly separated from the rest, which consist of a large number of sparsely populated cell barcodes. The cells barcodes have  $10^4$  reads on average, which in total account for 97% of reads. 850 cell barcodes were identified, agreeing with our experimental design (~900 cells are aliquoted). The sequencing result from H2B-ub is similar to that from H2B (**Supplementary Figure 7**). The cell barcodes identified from the H2Bub sample are almost the same as those from H2B (848 out of 850 barcodes). The clear cutoff for the cell barcode disappears when we analyze the library obtained from background on the gel (**Supplementary Figure 7**). Taken together, we concluded that those cell barcodes represent real single cells and the remaining sparsely populated barcodes are spurious results caused by PCR and sequencing errors.

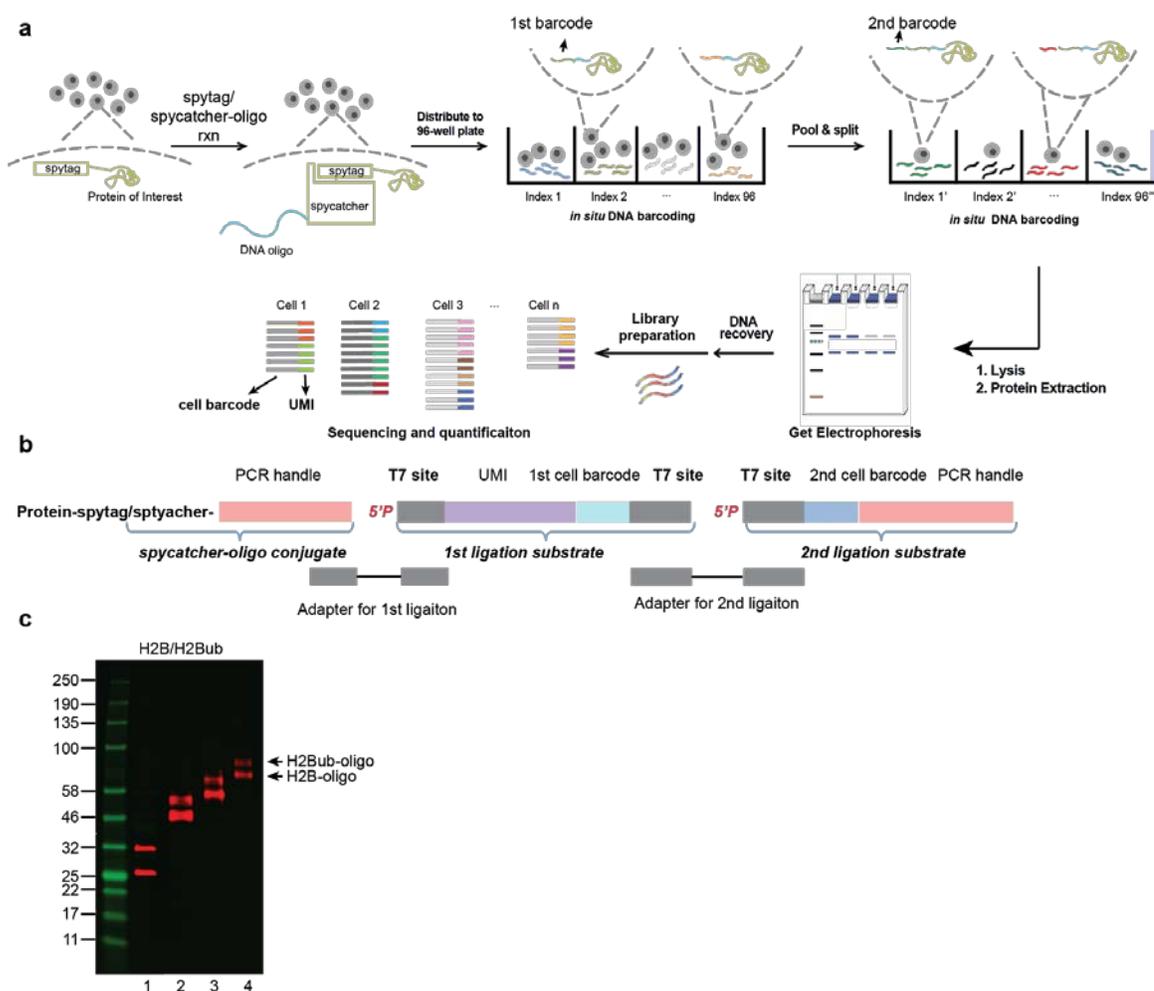
We confirmed that our UMI length could encode all the proteins in single cells. We also confirmed that all the possible UMIs were sufficiently sampled with current sequencing depth (**Supplementary Figure 8**). Based on the above results, the copy number of proteins in single cells could be quantified by directly counting UMIs associated with each cell's barcodes. It should be noted that since only a small fraction of the library is used in sequencing, the quantification will not count the absolute copy number of proteins in single cells, but instead their relative distribution.

The H2B copy number shows a well-known characteristic bimodal distribution according to different cell-cycle stages, which provides a nice standard to compare our method with. The histogram of the H2B copy number determined from our method shows a bimodal distribution, and the two peaks are roughly two-fold difference, demonstrating that we could accurately quantify the relative level of H2B in single cells. H2Bub also shows a bimodal distribution, but not as clearly as H2B (**Figure 2b**). The ratio between H2Bub/H2B was calculated in every single cell and the ratio shows a unimodal distribution varying from 0.08 to 0.18 with the mean ratio about 0.12, which agrees with both previous results<sup>18</sup> and the results from ensemble Western blot (**Figure 2c, Supplementary Figure 9**).

The H2Bub/H2B ratio was plotted versus the H2B copy number for each single cell in **Figure 2d**. Interestingly, cells at different cell cycle stages (as indicated by the H2B copy number) have different H2Bub/H2B ratio distribution. Specifically, at G1 stage, a subpopulation of cells is hyper-ubiquitinated (H2Bub/H2B ratio between 0.15–0.18), which is absent from the G2/M stages. Such heterogeneity is obscured from previous ensemble analysis (**Figure 2e**). To further investigate this, we applied our method on a yeast strain in which two de-ubiquitination enzymes (UBP8 and UBP10)<sup>19</sup> removing ubiquitin from H2B are knocked out. As expected, a large increase in the H2Bub level was observed in this strain (the H2Bub/H2B ratio is about 0.6, which is estimated from ensemble Western blot (**Supplementary Figure 9**) and also from the mean ratio of our single-cell data). However, we still observed that a group of cells are hyper-ubiquitinated (0.8–1.1) in G1 stage, compared to the cells at G2/M phase (H2Bub/H2B ratio: 0.5–0.8) (**Figure 2f**). Such heterogeneity suggest that the H2Bub level is dynamically regulated during the cell cycle, whose mechanism needs be further explored.

