Title

Single-cell and single-molecule methods for mapping protein-DNA interactions

Author

Altemose, Nicolas

Publication Date

2021

Peer reviewed|Thesis/dissertation

Single-cell and single-molecule methods for mapping protein-DNA interactions

By

Nicolas F Altemose

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Joint Doctor of Philosophy
with the University of California, San Francisco

in

Bioengineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Assistant Professor Aaron Streets, Chair
Professor James Wells
Associate Professor Nicholas Ingolia

Spring 2021

Abstract

Single-cell and single-molecule methods for mapping protein-DNA interactions

by

Nicolas F Altemose

Joint Doctor of Philosophy in Bioengineering

University of California, Berkeley and University of California, San Francisco

Assistant Professor Aaron Streets, Chair

The same two meters of DNA is carefully packed into the nucleus of nearly every cell in a human's body, where it encodes essentially all of the complex information required to build a complete human being. However, DNA by itself cannot give rise to life; it must be decoded and maintained by specialized macromolecules, including proteins that read, regulate, replicate, recombine, and repair DNA. Mapping where and how these life-giving proteins interact with DNA can provide key insights into how they function or malfunction in healthy and diseased cells.

High-throughput DNA sequencing technologies form the basis of several powerful methods for mapping protein-DNA interactions across the genome, but they often require researchers to blend together many thousands or millions of cells to provide enough material to make an accurate measurement. Due to this blending, these bulk methods cannot capture the dynamic and heterogeneous nature of protein-DNA interactions as they regulate the genome in individual cells. While newer methods are beginning to enable protein-DNA mapping in single cells, they are incompatible with high-resolution microscopy, which can provide rich orthogonal information about nuclear organization and other complex phenotypes in single cells. Furthermore, existing protein-DNA mapping approaches fail almost completely within highly repetitive DNA sequences, which constitute roughly 5-10% of the human genome and play indispensable roles in maintaining genome stability.

In this body of work, I have developed two new technologies to address each of these limitations in turn. Firstly, I designed an integrated microfluidic platform (μDamID) that combines high-resolution imaging and sequencing information in the same single

cells, allowing for the joint analysis of the nuclear localization, sequence identity, and variability of protein-DNA interactions in single cells. Secondly, I worked collaboratively to develop DiMeLo-seq (Directed Methylation with Long-read sequencing), which uses cutting-edge DNA sequencing technologies to map protein-DNA interactions on long, single molecules of DNA that retain endogenous DNA methylation marks and can be mapped to highly repetitive regions of the genome. Together, these new methods expand the toolkit available to researchers to study the fundamental processes that regulate the genome, with the potential to enhance our understanding of embryo development, stem cell differentiation, and diseases resulting from genome misregulation.

Table of Contents

List of Figures

Acknowledgements

Mentors, colleagues, and collaborators
I joined Dr. Aaron Streets's lab as his very first student before the lab had even opened. Along with Jonathan White, who was a technician at the time, we turned an empty room into a fully operational microfluidics, imaging, and sequencing lab. Helping to grow the lab has been an incredibly informative and fulfilling experience for me, and Aaron's leadership and mentorship have guided and inspired me to become a better scientist. Aaron exemplifies a blend of brilliance and kindness that can be rare among scholars. He cares deeply about both the science and his trainees' wellbeing and success. Aaron is uncompromising in his principles of fairness and collegiality, he is generous with his time, and he uses his sharp emotional intelligence to cheer on and coach trainees through the highs and many lows of research. I'm proud to be his first student, and I aspire to be as good a scientist, leader, and mentor as he is.

I'm also very grateful for the mentorship and support provided my other committee members, Professors Jim Wells and Nick Ingolia, as well as Dr. Ushma Neill and Professors Aaron Straight and Gary Karpen. I'm thankful for the continued mentorship of Dr. Karen Miga, who has been part of my scientific journey from the very beginning, since I was only 18, and whose scientific vision, tenacity, and leadership I admire greatly. Finally, I'm thankful for past mentors Professors Hunt Willard, David Reich, and Simon Myers, who cheered me on and shaped me as a scientist, and to whom I owe any success.

My closest colleague through this PhD has been Annie Maslan, who has become a dear friend and has contributed immeasurably to the work described herein. Helping to recruit Annie was perhaps my single biggest contribution to the Streets Lab. Her brilliance, skills, mastery of the literature, organization, grit, and overall efficaciousness make her unstoppable, and I can't wait to see what she'll do next. Annie is extremely creative, thoughtful, empathetic, and dependable, and she has brought forth such great insights and ideas (and helped gently rein in many of my bad ideas!). Together we've made such a great team, and I hope some of Annie's talent has rubbed off on me over the years. I consider myself so lucky to have met Annie, and it will take some extremely good fortune to find colleagues like her in the future.

The other members of the Streets Lab have made it such an exciting and welcoming work environment. I've had so many fun days hacking together equipment and learning about microscopy with Gabriel Dorlhiac. Anushka Gupta tells the best jokes at lab meeting, and it was great to work with her on several fun side projects. I'm so

happy that Zoë Steier joined our lab and carried out such impressive work while providing very thoughtful feedback on the other projects in the lab. Adam Gayoso has been a great help on all things statistical and computational and brings a great sense of humor to all of our lab events. Rodrigo Cotrim Chaves and Dr. Soohong Kim have carried out such impressive microfluidic tech development in the lab and are always game to discuss new ideas. I've been so lucky to work with talented undergraduates and technicians like Tyler Chen, Mansi Zalavadia, Jonathan White, and Xinyi Zhang, and I've taken great joy in watching the success of my undergraduate mentees Andre Lai, Carolina Rios-Martinez, and Romy Mastel.

Our collaborators Owen Kabnick Smith, Dr. Kousik Sundararajan, and Rachel Brown in Professor Aaron Straight's group at Stanford have been a joy to work with, and I look forward to publishing with them. Thanks also to Dr. Carolyn de Graaf for providing us with data, to Professors Sophie DuMont, Bo Huang and Bianxiao Cui for providing guidance and/or materials, and to Professor Bas van Steensel for providing us with plasmids. Many thanks to Professor Amy Herr's lab, Professor Jay Groves's lab, and others for loaning us equipment and expertise, especially when we were first getting our lab set up.

Funding bodies and support staff

Science, including the process of training scientists, requires a lot of financial, administrative, and infrastructural support. I am extremely grateful to Howard Hughes Medical Institute for making my graduate studies at UC Berkeley possible by funding me with a Gilliam Fellowship for Advanced Study, which provided me with the flexibility to pursue a second PhD, the freedom to choose a lab without regard to stipend funding (which freed up funds to cover research costs), and the gift of belonging to an amazing community of scholars, several of whom have become close friends. I am also extremely grateful to the Department of Bioengineering, the Murray Slater Foundation, and the Siebel Scholars Program for funding the remainder of my PhD. The research itself was funded by the Department of Bioengineering, the Chan Zuckerberg Biohub Investigator program, and the National Institute of General Medical Sciences of the National Institutes of Health [Grant Number R35GM124916]. I am very grateful to the donors and taxpayers who funded all of this, and I hope to continue paying back this investment in me, in the form of new fundamental knowledge and new technologies for advancing science.

I am especially grateful to the UC Berkeley employees who kept the lights on, removed our lab waste, delivered our packages, and supported our grants and purchases. The staff in Stanley Hall were always such a pleasure to work with, and I'm grateful to Claris

Garzon, April Alexander, Chris Hardin, Mike Bentley, Kevin Werk, Martin Moreno, Harry Stark, Thom Opal, Kris Thompson, Paul Lum, Mary West, Geoff Bingaman, Kristin Olson, Rocio Sanchez, and Dave Rogers for everything they did to keep our lab and building open, safe, stocked, accessible, funded, and operational, especially during COVID-19 lockdowns.

Friends & family

I moved to the Bay Area from the UK at the beginning of this PhD knowing almost no one in the area, and I'm grateful to Ron Dahl and to my PhD cohort for helping me to get my initial footing. In particular, thanks to Lindsey Osimiri, Zoë Steier, Thomas Carey, Anjali Gopal, Tiama Hamkins-Indik, Tanner Dixon, Max Armstrong, Katrina Kalantar, Andrew Ng, Devante Horne, and Will Lykins for being such a valuable support network. I'm also grateful to my close friends at UC Berkeley, Nick Angelides, Andoni Mordokoutas, and Ignacio Pérez-Pozuelo, for keeping me healthy and emotionally supported throughout the PhD. Many thanks to my dear friends Adam Ward and Blake van Grouw for helping me to take breaks during the writing of this dissertation. And a special thanks to Dr. Kyle Daniels for being an invaluable source of encouragement and understanding as we look to next stages of our careers.

I've been supported and cheered on throughout my life by my wonderful brother, Ryan Altemose, and by my parents, who encouraged me to follow my passions and gave me the freedom I needed to develop an intrinsic love of learning (thanks also for their patience as I carried out genuinely dangerous but formative experiments in our garage). I deeply appreciate the love and support of Dr. Julia Dahl, as well as Annie & Willie McCrudden, Charo & Martin Richards, and Bernie Cassidy. I'm grateful for our cats, Dagny and the late Hank, who added calm, whimsy, and love to our lives. Finally, thank you to my wonderful and loving husband, Dr. Garreth McCrudden, for helping me to be a functioning adult and a better person, for supporting my scientific obsessions, and for grounding me in the importance of my everyday life and the people in it.

Curriculum Vitae for Nicolas Altemose, DPhil

altemose@berkeley.edu | about.me/altemose

PRIOR EDUCATION

University of Oxford, Oxford, United Kingdom **2011-2015**
- DPhil in Statistics (Statistical Genetics)

Duke University, Durham, North Carolina **2007-2011**
- BS in Biology, with a Concentration in Genomics and a Minor in Computational Biology and Bioinformatics

HONORS AND AWARDS

- **Howard Hughes Medical Institute Hanna H. Gray Fellowship**, 2021
- **Siebel Scholarship**, 2020-2021
- **Howard Hughes Medical Institute Gilliam Fellowship**, 2013-2019
- **Marshall Scholarship**, United Kingdom, 2011-2013
- **Angier B. Duke Scholarship**, Duke University, 2007-2011
- **Edward C. Horn Memorial Prize for Excellence in Biology**, Duke University, 2011
- **Barry M. Goldwater Scholarship**, 2010
- **UC Berkeley Nominee for the Regeneron Prize for Creative Innovation**, 2017
- **Prize for Best Talk**, Quantitative Genomics Student Conference, London, 2014
- **Summa Cum Laude**, Duke University, 2011
- **Graduation with Distinction**, Duke University, 2011
- **Phi Beta Kappa Society**, Duke University, 2010

PRIOR RESEARCH EXPERIENCE

Lab of Prof. Simon Myers, DPhil, Dept. of Statistics, Univ. of Oxford **2011-2015**

Lab of Prof. David Reich, DPhil, Dept. of Genetics, Harvard Medical School **2011**

Lab of Prof. Hunt Willard, PhD, Dept. of Biology, Duke University **2007-2011**
under the direct supervision of **Karen H. Miga, PhD**

Chapter 1

Introduction

Motivation

Complex life depends on protein-DNA interactions that constitute and maintain the epigenome, including interactions between DNA and histone proteins, transcription factors, DNA (de)methylases, and chromatin remodeling complexes, among others. These interactions enable the static DNA sequence inside the nucleus to dynamically execute different gene expression programs that shape the cell's identity and behavior.

Methods for measuring protein-DNA interactions have proven indispensable for understanding the epigenome, though to date most of this knowledge has derived from experiments in bulk cell populations. By requiring large numbers of cells, these bulk methods can fail to capture critical epigenomic processes that occur in small numbers of dividing cells, including processes that influence embryo development, developmental diseases, stem cell differentiation, and certain cancers. By averaging together populations of cells, bulk methods also fail to capture important epigenomic dynamics occurring in asynchronous single cells during differentiation or the cell cycle. Because of this, bulk methods can overlook important biological heterogeneity within a tissue. It also remains difficult to pair bulk biochemical data with imaging data, which inherently provide information in single cells, and which can reveal the spatial location of protein-DNA interactions within the nuclei of living cells. These limitations underline the need for high-sensitivity single-cell methods for measuring protein-DNA interactions.

Another major limitation of existing methods for mapping protein-DNA interactions is the inability to map these interactions in highly repetitive regions of the genome, owing to the short length of DNA sequencing reads produced by next-generation DNA sequencing technologies. In the human genome, these highly repetitive regions account for 5-10% of the total length of the genome but have remained almost entirely missing from the human genome assembly for the last 20 years (I. H. G. S. Consortium 2001). These missing regions are composed primarily of centromeres, telomeres, and the short arms of the acrocentric chromosomes, all of which serve important biological functions but have been ignored by most sequencing-based functional studies in the

age of genomics. Although new long-read sequencing technologies (M. Jain et al. 2016, Wenger et al. 2019) are now enabling the completion of the human genome reference assembly to include these highly repetitive regions (Miga et al. 2020), it remains challenging to uniquely map short sequencing reads within highly homogenized DNA repeats. This highlights the need for new protein-DNA interaction mapping methods that fully leverage the power of new long-read sequencing technologies to study the regulation and function of these formerly missing regions of the genome.

Existing methods for mapping and visualizing protein-DNA interactions

Chromatin Immunoprecipitation

Most approaches for mapping protein-DNA interactions rely on chromatin immunoprecipitation (ChIP) (Solomon et al. 1988), in which protein-DNA complexes are physically isolated using a high-affinity antibody against the protein, then purified by washing and de-complexed so the interacting DNA can be amplified and measured. The most widely used among these methods is chromatin immunoprecipitation with sequencing (ChIP-seq) (Barski et al. 2007, Johnson et al. 2007, Robertson et al. 2007), which has formed the backbone of several large epigenome mapping projects (Celniker et al. 2009, T. E. P. Consortium 2012, Meuleman et al. 2015). I used ChIP-seq extensively in my past work to study the DNA-binding properties of the meiotic recombination initiation protein PRDM9 (Altemose et al. 2017, Davies et al. 2016, R. Li et al. 2019). Protein-DNA mapping data proved essential for us to discover how changes in PRDM9's DNA binding patterns can lead to infertility and the early stages of speciation in mice (Davies et al. 2016, R. Li et al. 2019).

One drawback of ChIP-seq is that protein-DNA complexes, which are often fragile, must survive the shearing or digestion of the surrounding DNA, as well as several intermediate washing and purification steps, in order to be amplified and sequenced. This often requires fixing the sample with formaldehyde to covalently crosslink proteins to DNA, then removing these crosslinks after shearing and immunoprecipitation, a process that can result in substantial artifacts (Teves et al. 2016). Overall, ChIP methods typically have inefficient recovery of DNA from on-target protein-DNA complexes, which is overcome by starting with a large amount of input material, usually from millions of cells.

In situ antibody-targeted cleavage

More recent immunoaffinity-based methods have lower input requirements relative to ChIP-seq, but they recover relatively few interactions in small numbers of cells or single cells (Carter et al. 2019, Grosselin et al. 2019, Harada et al. 2019, Jakobsen et al. 2015,

Kaya-Okur et al. 2019, Ku et al. 2019, Rotem et al. 2015, Shen et al. 2015, Skene & Henikoff 2017, Wu et al. 2012, B. Zhang et al. 2016). Specifically, the methods CUT&RUN (Skene & Henikoff 2017) and CUT&TAG (Kaya-Okur et al. 2019), which are based on the earlier ChIC method (Schmid et al. 2004), involve binding an antibody to the target protein in permeabilized nuclei *in situ*. Then, the primary antibody is bound by Protein-A, a protein that binds to certain antibodies, fused to an endonuclease or transposase enzyme. After many washing steps to reduce nonspecific binding, the enzyme is activated by adding calcium to the buffer, releasing small fragments of DNA near the target protein's binding sites, which can be recovered from the supernatant, amplified, and sequenced. These *in situ* approaches greatly reduce the input requirements relative to ChIP-seq, and they preserve local chromatin structure while eliminating the need for heavy crosslinking.

DamID plus short-read sequencing

An alternative method for probing protein-DNA interactions, called DNA adenine methyltransferase identification (DamID), relies not on physical separation of protein-DNA complexes (as in ChIP-seq), but on a sort of 'chemical recording' of protein-DNA interactions onto the DNA itself, which can later be selectively amplified (**Figure 1.1**) (Steensel & Henikoff 2000, Vogel et al. 2007). This method utilizes a small enzyme from *E. coli* called DNA adenine methyltransferase (Dam). When genetically fused to the protein of interest, Dam deposits methyl groups near the protein-DNA contacts at the N6 positions of adenine bases (^{m6}A) within GATC sequences (which occur once every 270 bp on average across the human genome). That is, wherever the protein contacts DNA throughout the genome, ^{m6}A marks are left at GATC sites in its trail. These ^{m6}A marks are highly stable in eukaryotic cells, which do not tend to methylate (or demethylate) adenines (O'Brown et al. 2019). Dam expression has been shown to have no discernable effect on gene expression in a human cell line, and its ^{m6}A marks were shown to be passed to daughter cells, halving in quantity each generation after Dam is inactivated (Park et al. 2019). These properties allow even transient protein-DNA interactions to be recorded as stable, biologically orthogonal chemical signals on the DNA, useful for integrating protein-DNA interactions over time, up to the length of a cell cycle.

DamID reads out these chemical recordings of protein-DNA interactions by specifically amplifying and then sequencing fragments of DNA containing the interaction site. First, genomic DNA is purified and digested with DpnI, a restriction enzyme that exclusively cleaves G^{m6}ATC sites (**Figure 1.1**). Then, universal adapters are ligated onto the fragment ends to allow for amplification using universal primers. Only regions with a high density of ^{m6}A produce DNA fragments short enough to be amplified by

Polymerase Chain Reaction (PCR) and quantified by microarray or high-throughput sequencing (Wu et al. 2016). DamID has been used to explore dynamic regulatory protein-DNA interactions such as transcription factor binding (Orian 2003) and RNA polymerase binding (Southall et al. 2013) as well as protein-DNA interactions that maintain large-scale genome organization. One frequent application of DamID is to study large DNA domains associated with proteins at the nuclear lamina, near the inner membrane of the nuclear envelope (Guelen et al. 2008, Pickersgill et al. 2006, Steensel & Belmont 2017). Because DamID avoids the limitations of antibody binding, physical separations, or intermediate purification steps, it lends itself to single-cell applications. DamID has been successfully applied to sequence interactions of the protein LaminB1 with DNA in single cells in a one-pot reaction, recovering hundreds of thousands of unique DNA fragments per cell (Kind et al. 2015).

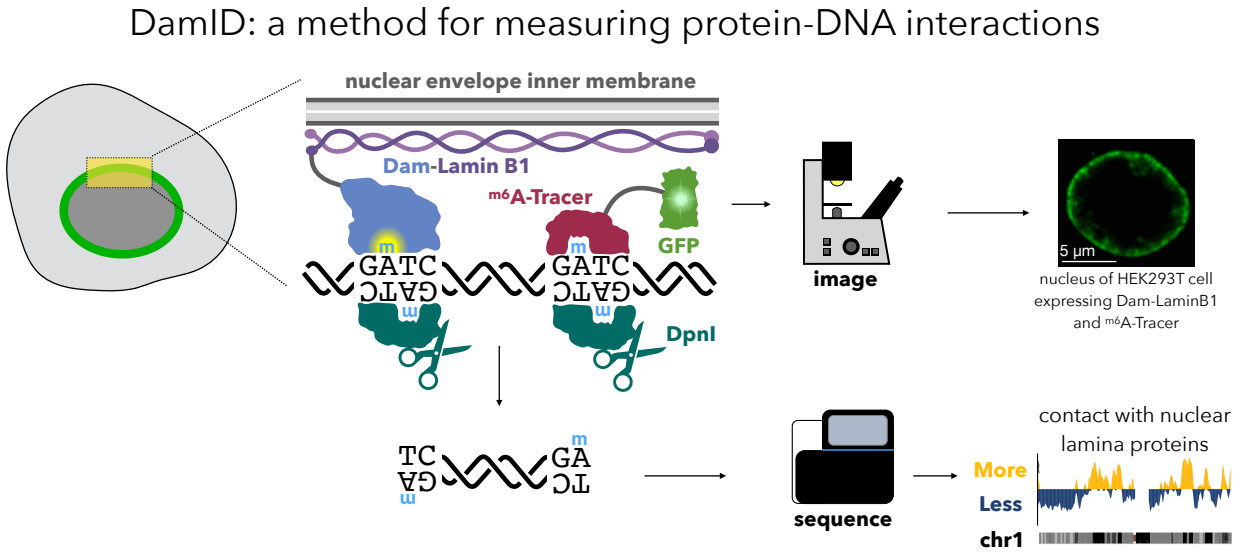


Figure 1.1. Overview of DamID and ^{m6}A-Tracer methods
This schematic illustrates how DamID can be used to map protein-DNA interactions, in this case, interactions between the Lamin B1 protein and lamina-associated domains. A fusion between Dam and the protein of interest causes DNA to be methylated at adenines within GATC sites nearby. These methyladenine marks can then be read out by imaging, using the ^{m6}A-Tracer protein, or by high-throughput sequencing after digesting the genome with a methyl-specific restriction enzyme, DpnI.

Variations of DamID
One variation of DamID, called MadID, involves fusing the nonspecific adenine methyltransferase EcoGII to the protein of interest *in vivo*, in lieu of Dam (Sobecki et al. 2018). EcoGII methylates adenines in any context, not just in GATC sites, which are rare

in certain repetitive parts of the genome. Instead of using DpnI to release and sequence small DNA fragments, methylated DNA is isolated by immunoprecipitation with an anti-mA antibody then sequenced by short-read sequencing. The authors demonstrated that this method can enrich for repetitive sequences bound by centromere and telomere binding proteins (Sobecki et al. 2018). MadID is a substantial improvement on the earlier method DamIP, which used a mutant version of Dam with less specificity than wild-type Dam, but more specificity than EcoGII (Xiao et al. 2010).

Another variation of DamID, called pA-DamID, involves targeting the Dam methyltransferase to a protein of interest *in situ*, in a similar fashion to CUT&RUN (Schaik et al. 2020). The methyl donor group S-adenosylmethionine (SAM) is withheld until the final activation step to prevent premature methylation. The cells are then processed according to the standard bulk DamID protocol. The authors show that this method generally has poorer signal-to-noise ratios than conventional DamID. This may owe to the longer effective incubation times of Dam fusion proteins expressed *in vivo*, and to the fact that active chromatin remodeling makes more DNA accessible to the Dam enzyme *in vivo*. Because methylation represents a cell's state at an instant in time, rather than an integrated signal of a protein's binding sites over time *in vivo*, this method allows for greater time resolution and was used to investigate cell cycle dynamics of LADs in synchronized cells (Schaik et al. 2020).

Researchers have also developed a split Dam enzyme to investigate protein co-localization in the genome (Hass et al. 2015). Others utilized tissue-specific promoters to constrain Dam-fusion protein expression to particular cell types in flies, allowing recovery of tissue-specific protein-DNA binding information from whole flies, without tissue isolation (Southall et al. 2013). Others have engineered different mutants of Dam with more specific activity and used them as synthetic controllers of gene expression and epigenetic inheritance (Park et al. 2019). Single-cell DamID has recently been combined with single-cell RNA sequencing in a method called scDam&T, which uses *in vitro* transcription to linearly amplify both RNA and Dam-methylated DNA from the same single cells in multiplexed batches (Rooijers et al. 2019).

DamID plus ^{m6}A-Tracer imaging

While DamID maps the sequence positions of protein-DNA interactions throughout the genome, the spatial location of these interactions in the nucleus can also play an important role in genome regulation (Bickmore & van Steensel 2013). A method related to DamID can be applied to specifically label and visualize protein-DNA interactions using fluorescence microscopy, revealing their spatial location within the nucleus in live cells (Kind et al. 2013). Visualization requires co-expression of a different

fusion protein called ^{m6}A-Tracer, which contains green fluorescent protein (GFP) and a domain that binds specifically to methylated GATC sites (**Figure 1.1**). Unlike other methods for imaging protein-DNA interactions, such as immunofluorescence (IF) plus fluorescence *in situ* hybridization (FISH), the ^{m6}A-Tracer approach allows visualization of everywhere where the protein has bound during the full incubation period, not just where the protein was bound at the moment of harvesting or fixation. This imaging technology has been applied to visualize the dynamics of LaminB1-DNA interactions within single cells (Kind et al. 2013). Both imaging and sequencing protein-DNA interactions can provide useful single-cell epigenomic information, but despite recent advances in single-cell sequencing technologies, it remains fundamentally difficult to track individual cells and pair their sequencing data with other measurements such as imaging data. While other DamID studies have performed imaging and sequencing in parallel (Borsos et al. 2019, Kind et al. 2015), they do not provide linked imaging and sequencing data for individual cells.

Approaches for combined single-cell imaging and sequencing

Microscopy is the original single-cell measurement platform, and imaging technologies still drive most cell biology studies. High-resolution imaging can provide important phenotypic information about single cells not achievable by other measurement approaches, including details about their morphology, metabolism, and subcellular protein localization over time. Using this information in combination with single-cell sequencing can provide useful insights into the connection between 'omics measurements and cellular phenotypes. Pairing imaging and sequencing data could be applied to study, for example, how the dynamic remodeling of chromatin proteins across the genome in developing cells relates to the localization of those proteins in the nucleus. Imaging prior to sequencing also allows for the identification and sorting of complex cytological phenotypes in cells, such as the presence of micronuclei and other nuclear abnormalities that would be difficult or impossible to measure using common fluorescence activated sorting methods.

However, it remains difficult to pair single-cell sequencing data with single-cell imaging data. Popular platforms for high-throughput single-cell sequencing, such as droplet-based sequencing methods (Macosko et al. 2015, Rotem et al. 2015, Satpathy et al. 2019, Zheng et al. 2017, Zilionis et al. 2017) or combinatorial indexing methods (Cao et al. 2018, C. Chen et al. 2017, Ramani et al. 2017, Vitak et al. 2017) assign sequencing barcodes at random to individual cells in a way that cannot be matched back to their individual images. It is possible to use single-cell sorting to combine single flow-cytometry-like fluorescent measurements with single-cell sequencing data (J. Q. Zhang et al. 2020), but this offers far less imaging information than obtainable on a

microscope. Some commercial microfluidic platforms enable combined imaging and sequencing, but their physical constraints make them incompatible with high-NA microscope objectives for sensitive, high-magnification imaging (Islam et al. 2014, Lane et al. 2017, Shalek et al. 2014, A. K. White et al. 2011). One possible option is to use a micromanipulator to move a cell into a well with a known barcode after imaging at high magnification (Saint et al. 2019), but this is labor intensive, slow, and prone to operator error or contamination. Cells can also be allowed to settle in or adhere to glass-bottom microwells (Yaron et al. 2014), but adhesion can take hours, and locating a cell within a microwell is nontrivial and inefficient.

To address these limitations, Streets et al. developed a custom microfluidic platform to enable combined high-resolution imaging and RNA sequencing of single cells (Streets et al. 2014, Streets & Huang 2013). This device uses active trapping of cells on a multilayer PDMS-based device with elastomeric valves and carries out each reaction step in physically separated nanoliter reaction chambers. Chen & Gupta et al. further developed this device to enable single-cell RNA library barcoding on chip (T. N. Chen et al. 2020). Other recent solutions to the problem of imaging and RNA-sequencing involve optical decoding of barcodes after random pairing of barcodes with single cells. In SCOPE-seq, combinatorial barcodes are attached to beads in microwells, and the sequence of each barcode is read out optically though iterative hybridization of fluorescently tagged oligonucleotide probes (Yuan & Sims 2016). In a different approach, barcodes are attached to beads whose identity is specified by spectral encoding with different ratios of lanthanide nanophosphors (Nguyen et al. 2017).

One limitation of all of these approaches is the need to dissociate cells into a suspension prior to trapping or encapsulation, which can be difficult for some tissues and destroys any spatial information in that tissue. Spatial 'omics approaches have begun to address this issue by enabling sequencing or detection of many nucleic acid targets in fixed cells *in situ*. Seq-FISH, mer-FISH, and related methods use sequential hybridization of carefully designed fluorescent probes to identify single RNA molecules from hundreds to thousands of pre-specified genes *in situ* (K. H. Chen et al. 2015, Lubeck et al. 2014). Slide-seq uses an array of barcoded beads with known coordinates on a glass slide to capture RNA molecules that diffuse out of a fine tissue slice, followed by multiplexed RNA sequencing, providing near-single-cell resolution (Rodriques et al. 2019). DBiT-seq deterministically patterns an array of barcodes onto a fixed tissue using a microfluidic device (Yang Liu et al. 2020). Lastly, *in situ de novo* RNA and DNA sequencing approaches have been developed, which require no pre-specification of target regions (J. H. Lee et al. 2015, Payne et al. 2021).

Fabrication and operation of microfluidic devices

Microfluidic devices allow for exquisite control of sub-microliter volumes of liquid and can exploit the efficiency gains possible when scaling laws favor small masses and volumes (reviewed by Streets & Huang 2013). By using small volumes, microfluidic devices enable orders of magnitude in cost savings, especially when using expensive biochemical reagents like enzymes. These cost savings, and the density of features that can be engineered into these small devices, can enable massively multiplexed measurements across thousands of reaction conditions in nanoliter-scale chambers in a single experiment (as a recent example, Aditham et al. 2021). Small reaction volumes also allow for rapid heating and cooling of reactions, rapid diffusion of solutes into or out of perfusion lines, and high effective concentrations of biomolecules or immobilized reagents. Microfluidic devices often take advantage of low Reynold's number conditions inside microfluidic channels, where viscous forces overtake inertial forces, enabling phenomena like laminar flow and inertial focusing (Squires & Quake 2005, Streets & Huang 2013).

A popular material for making microfluidic devices is polydimethylsiloxane (PDMS), a silicone polymer that is clear, flexible, gas-permeable, biocompatible, and hydrophobic. Because PDMS is gas-permeable, microchannels can be "dead-end filled" with liquids, meaning a closed chamber can be filled with a liquid by injecting it at high pressure, causing the gas already inside to diffuse through the PDMS and be displaced by the liquid (Hansen et al. 2002). PDMS can also be bonded to itself to form multilayer devices, and this allows for the creation of elastomeric valves (**Figure 1.2**; Unger et al. 2000). In an elastomeric valve device, two layers of microchannels are carefully aligned and bonded together. One layer is very thin, only ~30 microns taller than the tallest microchannel, and the other is thick, typically millimeters tall. Where channels overlap between the layers, they are separated only by a ~30 micron thick membrane (the difference between the thin layer total height and the thin layer channel height). If one channel is pressurized, the flexible membrane will deform into the other channel, closing it off like a valve in a plumbing system. Typically, one layer is used exclusively for fluid flow (the "flow layer"), and one is used exclusively for pressurizing the elastomeric valves (the "control layer"). In a low Reynold's number regime, elastomeric valves are analogous to transistors in electronic circuits, and they can be used to separate reagents, redirect fluid flow, trap cells/particles, create peristaltic pumps, and combine liquids in complex combinations (Streets & Huang 2013, Thorsen et al. 2002). Valve states are usually controlled by electronically actuated pneumatic valves in the air lines that pressurize each control layer microchannel.

PDMS is purchased as two separate liquid components, a base and a crosslinker, which, when mixed together, cure into a solid matrix in a concentration- and temperature-dependent manner. PDMS devices can be rapidly prototyped and fabricated in a process called soft lithography (Xia & Whitesides 1998), in which PDMS is cured on top of a solid mold, casting submicron features from the mold into the PDMS. The PDMS is typically peeled from the mold, holes are punched to provide access to any microchannels from the opposite side, and the feature-containing side is bonded to glass or another layer of PDMS to seal off microchannels. Molds are most often fabricated using photolithographic techniques borrowed from microchip manufacturing methods on silicon wafers. In a typical workflow, a two-dimensional pattern is first printed on transparency paper with a high-resolution printer, or etched into a thin chrome layer on glass, creating a mask. This mask is placed on top of a UV-activated epoxy resin, called photoresist, which has been spun and baked onto a flat silicon wafer to achieve a precise thickness. This assembly is exposed to an exact dose of collimated UV light then baked again, washed in a developer solution that removes unexposed resin, and baked a final time. The result is that the printed pattern has been transferred onto the silicon wafer as a set of permanently bonded epoxy features with a uniform height, usually tens of microns tall. More sophisticated versions of this protocol can produce features of different heights on the same mold, or allow for the creation of rounded or spherical features, which is critical for the creation of devices with elastomeric valves (Unger et al. 2000).

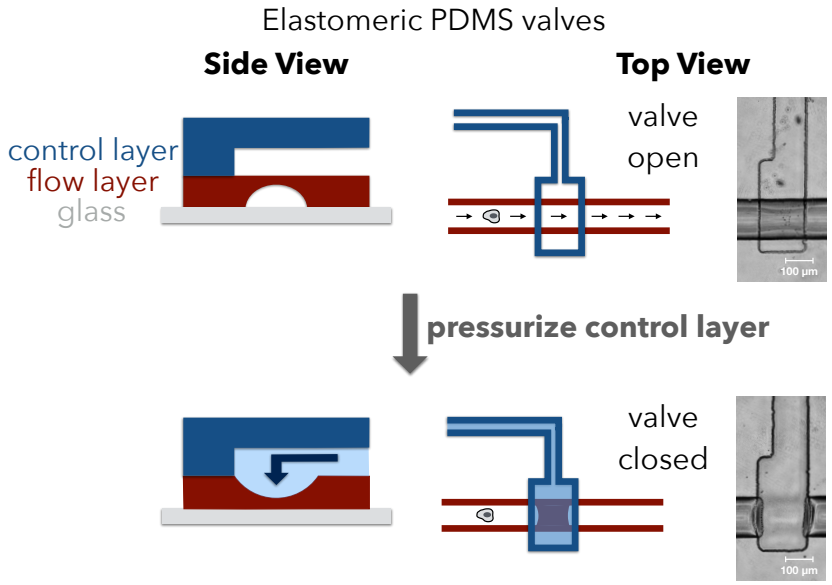


Figure 1.2. Schematic of elastomeric valve operation
Elastomeric valves can be created at the junctions of channels on overlapping layers of PDMS. Illustrated here is a "push-down" valve that is actuated by pressurizing a channel in the upper "control" layer, causing the thin PDMS membrane separating these two channels to deform and close off the "flow" layer channel below.

DNA Sequencing Technologies

Illumina

Illumina's sequencing-by-synthesis approach became dominant among peer "next generation sequencing" technologies beginning in the 2000's (Bentley 2006). In Illumina's sequencing approach, DNA fragments are first ligated to an adapter sequence, which is designed to hybridize onto immobilized DNA oligos patterned inside nanowells in a microfluidic flowcell. Once a DNA molecule diffuses into a nanowell and hybridizes onto the complementary adapter sequence, it is rapidly copied by an isothermal process of bridge amplification. The reaction kinetics favor the copying of only a single input DNA molecule per nanowell, ultimately resulting in many copies of the same single-stranded DNA molecule in one tight cluster. After amplification, a primer and a mix of DNA bases conjugated to distinct fluorophores are added in and ligated to a growing complementary strand on each molecule, but the chemistry guarantees only one base is added at a time. The flowcell is imaged, and the fluorescent signal present in each nanowell provides a readout of which base was incorporated into that growing molecule. The fluorophore is then cleaved off, and the cycle is repeated, one base at a time, until a fixed maximum number of cycles is reached, usually up to 250 bp with the latest chemistries. The whole process is then repeated using a different primer, to read from the opposite end of each molecule. The result is up to billions of paired 250 bp sequencing reads from each flow cell, typically with 99.9% base calling accuracy, at a basic marginal sequencing cost of ~$10-$35 per gigabase (Logsdon et al. 2020).

Pacific Biosciences Single Molecule, Real-Time Sequencing

Pacific Biosciences (PacBio) sequencing also sequences DNA by imaging the incorporation of fluorescently tagged nucleotides during DNA synthesis. However, this incorporation happens continuously in real time rather than in discrete, global hybridization and washing steps as with Illumina sequencing, allowing for much longer sequencing reads to be produced. There is also no amplification step; rather, the activity of a single polymerase enzyme copying a single DNA molecule is recorded in a movie, which is then decoded to produce a single sequencing read. To achieve the ability to rapidly detect photons from single fluorophores, the polymerase is immobilized above the holes of a zero-mode waveguide, a thin metal film with holes smaller than the wavelength of light being emitted, similar to the screen on the door of a microwave oven. This zero-mode waveguide only allows light from a tiny volume above each hole to pass through to the camera, effectively removing any background fluorescence from unincorporated nucleotides (Levene et al. 2003). Although the accuracy of a single basecall is only around 85-92% (Logsdon et al. 2020), the same ~20 kb DNA molecule can be circularized and read over and over, producing a

consensus sequence with median 99.8% accuracy, which is marketed as HiFi Sequencing (Wenger et al. 2019). The cost for HiFi sequencing is currently in the range $43-$86 per gigabase (Logsdon et al. 2020).

One additional benefit of PacBio's method is that information is not only derived from the fluorescent signals observed in each movie, but from the time intervals between fluorescent signals. Specifically, methylated adenine bases take significantly longer to pass through the polymerase during synthesis, and the presence of methyladenines can be called with over 90% accuracy in most sequence contexts (Flusberg et al. 2010, McIntyre et al. 2019). Methylcytosines produce a much weaker signal, but they can be called indirectly if first chemically converted to thymines (Yibin Liu et al. 2020)

Oxford Nanopore Technologies

Oxford Nanopore Technologies (ONT) has developed an altogether different sequencing approach, which reads out base sequences according to their impedance of ion flow as a single DNA strand is pushed through a tiny protein pore in a thin membrane (M. Jain et al. 2016). The raw data are recordings of picoamp currents over time for thousands of individual pores per flowcell, which are decoded by a deep learning algorithm into a sequence of bases. While the basecalling accuracy per read is only 87-98% with standard basecallers (Logsdon et al. 2020), this figure is substantially improving over time. The key advantage of nanopore sequencing is that reads up to several megabases long can be produced, with read length primarily determined by DNA library preparation methods. ONT sequencing hardware can be as small as a handheld device, in the case of the small-scale MinION device intended for individual use, or as large as a common benchtop centrifuge, in the case of the large-scale PromethION device intended for sequencing service provider use. This allows for a democratization of DNA sequencing in a way not possible before, as Illumina and PacBio sequencers can be the size of washing machines and involve massive up-front capital costs. Moreover, this technology allows for readout of both cytosine and adenine DNA base modifications (Yibin Liu et al. 2020, Rand et al. 2017, Simpson et al. 2017). ONT sequencing on a PromethION can cost as little as $21-$42 per gigabase (Logsdon et al. 2020).

The challenges of assembling and mapping to repetitive DNA

In 2001, the Human Genome Project announced that it had completed its draft reference sequence, but with an important caveat: due to technical limitations, it excluded highly repetitive sequences comprising roughly 5-10% of the human genome, found primarily in all centromeres, telomeres, and rRNA genes (**Figure 1.3**; I. H. G. S. Consortium 2001). These regions have remained almost entirely missing from

the reference sequence since then, owing to the fairly short lengths of sequencing reads produced by Sanger and Illumina sequencing. Genomes are typically assembled by shredding the whole genome, or portions of it, into small fragments, which are sequenced and stitched back together computationally based on unique sequence overlaps, a bit like a jigsaw puzzle. When short sequencing reads originate from highly repetitive regions, it becomes impossible to find unique stretches of sequence that allow those reads to be unambiguously stitched together. To carry on the jigsaw analogy, these repetitive regions are rather like the homogeneous sky pieces in a landscape, often left for the very end after all the more easily assembled pieces with unique features have been stitched together. The puzzle of the human genome sequence has remained unfinished for 20 years.

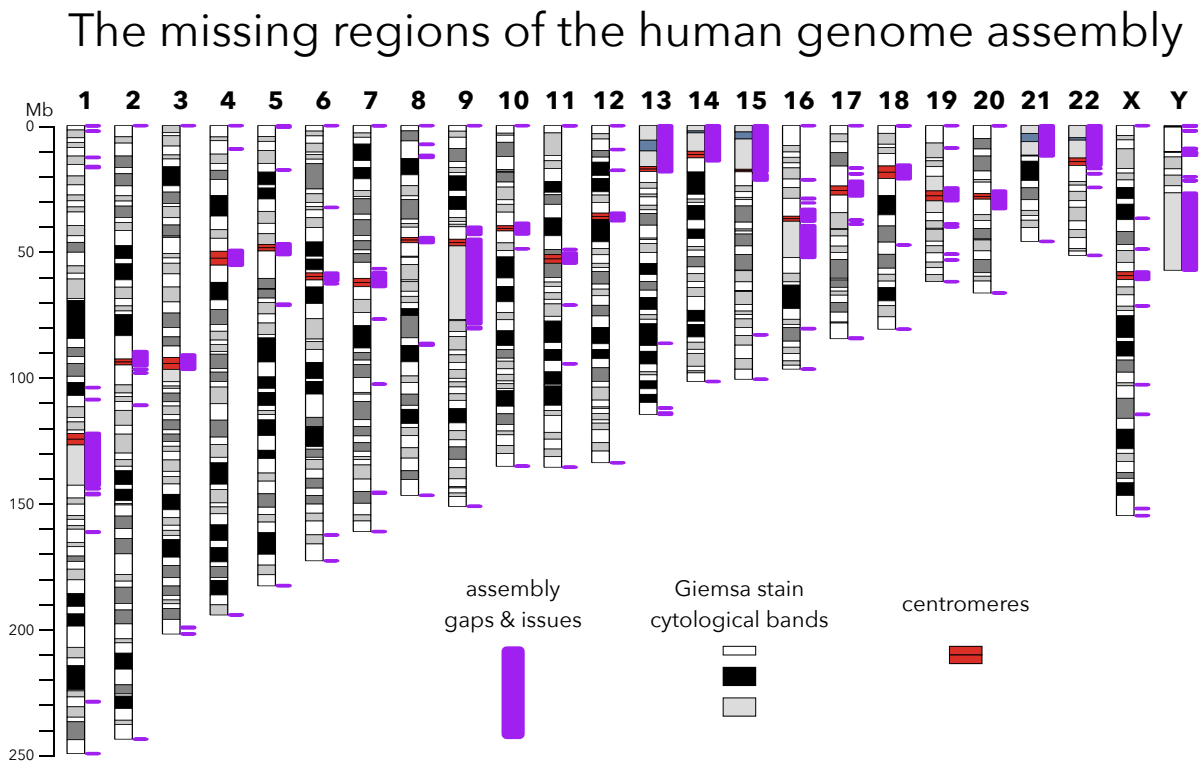


Figure 1.3. The missing regions of the hg38 human genome assembly
Chromosome ideograms show the scale and banding patterns of the 24 distinct human chromosomes. Purple bands in the adjacent track highlight the gaps and problematic areas in the latest hg38 human genome reference assembly, which are found at every centromere and telomere, on the short arms of the acrocentric chromosomes, in large pericentric heterochromatin blocks, and in various duplicated gene regions. The sequences in these gaps can be highly size variable among individuals, but on average they represent 5-10% of any person's genome.

These missing repetitive regions of the genome are known to play essential roles in chromosome segregation, nuclear architecture, and cell senescence, among others. However, without a reference sequence, studies of these regions have fallen behind the rapid advances of the genomics era, and many fundamental questions remain. Just this past year, the Telomere to Telomere (T2T) Consortium completed the first linear assembly of an entire human chromosome, repeats and all, owing to advances in long-read sequencing technologies (Miga et al. 2020). Two complete chromosomes (chrX and chr8) have been released so far (Logsdon et al. 2021, Miga et al. 2020), with the remainder of the genome expected to be released later this year. With this emerging reference, the formerly missing repetitive regions of the genome are ripe for investigation of their regulation, function, and evolution, but such studies also demand novel technological approaches.

Even with a complete linear reference sequence across the most repetitive regions of the genome, it remains challenging to map short high-throughput sequencing reads from existing protein-DNA mapping approaches, such as ChIP-seq or CUT&RUN reads, unambiguously within these regions. Occasionally a mutation within a repetitive region will create a unique marker that can anchor any overlapping reads to that exact site, but these can be rare, sometimes separated by tens of thousands of bases (**Figure 1.4**). It's also possible to identify "region-specific" markers that occur multiple times but with all instances contained in the same region of the genome. Both of these short-read mapping approaches within repetitive regions provide extremely limited resolution of individual protein-DNA interactions. This low resolution can be acceptable for some proteins with enormous DNA-binding footprints, but it remains challenging to achieve the resolution required to map the binding sites of rare histone variants or transcription factor-like proteins.

The formerly unassembled regions of the human genome

Centromeric alpha satellite

The largest component of the formerly missing regions of the genome is centromeric alpha satellite DNA, which occurs as multi-megabase arrays of 171 bp sequences repeated in tandem and organized into higher-order repeating units (Schueler et al. 2001). One or more large alpha satellite arrays occur in the centromere regions of every chromosome, near the middle of the chromosome for all chromosomes except the acrocentrics (13, 14, 15, 21, 22), which have a very short arm and a long arm. Each chromosome's alpha satellite sequence can be distinguished from other chromosomes by its sequence composition and repeat organization (McNulty & Sullivan 2018). Altogether, alpha satellites constitute about 3.1% of the genome,

although this varies from 1-5% among individuals due to large alpha satellite structural variations within populations (Miga 2019).

Roughly 35% of each alpha satellite array contains some amount of centromere protein A (CENP-A), a histone 3 variant that plays an essential role in the inner kinetochore, but the density of this protein throughout the centromere remains under dispute (Bodor et al. 2014, Sullivan et al. 2011). The CENP-A-containing regions of each alpha satellite array are also characterized by H3 nucleosomes with marks of open chromatin, such as H3K4me2, while chromatin outside this region is characterized by the constitutive heterochromatin mark H3K9me3 (McNulty & Sullivan 2018). Scores of inner and outer kinetochore proteins must assemble at the centromere every cell cycle. Proper kinetochore assembly is essential for proper chromosome segregation in mitosis and meiosis, and failure of this process can lead to cancer and birth defects.

Because long Oxford Nanopore sequencing reads also include information about endogenous CpG methylation, Miga et al. were able to investigate the DNA methylation landscape within the centromere of the telomere-to-telomere assembly of chrX (Miga et al. 2020). In doing so, they unexpectedly found a ~60 kb region of CpG hypomethylation within the repetitive sequences that constitute the centromere, and they suggest that it may overlap the region where a high density of inner kinetochore proteins bind to DNA; this pattern was also seen on chr8 (Logsdon et al. 2021, Miga et al. 2020). This highlights a need for the ability to map inner kinetochore proteins at high resolution to investigate how centromeres are epigenetically specified and inherited.

Pericentric satellite DNA, segmental duplications, ribosomal DNA loci, & telomeres
Apart from the alpha satellite sequences that encompass functional human centromeres, other families of tandemly repeated DNA are found in large arrays near centromeres, on the short arms of the acrocentric chromosomes, and on the long arm of the Y chromosome. The largest of these are Human Satellites 2 and 3 (HSat2 and HSat3), which occur in enormous arrays on chromosomes 1, 9, 16, Y, and the acrocentrics and altogether constitute about 2.1% (range 1-7% due to variation) of the human genome (Miga 2019). I studied the genomic organization of HSat2,3 extensively in the past, providing the first comprehensive catalog of these sequences in a human genome, along with tools to study their variation in populations; I showed that a large HSat3 array on chrY can vary in size from 7 to 98 Mb among a sample of 396 males (Altemose et al. 2014). The potential function of HSat2,3 remains elusive, although they have been shown to be transcribed in response to cellular stress and in certain cancers and senescent cells (Bersani et al. 2015, Landers et al. 2021, Swanson et al. 2013, Zhu

et al. 2018). It has been suggested that pericentromeric satellites in other organisms play a role in facilitating inter-chromosomal interactions to keep chromatin in one cohesive unit in the nucleus (Jagannathan et al. 2018). Other poorly characterized human pericentromeric satellite families include beta satellite, gamma satellite, Human Satellite 1, and others. Pericentromeric regions also tend to contain large stretches of segmental duplications, recent low-copy duplications of non-satellite sequences from other regions of the genome, which can include genes (Emanuel & Shaikh 2001). The ends of chromosomes, called the telomeres, contain simple repeated DNA sequences and play critical roles in nuclear organization, senescence, and cancer (Shay & Wright 2019).

Lastly, the human reference sequence has excluded ribosomal DNA (rDNA) loci, found on the short arms of the acrocentric chromosomes. These regions contain hundreds of tandem near-identical 43 kb repeats, each including a copy of the genes encoding the 47S rRNA component of the ribosome and a spacer region responsible for organizing the nucleolus (Salim & Gerton 2019). While nearly 60% of cellular transcription is devoted to rRNA production, only a subset of rDNA genes are active in any given cell, and large structural variations can occur between individuals and within cell populations (Salim & Gerton 2019). Studying the epigenetics of these formerly nebulous regions of the genome will benefit enormously from new T2T references as well as new technologies for mapping protein-DNA interactions in these regions.

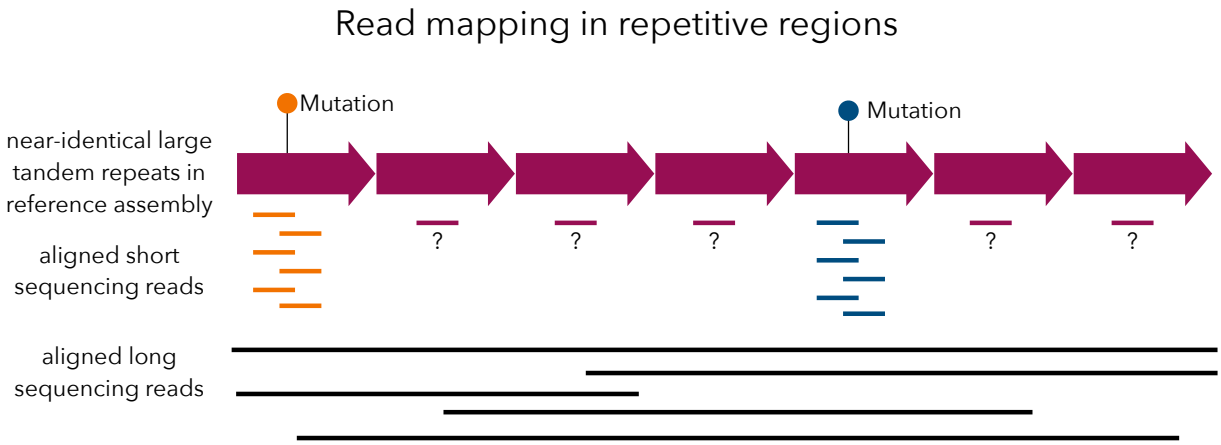


Figure 1.4. The challenges of mapping short reads in highly repetitive regions
This illustrates how short reads can only be mapped uniquely within repetitive regions when they overlap a unique sequence marker, while the intervening regions can only be mapped ambiguously. Long sequencing reads can bridge between these markers.

Measuring DNA accessibility with methyltransferases

In most conventional DamID studies, an unfused Dam-only control is used to account for regional differences in DNA accessibility, GATC frequency, and sequencing bias across the genome (Vogel et al. 2007). This Dam-only DamID approach can also be used on its own to investigate DNA accessibility *in vivo* (Aughey et al. 2018, Singh & Klar 1992, Wines et al. 1996) or *in situ* (Schaik et al. 2020). DNA accessibility has also been measured *in situ* using the GpC cytosine methyltransferase M.CviPI in conjunction with whole-genome bisulfite sequencing, which maps both endogenous CpG and exogenous GpC methylcytosines using short-read high-throughput sequencing (Kelly et al. 2012).