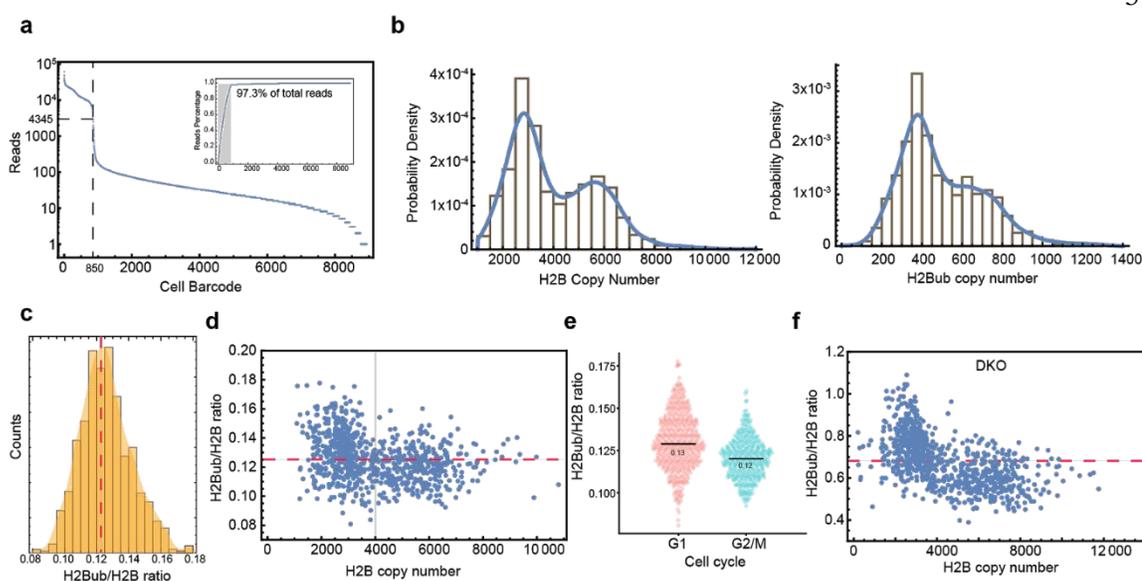
In conclusion, we have demonstrated a novel method to quantify a targeted protein and its isoforms in single cells using DNA barcoding and next-generation sequencing. The covalent DNA barcoding of a single-cell protein developed here allows us to differentiate protein isoforms based on their intrinsic properties (e.g., molecular weight), bypassing the need for high-quality antibodies. Our method can be extended to quantify other PTM isoforms, such as phosphorylation, by using isoelectronic electrophoresis<sup>20</sup> or phos-tag gels<sup>21</sup>. In addition, the multiplicity of our method could be further increased by adding additional ligation steps, so long as the targeted proteins could still be well separated during electrophoresis. Alternatively, orthogonal spyttag/spycatcher pairs could be used to further increase the multiplicity<sup>22</sup>.

## 3.4 Figures



**Figure 1.** a) Method scheme. *S. cerevisiae* strain containing spyttag at the C-terminal of protein of interest is constructed and the cells are reacted with spycatcher oligo to covalently attach DNA oligo to targeted proteins in situ. Then the targeted proteins of interest from each single cell are uniquely labeled with two rounds of “split-pool” barcoding. The cells are randomly distributed into wells, and well-specific first barcodes were ligated to the DNA oligo on the proteins via T7 ligation. Then the cells were pooled together and randomly distributed again into wells and second barcodes were ligated. After two rounds of split-pool barcoding, cells are pooled together, lysed to extract the protein-oligo conjugates, and analyzed by gel electrophoresis to separate

different protein isoforms. Gel bands corresponding to different protein isoforms are cut, and the DNA part of the protein-oligo conjugates are recovered and PCR amplified to generate a sequencing library. The oligos corresponding to single-cell proteins are quantified using next-generation sequencing. b) The oligo sequence design. Spycatcher-20nt oligo conjugate is synthesized in vitro. The oligo will serve as the PCR handle for sequencing library preparation. The first T7 ligation substrate oligo contains a T7 site sequence for first ligation, a random-base UMI sequence, a first cell barcode and another T7 site for second ligation. The second-ligation substrate oligo contains a T7 site for second ligation and a PCR handle for sequencing library preparation. The full length of the oligo is 112nt. c) The Western blot images of samples (1) before and (2) after spycatcher conjugation, (3) first barcoding and (4) second barcoding. Here, H2B (lower band) and its monoubiquitination isoform H2Bub (upper band) are used as an example, which show two bands in the gel. The bands shift up after each reaction, indicating successful ligation. The efficiency is estimated to be more than 90% at each step (**Supplementary Figure 2**).



**Figure 2.** a) Cell barcode identification from sequencing. The number of reads per cell barcode were plotted in descending order. A clear cutoff (dashed line) could be identified to separate barcodes with a high number of reads from the large number of sparsely populated cell barcodes. The inset shows accumulated reads percentage. The gray area corresponds to high-quality cells, which account for 97.3% of the total filtered reads. b) Left: histogram of H2B copy number in single cells. Right: histogram of H2Bub (mono-ubiquitinated H2B) copy number in single cells. c) Distribution of H2Bub/H2B ratio in single cells. The red dashed line is the average, which is in alignment with previous ensemble measurement. d) The H2Bub/H2B ratio as a function of H2B copy number in single cells. Each dot represents a single cell. The red dashed line is the population-average H2Bub/H2B ratio. The gray line divides the cells into G1 and G2/M cell-cycle stages. e) The distributions of H2Bub/H2B ratio for cells in G1 and G2/M stages, respectively. The two stages have different distributions: G1-stage cells have a subpopulation with a higher H2Bub/H2B ratio. f) In the UBP8 and UBP10 double knockout strain, the H2Bub/H2B ratio versus H2B copy number in single cells. Each dot represents a single cell. The red dashed line is the population-average H2Bub/H2B ratio.

### 3.5 Reference

1. Zhao, Y. & Garcia, B. A. Comprehensive Catalog of Currently Documented Histone Modifications. *Cold Spring Harb. Perspect. Biol.* 7, a025064 (2015).
2. Hafner, A. et al. p53 pulses lead to distinct patterns of gene expression albeit similar DNA-binding dynamics. *Nat. Struct. Mol. Biol.* 24, 840–847 (2017).
3. Zambrano, S., De Toma, I., Piffer, A., Bianchi, M. E. & Agresti, A. NF- $\kappa$ B oscillations translate into functionally related patterns of gene expression. *Elife* 5, e09100 (2016).
4. Cai, L., Dalal, C. K. & Elowitz, M. B. Frequency-modulated nuclear localization bursts coordinate gene regulation. *Nature* 455, 485–490 (2008).
5. Hughes, A. J. et al. Single-cell western blotting. *Nat. Methods* 11, 749–755 (2014).
6. Regot, S., Hughey, J. J., Bajar, B. T., Carrasco, S. & Covert, M. W. High-sensitivity measurements of multiple kinase activities in live single cells. *Cell* 157, 1724–1734 (2014).
7. Goltsev, Y. et al. Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell* 174, 968–981.e15 (2018).
8. Spitzer, M. H. & Nolan, G. P. Mass Cytometry: Single Cells, Many Features. *Cell* 165, 780–791 (2016).
9. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* 14, 865–868 (2017).
10. Shahi, P., Kim, S. C., Haliburton, J. R., Gartner, Z. J. & Abate, A. R. Abseq: Ultrahigh-throughput single cell protein profiling with droplet microfluidic barcoding. *Sci. Rep.* 7, 44447 (2017).
11. Acharya, P., Quinlan, A. & Neumeister, V. The ABCs of finding a good antibody: How to find a good antibody, validate it, and publish meaningful data. *F1000Res.* 6, 851 (2017).
12. Maecker, H. T. & Trotter, J. Flow cytometry controls, instrument setup, and the determination of positivity. *Cytometry A* 69, 1037–1042 (2006).
13. Voskuil, J. Commercial antibodies and their validation. *F1000Res.* 3, 232 (2014).
14. Zakeri, B. et al. Peptide tag forming a rapid covalent bond to a protein, through engineering a bacterial adhesin. *Proceedings of the National Academy of Sciences* 109, E690–E697 (2012).
15. Cusanovich, D. A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* 348, 910–914 (2015).
16. Kivioja, T. et al. Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* 9, 72–74 (2011).

17. Takemori, N. et al. Top-down/Bottom-up Mass Spectrometry Workflow Using Dissolvable Polyacrylamide Gels. *Anal. Chem.* [acs.analchem.7b00357](https://doi.org/10.1021/acs.analchem.7b00357) (2017).
18. Robzyk, K., Recht, J. & Osley, M. A. Rad6-dependent ubiquitination of histone H2B in yeast. *Science* 287, 501–504 (2000).
19. Schulze, J. M. et al. Splitting the task: Ubp8 and Ubp10 deubiquitinate different cellular pools of H2BK123. *Genes Dev.* 25, 2242–2247 (2011).
20. Tentori, A. M., Yamauchi, K. A. & Herr, A. E. Detection of Isoforms Differing by a Single Charge Unit in Individual Cells. *Angew. Chem. Int. Ed Engl.* 55, 12431–12435 (2016).
21. Kinoshita, E., Kinoshita-Kikuta, E. & Koike, T. Separation and detection of large phosphoproteins using Phos-tag SDS-PAGE. *Nat. Protoc.* 4, 1513–1521 (2009).
22. Liu, Y. et al. Tuning SpyTag-SpyCatcher mutant pairs toward orthogonal reactivity encryption. *Chem. Sci.* 8, 6577–6582 (2017).

## 3.6 Supporting information

### 3.6.1 Experiments

#### Yeast strains and plasmids

The *S. cerevisiae* strains used in this study were backgrounded by BY4741 (MATa his3 leu2 met15 ura3). The C-terminal tagging of spycatcher and 3XFlag was performed by PCR amplification of plasmid (pCHS81) with ~70–500 bp overhangs to be homologous to the target loci, and the resulting PCR product was chromosome integrated by standard LiOAc transformation protocol.

#### Spycatcher oligo synthesis

The spycatcher was C-terminal tagged with 6xHis-tag and sequences coding “LCTPSR” peptide sequence to be coupled with maleimide peg4 tetrazine. The plasmid (pCHS86) was transformed to NEB T7 express *E. coli* strain for expression. Transformed *E. coli* was grown on LB plate containing kanamycin overnight at 37°C. A fresh colony was diluted and cultured in 1L of LB containing kanamycin. When it reached OD600 of ~0.5, 500µM IPTG (RPI) was added for induction and incubated overnight at room temperature. Cells were harvested and suspended in 1x PBS containing 0.1mg/mL lysozyme (Sigma) and 0.1mg/mL DNase (Roche). After incubation on ice for 1hr, cells were lysed and the protein was extracted by French press. Imidazole was added to 20mM and the lysate was centrifuged for 10min. The supernatant was filtered using a 0.45µM pore filter and incubated in equilibrated Ni-NTA resin (Thermo) for at least 2hr at 4°C. The column was washed with 5 vol. of washing buffer (1x PBS with 25mM imidazole, pH7.4) and eluted with 5 vol. of elution buffer (1x PBS with 250mM imidazole, pH7.4). The eluate was buffer exchanged by MW cutoff spin column (Milipore) at 4°C 5000g with 20 vol. of 1x PBS to remove imidazole. Disulfide bonds in Spycatcher LCTPSR were reduced by 2X TECP (Thermo) for 30mins at room temperature and desalted by PD-10 column (GE

Healthcare). The eluate was reacted with maleimide-peg4-tetrazine for 2hr at room temperature and the reaction product was separated by PD-10 column from unreacted maleimide-peg4- tetrazine. The aliquots were prepared after measuring protein concentration by Pierce 660 reagent. TCO-labeled oligo was prepared by reaction between 5'-amine-modified oligonucleotides and TCO-PEG4-NHS ester (Click Chemistry Tools). After the reaction, the mixture was purified by HPLC and the final product was concentrated by speedvac. The concentration of TCO oligos were measured by nanodrop. Spycatcher-tetrazine was reacted with 2 vol. of TCO oligos overnight at 4°C with a constant shaking at 500rpm (**Supplementary Figure 1a**). Spycatcher-oligo conjugate was purified by ion-exchange chromatography to remove unreacted spycatcher and buffer exchanged by a MW cutoff spin column with 1x PBS. Their molecular weights were confirmed by SDS-PAGE gel (**Supplementary Figure 1b**). The aliquots were stored with 50% glycerol in -20°C until usage for cell ligation.