Recently, five similar methods have been published that use DNA methylases to methylate accessible regions of the genome *in situ*, followed by direct readout of DNA methylation using Oxford Nanopore or Pacific Biosciences long-read sequencing (Abdulhay et al. 2020, I. Lee et al. 2020, Shipony et al. 2020, Stergachis et al. 2020, Y. Wang et al. 2019). In the meSMLR method, permeabilized yeast cells were treated with M.CviPI then sequenced using Oxford Nanopore sequencing (Y. Wang et al. 2019). The nanoNOME approach is similar but also involves jointly measuring endogenous CpG methylation in human cells (I. Lee et al. 2020). SMAC-seq expanded on this and added two additional methyltransferases to improve resolution in yeast genomes, which lack endogenous methylation: EcoGII, which methylates adenines nonspecifically, and M.SssI, which methylates CpG sites. SAMOSA-seq used EcoGII to methylate adenines in reconstituted chromatin and in human cells, but it read out the methylation using PacBio sequencing instead of ONT (Shipony et al. 2020). Finally, Fiber-seq identified and used a different non-specific adenine methyltransferase, Hia5, and similarly read out adenine methylation using PacBio sequencing in human cells (Stergachis et al. 2020). Because they produce long reads, these methods can be used to investigate DNA accessibility in newly assembled repetitive regions of the human genome.

Lamina Associated Domains (LADs)

The nuclear lamina is a tangle of filamentous proteins, including Lamins A/C, B1, and B2, that lines the inside of the nuclear envelope and helps to maintain nuclear integrity. Lamina Associated Domains (LADs) are large segments of DNA (range 10 kb – 10 Mb, median 500 kb) that associate with the nuclear lamina in some or all cells (reviewed by van Steensel & Belmont 2017), comprising up to 30% of the genome in human cells (Guelen et al. 2008). The nuclear lamina and LADs serve both a structural function, underpinning the three-dimensional architecture of the genome in the nucleus, and a regulatory function, as LADs tend to be gene-poor, more heterochromatic, associated

with the "B compartment," and transcriptionally less active (reviewed by Buchwalter et al. 2019, Karoutas & Akhtar 2021, van Steensel & Belmont 2017). However, this regulatory role is complex, as programmed localization of DNA to the lamina is not sufficient to silence reporter genes (Kumaran & Spector 2008). Furthermore, this canonical nuclear organization, with heterochromatin at the periphery, can be inverted in some cell types, like rod photoreceptor cells (Solovei et al. 2009).

The establishment and inheritance of LADs through generations and cell divisions is closely tied to histone methylation, especially H3K9me2 and H3K9me3 marks (reviewed by Hoskins et al. 2021). However, histone modifications are not sufficient to cause lamina association of particular region of DNA (Karoutas & Akhtar 2021). It has been suggested that some LADs stochastically reshuffle between the nuclear lamina and nucleolus-associated heterochromatin every cell cycle (Kind et al. 2013). Additionally, some regions, often at the boundaries of LADs and the so-called "inter-LADs" (iLADs) that surround them, show variable contact with the nuclear lamina in single cells (Kind et al. 2015).

Because certain lamin proteins like Lamin B1 are found almost exclusively in the lamina, mapping their interactions with DNA can reveal how DNA is organized in the nucleus. LADs are frequently mapped using DamID with the LaminB1 protein (Borsos et al. 2019, Guelen et al. 2008, Kind et al. 2015, Lenain et al. 2017). These maps allow for comparison of nuclear organization between cell types (Lenain et al. 2017). While some regions are found associated with the nuclear lamina to some degree in all cell types (constitutive LADs, cLADs), others are associated in only some cell types (facultative LADs, fLADs), or in no cell types (constitutive inter-LADs, ciLADs).

Lamin proteins are critical for nuclear integrity and proper genome regulation, and their mutation or misregulation can lead to devastating diseases classed as laminopathies, including Hutchinson-Gilford progeria syndrome (Eriksson et al. 2003). Changes in lamin expression and laminar integrity are hallmarks of cellular senescence and aging (Freund et al. 2012). Disruptions of the nuclear lamina also contribute to changes in nuclear shape and size, which are commonly used as diagnostics for various cancers (Wolberg et al. 1999) and have been directly linked to malignancy and tumor progression (Bell & Lammerding 2016).

Dissertation Overview

This dissertation is broadly organized into two themes: developing methods for mapping protein-DNA interactions in single cells, and developing methods for mapping protein-DNA interactions in highly repetitive regions in bulk cell populations. Motivated by the need to pair imaging and sequencing information in single cells, I engineered an integrated microfluidic device that enables single-cell isolation, imaging, selection, and DamID processing, which I call "μDamID," for microfluidic DamID (now published in Altemose et al. 2020). Here I describe the theoretical and practical considerations that went into the design and implementation of this technology, as well as its benchmarking and potential applications. I applied this device to image and map lamina-associated domains in a transiently transfected human cell line co-expressing ^{m6}A-Tracer, and I validated these measurements against bulk DamID data from the same cell line as well as other human cell lines (Kind et al. 2015, Lenain et al. 2017). Then, I describe the ongoing development of a new method for mapping protein-DNA interactions using long, single-molecule reads, which we call DiMeLo-seq, for Directed Methylation with Long-read sequencing. I describe the development of an optimization pipeline for this method, an optimized protocol, and its application to new protein targets with an outlook to high-resolution mapping proteins in human centromeres. Together these methods advance our ability to map protein-DNA interactions in some of the most challenging settings.

Glossary of key terms and abbreviations

Dam: a DNA adenine methyltransferase that specifically methylates GATC sites

EcoGII: a non-specific DNA adenine methyltransferase

Hia5: a non-specific DNA adenine methyltransferase (with preference for A surrounded by G, C, or T)

M.CviPI: a GpC cytosine methyltransferase

M.SssI: a bacterial CpG cytosine methyltransferase

SAM: S-adenosylmethionine, a methyl donor substrate used by many methyltransferase enzymes

DpnI: a restriction enzyme that exclusively cleaves G^{m6}ATC in half, leaving blunt ends; cleaves hemimethylated sites much less efficiently than fully methylated sites

DpnII: a restriction enzyme that exclusively cleaves fully unmethylated GATC sites

^{m6}A-Tracer: a truncation of the DpnI protein, which binds exclusively to fully methylated G^{m6}ATC sites but does not cleave them; fluorescently tagged for imaging protein-DNA interaction sites

LMNB1: LaminB1, an intermediate filament protein found almost exclusively at the nuclear lamina

NES: nuclear export signal; a small peptide sequence bound by nuclear exportin proteins, which actively transport the NES-containing protein out of the nucleus

PCR: Polymerase Chain Reaction, a method for amplifying DNA that requires repeated thermal cycling

PDMS: polydimethylsiloxane, a flexible, clear, hydrophobic, gas-permeable silicone polymer used to make microfluidic devices

μDamID: a microfluidic device developed here, enabling the joint imaging and sequencing of protein-DNA interactions in single cells

DiMeLo-seq: **Di**rected **Me**thylation with **Lo**ng-read **seq**uencing, a protein-DNA mapping method developed here that uses long-read sequencing to directly read out methyl marks deposited by methyltransferases near a target protein's binding sites

pA or pAG: protein A or protein A/G; proteins that bind to IgG antibodies from common host organisms; pA binds best to rabbit IgG; pAG binds equally well to rabbit and mouse IgG

LAD: Lamina Associated Domain, a large region of DNA (median 0.5 Mb) associated with the nuclear lamina at the periphery of the nucleus; usually gene poor and transcriptionally quiet

iLAD: an "inter-LAD", i.e. the opposite of an LAD—a region occurring between two LADs that is not associated with the nuclear lamina

cLAD: a constitutive LAD, meaning it appears to associate with the nuclear lamina across all cell types (though not necessarily in every single cell of a given type)

fLAD: a facultative LAD, meaning it associates with the lamina in only a subset of cell types

ciLAD: a constitutive iLAD, meaning it never appears to associate with the nuclear lamina in any cell type

vLAD: a variable LAD; a term we use to describe LADs that only associate with the lamina in a subset of single cells in the same population; distinct from fLAD, which is a bulk cell property involving comparisons across cell types

Chapter 2

Design, fabrication, and operation of μDamID, a microfluidic device for imaging and sequencing protein-DNA interactions in single cells

Aims & overview

I set out to create a device that would allow us to combine imaging and sequencing measurements of protein-DNA interactions in single cells. After surveying existing single-cell protein-DNA mapping methods available at the time (Kind et al. 2015, Rotem et al. 2015), it became clear that single-cell DamID, which could be done in a one-pot reaction and recover hundreds of thousands of DNA fragments per cell (Kind et al. 2015), and which could be combined with live-cell ^{m6}A-Tracer imaging (Kind et al. 2013), would be best suited to porting onto a valve-based microfluidic device. One major concern was that the DamID method is limited in resolution and sensitivity by the uneven distribution of GATC sites throughout the genome. Before proceeding with device design, I performed computational simulations to estimate the theoretical maximum coverage of the genome with this method, and I began testing and optimizing cell transfection and ^{m6}A-Tracer imaging procedures. Once satisfied with the theoretical and practical performance of DamID and ^{m6}A-Tracer imaging, I proceeded to adapt a microfluidic device for carrying out single-cell DamID on chip, making key design optimizations critical to the final performance of the device. I further built up the fabrication infrastructure and protocols needed to fabricate the device with high reliability. Additionally, I designed, assembled, tested, and optimized the pneumatic and thermal control hardware needed to operate the device. In this chapter, I describe these preliminary steps and the final protocols for fabrication and operation of the device, highlighting common pitfalls and their solutions.

Simulated digestion at GATC sites

To examine some of the fundamental limitations of single-cell DamID for probing protein-DNA interactions across the genome, I performed several computational analyses to simulate elements of the scDamID protocol. DamID involves digesting the genome with the DpnI restriction enzyme, which cuts at adenine methylated GATC sites. After DpnI digestion, adapters are ligated to the resulting fragments, and DNA is

amplified by PCR, which favors shorter DNA fragments. Furthermore, Illumina library preparation involves a cleanup step that depletes very short fragments under ~150 bp, and cluster generation on most Illumina flow cells strongly disfavors fragments longer than ~750 bp (determined empirically from published sequencing data from Kind et al. 2015). Thus, scDamID can only probe regions of the genome producing fragments short enough to be PCR amplified and sequenced (~150-750 bp). However, GATC sites occur with an uneven distribution across the genome, owing to differences in regional GC content and repetitive sequences, and in some regions sequenceable fragments cannot be produced.

To survey where the "unsequenceable" regions of the genome are with respect to DamID, I simulated digestion at every GATC site in the reference sequence. While I initially used the incomplete hg38 reference for this analysis, I have now updated it using the Telomere-to-Telomere Consortium's complete chm13 v1.0 draft reference sequence. First, I quantified the proportion of the genome contained in fragments in the sequenceable range. **Figure 2.1a** shows the cumulative distribution of the proportion of the genome contained in GATC digestion fragments of increasing size; I found that 47.4% of the genome is contained in fragments in the sequenceable range. However, this should not be interpreted to suggest that only 47.4% of the genome can be probed by DamID. The Dam enzyme is reported to have a reach of up to 5 kb from a point binding site (Steensel & Henikoff 2000), so I sought to determine what fraction of the genome has a sequenceable fragment within reach, or in other words, what fraction of 5-kb bins overlap a sequenceable fragment. **Figure 2.1b** shows this proportion for bins of different sizes (1, 5, 10, 100 kb), revealing that 95.8% of 5-kb bins overlap a sequenceable fragment. To examine where the unsequenceable 4.2% might be, I plotted the relative coverage of 150-750 bp fragments across the chm13 reference (**Figure 2.1c**), showing that the most of the GATC-depleted regions occur in highly repetitive regions like centromeres and pericentric heterochromatin. Apart from these regions, the remainder of the genome appears to be sequenceable with scDamID, although this analysis has underlined the limitations of resolution possible with any method relying on DpnI digestion.

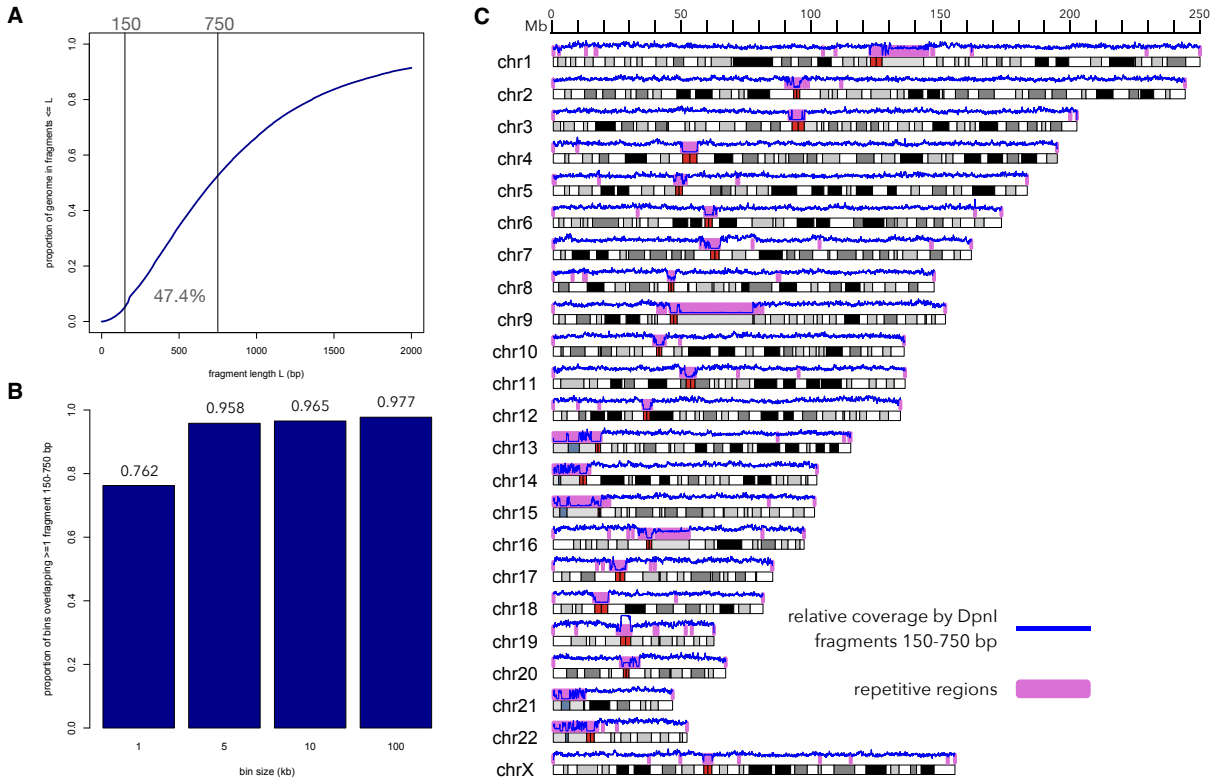


Figure 2.1. Simulated DpnI digestion of the T2T chm13 reference assembly
(**A**) the cumulative proportion of bases in the genome contained in simulated DpnI digest fragments of increasing length. 47.4% of bases exist in fragments in the Illumina-sequenceable range 150-750 bp. (**B**) The proportion of fixed non-overlapping bins, of various sizes, that overlap at least one fragment of sequenceable length. (**C**) A plot of relative coverage by sequenceable fragment lengths across each chromosome. GATC sites are often rare in repetitive regions, resulting in large, unsequenceable fragments from these regions.

Cloning, tissue culture, transfection, and imaging

HEK293T cells were chosen for their ease of growth, transfection, suspension, and isolation (CRL-3216, ATCC, Manassas, VA; validated by microsatellite typing, at passage number <10). I opted to express Dam proteins by transient transfection and optimized transfection conditions by varying concentrations of DNA and transfection reagent. Ultimately, cells were seeded in 24-well plates at 50000 cells per well in 0.5 ml media (DMEM plus 10% FBS). The next day, cells were transfected using FuGene HD transfection reagent according to their standard protocol for HEK293 cells (Promega, Madison, WI). DNA plasmids were cloned in Dam-negative *E. coli* to reduce sequencing reads originating from plasmid. Dam-LMNB1 and ^{m6}A-Tracer plasmids

were obtained from Bas van Steensel (from Kind et al. 2013); Dam-LMNB1 was modified to replace GFP with mCherry and to produce a Dam-only version, as well as to create a Dam-tdTomato-LMNB1 fusion for later experiments; their sequences are available as supplementary information (see https://github.com/altemose/microDamID). 250 ng Dam construct DNA plus 250 ng ^{m6}A-Tracer DNA were used per well. As controls to validate transfection, additional wells were left untransfected, transfected with ^{m6}A-Tracer only, or transfected with Dam construct only. The following day, successful transfection was validated by widefield fluorescence microscopy, seeing GFP signal in wells containing ^{m6}A-Tracer, and mCherry signal in all wells containing Dam construct only. Expression was highest at 72 hours post transfection, which is when cells were harvested. 20 hours before harvesting, the media was replaced and 0.5 µl Shield-1 ligand (0.5 mM stock, Takara Bio USA, Inc., Mountain View, CA; final concentration 0.5 µM) was added to each well to stabilize protein expression. I found that the use of polystyrene, not polypropylene, tissue culture plastic was critical for efficient transfection with FuGene HD, perhaps due to adhesion of the transfection reagent to polypropylene.

Fluorescence confocal imaging of cells was performed using an inverted scanning confocal microscope with a 488 nm Ar/Kr laser (Leica, Germany) for excitation, with a bandpass filter capturing backscattered light from 500-540 nm at the primary photomultiplier tube (PMT), with the pinhole set to 1 Airy unit, with a transmission PMT capturing widefield unfiltered forward-scattered light, and with a 63X 0.7 NA long-working-distance air objective with a correction collar, zoomed by scanning 4X. For later imaging, a 63X 1.2 NA water immersion objective was used, with a 6X scanning zoom. The focal plane was positioned in the middle of each nucleus, capturing the largest-circumference cross-section, and final images were averaged over 10 frames to remove noise. Confocal images confirmed the expected "ring-like" structures in a subset of cells expressing both Dam-LMNB1 and ^{m6}A-Tracer (**Figure 2.2**), consistent with proper DNA methylation. The presence of ^{m6}A was also validated by mass spectrometry (**Table 2.1**), with samples prepared as described in (Kriaucionis & Heintz 2009, Quinlivan & Gregory 2008).

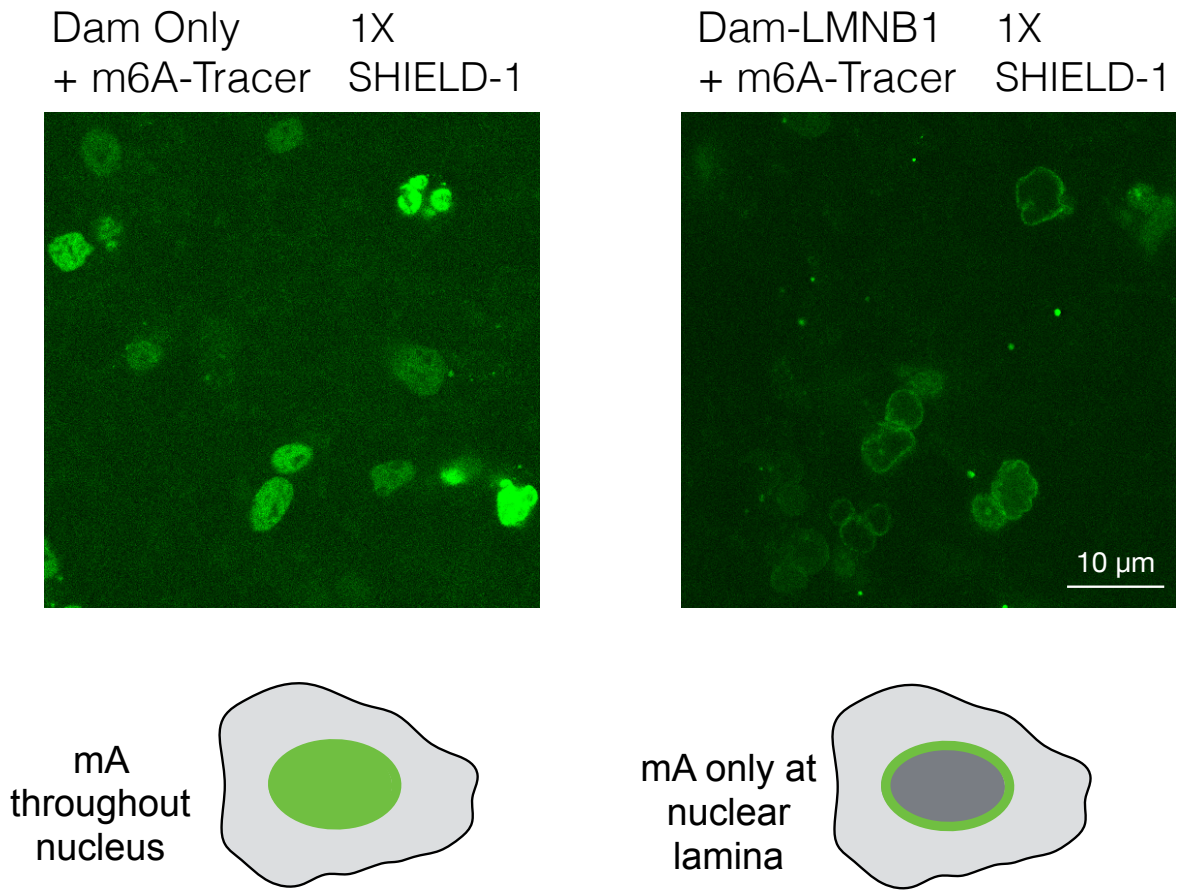


Figure 2.2. Tracking DNA methylation with ^{m6}A-Tracer in live cells
HEK293T cells were transiently transfected with constitutively expressed ^{m6}A-Tracer and shield-1-inducible Dam or Dam-LMNB1 then imaged 15 hours after induction with a scanning confocal microscope. Characteristic laminar rings indicate that DNA adenine methylation has been properly targeted to the nuclear lamina.

Sample	Methyl-deoxycytidine (% of all C)	Methyl-deoxyadenosine (% of all A)
1) DAM-LaminB1, induced (expect mA)	6.7%	**0.9%**
2) DAM-LaminB1, not induced (expect mA only if expression is leaky)	6.3%	**0.7%**
3) Dam only, induced (expect mA)	7.3%	**1.9%**
4) Dam only, not induced (expect mA only if expression is leaky)	7.0%	**1.3%**
5) Untransfected, with induction drug (expect no mA)	13.2%	<LOD
6) Untransfected, no induction drug (expect no mA)	10.2%	<LOD
7) no DNA mock control (expect no mA)	<LOD	<LOD

Table 2.1. ^{m6}**A mass spectrometry results**
DNA was harvested from untransfected cells or from transiently transfected cells with or without induction of Dam or Dam-LMNB1 expression, then digested to nucleosides (Kriaucionis & Heintz 2009, Quinlivan & Gregory 2008) and characterized by liquid chromatography mass spectrometry. This confirmed the presence of ^{m6}A following Dam or Dam-LMNB1 transfection, and an increase in induced relative to uninduced cells. However, ^{m6}A was still present to a lesser degree in uninduced cells, consistent with leaky expression.

Overview of device design and operation

I designed and fabricated a polydimethylsiloxane (PDMS) microfluidic device with integrated elastomeric valves to facilitate the various reaction stages of the DamID protocol in a single liquid phase within the same device to avoid sample loss prior to DNA amplification (**Figure 2.3**). The device is compatible with high-magnification imaging on inverted microscopes, enabling imaging prior to cell lysis. Each device was designed to process 10 cells in parallel, each in an individual reaction lane fed from a common set of inlets to avoid sample cross-contamination. Valves are controlled by

pneumatic actuators operated electronically via a programmable computer interface (J. A. White & Streets 2018).

Device operation was modified from a previously published single-cell RNA sequencing platform (Streets et al. 2014). A suspension of single cells is loaded into the cell inlet (**Figure 2.3b**) and cells are directed towards a trapping region by a combination of pressure-driven flow and precise peristaltic pumping. Cells enter the device in a wide filter region where dozens can be visualized and screened at once as they move towards a narrow channel leading to the trapping regions. As a cell crosses one of the 10 trapping regions, valves are actuated to immobilize the cell for imaging (**Figure 2.4**). The cell is imaged by confocal fluorescence microscopy to visualize the localization of ^{m6}A-Tracer, and after image acquisition, the user can choose whether to select the cell for DamID processing, or to reject it and send it out the waste outlet (**Figure 2.3b**).

Selected cells are injected from the trapping region into a holding chamber using pressure-driven flow from the reagent inlet (**Figure 2.3b**, **Figure 2.4**). Once all 10 holding chambers are filled with imaged cells, processing proceeds in parallel for all 10 cells by successively adding the necessary reagents for each step of the single-cell DamID protocol (Kind et al. 2015) and dead-end filling each of the subsequent reaction chambers. Reaction temperatures are controlled by placing the device on a custom-built thermoelectric control unit for dynamic thermal cycling. Enzymes are heat inactivated between each step (Kind et al. 2015) and a low concentration of mild detergent was added to all reactions to prevent enzyme adhesion to PDMS (Streets et al. 2014). Reaction dehydration due to water vapor diffusion through PDMS was prevented by pressurizing large hydration paddles overlapping each of the large reaction chambers during heating steps.

In the first reaction stage, a buffer containing detergent and proteinase pushes the cell into the lysis chamber, where its membranes are lysed and its proteins, including ^{m6}A-Tracer, are digested away (**Figure 2.3c**). Next, a DpnI reaction mix is added to digest the genomic DNA at Dam-methylated GATC sites in the digestion chamber. Then, a mix of DamID universal adapter oligonucleotides and DNA ligase is added to the ligation chamber. Finally, a PCR mix containing primers that anneal to the universal adapters is added and all valves within the lane are opened, creating a 120 nl cyclic reaction chamber. Contents are thoroughly mixed by peristaltic pumping around the reaction ring, then PCR is carried out on-device by thermocycling. Amplified DNA is collected from each individual lane outlet, and sequencing library preparation is carried out off-device.

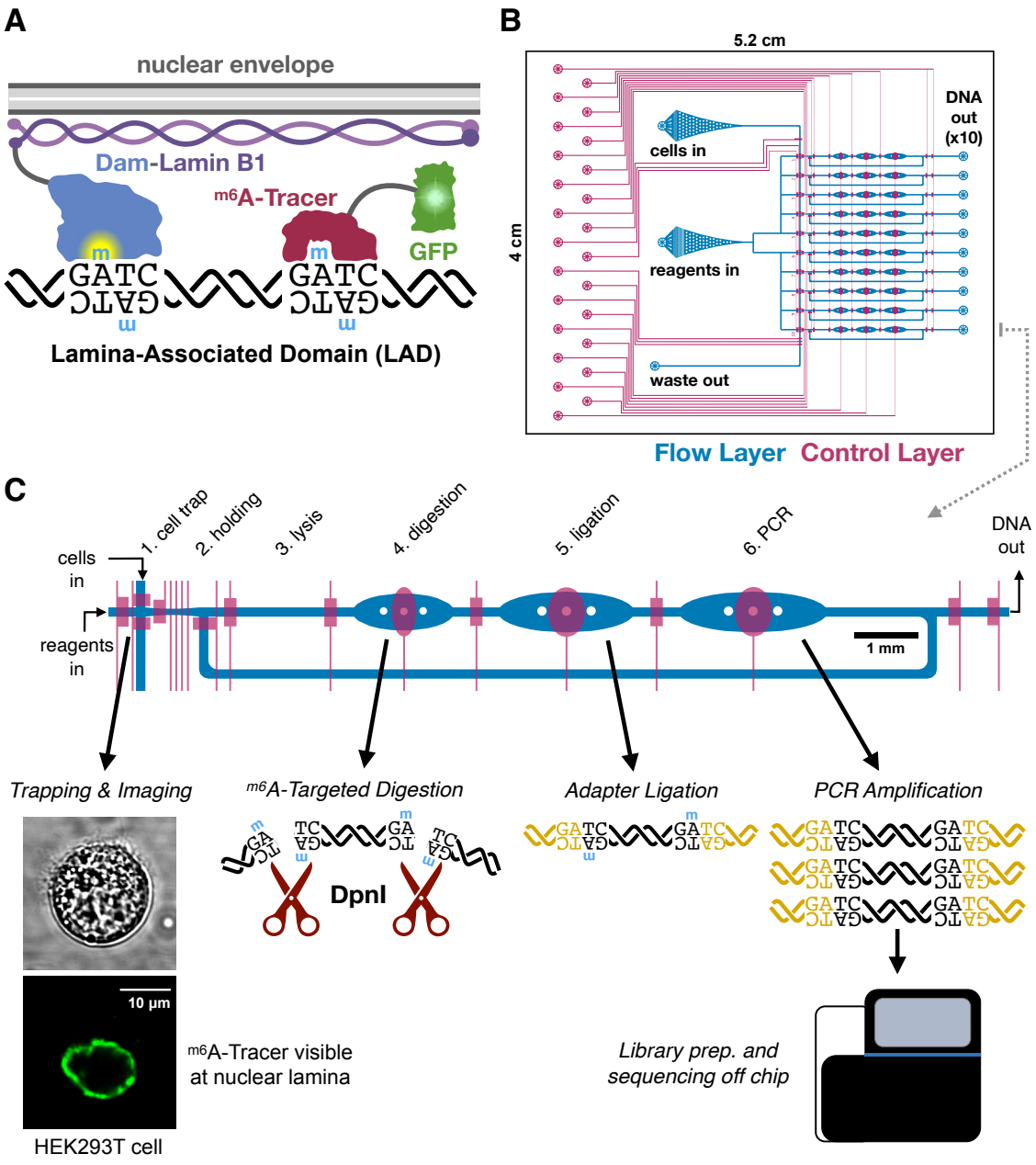


Figure 2.3. µDamID device design and function

(**A**) Overview of DamID (van Steensel & Henikoff 2000) and m6A-Tracer (Kind et al. 2013) technologies applied to study interactions between DNA and nuclear lamina proteins. (**B**) The overall design of the 10-cell device, showing the flow layer (blue, where cells and reagents enter channels) and the control layer (red, where elastomeric valves overlap the flow layer to control the flow of liquids). (**C**) A closer view of one lane explaining the DamID protocol and the function of each chamber of the device. 10 cells are trapped, imaged, and selected serially, one per lane, then all 10 cells are lysed and processed in parallel.

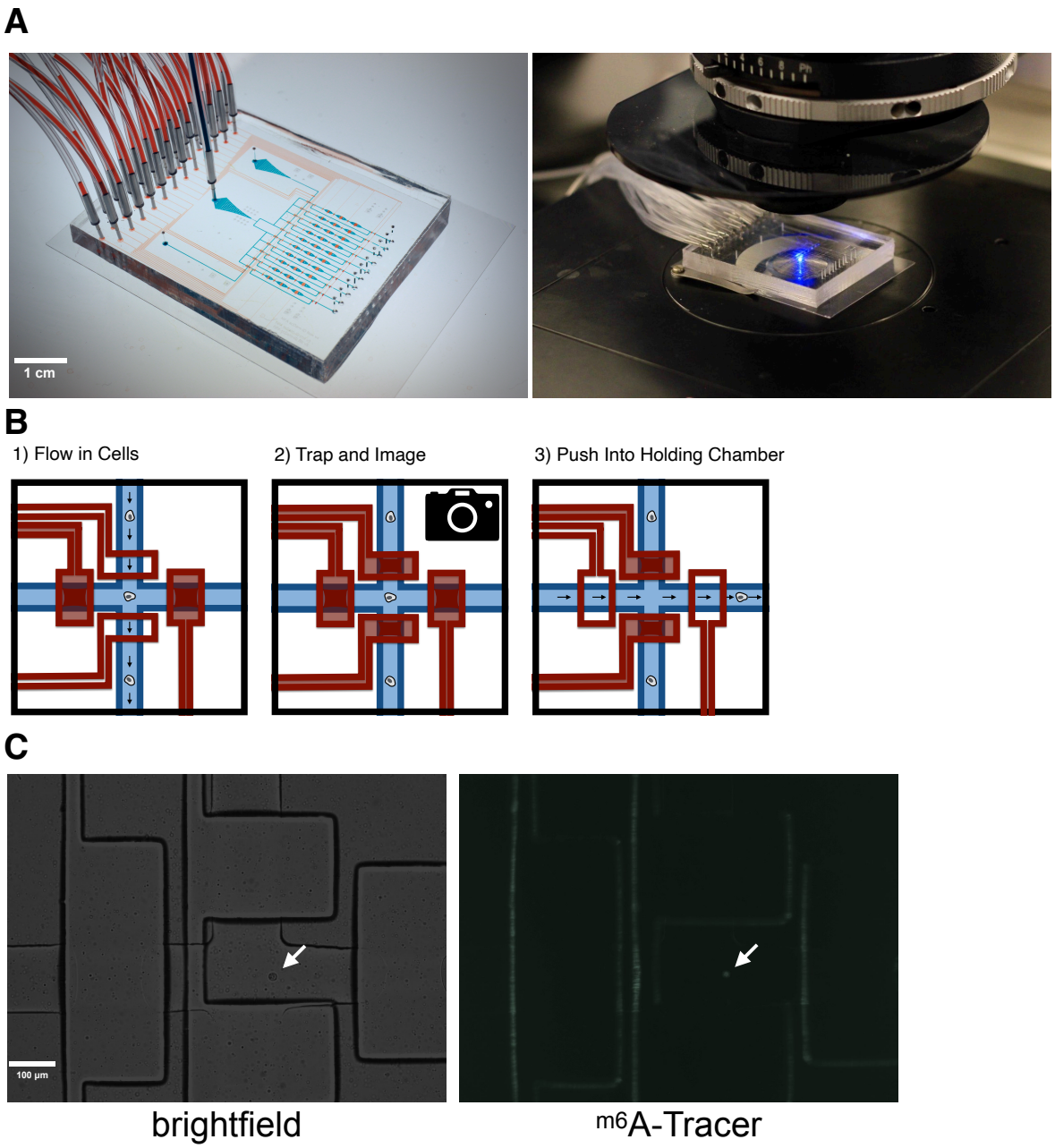


Figure 2.4. Illustration of cell trapping procedure
(**A**) an image of the µDamID device, with the flow layer filled with blue food coloring and the control layer filled with red food coloring, alongside an image of the device as operated on an inverted microscope. (**B**) Cells are driven through the device by peristaltic pumping or pressure-driven flow. Valves are actuated to confine the cell in the trapping region, where it is imaged, and if selected, it is pushed by dead-end filling into a holding chamber to the right of the trapping region. (**C**) 10X magnification images of an actual cell held in the trapping chamber prior to high-resolution imaging and sequencing (cell #018, expressing untethered Dam).

Key design elements and optimizations

The final design of the device incorporates several key improvements over previous versions, designed to address common pitfalls in device fabrication and operation. I highlight these here for consideration in future designs of similar devices. **Figure 2.5** illustrates five modifications from the first to the second major version of the device. Firstly, the device width was decreased, allowing for two device molds to be produced on a single 10 cm silicon wafer. Inlet ports on the flow layer, which were previously continuous circles, had "cartwheel" patterns added to improve their visibility when punching inlet holes. One common failure mode involved the large hydration paddles above each reaction chamber collapsing and fusing to the other layer of the device, blocking the channel and altering the chamber volume. To address this, I added a column directly beneath the hydration paddle and increased the chamber dimensions to keep the volume equivalent. I also increased the dimensions of valves on the control layer to allow for greater alignment tolerance, and I added large inlet filters to remove debris such as small pieces of PDMS that would frequently clog channels. These inlet filters also reduce the speed of fluid flowing through them, relative to the narrow channels they flow into, allowing many cells to be examined simultaneously under the microscope before they enter the channels and trapping regions. Finally, I redesigned the cell trap valves to make the right valve individually addressable, meaning each lane's right valve can be opened or closed independently of all the other lanes, while the left valve open or closes simultaneously for all lanes. In the previous, opposite configuration, residual pressure in the system would cause all holding chambers to fill with liquid, rendering them useless for cell trapping by dead-end filling. This configuration guarantees only the contents of the targeted trapping area plus clean buffer get loaded into the holding chamber.

Further changes to the device in version 3 are illustrated in **Figure 2.6**. Firstly, the geometry of the reagent input tree was adjusted so that all lanes have equivalent path lengths from the reagent inlet to either the cell inlet (used as an outlet in washing steps) or the waste outlet. This equalizes the volumetric flow rate between lanes and helps guarantee more efficient washing, and it was especially critical after moving the independently addressable valves from the left to the right side of each trapping chamber in version 2. There is still a potential "dead zone" between lanes 5 and 6, but this can be washed efficiently by alternately closing and opening valves above and below this region. Another common failure mode involved the right valve in each trapping area failing to close fully, due to the narrower geometry of the holding area and due to the effect of photoresist reflow causing channel junctions to rise, creating an aspect ratio incompatible with valve closure. To address this, I widened the holding chamber and the right valve. I also switched the chip configuration to be a "push-up"

style chip, with the control layer being the thin layer that gets bonded to glass, and the flow layer being the thick layer that is thermally bonded above that. This configuration allows valves to close with much less pressure, given the uniform thickness of the valve membrane relative to the "push-down" configuration. Given that the device was mounted on a #1.5 (170 micron thick) coverslip and the thin layer was spun to be 55 microns tall, with flow channels at 25 microns tall, the distance from the top of a channel to the bottom of the glass would be at max 250 microns, well within the 300 micron working distance of our 63X/1.2NA water immersion objective. However, for objectives with shorter working distances, fabricating the device in a "push-down" configuration is still possible, though it may require higher pressures for valve closure.

I also adjusted the AZ-40XT reflow protocol from a long reflow process to a 1-minute reflow at 140 °C (detailed below), reducing the degree of junction heightening. I also switched to exclusively use RTV615A PDMS with on-ratio PDMS thermal bonding (Lai et al. 2019), due to this PDMS formulation being more flexible, less prone to inlet splitting, and more predictable in terms of valve membrane material properties. Inlet splitting was also reduced by using wider, 690 micron TiN-coated punches (Accu-Punch MP10 with CR0420275N19R1 punch, Syneo, Angleton, TX). Finally, to reduce the chance that a misaligned final lane valve could lead to leakage during the final PCR step, I added a second valve at each lane outlet, controlled by the same control inlet.

Any future updates to the device design, software, or operation protocol will be posted to streetslab.com.

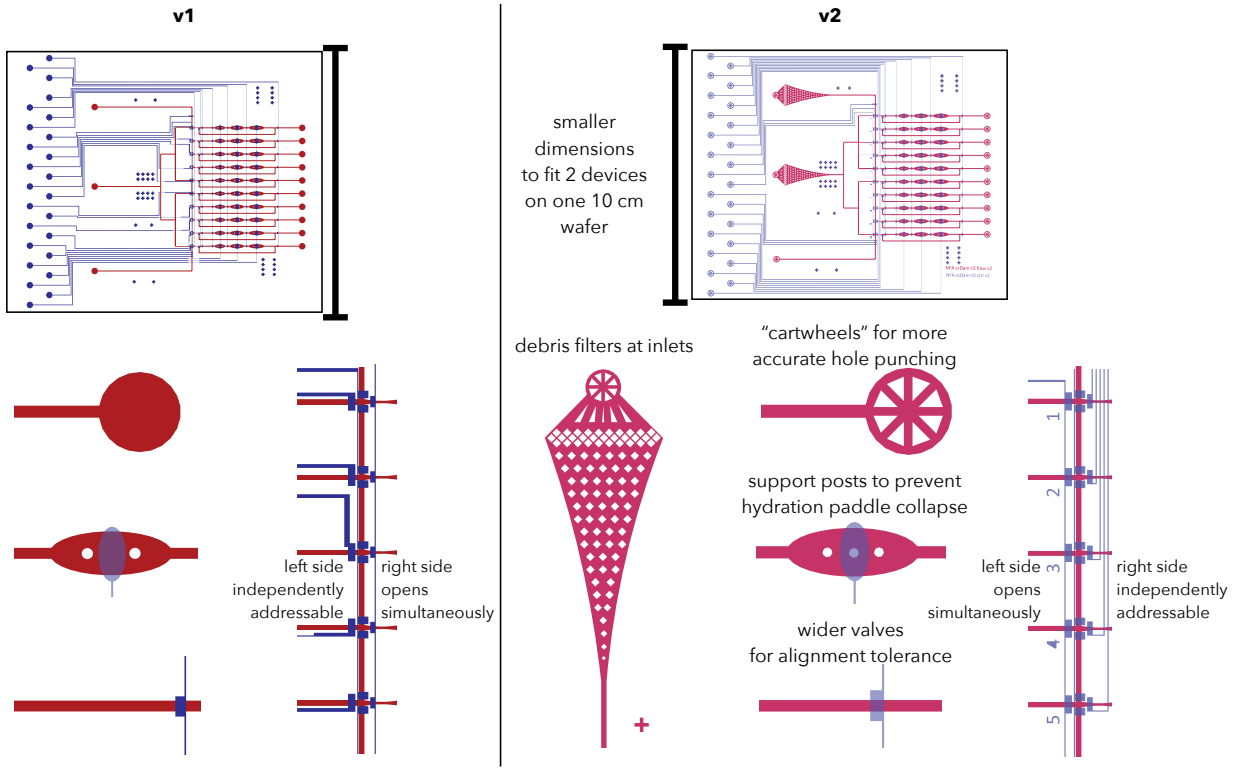


Figure 2.5. Key improvements to the device design

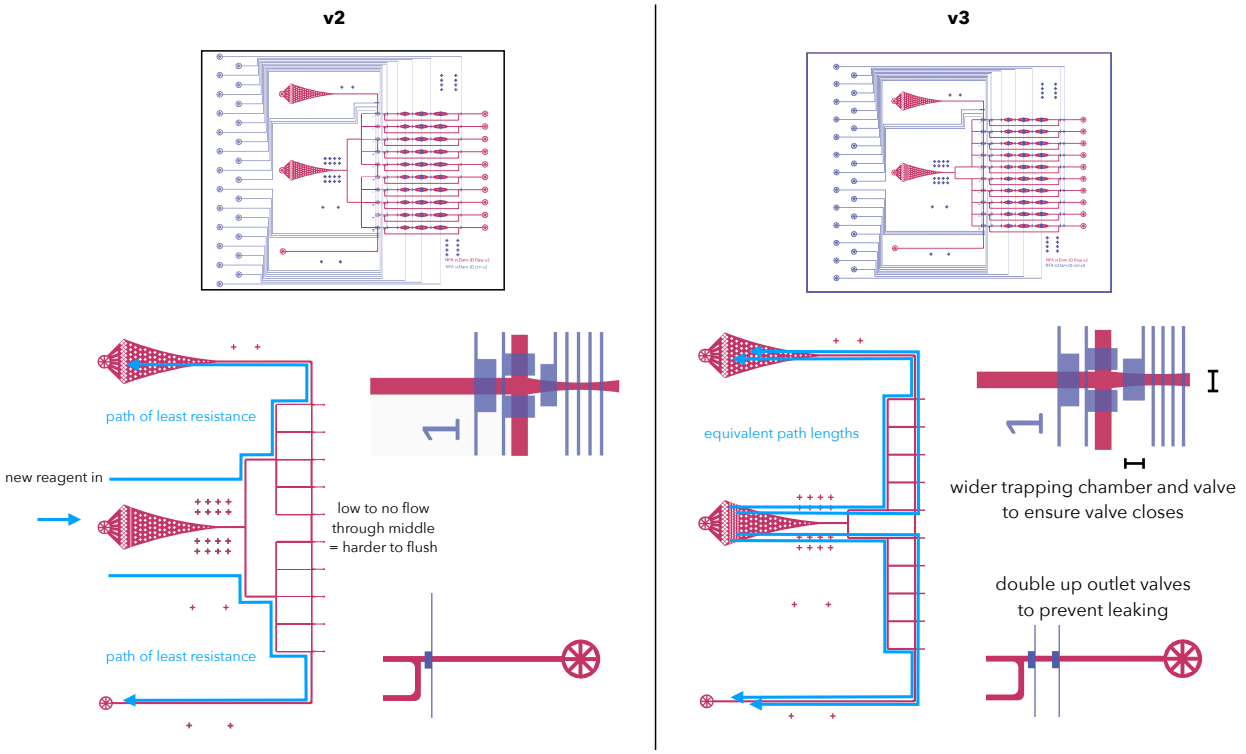


Figure 2.6. Additional improvements to the device design

Mold fabrication optimizations

After adding inlet filters to version 2 of the device design, I initially attempted to make them in the same mold layer as the rest of the flow channels, using AZ-40XT photoresist, which can be reflowed to produce rounded channels essential for valve closing (Unger et al. 2000). Unfortunately, the fine features in the filters tended to reflow into each other, resulting in filters either being overmerged, meaning they provided no filtration of fouling particles, or disconnected, meaning no fluid could pass (**Figure 2.7**). The solution to this was to create a multilayer mold for the flow layer, first by patterning on SU-8 photoresist, which does not reflow and can resolve fine features, to form the filters (which do not overlap valves and thus do not require rounded features), and then AZ-40XT photoresist to form the channels. The filters could also be made to be higher than the channels to increase the capacity of the filter regions (**Figure 2.7)**. The final mold fabrication protocol is below.

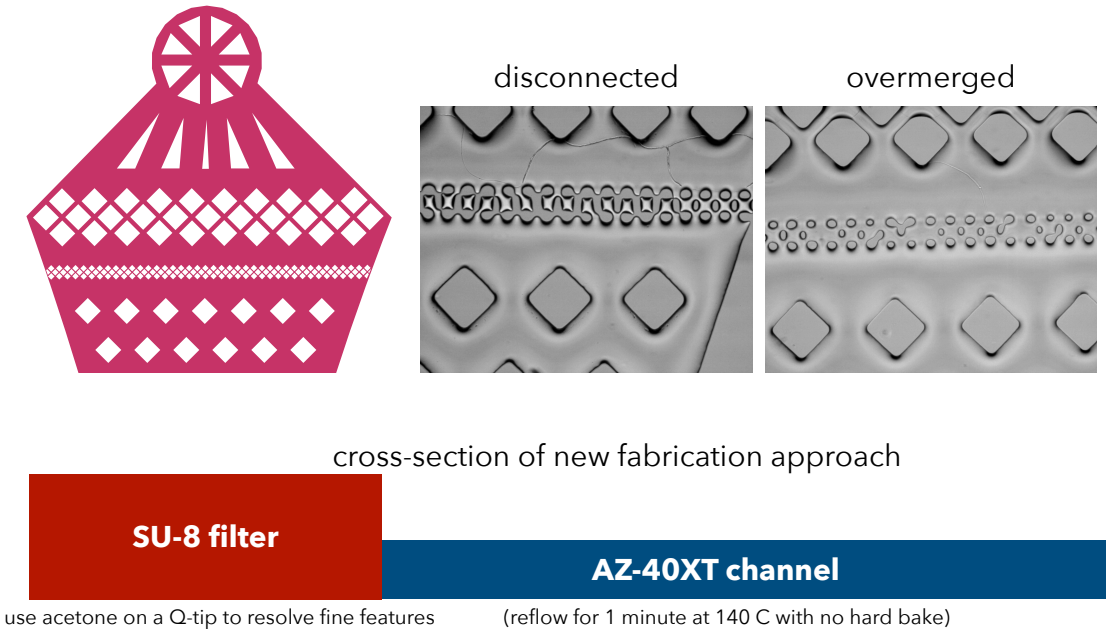


Figure 2.7. Solving filter fabrication issues using multilayer mold fabrication

Microfluidic filters were added to the flow layer cell and reagent inlets, which use finely spaced patterns to create a sieve that will catch any particles that could clog channels. These finely spaced patterns cannot be resolved using AZ-40XT photoresist, because they either merge or disconnect during the reflow step. Images show resulting PDMS channels using disconnected or overmerged filters, which block fluid flow completely or provide no filtering, respectively. To solve this, I first fabricated filters with SU-8 photoresist, which does not reflow, then patterned AZ on top of this. I made the filters taller than the channels to decrease their volumetric flow rate and provide more filtration surface area to avoid fouling.

Molds for casting each layer were fabricated on silicon wafers by standard photolithography. Photomasks for each layer were designed in AutoCAD and printed at 25400 DPI (CAD/Art Services, Inc., Bandon, Oregon). The mask for the thick layer, in this case the flow layer to make push-up valves, was scaled up in size uniformly by 1.5% to account for thick layer shrinkage. A darkfield mask was used for features made out of negative photoresist: the filters on the flow layer and the entire control layer; a brightfield mask was used for all flow layer channels, which were made out of positive photoresist (mask designs available on GitHub; see Data Availability section below). 10 cm diameter, 500 μm thick test-grade silicon wafers (item #452, University Wafer, Boston, MA) were cleaned by washing with 100% acetone, then 100% isopropanol, then DI water, followed by drying with an air gun, and heating at 200 °C for 5 minutes.

To make the control layer mold, SU-8 2025 negative photoresist (MicroChem Corp., Westborough, MA) was spin-coated to achieve 25 μm thickness (7 s at 500 rpm with 100 rpm/s ramp, then 30 s at 3500 rpm with 300 rpm/s ramp). All baking temperatures, baking times, exposure dosages, and development times followed the MicroChem data sheet. All baking steps were performed on pre-heated ceramic hotplates. After soft-baking, the wafer was exposed beneath the darkfield control layer mask using a UV aligner (OAI, San Jose, CA). After post-exposure baking and development, the mold was hard-baked at 150 °C for 5 minutes.

To make the flow layer mold, first the filters were patterned with SU-8 2025, which was required to make fine, high-aspect-ratio filter features that would not re-flow at high temperatures. SU-8 2025 was spin-coated to achieve 40 μm thickness (as above but with 2000 rpm final speed) and processed according to the MicroChem datasheet as above, followed by an identical hard-bake step. Next, AZ 40XT-11D positive photoresist (Integrated Micro Materials, Argyle, TX) was spin-coated on top of the SU-8 features to achieve 20 μm thickness across the wafer (as above but with 3000 rpm final speed). All baking temperatures, baking times, exposure dosages, and development times followed the AZ 40XT-11D data sheet. After development, the channels were rounded by reflowing the photoresist, achieved by placing the wafer at 65 °C for 1 min, then 95 °C for 1 min, then 140 °C for 1 min followed by cooling at room temperature. In our experience, reflowing for too long, or attempting to hard-bake the AZ 40XT-11D resulted in undesirable ‘beading’ of the resist, especially at channel junctions. Because it was not hard-baked, no organic solvents were used to clean the resulting mold. Any undeveloped AZ 40XT-11D trapped in the filter regions was carefully removed using 100% acetone applied locally with a cotton swab.

Soft lithography

Devices were fabricated by multilayer soft lithography (Unger et al. 2000). On-ratio 10:1 base:crosslinker RTV615A PDMS (Momentive Performance Materials, Inc., Waterford, NY) was used for both layers, and layer bonding was performed by partial curing, followed by alignment, then full curing (Lai et al. 2019). To prevent PDMS adhesion to the molds, the molds were silanized by exposure to trichloromethlysilane (Sigma-Aldrich, St. Louis, MO) vapor under vacuum for 20 min. PDMS base and crosslinker were thoroughly mixed by an overhead mixer for 2 minutes, then degassed under vacuum for 90 minutes. Degassed PDMS was spin-coated on the control layer mold (for the 'thin layer') to achieve a thickness of 55 μm (7 s at 500 rpm with 100 rpm/s ramp, then 60 s at 2000 rpm with 500 rpm/s ramp), then placed in a covered glass petri dish and baked for 10 minutes at 70 °C in a forced-air convection oven (Heratherm OMH60, Thermo Fisher Scientific, Waltham, MA) to achieve partial curing. The flow layer mold (for the 'thick layer') was placed in a covered glass petri dish lined with foil, and degassed PDMS was poured onto it to a depth of 5 mm. Any bubbles were removed by air gun or additional degassing under vacuum for 5 minutes, then the thick layer was baked for 19 minutes at 70 °C. Holes were punched using a precision punch with a 0.69 mm punch tip (Accu-Punch MP10 with CR0420275N19R1 punch, Syneo, Angleton, TX). The thick layer was peeled off the mold, cut to the edges of the device, and aligned manually under a stereoscope on top of the thin layer, which was still on its mold. The layers were then fully cured and bonded together by baking at 70 °C for 120 min. After cooling, the devices were peeled off the mold, and the inlets on the thin layer were punched. The final devices were bonded to 1 mm thick glass slides or to #1.5 coverglass, which were first cleaned by the same method as used for silicon wafers above, using oxygen plasma reactive ion etching (20 W for 23 s at 285 Pa pressure; Plasma Equipment Technical Services, Brentwood, CA), followed by heating at 100 °C on a ceramic hot plate for 5 minutes.