#### Cell culture and fixation

Fresh colonies of wild-type strain (WT) and UBP8 and UBP10 double-knockout strain (DKO) were grown in YPD media overnight. Cells were diluted in fresh YPD to OD600 of ~0.125. When they reached OD600 of ~0.5, cells were fixed by 1% formaldehyde by adding 32% formaldehyde with no methanol directly to the cell media and incubating for 30min at 30°C with a gentle shaking. Cells were then harvested and washed by buffer B (1.2M sorbitol/ 0.1M sodium phosphate, pH 7.4) three times. Spheroplasting was performed by 100µg zymolase and 10µL fresh beta-mercaptoethanol in 1mL of buffer B cell suspension for 10min at 37°C in a thermomixer with a gentle shaking. The cell suspension was inverted every 5min. The duration of the spheroplasting reaction was determined by the number of cells permeabilized. After the spheroplasting reaction, the cells were gently washed with buffer B three times. Cells were post-fixed in 1% formaldehyde in 1x PBS/0.6M KCl for 30min at RT with a gentle mixing by a slow mixing rotor. Cells were washed by buffer B for three times and spun at 1000g.

### In situ Spycatcher-oligo reaction

For spycatcher-oligo reaction,  $10^7$  cells were resuspended in 1ml 1xPBS/0.6M KCl solution containing spycatcher-oligo (final concentration: 10 $\mu$ M) and protease inhibitor cocktail (1x, Sigma). The reaction was incubated overnight at 4°C with a gentle shaking. After spycatcher-oligo reaction, cells were washed by buffer B three times.

### Split-pool barcoding with T7 ligation

For cell-specific barcoding, cells were distributed into a 96-well plate with T7 ligation reaction buffer containing T7 ligase (NEB) and first-ligation adaptor oligo (5 $\mu$ M). Barcoded ligation oligos (5 $\mu$ M) were added into each well. The reaction plate was incubated for 2hr at room temperature with a gentle shaking. Cells were pooled, washed by buffer B three times, and resuspended in T7 ligation reaction buffer B containing T7 ligase and second-ligation adaptor oligo (5 $\mu$ M). Cells were distributed into a 96-well plate and mixed with second-barcoded ligation oligos (5 $\mu$ M). After incubating the reaction plate for 2hr at room temperature, cells were pooled and washed with buffer B three times. The cell density was measured using hemocytometer and cell suspension corresponding to 900 cells were pipetted out.

To estimate the number of barcodes representing more than two cells, the “collision” rate was calculated and simulated (**Supplementary Figure 6**) as in previous work<sup>1</sup>. To keep expected collision rate lower than 10% with 9,216 possible barcode combinations, 900 cells were aliquoted in the experiment.

For “dummy” sample preparation, we synthesized spycatcher oligo with a dummy sequence as previously described (**Supplementary Figure 4**). After the spycatcher-dummy oligo reaction, all cell pellets were sequentially ligated to the first- and second-ligation oligos whose sequences will not be amplified by primers for Illumina sequencing library preparation. The 3' end of second-ligation oligo is one nucleotide shorter and modified by dTC6 amine and labeled with a rhodamine dye TAMRA-NHS,

in order to have a similar molecular weight and to enable visualization of the ligation bands in gel analysis by a typhoon scanner in order to cut target bands. The ligation efficiency was checked by Western blot against 3xFLAG tag.

#### Separation of H2B and H2Bub on SDS-PAGE and DNA recovery

2xlaemmli buffer (Bio-Rad) was added to both the barcoded sample and the dummy sample and boiled at 95°C for 10min. The barcoded sample and the dummy sample were mixed and loaded in a 10% dissolvable polyacrylamide gel. H2B and H2Bub were well separated due to the different molecular weight (~7kD). Using a typhoon scanner image with TAMRA fluorescence, the target protein-oligo conjugate bands were visualized and cut off, and the protein-oligo conjugates were recovered. The procedures for preparing dissolvable PAGE with crosslinker ethylene-glycol-diacrylate (EDA) and for gel recovery followed previous work<sup>2</sup> and allowed for high-recovery yield (**Supplementary Figure 5**).

#### Library Preparation and Sequencing

For next-generation sequencing library preparation, two rounds of PCR amplification were carried out. First, DNA was amplified using primers that bind to the oligo sequences from barcoded cells. Then sequencing adapters were appended using NEBNext Multiplex Oligos for Illumina (NEB) in second-round PCR. The amplification conditions for the first PCR were as follows: 95°C 1min, then 10–15 cycles at 95°C 10s/ 62°C 15s/ 65°C 30s, and a final extension at 65°C 3min. The number of cycles required for the first-round PCR can be determined from analyzing small aliquots of the sample on a qPCR machine. The number of cycles is determined from the point of exponential phase amplification. The optimal number of cycles can vary between different samples, but is usually in the range of 10–15 cycles for the first-round PCR. The PCR amplification condition for the second-round PCR was as follows: 95°C 1min, then four cycles at 95°C 10s/ 62°C 15s/ 65°C 30s, and a final extension at 65°C 3min. We found that fewer than five cycles for the second-round PCR showed the sharp bands on