One common failure mode during soft lithography resulted from reactive groups left over on the glass after plasma bonding. In the "push-down" valve configuration, this would frequently result in valve membranes irreversibly fusing to the glass once pressurized, even many days after fabrication. In the "push-up" configuration, this sometimes resulted in large valves or hydration paddles fusing to the glass, preventing them from closing when pressurized. With Anushka Gupta, I came up with a way of neutralizing the reactive groups on the glass by simply heating the device while forcing pressurized air through the thin layer's channels. That is, placing the device on a hotplate at 100 °C while forcing in air at 10 PSI for 1 hour stopped any unwanted PDMS-glass bonding from occurring (**Figure 2.8**).

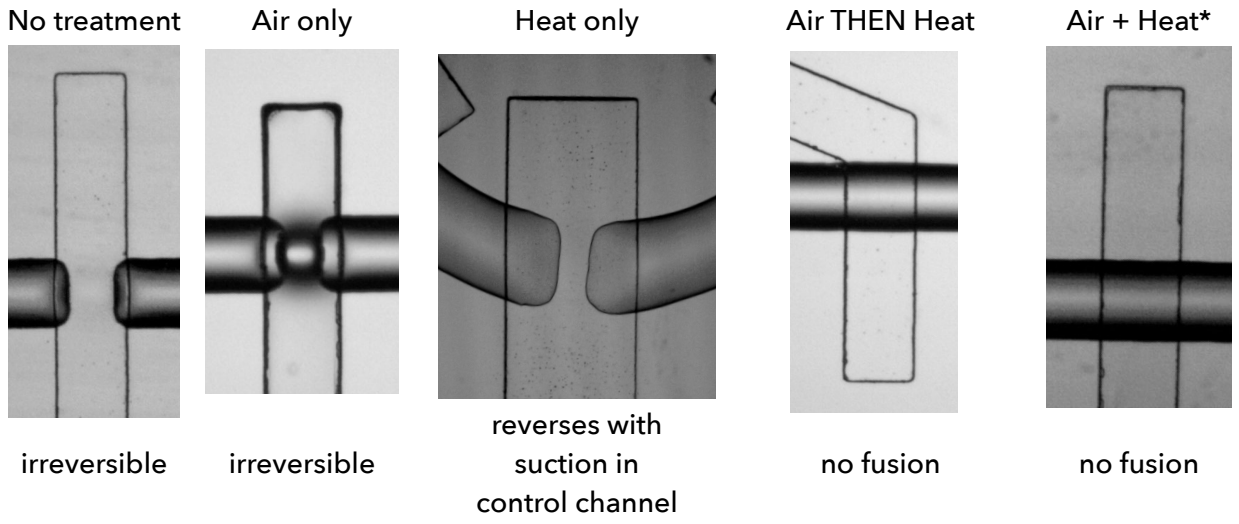


Figure 2.8. Preventing push-down valve fusion to plasma-treated glass
Micrographs show elastomeric valves in various regions of a push-down valved microfluidic device shortly after plasma bonding to glass. On the left is a failure mode caused by fusion of the membrane to the glass below after plasma bonding. This was solved by forcing air through the flow layer channels while placing the device on a hotplate at 100 °C for 1 hour.

Control hardware design and assembly

Valved microfluidic systems require specialized infrastructure for their fabrication and operation (**Figure 2.9**). Mold fabrication requires access to a clean-room facility with low-uv lighting, silicon wafer spin-coaters, precision hotplates, fume hoods for organic solvents, a mask aligner with a collimated and uniform UV light source, and a profilometer for evaluating mold geometry. These resources were available at the Biomolecular Nanotechnology Center in Stanley Hall at UC Berkeley, with assistance from Paul Lum and Naima Azgui. As described above, soft lithography requires a wafer spin-coater, a scale, a precision hole punch, a stereoscope, a scientific oven, a degassing chamber with vacuum pump, a dust-free hood, and an oxygen plasma source. A degassing mixer also helps to improve efficiency and consistency. Each instrument requires calibration and protocol optimization.

Once fabricated, devices were pneumatically controlled by a solenoid valve manifold (**Figure 2.9**; valves from Pneumadyne, Plymouth, MN). Each three-way, normally open solenoid valve switched between a regulated and filtered pressure source inlet at 25

psi (172 kPa) or ambient pressure to close or open microfluidic valves, respectively. Solenoid valves were controlled by the KATARA control board and software (J. A. White & Streets 2018). A detailed description of the hardware and electronic configurations is available in White & Streets 2018. Most operational steps were carried out on inverted microscopes using 4-10X objectives.

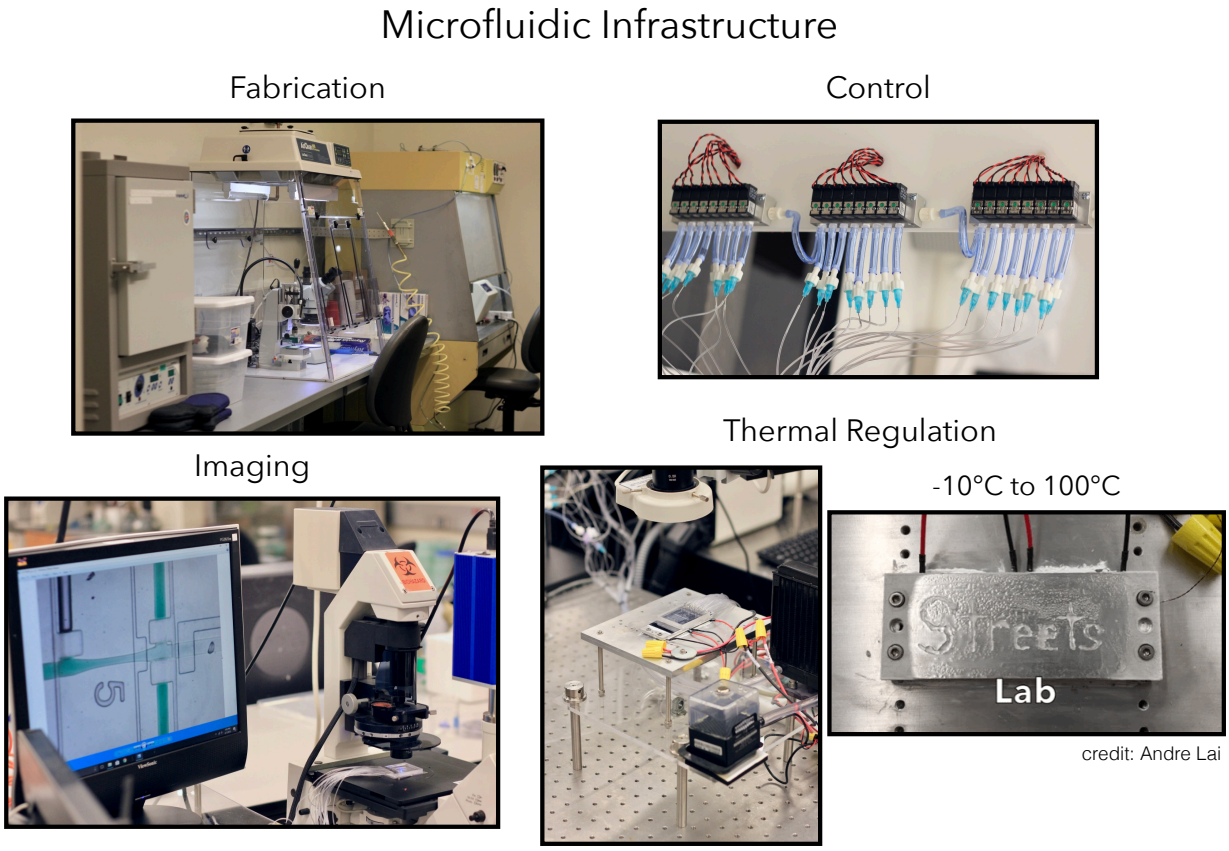


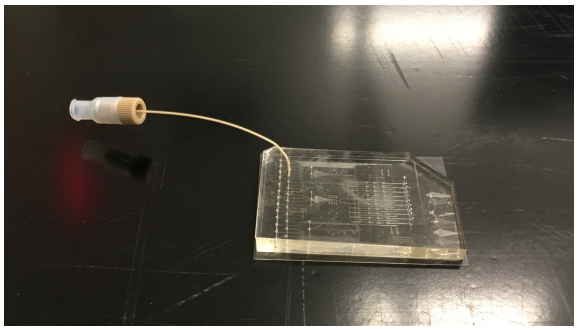
Figure 2.9. Hardware and infrastructure requirements for fabrication and operation These images show some of the required hardware for device operation. Pneumatic control valves (top right) allow for computer-controlled elastomeric valve actuation, and a custom-built thermocycler (bottom right) allows for precise reaction temperature control. Andre Lai played a key role in helping to set this all up, as did Jonathan White, who developed the KATARA hardware and software for valve control (not shown; White & Streets 2018) .

For incubation steps, the device was placed on a custom-built liquid-cooled thermoelectric temperature control module (**Figure 2.9** and **Figure 2.10**; TC-36-25-RS232 PID controller with a 36 V / 16 A power source and two serially connected VT-199-1.4-0.8P TE modules and an MP-3022 thermistor; TE technologies, Traverse City,

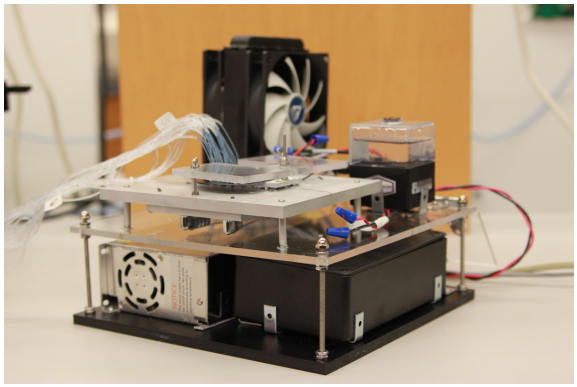
MI) controlled by a computer using a new KATARA GUI software module (available at https://github.com/altemose/microDamID). This thermal controller uses solid-state Peltier heat pumps to heat and cool a small aluminum plate topped with a piece of silicon wafer, which provides contrast for imaging with a stereoscope. Above this sample plate is a machined piece of acrylic that holds the microfluidic device flush against the sample plate and prevents bowing that occurs when the device is heated or cooled, which can lead to uneven thermal control (**Figure 2.10**). Below the thermoelectric heat pumps is a large aluminum heat sink to remove waste heat. This heat sink was custom machined to include a chamber for liquid cooling. It is connected in series with tubing to a small water pump and a radiator with an attached computer fan. The system is sealed with grease and filled with a 1:1 mix of ultrapure water and ethylene glycol, which prevents corrosion and microbe growth. Different heat sink configurations are possible, but a complete parts list is shown in **Table 2.2**. I worked with Anushka Gupta to optimize the power supply and thermoelectric plate configuration, and we showed than we can achieve both heating and cooling from 55 °C to 98 °C (the full operational range for PCR) in under 10 seconds each. A layer of mineral oil was applied between the device and the temperature controller to improve thermal conductivity and uniformity. A stereoscope was used to monitor the device while on the temperature controller. Critically, we replaced the standard metal pins and Tygon tubing at each control layer inlet with plastic PEEK tubing, which has higher thermal tolerance than Tygon, preventing pressure leaks that were common during the heating stages of PCR (**Figure 2.10**).

To set up each new device, each pneumatic valve was connected to one control inlet on the microfluidic device by serially connecting polyurethane tubing (3/32" ID, 5/32" OD; Pneumadyne) to Tygon tubing (0.5 mm ID, 1.5 mm OD) to >4 cm PEEK tubing (0.25 mm ID, 0.8 mm OD; IDEX Corporation, Lake Forest, IL). Solenoid valves were energized to de-pressurize the tubing and the tubing was primed by injecting water using a syringe connected to the end of the PEEK tubing, then the primed PEEK tubing was inserted directly into each punched inlet hole on the device. Solenoid valves were de-energized to pressurize the tubing until all control channels on the device were fully dead-end filled, then each microfluidic valve was tested and inspected by switching on and off its corresponding solenoid valve. All valves were opened and the device was passivated by filling all flow-layer channels with syringe-filtered 0.2% (w/w) Pluronic F-127 solution (P2443; MilliporeSigma, St. Louis, MO) from the reagent inlet and incubating at room temperature for 1 hour. The device was then washed by flowing through 0.5 ml of ultra-filtered water, followed by air to dry it.

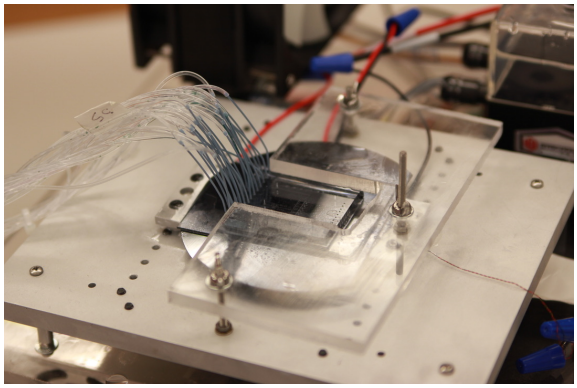
further optimization



use PEEK tubing instead of Tygon + metal pins; PEEK has a higher max operating temperature and is less likely to fail during PCR



Improved thermal controller uses 36V power supply, offering faster ramp rates



added acrylic clamp to hold chip down and prevent warping at high/low temperatures

Figure 2.10. Additional optimizations to device operation

Table 2.2. Detailed parts list for custom-built thermocycler

Supplier	Cat #	Description	#	Price Each	NOTES	URL
TE Technology	VT-199-1.4-0.8P	Thermoelectric Module-200C 120mm Teflon wire leads potted	2-4	$64.10	BE SURE TO ORDER POTTED!!! wire 2 in series	https://tetech.com/peltier-thermoelectric-cooler-modules/high-temperature/
TE Technology	MP-3022	Thermistor with a 0.9 mm diameter sensor heat and 32 gauge wire. 50K ohms at 25 °C, for measuring from 0 °C to +150 °C.	1	$17	good idea to buy a spare	https://tetech.com/product/mp-3022/
TE Technology	TC-36-25-RS232	PID temp controller board that can modulate power input from 12 V up to 36 V, or from 0 V to 36 V with a second power supply, up to 25 A.	1	$493	never supply with wrong voltage polarization	https://tetech.com/product/tc-36-25-rs232/
Eyeboot	EYE-600W-36V-22	Power Supply, Output Power: 36 Vdc at 16 A max.	1	$36	or similar (>= 16 Amp rating)	https://www.amazon.com/gp/product/B01EY6BALS/ref=ppx_od_dt_b_asin_title_s00?ie=UTF8&psc=1
TE Technology	RS232 Adapter	USB-to-RS232 converter needed when connecting the PID controller to a host computer with an available USB port.	1	$43	potentially cheaper from a different supplier	https://tetech.com/product/rs232-adapter/
Amazon	SC-300T	DC 12V Ultra-Quiet Water Cooling Pump Tank 4W Reservoir max. 300L/h	1	$21.59	Amazon	https://www.amazon.com/gp/product/B00MB9EP3G/ref=ppx_yo_dt_b_search_asin_title?ie=UTF8&psc=1
Amazon	AFACO-12000-GBA01	ARCTIC F12 - 120 mm Standard Case Fan, very quiet motor, Computer, Push- or Pull Configuration, Fan Speed: 1350 RPM	2	$8	Buy 0.38 Amps	https://www.amazon.com/gp/product/B002KTVFTE/ref=ppx_yo_dt_b_search_asin_title?ie=UTF8&psc=1
LowBrow Customs	1224	5/16 inch Hose Barb 90 Elbow x 1/8 inch NPT - Chrome	2	$11.10	buy spares	https://www.lowbrowcustoms.com/products/cycle-standard-5-16-hose-barb-90-elbow-x-1-8-npt-chrome
Amazon	B08FSPJ3HG	120mm Aluminum Liquid Cooling Radiator Heat Sink	1	$17.99	or similar	https://www.amazon.com/DIYhz-Computer-Radiator-Aluminum-Exchanger/dp/B08FSPJ3HG/ref=sr_1_7?dchild=1&keywords=liquid%2Bcooling%2Bradiator&qid=1618950196&sr=8-7&th=1
various		custom machined aluminum heat sink with liquid cooling chamber		$1,000	alternative heat sinks can be used	design available at streetslab.berkeley.edu
various		potting box, wiring		$30	to hold electronics	
various		acrylic and hardware for housing		$50	custom housings can vary	
various		ethylene glycol		$10	mix 1:1 with water for coolant	
				$1,802	**TOTAL**	

Device operation

Initially, all chamber valves and reagent inlet valves were closed. Gel-loading pipette tips were used to inject reagents and cells into the flow channels. To prepare the device for operation, pick buffer was injected into the reagent inlet and pressurized at 5-10 psi to dead-end fill the reagent inlet channels. Negative controls were generated by injecting pure pick buffer into one of the holding chambers before trapping and sorting cells into the other lanes. 50 µl of cell suspension was then loaded into a gel-loading pipette tip, and injected directly into the cell inlet. A high-precision pressure regulator was used to load the single-cell suspension at 1 psi (7 kPa). Cells were observed in the filter region with brightfield and epifluorescence using a 10X objective to identify candidate cells. These were then tracked through the device until they approached the trapping chamber for an empty lane. To trap a candidate cell, the device's peristaltic pump was operated at 1 Hz to deliver that cell to the trap region. The trap valves (above and below the trap region; see **Figure 2.4**) were closed and the cell was imaged with scanning confocal microscopy as described above. If the cell was rejected after imaging, the trap valves were opened and it was flushed to the waste outlet. Otherwise, the cell was injected into the holding chamber by dead-end filling. This process was repeated to fill each lane with single cells for DamID. To test background DNA levels, we filled several lanes with only cell suspension buffer. Nearly undetectable levels of amplified DNA were recovered from these lanes.

After filling all 10 lanes, the reagent inlet and cell trapping channels were flushed with 0.5 ml of water, which exited through the waste outlet and the cell inlet, to remove any remaining Pick buffer or cell debris, then air dried. The same washing and drying was repeated between each reaction step. To inject reagents for each reaction of the DamID protocol, the trap valves were closed, the reagent channels were dead-end filled with freshly prepared and syringe-filtered reagent, then the reagent inlet valves and the valves for the necessary reaction chambers were opened, and each lane was dead-end filled individually to prevent any possible cross-contamination. Reaction contents are described in **Table 2.3**.

Reagents were mixed by actuating the chamber valves at 5 Hz for 5 minutes. At the PCR step, rotary mixing was achieved by using the chamber valves to make a peristaltic pump driving fluid around the full reaction ring. For each reaction step, the device was placed on the thermal controller and reactions were with times and temperatures described in **Table 2.3**. PCR thermocycling conditions are described in **Table 2.4**. To ensure adequate hydration during PCR, all valves were pressurized. Amplified DNA was flushed out of each lane individually using purified water from the reagent inlet,

collected into a gel loading tip placed in the lane outlet to a final volume of 5 µl then transferred to a 0.2 ml PCR strip tube.

Table 2.3. Reaction buffers and conditions

Reaction Stage	Buffer	Incubation
Trapping & Holding	Pick Buffer: 50mM Tris-HCl pH 8.3 75mM KCl, 3mM $MgCl_2$ 137mM NaCl	RT
Lysis	10mM TRIS acetate pH 7.5 10mM magnesium acetate 50mM potassium acetate 0.67% Tween-20 0.67% Igepal 0.67 mg/ml proteinase K	42 °C for 4 hours then 80 °C for 10 min
Digestion	mix 7µl 10X Cutsmart buffer 1 µl DpnI (New England Biolabs, Ipswich, MA) 62 µl H_2O	37 °C for 4 hours then 80 °C for 20 min
Ligation	mix 6 µl 10X NEB T4 ligase buffer 1 µl DamID adapter stock at 25 µM **0.2 µl NEB T4 ligase at 400 U/µl*** 21.8 µl H_2O 1 µl 2% w/v Tween-20	16 °C overnight then 65 °C for 10 min
PCR	from Takara Clontech Advantage 2 kit: mix 5 µl 10X PCR buffer 1 µl dNTPs at 10 mM each 1 µl polymerase mix 0.63 µl DamID primer 21.37 µl H_2O 1 µl 2% Tween-20	See **Table 2.4**

*Ligase unit definitions can differ among different suppliers (e.g. Roche uses Weiss Units, which are equivalent to 200 of NEB's Cohesive End Units). We predict that increasing the amount of NEB T4 ligase from 80 NEB units to 1500 NEB units at this step may improve ligation efficiency further.

Table 2.4. PCR thermocycling conditions

PCR Step	Incubation
1	68 °C for 10 min
2	94 °C for 1 min
3	65 °C for 5 min
4	68 °C for 15 min
5	94 °C for 1 min
6	65 °C for 1 min
7	68 °C for 10 min
8	Go to step 5 (x 3)
9	94 °C for 1 min
10	65 °C for 1 min
11	68 °C for 2 min
12	Go to step 9 (x 22)
13	Hold 10 °C

Oligonucleotides

>AdRt
CTAATACGACTCACTATAGGGCAGCGTGGTCGCGGCCGAGGA
>AdRb
TCCTCGGCCG
>AdR_PCR
NNNNGTGGTCGCGGCCGAGGATC

To anneal DamID adapter (Vogel et al. 2007): mix equal volumes of 50 μM AdRt and 50 μM AdRb in a microcentrifuge tube, then fully submerge it in a beaker of boiling water, and allow the water to equilibrate to room temperature slowly.

Chapter 3

Validation of μDamID and application to study single-cell genome organization

Aims & overview

I sought to evaluate the performance of the μDamID platform by mapping the sequence and spatial location of lamina-associated domains in a human cell line, allowing us to compare our data to previously published LAD maps from DamID experiments in human cell lines (Kind et al. 2015, Lenain et al. 2017). LADs are large (median 500 kb) and comprise up to 30% of the genome in human cells (Guelen et al. 2008). ^{m6}A-Tracer has previously been applied to visualize LADs, which appear as a characteristic ring around the nuclear periphery in confocal fluorescence microscopy images (Kind et al. 2013; **Figure 2.3c**).

I carried out experiments in HEK293T cells for their ease of growth, transfection, suspension, and isolation. To enable rapid expression of Dam and ^{m6}A-Tracer transgenes, I transiently transfected cells with DNA plasmids containing genes for a drug-inducible Dam-LMNB1 fusion protein as well as constitutively expressed ^{m6}A-Tracer. I then induced Dam-LMNB1 expression, optimizing the expression times to maximize the proportion of cells with fluorescent laminar rings (an example of which is visible in **Figure 2.3c**). Because transient transfection yields a heterogeneous population of cells, each with potentially variable copies of the transgenes, it was important for me to be able to take high-resolution confocal images of cells and select only those with visible laminar rings, which were more likely to have the correct expression levels, and which were unlikely to be in the mitosis phase of the cell cycle. This kind of complex sorting would not be possible with sorting methods like fluorescence-activated cell sorting (FACS) but is straightforward in our microfluidic platform.

In addition to processing Dam-LMNB1 cells, I transfected cells with the Dam gene alone, not fused to LMNB1, to provide a negative control demonstrating where the unfused Dam enzyme would mark the genome if not tethered to the nuclear lamina (Vogel et al. 2007). This control is useful for estimating the background propensity for each genomic region to be methylated, since Dam preferentially methylates more

accessible regions of the genome, including gene-rich regions (Aughey et al. 2018, Lenain et al. 2017, Singh & Klar 1992). I selected Dam-only cells that had high fluorescence levels across the nucleus and did not appear mitotic. We also performed bulk DamID (Vogel et al. 2007) in populations of transiently transfected HEK293T cells for validation (credit: Annie Maslan). We used a mutant of Dam (V133A; Elsawy & Chahar 2014), which is predicted to have weaker methylation activity than the wild-type allele on unmethylated DNA, and we hypothesized that it would reduce background methylation, similar to weakened Dam mutants previously engineered to improve methylation specificity (Park et al. 2019). To test this, we performed bulk DamID experiments comparing the mutant and wild-type alleles and found that the V133A mutant allele provides more than twofold greater signal-to-background compared to the wild-type allele (**Figure 3.1**). We also performed RNA sequencing in bulk cells that were untreated or transfected with Dam-only, Dam-LMNB1, or ^{m6}A-Tracer, and we found only two differentially expressed genes (**Figure 3.1**; credit: Annie Maslan). This corroborates similar published findings by others showing that Dam expression and adenine methylation have little or no effect on gene expression in HEK293T cells (Vogel et al. 2007).

I first ran three devices containing 25 imaged cells total, with five empty lanes left as negative controls that did not yield sequenceable quantities of DNA. From these 25 cells, I selected a batch of 18 cells for multiplexed sequencing based on imaging quality and DNA yield in order to achieve a desired level of coverage per cell (**Figure 3.2**). To increase throughput modestly, I built a second microfluidic control system, enabling me to run two devices in parallel, processing up to 20 cells in one experiment. Using four additional devices, I processed a second batch of 40 Dam-LMNB1 cells, with several experimental changes discussed below (**Figure 3.3**). I found that 34 of 38 cells with visible laminar ^{m6}A-Tracer 'rings' yielded sequenceable DNA quantities (89% yield), and I proceeded to sequence those 34 cells plus two more: one cell with no ring, and one cell with a ring but low sequencing yield (D09 and D10, respectively, **Figure 3.3**). In total I sequenced 54 cells from both batches. I obtained a mean of 4 million raw read pairs per cell (range 300k-8M), covering a mean of 140,000 unique DpnI fragments per cell (10k-370k), which falls in the range of previous DamID results from single cells (Kind et al. 2015; **Figure 3.4a**).

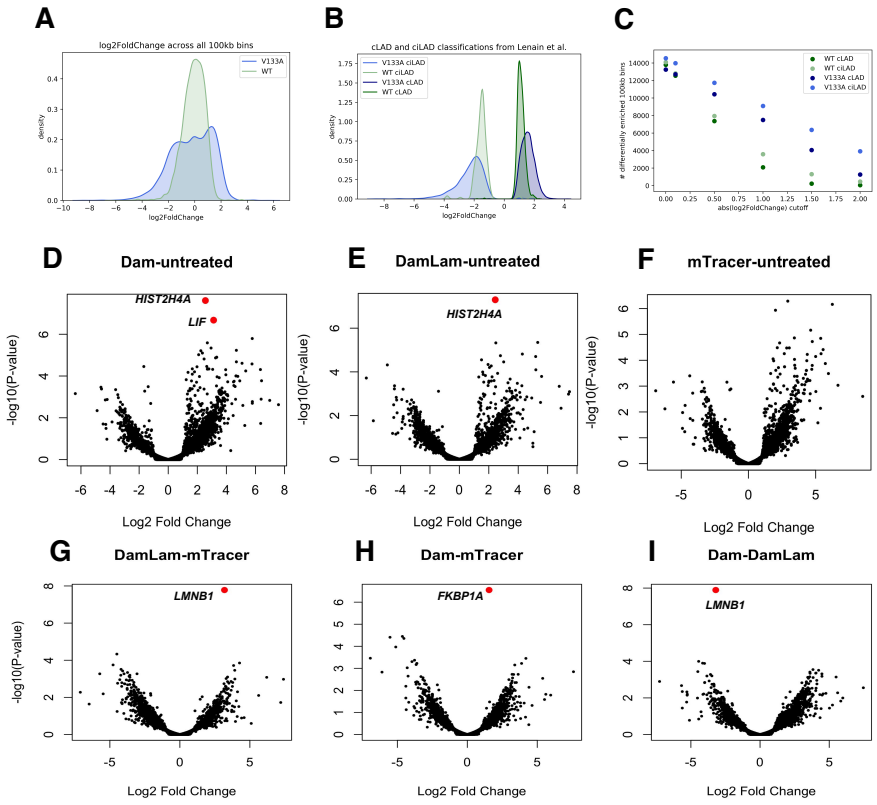
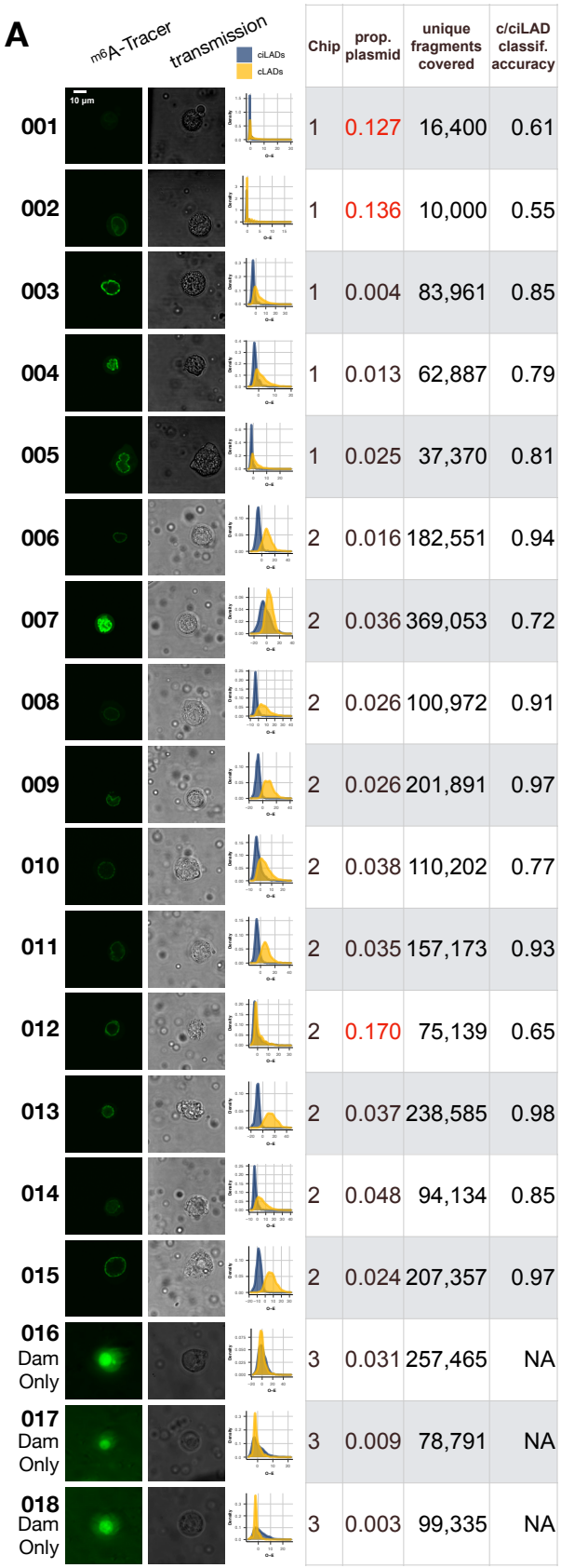


Figure 3.1. Comparing Dam mutants & examining effect of Dam on gene expression (**A**) Kernel density estimate of log_2FoldChange from DESeq2 differential enrichment analysis of Dam-LMNB1 coverage compared to Dam-only as reference. With V133A, more extreme log_2FoldChange values are observed with greater separation between the positive and negative log_2FoldChange peaks. In other words, compared to wild-type, the V133A Dam-LMNB1 and Dam-only signals are more distinct. (**B**) Kernel density estimate of log_2 Fold Change, with cLAD/ciLAD classification from Lenain et al. 2017 indicated, shows greater separation for cLAD and ciLAD signal with V133A.
(**C**) V133A has higher sensitivity than WT, with more differentially enriched regions at each log_2FoldChange threshold for calling significant differential enrichment. (**D-I**) Significantly differentially expressed genes (logFC significantly > 1 and adjusted p-value < 0.01) are indicated in red for bulk HEK293T cells transfected with Dam, Dam-LMNB1, ^{m6}A-Tracer, or no treatment control. Differentially expressed genes compared to no treatment control are *HIST2H4A* and *LIF* for Dam, *HIST2H4A* for Dam-LMNB1, and no genes for ^{m6}A-Tracer. When comparing Dam to ^{m6}A-Tracer, the only differentially expressed gene is *FKBP1A*, which is expected given the mutated FKBP1A-derived destabilization domain tethered to Dam in our construct. When comparing Dam-LMNB1 to ^{m6}A-Tracer, the only differentially expressed gene is *LMNB1*, which is again expected given *LMNB1* is expressed from the *Dam-LMNB1* construct itself. Data, analysis, and figure were generated by Annie Maslan.



	Chip	prop. plasmid	unique fragments covered	c/ciLAD classif. accuracy
001	1	0.127	16,400	0.61
002	1	0.136	10,000	0.55
003	1	0.004	83,961	0.85
004	1	0.013	62,887	0.79
005	1	0.025	37,370	0.81
006	2	0.016	182,551	0.94
007	2	0.036	369,053	0.72
008	2	0.026	100,972	0.91
009	2	0.026	201,891	0.97
010	2	0.038	110,202	0.77
011	2	0.035	157,173	0.93
012	2	0.170	75,139	0.65
013	2	0.037	238,585	0.98
014	2	0.048	94,134	0.85
015	2	0.024	207,357	0.97
016 Dam Only	3	0.031	257,465	NA
017 Dam Only	3	0.009	78,791	NA
018 Dam Only	3	0.003	99,335	NA

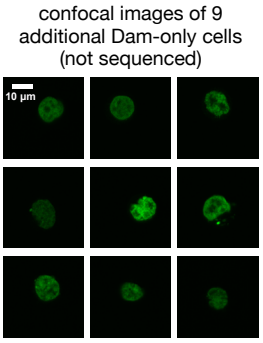
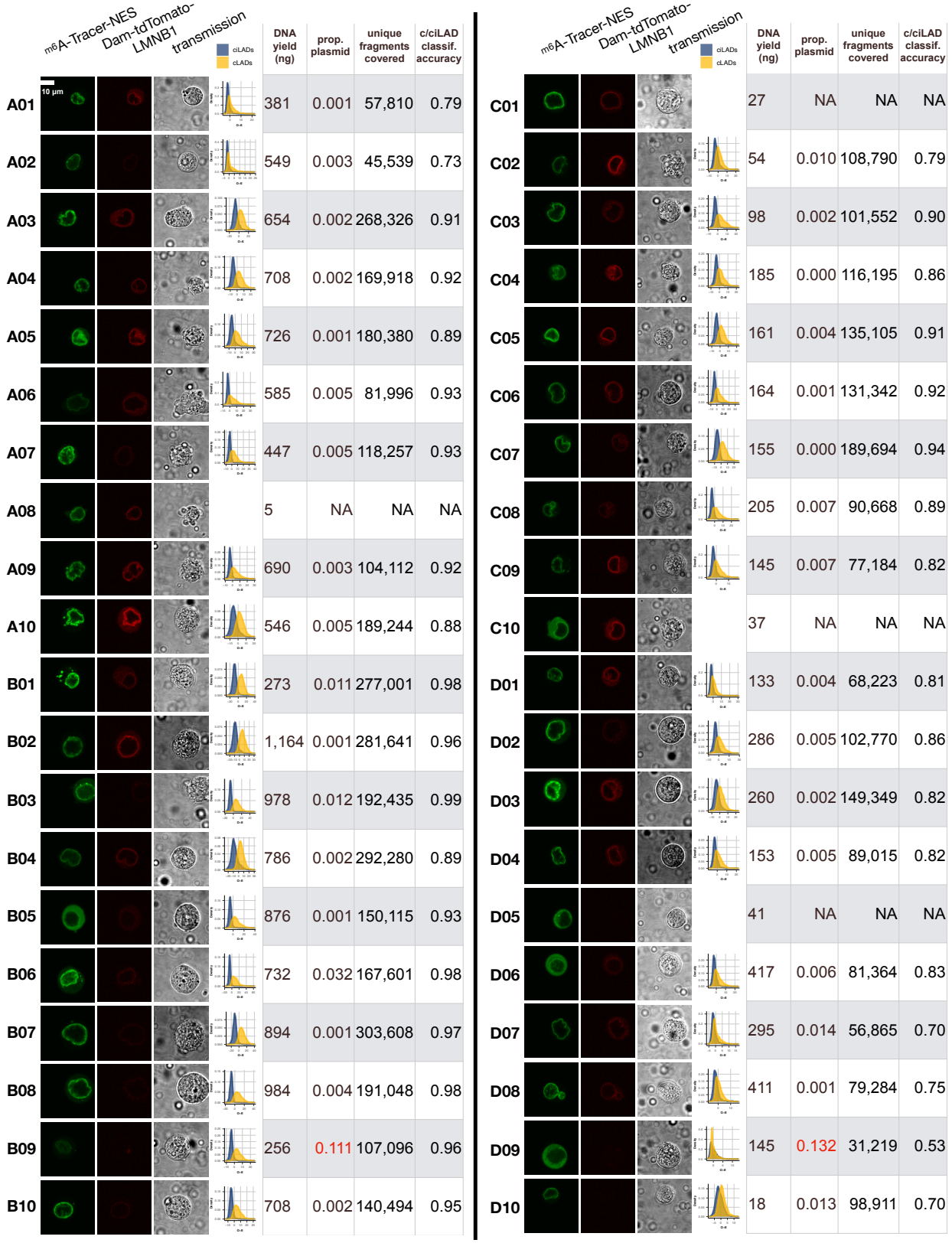


Figure 3.2. Images and sequencing statistics for each batch 1 cell

(**A**) Each row corresponds to a single batch 1 cell, showing its ^{m6}A-Tracer image, transmission image, coverage distributions in c/ciLAD control regions, identifier for the device (chip) it was sequenced on, proportion of reads mapping to the transfected plasmid, number of unique DpnI fragments covered in the genome, and classification accuracy on the c/ciLAD control regions. Nine confocal ^{m6}A-Tracer images from unsequenced Dam-only cells are provided for comparison to the widefield images acquired for cells 015, 016, and 018.

Figure 3.3 (on next page).

As in **Figure 3.2** but for 40 batch 2 cells, with Dam-tdTomato-LMNB1 images and library DNA yields added. Letters in each cell identifier indicate which device they were processed on.



	DNA yield (ng)	prop. plasmid	unique fragments covered	c/ciLAD classif. accuracy
A01	381	0.001	57,810	0.79
A02	549	0.003	45,539	0.73
A03	654	0.002	268,326	0.91
A04	708	0.002	169,918	0.92
A05	726	0.001	180,380	0.89
A06	585	0.005	81,996	0.93
A07	447	0.005	118,257	0.93
A08	5	NA	NA	NA
A09	690	0.003	104,112	0.92
A10	546	0.005	189,244	0.88
B01	273	0.011	277,001	0.98
B02	1,164	0.001	281,641	0.96
B03	978	0.012	192,435	0.99
B04	786	0.002	292,280	0.89
B05	876	0.001	150,115	0.93
B06	732	0.032	167,601	0.98
B07	894	0.001	303,608	0.97
B08	984	0.004	191,048	0.98
B09	256	0.111	107,096	0.96
B10	708	0.002	140,494	0.95

	DNA yield (ng)	prop. plasmid	unique fragments covered	c/ciLAD classif. accuracy
C01	27	NA	NA	NA
C02	54	0.010	108,790	0.79
C03	98	0.002	101,552	0.90
C04	185	0.000	116,195	0.86
C05	161	0.004	135,105	0.91
C06	164	0.001	131,342	0.92
C07	155	0.000	189,694	0.94
C08	205	0.007	90,668	0.89
C09	145	0.007	77,184	0.82
C10	37	NA	NA	NA
D01	133	0.004	68,223	0.81
D02	286	0.005	102,770	0.86
D03	260	0.002	149,349	0.82
D04	153	0.005	89,015	0.82
D05	41	NA	NA	NA
D06	417	0.006	81,364	0.83
D07	295	0.014	56,865	0.70
D08	411	0.001	79,284	0.75
D09	145	0.132	31,219	0.53
D10	18	0.013	98,911	0.70

Figure3.3. Images and sequencing statistics for each batch 2 cell

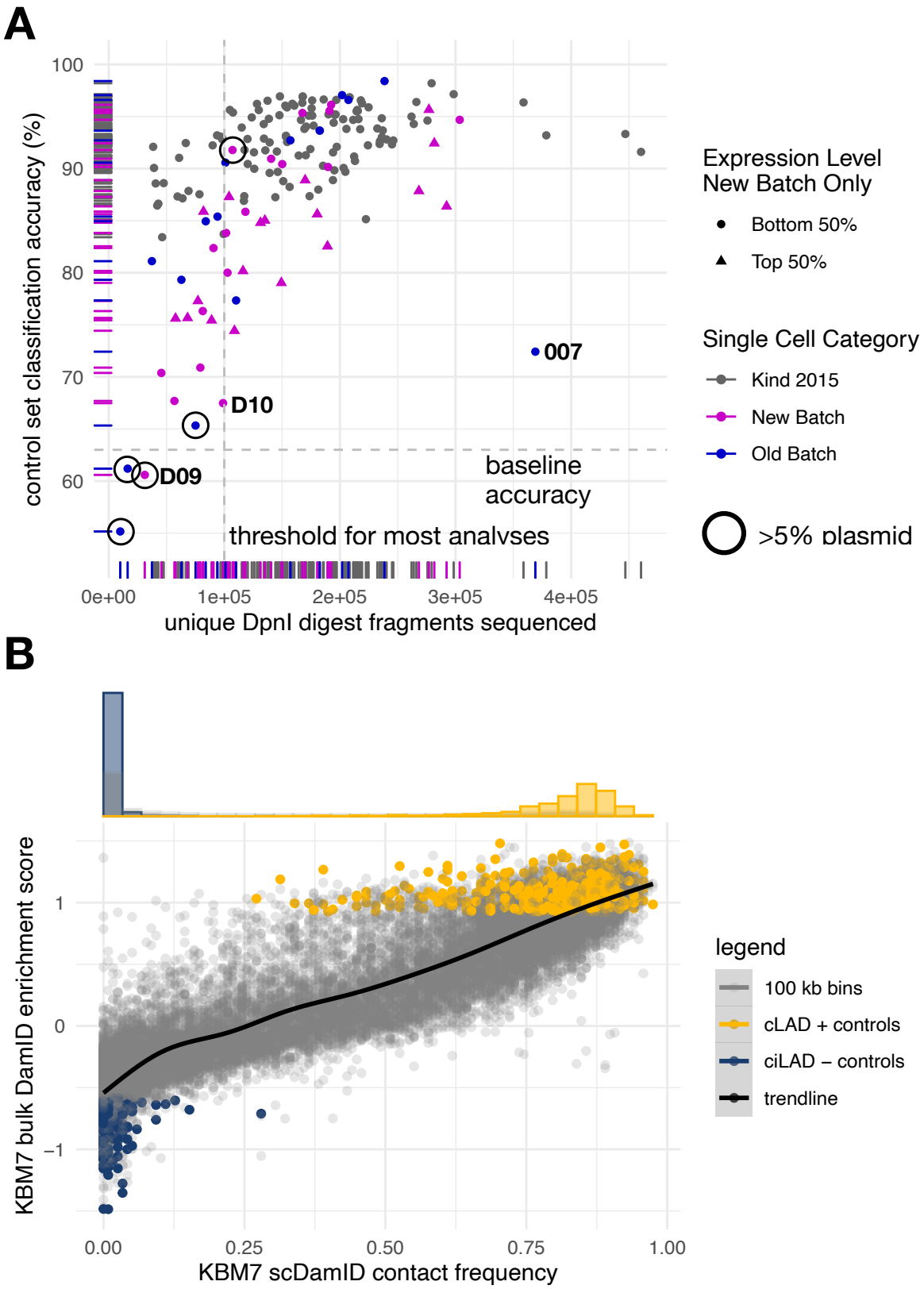

A
100
90
80
70
60
control set classification accuracy (%)
007
D10
D09
baseline
accuracy
threshold for most analyses
0e+00
1e+05
2e+05
3e+05
4e+05
unique DpnI digest fragments sequenced
Expression Level
New Batch Only
Bottom 50%
Top 50%
Single Cell Category
Kind 2015
New Batch
Old Batch
>5% plasmid
B
1
0
−1
KBM7 bulk DamID enrichment score
0.00
0.25
0.50
0.75
1.00
KBM7 scDamID contact frequency
legend
100 kb bins
cLAD + controls
ciLAD – controls
trendline

Figure 3.4. Library complexity, and published LAD contact frequencies
(**A**) Similar to **Figure 3.5d**, a plot of classification accuracy vs library complexity (number of unique DpnI fragments covered), for all cells compared to single KBM7 cells from Kind et al. 2015 (gray points). Cell colors indicate which batch the cells were sequenced in, and batch 2 cells in the top half of Dam-tdTomato-LMNB1 expression levels are indicated as triangles. High-expression cells tend to have lower classification accuracies, as expected. Cells with fewer than 100k unique fragments show a drop in classification accuracy and were excluded from most downstream analyses. Cells with unusually high proportions of reads mapping to the transfection plasmid are circled. Rug plots are drawn on each axis. Cells D09 (no fluorescence at lamina or nuclear interior) and D10 (low DNA yield) are labeled, along with cell 007 (high m6A-Tracer signal in the nuclear interior, shown in **Figure 3.6b**). (**B**) A plot of contact frequency vs. bulk DamID signal using data from KBM7 cells alone (Kind et al. 2015, Lenain et al. 2017), showing high correlation (r=0.94). Sets of stringent cLADs (gold points) and ciLADs (blue points) were identified in a similar fashion to those in HEK293T cells (top 1200 by ranked bulk DamID enrichment scores, but without a p-value cutoff since none was available, which may explain larger variance in CF values). Above is a histogram showing the distribution of contact frequencies within these control sets.

Validating single-cell LAD maps

To assess whether the single-cell μDamID sequencing data provide accurate measurements of lamina-associated domains, I first compared my single-cell results to those we obtained from bulk DamID in the same cell line. DamID results are reported as a difference or log ratio between the observed coverage from Dam-LMNB1 expressing cells and the expected coverage from background methylation, estimated using coverage from Dam-only expressing cells (see Methods). This measure is computed within fixed 100 kb bins across the genome, as reported previously (Kind et al. 2015). For each cell, I made binary calls of whether a bin was in contact with the lamina in that cell (Methods), and the broad-scale organization of these single-cell binary LAD maps largely agrees with the bulk data (**Figure 3.5a**). By aggregating the raw coverage from our Dam-LMNB1 expressing single cells, I found excellent correspondence with the bulk coverage obtained from millions of cells (**Figure 3.5b**, r=0.84).

In order to create the binary contact maps across the genome within single cells, I trained a classifier on a set of stringent positive and negative controls: regions confidently known to be strongly associated with the lamina or strongly unassociated

with the lamina based on bulk DamID data from our study and others (Lenain et al. 2017; see Methods). Positive controls were derived from 100 kb bins across the genome that were previously annotated in other human cell lines to be strongly associated with the nuclear lamina (referred to as constitutive LADs, or cLADS), and further filtered to have the highest bulk DamID scores in HEK293T cells. These bins are therefore most likely to have high contact frequencies in individual cells (Kind et al. 2015; **Figure 3.4b**). Negative controls were similarly determined using bulk data to be consistently not associated with the nuclear lamina across cell types and in our cells (referred to as constitutive inter-LADs, or ciLADS), making them most likely to have low contact frequencies in individual cells (Kind et al. 2015; **Figure 3.4b**). These stringent control sets constitute roughly 4% of the genome each.

For each single cell expressing Dam-LMNB1, I computed the distribution of its normalized sequencing coverage in bins from the positive (cLAD) and negative (ciLAD) control regions (**Figure 3.5c**), with the expectation that cLADs have high coverage and ciLADs have little or no coverage in each cell. Given these control distributions, I chose a coverage threshold to maximally separate the known cLADs and ciLADs. Across the 51 Dam-LMNB1 cells, I determined thresholds that distinguish the known cLADs and ciLADs with a median accuracy of 85% before any filtering (vs. 63% if all bins are scrambled), which correlates positively with the number of unique DpnI fragments sequenced per cell, a measure of library complexity (**Figure 3.5d** and **Figure 3.4a**). Because I used a transient transfection system, expression levels of Dam-LMB1 varied widely from cell-to-cell, reducing classification accuracy in some cells with high noise levels due to background methylation. I filtered higher-noise cells using a threshold of unique covered fragments, leaving 31 Dam-LMNB1 cells with a median classification accuracy of 90% (range 74%-98%, **Figure 3.5d**). This classification approach enables inference of expected error rates for each bin's coverage level in each cell, providing a framework for data normalization, interpretation, and further inference. These error rates can be represented with receiver operating characteristic (ROC) curves for each cell, showing the empirical tradeoff between false-positive and false-negative classifications at varying normalized coverage thresholds (**Figure 3.5e**).

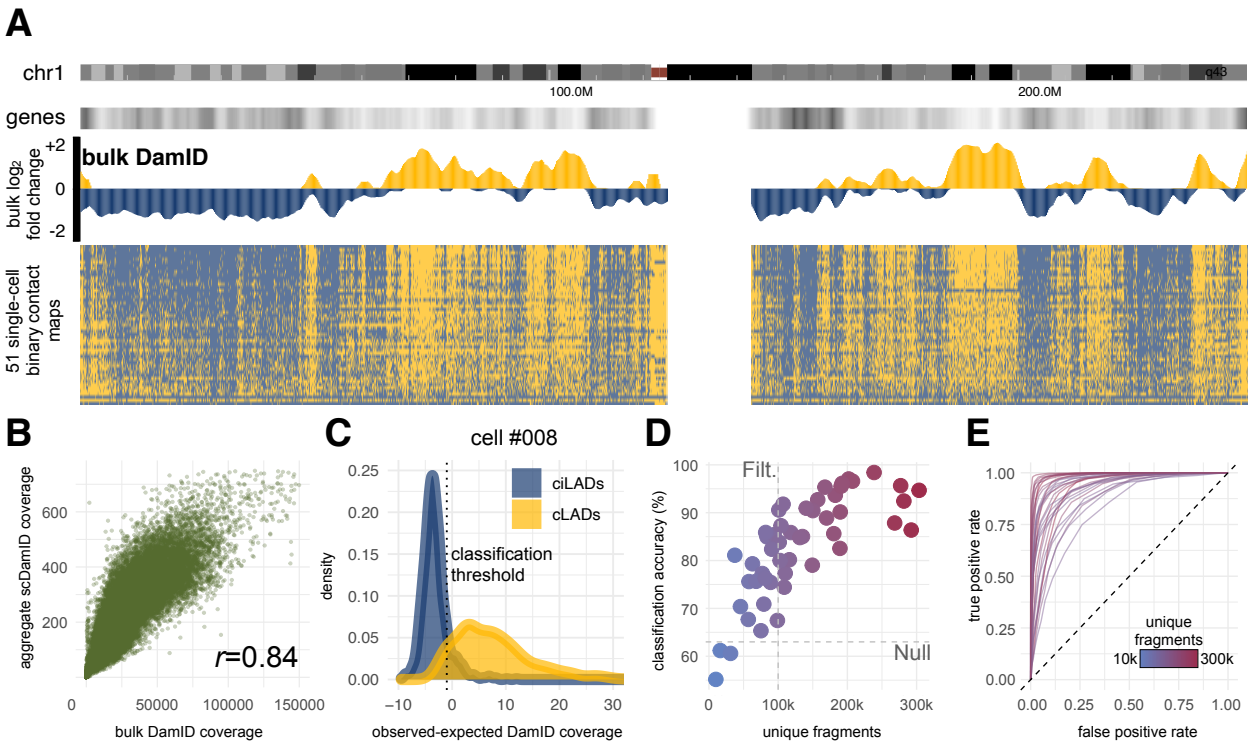


Figure 3.5. Validation of μDamID sequencing data

(**A**) Comparison of bulk DamID sequencing data and single-cell sequencing data across all of human chromosome 1 for Dam-LMNB1 data normalized to bulk Dam-only data. Positive values (gold) represent regions associated with the nuclear lamina, which tend to have lower gene density (second track from top). Each row of the binary contact map represents a single cell, sorted from top to bottom by genome-wide classification accuracy. (**B**) Scatterplot comparing raw Dam-LMNB1 sequencing coverage in bulk vs. aggregated single-cell samples. (**C**) Normalized coverage distributions within positive (cLADs, gold) and negative (ciLADs, blue) control sets in one cell (#008) expressing Dam-LMNB1. The threshold that distinguishes these sets with maximal accuracy is shown as a vertical dotted line. (**D**) The maximum control set classification accuracy for each of 50 Dam-LMNB1 cells vs. the number of unique DpnI fragments sequenced for each cell (also indicated by color gradient; outlier cell #007 was excluded). A coverage threshold of 100k fragments used for downstream analyses is indicated, as well as the null accuracy achieved after scrambling values in all bins across the genome (63%). (**E**) Receiver-Operator Characteristic curves for 31 Dam-LMNB1 cells above the 100k coverage threshold.