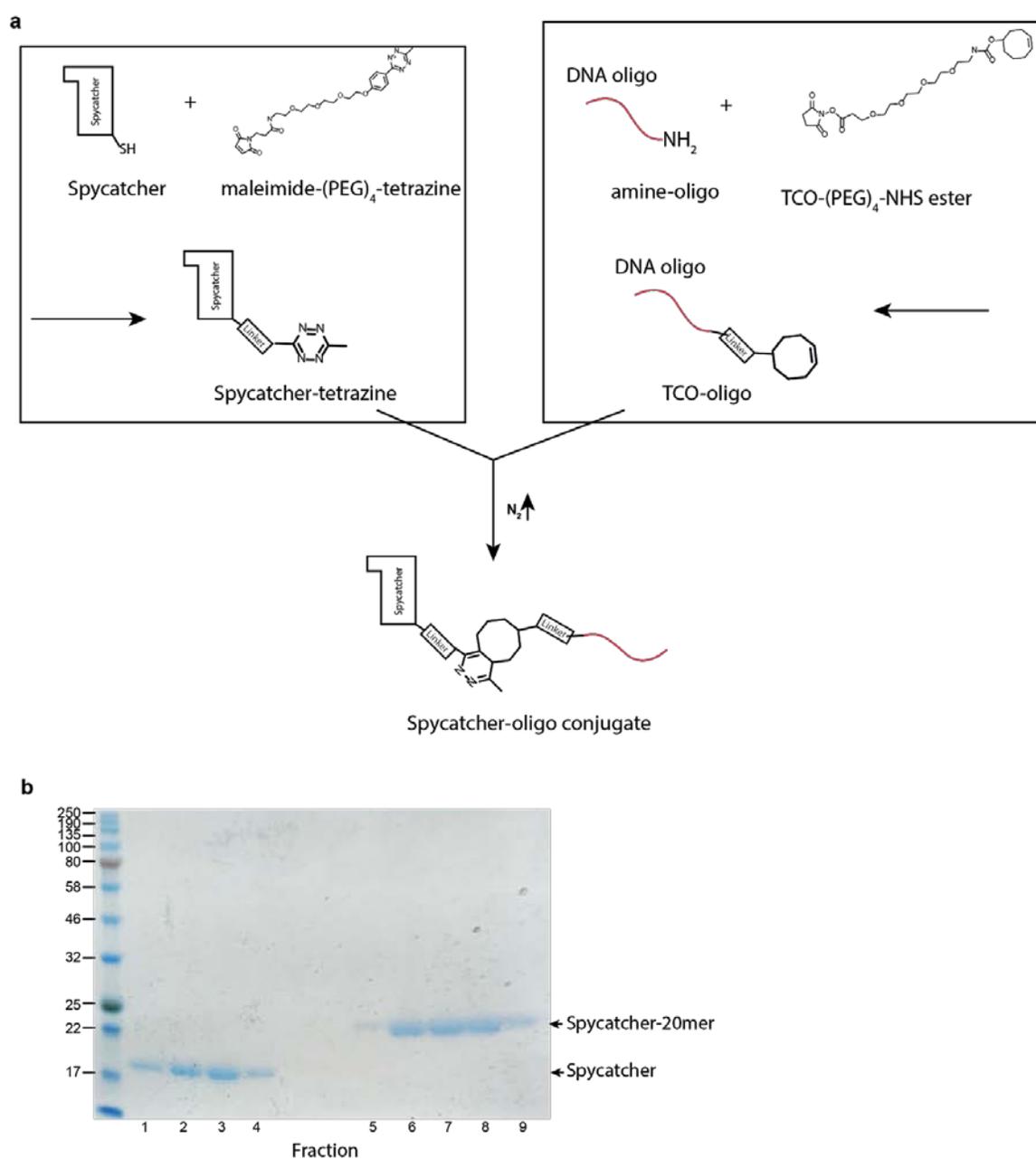
an agarose gel and avoided PCR overamplification to minimize PCR amplification bias. After each round of PCR, PCR amplicons were run on 3% agarose gel and purified using a gel extraction kit (Qiagen). The PCR-amplified library was quantified using a Qubit High-sensitivity DNA kit (Invitrogen). The final purified amplicons were sequenced on a HiSeq 2500 (Illumina) with the targeted read depth of 5–25 million per gel band.

### 3.6.2 Data analysis

The sequencing reads were first filtered based on the constant fixed region in the oligo (the constant region includes the PCR handle, the first T7 ligation site, and the second T7 ligation site). Reads that had more than one mismatch against a constant region were disregarded. Then, the first-cell barcode and second-cell barcode were connected to generate the full-cell barcode. Reads with cell barcodes which did not match the set of barcode combinations (96\*96 in total) were disregarded. The number of reads associated with each barcode were then calculated and the real-cell barcodes were identified as those with much higher number of reads than non-real-cell barcodes (**Figure 2a**, **Supplementary Figure 7**). For the H2B/H2Bub sample, the common cell barcodes identified from H2B and H2Bub (848 out of 850) were used for further analysis.

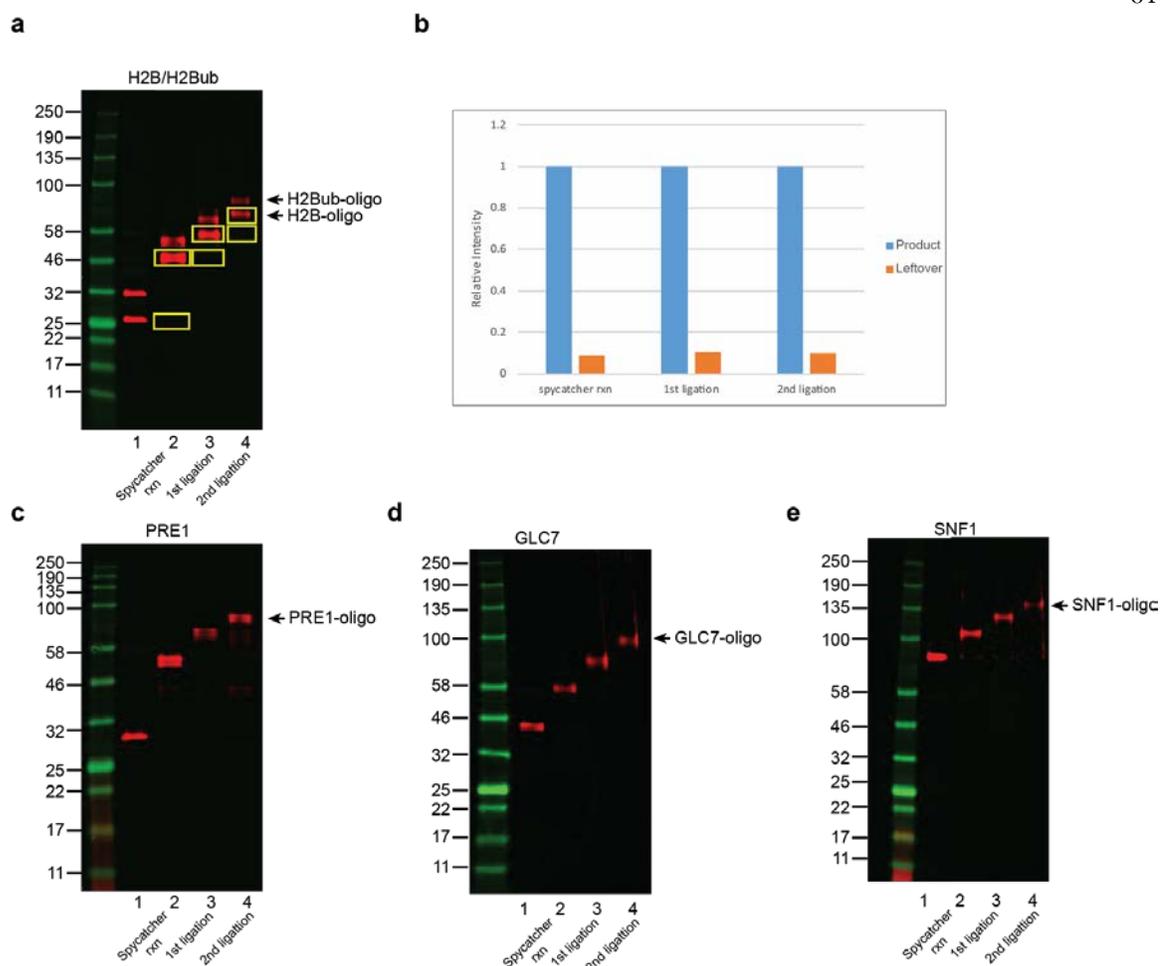
To verify that the UMIs had enough coding space to encode all the proteins in single cells, we computationally shortened the UMIs and counted how many unique UMIs we could identify from sequencing results (**Supplementary Figure 8**). To verify that the sequencing depth was high enough to sample all the UMIs, we computationally subsampled the sequencing reads and calculated how many UMIs observed were associated with single-cell barcodes (**Supplementary Figure 8**). We found that different sequencing depths were needed to saturate the library. For example, for the H2B sample, 25 million reads were required under current conditions, while for the H2Bub sample, only 5 million reads were required for library saturation. This reflects the different complexity of these two libraries, which corresponds to different protein copy numbers inside the cells.