I next computed pairwise correlations between the raw coverage for all single cells with each other, with the bulk data, with aggregated published single-cell DamID data (from Kind et al. 2015), and with the number of annotated genes in each 100 kb bin genome-wide. After removing low-complexity cells, I performed unsupervised hierarchical clustering on these datasets and produced a heatmap of their pairwise correlations (**Figure 3.6a**). I found that the 3 Dam-only single cells cluster with each other, along with the bulk Dam-only data, with the Kind et al. Dam-only data, and with the number of genes, as expected. The Dam-LMNB1 cells cluster separately with each other, with the bulk Dam-LMNB1 data, and with the Kind et al. Dam-LMNB1 data, confirming that these sequencing data are measuring meaningful biological patterns in single cells. These clusters also reflect expected nuclear spatial distributions of methylation reported by ^{m6}A-Tracer fluorescence (**Figure 3.6b-d**). Interestingly, one Dam-LMNB1 cell with unexpectedly high fluorescence signal in the nuclear interior contained a methylation profile that appeared intermediate between the Dam-Only and other Dam-LMNB1 cells, perhaps owing to high Dam-LMNB1 expression (**Figure 3.6a**). This illustrates how spatial information can be used to validate DamID with joint single-cell imaging and sequencing measurements.

Identifying variable LADs

In any given cell, only a subset of potential LADs come into contact with the lamina, and this subset can vary stochastically between cells (Kind et al., 2013). While most LADs at the lamina appear to remain in stable contact with the lamina throughout interphase, some LADs have been shown to move dynamically short distances towards and away from the lamina within the same cell over time (Kind et al. 2013), also potentially contributing to cell-to-cell variability in LADs. Single-cell DamID provides a unique opportunity to identify LADs that vary *within* a population of cells of the same type.

To measure this variability, at each bin in the genome, I counted the number of Dam-LMNB1 cells in which that bin was classified as having laminar contact (out of 31 total cells) to estimate its contact frequency (Kind et al. 2015), and I developed a method for propagating measurement and sampling uncertainty when inferring the true contact frequency of each bin (Methods, **Figure 3.7**, **Figure 3.8**). As expected, bins belonging to the cLAD control sets have high contact frequencies and lower gene expression while those in the ciLAD control sets have low contact frequencies and higher gene expression (**Figure 3.7**, **Figure 3.9**). Furthermore, I found that contact frequencies for each bin correlated well overall with published single-cell contact frequencies from a different cell line, KBM7 (r=0.8, **Figure 3.9**; Kind et al., 2015).

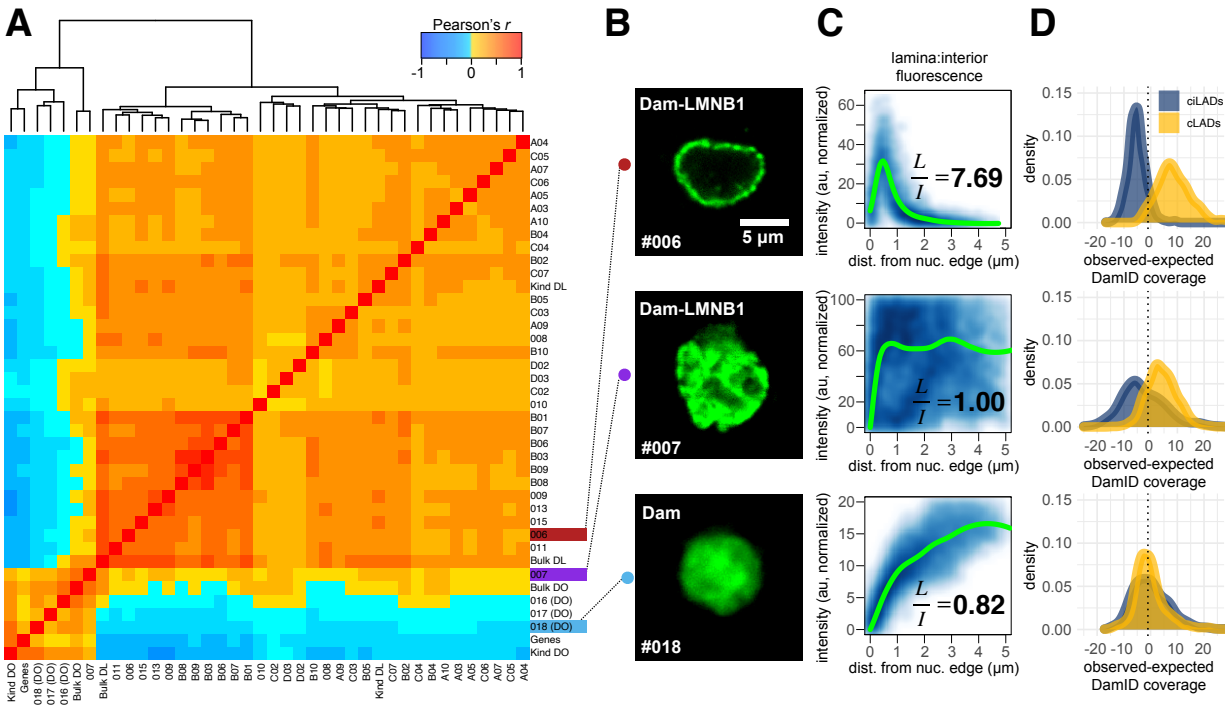


Figure 3.6. Genome-wide comparisons of sequencing data and relation to imaging data

(**A**) Pairwise Pearson correlation heatmap for raw sequencing coverage in 100 kb bins genome-wide, with dendrogram indicating hierarchical clustering results. Cell identifiers label each row (first batch 00**, second batch A-D**). DL = Dam-LMNB1. DO = Dam-only. Genes = number of Refseq genes in each bin. Kind = aggregated single-cell data from Kind et al. 2015. Bulk = bulk HEK293T DamID data from this study. (**B**) Confocal fluorescence microscopy images of ^{m6}A-Tracer GFP signal from 3 cells: one expressing Dam-only (#018), one expressing Dam-LMNB1 but showing high interior fluorescence (#007), and one expressing Dam-LMNB1 and showing the expected ring-like fluorescence at the nuclear lamina (#006). (**C**) Normalized pixel intensity values plotted as a function of their distance from the nuclear edge (blue), with a fitted loess curve overlaid (green). Ratios of the mean normalized pixel intensities in the Lamina (<1 micron from the edge) versus the Interior (>3.5 microns from the edge) are printed on each plot. (**D**) DamID sequencing coverage distributions for each of the cLAD or ciLAD control sets (as in **Figure 3.5c**).

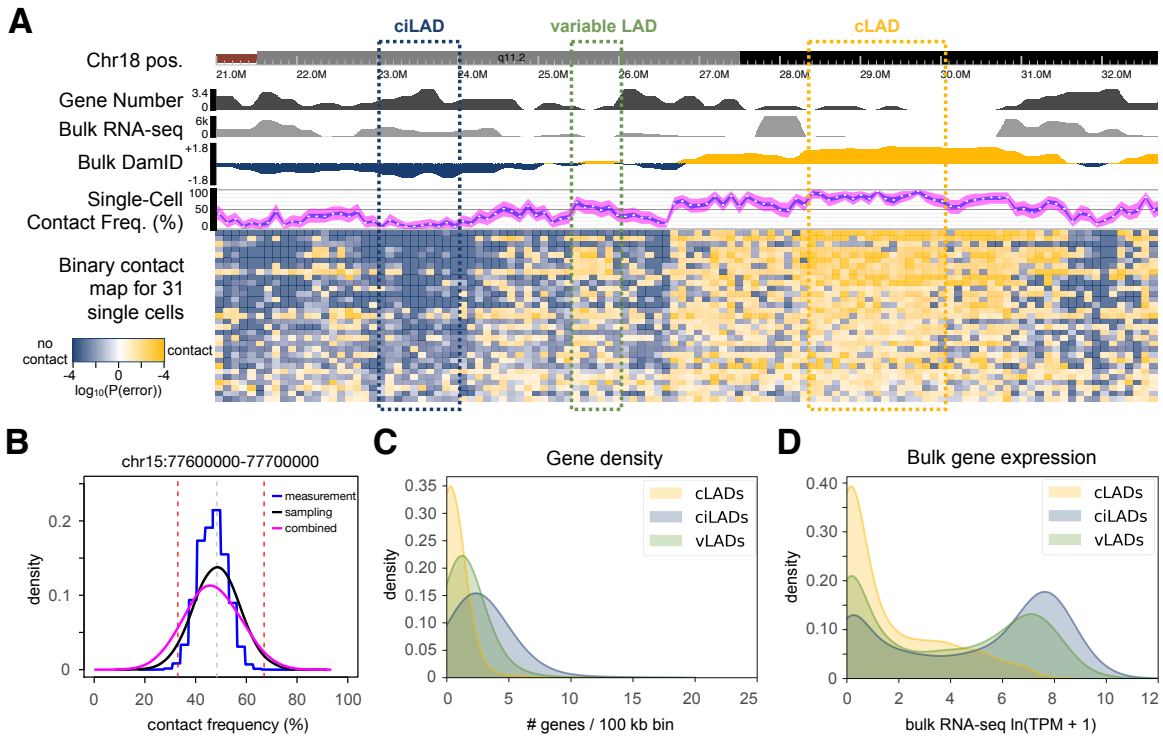


Figure 3.7. Identification and characterization of variable LADs in HEK293T cells (**A**) A browser screenshot from chr18:21-33 Mb. The first track shows the chromosome ideogram and coordinates. The second track reports the number of Refseq genes falling in each 100 kb bin. The third track reports the mean Transcripts Per Million (TPM) value for each gene within each bin from bulk RNA-seq data from untreated HEK293T cells. The fourth track reports the bulk DamID $\log_2$FoldChange values as in **Figure 3.4a**. The fifth track indicates the contact frequency (CF) estimate for each bin (white point), with a blue ribbon indicating the 95% confidence interval for the sample CF (measurement error), and the magenta ribbon indicating the 95% confidence interval for the population CF (measurement + sampling error). The sixth track shows binary contact calls for each bin (columns) in each cell (rows). Shades of gold and blue indicate bins classified as having lamina contact or no lamina contact, respectively, with darker shades indicating higher confidence in the classification. Annotated cLADs and ciLADs are indicated by gold and blue boxes, respectively, with a variable LAD region (vLAD) in green.(**B**) For one bin in a different region, a comparison of measurement (blue) and sampling (black) distributions, along with a combined distribution (magenta) used for contact frequency inference with propagated measurement uncertainty (as shown in (A) track 5). The gray vertical dotted line is the point estimate for that bin, and red dotted vertical lines are drawn at the vLAD CF thresholds (33%, 66%). (**C-D**) Distributions of the number of genes (C) or mean TPM per gene (D) per 100 kb bin for each of the sets of cLADs, ciLADs, or vLADs.

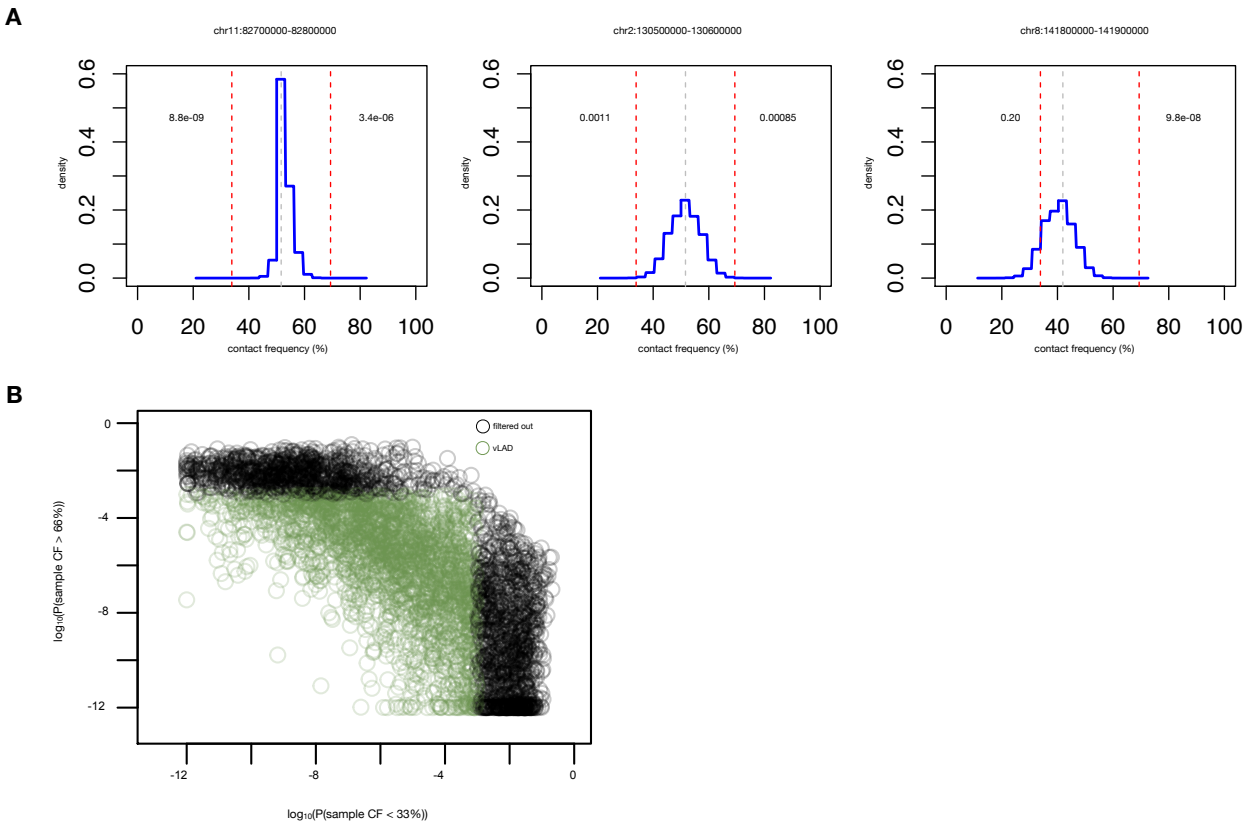


Figure 3.8. Modeling and comparing single-cell contact frequencies between cell types

(**A**) For 3 individual bins in the genome (coordinates listed above each plot), a Poisson-Binomial distribution representing uncertainty in its sample contact frequency estimate after accounting for noise in the sequencing data (classification error rates on the control regions). Gray vertical dotted lines are the point estimate for each bin, and red dotted vertical lines are drawn at 11 and 21 out of 31 cells (CF roughly 33%-66%). Estimated probabilities of lying above or below this interval are indicated on each side of the plot. Note the difference in uncertainty between bins, as well as the difference in probabilities of lying outside the intermediate contact frequency interval. (**B**) When filtering intermediate-contact-frequency bins to choose a final set of high-confidence variable LADs (vLADs), the measurement error distributions were used to select bins with the smallest probabilities of lying above or below the 33-66% CF interval ($p<0.001$ for each test, plotted in green).

To identify variable LADs, I defined a conservative set of bins with intermediate contact frequencies between 33 and 66 percent (Methods, **Figure 3.8**). I hypothesized that these stringently defined regions, which comprise 8% of the genome, would be more gene rich and have higher gene expression than cLADs, given their dynamic positioning in cells. Indeed, these variable LADs show intermediate gene density and intermediate bulk gene expression levels compared to the control sets of cLADs and ciLADs (**Figure 3.7**), consistent with these regions being variably active within different cells.

I then explored whether these variable LADs were conserved in another human cell type. I found that the contact frequencies of bins containing variable LADs identified in HEK293T cells varied widely in KBM7 cells (**Figure 3.9a**), suggesting only a small subset of these LADs are variable in both cell types, consistent with prior observations that regions with intermediate contact frequencies are more likely to have different bulk DamID signals across cell types (Kind et al., 2015). Comparison of bulk RNA expression levels in bins that were classified as high, intermediate, or low contact frequency in each cell type corroborated the inverse relationship between single-cell contact frequency and bulk gene expression observed previously (Kind et al. 2015; **Figure 3.9b-h**; bulk RNA data analyzed by Annie Maslan). For example, as regions shift from intermediate contact frequencies to high contact frequencies in one cell type as compared to the other, we observe a corresponding decrease of gene expression (**Figure 3.9e,h**). These observations support the notion that the nuclear lamina serves as a dynamic regulatory element, not only between cell types but within a given cell type (Rooijers et al. 2019).

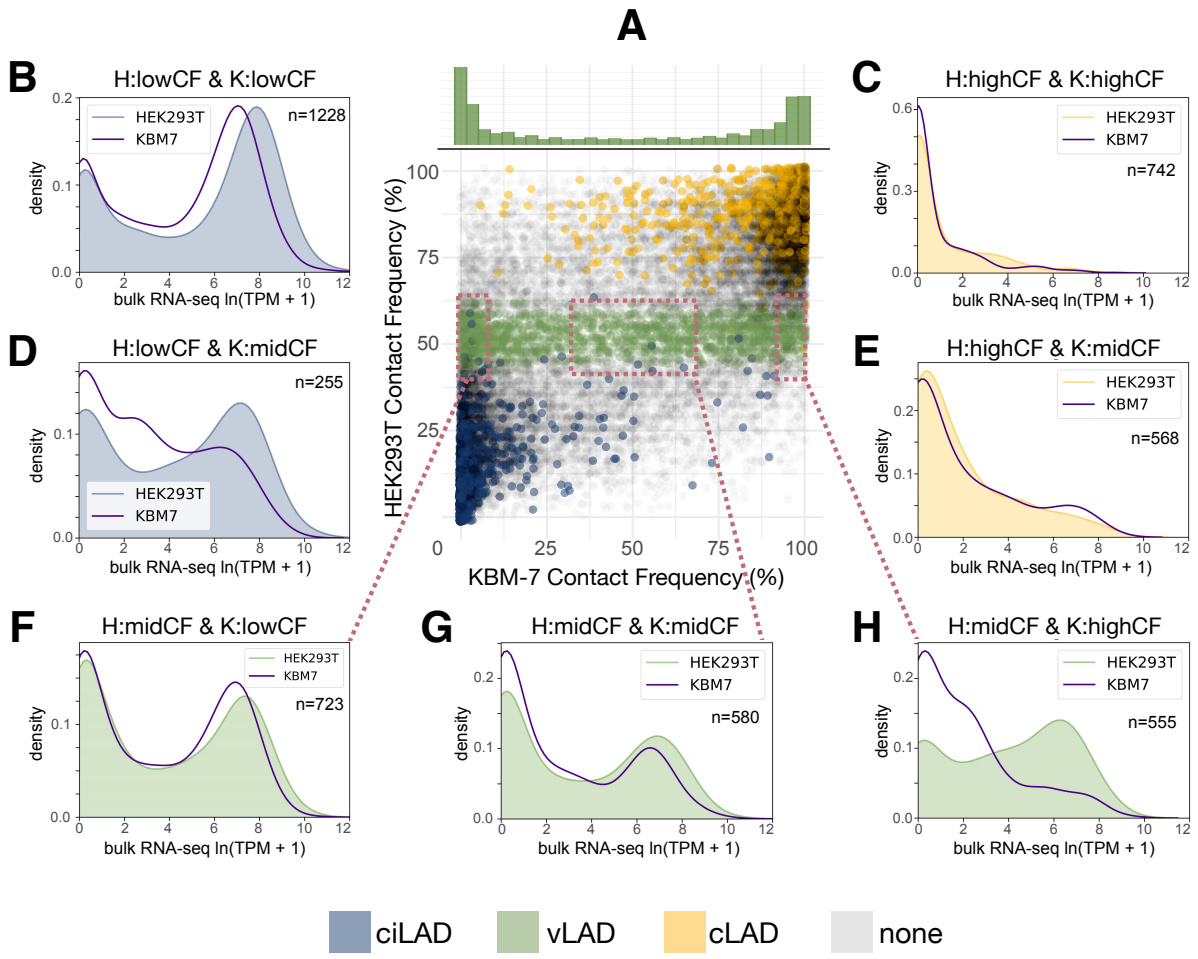


Figure 3.9. Comparing single-cell contact frequencies between cell types
(**A**) A scatterplot of the contact frequency estimates in HEK293T cells (this study) vs. KBM-7 cells (Kind et al., 2015) across all bins in the genome. Each point is colored if the corresponding bin belongs to the cLAD (gold), ciLAD (blue), or vLAD (green) sets defined in HEK293T. Above the scatterplot is a histogram showing the KBM-7 contact frequency (CF) distribution for all bins defined as vLADs in HEK293T, illustrating vLAD differences between cell types. (**B-H**) Density plots indicating the relative distributions of bulk RNA-seq coverage in each cell type, within bins classified as low CF (<5% CF, with high expression), middle CF (33-66% CF, with intermediate expression), or high CF (>90% CF, with low expression) in each cell type. For example, (D) shows the RNA-seq TPM distribution for 255 bins classified as low CF in HEK293T (higher expression) and as middle CF in KBM7 (lower expression).

Imaging LADs in the μDamID device using ^{m6}A-Tracer-NES

I aimed to use fluorescence microscopy to quantify the spatial distribution of LADs in the μDamID device prior to DamID processing. In the first batch of 18 cells, I imaged ^{m6}A-Tracer to identify the localization of lamina-interacting DNA in the nucleus. I selected Dam-LMNB1-expressing cells that had laminar rings consistent with effective LAD methylation, as well as one anomalous Dam-LMNB1 cell with high signal in the nuclear interior (**Figure 3.6c**). I also observed fairly uniform fluorescence across the nucleus in cells expressing untethered Dam. These imaging patterns were largely predictive of their respective sequencing coverage distributions (**Figure 3.6d**). However, this investigation revealed an important limitation of the ^{m6}A-Tracer technology, which is that the ^{m6}A-Tracer protein localizes to the nucleus even in cells expressing no Dam (**Figure 3.10a-b**). One consequence is that cells with Dam and cells without Dam are nearly indistinguishable (**Figure 3.10b**), and cells with overexpressed ^{m6}A-Tracer show high background fluorescence levels in the nuclear interior even when co-expressing Dam-LMNB1 (**Figure 3.10b**). The only way to prevent this background issue is to carefully tune the expression level of ^{m6}A-Tracer so that the copy number of ^{m6}A-Tracer proteins does not exceed the number of available methylated GATC sites. This tuning would have to occur separately for any new Dam fusion protein. In a heterogeneous expression system like the one used here, since ^{m6}A-Tracer and Dam are expressed from separate plasmids, only a small fraction of cells have the correct ratios of expression to produce sharp laminar rings with low background in the nuclear interior (**Figure 3.10b**).

No cryptic nuclear localization sequences were detected in ^{m6}A-Tracer (Methods), nor are human cells likely to contain any significant background levels of ^{m6}A without Dam (O'Brown et al. 2019). Instead, its default nuclear localization may arise from a weak interaction between genomic DNA and the DNA binding domain of ^{m6}A-Tracer, combined with the ability of ^{m6}A-Tracer to diffuse freely through nuclear pores given its small size (**Figure 3.10a**). Annie Maslan and I hypothesized that adding a Nuclear Export Signal (NES) to ^{m6}A-Tracer might overcome its weak affinity for DNA and keep any unbound copies of the protein sequestered in the cytoplasm. With Carolina Rios-Martinez, we found that the HIV-1 Rev NES sequence fused to either terminus resulted in robust localization of ^{m6}A-Tracer to the cytoplasm in cells not expressing Dam (**Figure 3.10**, **Figure 3.11**), and for downstream experiments we proceeded to use the C-terminal fusion, which we call ^{m6}A-Tracer-NES.

While the NES appears to prevent nonspecific ^{m6}A-Tracer interactions with DNA, it does not overcome on-target binding to Dam-methylated DNA. When Dam was co-expressed, the localization of ^{m6}A-Tracer-NES shifted almost entirely from the

cytoplasm to the nucleus (**Figure 3.10b**). When Dam-LMNB1 was co-expressed, ^{m6}A-Tracer-NES shifted to the nuclear lamina, with excess copies remaining in the cytoplasm in a subset of cells with especially high expression (**Figure 3.10b**, **Figure 3.11**). This shift in localization began within 2-3 hours of Dam-LMNB1 induction and produced visible rings in the majority of transfected cells within 5 hours (**Figure 3.11**). Because ^{m6}A-Tracer-NES only binds methylated sites in the nucleus, it solves two major problems: 1) ^{m6}A-Tracer fluorescence in the nucleus is no longer ambiguous and can be interpreted as a signal of methylation, and 2) high contrast between the nuclear lamina and the nuclear interior can be achieved for a much wider range of ^{m6}A-Tracer expression levels.

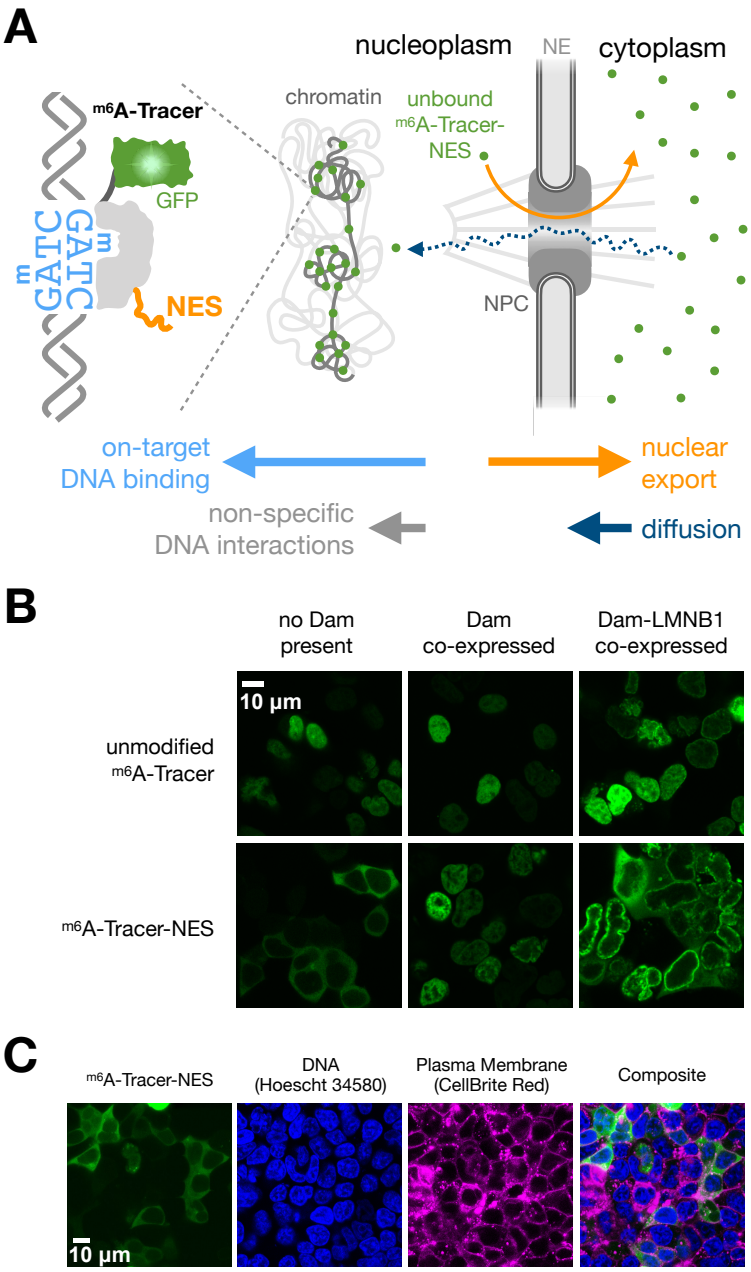


Figure 3.10. Improved imaging of protein-DNA interactions with ^{m6}A-Tracer-NES
(**A**) Illustration of potential mechanism by which ^{m6}A-Tracer-NES (^{m6}A-Tracer with a C-terminal HIV-1 Rev Nuclear Export Signal) reduces background fluorescence in the nucleus caused by non-specific DNA interactions, due to the relative rates of export, diffusion, and DNA binding (indicated by horizontal arrows). (**B**) Confocal images of ^{m6}A-Tracer-NES expressing cells co-stained with Hoescht 34580 to label DNA and CellBrite Red to label plasma membranes, showing cytoplasmic localization without Dam co-expression. (**C**) Confocal fluorescent microscope images revealing the different localization patterns of ^{m6}A-Tracer (Kind et al., 2013) with/without a NES, and with/without Dam or Dam-LMNB1 co-expression.

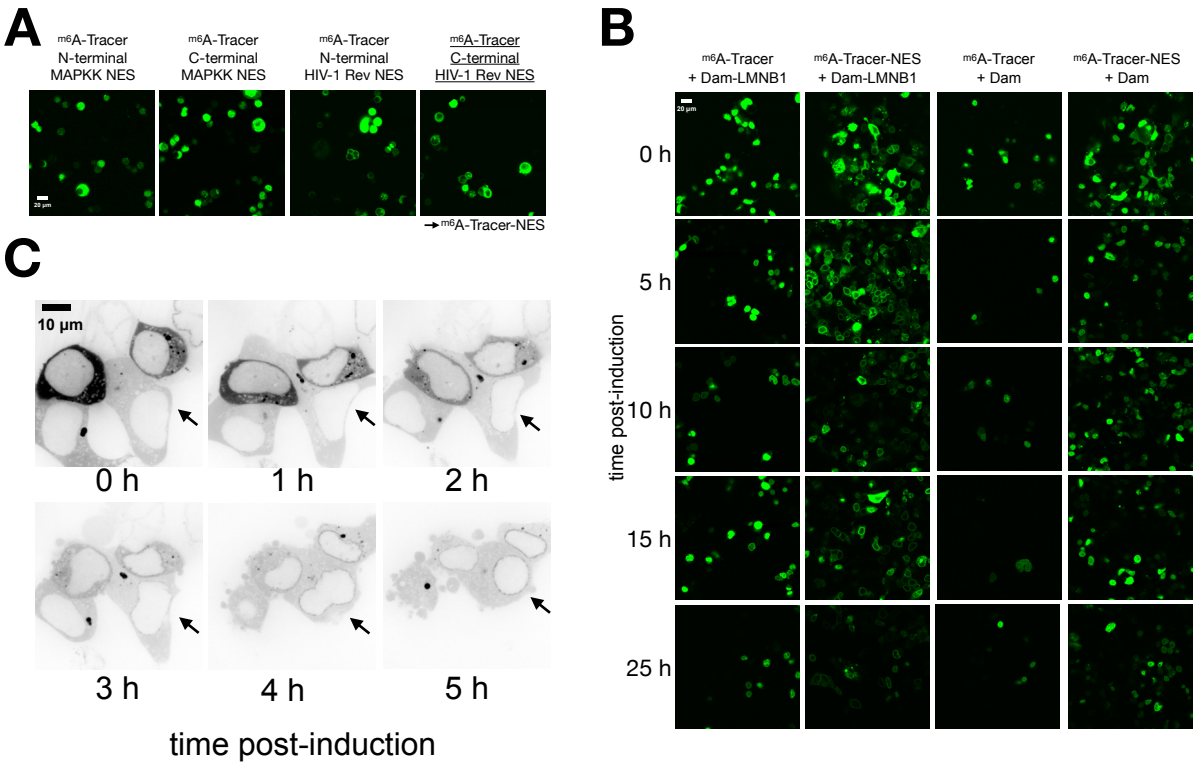


Figure 3.11. Additional characterization of ^{m6}A-Tracer-NES constructs

(**A**) Confocal microscope images showing the localization of ^{m6}A-Tracer fluorescence when fused to one of two different Nuclear Export Signals on either terminus, in cells not expressing Dam. The HIV-1 Rev NES worked on either terminus and the C-terminal fusion was selected for downstream experiments. (**B**) Time-lapse confocal images of ^{m6}A-Tracer-NES or unmodified ^{m6}A-Tracer fluorescence in different fields of cells, in cells co-expressing either Dam or Dam-LMNB1. Some nuclear localization is visible at time 0 in ^{m6}A-Tracer-NES + Dam cells, likely owing to leaky expression prior to induction. (**C**) Time-lapse confocal microscope images of ^{m6}A-Tracer-NES fluorescence in the same field of cells at timepoints after Dam-LMNB1 expression. An inverted lookup table is used, and an arrow points to the nucleus of the same cell, which begins to show laminar signal around 2h post-induction.

Joint imaging and sequencing analysis

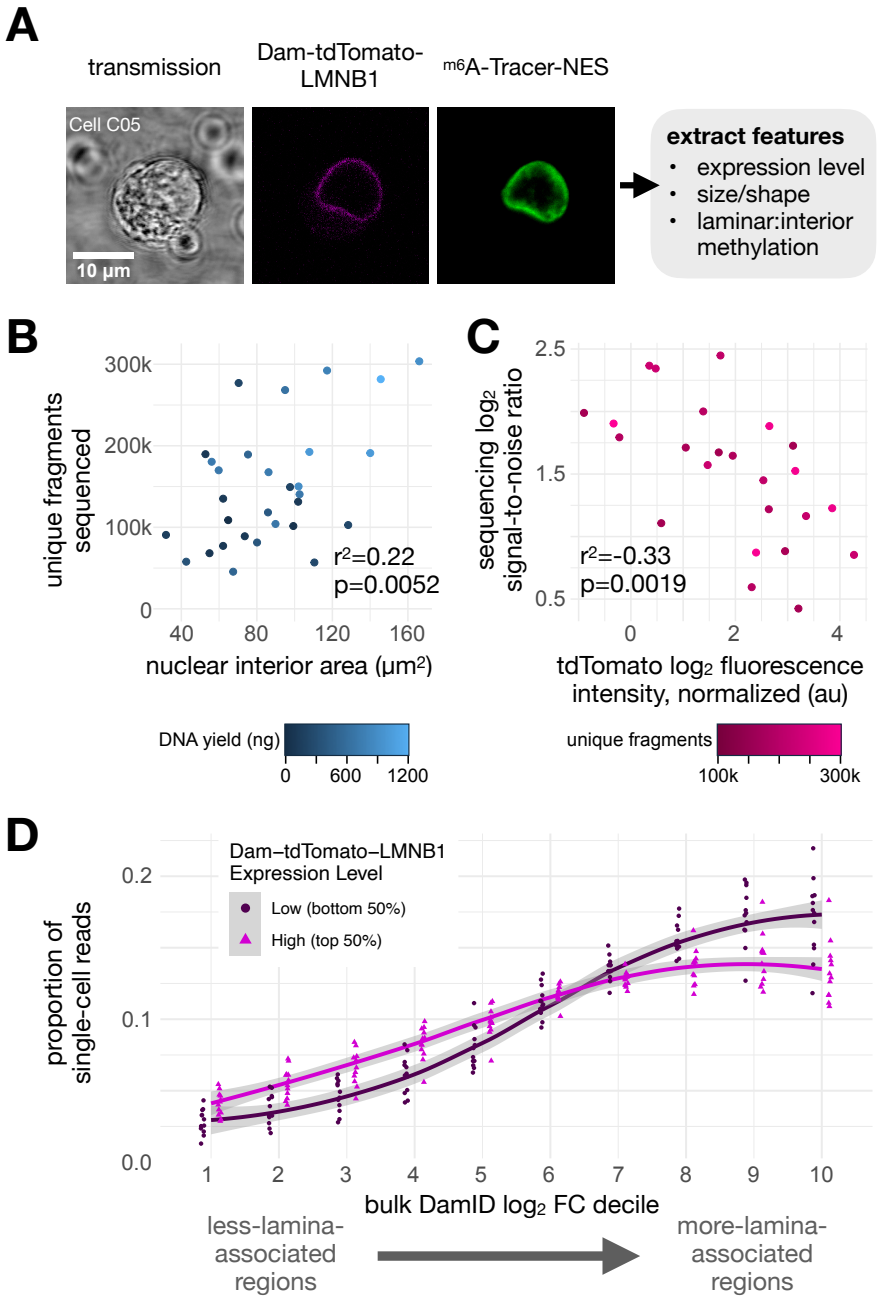
μDamID enables the joint analysis of the nuclear localization and sequence identity of protein-DNA interactions within each cell and between cells. Because the nuclear localization of LADs is well characterized, one could generate and test hypotheses about the sequencing data given the imaging data for each cell in this study. Because the first batch of cells expressed unmodified ^{m6}A-Tracer, it is possible that high ^{m6}A-Tracer expression could explain why some anomalous Dam-LMNB1 cells have high fluorescence in the nuclear interior (**Figure 3.6b**). Furthermore, because the first batch of cells lacked any direct readouts of Dam-LMNB1 expression levels, excessive Dam-LMNB1 expression could explain why some cells have high and unexpected sequencing coverage in ciLADs, leading to lower classification accuracy. To test this, in our second batch of cells I tagged the Dam-LMNB1 fusion protein with the red fluorescing protein tdTomato to enable monitoring of relative expression levels and the precise location of the nuclear lamina, and we used ^{m6}A-Tracer-NES to track the physical locations of lamina-associated DNA in the nucleus (**Figure 3.12**).

For each cell, I extracted a rich set of quantitative features from its images, including: nuclear lamina size/roundness, cell size/roundness, overall tdTomato intensity, and ^{m6}A-Tracer-NES intensity in each compartment. I then compared these imaging features to sequencing features for each cell: library DNA yield, unique fragment number, signal-to-noise ratio and accuracy on control sets, and raw coverage distribution in genomic bins ranked by lamina association from bulk data. Several strong associations stood out from the data (**Figure 3.13**). Firstly, I found that cells with larger nuclei tended to yield more DNA in their libraries, and resulted in more unique fragments sequenced, indicating greater library complexity (**Figure 3.12b**). This matches expectations, given that larger nuclei in this asynchronous, heterogeneous population are likely to have more DNA, either due to ploidy differences or cell cycle phase.

Secondly, I found that, among cells with high library complexity (over 100,000 unique fragments), cells with greater expression of Dam-tdTomato-LMNB1 showed diminished sequencing signal-to-noise ratios in cLADs vs ciLADs (**Figure 3.12c**) and generally showed higher coverage in less lamina-associated regions of the genome (**Figure 3.12d**). This is consistent with our hypothesis, as higher Dam fusion protein expression is expected to produce higher background methylation that is not specific to the protein-DNA interaction of interest. Notably, among Dam-LMNB1-expressing cells I did not find a strong association between ^{m6}A-Tracer-NES signal in the nuclear interior and background methylation (**Figure 3.13**). This may be because variation in expression of ^{m6}A-Tracer obscures biological variation in methylation at the lamina.

Imaging data did reveal, however, that two cells without bright laminar rings produced low-complexity sequencing libraries (cells D05 and D09, **Figure 3.3**), and these would be difficult to filter out by other sorting approaches and would lead to low-quality outliers in DamID sequencing data. This series of measurements serves as a proof of principle that μDamID can be used to sort cells based on visual phenotypes that are correlated with sequencing measurements. Here I use this capability to predict sequencing information content from imaging phenotypes in single cells.

Figure 3.12. Joint image and sequence analysis



Legend for Figure 3.12. Joint image and sequence analysis

(**A**) An example of raw imaging data acquired for a single cell co-expressing ^{m6}A-Tracer-NES and Dam-tdTomato-LMNB1 used for imaging feature extraction. Integrated tdTomato intensity in the cell serves as a measure of relative Dam-LMNB1 expression between cells. (**B**) Scatterplot of the number of unique fragments covered by sequencing data for each cell compared to its nuclear area determined from its imaging data, colored by the library DNA yield for that cell (n=30 batch 2 cells with detectable tdTomato signal). r^2 and p(slope>0) are provided from an ordinary least squares linear model. (**C**) Scatterplot of the sequencing signal:noise ratio for each cell (computed using coverage in control regions) compared to its relative Dam-tdTomato-LMNB1 expression (determined by imaging), colored by the number of unique fragments for that cell (n=24 batch 2 cells with detectable tdTomato signal and >100k unique fragments). r^2 and p(slope<0) are provided as in (B). (**D**) a comparison of the distribution of raw single-cell sequencing coverage across deciles of increasing bulk DamID signal in the genome, for the group of 12 cells with the highest (magenta) or lowest (dark red) Dam-tdTomato-LMNB1 expression levels. Values for individual cells are plotted as points in each decile, and loess curves are overlaid with 99% confidence interval ribbons in gray. Higher coverage in the left-hand side of the plot is consistent with greater background methylation.

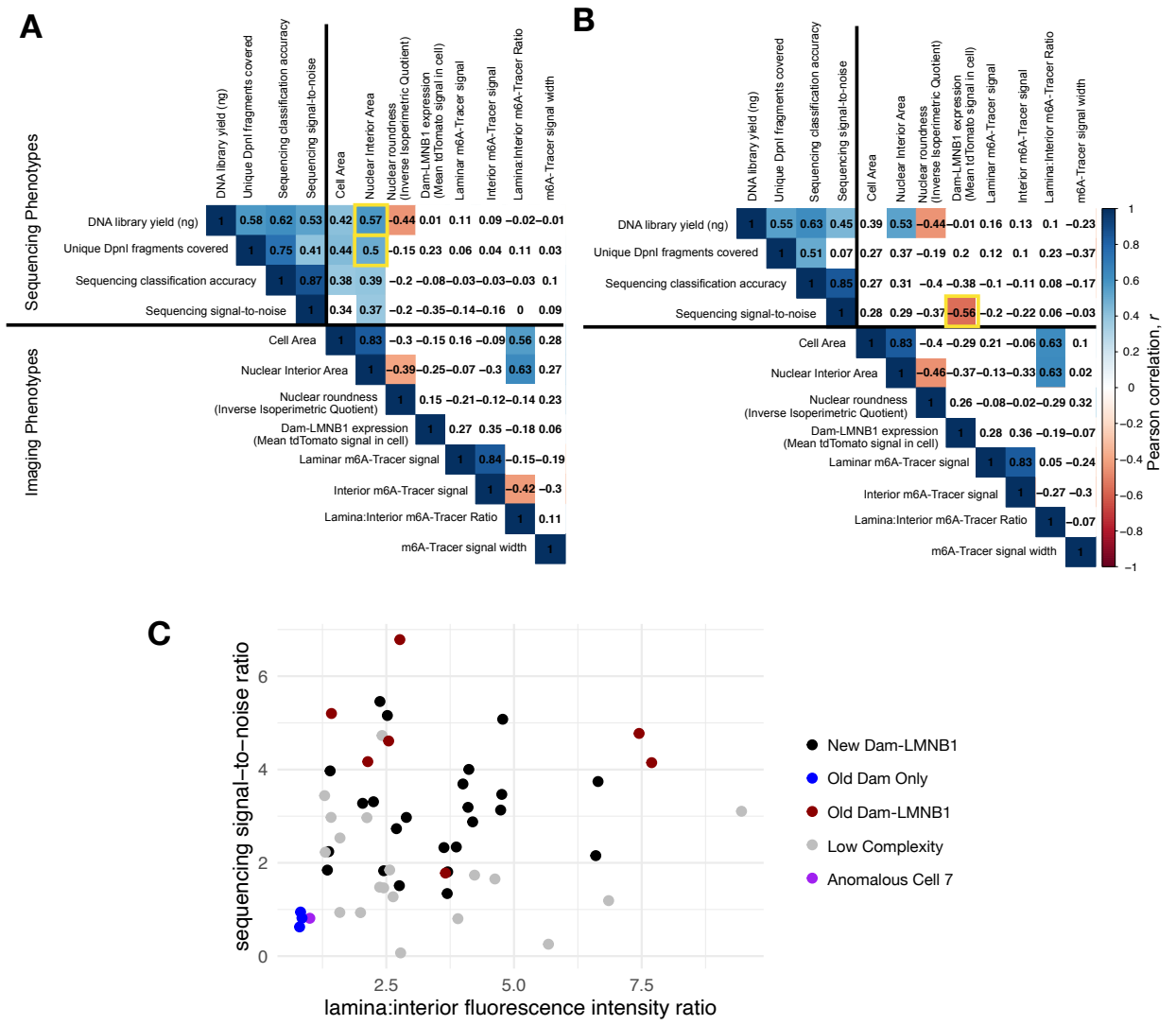


Figure 3.13. Correlations of sequencing and imaging phenotypes

(**A**) A correlation matrix showing the relationships between imaging measures (see Methods) and sequencing measures for 30 batch 2 cells with nuclear areas definable by tdTomato imaging. Correlations with $p<0.05$ are shaded white, while significant correlations are colored by the strength of their positive (blue) or negative (red) correlation. Associations that were further explored in **Figure 3.12** are highlighted in yellow. (**B**) As in (A) but after filtering cells to remove those with <100k unique fragments, which confound estimates of classification accuracy. (**C**) Imaging ratios are reported for each cell as in **Figure 3.6**. Dark blue points represent Dam-only cells, and dark red points and black points represent Dam-LMNB1 cells from batch 1 and batch 2, respectively. The anomalous Dam-LMNB1 cell #007 (shown in **Figure 3.6**) is highlighted in purple. Cells with fewer than 100k unique fragments are grayed out.

Discussion and future directions

Here I have demonstrated the use of μDamID, an integrated microfluidic device for single-cell isolation, imaging, and sorting, followed by DamID. This system enables the acquisition of paired imaging and sequencing measurements of protein-DNA interactions within single cells, giving a readout of both the 'geography' and identity of these interactions in the nucleus. Specifically, I tested the device by mapping well-characterized interactions between DNA and proteins found at the nuclear lamina, providing a measure of genome regulation and 3D chromatin organization within the cell, and recapitulating similar maps in other cell types. We also improved the method of imaging protein-DNA interactions with ^{m6}A-Tracer by attaching a nuclear export signal. This modification greatly reduces background fluorescence due to nonspecific interactions with unmethylated DNA, providing a more universal readout of the ^{m6}A methylation status of the nucleus. ^{m6}A-Tracer-NES will allow for more sensitive imaging of other classes of protein-DNA interactions in the nucleus, and it could potentially also be utilized in synthetic genetic and epigenetic circuits (Park et al. 2019) to reduce off-target effects, or to serve as a nuclear localization switch.

I am now working with collaborators to apply μDamID to study protein-DNA interactions that are critical for the process of DNA repair. In the future, I hope to apply μDamID or similar methods to study protein-DNA interactions in single meiotic cells. Meiosis is fundamentally a single-cell process, in which a single cell produces four unique gametes. The uniqueness of these gametes owes in large part to the unique combination of protein-DNA interactions that mediate the process of meiotic recombination in each cell. I studied meiotic recombination extensively in my previous work (Altemose et al. 2017, Davies et al. 2016, R. Li et al. 2019), so I know firsthand how much this field could benefit from the ability to image and map protein-DNA interactions in single cells in this system. μDamID is also well-suited to studying abnormalities in nuclear morphology, such as micronuclei or the changes associated with aging and diseases like progeria (Karoutas & Akhtar 2021).

Detailed Materials and Methods

A full list of key resources attached as **Appendix 1**. Any updates to the protocol will be posted to streetslab.com.

Harvesting and imaging cells

Cells were harvested 72 hours after transfection. 20 hours before harvesting, the media was replaced and 0.5 µl Shield-1 ligand (0.5 mM stock, Takara Bio USA, Inc., Mountain View, CA) was added to each well to stabilize protein expression. Cells transfected with Dam-LMNB1 were inspected by fluorescence microscopy to look for the characteristic signal at the nuclear lamina, indicating proper expression and protein activity. To harvest the cells and prepare them for loading on the device, the cells were washed with PBS, then incubated at room temperature with 1X TrypLE Select (ThermoFisher Scientific, Waltham, MA) for 5 minutes to dissociate them from the plate. Cells were pipetted up and down to break up clumps, then centrifuged at 300xg for 5 minutes, resuspended in PBS, centrifuged again, and resuspended in 500 µl Pick Buffer (50 mM Tris-HCl pH 8.3, 75 mM KCl, 3 mM MgCl2, 137 mM NaCl), achieving a final cell concentration of roughly 500,000 cells per ml. Cells were passed through a 40 µm cell strainer before loading onto the device.

For batch 2 cells, 10 confocal z slices were taken for each cell, and the slice with the largest nuclear perimeter was selected for image processing (see **Chapter 2** for microscope configuration. The 3 cells expressing Dam-only that were sequenced in this study were imaged with a widefield CCD camera. Other Dam-only cells were imaged with confocal microscopy and showed similar relatively homogenous fluorescence throughout the nucleus, and never the distinct 'ring' shape found in Dam-LMNB1 expressing cells (Kind et al. 2013; **Figure 3.4a**). No image enhancement methods were used prior to quantitative image processing. Images in **Figure 3.6** have been linearly thresholded to diminish background signal.

Quality control, library prep, and sequencing

Samples were diluted to 10 µl total volume and two replicates of qPCR were performed using the DamID PCR primer to estimate DNA quantities relative to the pick-buffer-only negative control (1 µl DNA per replicate in 10 µl reaction volume). I also used 1 µl of sample to measure DNA concentration using a Qubit fluorometer with a High-Sensitivity DNA reagent kit (quantitative range 0.2 ng – 100 ng; ThermoFisher Scientific). Samples with the lowest Ct values and highest Qubit DNA measurements were selected for library preparation and sequencing. Library preparation was carried out using an NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB E7645) with dual-indexed multiplex i5/i7 oligo adapters. Size selection was not performed; PCR was

carried out for 9 cycles. Libraries were quantified again by Qubit and size profiled on a TapeStation 4200 with a D5000 HS kit (Agilent, Santa Clara, CA), then mixed to achieve equimolar amounts of each library. DNA was sequenced on an Illumina MiniSeq with a 150-cycle high output kit, to produce paired 75 bp reads, according to manufacturer instructions (Illumina, San Diego, CA). Roughly 13 million read pairs were obtained for batch 1 cells. For batch 2, I performed library preparation using an NEBnext UltraII FS kit (NEB E7805) and obtained 200 million total read pairs with an Illumina NextSeq 550 High Output run, to guarantee sequencing of nearly the full available library complexity.

Bulk DamID, *Credit: Annie Maslan*
Genomic DNA was isolated from ~3.7 x 10^6 transfected HEK293T cells using the DNeasy Blood & Tissue kit (Qiagen) following the protocol for cultured animal cells with the addition of RNAse A. The extracted gDNA was then precipitated by adding 2 volumes of 100% ethanol and 0.1 volume of 3 M sodium acetate (pH 5.5) and storing at -20 °C for 30 minutes. Next, centrifugation for 30 minutes at 4 °C, >16,000 x g was performed to spin down the gDNA. The supernatant was removed, and the pellet was washed by adding 1 volume of 70% ethanol. Centrifugation for 5 minutes at 4 °C, >16,000 x g was performed, the supernatant was removed, and the gDNA pellets were air-dried. The gDNA was dissolved in 10 mM Tris-HCl pH 7.5, 0.1 mM EDTA to 1 µg/µl, incubating at 55 °C for 30 minutes to facilitate dissolving. The concentration was measured using Nanodrop.

The following DpnI digestion, adaptor ligation, and DpnII digestion steps were all performed in the same tube. Overnight DpnI digestion at 37 °C was performed with 2.5 µg gDNA, 10 U DpnI (NEB), 1X CutSmart (NEB), and water to 10 µl total reaction volume. DpnI was then inactivated at 80 °C for 20 minutes. Adaptors were ligated by combining the 10 µl of DpnI-digested gDNA, 1X ligation buffer (NEB), 2 µM adaptor dsAdR, 5 U T4 ligase (NEB), and water for a total reaction volume of 20 µl. Ligation was performed for 2 hours at 16 °C and then T4 ligase was inactivated for 10 minutes at 65 °C. DpnII digestion was performed by combining the 20 µl of ligated DNA, 10 U DpnII (NEB), 1X DpnII buffer (NEB), and water for a total reaction volume of 50 µl. The DpnII digestion was 1 hour at 37 °C followed by 20 minutes at 65 °C to inactivate DpnII.

Next, 10 µl of the DpnII-digested gDNA was amplified using the Takara Advantage 2 PCR Kit with 1X SA PCR buffer, 1.25 µM Primer Adr-PCR, dNTP mix (0.2 mM each), 1X PCR advantage enzyme mix, and water for a total reaction volume of 50 µl. PCR was performed with an initial extension at 68 °C for 10 minutes; one cycle of 94 °C for 1 minute, 65 °C for 5 minutes, 68 °C for 15 minutes; 4 cycles of 94 °C for 1 minute, 65 °C

for 1 minute, 68 °C for 10 minutes; 21 cycles of 94 °C for 1 minute, 65 °C for 1 minute, 68 °C for 2 minutes. Post-amplification DpnII digestion was performed by combining 40 µl of the PCR product with 20 U DpnII, 1X DpnII buffer, and water to a total volume of 100 µl. The DpnII digestion was performed for 2 hours at 37 °C followed by inactivation at 65 °C for 20 minutes. The digested product was purified using QIAquick PCR purification kit. The purified PCR product (1 µg brought up to 50 µl in TE) was sheared to a target size of 200 bp using the Bioruptor Pico with 13 cycles with 30"/30" on/off cycle time. DNA library preparation of the sheared DNA was performed using NEBNext Ultra II DNA Library Prep Kit for Illumina using AMPure XP beads (Beckman Coulter Life Sciences, Indianapolis, IN).

Bulk DamID, comparing Dam mutants, *Credit: Annie Maslan*
Bulk DamID for comparing the wild-type allele and V133A mutant allele was performed as outlined in the Bulk DamID section above with the following modifications. Genomic DNA was extracted from ~ 2.4 x 10^5 transfected HEK293T cells. A cleanup before methylation-specific amplification was included to remove unligated Dam adapter before PCR. The Monarch PCR & DNA Cleanup Kit with 20 µl DpnII-digested gDNA input and an elution volume of 10 µl was used. Shearing with the Bioruptor Pico was performed for 20 total cycles with 30"/30" on/off cycle time. Paired-end 2 x 75 bp sequencing was performed on an Illumina NextSeq with a mid output kit. Approximately 3.8 million read pairs per sample were obtained.

Bulk RNA-seq, *Credit: Annie Maslan*
RNA was extracted from ~1.9 x 10^6 transfected HEK293T cells using the Rneasy Mini Kit from Qiagen with the QIAshredder for homogenization. RNA library preparation was performed using the NEBNext Ultra II RNA Library Prep Kit for Illumina with the NEBNext Poly(A) mRNA Magnetic Isolation Module. Paired-end 2 x 150 bp sequencing for both DamID-seq and RNA-seq libraries was performed on 1 lane of a NovaSeq S4 run. Approximately 252 million read pairs were obtained for each DamID-seq sample, and roughly 64 million read pairs for each RNA sample.

^{m6}A-Tracer-NES, *Additional credit: Carolina Rios-Martinez & Annie Maslan*
To reduce background fluorescence due to ^{m6}A-Tracer, we fused its N or C terminus to one of two different nuclear export signals (NES): HIV-1 Rev (LQLPPLERLTLD) or MAPKK (LQKKLEELEL) (Kakar et al. 2007). We compared the localization of each of the 4 resulting constructs by imaging HEK293T cells transiently transfected with ^{m6}A-tracer-NES by itself or with Dam. Negative controls included transfection with unmodified ^{m6}A-Tracer only or Dam only, and no transfection. The MAPKK NES did not appreciably reduce nuclear localization of ^{m6}A-tracer-NES in the absence of Dam (**Figure 3.11**).

However, the HIV-1 Rev NES, in either the N- or C-terminal configuration, showed significant improvement in localizing signal to the cytoplasm in the absence of Dam, while permitting nuclear localization in the presence of Dam (**Figure 3.10**, **Figure 3.11**). We proceeded to use the C-terminal HIV-1 Rev ^{m6}A-Tracer construct for downstream experiments. Co-transfection with Dam-LMNB1 resulted in a greater proportion of transiently transfected cells having visible laminar rings than with unmodified ^{m6}A-Tracer. Timelapse imaging of the same field of Dam-LMNB1 + ^{m6}A-Tracer-NES cells over time or different fields at each timepoint (**Figure 3.11**) demonstrated that laminar rings become visible within 2-3 hours and reach full intensity around 5 hours after Dam-LMNB1 induction with Shield-1 ligand. To test the possibility that unmodified ^{m6}A-Tracer localizes to the nucleus due to a cryptic Nuclear Localization Signal, I searched for NLS motifs using NLSdb (Bernhofer et al. 2018) but found no matches.

Quantification and statistical analysis

Bulk RNA-seq

Adapters were trimmed using trimmomatic (v0.39; Bolger et al., 2014; ILLUMINACLIP:adapters-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36, where adapters-PE.fa is:
>PrefixPE/1
TACACTCTTTCCCTACACGACGCTCTTCCGATCT
>PrefixPE/2
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT).

Transcript quantification was performed using Salmon (Patro et al. 2017) with the GRCh38 transcript reference. Differential expression analysis was performed using the voom function in limma (Ritchie et al. 2015). Differential expression was called based on logFC significantly greater than 1 and adjusted p-value < 0.01.

For KBM7 bulk gene expression analysis, publicly available single-end RNA sequencing data (SRA accession SRP044391, Essletzbichler et al. 2014) from two replicates were processed. For adapter trimming, trimmomatic was used in the SE mode with the adapter file ILLUMINACLIP:TruSeq3-SE. All other trimmomatic parameters were the same as were used in the HEK293T RNA-seq data processing, and Salmon was used for transcript quantification in single-end mode. Credit: Annie Maslan.

Bulk and single-cell DamID

Bulk and single-cell DamID reads were demultiplexed using Illumina's BaseSpace platform to obtain fastq files for each sample. DamID and Illumina adapter sequences were trimmed off using trimmomatic (v0.39; Bolger et al. 2014; ILLUMINACLIP:adapters-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:20, where adapters-PE.fa is:

>PrefixPE/1
TACACTCTTTCCCTACACGACGCTCTTCCGATCT
>PrefixPE/2
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
>Dam
GGTCGCGGCCGAGGA
>Dam_rc
TCCTCGGCCGCGACC

Trimmed reads were aligned to a custom reference (hg38 reference sequence plus the Dam-LMNB1 and ^{m6}A-Tracer plasmid sequences) using BWA-MEM (v0.7.15-r1140, (Heng Li 2013)). Alignments with mapping quality 0 were discarded using samtools (v1.9, (H. Li et al. 2009)). The hg38 reference sequence was split into simulated DpnI digestion fragments by reporting all intervals between GATC sites (excluding the GATC sites themselves), yielding 7180359 possible DpnI fragments across the 24 chromosome assemblies. The number of reads overlapping each fragment was counted using bedtools (v2.28; (Quinlan & Hall 2010)). For single-cell data, the number of DpnI fragments with non-zero coverage was reported within each non-overlapping bin in the genome (28163 total 100 kb bins, after excluding unmappable regions with zero coverage in any cell). For bulk data, the number of read pairs overlapping each 100 kb bin was reported. The same exact pipeline was applied to the raw reads from Kind et al. 2015 (GEO accession GSE69423). RefSeq gene positions were downloaded from the UCSC Genome Browser and counted in each bin. For bulk data, Dam-LMNB1 vs DamOnly enrichment was computed using Deseq2 in each 100 kb bin (Love et al. 2014). For single-cell data, the expected background coverage in each bin was computed as $n(m/t)$, where n is the number of unique fragments sequenced from that cell, m is the number of bulk Dam-only read pairs mapping to that bin, and t is the total number of mapped bulk Dam-only read pairs. Single-cell normalization was computed either as a ratio of observed to expected coverage (for browser visualization and comparison to bulk data), or as their difference (for classification and coverage distribution plotting). Positive and negative control sets of cLAD and ciLAD bins were defined under the assumption that genomic regions that have high bulk DamID signal and that are lamina associated across many cell types are likely to be in contact with the lamina in the vast majority of single cells, which is supported by previous scDamID

data (Kind et al. 2015; **Figure 3.4**). Specifically, we defined them as bins with a bulk Dam-LMNB1:Dam-only Deseq2 p-value smaller than 0.05/28760, that intersected published cLADs and ciLADs in other cell lines (Lenain et al. 2017), and that were among the top 1200 most differentially enriched bins in either direction (positive or negative log fold change for cLADs and ciLADs, respectively). Normalized coverage thresholds for LAD/iLAD (i.e. contact vs. no contact) classification were computed for each cell to maximize accuracy on the cLAD and ciLAD control sets. To examine whether using the full control sets to set thresholds and define classification error was resulting in substantial overfitting, we split the control sets into training and test sets for threshold setting and accuracy determination, respectively, and only observed a 0.7% mean drop in accuracy relative to using the full sets. Signal-to-noise ratios were computed for each cell using the normalized coverage distributions in the cLAD and ciLAD control sets as $(\mu_{cLAD} - \mu_{ciLAD})/\sigma_{ciLAD}$. For most downstream analyses, we chose to exclude 20 cells with fewer than 100,000 unique covered fragments, which includes cells with poor laminar rings and lower DNA yields (**Figure 3.4** and **Figure 3.5**). For any given application of μDamID, this threshold will depend on the level of noise due to background methylation in the biological system being used, which is expected to depend in part on the expression level of the Dam fusion protein. In a transiently transfected cell population, this expression level is expected to vary widely, which motivated the use of data to explore this as a cause of variable classification accuracy between cells. The remaining 31 Dam-LMNB1 cells had a median classification accuracy of 90% (range 74%-98%). Bulk analysis credit: Annie Maslan.

Calling vLADs

Variable LADs were defined as bins called as LADs in 33-66% of cells and conservatively filtered to remove regions resulting from sampling error. This was done by computing, for each bin and for each cell, the probability that the true sample contact frequency lies outside the interval (33%, 66%). I estimated this probability using a Poisson-binomial distribution, a generalization of the binomial distribution allowing individual samples to have varying success probabilities. Specifically, each bin in the genome has k cells called as LADs and n-k cells called as iLADs, with n=31 in this study. For the k LADs I generated a vector of k false-positive probabilities, with each probability estimated as the fraction of negative-control ciLADs with coverage greater than the observed coverage in that bin. I used this probability vector to parameterize a Poisson-binomial distribution with k draws, providing the distribution of false-positive calls in the bin. I repeated this for the n-k iLAD bins, with each false-negative probability estimated as the fraction of positive-control cLADs with coverage lower than the observed coverage in that bin. These two distributions were combined into a single density by reflecting the false-positive distribution about the y axis, scaling each one

according to its mean, and adding *k* to produce the plots in **Figure 3.8**). Only regions with $p<10^{-3}$ for both tails were called as variable LADs. I then generated 10,000 samples of the sample contact frequency, *c*, from this distribution and used each one to generate a single binomial (n=31, $p=c/31$) sample, generating a combined measurement and sampling distribution with greater variance than either alone (**Figure 3.7**), from which I generated 95% confidence intervals for the population contact frequency in each bin (**Figure 3.7**). Statistical analyses and plots were made in R (v4.0.0) using the ggplot2 (v3.3.0), gplots (v3.0.3), colorRamps (v2.3), reshape2 (v1.4.4), ggextra (v0.9) and poisbinom (v1.0.1) packages. Browser figures were generated using the WashU Epigenome Browser (D. Li et al. 2019).

Image processing

Images were processed in R (v4.0.0) and plots were produced using the reshape2 (v1.4.3), SDMTools (v1.1-221.1), spatstat (v1.59-0), magick (v2.0), and ggplot2 (v3.3.0) packages. Grayscale images were converted to numeric matrices and edge detection was performed using Canny edge detection using the image_canny function in magick, varying the geometry parameters manually for each cell. The center of mass of all edge points was obtained, and all edge points were plotted in Cartesian coordinated with this center of mass as the origin. Noise was removed by removing points with a nearest neighbor more than 2 microns away. Edge point coordinates were converted to polar coordinates, and the farthest points from the origin in each 10 degree arc were reported. Within each 10 degree arc, all pixel intensities from the original image within the edges of the nucleus were reported as a function of their distance from the farthest edge point in that arc to make **Figure 3.6c**. For each cell a loess curve (span 0.3) was fitted to the data after subtracting the minimum intensity value within 3.5 microns of the edge. The Lamina:Interior ratio was computed as the ratio of mean intensity of pixels within 1 micron of the edge to the mean intensity of pixels more than 3.5 microns from the edge, after subtracting the minimum value of the loess curve for that cell. To provide an additional metric, I computed the distance from the laminar edge where the fluorescence intensity decays to 10% of the peak laminar intensity. A nuclear mask was created by drawing a polygon with vertices as the farthest point in each arc from the center of mass, and a similar cell mask polygon was generated using the transmission image—these masks enabled the computation of cell and nuclear area and perimeter as well as the mean fluorescence intensity in different compartments of the cell. For batch 2 cells, I used the tdTomato image for each cell to infer the laminar boundary by thresholding the images iteratively and performing morphological dilations until a closed loop formed, providing a new nuclear mask. As a measure of nuclear roundness, I computed the inverse isoperimetric quotient for each cell: the ratio of the area of a circle with the equivalent perimeter to

the observed area of the nucleus, computed as $P^2/(4\pi A)$ (domain 1 to infinity). All image processing steps are deterministic and reproducible, with all R code and necessary metadata files published in our github repository.

Co-author contributions

Nicolas Altemose and Aaron Streets conceived of and designed the study and the microfluidic device. Nicolas Altemose and Andre Lai fabricated and optimized operation of the device. Annie Maslan performed bulk cell experiments and data processing, Carolina Rios-Martinez performed ^{m6}A-Tracer-NES experiments with supervision from Annie Maslan and Nicolas Altemose, and Nicolas Altemose performed all other experiments, analysis, and pneumatic/thermoelectric hardware construction. Jonathan A. White developed the microfluidic control platform and thermal cycling software, with minor modifications by Nicolas Altemose. Nicolas Altemose wrote the manuscript with contributions from Annie Maslan, Carolina Rios-Martinez, and Aaron Streets. Aaron Streets supervised the study.

Materials, Data, and Code Availability

Materials Availability: Plasmids generated in this study have been deposited to Addgene (https://www.addgene.org/browse/article/28211957/).

Data and Code Availability: The sequencing data generated during this study are available at GEO (accession GSE156150). The imaging data generated during this study are available at FigShare: https://doi.org/10.6084/m9.figshare.12798158. Analysis code, control software, device design files, and plasmid sequences are freely available for download on GitHub: https://github.com/altemose/microDamID. Source data for bulk KBM-7 RNA-seq were obtained from SRA (accession SRP044391), and source data for KBM-7 scDamID were obtained from GEO (accession GSE69423).

This work is now published as Altemose et al. 2020:

Altemose N, Maslan A, Rios-Martinez C, Lai A, White JA, & Streets A. (2020). μDamID: a microfluidic approach for joint imaging and sequencing of protein-dna interactions in single cells. *Cell Systems*, 11(4), 354-366.e9. https://doi.org/10.1016/j.cels.2020.08.015

Chapter 4

Development, optimization, and validation of DiMeLo-seq, a single-molecule method for mapping protein-DNA interactions *in situ*

Aims & overview

Measuring protein-DNA interactions is key to understanding how the information coded in DNA is brought to life. Making these measurements with increased sensitivity has proven useful for mapping protein-DNA interactions in single cells (**Chapter 3**), but these efforts also highlight an additional blind spot for protein-DNA interaction mapping: the repetitive regions of the genome. Due to the lack of GATC sites in many repetitive regions (**Figure 2.1**, Sobecki et al. 2018), even with the forthcoming complete Telomere-to-Telomere genome assembly, these GATC-poor repetitive regions cannot be probed by conventional DamID. Other protein-DNA mapping approaches, such as MadID, ChIP-seq, and CUT&RUN, can recover information from these repetitive regions, but they produce short sequencing reads that cannot be mapped unambiguously within highly repeated sequences. This underlines the need for a technology that is able to map protein-DNA interactions using long sequencing reads, which can map to and cover repetitive regions more comprehensively.

To address this need, I devised and piloted the idea of using long-read sequencing to directly read out methyladenines deposited by an adenine methyltransferase (Dam or EcoGII) fused to a protein of interest (such as centromere protein C) and, with Annie Maslan, I extended the idea to enable this sort of mapping *in situ*, as in the pA-DamID method (Schaik et al. 2020). This makes the approach much more versatile and widely useful, as it can map post-translational marks like histone modifications and can be used in primary tissue samples, unlike conventional *in vivo* DamID methods. We then joined forces with Professor Aaron Straight's research group at Stanford, who were working on a similar idea. Our collective efforts have produced a new method for mapping protein-DNA interactions, which we call **DiMeLo-seq**, for **Di**rected **Me**thylation with **Lo**ng-read **seq**uencing (**Figure 4.1**). The name pays tribute to my Bolivian heritage: in Spanish, *dímelo* means "tell me it."

Here, I describe my conception of the idea for DiMeLo-seq and early results from *in vivo* expression experiments, followed by the development of a rapid pipeline for evaluating protocol performance with Annie Maslan. Next, I discuss the results of a long process of *in situ* protocol optimization using our pipeline, in collaboration with Owen Kabnick Smith, Dr. Kousik Sundararajan, and Rachel Brown in Aaron Straight's group at Stanford. I also describe an approach I developed to enrich repetitive centromeric sequences prior to sequencing, along with early results and immediate plans for the application of these methods to probe important biological questions in the formerly missing regions of the human genome.

DiMeLo-seq is useful for mapping protein-DNA interactions in repetitive regions, but it also provides additional single-molecule information that can be leveraged in several ways. For example, endogenous CpG methylation can be jointly measured along with protein-DNA interaction sites on the same single molecules of DNA. This is useful when studying how DNA methylation and protein binding interact, for example when DNA methylation abolishes the binding of certain transcription factors. Additionally, because methyltransferases favor accessible linker DNA between nucleosomes, nucleosome positioning can be inferred based on the density of methylation marks, as with existing long-read accessibility measurement technologies (Abdulhay et al. 2020, I. Lee et al. 2020, Shipony et al. 2020, Stergachis et al. 2020, Y. Wang et al. 2019). Because it is an amplification-free method, it also allows one to linearly infer the frequency of binding of a protein at a particular site in a population of cells. Furthermore, since the enzyme reach is on the order of 100-200 bp, and reads can regularly be as long as hundreds of kb, we can infer multiple protein-DNA binding events on single molecules. This is useful for exploring the density of a protein along a stretch of chromatin, or the exact joint binding profile of proteins to proximal sites. It also lends itself to examining the joint distribution of multiple proteins, each fused to a distinct DNA-modifying enzyme, on the same long single molecule of DNA. Finally, it is also feasible to extend the method to look at DNA-DNA or RNA-DNA interactions.

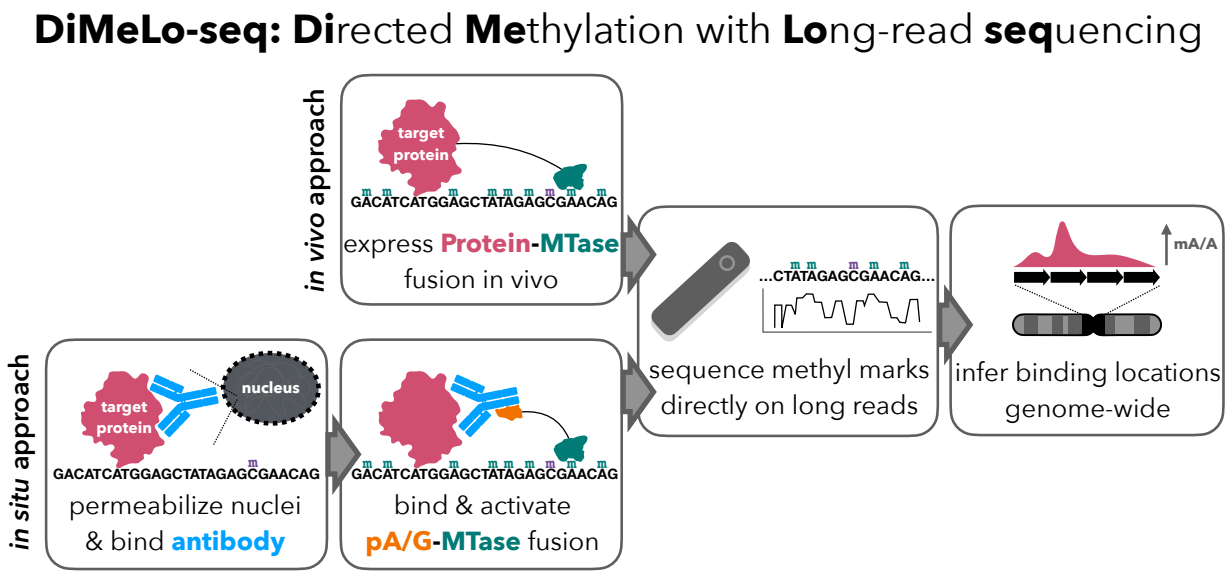


Figure 4.1. Schematic of DiMeLo-seq workflows

Initial considerations

I initially wanted to investigate whether DiMeLo-seq could faithfully reproduce existing DamID results with the protein LMNB1 in the HEK293T human cell line, and whether using the nonspecific adenine methyltransferase EcoGII would improve the accuracy and resolution of the method compared to Dam. I chose LMNB1 as an initial target not only because we had abundant reference data for it in HEK293T cells from our μDamID study (Altemose et al. 2020), but also because LMNB1 has an enormous binding footprint (median 500 kb), allowing us to meaningfully assess the performance of the method with low-coverage sequencing.

I was concerned that conventional DamID and MadID might only achieve useful signal-to-noise ratios because they enrich for methylated DNA (by PCR or immunoprecipitation), while DiMeLo-seq involves no enrichment or amplification at all. To emphasize, the readout for short-read DamID-like methods is the amplified sequence coverage of any particular region of the genome, whereas the readout for DiMeLo-seq is the proportion of adenines that are methylated in any particular region of the genome. That is, DiMeLo-seq sequences the entire genome uniformly, with protein-DNA information occurring as metadata along each read; if the per-cell methylation level is too low, then DiMeLo-seq as described would only be useful at impractically high sequencing depths, with only a tiny subset of reads providing useful protein-DNA binding information. While it would be possible to enrich for long, methylated DNA fragments by immunoprecipitation, this would introduce biases and

destroy the linear relationship between protein-DNA interaction frequency and adenine methylation levels. This would especially complicate the inference of joint binding events on single molecules.

Another concern was that the mA vs A calling accuracy would be too low to provide useful single-molecule information without substantial binning. While PacBio sequencing has been shown to provide >95% mA calling accuracy in most contexts examined (McIntyre et al. 2019), this information was not publicly available for ONT MinION flowcells using their experimental release of an all-contexts mA calling model. To obtain a rough estimate of the accuracy, I contacted Marcus Stoiber, Senior Data Analyst at ONT, who helpfully shared some internal results showing that they achieve mA/A classification F1-scores between 0.7 and 0.95 across 16 sequence contexts, with most contexts in the range of 0.75-0.85 (Marcus Stoiber, personal correspondence). This gave us additional confidence to continue with ONT sequencing.

In vivo DiMeLo-seq

To examine the feasibility of DiMeLo-seq and the quality of the resulting data, I began by sequencing DNA that had been methylated at adenines *in vivo*, as in conventional DamID. To do so, I created stable HEK293T human cell lines that could inducibly express Dam, Dam-LMNB1, EcoGII, or EcoGII-LMNB1. I harvested their DNA after induction and incubation, as one would do for conventional DamID or MadID (Sobecki et al. 2018, Vogel et al. 2007), and then I sequenced them using an Oxford Nanopore Technologies (ONT) MinION device in our lab. I completed one sequencing run in March 2020, just before COVID-19-related shelter-in-place orders began, and I resumed these efforts in November 2020.

I began by comparing DNA methylated *in vivo* by Dam-LMNB1 or EcoGII-LMNB1 to our previously published Dam-LMNB1 conventional bulk DamID data (**Figure 4.2**; Altemose et al. 2020). When using a classification model that exclusively examined GATC sites, it appeared that the Dam-LMNB1 sample outperformed the EcoGII-LMNB1 sample, but when the EcoGII-LMNB1 sample was called with an all-contexts model, it appeared to reproduce the *in vivo* DamID data the best (**Figure 4.2**). To estimate signal and background in order to evaluate the performance of the method, I utilized the positive cLAD and negative ciLAD control regions identified in our μDamID study (Altemose et al. 2020). By reporting the fraction of A's methylated in cLADs (where we expect LMNB1 to contact DNA in nearly all cells), the fraction of A's methylated in ciLADs (where we almost never expect LMNB1 to contact DNA), and the ratio between these, we can obtain a genome-wide summary of the on-target vs background

methylation detected in each sample, even when using extremely low (0.1x) sequencing depth.

Each adenine basecall is reported with an 8-bit probability score, where 255 represents a methylation probability of 1, and 0 represents a methylation probability of 0. The choice of threshold for binary mA calling can affect the signal:background ratio, as less stringent thresholds result in more false positive calls. For the EcoGII-LMNB1 sample, this signal:background ratio was 10.5 when using a mA probability threshold of 0.99, and 9.3 when using a threshold of 0.5. Nearly 1.5% of adenines on reads mapped to cLADs were called as methylated at the 0.5 threshold, while only 0.06% were called as methylated at the 0.99 threshold. These probability scores also depend on the algorithm used for modification calling (see below).

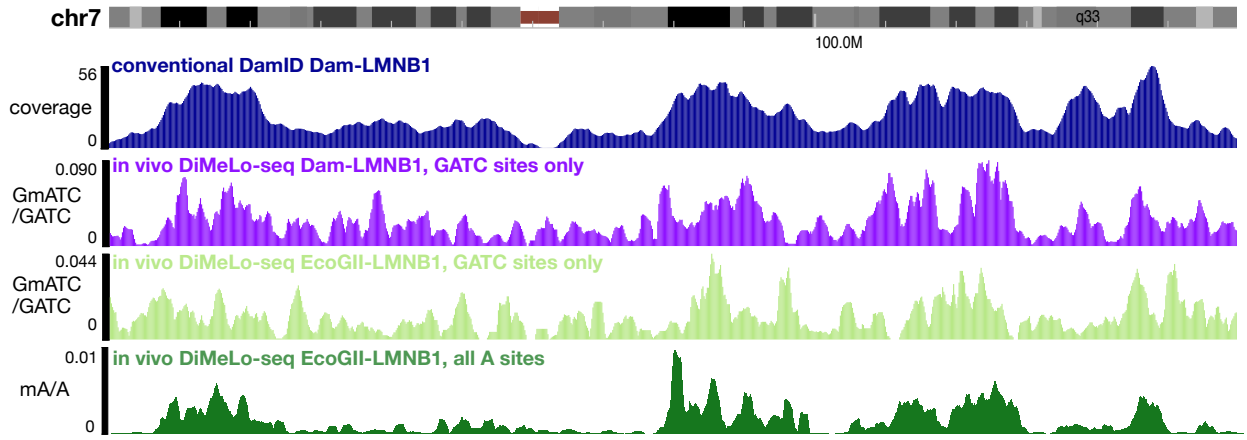


Figure 4.2. ***In vivo*** **DiMeLo-seq recapitulates** ***in vivo*** **DamID results**
Comparison of the coverage from conventional Dam-LMNB1 DamID (1st track) to the proportion of adenines methylated in Dam-LMNB1 or EcoGII-LMNB1 expressing cells sequenced with ONT, using either an m6A classification model restricted to GATC sites (2nd and 3rd tracks), or an all-context m6A model (4th track) across human chromosome 7. Sequencing depth was approximately 0.1x for each ONT sample.

Preliminary specificity estimates of modification calling algorithms
ONT's Guppy software uses a neural network to call bases and base modifications from raw sequencing data, while their Megalodon software calls base modifications by first

aligning reads to a reference sequence. To compare these algorithms and estimate the false positive rate at different probability score thresholds, I sequenced DNA from untreated cells as a negative control, and DNA from cells expressing untethered EcoGII as a positive control, and then plotted the distributions of modification probability scores across all adenines on all reads (**Figure 4.3**). Notably, for Guppy basecalls, I exclusively looked at adenines with a base quality phred score of at least 10, which indicates a predicted basecalling error rate of 10%; Megalodon calls are reference-anchored so do not require a similar base quality filter. These results show that Megalodon appears to have a higher False Discovery Rate (FDR) than Guppy at nearly all probability thresholds, but it may also be more sensitive (**Figure 4.3**). Megalodon shows an anomalous spike at a probability threshold of 0.5, likely an artifact of the classification algorithm used. For most analyses, I compute signal:background ratios and mA fractions at a mA probability threshold of 0.9 (which has a predicted FDR of <0.6%), which provides a suitable balance between sensitivity and specificity for protocol evaluation in large cLADs vs ciLADs. More sensitive thresholds are needed for high-resolution mapping.

The untethered EcoGII *in vivo* sample has 0.5% of all adenines called as methylated by Guppy at an FDR of 1%, which rises to 12.8% called as methylated at an FDR of 8% (**Figure 4.3**). Because overall methylation depends on DNA accessibility and enzyme efficiency, we would not expect to detect 100% methylation on any strand of DNA. It has also been suggested that fully adenine methylated DNA can produce errors in nanopore devices, biasing against the most heavily methylated strands (Shipony et al. 2020). Speculatively, Guppy and Megalodon may perform less accurately when a strand contains many methylated adenines, since they were trained on microbial reference DNA with relatively dispersed methyladenines in a relatively small number of sequence motif contexts. Because we do not know the ground-truth proportion of adenines that are methylated in any of our samples methylated by nonspecific adenine methyltransferases, it is difficult to place an upper bound on the overall sensitivity of our mA detection pipeline. In ongoing work, we are trying to estimate this ground truth using a ^{m6}A ELISA kit (Epigentek P-9010-96) and mass spectrometry (Kriaucionis & Heintz 2009, Quinlivan & Gregory 2008). For now, given the observed positivity rate of 12.8% at the most lenient threshold (at which the FPR is 1%), we can correct it for the false positivity rate and place a conservative lower bound on the sensitivity at 1-(1-0.128)/(1-0.01) = 11.9%. However, in light of the mean ^{m6}A classification F-1 score of 0.8 communicated to me by ONT (Marcus Stoiber, personal correspondence), we can infer that the actual global sensitivity should be at least 66.7%. We can use this to solve the confusion matrix and estimate that at most 18% of adenines are methylated by untethered EcoGII *in vivo*.

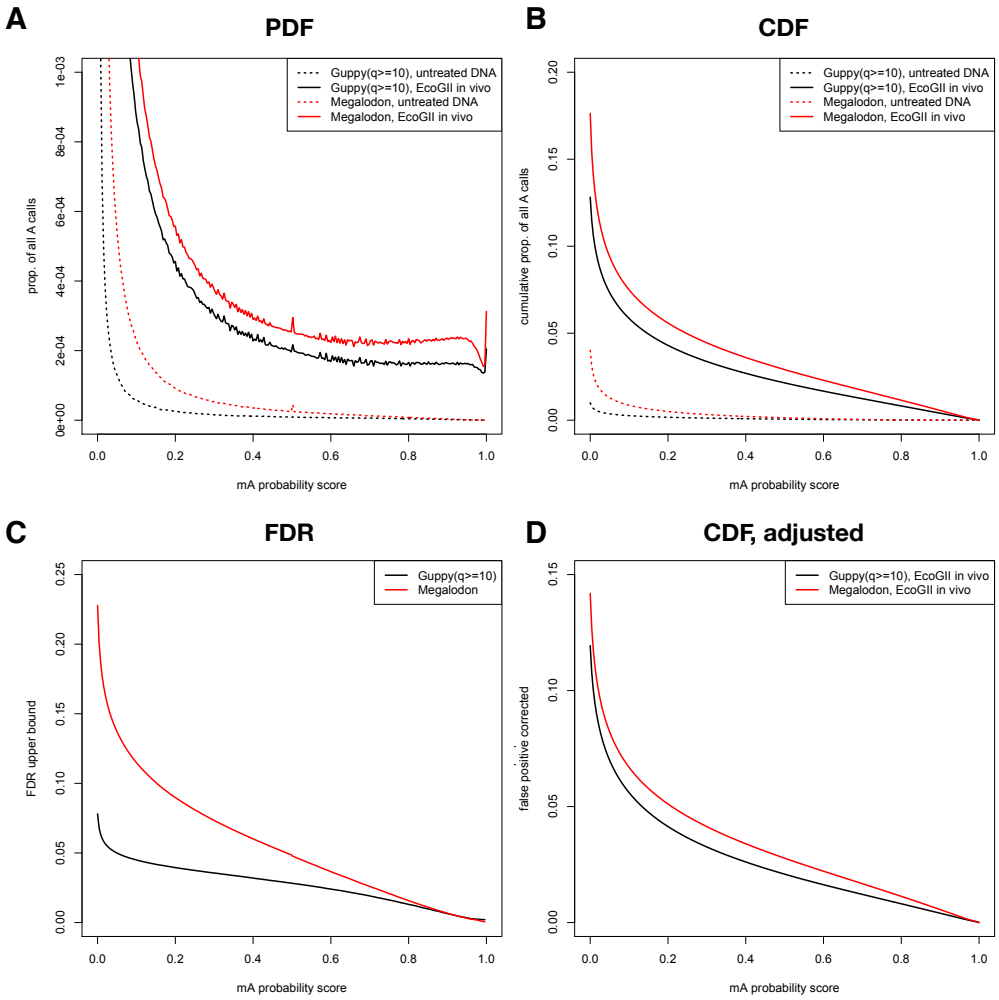


Figure 4.3. Estimation of false positive methyladenine calling rates
(**A**) The proportion of all adenines called as methylated at each possible probability threshold using two different software packages on ONT reads from two HEK293T DNA samples: untreated genomic DNA, and DNA methylated by untethered EcoGII *in vivo*. The untreated DNA provides a measure of the false positive rate at each threshold, since it contains few or no methyladenines. (**B**) The same as A, but showing the cumulative proportion of all adenines called as methylated above each probability threshold (exclusive of 0). (**C**) An upper bound on the False Discovery Rate (FDR), computed as the ratio of mA/A for the untreated sample to mA/A for the EcoGII sample. Specifically, this estimates the proportion of positive mA calls in the EcoGII sample that are false positives at each threshold. The EcoGII sample is not 100% methylated, so it is not a perfect positive ground-truth set, making the FDR only an upper bound. (**D**) This shows the proportion of positive mA calls in the EcoGII sample after adjusting for the estimated false positive rate at each threshold. Effectively, this is an estimate of the true positive proportion when assuming 100% sensitivity.

In situ protocol development and optimization

My preliminary data and analytical pipeline provided confidence that methyladenines were frequent enough, that the mod calling algorithms were sensitive and specific enough, and that background methylation was low enough for the general DiMeLo-seq approach to work successfully on samples methylated *in vivo*. However, the *in vivo* method, like most DamID methods, relies on genetically manipulating the sample of interest and tuning the expression of the MTase-fusion protein to achieve useful signal:background methylation. This genetic manipulation is time consuming and difficult or impossible for some sample types, including primary human tissue samples. Furthermore, it does not enable the detection of post-translational modifications like histone marks. To address these limitations, Annie Maslan and I set out to develop an efficient protocol to target a nonspecific adenine methyltransferase to a protein of interest using antibodies in permeabilized nuclei *in situ*, using a similar workflow to CUT&RUN or pA-DamID (Schaik et al. 2020, Skene & Henikoff 2017).

To begin, I designed a fusion construct between EcoGII and protein A/G (which combines domains from protein A and protein G to bind IgG antibodies from most host organisms), connected by a DDDKEF(GGGGS)x4 linker, similar to the pAG-Tn5 construct used in CUT&RUN (Kaya-Okur et al. 2019), and I had it synthesized and purified (Genscript customized protein production). We decided to combine elements of the published *in situ* antibody targeting protocols CUT&RUN, CUT&TAG, and pA-DamID (Kaya-Okur et al. 2019, Schaik et al. 2020, Skene & Henikoff 2017). To the standard CUT&RUN wash buffer (20 mM HEPES pH 7.5, 150 mM NaCl, 0.5 mM spermidine, 1 Roche protease inhibitor tablet), we added 0.1% Tween-20 (a mild non-ionic detergent) to reduce cell adhesion to tubes and promote plasma membrane permeabilization, and 0.1% Bovine Serum Albumin (BSA) to block nonspecific antibody interactions. For the cell/nucleus permeabilization buffer, we added 0.02% digitonin to the wash buffer. We opted to include multiple washing steps after the antibody and pAG-EcoGII binding steps in order to reduce background binding, unlike the pA-DamID protocol. Much like pA-DamID, we activated the pAG-EcoGII by adding the methyl donor S-adenosylmethionine (SAM) and incubating at 37 °C for 30 minutes.

In vitro methylation comparisons

First, we checked that the pAG-EcoGII fusion construct was capable of methylating DNA *in vitro* by a restriction enzyme blocking assay, and then we tested it with various buffers and additives *in vitro* (**Figure 4.4**). We also compared it to a commercially available EcoGII enzyme from New England Biolabs and to purified Hia5 nonspecific adenine methyltransferase courtesy of Aaron Straight's lab (Hia5 was used in Fiber-seq, Stergachis et al. 2020). We incubated each enzyme with a double-stranded DNA oligo

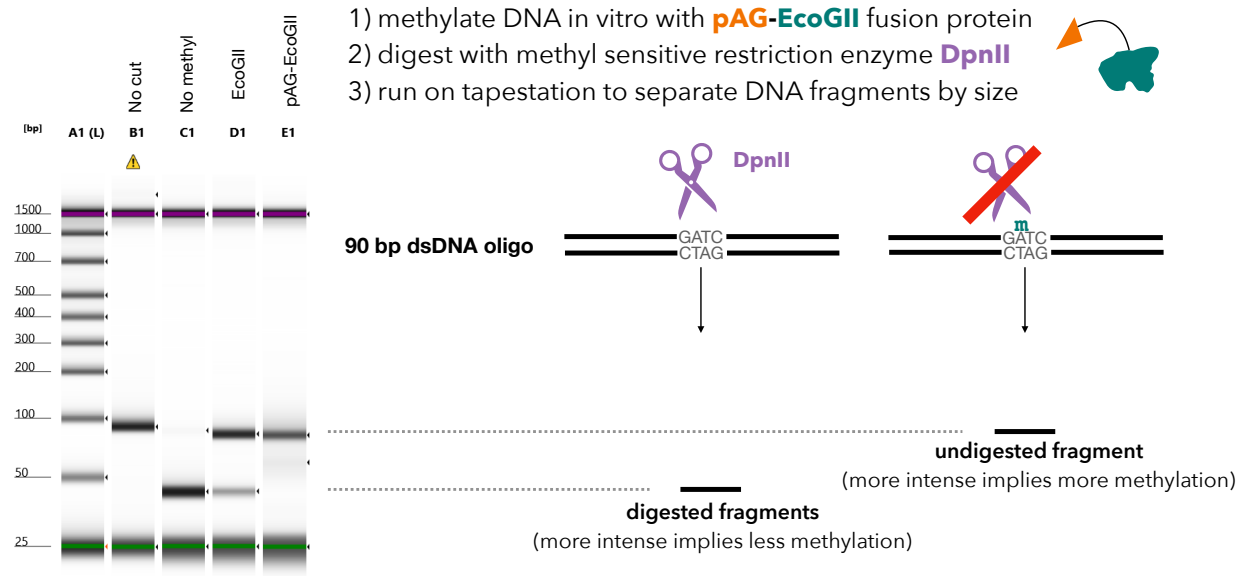
containing a single GATC in the middle, then digested this methylated oligo with DpnII, a restriction enzyme that is blocked by adenine methylation. We ran the resulting fragments on a TapeStation machine, which separates DNA fragments by size, and compared the intensity of bands corresponding to digested and undigested DNA—more undigested DNA implies higher methylation levels. Using this approach, we confirmed that pAG-EcoGII does have strong methylation activity *in vitro*, it performs indistinguishably from Hia5, and it appeared to perform better than the commercially available enzyme. However, in this experiment we could not guarantee that the NEB EcoGII concentration was perfectly matched to pAG-EcoGII and Hia5.

We also examined pAG-EcoGII's methylation ability in various buffers and with various additives commonly found in antibody binding buffers: BSA, Tween-20, Digitonin (a detergent that permeabilizes cell membranes), Spermidine (a polyamine that stabilizes the charge of DNA and preserves chromatin structure when divalent cations are removed), HEPES (a buffer), and Roche protease inhibitor tablets. We also tested the reaction in Hank's Balanced Salt Solution, an isotonic medium containing glucose, and in a low-salt buffer used by the Straight Lab for enzyme activation. We found a dramatic improvement in pAG-EcoGII's methylation activity in our wash buffer when BSA was included (**Figure 4.4**). We speculate that it acts as a molecular crowding agent or reduces protein adhesion to tube walls. However, we also saw high methylation levels in HBSS and in Straight Lab buffer, which do not contain BSA. We saw no appreciable effect of spermidine, tween, Roche tablets, or digitonin on methylation activity.

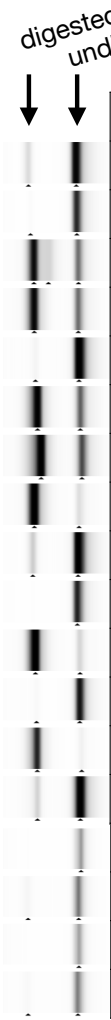
Immunofluorescence assays

Satisfied that the purified methyltransferases were highly active *in vitro*, we next ran through various stages of the initial *in situ* DiMeLo-seq protocol and used microscopy and immunofluorescence to qualitatively evaluate cell permeabilization, nuclear integrity, primary antibody on-target and background binding, and the effects of using a secondary antibody to recruit many methyltransferases to each primary antibody (**Figure 4.5**). Alongside 0.02% digitonin, we decided to test a different detergent, 0.5% NP-40, which is frequently used in nuclear prep protocols. For detection of pAG-EcoGII binding, I used two different secondary antibodies: a goat anti-mouse IgG antibody not expected to bind to the rabbit primary or goat secondary antibodies but is expected to be bound by pAG, and a goat anti-V5 antibody expected to bind to the C-terminal V5 tag on pAG-EcoGII. These ensure that we are visualizing the pAG-EcoGII localization and not just the primary or secondary antibody localization.

A



B



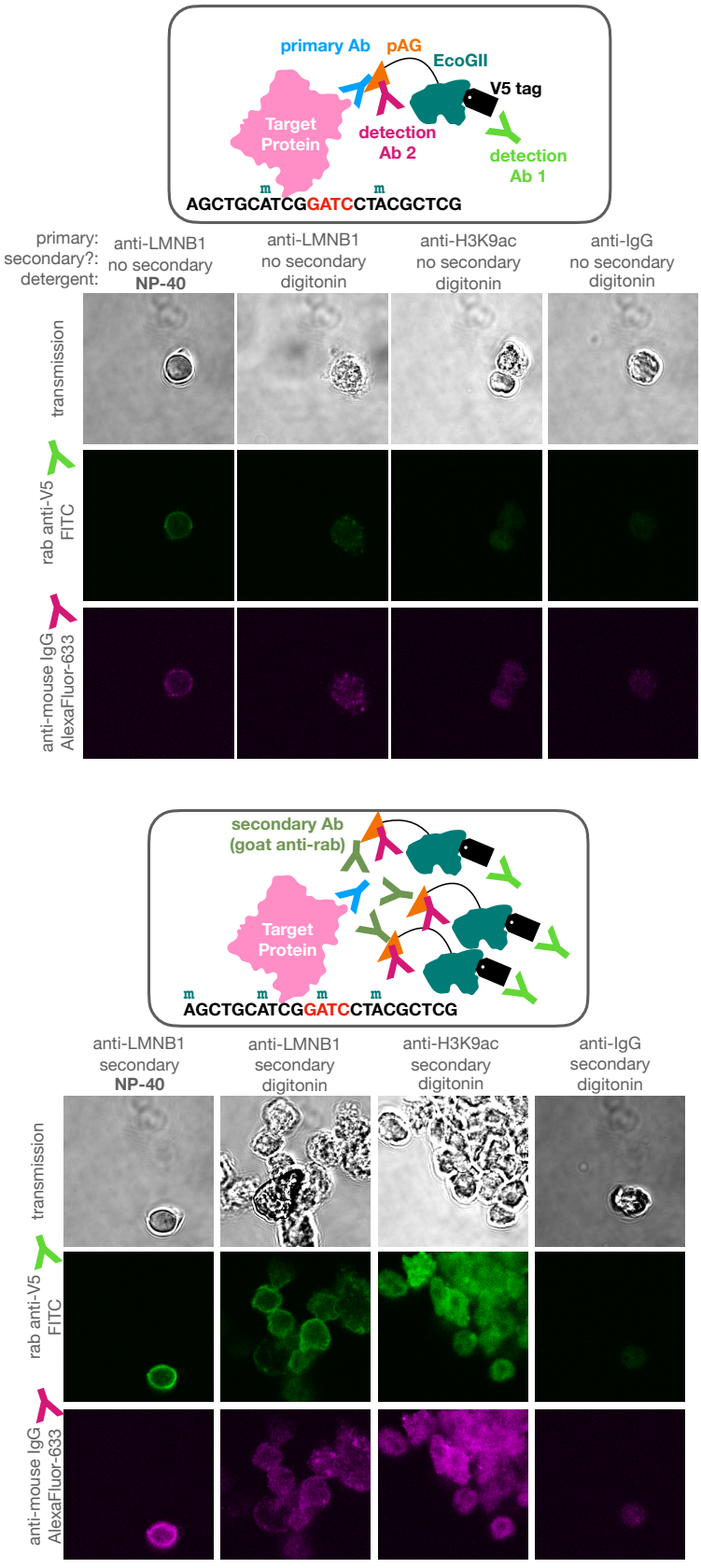
enzyme	buffer	fraction protected
pAG-EcoGII	**CutSmart (contains BSA)**	**0.946**
pAG-EcoGII	**full wash (HEPES, NaCl, BSA, Roche, tween, spermidine)**	**0.984**
pAG-EcoGII	baseline wash (spermidine, NaCl, HEPES)	0.290
pAG-EcoGII	baseline + Roche complete protease inhibitor tablet -EDTA	0.284
pAG-EcoGII	**baseline + BSA**	**0.982**
pAG-EcoGII	baseline + tween	0.270
pAG-EcoGII	baseline + digitonin	0.281
NEB EcoGII	CutSmart (contains BSA)	0.057
pAG-EcoGII	**CutSmart (contains BSA) [90 min]**	**0.941**
pAG-EcoGII	**full wash (BSA, roche, tween, spermidine) [90 min]**	**0.990**
NEB EcoGII	CutSmart (contains BSA) [90 min]	0.064
pAG-EcoGII	**full wash (BSA, roche, tween, spermidine) + older SAM**	**0.979**
pAG-EcoGII	full wash (BSA, roche, tween, spermidine) -SAM	0.047
pAG-EcoGII	**HBSS**	**0.942**
pAG-EcoGII	**full wash (HEPES, NaCl, BSA, Roche, tween, spermidine)**	**0.939**
pAG-EcoGII	**Straight Lab Buffer (Tris, NaCl, KCl, EDTA, EGTA, spermidine)**	**0.904**
Hia5	**full wash (HEPES, NaCl, BSA, Roche, tween, spermidine)**	**0.920**
Hia5	**Straight Lab Buffer (Tris, NaCl, KCl, EDTA, EGTA, spermidine)**	**0.918**

Figure 4.4. ***In vitro*** **methylation assay confirms enzyme activity**
(**A**) Initial demonstration of the restriction enzyme blocking assay to test *in vitro* methylation efficiency. Both commercially available EcoGII and our purified pAG-EcoGII are capable of methylating DNA *in vitro*. (**B**) Results from *in vitro* methylation assay in a variety of different buffer conditions, and comparing EcoGII with Hia5. Both pAG-EcoGII and Hia5 offer near-perfect protection from DpnII digestion. BSA improves efficiency methylation efficiency in our wash buffer.

The confocal fluorescent images shown in **Figure 4.5** reveal that the pAG-EcoGII protein is able to diffuse into the permeabilized nucleus. The anti-LMNB1 samples show the expected ring patterns consistent with proper primary antibody binding to the nuclear lamina, which is not seen for an anti-H3K9ac antibody or an IgG isotype control. Both the transmission images and the anti-LMNB1 images would suggest NP-40 is better than digitonin at preserving nuclear integrity and removing cytoplasmic debris. Comparing fluorescence intensities confirms that there is an amplifying effect due to the use of secondary antibody. That is, it would appear that more molecules of pAG-EcoGII are recruited to the targeted regions when a secondary antibody is used, providing greater contrast between the nuclear lamina and the nuclear interior in the case of the anti-LMNB1 samples. Given these results, we hypothesized that using NP-40 detergent and amplifying signal using a secondary antibody should yield the highest methylation levels and best signal-to-background ratios when the DNA was sequenced.

Protocol optimization using sequencing results
Because we were able to detect substantially different methylation levels in cLADs and ciLADs from *in vivo* EcoGII-LMNB1 samples that were sequenced at low coverage, Annie Maslan and I proceeded to optimize our *in situ* protocol using an anti-LMNB1 antibody. Specifically, we were aiming to optimize the proportion of adenines methylated in cLADs, as a measure of on-target signal, as well as the ratio of this proportion in cLADs and ciLADs, as a measure of signal-to-background. We developed a pipeline that allowed us to multiplex up to 24 samples per flowcell, and to go from cells harvested on Monday morning to fully analyzed sequencing data by Wednesday night.

Figure 4.5. Immunofluorescence confirms proper targeting of pAG-EcoGII



Given this rapid optimization pipeline, we began to test variations of many components of the protocol: input cell numbers, detergents, primary antibody concentrations, the use of secondary antibodies, enzyme types, enzyme concentrations, incubation temperatures, methylation incubation times, methylation buffers, and methyl donor concentrations. We also tested whether cryopreserved or lightly fixed cells could be used as input and whether concanavalin-A coated magnetic beads could be used to carry out cell washing steps on a magnet instead of by centrifugation. While optimization was carried out in HEK293T cells, we validated that it worked in other human cell lines as well: Hap1, GM12878, HG002. We also performed IgG isotype controls and free-floating MTase controls to measure background methylation and DNA accessibility. These results are summarized in **Table 4.1** and **Figure 4.6**. Our final optimized protocol is included as **Appendix 2**.