## 3.6.3 Supplementary Figures

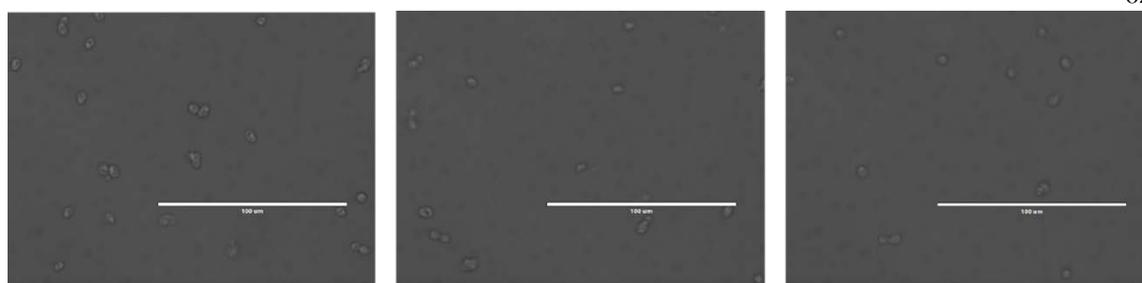


**Supplementary Figure 1.** a) Scheme for spycatcher-oligo conjugate synthesis. Spycatcher with a cysteine at the C-terminal region was reacted with the maleimide-PEG4-tetrazine to generate spycatcher-tetrazine. 5' amine-modified oligo was reacted with NHS ester-PEG4-TCO to generate oligo-TCO. The spycatcher-tetrazine and oligo-TCO were then conjugated together via click chemistry. b) The gel

electrophoresis results for spycatcher-oligo (20mer) purification. Spycatcher oligo (20mer) was separated from unreacted spycatcher and purified using ion-exchange chromatography. The number from 1 to 9 indicates the different fraction from ion-exchange chromatography. Fractions were mixed with laemmli buffer and analyzed after boiling by SDS-PAGE with Coomassie staining.



**Supplementary Figure 2.** a) The Western blot images of H2B samples before and after spycatcher reaction, after first ligation, and after second ligation. The bands shift up, indicating a successful reaction. b) The intensities of the regions (indicated by yellow boxes) were determined from images after removing background in order to estimate the reaction efficiency. The intensity for leftover signal is less than 10% of the product, indicating that the reaction efficiency is about 90%. c–e) The reaction and ligation efficiency of different target proteins to demonstrate the generality of our labeling method.