Our very first *in situ* tests yielded two orders of magnitude less methylation than the *in vivo* EcoGII-LMNB1 sample, with a signal-to-background ratio between 1.6-3. After months of optimization in which we tested over 80 conditions, we now routinely exceed *in vivo* methylation levels (2% of cLAD adenines methylated at a Guppy probability cutoff of 0.5) and regularly achieve a signal-to-background ratio in the range of 20-30. The most important single condition was the enzyme choice. While EcoGII and Hia5 appeared to have comparable activity in our *in vitro* assay, pA-Hia5 (kindly provided by Aaron Straight's lab) greatly outperformed pAG-EcoGII *in situ*. To confirm that this was not due to the choice of pA vs pAG, Owen Smith and Kousik Sundararajan in Aaron Straight's lab cloned and purified pAG-Hia5 and provided it to us for comparison. We found that pA-Hia5 and pAG-Hia5 performed almost identically, with a slight advantage for pA-Hia5, perhaps owing to its smaller size. We speculate that perhaps EcoGII and Hia5 have comparable efficiencies when they are free-floating and methylating naked DNA, but Hia5 has an advantage when tethered to a pA-Ab complex; for example, this might be the case if EcoGII requires dimerization but Hia5 does not. We also tested two different linker lengths between pA and Hia5, which were produced by the Straight Lab, one short (DDDKEF) and one long (DDDKEF(GGGGS)x4), and we did not detect an appreciable difference between them. Furthermore, we tried mixing these linker lengths, and mixing EcoGII with Hia5, but none of these mixtures improved results.

To our great surprise, some of the worst performing samples were those permeabilized in NP-40 detergent (**Figure 4.6**), despite the improvements seen in our immunofluorescence experiments (**Figure 4.5**). This is especially confusing in light of the fact that the detergent is used only for 5 minutes at the start of the protocol, followed by hours of incubations and washes without it prior to enzyme activation—the

detergent never touches the pA/G-MTase. The IF data confirm that this is not due to a failure of permeabilization or a failure of antibody or pA/G binding (**Figure 4.5**). We observe this inhibitory effect of NP-40 on both EcoGII and Hia5, but this effect has not been reported by others when NP-40 has been used for CUT&RUN (MNase enzyme) or CUT&Tag (Tn5 transposase) (based on author correspondences on the respective protocols.io entries for these protocols). We observed a similar effect when we used 0.1% Triton X-100 (**Figure 4.6**), but we have not used imaging to formally rule out that this could be due to lower permeabilization efficiency rather than methylation inhibition (note: 0.1% Tween-20 is present in all of our wash buffers). We can only conclude that somehow NP-40 causes a change in the substrate chromatin that is not reversed by washing, and which specifically inhibits DNA methylation downstream in the protocol. We can only speculate about mechanisms to explain this, and it warrants further investigation.

Another surprise was that secondary antibody amplification does not increase on-target methylation rates or signal-to-background ratios (**Figure 4.6**), despite IF results showing that more pA/G-MTase is recruited to the nuclear lamina when a secondary is used (**Figure 4.5**). This result holds for both pAG-EcoGII and pA-Hia5: for both linker lengths individually, for both linkers mixed together, and for a mixture of pAG-EcoGII and pA-Hia5 with both linkers. Later we show that secondary antibodies provide no improvement for other target proteins as well, nor does using a guinea-pig-derived secondary antibody, which is predicted to have higher affinity for protein A and protein AG than goat-derived antibodies. We speculate that the secondary antibodies have some sort of steric effect resulting in lower methylation rates; the bulky complex of secondary antibodies may block access to the DNA or move the pA/G-MTase too far away from the DNA to methylate efficiently. Perhaps an even longer linker could help to resolve this, but it might result in lower binding-site resolution and more trans methylation. Because the secondary antibody incubation and washes add substantial time to the protocol, we conclude that it is more efficient to leave them out.

These additional factors improved on-target methylation rates and signal-to-background ratios (**Table 4.1**):

-Straight Lab activation buffer + BSA (>Straight Lab activation buffer without BSA > Streets Lab buffer with BSA and low salt > Streets Lab buffer with BSA),
-increased primary antibody concentration (1:50>1:100>1:500),
-increased pA-Hia5 concentration (527 nM = 200 nM > 50 nM),
-room temperature pA/G-MTase incubation (better than 4 °C),
-increased SAM concentration during methylation (800 uM > 500 uM).

These additional factors did not improve performance:

-longer or shorter methylation incubation time (90, 60, 15 mins vs standard 30),
-methylation incubation at 30C instead of 37C (while this did help for pAG-EcoGII, it did not help for pA-Hia5),
-replenishment of SAM during methylation,
-use of higher salt concentration (300 nM) in wash buffer (which has been reported to help in CUT&Tag protocols),
-pre-treatment of DNA with RNAse to potentially increase DNA accessibility.

These conditions did not appreciably harm performance:

-using cells lightly fixed in 0.1% PFA for 2 mins,
-using freshly thawed cells that were cryopreserved in DMSO-containing freezing medium and stored in liquid nitrogen (though anecdotally we observed shorter read lengths for these samples),
-using concanavalin-A coated magnetic beads for cell processing (though this may limit capacity per tube and makes IF quality controls difficult),
-starting with 5 million cells vs 1 million cells per tube (we observed some loss of performance with 10 million cells though).

Table 4.1. Summary of all DiMeLo-seq sequencing runs

index & name	seq. date	barcode	cell line	detergent	Ab	Ab dilution factor	misc	pA/G	linker len.	MTase	[pA/G-MTase]	2Ab time & temp	pA/G time & temp	act. buffer	[SAM]	act. time	act. Temp	read number	total bases sequenced	mean read len	cLAD:ciLAD mA ratio	cLAD mA/A	all reads mA/A
1-in vivo dam-lmnb1 full	20200320	1	HEK293T	-	-	-	in vivo Dam-LMNB1 full ind, g-tube + SRE XS	-	-	Dam	-	-	-	-	-	-	37C	53,288	418,298,496	7,850	2.628	2.49E-05	1.51E-05
2-in vivo ecogii-lmnb1 full	20200320	2	HEK293T	-	-	-	in vivo EcoGII-LMNB1 full ind, g-tube+SRE XS	-	-	EcoGII	-	-	-	-	-	-	37C	61,294	445,922,529	7,275	10.464	3.97E-03	1.43E-03
3-in vivo ecogii full	20201117	3	HEK293T	-	-	-	in vivo EcoGII full ind, SRE XL	-	-	EcoGII	-	-	-	-	-	-	37C	22,706	499,613,882	22,004	0.682	4.28E-03	5.30E-03
4-in vivo ecogii 10 ind	20201117	4	HEK293T	-	-	-	in vivo EcoGII 10% ind, SRE XL	-	-	EcoGII	-	-	-	-	-	-	37C	45,074	1,059,489,578	23,506	0.673	4.13E-04	5.18E-04
5-in vivo ecogii-lmnb1 10 ind	20201117	5	HEK293T	-	-	-	in vivo EcoGII-LMNB1 10% ind, SRE XL	-	-	EcoGII	-	-	-	-	-	-	37C	23,326	626,118,574	26,842	12.398	3.96E-04	1.23E-04
6-pag-ecogii-lmnb1	20201117	6	HEK293T	digitonin 0.02%	LMNB1	500	SRE XL	pAG	29 aa	EcoGII	50 nM	-	1 h, 4C	Streets	500 uM	30 min	37C	86,481	1,976,058,348	22,850	1.668	4.46E-05	3.37E-05
7-pag-ecogii-igg	20201117	7	HEK293T	digitonin 0.02%	IgG	500	SRE XL	pAG	29 aa	EcoGII	50 nM	-	1 h, 4C	Streets	500 uM	30 min	37C	39,167	872,653,096	22,280	1.113	2.22E-05	2.46E-05
8-unmethylated	20201117	8	HEK293T	-	-	-	unmethylated DNA, SRE XL	-	-	-	-	-	-	-	-	-	37C	11,686	302,257,947	25,865	2.187	3.51E-05	2.51E-05
9-pag-ecogii-lmnb1-2ab	20201119	9	HEK293T	digitonin 0.02%	LMNB1	500	SRE XL	pAG	29 aa	EcoGII	150 nM	1 h, 4C	1 h, 4C	Streets	500 uM	30 min	37C	38,539	847,065,308	21,979	2.929	1.57E-04	1.00E-04
10-pag-ecogii-lmnb1-1ab	20201119	10	HEK293T	digitonin 0.02%	LMNB1	500	SRE XL	pAG	29 aa	EcoGII	150 nM	-	1 h, 4C	Streets	500 uM	30 min	37C	36,921	856,905,359	23,209	2.499	1.37E-04	8.96E-05
11-pag-ecogii-h3k9ac-2ab	20201119	11	HEK293T	digitonin 0.02%	H3K9ac	100	SRE XL	pAG	29 aa	EcoGII	150 nM	1 h, 4C	1 h, 4C	Streets	500 uM	30 min	37C	44,855	1,100,797,612	24,541	0.621	8.84E-05	1.11E-04
12-pag-ecogii-h3k9ac-1ab	20201119	12	HEK293T	digitonin 0.02%	H3K9ac	100	SRE XL	pAG	29 aa	EcoGII	150 nM	-	1 h, 4C	Streets	500 uM	30 min	37C	42,793	992,142,666	23,185	0.783	5.63E-05	6.78E-05
13-pag-ecogii-igg-1ab	20201119	13	HEK293T	digitonin 0.02%	IgG	500	SRE XL	pAG	29 aa	EcoGII	150 nM	-	1 h, 4C	Streets	500 uM	30 min	37C	66,878	1,500,527,829	22,437	0.982	4.86E-05	4.68E-05
14-pag-ecogii-lmnb1-2ab-np40	20201119	14	HEK293T	NP40 0.5%	LMNB1	500	SRE XL	pAG	29 aa	EcoGII	150 nM	1 h, 4C	1 h, 4C	Streets	500 uM	30 min	37C	60,602	1,490,024,882	24,587	1.208	1.18E-04	1.04E-04
15-pag-ecogii-lmnb1-1ab-np40	20201119	15	HEK293T	NP40 0.5%	LMNB1	500	SRE XL	pAG	29 aa	EcoGII	150 nM	-	1 h, 4C	Streets	500 uM	30 min	37C	38,772	943,586,111	24,337	1.351	7.23E-05	5.95E-05
16-pag-ecogii-lmnb1-1ab-7.3ug-RT	20201207	16	HEK293T	digitonin 0.02%	LMNB1	500		pAG	29 aa	EcoGII	527 nM	-	1 h, RT	Streets	500 uM	30 min	37C	162,252	1,752,177,765	10,799	1.377	5.32E-05	4.83E-05
17-pag-ecogii-lmnb1-2ab-7.3ug-RT	20201207	17	HEK293T	digitonin 0.02%	LMNB1	500		pAG	29 aa	EcoGII	527 nM	1 h, RT	1 h, RT	Streets	500 uM	30 min	37C	198,264	2,102,108,611	10,603	3.034	9.52E-05	6.15E-05
18-pag-ecogii-lmnb1-1ab-2.2ug-RT	20201207	18	HEK293T	digitonin 0.02%	LMNB1	500		pAG	29 aa	EcoGII	150 nM	-	1 h, RT	Streets	500 uM	30 min	37C	106,892	1,227,114,851	11,480	2.046	2.99E-05	2.36E-05
19-pag-ecogii-lmnb1-1ab-1:100-7.3ug-RT	20201207	19	HEK293T	digitonin 0.02%	LMNB1	100		pAG	29 aa	EcoGII	527 nM	-	1 h, RT	Streets	500 uM	30 min	37C	238,437	2,415,676,861	10,131	4.795	1.74E-04	9.03E-05
20-pag-ecogii-lmnb1-2ab-1:100-7.3ug-RT	20201211	20	HEK293T	digitonin 0.02%	LMNB1	100		pAG	29 aa	EcoGII	527 nM	1 h, RT	1 h, 4C	Streets	500 uM	30 min	37C	121,419	1,261,478,986	10,389	4.729	4.01E-04	2.06E-04
21-pag-ecogii-lmnb1-1ab-1:100-7.3ug-4C	20201211	21	HEK293T	digitonin 0.02%	LMNB1	100		pAG	29 aa	EcoGII	527 nM	-	1 h, 4C	Streets	500 uM	30 min	37C	123,908	1,276,684,561	10,303	3.982	2.99E-04	1.65E-04
22-pag-ecogii-lmnb1-1ab-1:100-7.3ug-4C-noSAM	20201211	22	HEK293T	digitonin 0.02%	LMNB1	100	no SAM	pAG	29 aa	EcoGII	527 nM	-	1 h, 4C	Streets	0 uM	30 min	37C	144,251	1,297,558,669	8,995	1.216	3.83E-05	3.69E-05
23-pag-ecogii-lmnb1-1ab-1:100-7.3ug-4C-60min	20201211	23	HEK293T	digitonin 0.02%	LMNB1	100	60 min act	pAG	29 aa	EcoGII	527 nM	-	1 h, 4C	Streets	500 uM	60 min	37C	112,198	1,162,823,359	10,364	3.728	3.23E-04	1.87E-04
24-pag-ecogii-lmnb1-1ab-1:100-7.3ug-4C-60min-replenishSAM	20201211	24	HEK293T	digitonin 0.02%	LMNB1	100	SAM replenished	pAG	29 aa	EcoGII	527 nM	-	1 h, 4C	Streets, SAM replenished	500 uM	60 min	37C	119,153	1,207,239,004	10,132	3.049	3.13E-04	1.89E-04
25-pag-ecogii-lmnb1-1ab-1:100-7.3ug-4C-90min-replenishSAM	20201211	1	HEK293T	digitonin 0.02%	LMNB1	100	SAM replenished	pAG	29 aa	EcoGII	527 nM	-	1 h, 4C	Streets, SAM replenished	500 uM	90 min	37C	96,385	1,076,704,141	11,171	2.650	2.87E-04	1.73E-04
26-pag-ecogii-igg-2ab	20201211	2	HEK293T	digitonin 0.02%	IgG	100		pAG	29 aa	EcoGII	527 nM	1 h, RT	1 h, 4C	Streets	500 uM	30 min	37C	165,186	1,608,748,957	9,739	0.818	6.05E-05	8.06E-05
27-pag-ecogii-igg-1ab	20201211	3	HEK293T	digitonin 0.02%	IgG	100		pAG	29 aa	EcoGII	527 nM	-	1 h, 4C	Streets	500 uM	30 min	37C	95,567	982,777,929	10,284	0.835	7.71E-05	9.15E-05
28-baseline, rep1	20201223 + 20201228	4	HEK293T	digitonin 0.02%	LMNB1	100		pAG	29 aa	EcoGII	527 nM	1 h, RT	1 h, RT	Streets	500 uM	30 min	37C	80,135	839,079,448	10,471	3.564	4.22E-04	2.36E-04
29-15 min act	20201223 + 20201228	5	HEK293T	digitonin 0.02%	LMNB1	100	15 min act	pAG	29 aa	EcoGII	527 nM	1 h, RT	1 h, RT	Streets	500 uM	15 min	37C	101,911	1,067,611,582	10,476	3.872	3.25E-04	1.74E-04
30-30C act	20201223 + 20201228	6	HEK293T	digitonin 0.02%	LMNB1	100	30C act	pAG	29 aa	EcoGII	527 nM	1 h, RT	1 h, RT	Streets	500 uM	30 min	30C	68,779	651,361,202	9,470	7.536	4.95E-04	2.13E-04
31-baseline, rep2	20201223 + 20201228	7	HEK293T	digitonin 0.02%	LMNB1	100		pAG	29 aa	EcoGII	527 nM	1 h, RT	1 h, RT	Streets	500 uM	30 min	37C	83,487	902,381,502	10,809	4.854	4.40E-04	2.35E-04
32-pA-Hia5-short-2ab	20201223	8	HEK293T	digitonin 0.02%	LMNB1	100		pA	7 aa	Hia5	527 nM	1 h, RT	1 h, RT	Streets	500 uM	30 min	37C	26,351	227,185,814	8,622	5.641	7.04E-05	2.87E-05

Table 4.1. Summary of all DiMeLo-seq sequencing runs [continued]

index & name	seq. date	barcode	cell line	detergent	Ab	Ab dilution factor	misc	pA/G	linker len.	MTase	[pA/G-MTase]	2Ab time & temp	pA/G time & temp	act. buffer	[SAM]	act. time	act. Temp	read number	total bases sequenced	mean read len	cLAD:ciLAD mA ratio	cLAD mA/A	all reads mA/A
33-pA-Hia5-long-2ab	20201223	9	HEK293T	digitonin 0.02%	LMNB1	100		pA	26 aa	Hia5	527 nM	1 h, RT	1 h, RT	Streets	500 uM	30 min	37C	19,112	169,551,133	8,871	7.415	8.39E-05	4.49E-05
34-pA-Hia5-both-2ab	20201223	10	HEK293T	digitonin 0.02%	LMNB1	100		pA	mix	Hia5	527 nM	1 h, RT	1 h, RT	Streets	500 uM	30 min	37C	15,778	151,798,963	9,621	9.844	7.88E-05	3.82E-05
35-pAG-EcoGII-pA-both-2ab	20201223	11	HEK293T	digitonin 0.02%	LMNB1	100	pAG-EcoGII+pA-Hia5-both	mix	mix	EcoGII +Hia5	527 nM	1 h, RT	1 h, RT	Streets	500 uM	30 min	37C	15,975	149,309,499	9,346	11.823	2.65E-04	1.45E-04
36-pAG-EcoGII-NP40	20201223 + 20201228	12	HEK293T	NP40 0.1%	LMNB1	100	NP40	pAG	29 aa	EcoGII	527 nM	1h, RT	1 h, RT	Streets	500 uM	30 min	37C	86,033	860,416,335	10,001	2.479	2.50E-04	1.65E-04
37-pAG-EcoGII-freeFloating	20201223 + 20201228	13	HEK293T	digitonin 0.02%	No Ab	-	free floating pAG-EcoGII	pAG	29 aa	EcoGII	527 nM	1h, RT	1 h, RT	Streets	500 uM	30 min	37C	102,650	983,478,388	9,581	0.572	2.80E-04	4.49E-04
38-baseline	20201223 + 20201228	14	HEK293T	digitonin 0.02%	LMNB1	100	baseline	pAG	29 aa	EcoGII	527 nM	-	1 h, RT	Streets	500 uM	30 min	37C	101,662	993,731,709	9,775	2.513	2.97E-04	1.98E-04
39-pA-Hia5-short dig Streets	20201223 + 20201228	15	HEK293T	digitonin 0.02%	LMNB1	100		pA	7 aa	Hia5	527 nM	-	1 h, RT	Streets	500 uM	30 min	37C	106,830	985,205,512	9,222	12.427	7.89E-04	3.26E-04
40-pA-Hia5-long dig Streets	20201223 + 20201228	16	HEK293T	digitonin 0.02%	LMNB1	100		pA	26 aa	Hia5	527 nM	-	1 h, RT	Streets	500 uM	30 min	37C	74,873	793,137,043	10,593	11.724	1.06E-03	4.23E-04
41-pA-Hia5-both dig Streets	20201223 + 20201228	17	HEK293T	digitonin 0.02%	LMNB1	100		pA	mix	Hia5	527 nM	-	1 h, RT	Streets	500 uM	30 min	37C	93,269	988,096,097	10,594	17.681	8.36E-04	3.43E-04
42-pAG-EcoGII-pA-both dig Streets	20201223 + 20201228	18	HEK293T	digitonin 0.02%	LMNB1	100	pAG-EcoGII+pA-Hia5-both	mix	mix	EcoGII +Hia5	527 nM	-	1 h, RT	Streets	500 uM	30 min	37C	92,725	945,637,416	10,198	6.085	7.71E-04	3.27E-04
43-pag-ecogii-lightFixation	20201228	19	HEK293T	digitonin 0.02%	LMNB1	100	light fixation	pAG	29 aa	EcoGII	527 nM	1 h, RT	1 h, RT	Streets	500 uM	30 min	37C	91,520	686,940,638	7,506	5.370	4.31E-04	2.16E-04
44-pag-ecogii-lightFixation-noAb	20201228	20	HEK293T	digitonin 0.02%	No Ab	-	light fixation	pAG	29 aa	EcoGII	527 nM	1 h, RT	1 h, RT	Streets	500 uM	30 min	37C	118,802	260,574,348	2,193	0.573	8.96E-05	1.38E-04
45-pA-Hia5 long triton Streets	20201230	21	HEK293T	Triton 0.1%	LMNB1	100	triton	pA	26 aa	Hia5	527 nM	-	1 h, RT	Streets	500 uM	30 min	37C	100,177	1,126,513,813	11,245	4.664	9.10E-05	4.20E-05
46-pA-Hia5 long triton Straight	20201230	22	HEK293T	Triton 0.1%	LMNB1	100	triton	pA	26 aa	Hia5	527 nM	-	1 h, RT	Straight	500 uM	30 min	37C	114,462	1,105,844,974	9,661	21.179	3.68E-04	1.36E-04
47-pA-Hia5 long np40	20201230	23	HEK293T	NP40 0.1%	LMNB1	100	NP40	pA	26 aa	Hia5	527 nM	-	1 h, RT	Streets	500 uM	30 min	37C	105,547	934,810,760	8,857	3.679	6.12E-05	2.82E-05
48-pAG-EcoGII NP40	20201230	24	HEK293T	NP40 0.1%	LMNB1	100	NP40	pAG	29 aa	EcoGII	527 nM	-	1 h, RT	Streets	500 uM	30 min	37C	84,310	868,314,441	10,299	4.312	1.11E-04	5.22E-05
49-baseline	20201230	1	HEK293T	digitonin 0.02%	LMNB1	100	baseline	pAG	29 aa	EcoGII	527 nM	-	1 h, RT	Streets	500 uM	30 min	37C	91,978	1,189,823,234	12,936	3.411	1.49E-04	7.60E-05
50-pA-Hia5 long dig Streets	20201230	2	HEK293T	digitonin 0.02%	LMNB1	100		pA	26 aa	Hia5	527 nM	-	1 h, RT	Streets	500 uM	30 min	37C	105,110	1,264,716,882	12,032	24.212	5.33E-04	1.78E-04
51-pA-Hia5 long dig Straight	20201230	3	HEK293T	digitonin 0.02%	LMNB1	100		pA	26 aa	Hia5	527 nM	-	1 h, RT	Straight	500 uM	30 min	37C	103,978	1,311,352,305	12,612	38.048	2.06E-03	6.93E-04
52-pA-Hia5 long Straight+BSA 37C	20210119	4	HEK293T	digitonin 0.02%	LMNB1	100		pA	26 aa	Hia5	527 nM	-	1 h, RT	Straight + BSA	500 uM	30 min	37C	90,785	752,583,312	8,290	26.474	4.08E-03	1.47E-03
53-pAG-Hia5 Straight+BSA 37C	20210119	5	HEK293T	digitonin 0.02%	LMNB1	100		pAG	7 aa	Hia5	527 nM	-	1 h, RT	Straight + BSA	500 uM	30 min	37C	48,122	459,680,632	9,552	29.831	2.23E-03	8.67E-04
54-pAG-EcoGII Straight+BSA 37C	20210119	6	HEK293T	digitonin 0.02%	LMNB1	100		pAG	29 aa	EcoGII	527 nM	-	1 h, RT	Straight + BSA	500 uM	30 min	37C	81,695	727,342,635	8,903	7.484	1.21E-03	5.25E-04
55-pA-Hia5 long Straight 37C	20210119	7	HEK293T	digitonin 0.02%	LMNB1	100		pA	26 aa	Hia5	527 nM	-	1 h, RT	Straight	500 uM	30 min	37C	65,117	583,081,677	8,954	23.486	3.05E-03	1.20E-03
56-pAG-Hia5 Straight 37C	20210119	8	HEK293T	digitonin 0.02%	LMNB1	100		pAG	7 aa	Hia5	527 nM	-	1 h, RT	Straight	500 uM	30 min	37C	56,317	540,388,209	9,595	29.747	2.08E-03	8.06E-04
57-pAG-EcoGII Straight 37C	20210119	9	HEK293T	digitonin 0.02%	LMNB1	100		pAG	29 aa	EcoGII	527 nM	-	1 h, RT	Straight	500 uM	30 min	37C	47,684	449,643,915	9,430	8.223	1.15E-03	4.77E-04
58-pA-Hia5 long Straight 30C	20210119	10	HEK293T	digitonin 0.02%	LMNB1	100		pA	26 aa	Hia5	527 nM	-	1 h, RT	Straight	500 uM	30 min	30C	58,489	523,464,949	8,950	29.977	2.24E-03	8.07E-04
59-pAG-Hia5 Straight 30C	20210119	11	HEK293T	digitonin 0.02%	LMNB1	100		pAG	7 aa	Hia5	527 nM	-	1 h, RT	Straight	500 uM	30 min	30C	57,609	511,675,484	8,882	17.415	1.26E-03	4.72E-04
60-pAG-EcoGII Straight 30C	20210119	12	HEK293T	digitonin 0.02%	LMNB1	100		pAG	29 aa	EcoGII	527 nM	-	1 h, RT	Straight	500 uM	30 min	30C	61,562	551,698,966	8,962	4.400	5.33E-04	2.44E-04
61-pAG-EcoGII Streets 30C	20210119	13	HEK293T	digitonin 0.02%	LMNB1	100		pAG	29 aa	EcoGII	527 nM	-	1 h, RT	Streets	500 uM	30 min	30C	71,900	628,811,974	8,746	8.180	3.02E-04	1.31E-04
62-pA-Hia5 long 2ab Straight 37C	20210119	14	HEK293T	digitonin 0.02%	LMNB1	100	secondary (GP)	pA	26 aa	Hia5	527 nM	1h, RT	1 h, RT	Straight	500 uM	30 min	37C	99,602	843,096,447	8,465	21.368	2.33E-03	8.98E-04
63-pA-Hia5 long 2ab Streets low salt 37C	20210119	15	HEK293T	digitonin 0.02%	LMNB1	100	secondary (GP), low salt	pA	26 aa	Hia5	527 nM	1h, RT	1 h, RT	Streets Low Salt	500 uM	30 min	37C	54,677	484,736,060	8,865	8.614	4.85E-04	2.04E-04
64-pAG-Hia5 2ab Straight 37C	20210119	16	HEK293T	digitonin 0.02%	LMNB1	100	secondary (GP)	pAG	7 aa	Hia5	527 nM	1h, RT	1 h, RT	Straight	500 uM	30 min	37C	63,400	571,575,224	9,015	13.879	1.43E-03	5.34E-04
65-pAG-EcoGII 2ab Streets 30C	20210119	17	HEK293T	digitonin 0.02%	LMNB1	100	secondary (GP)	pAG	29 aa	EcoGII	527 nM	1h, RT	1 h, RT	Streets	500 uM	30 min	30C	60,290	517,001,320	8,575	7.268	3.43E-04	1.68E-04
66-pA-Hia5 S+L Straight 37C	20210119	18	HEK293T	digitonin 0.02%	LMNB1	100	pA-Hia5 short+long	pA	mix	Hia5	527 nM	-	1 h, RT	Straight	500 uM	30 min	37C	47,923	468,996,420	9,786	21.990	3.19E-03	1.31E-03
67-pA-Hia5 200 nM 2ab Straight 37C	20210119	19	HEK293T	digitonin 0.02%	LMNB1	100	pA-Hia5 200 nM, secondary (GP)	pA	26 aa	Hia5	200 nM	1h, RT	1 h, RT	Straight	500 uM	30 min	37C	52,514	513,948,158	9,787	19.589	2.82E-03	1.06E-03
68-pA-Hia5 50 nM 2ab Straight 37C	20210119	20	HEK293T	digitonin 0.02%	LMNB1	100	pA-Hia5 50 nM, secondary (GP)	pA	26 aa	Hia5	50 nM	1h, RT	1 h, RT	Straight	500 uM	30 min	37C	48,723	441,526,404	9,062	20.303	2.55E-03	9.19E-04
69-click conj	20210119	21	HEK293T	digitonin 0.02%	LMNB1	100	click conjugation attempt	-	-	Hia5	527 nM		1 h, RT	Streets	500 uM	30 min	37C	53,990	517,552,447	9,586	1.943	6.80E-05	5.19E-05

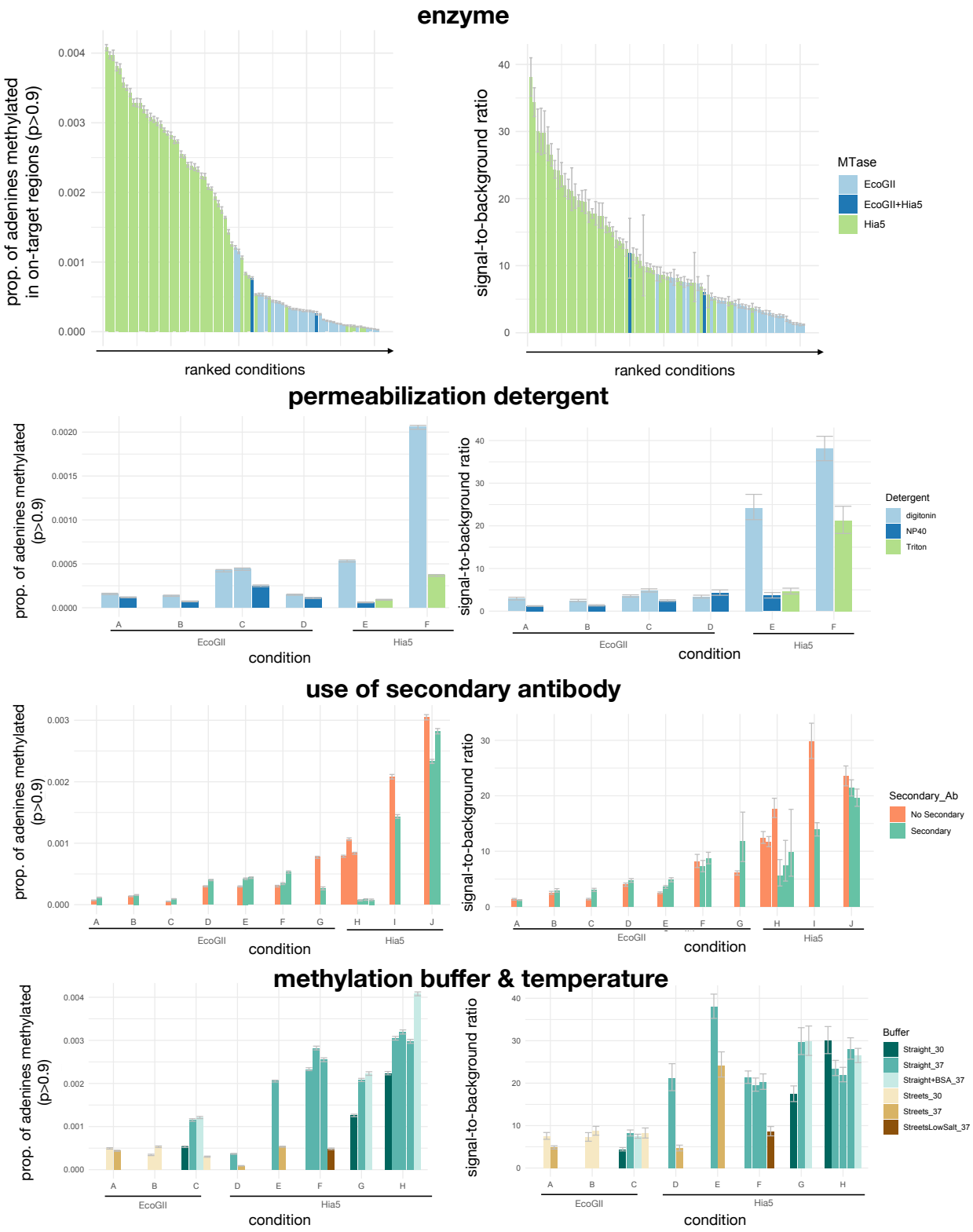
Table 4.1. Summary of all DiMeLo-seq sequencing runs [continued]

index & name	seq. date	barcode	cell line	detergent	Ab	Ab dilution factor	misc	pA/G	linker len.	MTase	[pA/G-MTase]	2Ab time & temp	pA/G time & temp	act. buffer	[SAM]	act. time	act. Temp	read number	total bases sequenced	mean read len	cLAD:cLAD mA ratio	cLAD mA/A	all reads mA/A
70-pA-Hia5 Straight 37C frozen	20210119	22	HEK293T	digitonin 0.02%	LMNB1	100	frozen	pA	26 aa	Hia5	527 nM	-	1 h, RT	Straight	500 uM	30 min	37C	71,276	488,718,798	6,857	28.059	2.98E-03	1.09E-03
71-pAG-EcoGII goat 2ab Streets 30C	20210119	23	HEK293T	digitonin 0.02%	LMNB1	100	secondary (Goat)	pAG	29 aa	EcoGII	527 nM	1h, RT	1 h, RT	Streets	500 uM	30 min	30C	61,444	554,745,469	9,028	8.711	5.32E-04	2.22E-04
72-LMNB1 HEK	20210201	1	HEK293T	digitonin 0.02%	LMNB1	100		pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	45,362	345,941,926	7,626	7.676	2.36E-03	1.09E-03
73-LMNB1 HEK	20210201	2	HEK293T	digitonin 0.02%	LMNB1	100		pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	40,187	353,173,477	8,788	6.865	2.38E-03	1.06E-03
74-LMNB1 1:50 HEK	20210201	3	HEK293T	digitonin 0.02%	LMNB1	50		pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	33,387	278,128,346	8,330	9.621	3.97E-03	1.73E-03
75-H3K9ac 1:100 HEK	20210201	4	HEK293T	digitonin 0.02%	H3K9ac	100		pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	40,594	335,773,580	8,272	1.615	5.57E-04	4.51E-04
76-H3K9ac 1:50 HEK	20210201	5	HEK293T	digitonin 0.02%	H3K9ac	50		pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	33,494	295,660,205	8,827	1.097	6.40E-04	5.43E-04
77-CENP-A HEK	20210201	6	HEK293T	digitonin 0.02%	CENP-A	100		pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	45,117	373,778,336	8,285	1.925	3.94E-04	3.57E-04
78-LMNB1 GM12878	20210201	7	GM12878	digitonin 0.02%	LMNB1	100		pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	55,380	462,838,466	8,358	8.386	2.75E-03	1.24E-03
79-LMNB1 HG002	20210201	8	HG002	digitonin 0.02%	LMNB1	100		pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	59,508	439,509,594	7,386	9.339	3.13E-03	1.30E-03
80-LMNB1 Hap1	20210201	9	Hap1	digitonin 0.02%	LMNB1	100		pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	59,648	424,749,575	7,121	7.473	3.43E-03	1.78E-03
81-IgG GM12878	20210201	10	GM12878	digitonin 0.02%	IgG	100		pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	35,649	304,927,237	8,554	1.773	4.57E-04	3.72E-04
82-IgG HG002	20210201	11	HG002	digitonin 0.02%	IgG	100		pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	58,023	418,761,436	7,217	1.448	4.44E-04	3.67E-04
83-IgG Hap1	20210201	12	Hap1	digitonin 0.02%	IgG	100		pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	73,437	453,120,139	6,170	1.865	4.77E-04	3.78E-04
84-IgG HEK	20210201	13	HEK293T	digitonin 0.02%	IgG	100		pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	50,184	395,339,186	7,878	1.563	4.15E-04	3.44E-04
85-LMNB1 HEK no SAM	20210201	14	HEK293T	digitonin 0.02%	LMNB1	100	no SAM	pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	0 uM	30 min	37C	37,978	301,486,227	7,938	2.585	5.13E-04	3.57E-04
86-LMNB1 HEK fixation	20210201	15	HEK293T	digitonin 0.02%	LMNB1	100	light fixation	pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	52,734	315,838,720	5,989	9.736	3.58E-03	1.48E-03
87-free Hia5	20210201	16	HEK293T	digitonin 0.02%	- -	-	free floating Hia5	-	-	Hia5	200 nM	-	-	Straight+ BSA	500 uM	30 min	37C	49,149	363,924,351	7,405	1.102	7.55E-03	7.47E-03
88-LMNB1 Hap1 RT	20210201	17	Hap1	digitonin 0.02%	LMNB1	100	primary at RT	pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	54,120	354,171,388	6,544	8.120	3.29E-03	1.72E-03
89-CENP-A Hap1	20210201	18	Hap1	digitonin 0.02%	CENP-A	100		pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	55,095	385,496,401	6,997	1.846	4.14E-04	3.56E-04
90-LMNB1 HEK conA	20210201	19	HEK293T	digitonin 0.02%	LMNB1	100	conA beads	pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	61,614	494,542,424	8,026	8.561	3.29E-03	1.41E-03
91-LMNB1 GM12878 conA	20210201	20	GM12878	digitonin 0.02%	LMNB1	100	conA beads	pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	64,420	461,182,886	7,159	5.146	3.08E-03	1.52E-03
92-LMNB1 Hap1 conA	20210201	24	Hap1	digitonin 0.02%	LMNB1	100	conA beads	pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	54,354	404,350,726	7,439	4.455	3.01E-03	1.65E-03
93-LMNB1 HEK bad	20210210	1	HEK293T	digitonin 0.02%	LMNB1	100		pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	45,503	360,723,378	7,927	10.651	1.84E-03	7.29E-04
94-LMNB1 HEK good	20210210	4	HEK293T	digitonin 0.02%	LMNB1	100		pA	26 aa	Hia5	527 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	62,799	407,120,248	6,483	16.014	3.81E-03	1.55E-03
95-LMNB1 HEK pA S	20210210	5	HEK293T	digitonin 0.02%	LMNB1	100		pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	33,562	297,569,878	8,866	18.086	3.28E-03	1.23E-03
96-LMNB1 HEK pA L	20210210	6	HEK293T	digitonin 0.02%	LMNB1	100		pA	26 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	27,725	284,031,573	10,245	19.494	3.78E-03	1.42E-03
97-LMNB1 HEK pAG S	20210210	7	HEK293T	digitonin 0.02%	LMNB1	100		pAG	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	31,665	307,111,011	9,699	11.343	1.75E-03	6.91E-04
98-LMNB1 HEK pAG S 1:50	20210210	8	HEK293T	digitonin 0.02%	LMNB1	50		pAG	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	40,276	349,442,835	8,676	17.398	1.94E-03	7.89E-04
99-H3K9ac HEK pAG S	20210210 +20210212	9	HEK293T	digitonin 0.02%	H3K9ac	100		pAG	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	30,374	284,046,721	9,352	1.283	3.86E-04	3.63E-04
100-cenB HEK pAG S	20210210 +20210212	10	HEK293T	digitonin 0.02%	CENP-B	100		pAG	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	29,082	273,908,986	9,419	1.603	3.88E-04	3.48E-04
101-cenC HEK pAG S	20210210 +20210212	11	HEK293T	digitonin 0.02%	CENP-C	100		pAG	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	29,845	307,968,146	10,319	1.541	4.22E-04	3.78E-04
102-centromere enrichment	20210210	12	HEK293T	digitonin 0.02%	-	-	cen enrichment	-	-	-	-	-	-	-	-	-	-	307,381	1,098,607,718	3,574	1.819	1.57E-04	1.31E-04
103-failed (run overloaded, rerun)	20210304	21	HEK293T	digitonin 0.02%	-	-	free floating Hia5	-	-	Hia5	200 nM	-	-	Straight+ BSA	500 uM	30 min	37C	87,050	355,217,390	4,081	1.099	1.16E-02	1.10E-02
104-failed (run overloaded, rerun)	20210304	22	HEK293T	digitonin 0.02%	-	-	free floating Hia5 RNAse	-	-	Hia5	200 nM	-	-	Straight+ BSA	500 uM	30 min	37C	85,984	395,520,007	4,600	1.115	1.18E-02	1.15E-02
105-failed (run overloaded, rerun)	20210304	23	HEK293T	digitonin 0.02%	-	-	free floating Hia5 SDS	-	-	Hia5	200 nM	-	-	Straight+ BSA	500 uM	30 min	37C	100,394	446,330,860	4,446	1.168	8.99E-03	8.33E-03
106-failed (run overloaded, rerun)	20210304	24	HEK293T	digitonin 0.02%	CENP-A	500	baseline	pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	65,318	350,707,424	5,369	1.787	1.26E-03	1.03E-03
107-failed (run overloaded, rerun)	20210304	2	HEK293T	digitonin 0.02%	CENP-A	500	RNAse	pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	121,755	632,710,163	5,197	1.712	1.32E-03	1.07E-03
108-failed (run overloaded, rerun)	20210304	3	HEK293T	digitonin 0.02%	CENP-A	500	SDS	pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	82,013	391,736,446	4,777	1.915	1.21E-03	1.01E-03
109-failed (run overloaded, rerun)	20210304	13	HEK293T	digitonin 0.02%	CENP-A	500	salt wash	pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	110,309	626,710,796	5,681	1.576	1.12E-03	1.00E-03
110-failed (run overloaded, rerun)	20210304	14	HEK293T	digitonin 0.02%	LMNB1	100	baseline	pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	88,855	450,221,395	5,067	5.291	5.27E-03	2.66E-03
111-failed (run overloaded, rerun)	20210304	15	HEK293T	digitonin 0.02%	LMNB1	100	RNAse	pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	147,418	666,882,572	4,524	4.093	3.90E-03	2.08E-03

Table 4.1. Summary of all DiMeLo-seq sequencing runs [continued]

index & name	seq. date	barcode	cell line	detergent	Ab	Ab dilution factor	misc	pA/G	linker len.	MTase	[pA/G-MTase]	2Ab time & temp	pA/G time & temp	act. buffer	[SAM]	act. time	act. Temp	read number	total bases sequenced	mean read len	cLAD:cLAD mA ratio	cLAD mA/A	all reads mA/A
112-failed (run overloaded, rerun)	20210304	16	HEK293T	digitonin 0.02%	LMNB1	100	SDS	pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	108,548	494,216,880	4,553	2.417	2.13E-03	1.46E-03
113-failed (run overloaded, rerun)	20210304	17	HEK293T	digitonin 0.02%	LMNB1	100	salt wash	pA	26 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	110,583	603,642,079	5,459	5.751	5.68E-03	2.96E-03
114-failed (run overloaded, rerun)	20210304	18	GM12878	-	-	-	in vivo EcoGII-CenpC	-	-	EcoGII	-	-	-	-	-	-	-	33,052	123,561,248	3,738	1.460	1.25E-03	9.95E-04
115-failed (run overloaded, rerun)	20210304	19	GM12878	-	-	-	in vivo EcoGII	-	-	EcoGII	-	-	-	-	-	-	-	97,583	261,242,129	2,677	1.375	1.20E-03	1.10E-03
116-Hia5 only	20210309	21	HEK293T	digitonin 0.02%	-	-	free floating Hia5	-	-	Hia5	200 nM	-	-	Straight+ BSA	500 uM	30 min	37C	224,739	1,954,375,136	8,696	1.168	6.34E-03	6.06E-03
117-Hia5 only RNAse	20210309	22	HEK293T	digitonin 0.02%	-	-	free floating Hia5 RNAse	-	-	Hia5	200 nM	-	-	Straight+ BSA	500 uM	30 min	37C	192,453	1,755,187,759	9,120	1.228	6.96E-03	6.45E-03
118-pA baseline CENP-A	20210309	24	HEK293T	digitonin 0.02%	CENP-A	500	baseline	pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	173,211	1,736,978,663	10,028	2.003	8.43E-05	6.12E-05
119-pA RNAse CENP-A	20210309	2	HEK293T	digitonin 0.02%	CENP-A	500	RNAse	pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	253,416	2,446,404,601	9,654	1.812	7.47E-05	6.09E-05
120-pA salt CENP-A	20210309	13	HEK293T	digitonin 0.02%	CENP-A	500	salt wash	pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	269,783	2,558,490,435	9,484	1.903	7.32E-05	6.17E-05
121-pA baseline LMNB1	20210309	14	HEK293T	digitonin 0.02%	LMNB1	100	baseline	pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	202,447	1,902,202,299	9,396	19.705	2.52E-03	9.67E-04
122-pA RNAse LMNB1	20210309	15	HEK293T	digitonin 0.02%	LMNB1	100	RNAse	pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	376,220	2,775,298,200	7,377	13.630	1.64E-03	6.31E-04
123-pA salt LMNB1	20210309	17	HEK293T	digitonin 0.02%	LMNB1	100	salt wash	pA	26 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	226,875	1,978,617,616	8,721	17.773	2.73E-03	1.13E-03
124-in vivo EcoGII-CenpC GM12878	20210309	18	GM12878	-	-	-	in vivo EcoGII-CenpC	-	-	EcoGII	-	-	-	-	-	-	-	84,905	705,554,640	8,310	1.623	8.83E-05	6.61E-05
125-in vivo EcoGII only GM12878	20210309	19	GM12878	-	-	-	in vivo EcoGII	-	-	EcoGII	-	-	-	-	-	-	-	186,218	1,222,738,245	6,566	1.071	7.38E-05	7.58E-05
126-LMNB1-500u MSAM	20210323	20	HEK293T	digitonin 0.02%	LMNB1	100	500uM SAM	pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	129,920	1,273,519,587	9,802	24.331	2.85E-03	1.04E-03
127-LMNB1-800u MSAM	20210323	21	HEK293T	digitonin 0.02%	LMNB1	100	800uM SAM	pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	800 uM	30 min	37C	138,407	1,318,014,140	9,523	34.315	2.90E-03	1.05E-03
128-H3K27ac 1:50	20210323	22	GM12878	digitonin 0.02%	H3K27ac	50		pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	130,590	1,367,470,162	10,471	0.711	7.88E-05	8.78E-05
129-H3K9ac 1:50	20210323	23	GM12878	digitonin 0.02%	H3K9ac	50		pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	202,096	1,926,148,906	9,531	1.550	2.07E-04	1.65E-04
130-H3K9ac 1:500	20210323	24	GM12878	digitonin 0.02%	H3K9ac	500		pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	135,495	1,418,334,592	10,468	1.536	3.22E-05	2.99E-05
131-H3K9me3 1:50	20210323	1	GM12878	digitonin 0.02%	H3K9me3	50		pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	187,378	1,894,695,703	10,112	1.967	1.65E-03	1.50E-03
132-H3K9me3 1:500	20210323	2	GM12878	digitonin 0.02%	H3K9me3	500		pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	197,488	1,908,431,833	9,664	1.808	1.18E-04	1.03E-04
133-CTCF 1:50	20210323	3	GM12878	digitonin 0.02%	CTCF	50		pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	226,621	2,037,821,065	8,992	0.476	1.12E-04	1.74E-04
134-CTCF 1:500	20210323	4	GM12878	digitonin 0.02%	CTCF	500		pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	205,080	2,000,935,224	9,757	0.591	7.02E-05	9.67E-05
135-nucleolin 1:50	20210323	5	GM12878	digitonin 0.02%	nucleolin	50		pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	201,905	1,911,021,346	9,465	1.449	4.35E-05	3.48E-05
136-nucleolin 1:500	20210323	6	GM12878	digitonin 0.02%	nucleolin	500		pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	500 uM	30 min	37C	176,531	1,826,421,927	10,346	1.999	3.95E-05	2.82E-05
137-LMNB1 1:50 2M	20210405	7	HEK293T	digitonin 0.02%	LMNB1	50	2M cells	pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	800 uM	30 min	37C	105,855	977,322,733	9,233	15.731	3.50E-03	1.41E-03
138-LMNB1 1:50 5M	20210405	8	HEK293T	digitonin 0.02%	LMNB1	50	5M cells	pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	800 uM	30 min	37C	80,057	694,123,573	8,670	15.069	3.97E-03	1.54E-03
139-LMNB1 1:50 10M	20210405	9	HEK293T	digitonin 0.02%	LMNB1	50	10M cells	pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	800 uM	30 min	37C	101,673	861,972,532	8,478	13.248	2.40E-03	1.03E-03
140-CTCF 1:50	20210405	10	GM12878	digitonin 0.02%	CTCF	50		pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	800 uM	30 min	37C	139,694	1,170,977,786	8,382	0.603	2.60E-04	3.49E-04
141-CTCF 1:50 secondary	20210405	11	GM12878	digitonin 0.02%	CTCF	50	secondary (GP)	pA	7 aa	Hia5	200 nM	1h RT	1 h, RT	Straight+ BSA	800 uM	30 min	37C	188,343	1,616,286,164	8,582	0.657	2.41E-04	3.04E-04
142-H3K27me3 1:50	20210405	12	GM12878	digitonin 0.02%	H3K27me3	50		pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	800 uM	30 min	37C	133,870	1,132,136,243	8,457	0.758	3.01E-04	4.87E-04
143-H3K27me3 1:50 secondary	20210405	13	GM12878	digitonin 0.02%	H3K27me3	50	secondary (GP)	pA	7 aa	Hia5	200 nM	1h RT	1 h, RT	Straight+ BSA	800 uM	30 min	37C	149,383	1,291,383,161	8,645	0.751	3.08E-04	4.47E-04
144-H3K9me3 1:50	20210405	14	GM12878	digitonin 0.02%	H3K9me3	50		pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	800 uM	30 min	37C	110,815	971,889,773	8,770	1.897	7.65E-04	7.32E-04
145-H3K9me3 1:50 secondary	20210405	15	GM12878	digitonin 0.02%	H3K9me3	50	secondary (GP)	pA	7 aa	Hia5	200 nM	1h RT	1 h, RT	Straight+ BSA	800 uM	30 min	37C	176,896	1,409,688,249	7,969	0.961	4.36E-04	5.05E-04
146-IgG 1:50	20210405	16	GM12878	digitonin 0.02%	IgG	50		pA	7 aa	Hia5	200 nM	-	1 h, RT	Straight+ BSA	800 uM	30 min	37C	169,811	1,344,179,912	7,916	1.473	1.50E-04	1.16E-04
147-Hia5 only	20210405	17	GM12878	digitonin 0.02%	-	-	free floating Hia5	-	-	Hia5	-	-	-	Straight+ BSA	800 uM	30 min	37C	94,126	831,589,618	8,835	1.123	7.98E-03	7.66E-03
																		TOTAL	**130 Gb**				

Figure 4.6. Key conditions in DiMeLo-seq protocol optimizations



Legend for Figure 4.6. Key conditions in DiMeLo-seq protocol optimizations
Barplots comparing on-target (cLAD) methylation efficiency and signal-to-background ratios for different DiMeLo-seq protocol conditions targeting LMNB1, with colors highlighting key parameters for comparison. Across the supermajority of the over 80 conditions tested, Hia5 outperforms EcoGII. In plots where conditions are labeled (arbitrarily with 'A', 'B', 'C', etc.), each condition represents identical (or effectively identical) protocol conditions run on the same day, with the only variable being the parameter investigated. Thus, it is valid to compare bars grouped into the same condition, but not across conditions. Digitonin greatly outperforms NP40 or Triton. Secondary antibodies do not help performance. Straight Lab activation buffer with BSA at 37 °C provides the best on-target methylation efficiency.

DiMeLo-seq reveals lamina association of centromeric regions
The combined results from the most optimal anti-LMNB1 conditions can be seen plotted along chr7 in **Figure 4.7**, and they correspond well with *in vivo* conventional DamID and *in vivo* DiMeLo-seq (shown in **Figure 4.2**). We can also see relatively uniform DiMeLo-seq coverage across the centromere given its long ONT reads, which is not the case for the short Illumina reads used in conventional DamID. The bottom two tracks show two background controls that indicate the relative accessibility, sequenceability, and mappability of each region of the genome. For DiMeLo-seq, this background control represents permeabilized nuclei that were treated with free-floating Hia5 (as in Fiber-seq, Stergachis et al. 2020). For conventional DamID, this represents cells that expressed untethered Dam *in vivo*. In this chromosome's centromere, the advantage of long reads is clear, and it is also clear that the centromere is sufficiently accessible to be methylated by Hia5. The anti-LMNB1 methylation data suggest that this centromere is not strongly lamina associated, which aligns with broad observations that centromeres are often not preferentially associated with the nuclear lamina in mammalian cells, unlike certain other clades (reviewed by Hoskins et al. 2021).

If we examine the centromere of a different chromosome, chromosome 3 (**Figure 4.7b**), we see evidence of strong lamina association as well as a more pronounced dip in mappability and a dip in apparent accessibility at the centromere. This underlines the need to produce longer reads at higher coverage in centromeric regions. However, this centromere does exhibit a robust signal of lamina association, apparent in the DiMeLo-seq data, which cannot be ascertained from the DamID data. This appears to be the strongest signal of lamina association at any centromere, and on close examination it appears that this may be related to the unusual nature of this

centromere's organization. The alpha satellite of centromere 3 does not occur in one contiguous block but is divided into two pieces by a 2.5 Mb array of a different satellite DNA family, Human Satellite 1A (HSat1A), which is not known to be directly related to centromere function. Chromosome 4 has a similar centromeric organization, and it also appears to have a peak in lamina association in its own intervening HSat1A array, which diminishes inside the alpha satellite arrays. Obtaining higher coverage with even longer and more mappable reads in these regions will allow us to dissect these differences at much finer resolution.

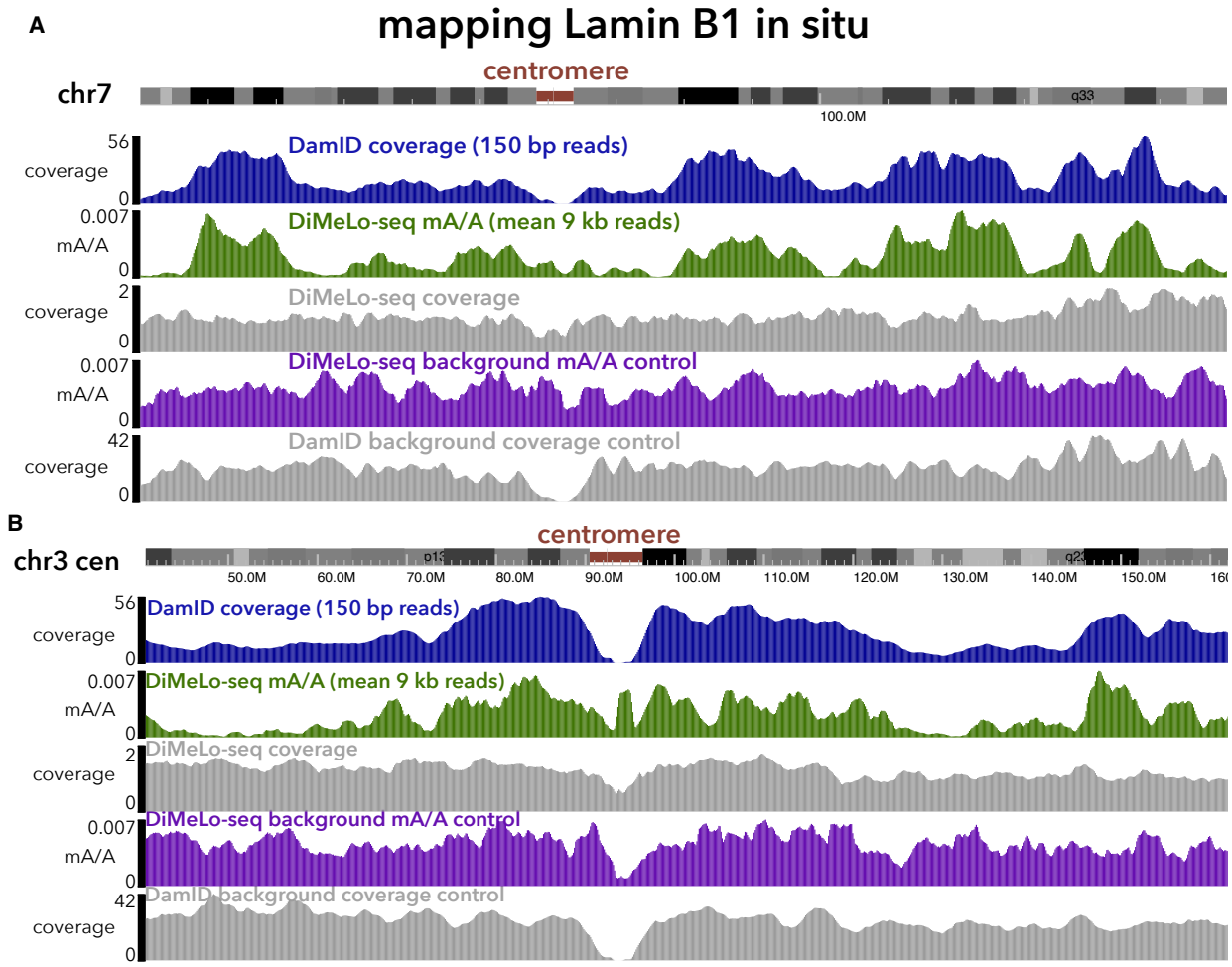


Figure 4.7. Browser tracks showing DiMeLo-seq LMNB1 results
(**A**) A browser screenshot across all of chr7 from the T2T-chm13 reference sequence, comparing conventional bulk short-read DamID coverage to DiMeLo-seq methylation levels. The centromere region has far more coverage in the centromere region with DiMeLo-seq compared to DamID. The purple track provides a measure of DNA accessibility, similar to Fiber-seq. (**B**) A similar browser view, but of the central region of chr3, illustrating a centromeric LAD detectable only by DiMeLo-seq but not DamID.

Moving toward new protein targets

Satisfied that our *in situ* DiMeLo-seq protocol allowed us to methylate on-target DNA regions with high efficiency while maintaining low background methylation, we next wanted to test that this protocol would work well for target proteins other than LMNB1. We chose several histone marks to test in GM12878 cells, for which an abundance of ChIP-seq data are available for comparison (T. E. P. Consortium 2012): H3K9me3, H3K27me3, H3K9ac, and H3K27ac. We sequenced these at low coverage, with the hope that by looking at averaged profiles around many ChIP-seq peaks, we could see if there was specific localization (**Figure 4.8**). We also ran IgG isotype controls and Hia5-treated nuclei to probe background and DNA accessibility. To do this, we compared the methyl and acetyl marks of each type, which are not expected to overlap. The acetyl marks (H3K9ac & H3K27ac) are found in active promoters and enhancers, while the methyl marks (H3K9me3 & H3K27me3) are found in transcriptionally silenced regions (T. E. P. Consortium 2012).

When the average adenine methylation probability profile for H3K27me3 is plotted around H3K27me3 ChIP-seq peaks from GM12878 (**Figure 4.8**), we see a broad increase over the methylation levels observed for IgG around those same sites, but we do not see this same magnitude of increase when H3K27me3 is centered around H3K27ac peaks. This is consistent with H3K27me3 having a broad, spreading distribution. We also do not see an increase in DNA accessibility (Hia5 methylation) around H3K27me3 peaks, as expected. Similarly, for H3K27ac we see a strong peak above its baseline when centered at H3K27ac ChIP-seq peaks, but not when centered at H3K27me3 peaks. However, we do see an increase in IgG and Hia5 signal at H3K27ac peaks, as expected since they tend to be more accessible. We see similar patterns for H3K9me3 vs H3K9ac. This does confirm that DNA accessibility plays a role in driving background methylation in DiMeLo-seq, as it does for DamID, and this underlines the need to develop a method for normalizing to an IgG or Hia5 control, just as DamID is normalized to an untethered Dam control, or ChIP-seq is normalized to an input control. We are continuing to explore the best method for doing this.

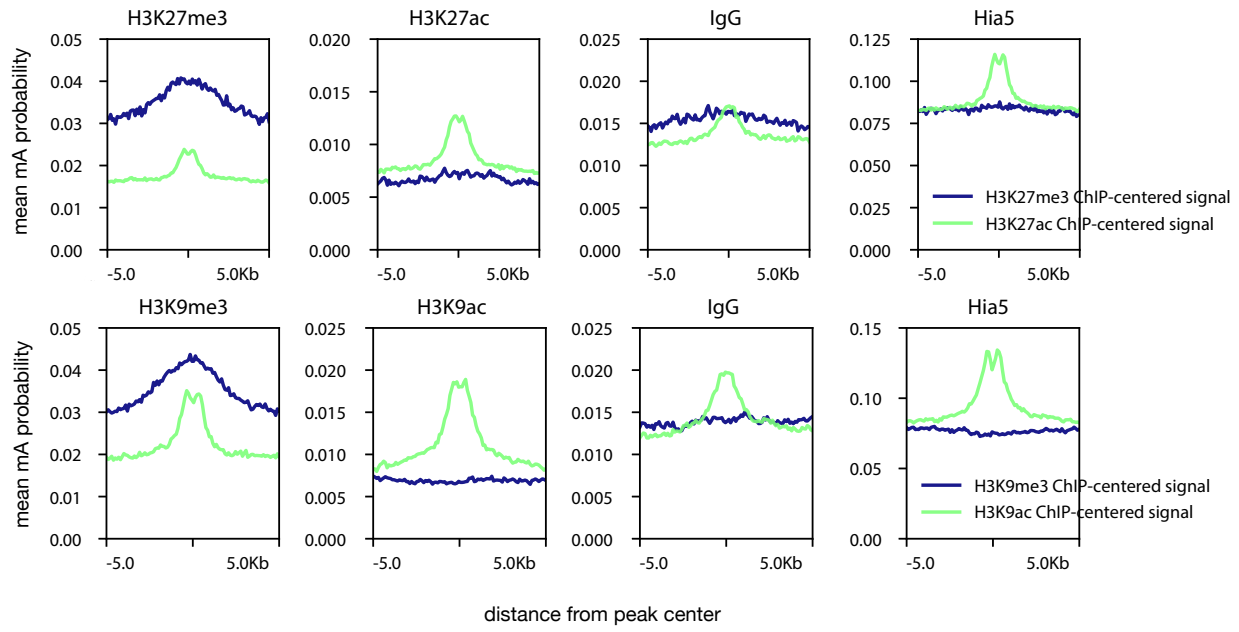


Figure 4.8. Averaged DiMeLo-seq profiles for 4 histone marks

Averaged profile plots of DiMeLo-seq methylation probability scores centered at all ENCODE ChIP-seq peaks for H3K27me3 (top row, blue lines), H3K27ac (top row, green lines), H3K27me3 (bottom row, blue lines), and H3K9ac (bottom row, green lines) in GM12878 cells (the same cell line that DiMeLo-seq was performed in for these targets). The DiMeLo-seq target protein is printed above each plot. "IgG" is an isotype control antibody that provides a measure of nonspecific binding, and shows a small increase at H3K27ac and H3K9ac ChIP-seq peaks, which are expected to be more accessible. "Hia5" (last column) represents untargeted, free-floating Hia5 methylation (as in Fiber-seq), that serves as a measure of DNA accessibility. Plot credit: Annie Maslan.

We next explored the genome-wide distribution of the H3K9me3 mark, since it is expected to be found at centromeres (**Figure 4.9**). Indeed, we do see a strong peak of H3K9me3 overlapping the centromeric alpha satellite regions on every chromosome. Notably, there are still some regions of the genome that are not well covered in this experiment; these mostly correspond to large regions occupied by Human Satellites 1, 2, and 3, which are among the most unmappable regions of the genome and are found on chromosomes 1, 3, 4, 9, 16, and the acrocentrics. This low coverage stems from the fact that we traded shorter read lengths for higher throughput in these initial experiments, achieving a mean read length of only ~9 kb, and we mapped reads to the reference using a mapping quality threshold (mapq) of 10. For reads mapped with mapq 10, we can expect 10% of them to be mismapped, but it is certain that each read above this threshold does not have more than one mapping location tied for the top score. We plan to adjust our library prep strategy to increase average read lengths substantially, which should improve mapping to these regions. Poor mapping quality can also result from basecalling errors. To address this, we also plan to re-basecall reads using the more accurate Bonito algorithm, which is available for beta testing but is far slower than Guppy or Megalodon.

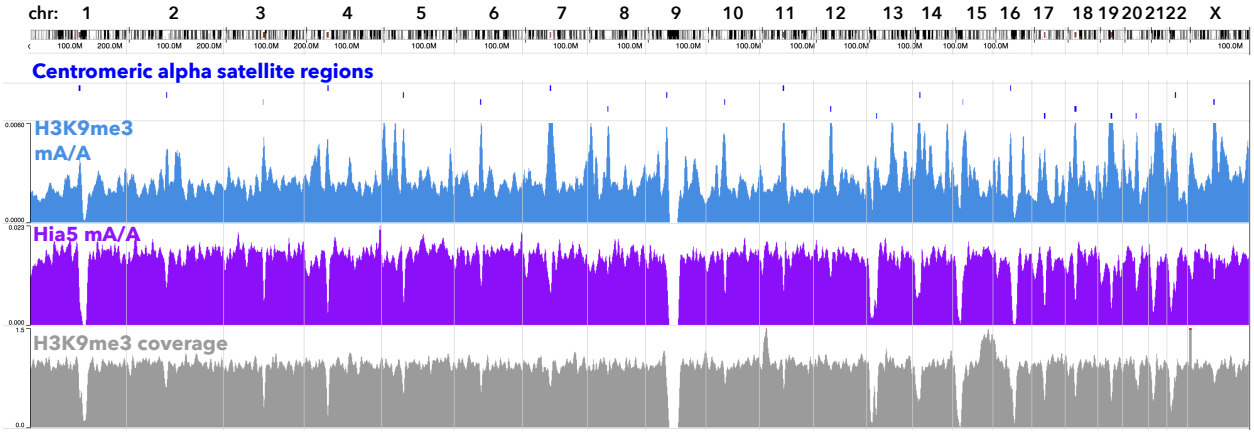


Figure 4.9. DiMeLo-seq confirms H3K9me3 enrichment at centromeres
Browser tracks across the entire genome showing the DiMeLo-seq methylation levels for H3K9me3 (blue track) and free-floating Hia5 (purple track). Chromosome ideograms are displayed on the top track, and the locations of centromeric alpha satellite arrays are displayed on the second track. The coverage track (gray) shows the coverage obtained after stringent alignment filtering (mapq >10). Since reads in this sample only have a mean length of 9 kb, some of the large pericentric satellite arrays remain unmappable.

To test DiMeLo-seq on a protein with a very small binding footprint (in contrast to LMNB1), we applied it to map the locations of CTCF, a zinc finger protein that binds small motifs at thousands of sites in the genome and plays an important role in nuclear architecture. We examined averaged DiMeLo-seq profiles around CTCF binding sites identified by ChIP-seq in GM12878 cells, and we broke these ChIP-seq peaks into four quartiles by their ChIP-seq signal (**Figure 4.10a**). Because CTCF tends to strongly phase the nucleosomes around it, and because Hia5 preferentially methylates linker DNA between nucleosomes, we observe a characteristic decaying wave pattern indicating the positions of these nucleosomes—a pattern commonly seen in ChIP-seq, CUT&RUN, and ATAC-seq data around CTCF binding sites. We did not observe the same pattern for the IgG control, but we did see it for the Hia5-only sample, as expected given the increased accessibility of these regions. The strength of methylation also increased with the strength of ChIP-seq signal, consistent with the expectation that DiMeLo-seq signal should quantitatively capture protein-DNA interaction frequency.

Finally, to examine the spatial resolution achievable with DiMeLo-seq, we plotted methylation probability scores as a heat map, with rows corresponding to individual CTCF binding sites with the highest methylation probabilities in the surrounding 2 kb region (**Figure 4.10b**). Because coverage is so low in this sample (1X), we only expect 1 read to overlap most of these peaks, providing a single-molecule view of methylation. It is clear that the reach of the methyltransferase decays completely to baseline around 500 bp from the peak center, but 60% of this decay happens within 100 bp. This sharp decay will likely help us to resolve the peak center much more finely as we begin to develop *de novo* peak calling algorithms for these data. These results encouragingly show that DiMeLo-seq can provide binding site resolution on the order of hundreds of base pairs, increasing the domain of proteins that can be usefully mapped with this method.

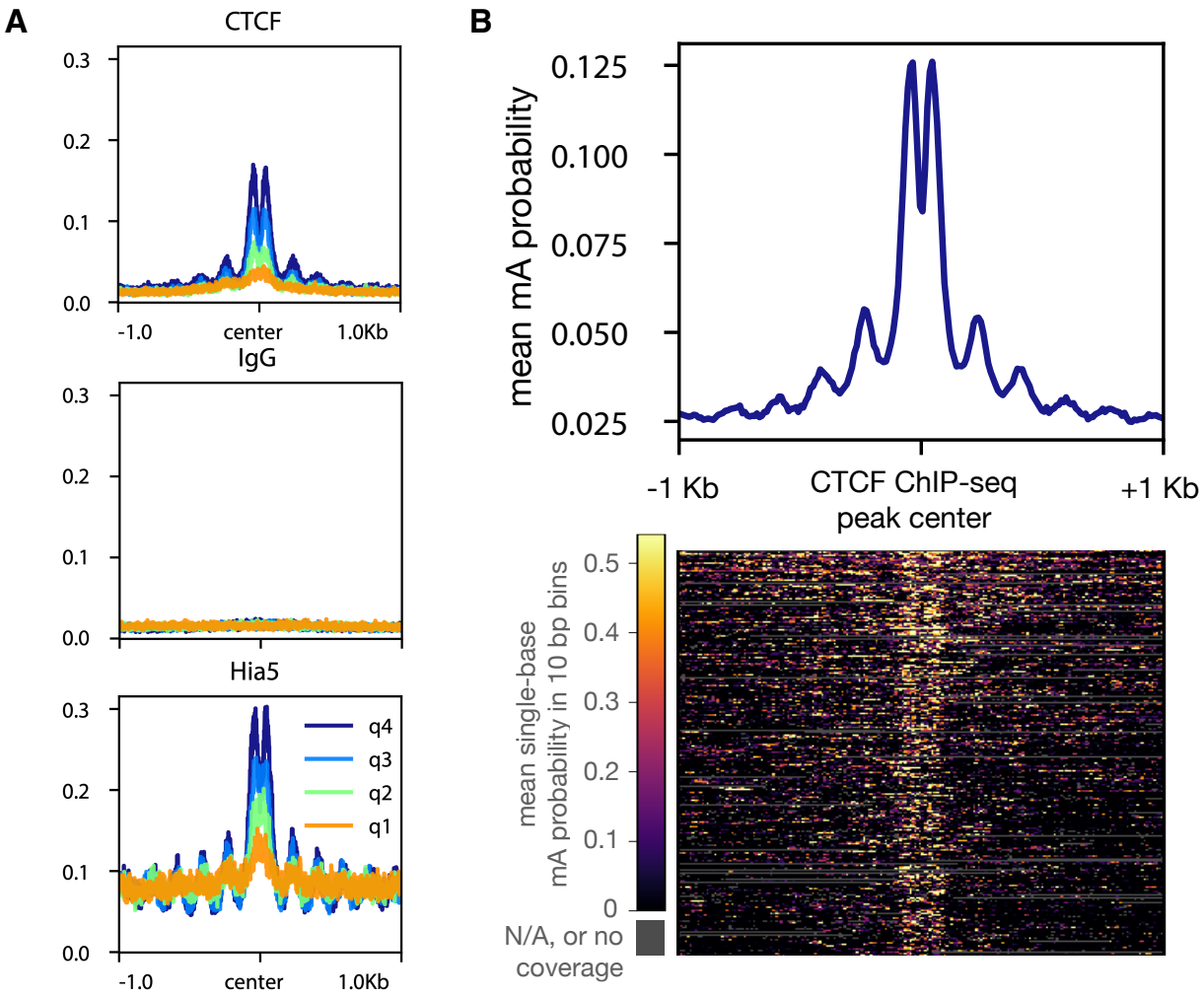


Figure 4.10. DiMeLo-seq profiles around CTCF binding sites
(**A**) Averaged DiMeLo-seq profile plots surrounding all ENCODE CTCF ChIP-seq peaks in GM12878 cells, broken into quartiles by ChIP-seq peak strength (q4 being the strongest ChIP-seq peaks; dark blue lines). Results around the same sites are shown for DiMeLo-seq targeting CTCF (top), IgG (middle), or untargeted free-floating Hia5 (bottom) in GM12878 cells. The strongest ChIP-seq peaks are also the strongest DiMeLo-seq peaks, though some of this correlation is attributable to DNA accessibility. (**B**) A similar profile plot, but with the top quartile of individual CTCF ChIP-seq peaks displayed below, one site per row, and colored according to the mean mA probability score in 10-bp bins. Because coverage is low (1X), most of these rows (expected 70%) represent single reads, providing a single-molecule view of DiMeLo-seq signal and resolution. Plot credit: Annie Maslan.

Centromere enrichment strategy

Alpha satellite repeats constitute only 1-5% of the human genome (Miga 2019). If one used the standard DiMeLo-seq approach to map proteins binding within alpha satellite arrays, like CENPA, CENPB, and CENPC, as we plan to do, then the remaining 95% of the genome would be sequenced needlessly, increasing the time and cost of sequencing. Because DiMeLo-seq relies on ^{m6}A marks, which cannot be replicated, specific regions of the genome cannot be targeted and amplified while retaining protein-DNA interaction information. While there are several targeted nanopore sequencing methods that do not require amplification, they are either not well suited to targeting large repetitive regions (Gilpatrick et al. 2020), or they require advanced hardware to basecall and align reads in real time while rejecting off-target reads, which is likely to result in lower throughput (Kovaka et al. 2021). I wanted to develop a method of enriching the input material itself for alpha satellite DNA, and I devised a way to do it that leverages the repetitive nature of satellites. Because satellite repeats are relatively short and homogeneous, short DNA k-mers are not uniformly distributed throughout these regions. In fact, some k-mers are completely absent from some families of repeats; for example, GATC is missing from many large repetitive regions, as seen in **Figure 2.1**. If one could digest the genome with a restriction enzyme that cuts motifs found commonly outside alpha satellite regions, but rarely inside them, then one could remove short digested DNA fragments by size selection, and mostly long, undigested alpha satellite DNA would remain (**Figure 4.11a**).

To see if this approach would be feasible, I simulated digestion of the T2T chm13 reference sequence with a set of all restriction enzymes available from New England Biolabs that had 4-6 bp cut sites and that were annotated as being insensitive to CpG or Dam methylation. Of those, I selected 28 enzymes for which fewer than 5% of fragments mapped to alpha satellite, and for which the genome was digested into at least 200,000 total fragments. I then removed all simulated fragments under 10 kb, to simulate a size selection process, and I computed the fraction of remaining fragments mapping to centromeres. This allowed me to estimate the theoretical enrichment of centromeric sequences as well as the systematic loss of centromeric sequences predicted to be digested into fragments under the size cutoff. I also tested double and triple digest combinations of the top 3 enzymes with >4-fold alpha satellite enrichment and <3% alpha satellite loss (MscI, AseI, PvuII), and I found the best possible overall enrichment regime to be a double digest with MscI and AseI, predicted to yield 18-fold enrichment of alpha satellite with only 0.8% systematic loss of alpha satellite (**Figure 4.11b**).

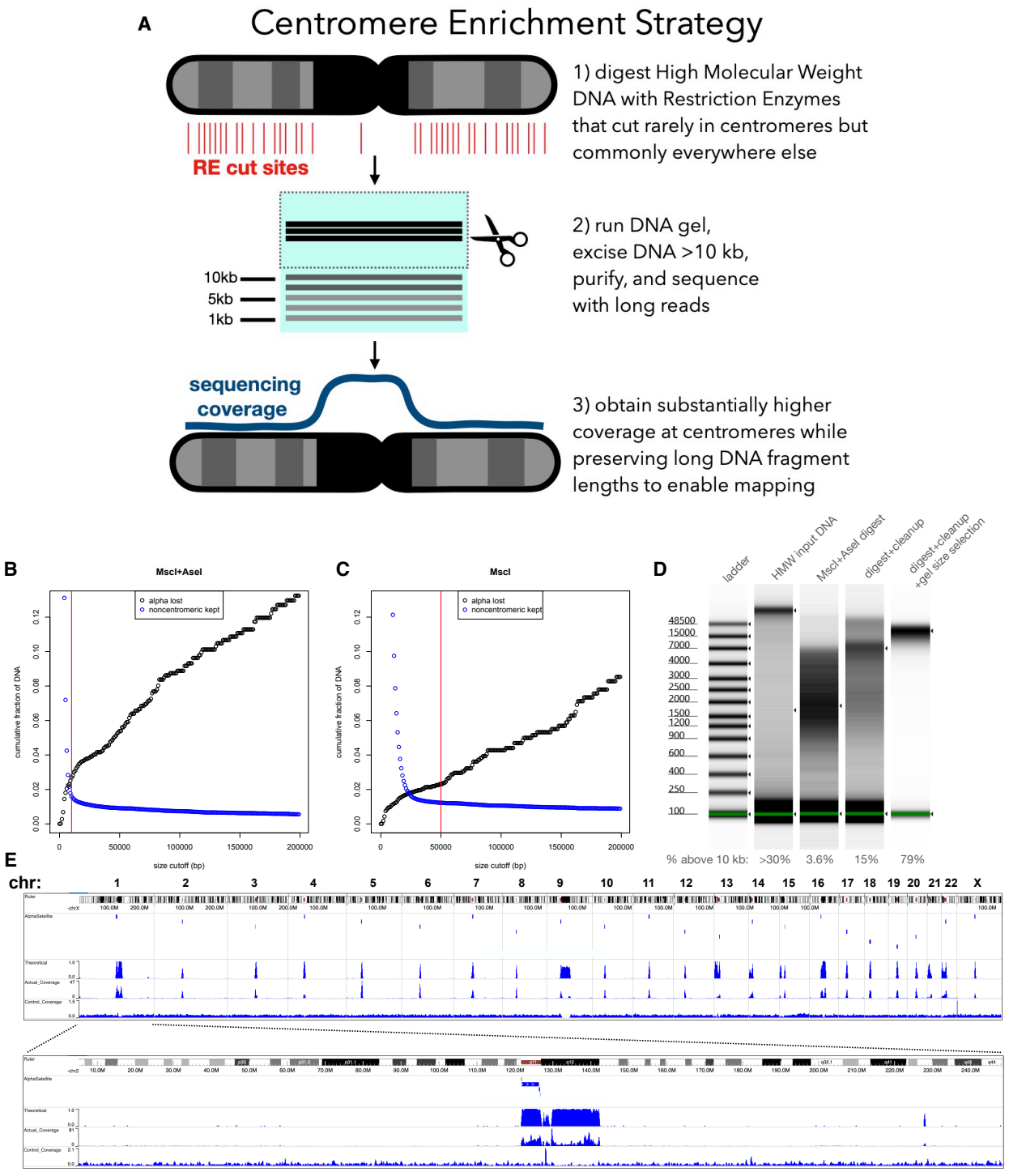


Figure 4.11. Centromere enrichment by restriction digestion and size selection (**A**) Illustration of the overall centromere enrichment strategy. (**B**) Simulated tradeoff between loss of sensitivity (% of alpha satellite lost) and gain of specificity (% of non-centromeric sequences not removed) as different size cutoff thresholds are used on genomic DNA digested with MscI+AseI (red line at 10 kb cutoff). (**C**) As in B but for

gDNA digested with MscI only. Red line is shown at 50 kb. (**D**) Tapestation results illustrating the change in size distribution through the steps of the enrichment protocol. (**E**) Browser tracks illustrating the location of alpha satellite higher order repeat arrays (1st track); theoretical coverage assuming perfect size selection, perfect recovery, perfect mappability, and no sequencing bias (2nd track); actual coverage from DNA isolated by this strategy (3rd track); coverage from un-enriched gDNA sequencing (4th track). Overall, there is 20-fold higher coverage in centromere regions in the enriched track vs control track.

To test this approach in practice, I digested ~100 µg high-molecular-weight (HWM) DNA isolated from ~25M HEK293T cells overnight with MscI and AseI, then cleaned up the digest with a column that depletes fragments under 3 kb (Zymo gDNA Clean & Concentrator Kit), yielding 15 µg. Early attempts to perform size selection with a Circulomics Short Read Eliminator XS kit resulted in extremely low yields. Instead, I loaded this onto a 0.3% TAE + agarose gel (using SeaKem Gold agarose, which is specialized for large fragment separations) and ran it at low voltage (2 V/cm) until fragments under 10 kb were visibly separated from a visible HMW band (~1 h). I cut out everything above 10 kb, including the loading well, and purified the DNA using a Zymo Large DNA Fragment Recovery kit (with modifications detailed in Methods). This yielded ~1.8 µg of DNA, which we library prepped and sequenced on a MinION device. By mapping reads back to the reference sequence, I observed ~20-fold enrichment of alpha satellite sequences (**Figure 4.11e**). Specifically, while alpha satellite higher order repeats constitute only 2.3% of the genome, reads overlapping these regions represented 46.2% of bases on all mapped reads. This means a single 72 hour, <$1500, 20 Gb run on a MinION flowcell could yield ~130X coverage of alpha satellite regions, which is enough to split over many DiMeLo-seq samples. Without enrichment, obtaining this same coverage on a single MinION would require 2 months and $30k.

While this approach does produce longer reads compared to an average sequencing run (empirical N50=19 kb vs typical run with N50=15 kb), ideally these reads would be longer to provide greater mappability and more joint single-molecule information for centromere-mapping reads. Some of this loss of read length is due to the choice of a double digestion followed by 10 kb size selection. While most alpha satellite DNA is predicted to be in fragments well over 10-kb, the nanopore sequencing device tends to favor sequencing shorter DNA fragments, so the final distribution of read lengths is almost always shifted to the left relative to the distribution of input DNA fragment

lengths. If I were to cut with MscI alone and perform size selection at 50 kb, I could theoretically achieve 25-fold enrichment with only 2% systematic loss of alpha satellite DNA (**Figure 4.11c**), while likely increasing the median read length substantially.

Loss of fragment length can also occur during gel purification or ligation-based library prep, which can shear DNA during column-based or bead-based purification steps. Another issue is that the reaction cleanup prior to gel loading, which depletes fragments smaller than 3 kb, would not deplete as much of the single-digest DNA compared to the double-digest DNA. Thus, in order to prevent anomalous band migration due to gel overloading, more gel lanes would need to be used, which can result in lower yields. To address these issues, moving forward I will test electroelution as an alternative method for recovering DNA from extracted gel fragments while preserving long fragment lengths (Strong et al. 1997). I further plan to test ONT's ligation-free library prep kits recommended for ultralong read library prep. I am also interested in developing methods for performing library preparation reactions on HMW DNA trapped in an agarose gel slice, followed by electroelution of purified, library-prepped DNA ready for sequencing. Because buffer exchange is simplified by immobilizing DNA in a hydrogel, this would remove any need to purify the DNA by precipitation and long rehydration steps, as is done for existing ultralong library prep protocols.

Discussion and next steps

Here, we have developed, optimized, extended, and validated DiMeLo-seq, a new method for mapping protein-DNA interactions genome-wide. DiMeLo-seq can map a protein's binding sites within hundreds of base pairs on single molecules of sequenced DNA up to hundreds of kilobases in length. This long read length improves mappability in repetitive regions of the genome, opening them up to new studies of their regulation. Because DiMeLo-seq involves no amplification, it can provide a linear readout of protein-DNA binding frequency at every site. It can also provide joint CpG and protein-DNA interactions on the same long single molecules. Now that the long process of basic protocol development has concluded, we are excited to explore all of the advantages of DiMeLo-seq and apply it to some interesting biological questions.

This work has thrived as a joint effort with Owen Kabnick Smith, Dr. Kousik Sundararajan, and Rachel Brown in Aaron Straight's Lab, who provided us with copious amounts of Hia5/pA-Hia5/pAG-Hia5, and with key reagent formulas used in the final protocol, like their activation buffer. They have also performed a tremendous amount of work characterizing the DiMeLo-seq workflow on reconstituted chromatin *in vitro*, generating high-coverage data from an extremely well-controlled system. In doing so,

they proved the specificity of antibody-targeted methylation, developed a method to detect targeted nucleosome positions on single molecules, and performed important controls to validate the methylation calling pipeline. Because they have expertise in centromere biology, we are excited to partner with them to map centromere proteins at high resolution using DiMeLo-seq. To begin, we are generating high-coverage, long-read ONT sequencing data for H3K9me3 and CENP-A, which are both found in centromeres and are expected to be non-overlapping (McNulty & Sullivan 2018). We hope to discover how CENP-A nucleosomes are distributed throughout different subregions of active alpha satellite arrays in every human centromere, as well as how boundaries between H3K9me3 and CENP-A vary among centromeres and between cells and cell types. The ability to jointly map multiple CENP-A nucleosomes on the same single molecules will prove essential for this endeavor.

We are also generating high-coverage data for CTCF, to provide ample training material that will allow us to develop new *de novo* single-molecule peak calling algorithms for this new data type. We plan to jointly analyze endogenous CpG methylation on these reads, as we expect CpG methylation in CTCF's binding motif to abolish its binding (H. Wang et al. 2012). In parallel, we are sequencing these samples with PacBio sequencing, which is expected to provide more accurate basecalls and methylation calls, but on reads only up to 20 kb, and without joint CpG information. We have also begun developing a way to multiplex DiMeLo-seq, by fusing two different proteins to two different methyltransferases, for example, one that methylates adenines and one that methylates GpC cytosines. This would allow us to jointly map two proteins and endogenous CpG methylation on the same long single molecules of DNA, which could be used to study phenomena like heterochromatin spreading. Future extensions of this work would enable us to map RNA-DNA interactions, as in ChIRP or RNA-DamID (Cheetham & Brand 2018, Chu et al. 2011), or DNA-DNA interactions, as in 4 °C (Zhao et al. 2006).

Detailed materials and methods

Creation and induction of stable cell lines for in vivo DiMeLo-seq

Stable HEK293T and GM12878 cell lines were created by retroviral transduction followed by drug selection. Retroviral plasmids containing EcoGII, EcoGII-LMNB1, and EcoGII-CENPC were obtained from Addgene (#122082, #122083, #122085; Sobecki et al. 2018). These plasmids were modified to create versions with Dam or Dam-LMNB1. Retroviruses were produced in the Phoenix Ampho packaging cell line (obtained from the UC Berkeley cell culture facility). Phoenix cells were seeded in standard growth medium (DMEM with 10% FBS and 1X P/S) in a T75 flask 24 hours before transfection, aiming for 70% confluence at the time of transfection. 25 µg of plasmid DNA was combined with 75 µl FUGENE-HD transfection reagent in 1200 µl optiMEM and incubated for 10 minutes, then added to the media. After 12 hours, the media was replaced with fresh media, and the cells were incubated at 32 °C with 5% CO_2 and 100% humidity to help preserve viral particles. 36 hours later, the virus-containing media was harvested and centrifuged at 1800 rpm for 5 minutes to remove any Phoenix cells. The media was supplemented with 10 µl/ml of 1 M HEPES and 4 µg/ml of polybrene. For HEK293T cells, 2.5 ml of this media was added to each well of a 6-well plate containing adhered cells at 40-50% confluence. For GM12878 cells, 1.5 million cells were resuspended in 3 ml of virus-containing media and added to each well. Plates were spinoculated in a centrifuge with a swinging-bucket plate rotor at 1300x*g* for 1 hour at room temperature, then incubated at 37 °C overnight. The media was replaced the next morning. After 24 hours, puromycin was added to the media at a concentration of 1µg/ml and the media was replenished every 48 hours for 10 days. Surviving cells were expanded and frozen for later use. 15 hours prior to harvesting, Shield-1 reagent was added to the media to stabilize protein expression, either with 1 µM Shield-1 for full-induction, or 10 nM Shield-1 aiming for ~10% of full induction levels.

In situ DiMeLo-seq and immunofluorescence

Please see **Appendix 2** for detailed protocol.

Basecalling, modification calling, and data analysis

All sequencing was performed on ONT MinION v9.4 flow cells. Basecalling and modification calling were performed on Amazon Web Services g4dn.metal instances, which have 8 NVIDIA T4 GPUs, 96 CPUs, 384 Gb memory, and 2x900 Gb local solid-state storage; this configuration allows for efficient parallelization and high basecalling speed. Basecalling was first performed using Oxford Nanopore Technologies's Guppy software (v4.5.4), using a Rerio res_dna_r941_min_modbases-all-context_v001.cfg basecalling model, and demultiplexing when appropriate. Modification calls were

extracted from fast5 output files using ont-pyguppy-client-api. Basecalled reads were aligned to the T2T-chm13v1 reference sequence using Winnowmap (v2.03), which is adapted to perform better than other long-read aligners in repetitive regions (C. Jain et al. 2020). Fast5 files were split by barcode using fast5_subset then re-basecalled using ONT's Megalodon software (v2.3.1), using the same reference and model file. Custom code was used to parse output files. To evaluate performance, cLAD and ciLAD coordinates were lifted over from hg38 to the chm13 reference. A read was assigned to a cLAD or ciLAD bin if it overlapped the bin with more than 50% of its length, and any mA calls on that read were assigned to that bin. Profile plots were made using deepTools2 (Ramírez et al. 2016), after lifting over ENCODE GM12878 ChIP-seq peaks from hg38 to chm13. Browser plots were made using the WashU Epigenome Browser (D. Li et al. 2019).

Centromere Enrichment

Genomic DNA was extracted from ~25 million cells using an NEB HMW DNA extraction kit (#T3050L). The DNA was eluted in a total of 300 µl elution buffer and allowed to relax at 4 °C for 2 days, although it remained viscous until it was digested. 37 µl NEBuffer 2.1 was added, along with 100 units of MscI and 100 units of AseI (NEB #R0534M and #R0526M) to a total volume of 370 µl in a 1.5 ml lo-bind Eppendorf tube. This was placed on a rotator at 12 rpm at 37 °C overnight. DNA was purified from the reaction using a Zymo genomic DNA clean & concentrator 10 kit (Zymo #D4010), then quantified using a Qubit Broad Range DNA kit (Thermo Fisher #Q32850). DNA was then mixed with orange loading buffer and loaded on a 0.3% TAE agarose gel made with Lonza SeaKem Gold agarose (# 50512) and 15 µl SYBRSafe gel stain (Thermo Fisher #S33102) per 100 ml gel. A GeneRuler High Range DNA Ladder (Thermo Fisher SM1351) was loaded in an adjacent lane. To avoid overloading, DNA was loaded with no more than 250 ng per mm of lane width. The gel was run at 2 V/cm for 1 hour and imaged over a blue light transilluminator. The gel was cut to remove fragments smaller than 10 kb, while keeping everything larger, up to the well itself. DNA was purified from the resulting gel slice using a Zymoclean Large Fragment DNA Recovery Kit (Zymo # D4045), with modifications: the gel slice was melted at room temperature on a rotator at 12 rpm, and DNA was eluted from the column twice with elution buffer heated to 70 °C. The DNA was then quantified by Qubit again. DNA was prepared for sequencing using an ONT LSK-110 native library prep kit, and sequenced on a v9.4 MinION flow cell.

Chapter 5

Conclusion

Our collective understanding of genome regulation has advanced greatly in the last two decades, owing to the Human Genome Project and to powerful sequencing-based methods for measuring gene expression, DNA accessibility, and protein-DNA interactions. More recently, single-cell methods have enabled researchers to dissect differentiation processes and single-cell heterogeneity with unprecedented granularity, although these studies have primarily employed methods like single-cell RNA-seq and single-cell ATAC-seq, which have matured faster than technologies designed to measure specific protein-DNA interactions in single cells. Nearly all studies of the human genome, be they in bulk tissues or single cells, have ignored repetitive regions, which have remained missing from the human genome assembly until now. In the next decade, telomere-to-telomere genome assemblies will become routine, and the repetitive regions of the genome can no longer be ignored. Studying these regions will require new tools that can fully leverage the power of new long-read sequencing technologies. In this context, I have engineered new methods to advance our ability to measure protein-DNA interactions in single cells and in repetitive regions of the genome.

Outlook for μDamID

Firstly, I engineered μDamID, an integrated microfluidic device that allows the user to isolate single cells, image them at high resolution, sort them for processing, and then perform single-cell DamID on them to map protein-DNA interactions. The output data are paired imaging and sequencing measurements of protein-DNA interactions within single cells, giving a readout of both the spatial and sequence coordinates of these interactions in the nucleus. μDamID is compatible with any imaging modality that can be implemented with a standard inverted microscope, including 2-photon fluorescence and other nonlinear optical microscopy techniques, in addition to many common super-resolution microscope configurations that take advantage of photoactivatable fluorescent proteins (reviewed by Huang et al. 2010). However, the thickness of the PDMS device may be incompatible with optics above the sample, such as condensers, that have short working distances, in which case a thin-chip design might be considered (Kim et al. 2021). The device can be mounted on a coverslip and

therefore can be imaged using high-NA objectives, as done here. In this study, we demonstrate short-term live-cell imaging with μDamID; however, imaging modalities that require fixed cells or nuclei are also compatible, although fixation may affect sequencing yields. The flexibility of integrated microfluidic circuits provides compatibility with imaging techniques that require multiple wash steps such as *in situ* hybridization (reviewed by Rodriguez-Mateos et al. 2020), as well as time-lapse imaging of live cells prior to DamID processing (Ramalingam et al. 2016). However, these implementations would require modifications to the device design.

μDamID can also be applied to study many other types of protein-DNA interactions in single cells, and it could be combined with other sequencing and/or imaging modalities to gather even richer information from each cell. For example, the nuclear localization of specific proteins such as heterochromatin-associated proteins or nucleolus-associated proteins can be visualized by fluorescent tagging, and then DamID can be used to sequence and identify nearby genomic regions. This device could readily be applied to study chromatin organization in micronuclei and other abnormal nuclei by imaging and selectively sorting these nuclear phenotypes and performing DamID, which would be infeasible by bulk or FACS-based methods. Recent advances allow for simultaneous DamID and transcriptome sequencing in single cells (Rooijers et al. 2019). The integrated valves and modular reaction chambers in the μDamID device could be leveraged to extend this platform to such multi-omic protocols. This would allow for joint analysis of spatial chromatin organization, protein-DNA interactions, and gene expression within single cells. Further improvements to the DamID protocol may also increase its sensitivity and specificity.

Although the sequencing throughput of this particular design is limited to 10 cells per μDamID device, many hundreds of cells can be rapidly screened and rejected from the device by monitoring a wide field of cells entering the input filter area. Thus, even relatively rare cell phenotypes can be enriched and sequenced on the device. We note that the rate-limiting step is often high-resolution image acquisition, which can take minutes per cell depending on the imaging method. The throughput of this platform can be increased to hundreds of cells per device by scaling up the design and incorporating features like multiplexed valve control (Kim et al. 2017) and automated image processing and sorting. To scale this technology further, paired imaging and sequencing data could be obtained using spatially or optically registered DNA barcodes (T. N. Chen et al. 2020, Cole et al. 2017, Nguyen et al. 2017, Yuan et al. 2018).

One important limitation of μDamID is that, like any single-cell methods requiring cell suspensions as input, it destroys any spatial information related to the positioning of

cells within a tissue. The next frontier will be to develop spatial 'omic methods for mapping protein-DNA interactions *in situ*. One can imagine, for example, using antibody targeting to insert sequencing adapters near a protein's binding sites *in situ*, as in CUT&Tag, but then performing *in situ* DNA sequencing (Payne et al. 2021).

Outlook for DiMeLo-seq

The idea for DiMeLo-seq stemmed from my experience with DamID (Altemose et al. 2020), my expertise in characterizing repetitive DNA sequences computationally (Altemose et al. 2014), and my involvement as a member of the Telomere-to-Telomere (T2T) Consortium, which is using long-read sequencing data to complete the human genome reference assembly (Miga et al. 2020). As I began to annotate newly assembled repetitive regions in the new T2T reference assembly as a side project, it became clear that there was a strong need for a long-read sequencing method to map protein-DNA interactions in highly repetitive regions of the human genome. To address this need, I conceived of the DiMeLo-seq method, and I have worked collaboratively to implement, optimize, and extend it over the past 6 months. After developing a rapid experimental and computational pipeline to evaluate method performance with low-coverage sequencing, we iterated through scores of conditions and improved the efficiency of the method by multiple orders of magnitude. We also demonstrated its efficacy and utility on several different protein targets, showing that we can achieve binding site resolution on the order of hundreds of base pairs. Now we are applying the method to map the locations of important protein-DNA interactions within centromeric alpha satellite repeats, aided by an approach I developed to enrich for centromeric sequences. In the process, I am developing the analytical and computational tools we need to normalize our data, call peaks *de novo*, and estimate our sensitivity and specificity, as I have done in the past for other sequencing data types (Altemose et al. 2017).

Although development of DiMeLo-seq was originally motivated by a desire to study repetitive regions of the genome, it offers a number of powerful advantages likely to make it useful for other applications. Firstly, it is a single-molecule method, which allows the user to jointly map proteins on long (>100 kb) single molecules of DNA. This can allow one to explore the density of a protein's binding along a single chromatin fiber from a single cell; for example, to study how "spreading" chromatin states travel along DNA, or perhaps how the stoichiometry of a DNA-binding protein in an enhancer affects the transcription of nearby genes. Secondly, because the approach involves no amplification, it should provide a linear readout of a protein's binding frequency at any site in the genome. That is, at any one locus, each overlapping read represents a single cell, so one can estimate things like single-cell contact frequency locally based on bulk

data. Thirdly, the method provides joint endogenous CpG methylation information, which can provide insights into how CpG methylation affects protein-DNA binding and local chromatin states. One can also imagine adding exogenous GpC methylation marks to provide information about DNA accessibility or a second protein's joint binding profile.

One limitation of DiMeLo-seq is that it cannot map protein-DNA interactions genome-wide in single cells. Long-read sequencers require a lot of input DNA, only a tiny fraction of which actually gets sequenced, and DNA methylation marks cannot currently be amplified. Thus, DiMeLo-seq, for now, is a bulk sequencing method only. To address this, it may be possible to convert these marks into mutated bases then amplify long, barcoded DNA from single cells (Yibin Liu et al. 2020). It may also be possible to develop methods to copy methylation marks to newly replicated DNA strands, mimicking the process of epigenetic inheritance that occurs *in vivo*. As nanopore sequencing technologies continue to advance, it may also be possible to sequence tiny amounts of unamplified DNA in the near future. DiMeLo-seq can also feasibly be extended to map RNA-DNA interactions and DNA-DNA interactions. To do so, we first need to explore the *trans* methylation activity of DiMeLo-seq with carefully controlled experiments.

Overall, it is my hope that μDamID and DiMeLo-seq will prove to be useful research tools that enable biologists to probe new fundamental questions about genome regulation and chromatin biology, and I hope they inspire further technology development towards these aims.