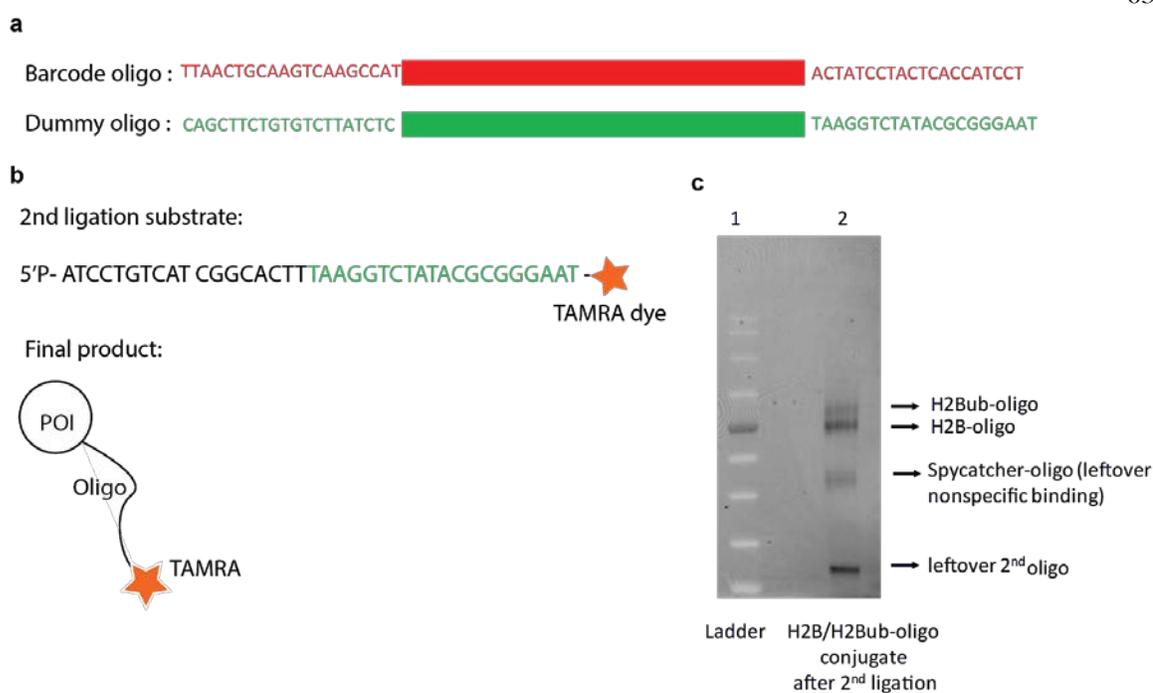


After Spycatcher-oligo conjugation

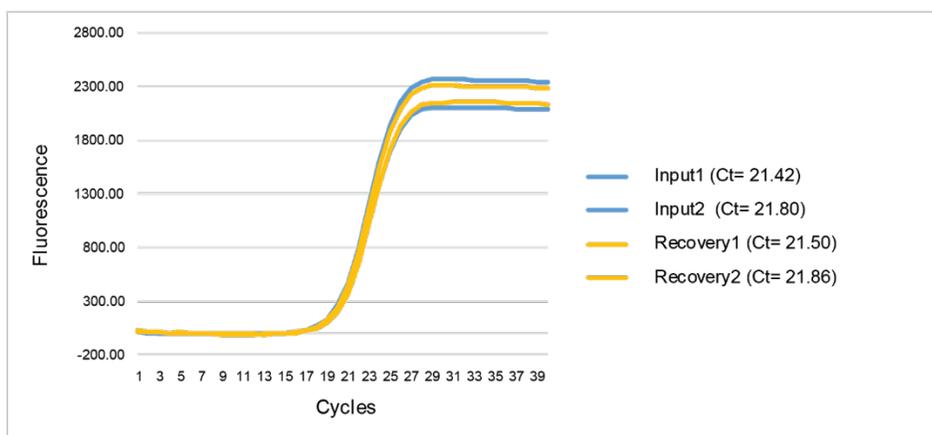
After 1st ligation

After 2nd ligation

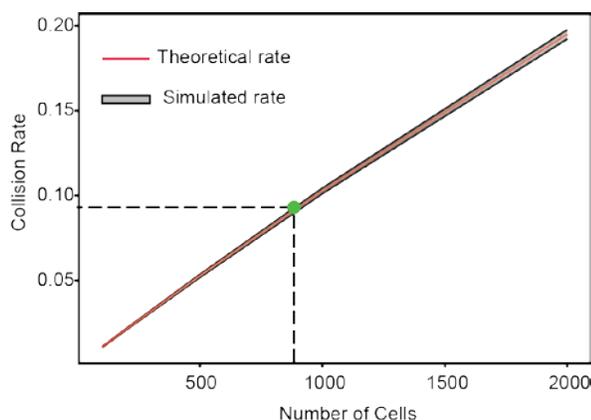
**Supplementary Figure 3.** Cell morphology check after spycatcher reaction, first ligation, and second ligation under the microcopy. Individual intact cells can be observed after each step without morphological changes, affirming that each single cell could be used as compartments during “split-pool” barcoding. (Scale bar: 100um)



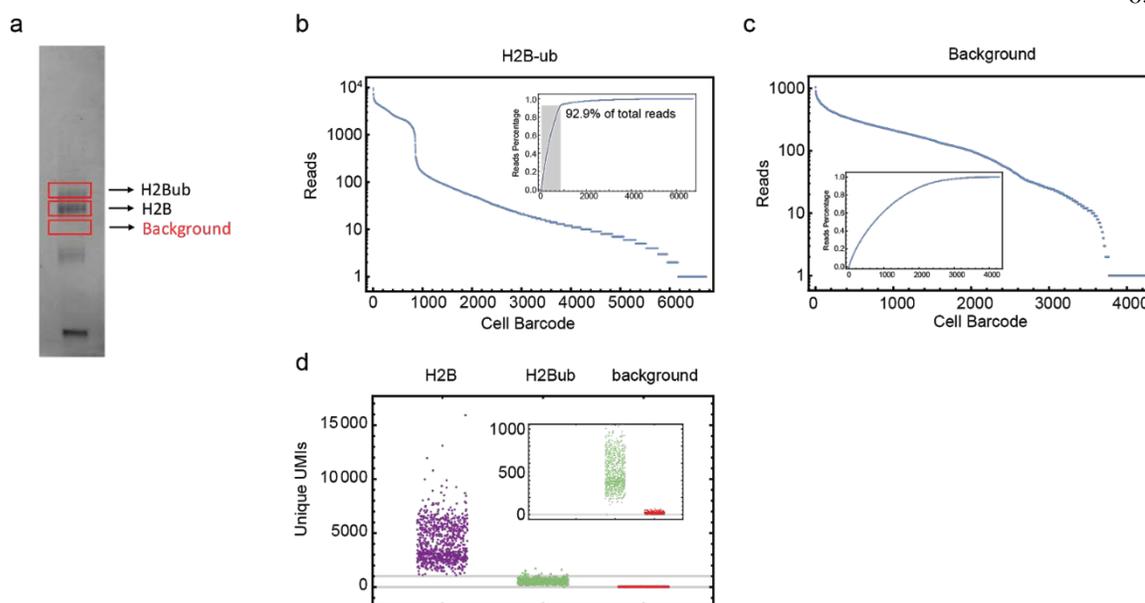
**Supplementary Figure 4.** a) Design of dummy oligo. The dummy oligo has the same length as the barcode oligo but has different sequences in the PCR handle. Therefore, proteins labeled with dummy oligo will co-immigrate with proteins labeled with barcode oligo during gel electrophoresis, but will not be amplified during library preparation. b) The second-ligation substrate used in the dummy sample preparation has a TAMRA dye at its end. The protein-oligo conjugate can then be visualized on the gel using fluorescence. c) The fluorescent gel image to visualize TAMRA dye from typhoon scanner. H2Bub-oligo, and H2B-oligo conjugated with TAMRA dye can be identified. The band whose size corresponds to spycatcher oligo after two rounds of barcoding ligation can also be observed. This may come from unreacted, leftover spycatcher-oligos that bind non-specifically inside the cells and that are further barcoded during pool-split barcoding. Since those spycatcher-full-length oligo products can be PCR amplified and thus interfere with quantification, this result means that, by running the gel, we can further remove non-specific background, resulting in accurate quantification.



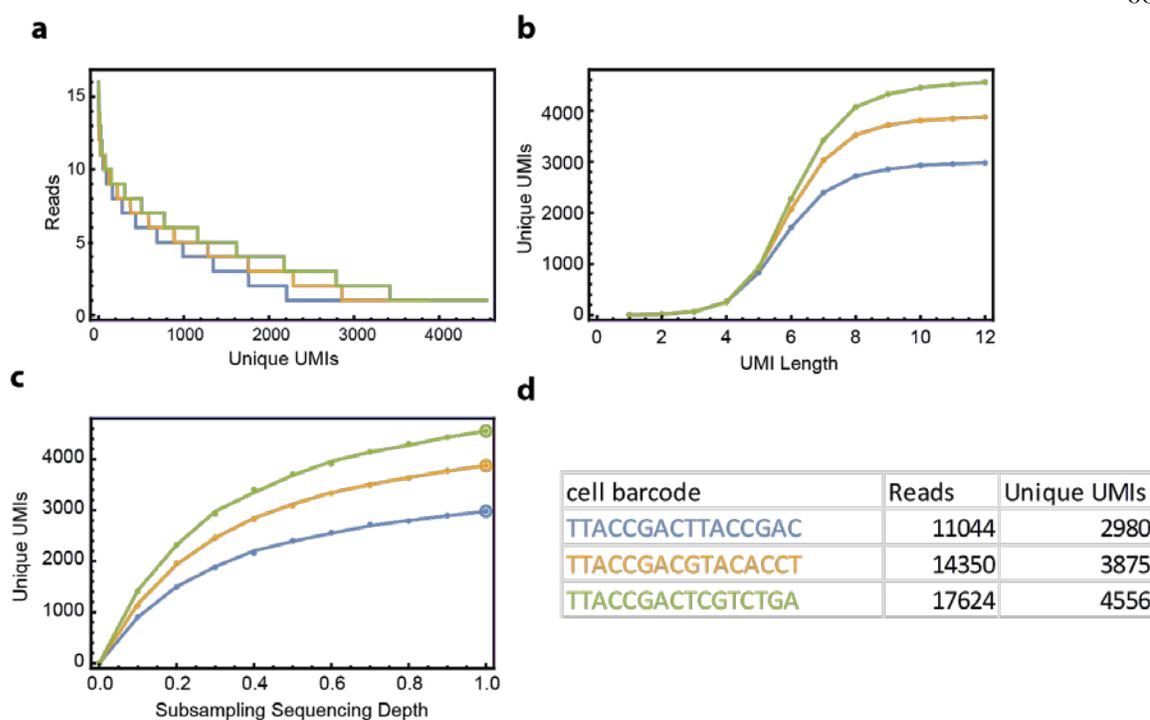
**Supplementary Figure 5.** The qPCR result for measuring gel extraction yield. A DNA oligo (112nt) sample was ran through the dissolvable PAGE gel and the corresponding band was cut from the gel, followed by the oligo extraction. The input sample and the sample from gel extraction were quantified by qPCR. The Ct values are almost the same within qPCR errors, which indicates the high efficiency of oligo recovery from the gel. It should be noted that here the 112nt oligo is used to test the recovery efficiency. We reasoned that the result for protein-oligo complex could be qualitatively similar.



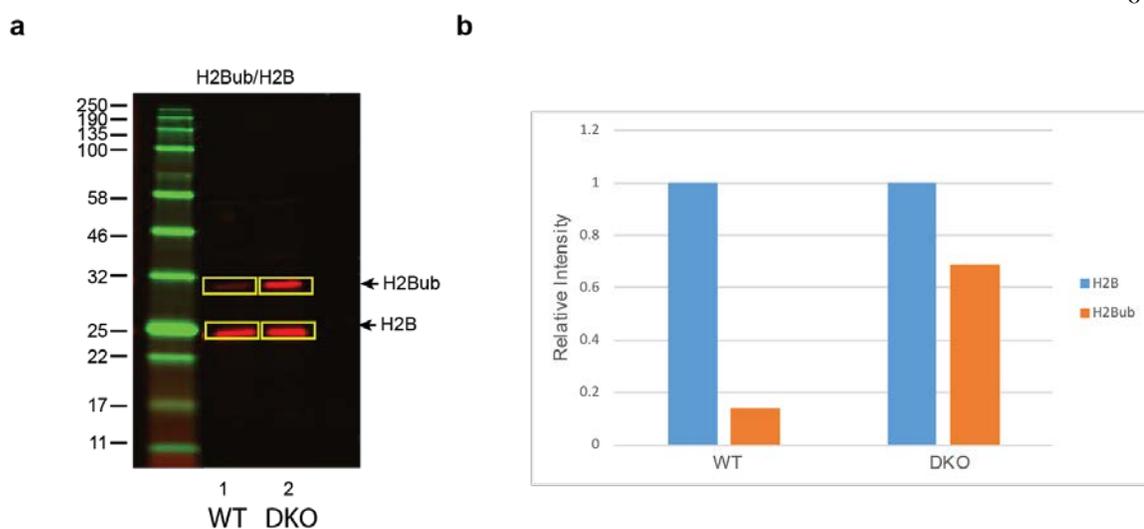
**Supplementary Figure 6.** Expected collision rate. The orange line shows the expected theoretical collision rate (supplementary method). The gray ribbon shows estimated 95% confidence interval of simulated collision rates. The green dot represents the expected collision rate (<10%) when 900 cells are sampled, as in our experiment.



**Supplementary Figure 7.** a) Targeted protein showing up as fluorescent bands in the TAMRA channel using the typhoon scanner. Bands corresponding to different targeted proteins were cut off and DNA-protein conjugation was extracted from the gel slab. In addition to targeted proteins, a blank gel piece was also cut as background. b) Real-cell barcodes could be identified from H2Bub band, as a clear cutoff separated significant barcodes with a high number of reads from spurious barcodes with low number of reads. The same results were obtained for the H2B band, as shown in Figure 2a. The high-quality cell barcodes identified from H2B and from H2Bub are essentially the same (848 out of 850), which further affirms that those barcodes represent real single cells. c) By contrast, barcodes from the background do not show the clear cutoff. d) The unique UMIs (aka, protein copy numbers) associated with the real barcodes from targeted bands and from the background. This result shows that protein is clearly resolved during electrophoresis and the gel has low backgrounds.



**Supplementary Figure 8.** a) The number of reads associated with UMIs. The data for three barcodes (shown in d)) is plotted here. The number of reads varies from 16 to 1, which reflects the PCR biases. b) The number of unique UMI identified when sublength of UMIs is taken. The number of UMIs increased dramatically with the length of the UMIs and reached a plateau after around 10nt, meaning the number of unique UMIs will not increase even if a longer sequence is used. This result indicates that our current UMIs (12nt) have enough coding space, which assured a successful digital counting scheme by UMIs. c) and d) The number of unique UMIs as a function of sequencing depth (reads). As sequencing depth increases, the number of uniquely identified UMIs increases. With current full sequencing depth (1.0), the number of uniquely identified UMIs reaches a plateau, implying that all the UMIs are sufficiently sampled. Based on these, the protein copy number from single cells can be determined by the number of unique UMIs, as shown in d).



**Supplementary Figure 9.** a) Western blot images of H2B for wild-type strain (WT) and double knockout of *du8* and *du10* strain (DKO). b) The ratio between H2Bub/H2B measured from images. Relative H2Bub ratio is ~12% in WT and ~69% in DKO. Note the ensemble value agrees with the mean of the population determined from our single-cell measurement.

### 3.6.4 Reference

1. Cusanovich, D. A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* 348, 910–914 (2015).
2. Bornemann, S., Rietschel, B., Baltruschat, S., Karas, M. & Meyer, B. A novel polyacrylamide gel system for proteomic use offering controllable pore expansion by crosslinker cleavage. *Electrophoresis* 31, 585–592 (2010).