References

Abdulhay NJ, McNally CP, Hsieh LJ, Kasinathan S, Keith A, Estes LS, Karimzadeh M, Underwood JG, Goodarzi H, Narlikar GJ, & Ramani V. (2020). Massively multiplex single-molecule oligonucleosome footprinting. *ELife*, 9, e59404. https://doi.org/10.7554/elife.59404

Aditham AK, Markin CJ, Mokhtari DA, DelRosso N, & Fordyce PM. (2021). High-Throughput Affinity Measurements of Transcription Factor and DNA Mutations Reveal Affinity and Specificity Determinants. *Cell Systems*, 12(2), 112-127.e11. https://doi.org/10.1016/j.cels.2020.11.012

Altemose N, Maslan A, Rios-Martinez C, Lai A, White JA, & Streets A. (2020). μDamID: a microfluidic approach for joint imaging and sequencing of protein-dna interactions in single cells. *Cell Systems*, 11(4), 354-366.e9. https://doi.org/10.1016/j.cels.2020.08.015

Altemose N, Miga KH, Maggioni M, & Willard HF. (2014). Genomic characterization of large heterochromatic gaps in the human genome assembly. *PLoS Computational Biology*, 10(5), e1003628. https://doi.org/10.1371/journal.pcbi.1003628

Altemose N, Noor N, Bitoun E, Tumian A, Imbeault M, Chapman JR, Aricescu AR, & Myers SR. (2017). A map of human PRDM9 binding provides evidence for novel behaviors of PRDM9 and other zinc-finger proteins in meiosis. *ELife*, 6, e28383. https://doi.org/10.7554/elife.28383

Aughey GN, Gomez AE, Thomson J, Yin H, & Southall TD. (2018). CATaDa reveals global remodelling of chromatin accessibility during stem cell differentiation in vivo. *ELife*, 7, e32341. https://doi.org/10.7554/elife.32341

Barski A, Cuddapah S, Cui K, Roh T-Y, Schones DE, Wang Z, Wei G, Chepelev I, & Zhao K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, 129(4), 823–837. https://doi.org/10.1016/j.cell.2007.05.009

Bell ES, & Lammerding J. (2016). Causes and consequences of nuclear envelope alterations in tumour progression. *European Journal of Cell Biology*, 95(11), 449–464. https://doi.org/10.1016/j.ejcb.2016.06.007

Bentley DR. (2006). Whole-genome re-sequencing. *Current Opinion in Genetics & Development*, 16(6), 545–552. https://doi.org/10.1016/j.gde.2006.10.009

Bernhofer M, Goldberg T, Wolf S, Ahmed M, Zaugg J, Boden M, & Rost B. (2018). NLSdb—major update for database of nuclear localization signals and nuclear export signals. *Nucleic Acids Research*, 46(D1), D503–D508. https://doi.org/10.1093/nar/gkx1021

Bersani F, Lee E, Kharchenko PV, Xu AW, Liu M, Xega K, MacKenzie OC, Brannigan BW, Wittner BS, Jung H, Ramaswamy S, Park PJ, Maheswaran S, Ting DT, & Haber DA. (2015). Pericentromeric satellite repeat expansions through RNA-derived DNA intermediates in cancer. *Proceedings of the National Academy of Sciences*, 112(49), 15148–15153. https://doi.org/10.1073/pnas.1518008112

Bickmore WA, & van Steensel B. (2013). Genome architecture: domain organization of interphase chromosomes. *Cell*, 152(6), 1270–1284. https://doi.org/10.1016/j.cell.2013.02.001

Bodor DL, Mata JF, Sergeev M, David AF, Salimian KJ, Panchenko T, Cleveland DW, Black BE, Shah JV, & Jansen LE. (2014). The quantitative architecture of centromeric chromatin. *ELife*, 3, e02137. https://doi.org/10.7554/elife.02137

Bolger AM, Lohse M, & Usadel B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Borsos M, Perricone SM, Schauer T, Pontabry J, Luca KL de, Vries SS de, Ruiz-Morales ER, Torres-Padilla M-E, & Kind J. (2019). Genome–lamina interactions are established de novo in the early mouse embryo. *Nature*, 569(7758), 729–733. https://doi.org/10.1038/s41586-019-1233-0

Buchwalter A, Kaneshiro JM, & Hetzer MW. (2019). Coaching from the sidelines: the nuclear periphery in genome regulation. *Nature Reviews Genetics*, 20(1), 39–50. https://doi.org/10.1038/s41576-018-0063-5

Cao J, Cusanovich DA, Ramani V, Aghamirzaie D, Pliner HA, Hill AJ, Daza RM, McFaline-Figueroa JL, Packer JS, Christiansen L, Steemers FJ, Adey AC, Trapnell C, & Shendure J. (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science*, 361(6409), eaau0730. https://doi.org/10.1126/science.aau0730

Carter B, Ku WL, Tang Q, Kang JY, & Zhao K. (2019). *Mapping histone modifications in low cell number and single cells using antibody-guided chromatin tagmentation(ACT-seq)*. Genomics. http://biorxiv.org/lookup/doi/10.1101/571208

Celniker SE, Dillon LAL, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM, Micklem G, Piano F, Snyder M, Stein L, White KP, Waterston RH, & Consortium modENCODE. (2009). Unlocking the secrets of the genome. *Nature*, 459(7249), 927–930. https://doi.org/10.1038/459927a

Cheetham SW, & Brand AH. (2018). RNA-DamID reveals cell-type-specific binding of roX RNAs at chromatin-entry sites. *Nature Structural & Molecular Biology*, 25(1), 109–114. https://doi.org/10.1038/s41594-017-0006-4

Chen C, Xing D, Tan L, Li H, Zhou G, Huang L, & Xie XS. (2017). Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). *Science (New York, NY)*, 356(6334), 189–194. https://doi.org/10.1126/science.aak9787

Chen KH, Boettiger AN, Moffitt JR, Wang S, & Zhuang X. (2015). Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, 348(6233), aaa6090. https://doi.org/10.1126/science.aaa6090

Chen TN, Gupta A, Zalavadia MD, & Streets A. (2020). μCB-seq: microfluidic cell barcoding and sequencing for high-resolution imaging and sequencing of single cells. *Lab on a Chip*, 20(21), 3899–3913. https://doi.org/10.1039/d0lc00169d

Chu C, Qu K, Zhong FL, Artandi SE, & Chang HY. (2011). Genomic maps of long noncoding rna occupancy reveal principles of rna-chromatin interactions. *Molecular Cell*, 44(4), 667–678. https://doi.org/10.1016/j.molcel.2011.08.027

Cole RH, Tang S-Y, Siltanen CA, Shahi P, Zhang JQ, Poust S, Gartner ZJ, & Abate AR. (2017). Printed droplet microfluidics for on demand dispensing of picoliter droplets and cells. *Proceedings of the National Academy of Sciences*, 114(33), 8728–8733. https://doi.org/10.1073/pnas.1704020114

Consortium IHGS. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822), 860–921. https://doi.org/10.1038/35057062

Consortium TEP. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), 57–74. https://doi.org/10.1038/nature11247

Davies B, Hatton E, Altemose N, Hussin JG, Pratto F, Zhang G, Hinch AG, Moralli D, Biggs D, Diaz R, Preece C, Li R, Bitoun E, Brick K, Green CM, Camerini-Otero RD, Myers SR, & Donnelly P. (2016). Re-engineering the zinc fingers of PRDM9 reverses hybrid sterility in mice. *Nature*, 530(7589), 171–176. https://doi.org/10.1038/nature16931

Elsawy H, & Chahar S. (2014). Increasing DNA substrate specificity of the EcoDam DNA-(Adenine n6)-methyltransferase by site-directed mutagenesis. *Biochemistry (Moscow)*, 79(11), 1262–1266. https://doi.org/10.1134/s0006297914110145

Emanuel BS, & Shaikh TH. (2001). Segmental duplications: an "expanding" role in genomic instability and disease. *Nature Reviews Genetics*, 2(10), 791–800. https://doi.org/10.1038/35093500

Eriksson M, Brown WT, Gordon LB, Glynn MW, Singer J, Scott L, Erdos MR, Robbins CM, Moses TY, Berglund P, Dutra A, Pak E, Durkin S, Csoka AB, Boehnke M, Glover TW, & Collins FS. (2003). Recurrent de novo point mutations in lamin A cause Hutchinson–Gilford progeria syndrome. *Nature*, 423(6937), 293–298. https://doi.org/10.1038/nature01629

Essletzbichler P, Konopka T, Santoro F, Chen D, Gapp BV, Kralovics R, Brummelkamp TR, Nijman SMB, & Bürckstümmer T. (2014). Megabase-scale deletion using CRISPR/Cas9 to generate a fully haploid human cell line. *Genome Research*, 24(12), 2059–2065. https://doi.org/10.1101/gr.177220.114

Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, & Turner SW. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods*, 7(6), 461–465. https://doi.org/10.1038/nmeth.1459

Freund A, Laberge R-M, Demaria M, & Campisi J. (2012). Lamin B1 loss is a senescence-associated biomarker. *Molecular Biology of the Cell*, 23(11), 2066–2075. https://doi.org/10.1091/mbc.e11-10-0884

Gilpatrick T, Lee I, Graham JE, Raimondeau E, Bowen R, Heron A, Downs B, Sukumar S, Sedlazeck FJ, & Timp W. (2020). Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nature Biotechnology*, 38(4), 433–438. https://doi.org/10.1038/s41587-020-0407-5

Grosselin K, Durand A, Marsolier J, Poitou A, Marangoni E, Nemati F, Dahmani A, Lameiras S, Reyal F, Frenoy O, Pousse Y, Reichen M, Woolfe A, Brenan C, Griffiths AD, Vallot C, & Gérard A. (2019). High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nature Genetics*, 51(6), 1060–1066. https://doi.org/10.1038/s41588-019-0424-9

Guelen L, Pagie L, Brasset E, Meuleman W, Faza MB, Talhout W, Eussen BH, Klein A de, Wessels L, Laat W de, & Steensel B van. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, 453(7197), 948–951. https://doi.org/10.1038/nature06947

Hansen CL, Skordalakes E, Berger JM, & Quake SR. (2002). A robust and scalable microfluidic metering method that allows protein crystal growth by free interface diffusion. *Proceedings of the National Academy of Sciences*, 99(26), 16531–16536. https://doi.org/10.1073/pnas.262485199

Harada A, Maehara K, Handa T, Arimura Y, Nogami J, Hayashi-Takanaka Y, Shirahige K, Kurumizaka H, Kimura H, & Ohkawa Y. (2019). A chromatin integration labelling method enables epigenomic profiling with lower input. *Nature Cell Biology*, 21(2), 287–296. https://doi.org/10.1038/s41556-018-0248-3

Hass MR, Liow H, Chen X, Sharma A, Inoue YU, Inoue T, Reeb A, Martens A, Fulbright M, Raju S, Stevens M, Boyle S, Park J-S, Weirauch MT, Brent MR, & Kopan R. (2015). SpDamID: Marking DNA Bound by Protein Complexes Identifies Notch-Dimer Responsive Enhancers. *Molecular Cell*, 59(4), 685–697. https://doi.org/10.1016/j.molcel.2015.07.008

Hoskins VE, Smith K, & Reddy KL. (2021). The shifting shape of genomes: dynamics of heterochromatin interactions at the nuclear lamina. *Current Opinion in Genetics & Development*, 67, 163–173. https://doi.org/10.1016/j.gde.2021.02.003

Huang B, Babcock H, & Zhuang X. (2010). Breaking the diffraction barrier: super-resolution imaging of cells. *Cell*, 143(7), 1047–1058. https://doi.org/10.1016/j.cell.2010.12.002

Islam S, Zeisel A, Joost S, Manno GL, Zajac P, Kasper M, Lönnerberg P, & Linnarsson S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, 11(2), 163–166. https://doi.org/10.1038/nmeth.2772

Jagannathan M, Cummings R, & Yamashita YM. (2018). A conserved function for pericentromeric satellite DNA. *ELife*, 7, e34122. https://doi.org/10.7554/elife.34122

Jain C, Rhie A, Hansen N, Koren S, & Phillippy AM. (2020). A long read mapping method for highly repetitive reference sequences. *BioRxiv*, 2020.11.01.363887. https://doi.org/10.1101/2020.11.01.363887

Jain M, Olsen HE, Paten B, & Akeson M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, 17(1), 239. https://doi.org/10.1186/s13059-016-1103-0

Jakobsen JS, Bagger FO, Hasemann MS, Schuster MB, Frank A-K, Waage J, Vitting-Seerup K, & Porse BT. (2015). Amplification of pico-scale DNA mediated by bacterial carrier DNA for small-cell-number transcription factor ChIP-seq. *BMC Genomics*, 16(1), 46. https://doi.org/10.1186/s12864-014-1195-4

Johnson DS, Mortazavi A, Myers RM, & Wold B. (2007). Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830), 1497–1502. https://doi.org/10.1126/science.1141319

Kakar M, Davis JR, Kern SE, & Lim CS. (2007). Optimizing the protein switch: Altering nuclear import and export signals, and ligand binding domain. *Journal of Controlled Release*, 120(3), 220–232. https://doi.org/10.1016/j.jconrel.2007.04.017

Karoutas A, & Akhtar A. (2021). Functional mechanisms and abnormalities of the nuclear lamina. *Nature Cell Biology*, 23(2), 116–126. https://doi.org/10.1038/s41556-020-00630-5

Kaya-Okur HS, Wu SJ, Codomo CA, Pledger ES, Bryson TD, Henikoff JG, Ahmad K, & Henikoff S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature Communications*, 10(1), 1930.

Kelly TK, Liu Y, Lay FD, Liang G, Berman BP, & Jones PA. (2012). Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Research*, 22(12), 2497–2506. https://doi.org/10.1101/gr.143008.112

Kim S, Dorlhiac G, Chaves RC, Zalavadia M, & Streets A. (2021). Paper-thin multilayer microfluidic devices with integrated valves. *Lab on a Chip*, 21(7), 1287–1298. https://doi.org/10.1039/d0lc01217c

Kim S, Jonghe JD, Kulesa AB, Feldman D, Vatanen T, Bhattacharyya RP, Berdy B, Gomez J, Nolan J, Epstein S, & Blainey PC. (2017). High-throughput automated microfluidic sample preparation for accurate microbial genomics. *Nature Communications*, 8(1), 13919. https://doi.org/10.1038/ncomms13919

Kind J, Pagie L, de Vries SS, Nahidiazar L, Dey SS, Bienko M, Zhan Y, Lajoie B, de Graaf CA, Amendola M, Fudenberg G, Imakaev M, Mirny LA, Jalink K, Dekker J, van Oudenaarden A, & van Steensel B. (2015). Genome-wide maps of nuclear lamina interactions in single human cells. *Cell*, 163(1), 134–147. https://doi.org/10.1016/j.cell.2015.08.040

Kind J, Pagie L, Ortabozkoyun H, Boyle S, de Vries SS, Janssen H, Amendola M, Nolen LD, Bickmore WA, & van Steensel B. (2013). Single-cell dynamics of genome-nuclear lamina interactions. *Cell*, 153(1), 178–192. https://doi.org/10.1016/j.cell.2013.02.028

Kovaka S, Fan Y, Ni B, Timp W, & Schatz MC. (2021). Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. *Nature Biotechnology*, 39(4), 431–441. https://doi.org/10.1038/s41587-020-0731-9

Kriaucionis S, & Heintz N. (2009). The Nuclear DNA Base 5-Hydroxymethylcytosine Is Present in Purkinje Neurons and the Brain. *Science*, 324(5929), 929–930. https://doi.org/10.1126/science.1169786

Ku WL, Nakamura K, Gao W, Cui K, Hu G, Tang Q, Ni B, & Zhao K. (2019). Single-cell chromatin immunocleavage sequencing (scChIC-seq) to profile histone modification. *Nature Methods*, 16(4), 323–325. https://doi.org/10.1038/s41592-019-0361-7

Kumaran RI, & Spector DL. (2008). A genetic locus targeted to the nuclear periphery in living cells maintains its transcriptional competence. *Journal of Cell Biology*, 180(1), 51–65. https://doi.org/10.1083/jcb.200706060

Lai A, Altemose N, White JA, & Streets AM. (2019). On-ratio PDMS bonding for multilayer microfluidic device fabrication. *Journal of Micromechanics and Microengineering*, 29(10), 107001. https://doi.org/10.1088/1361-6439/ab341e

Landers CC, Rabeler CA, Ferrari EK, D'Alessandro LR, Kang DD, Malisa J, Bashir SM, & Carone DM. (2021). Ectopic expression of pericentric HSATII RNA results in nuclear RNA accumulation, MeCP2 recruitment, and cell division defects. *Chromosoma*, 130(1), 75–90. https://doi.org/10.1007/s00412-021-00753-0

Lane K, Valen DV, DeFelice MM, Macklin DN, Kudo T, Jaimovich A, Carr A, Meyer T, Pe'er D, Boutet SC, & Covert MW. (2017). Measuring Signaling and RNA-Seq in the Same Cell Links Gene Expression to Dynamic Patterns of NF-κB Activation. *Cell Systems*, 4(4), 458-469.e5. https://doi.org/10.1016/j.cels.2017.03.010

Lee I, Razaghi R, Gilpatrick T, Molnar M, Gershman A, Sadowski N, Sedlazeck FJ, Hansen KD, Simpson JT, & Timp W. (2020). Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *Nature Methods*, 17(12), 1191–1199. https://doi.org/10.1038/s41592-020-01000-7

Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Ferrante TC, Terry R, Turczyk BM, Yang JL, Lee HS, Aach J, Zhang K, & Church GM. (2015). Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. *Nature Protocols*, 10(3), 442–458. https://doi.org/10.1038/nprot.2014.191

Lenain C, Graaf CA de, Pagie L, Visser NL, Haas M de, Vries SS de, Peric-Hupkes D, Steensel B van, & Peeper DS. (2017). Massive reshaping of genome–nuclear lamina interactions during oncogene-induced senescence. *Genome Research*, 27(10), 1634–1644. https://doi.org/10.1101/gr.225763.117

Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, & Webb WW. (2003). Zero-Mode Waveguides for Single-Molecule Analysis at High Concentrations. *Science*, 299(5607), 682–686. https://doi.org/10.1126/science.1079700

Li D, Hsu S, Purushotham D, Sears RL, & Wang T. (2019). WashU Epigenome Browser update 2019. *Nucleic Acids Research*, 47(W1), W158–W165. https://doi.org/10.1093/nar/gkz348

Li H., Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, & Subgroup 1000 Genome Project Data Processing. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li Heng. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. https://arxiv.org/abs/1303.3997v2

Li R, Bitoun E, Altemose N, Davies RW, Davies B, & Myers SR. (2019). A high-resolution map of non-crossover events reveals impacts of genetic diversity on mammalian meiotic recombination. *Nature Communications*, 10(1), 3900. https://doi.org/10.1038/s41467-019-11675-y

Liu Yang, Yang M, Deng Y, Su G, Enninful A, Guo CC, Tebaldi T, Zhang D, Kim D, Bai Z, Norris E, Pan A, Li J, Xiao Y, Halene S, & Fan R. (2020). High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue. *Cell*, 183(6), 1665-1681.e18. https://doi.org/10.1016/j.cell.2020.10.026

Liu Yibin, Cheng J, Siejka-Zielińska P, Weldon C, Roberts H, Lopopolo M, Magri A, D'Arienzo V, Harris JM, McKeating JA, & Song C-X. (2020). Accurate targeted long-read DNA methylation and hydroxymethylation sequencing with TAPS. *Genome Biology*, 21(1), 54. https://doi.org/10.1186/s13059-020-01969-6

Logsdon GA, Vollger MR, & Eichler EE. (2020). Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, 21(10), 597–614. https://doi.org/10.1038/s41576-020-0236-x

Logsdon GA, Vollger MR, Hsieh P, Mao Y, Liskovykh MA, Koren S, Nurk S, Mercuri L, Dishuck PC, Rhie A, Lima LG de, Dvorkina T, Porubsky D, Harvey WT, Mikheenko A, Bzikadze AV, Kremitzki M, Graves-Lindsay TA, Jain C, … Eichler EE. (2021). The structure, function and evolution of a complete human chromosome 8. *Nature*, 1–7. https://doi.org/10.1038/s41586-021-03420-7

Love MI, Huber W, & Anders S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Lubeck E, Coskun AF, Zhiyentayev T, Ahmad M, & Cai L. (2014). Single-cell in situ RNA profiling by sequential hybridization. *Nature Methods*, 11(4), 360–361. https://doi.org/10.1038/nmeth.2892

Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, Trombetta JJ, Weitz DA, Sanes JR, Shalek AK, Regev A, & McCarroll SA. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5), 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002

McIntyre ABR, Alexander N, Grigorev K, Bezdan D, Sichtig H, Chiu CY, & Mason CE. (2019). Single-molecule sequencing detection of N6- methyladenine in microbial reference materials. *Nature Communications*, 1–11. https://doi.org/10.1038/s41467-019-08289-9

McNulty SM, & Sullivan BA. (2018). Alpha satellite DNA biology: finding function in the recesses of the genome. *Chromosome Research*, 26(3), 115–138. https://doi.org/10.1007/s10577-018-9582-3

Meuleman W, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Pfenning A, Wang X, Liu MC, … Wang T. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539), 317–330. https://doi.org/10.1038/nature14248

Miga KH. (2019). Centromeric Satellite DNAs: Hidden Sequence Variation in the Human Population. *Genes*, 10(5), 352. https://doi.org/10.3390/genes10050352

Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA, Schneider VA, Potapova T, Wood J, Chow W, Armstrong J, Fredrickson J, Pak E, Tigyi K, Kremitzki M, … Phillippy AM. (2020). Telomere-to-telomere assembly of a complete human X chromosome. *Nature*, 585(7823), 79–84. https://doi.org/10.1038/s41586-020-2547-7

Nguyen HQ, Baxter BC, Brower K, Diaz-Botia CA, DeRisi JL, Fordyce PM, & Thorn KS. (2017). Programmable microfluidic synthesis of over one thousand uniquely identifiable spectral codes. *Advanced Optical Materials*, 5(3), 1600548. https://doi.org/10.1002/adom.201600548

O'Brown ZK, Boulias K, Wang J, Wang SY, O'Brown NM, Hao Z, Shibuya H, Fady P-E, Shi Y, He C, Megason SG, Liu T, & Greer EL. (2019). Sources of artifact in measurements of 6mA and 4mC abundance in eukaryotic genomic DNA. *BMC Genomics*, 20(1), 445. https://doi.org/10.1186/s12864-019-5754-6

Orian A. (2003). Genomic binding by the drosophila myc, max, mad/mnt transcription factor network. *Genes & Development*, 17(9), 1101–1114. https://doi.org/10.1101/gad.1066903

Park M, Patel N, Keung AJ, & Khalil AS. (2019). Engineering epigenetic regulation using synthetic read-write modules. *Cell*, 176(1–2), 227-238.e20. https://doi.org/10.1016/j.cell.2018.11.002

Patro R, Duggal G, Love MI, Irizarry RA, & Kingsford C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4), 417–419. https://doi.org/10.1038/nmeth.4197

Payne AC, Chiang ZD, Reginato PL, Mangiameli SM, Murray EM, Yao C-C, Markoulaki S, Earl AS, Labade AS, Jaenisch R, Church GM, Boyden ES, Buenrostro JD, & Chen F. (2021). In situ genome sequencing resolves DNA sequence and structure in intact biological samples. *Science*, 371(6532), eaay3446. https://doi.org/10.1126/science.aay3446

Pickersgill H, Kalverda B, Wit E de, Talhout W, Fornerod M, & Steensel B van. (2006). Characterization of the Drosophila melanogaster genome at the nuclear lamina. *Nature Genetics*, 38(9), 1005–1014. https://doi.org/10.1038/ng1852

Quinlan AR, & Hall IM. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Quinlivan EP, & Gregory JF. (2008). *DNA digestion to deoxyribonucleoside: a simplified one-step procedure*. 373(2), 383–385. https://doi.org/10.1016/j.ab.2007.09.031

Ramalingam N, Fowler B, Szpankowski L, Leyrat AA, Hukari K, Maung MT, Yorza W, Norris M, Cesar C, Shuga J, Gonzales ML, Sanada CD, Wang X, Yeung R, Hwang W, Axsom J, Devaraju NSGK, Angeles ND, Greene C, … West JAA. (2016). Fluidic logic used in a systems approach to enable integrated single-cell functional analysis. *Frontiers in Bioengineering and Biotechnology*, 4. https://doi.org/10.3389/fbioe.2016.00070

Ramani V, Deng X, Qiu R, Gunderson KL, Steemers FJ, Disteche CM, Noble WS, Duan Z, & Shendure J. (2017). Massively multiplex single-cell Hi-C. *Nature Methods*, 14(3), 263–266. https://doi.org/10.1038/nmeth.4155

Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F, & Manke T. (2016). deepTools2: a next generation web server for deep-

sequencing data analysis. *Nucleic Acids Research*, 44(W1), W160–W165. https://doi.org/10.1093/nar/gkw257

Rand AC, Jain M, Eizenga JM, Musselman-Brown A, Olsen HE, Akeson M, & Paten B. (2017). Mapping DNA methylation with high-throughput nanopore sequencing. *Nature Methods*, 14(4), 411–413. https://doi.org/10.1038/nmeth.4189

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, & Smyth GK. (2015). Limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research*, 43(7), e47–e47. https://doi.org/10.1093/nar/gkv007

Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, & Jones S. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature Methods*, 4(8), 651–657. https://doi.org/10.1038/nmeth1068

Rodriguez-Mateos P, Azevedo NF, Almeida C, & Pamme N. (2020). FISH and chips: a review of microfluidic platforms for FISH analysis. *Medical Microbiology and Immunology*, 209(3), 373–391. https://doi.org/10.1007/s00430-019-00654-1

Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR, Welch J, Chen LM, Chen F, & Macosko EZ. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434), 1463–1467. https://doi.org/10.1126/science.aaw1219

Rooijers K, Markodimitraki CM, Rang FJ, Vries SS de, Chialastri A, Luca KL de, Mooijman D, Dey SS, & Kind J. (2019). Simultaneous quantification of protein–DNA contacts and transcriptomes in single cells. *Nature Biotechnology*.

Rotem A, Ram O, Shoresh N, Sperling RA, Goren A, Weitz DA, & Bernstein BE. (2015). Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nature Biotechnology*, 33(11), 1165–1172. https://doi.org/10.1038/nbt.3383

Saint M, Bertaux F, Tang W, Sun X-M, Game L, Köferle A, Bähler J, Shahrezaei V, & Marguerat S. (2019). Single-cell imaging and RNA sequencing reveal patterns of gene expression heterogeneity during fission yeast growth and adaptation. *Nature Microbiology*, 4(3), 480–491. https://doi.org/10.1038/s41564-018-0330-4

Salim D, & Gerton JL. (2019). Ribosomal DNA instability and genome adaptability. *Chromosome Research*, 27(1–2), 73–87. https://doi.org/10.1007/s10577-018-9599-7

Satpathy AT, Granja JM, Yost KE, Qi Y, Meschi F, McDermott GP, Olsen BN, Mumbach MR, Pierce SE, Corces MR, Shah P, Bell JC, Jhutty D, Nemec CM, Wang J, Wang L, Yin Y, Giresi PG, Chang ALS, … Chang HY. (2019). Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nature Biotechnology*, 37(8), 925–936. https://doi.org/10.1038/s41587-019-0206-z

Schaik T van, Vos M, Peric-Hupkes D, Celie PH, & Steensel B van. (2020). Cell cycle dynamics of lamina-associated DNA. *EMBO Reports*, 21(11), e50636. https://doi.org/10.15252/embr.202050636

Schmid M, Durussel T, & Laemmli UK. (2004). Chic and chec. *Molecular Cell*, 16(1), 147–157. https://doi.org/10.1016/j.molcel.2004.09.007

Schueler MG, Higgins AW, Rudd MK, Gustashaw K, & Willard HF. (2001). Genomic and Genetic Definition of a Functional Human Centromere. *Science*, 294(5540), 109–115. https://doi.org/10.1126/science.1065042

Shalek AK, Satija R, Shuga J, Trombetta JJ, Gennert D, Lu D, Chen P, Gertner RS, Gaublomme JT, Yosef N, Schwartz S, Fowler B, Weaver S, Wang J, Wang X, Ding R, Raychowdhury R, Friedman N, Hacohen N, … Regev A. (2014). Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, 510(7505), 363–369. https://doi.org/10.1038/nature13437

Shay JW, & Wright WE. (2019). Telomeres and telomerase: three decades of progress. *Nature Reviews Genetics*, 20(5), 299–309. https://doi.org/10.1038/s41576-019-0099-1

Shen J, Jiang D, Fu Y, Wu X, Guo H, Feng B, Pang Y, Streets AM, Tang F, & Huang Y. (2015). H3K4me3 epigenomic landscape derived from ChIP-Seq of 1 000 mouse early embryonic cells. *Cell Research*, 25(1), 143–147. https://doi.org/10.1038/cr.2014.119

Shipony Z, Marinov GK, Swaffer MP, Sinnott-Armstrong NA, Skotheim JM, Kundaje A, & Greenleaf WJ. (2020). Long-range single-molecule mapping of chromatin

accessibility in eukaryotes. *Nature Methods*, 17(3), 319–327. https://doi.org/10.1038/s41592-019-0730-2

Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, & Timp W. (2017). Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods*, 14(4), 407–410. https://doi.org/10.1038/nmeth.4184

Singh J, & Klar AJ. (1992). Active genes in budding yeast display enhanced in vivo accessibility to foreign DNA methylases: a novel in vivo probe for chromatin structure of yeast. *Genes & Development*, 6(2), 186–196. https://doi.org/10.1101/gad.6.2.186

Skene PJ, & Henikoff S. (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *ELife*, 6, e21856. https://doi.org/10.7554/elife.21856

Sobecki M, Souaid C, Boulay J, Guerineau V, Noordermeer D, & Crabbe L. (2018). Madid, a versatile approach to map protein-dna interactions, highlights telomere-nuclear envelope contact sites in human cells. *Cell Reports*, 25(10), 2891-2903.e5. https://doi.org/10.1016/j.celrep.2018.11.027

Solomon MJ, Larsen PL, & Varshavsky A. (1988). Mapping proteinDNA interactions in vivo with formaldehyde: Evidence that histone H4 is retained on a highly transcribed gene. *Cell*, 53(6), 937–947. https://doi.org/10.1016/s0092-8674(88)90469-2

Solovei I, Kreysing M, Lanctôt C, Kösem S, Peichl L, Cremer T, Guck J, & Joffe B. (2009). Nuclear Architecture of Rod Photoreceptor Cells Adapts to Vision in Mammalian Evolution. *Cell*, 137(2), 356–368. https://doi.org/10.1016/j.cell.2009.01.052

Southall TD, Gold KS, Egger B, Davidson CM, Caygill EE, Marshall OJ, & Brand AH. (2013). Cell-type-specific profiling of gene expression and chromatin binding without cell isolation: assaying rna pol ii occupancy in neural stem cells. *Developmental Cell*, 26(1), 101–112. https://doi.org/10.1016/j.devcel.2013.05.020

Squires TM, & Quake SR. (2005). Microfluidics: Fluid physics at the nanoliter scale. *Reviews of Modern Physics*, 77(3), 977–1026. https://doi.org/10.1103/revmodphys.77.977

Steensel B van, & Belmont AS. (2017). Lamina-associated domains: links with chromosome architecture, heterochromatin, and gene repression. *Cell*, 169(5), 780–791. https://doi.org/10.1016/j.cell.2017.04.022

Steensel B van, & Henikoff S. (2000). Identification of in vivo DNA targets of chromatin proteins using tethered Dam methyltransferase. *Nature Biotechnology*, 18(4), 424–428. https://doi.org/10.1038/74487

Stergachis AB, Debo BM, Haugen E, Churchman LS, & Stamatoyannopoulos JA. (2020). Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science*, 368(6498), 1449–1454. https://doi.org/10.1126/science.aaz1646

Streets AM, & Huang Y. (2013). Chip in a lab: Microfluidics for next generation life science research. *Biomicrofluidics*, 7(1), 011302–011324. https://doi.org/10.1063/1.4789751

Streets AM, Zhang X, Cao C, Pang Y, Wu X, Xiong L, Yang L, Fu Y, Zhao L, Tang F, & Huang Y. (2014). Microfluidic single-cell whole-transcriptome sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 111(19), 7048–7053. https://doi.org/10.1073/pnas.1402030111

Strong SJ, Ohta Y, Litman GW, & Amemiya CT. (1997). Marked Improvement of PAC and BAC Cloning Is Achieved Using Electroelution of Pulsed-Field Gel-Separated Partial Digests of Genomic DNA. *Nucleic Acids Research*, 25(19), 3959–3961. https://doi.org/10.1093/nar/25.19.3959

Sullivan LL, Boivin CD, Mravinac B, Song IY, & Sullivan BA. (2011). Genomic size of CENP-A domain is proportional to total alpha satellite array size at human centromeres and expands in cancer cells. *Chromosome Research*, 19(4), 457. https://doi.org/10.1007/s10577-011-9208-5

Swanson EC, Manning B, Zhang H, & Lawrence JB. (2013). Higher-order unfolding of satellite heterochromatin is a consistent and early event in cell senescence. *Journal of Cell Biology*, 203(6), 929–942. https://doi.org/10.1083/jcb.201306073

Teves SS, An L, Hansen AS, Xie L, Darzacq X, & Tjian R. (2016). A dynamic mode of mitotic bookmarking by transcription factors. *ELife*, 5, e22280. https://doi.org/10.7554/elife.22280

Thorsen T, Maerkl SJ, & Quake SR. (2002). Microfluidic Large-Scale Integration. *Science*, 298(5593), 580–584. https://doi.org/10.1126/science.1076996

Unger MA, Chou H-P, Thorsen T, Scherer A, & Quake SR. (2000). Monolithic Microfabricated Valves and Pumps by Multilayer Soft Lithography. *Science*, 288(5463), 113–116. https://doi.org/10.1126/science.288.5463.113

Vitak SA, Torkenczy KA, Rosenkrantz JL, Fields AJ, Christiansen L, Wong MH, Carbone L, Steemers FJ, & Adey A. (2017). Sequencing thousands of single-cell genomes with combinatorial indexing. *Nature Methods*, 14(3), 302–308. https://doi.org/10.1038/nmeth.4154

Vogel MJ, Peric-Hupkes D, & Steensel B van. (2007). Detection of in vivo protein–DNA interactions using DamID in mammalian cells. *Nature Protocols*, 2(6), 1467–1478. https://doi.org/10.1038/nprot.2007.148

Wang H, Maurano MT, Qu H, Varley KE, Gertz J, Pauli F, Lee K, Canfield T, Weaver M, Sandstrom R, Thurman RE, Kaul R, Myers RM, & Stamatoyannopoulos JA. (2012). Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Research*, 22(9), 1680–1688. https://doi.org/10.1101/gr.136101.111

Wang Y, Wang A, Liu Z, Thurman AL, Powers LS, Zou M, Zhao Y, Hefel A, Li Y, Zabner J, & Au KF. (2019). Single-molecule long-read sequencing reveals the chromatin basis of gene expression. *Genome Research*, 29(8), 1329–1342. https://doi.org/10.1101/gr.251116.119

Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, Töpfer A, Alonge M, Mahmoud M, Qian Y, Chin C-S, Phillippy AM, Schatz MC, Myers G, DePristo MA, … Hunkapiller MW. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, 37(10), 1155–1162. https://doi.org/10.1038/s41587-019-0217-9

White AK, VanInsberghe M, Petriv OI, Hamidi M, Sikorski D, Marra MA, Piret J, Aparicio S, & Hansen CL. (2011). High-throughput microfluidic single-cell RT-qPCR. *Proceedings of the National Academy of Sciences*, 108(34), 13999–14004. https://doi.org/10.1073/pnas.1019446108

White JA, & Streets AM. (2018). Controller for microfluidic large-scale integration. *HardwareX*, 3, 135–145. https://doi.org/10.1016/j.ohx.2017.10.002

Wines DR, Talbert PB, Clark DV, & Henikoff S. (1996). Introduction of a DNA methyltransferase into Drosophila to probe chromatin structure in vivo. *Chromosoma*, 104(5), 332–340. https://doi.org/10.1007/bf00337221

Wolberg WH, Street WN, & Mangasarian OL. (1999). Importance of nuclear morphology in breast cancer prognosis. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, 5(11), 3542–3548.

Wu AR, Kawahara TLA, Rapicavoli NA, Riggelen J van, Shroff EH, Xu L, Felsher DW, Chang HY, & Quake SR. (2012). High throughput automated chromatin immunoprecipitation as a platform for drug screening and antibody validation. *Lab on a Chip*, 12(12), 2190. https://doi.org/10.1039/c2lc21290k

Xia Y, & Whitesides GM. (1998). Soft Lithography. *Angewandte Chemie International Edition*, 37(5), 550–575. https://doi.org/10.1002/(sici)1521-3773(19980316)37:5<550::aid-anie550>3.0.co;2-g

Xiao R, Roman-Sanchez R, & Moore DD. (2010). DamIP: a novel method to identify DNA binding sites in vivo. *Nuclear Receptor Signaling*, 8, e003. https://doi.org/10.1621/nrs.08003

Yaron JR, Ziegler CP, Tran TH, Glenn HL, & Meldrum DR. (2014). A convenient, optimized pipeline for isolation, fluorescence microscopy and molecular analysis of live single cells. *Biological Procedures Online*, 16(1), 9. https://doi.org/10.1186/1480-9222-16-9

Yuan J, Sheng J, & Sims PA. (2018). SCOPE-Seq: a scalable technology for linking live cell imaging and single-cell RNA sequencing. *Genome Biology*, 19(1), 227. https://doi.org/10.1186/s13059-018-1607-x

Yuan J, & Sims PA. (2016). *An Automated Microwell Platform for Large-Scale Single Cell RNA-Seq* (pp. 1–23). https://mail.google.com/mail/u/0/

Zhang B, Zheng H, Huang B, Li W, Xiang Y, Peng X, Ming J, Wu X, Zhang Y, Xu Q, Liu W, Kou X, Zhao Y, He W, Li C, Chen B, Li Y, Wang Q, Ma J, … Xie W. (2016). Allelic reprogramming of the histone modification H3K4me3 in early mammalian development. *Nature*, 537(7621), 553–557. https://doi.org/10.1038/nature19361

Zhang JQ, Siltanen CA, Liu L, Chang K-C, Gartner ZJ, & Abate AR. (2020). Linked optical and gene expression profiling of single cells at high-throughput. *Genome Biology*, 21(1), 49. https://doi.org/10.1186/s13059-020-01958-9

Zhao Z, Tavoosidana G, Sjölinder M, Göndör A, Mariano P, Wang S, Kanduri C, Lezcano M, Sandhu KS, Singh U, Pant V, Tiwari V, Kurukuti S, & Ohlsson R. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature Genetics*, 38(11), 1341–1347. https://doi.org/10.1038/ng1891

Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, Gregory MT, Shuga J, Montesclaros L, Underwood JG, Masquelier DA, Nishimura SY, Schnall-Levin M, Wyatt PW, Hindson CM, … Bielas JH. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1), 14049. https://doi.org/10.1038/ncomms14049

Zhu Q, Hoong N, Aslanian A, Hara T, Benner C, Heinz S, Miga KH, Ke E, Verma S, Soroczynski J, Yates JR, Hunter T, & Verma IM. (2018). Heterochromatin-Encoded Satellite RNAs Induce Breast Cancer. *Molecular Cell*, 70(5), 842-853.e7. https://doi.org/10.1016/j.molcel.2018.04.023

Zilionis R, Nainys J, Veres A, Savova V, Zemmour D, Klein AM, & Mazutis L. (2017). Single-cell barcoding and sequencing using droplet microfluidics. *Nature Protocols*, 12(1), 44–73. https://doi.org/10.1038/nprot.2016.154

Appendix 1: Key Resources Table for uDamID

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and Virus Strains		
dam-/dcm- Competent *E. Coli*	New England Biolabs	Cat#C2925I
Chemicals, Peptides, and Recombinant Proteins		
DMEM with GlutaMAX	Gibco	Cat#10566-016
Pen Strep	Gibco	Cat#15140-122
FBS (Seradigm Select Grade)	Avantor	Cat#89510-186
FuGene HD	Promega	Cat#E2311
Shield-1 Ligand	Takara Bio USA, Inc.	Cat#632189
TrypLE Select	Gibco	Cat#12563-011
SU-8 2025 negative photoresist	Microchem	SU-8 2025
AZ 40XT-11D positive photoresist	Integrated Micro Materials	AZ40XT-11D
PDMS	Momentive Performance Materials	RTV615A
Trichloromethylsilane	Millipore Sigma	Cat#M85301-5G
Pluronic F-127	Millipore Sigma	Cat#P2443
Proteinase K	New England Biolabs	Cat#P8107S
common salts & detergents: KCl, MgCl2, NaCl, TRIS-HCl, TRIS acetate, magnesium acetate, potassium acetate, sodium acetate, EDTA, Tween-20, IGEPAL	Various	Various
T4 DNA ligase and T4 ligase buffer	New England Biolabs	Cat#M0202S
DpnI and CutSmart buffer	New England Biolabs	Cat#R0176S
Takara Advantage 2 PCR Kit	Takara Bio USA, Inc.	Cat#639207
DpnII and DpnII Buffer	New England Biolabs	Cat#R0543S
DNeasy Blood & Tissue Kit	Qiagen	Cat#69504
QIAquick PCR Purification Kit	Qiagen	Cat#28104
RNeasy Mini Kit with QIAshredder	Qiagen	Cat#74104

AMPure XP magnetic SPRI beads	Beckman Coulter	Cat#A63880
Monarch PCR & DNA Cleanup Kit	New England Biolabs	Cat#T1030S
solvents: SU-8 developer, AZ 300MIF developer, molecular biology grade 100% ethanol, IPA, acetone	Various	Various
RNase A	Qiagen	Cat#19101
Critical Commercial Assays		
Qubit dsDNA HS Assay Kit	ThermoFisher Scientific	Cat#Q32851
TapeStation D5000 HS Ladder, Reagents, and ScreenTape	Agilent	Cat#5067-5594, Cat#5067-5593, Cat#5067-5592
NEBNext Ultra II DNA Library Prep Kit for Illumina	New England Biolabs	Cat#E7645
NEBnext Ultra II FS DNA Library Prep Kit for Illumina	New England Biolabs	Cat#E7805
NEBNext Ultra II RNA Library Prep Kit for Illumina	New England Biolabs	Cat#E7770S
NEBNext Poly(A) mRNA Magnetic Isolation Module	New England Biolabs	Cat#E7490S
Deposited Data		
KBM-7 bulk RNA-seq	Essletzbichler et al., 2014	SRA: SRP044391
KBM-7 single-cell DamID	Kind et al., 2015	GEO: GSE69423
HEK293T cell bulk & single-cell DamID sequencing reads and bulk RNA-seq reads	this study	GEO: GSE156150
HEK293T cell raw images	this study	available on FigShare: https://doi.org/10.6084/m9.figshare.12798158
Experimental Models: Cell Lines		
HEK293T cells	ATCC	Cat#CRL-3216; RRID:CVCL_0063
Oligonucleotides		

AdRt: CTAATACGACTCACTA TAGGGCAGCGTGGT	Integrated DNA Technologies	CustomOrder
CGCGGCCGAGGA		
AdRb: TCCTCGGCCG	Integrated DNA Technologies	CustomOrder
AdR_PCR: NNNNGTGGTCGCGGCCGA GGATC	Integrated DNA Technologies	CustomOrder
NEBNext Multiplex Oligos for Illumina (96 Unique Dual Index Primer Pairs)	New England Biolabs	Cat#E6440S
Recombinant DNA		
plasmid: DD-DamV133A-LMNB1-IRES2-mCherry	this study (modified from a gift from Bas van Steensel)	Deposited to Addgene (#159599), sequence also on GitHub https://github.com/altemose/microDamID
plasmid: DD-DamV133A-IRES2-mCherry	this study (modified from a gift from Bas van Steensel)	Deposited to Addgene (#159600), sequence also on GitHub https://github.com/altemose/microDamID
plasmid: DD-DamWT-LMNB1-IRES2-mCherry	this study (modified from a gift from Bas van Steensel)	Deposited to Addgene (#159601), sequence also on GitHub https://github.com/altemose/microDamID
plasmid: DD-DamWT-IRES2-mCherry	this study (modified from a gift from Bas van Steensel)	Deposited to Addgene (#159602), sequence also on GitHub https://github.com/altemose/microDamID
plasmid: DD-DamV133A-tdTomato-LMNB1	this study (modified from a gift from Bas van Steensel)	Deposited to Addgene (#159604), sequence also on

		GitHub https://github.com/altemose/microDamID
plasmid: m6A-Tracer	gift from Bas van Steensel	Kind et al., 2013
plasmid: m6A-Tracer-NES	this study	Deposited to Addgene (#159607), sequence also on GitHub https://github.com/altemose/microDamID
Software and Algorithms		
trimmomatic v0.39	Bolger et al., 2014	N/A
Salmon	Patro et al., 2017	N/A
limma (R package)	Ritchie et al., 2015	N/A
BWA-MEM v0.7.15-r1140	Li, 2013	N/A
samtools v1.9	Li et al., 2009	N/A
bedtools v2.28	Quinlan and Hall, 2010	N/A
DESeq2	Love et al., 2014	N/A
R (v4.0.0)	The R Project for Statistical Computing	N/A
Additional R packages: ggplot2 (v3.3.0), gplots (v3.0.3), colorRamps (v2.3), reshape2 (v1.4.4), ggextra (v0.9), poisbinom (v1.0.1), SDMTools (v1.1-221.1), spatstat (v1.59-0), magick (v2.0)	CRAN & Bioconductor repositories	N/A
Wash U Epigenome Browser	Li et al., 2019	N/A
In-house code for file parsing (bash, perl, & python) and data analysis (python & R)	this study	All code is available on this study's GitHub repository: https://github.com/altemose/microDamID
Other		
10 cm diameter, 500 μm thick test-grade silicon wafers	University Wafer	Cat#452

Photomasks (25400 DPI)	CAD/Art Services	Design files are available on this study's GitHub repository: https://github.com/altemose/microDamID
Gold SEAL 48x65 mm No.1 coverglass	ThermoFisher Scientific	Cat#3335
PEEK tubing (0.25 mm ID, 0.8 mm OD)	IDEX Corporation	Cat#1581
Gel loading tips	Cole-Parmer	Cat#UX-25713-12

Appendix 2: Full DiMeLo-seq Protocol

Prepared by Annie Maslan. An updated version of this protocol can be found at streetslab.com

Materials

- HEPES-KOH, 1 M, pH 7.5 (BBH-75-K)
- NaCl, 5 M (59222C-500ML)
- Spermidine, 6.4 M (S0266-5G)
- Roche cOmplete tablet -EDTA (11873580001)
- BSA (A6003-25G)
- Digitonin (300410-250MG)
- Tween-20 (P7949-100ML)
- KCl (PX1405-1)
- EDTA, 0.5 M, pH 8.0 (Invitrogen 15575-038)
- EGTA, 0.5 M, pH 8.0 (Fisher 50-255-956)
- SAM, 32 mM (NEB B9003S)
- PFA, 16% (if fixing) (EMS 15710)
- Glycine (if fixing) (BP381-1)
- Eppendorf DNA LoBind tubes, 1.5 ml (022431021)
- Wide bore 200 µl and 1000 µl tips (e.g. USA Scientific 1011-8810, VWR 89049-168)
- pA-Hia5
- Primary antibody for protein target of interest, from species compatible with pA (e.g. ab16048)
- Secondary antibody for IF QC (e.g. ab3554)
- Trypan Blue (T10282)
- Monarch Genomic DNA Purification Kit (T3010S)
- Qubit dsDNA BR Assay Kit (Q32850)
- Agencourt AMPure XP beads (A63881)
- Blunt/TA Ligase Master Mix (NEB M0367S)
- NEBNext quick ligation module (NEB E6056S)
- NEBNext End Repair dA-tailing Module (NEB E7546S)
- NEBNext FFPE DNA repair kit (NEB M6630S)
- Ligation Sequencing Kit (ON SQK-LSK109)
- Native Barcoding Expansion 1-12 (ON EXP-NBD104)
- Native Barcoding Expansion 13-24 (ON EXP-NBD114)
- Flow Cell Wash Kit (ON EXP-WSH004)
- Flow cells (ON FLO-MIN106D)

Timeline

Day 1: Perform *in situ* targeted methylation and DNA extraction
Day 2: Perform library preparation and start sequencing
Day 2-4: sequence

Perform *in situ* targeted methylation and DNA extraction

Reagent preparation

Prepare all reagents fresh and keep on ice. Syringe filter all solutions through a 0.2 µM filter.

1. 5% digitonin solution

Solubilize digitonin in preheated 95 °C Milli-Q water to create a 5% digitonin solution (e.g. 10mg/200µl).

2. Wash buffer

component	amount	concentration
HEPES-KOH, 1 M, pH 7.5	1 ml	20 mM
NaCl, 5 M	1.5 ml	150 mM
Spermidine, 6.4 M	3.91 µl	0.5 mM
Roche Complete tablet -EDTA	1 tablet	-
BSA	50 mg	0.1%
H2O	up to 50 ml	-

3. Dig-Wash buffer

Add 0.02% digitonin to wash buffer. For example, add 20 µl of 5% digitonin solution to 5 ml wash buffer.

4. Tween-Wash buffer

Add 0.1% Tween-20 to wash buffer. For example, add 50 µl Tween-20 to 50 ml wash buffer.

5. Activation buffer

Create the activation buffer but wait to add SAM until the activation step.

component	amount	concentration
Tris, pH 8.0 1 M	750 µl	15 mM
NaCl 5 M	150 µl	15 mM
KCl 1 M	3 mL	60 mM
EDTA, pH 8.0 0.5 M	100 µl	1 mM
EGTA, pH 8.0 0.5 M	50 µl	0.5 mM
Spermidine, 6.4 M	3.91µl	0.5 mM
BSA	50 mg	0.1%
H2O	up to 50 mL	-
SAM, 32 mM	(add at activation step)	800 µM

Protocol

General notes

- All spins are at 4 °C for 3 minutes at 500 x g.
- To prevent nuclei from lining the side of the tube, break all spins into two parts: 2 minutes with the tube hinge facing inward, followed by 1 minute with the tube hinge facing outward.
- Use wide bore tips when working with nuclei.
- Do not use Triton (0.1%) or NP-40. Both appear to dramatically reduce methylation activity.
- The best digitonin concentration may vary by cell type. For HEK293T, GM12878, HG002, and Hap1 cells, 0.02% works well. You can test different concentrations of digitonin and verify permeabilization and nuclear integrity by Trypan blue staining. For example, you may try 0.02% to 0.1% digitonin.
- We use Tween to reduce hydrophilic non-specific interactions and BSA to reduce hydrophobic non-specific interactions. We've also found BSA at the activation step significantly increases methylation activity as well.
- The best primary antibody concentration may vary by protein target of interest. A 1:50 dilution works well for targeting LMNB1 and is likely a good starting point for most antibodies.
- A secondary antibody binding step following primary antibody binding and before pA-Hia5 binding reduced total methylation and specificity. Including a secondary antibody binding step is not recommended.

(Optional fixation)

1. Resuspend cells in PBS.
2. Add PFA to 0.1% (e.g. 6.2 µl of 16% PFA to 1 ml cells) for 2 minutes while gently vortexing.
3. Add 1.25 M glycine (sterile; 0.938 g in 10 ml) to twice the molar concentration of PFA to stop the crosslinking (e.g. 60 µl of 1.25 M glycine to 1 ml)
4. Centrifuge 3 minutes at 500 x g at 4 °C and remove the supernatant
5. Resuspend the fixed cells in Dig-Wash buffer (A. Nuclear isolation, step 3).

A. Nuclear isolation

1. Prepare cells (1M-5M per condition)
2. Wash cells in PBS. Spin and remove supernatant.
3. Resuspend cells in 1 ml Dig-Wash buffer. Incubate for 5 minutes on ice.
4. Split nuclei suspension into separate tubes for each condition.
5. Spin and remove supernatant.

QC: Check permeabilization was successful by taking 1 µl of the nuclei following the 5-minute incubation on ice, diluting to 10 µl with PBS, and staining with Trypan Blue. Alternatively, fix with 1% PFA for 2 minutes at room temperature and wash with Tween-Wash. Spin and remove supernatant. Resuspend in fluoromount with 1:500 Hoechst.

B. Primary antibody binding

1. Gently resolve each pellet in 200 µl Tween-Wash containing primary antibody at 1:50.
2. Place on rotator at 4 °C for ~2 hr.
3. Spin and remove supernatant.
4. Wash twice with 0.95 ml Tween-Wash. For each wash, gently and completely resolve the pellet. This may take pipetting up and down ~10 times. Following resuspension, place on rotator at 4 °C for 5 minutes before spinning down.

C. pA-Hia5 binding

1. Gently resolve pellet in 200 µl Tween-Wash containing 200 nM pA-Hia5. See protein quantification protocol below.
2. Place on rotator at room temperature for ~1 hr.
3. Spin and remove supernatant.
4. Wash with 0.95 ml Tween-Wash. For each wash, gently and completely resolve the pellet. Following resuspension, place on rotator at 4 °C for 5 minutes before spinning down.

Protein quantification protocol:

1. Thaw protein from -80 °C at room temperature and then move to ice immediately.
2. Spin at 4 °C for 10 minutes at 10,000 x g or higher.

3. Transfer supernatant to a new tube.
4. Use Qubit with 2 µl sample volume to quantify protein.

QC:

1. Add 1.6 µl of 16% PFA to 25 µl of nuclei in Tween-Wash (taken from the 0.95 ml final wash) for 1% total PFA concentration.
2. Incubate at room temperature for 5 minutes.
3. Add 975 µl of Tween-Wash to stop the fixation by dilution.
4. Add 1 µl secondary antibody.
5. Put on rotator for 30 minutes at room temperature, protected from light.
6. Wash 2 times (or just once). Pellet likely will not be visible.
7. Resuspend in mounting media after last wash. Use as little as possible, ideally 5 µl.
8. Put 5 µl on a slide, make sure there are no bubbles, and put on a coverslip.
9. Seal with nail polish along the edges.
10. Image or put at -20 °C once the nail polish has dried.

D. Activation

1. Gently resolve pellet in 100 µl of Activation Buffer per sample. Be sure to add SAM to 800 µM to the activation buffer at this step!
2. Incubate at 37 °C for 30 minutes.
3. Spin and remove supernatant.
4. Resuspend in 100 µl cold PBS.

QC: Check nuclei by Trypan blue staining to determine recovery and check integrity of nuclei if desired.

E. DNA extraction

Use the Monarch Genomic DNA Purification Kit. Follow protocol for genomic DNA isolation using cell lysis buffer. Include RNase A. NB. If fixation was performed, be sure to do the 56 °C incubation for lysis for 1 hour (not just 5 minutes) to reverse crosslinks. Perform two elutions: 100 µl and then 35 µl. Quantify DNA yield by Qubit dsDNA BR Assay Kit. Concentrate by speedvac if necessary for 3 µg DNA in 48 µl for input to library prep.

Perform library preparation and start the sequencing run

Follow Nanopore protocol for Native Barcoding Ligation Kit with the following modifications:

1. Load ~3 µg DNA into end repair.
2. Incubate for 10 minutes at 20 °C for end repair instead of 5 minutes.
3. Load ~ 1 µg of end repaired DNA into barcode ligation.
4. Double the ligation incubation time to at least 20 minutes.
5. Elute in 18 µl instead of 26 µl following barcode ligation reaction cleanup to allow for more material to be loaded into the final ligation.

6. Load ~3 µg of pooled barcoded material into the final ligation. If needed, concentrate using speedvac to be able to load 3 µg into the final ligation.
7. Double the ligation incubation time to at least 20 minutes.
8. Make sure to use LFB (NOT ethanol) for the final cleanup.
9. Perform final elution in 13 µl EB. Take out 1 µl to dilute 1:5 for quantification by Qubit (and size distribution analysis by TapeStation / Bioanalyzer if desired).
10. Load ~1 µg of DNA onto the sequencer.
11. Bubbles will absolutely destroy pores and ruin runs; mix and spin down all flush/wash solutions really well to eliminate bubbles.
12. The Flow Cell Wash Kit can increase the throughput per flowcell with <1% carryover of pre-wash barcodes.
13. Spiking in more library + SQB + LB during a run, without a wash step, can also increase pore occupancy if it is low.