Title
Low-input library preparation methods for single molecule sequencing

Author
Nanda, Arjun Scott

Publication Date
2023

Low-input library preparation methods for single molecule sequencing

by
Arjun Scott Nanda

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
LUKE GILBERT
71F73C69F83C48B...

LUKE GILBERT
Chair

DocuSigned by:
Hani Goodarzi
DocuSigned by:
Bjoern Schwer
9F6FFBE782AC485...

Hani Goodarzi

Bjoern Schwer

Committee Members

For my parents, Rema and Rajiv Nanda

Acknowledgements

We stand on the shoulders of those who came before us. I was extraordinarily lucky to have been raised by two driven, hardworking, and ambitious women – my mother, and my grandmother – who pushed me to pursue my passions and never compromise in the face of adversity. My mother, Dr. Rema Nanda, who navigated foreign shores during her graduate training, has guided me through this complex journey and constantly inspires me to go farther in the pursuit of knowledge. I would not be where I am today without her. My academic journey has been one with ups and downs, but I have been blessed with the constant love and support of my partner and her wonderful family, who have adopted me as one of their own. Saloni, your love and your enduring spirit has kept me moving forward, even at the toughest of times. I am fortunate to also have true friends who have always been a phone call away – Will Palmisano, you have kept my curiosity and humor alive over these tumultuous pandemic years.

UCSF has been a phenomenal environment to complete my doctoral training, from the excellent faculty mentors, including program directors, Dr. Ryan Hernandez and Dr. Tony Capra, to my supportive committee, Dr. Luke Gilbert and Dr. Bjoern Schwer. Sincere thanks as well to my advisors Dr. Hani Goodarzi and Dr. Vijay Ramani for supporting my work and providing valuable insights into the science at hand. I strongly believe science is a collaborative practice, and I owe a great deal to many individuals including the exceptionally talented Coco Wu, as well as Nour Abudulhay, and Colin McNally, who were instrumental in developing foundational methods critical for my own work. Special thanks as well to Camille Moore, who was always game to bounce around some big ideas, and the other members of the lab who do great science, including Megan Ostrowski, Sean Wang, Iryna Irkliyenko, Marty Yang and Hannah Richter. I was fortunate to have been adopted by Dr. Serena Nik-Zainal, across the pond at the University of Cambridge, during my Masters, and it has been a wonderful experience continuing to work on exciting projects with members of her team including Dr. Xueqing Zou and Dr. Gene C.C. Koh. Finally, to my once IPQB now BMI cohort including Aiden Winters, Erin Gilbertson, Laura Shub, Hassan Alkhairo, Tianna Grant, and Zach Cutts – thank you for coming along with me on this journey, I can't wait to see where your work takes you next!

Contributions

This dissertation was supervised by Dr. Vijay Ramani and Dr. Hani Goodarzi. Chapter 2 contains material from a manuscript currently under review, and available in open access format:

> Nanda, A. S. *et al.* Sensitive multimodal profiling of native DNA by transposase-mediated single-molecule sequencing. Preprint at https://doi.org/10.1101/2022.08.07.502893 (2022).

Chapter 3 and Chapter 4 contain unpublished material.

Low-input library preparation methods for single molecule sequencing

Arjun Scott Nanda

Abstract

Over the past two decades, high-throughput DNA sequencing has revolutionized our understanding of epigenetics. The development of quantitative assays that use sequencing to measure chromatin accessibility and methylation state have provided key insights into oncogenesis and cancer metastasis. Recently developed 3rd generation technologies offer significant improvements over current sequencing platforms, including kilobase-scale read lengths and native detection of epigenetic marks on single molecules. Profiling techniques built on these platforms can therefore capture the epigenetic states of single chromatin fibers with unprecedented resolution. However, the inherently higher input requirements for these assays and sequencing platforms have limited their general applicability in medical settings, where sample material is constrained.

In this dissertation, we address this problem by developing a new generalizable strategy for preparing native 3rd generation sequencing libraries from low-input samples using a hyperactive transposase. In Chapter 1, we motivate and contextualize this work by discussing how the epigenome is dysregulated in cancer progression and the current tools we use to study it. Then, in Chapter 2 we discuss how our transposase-mediated method enables the study of chromatin fibers in a range of clinically relevant samples including cancer cell lines and patient derived xenograft models. In Chapter 3, we present further methodological improvements that enable direct library preparation from cells and nuclei, collectively lowering input requirements ~20X and making native single molecule profiling studies competitive with existing epigenomic assays. Finally, we conclude in Chapter 4 by considering how our method is a general tool for using 3rd generation sequencing as a read out for biological assays. We demonstrate this for two specific cases: high-depth profiling of targeted genomic regions and resolving chromatin states in single cells.

Table of Contents

List of Figures

List of Tables

Chapter 1: Epigenomic regulation of gene expression

1.1. Introduction

In healthy cells, DNA is compacted into chromatin and marked by epigenetic modifications. These dual features, which together comprise the epigenome, work in tandem to regulate gene expression by tightly controlling the accessibility of DNA to gene regulatory factors and transcriptional machinery. In tumor cells undergoing neoplastic transformation or metastasis the epigenome is strongly dysregulated, resulting in alterations to both chromatin and epigenetic marks that drive aberrant transcriptional programs. Our understanding of how this dysregulation occurs, how exactly the epigenome is altered, derives primarily from assays that utilize 2nd generation sequencing. While capable of generating vast amounts of data measuring features such as the genome-wide accessibility of chromatin or the distribution of epigenetic marks, these assays generally require fragmenting the constituent chromatin fibers in a sample, obfuscating the single-molecule nature of chromatin and producing population-averaged measurements. Therefore, to obtain high-resolution information on how epigenomic state is altered in cancer progression, we have turned to recently developed footprinting assays that leverage 3rd generation sequencing technologies to sequence single chromatin fibers. However, applying these assays to informative samples derived from patients or primary models is difficult because of the high input requirements for 3rd generation sequencing.

In this dissertation, to resolve this issue and make 3rd generation sequencing more accessible, we develop a novel method to prepare sequencing libraries from small amounts of cells or DNA using a hyperactive transposase Tn5. We then use our method to probe how the single molecule accessibility landscape changes in various cancer models. In Chapter 1, we motivate this work by summarizing how the healthy epigenome regulates gene expression, as well as the consequences of epigenomic dysregulation in tumors. We then describe current 2nd generation assays that rely on fragmentation of chromatin fibers and short-read sequencing and discuss how 3rd generation single molecule sequencing

works to yield higher resolution measurements. Finally, we examine the limitations in applying 3rd generation sequencing to primary samples, primarily the high input requirements, and highlight the contributions of this dissertation presented in Chapters 2, 3 and 4.

1.2. The epigenome dynamically regulates gene expression

The nucleosome is the fundamental unit of chromatin

The building block of chromatin is the nucleosome – ~147 base pairs (bp) of double stranded DNA (dsDNA) wound ~1.65 times around a multiprotein complex consisting of two histone H2A-H2B dimers and one histone $(H3\text{-}H4)_2$ tetramer[1]. Extensive structural studies using X-ray crystallography[2] and cryo-electron microscopy[3] have demonstrated that core histone proteins (11 – 15 kDa each) are organized into a compact octamer through close interactions via α-helical C-terminal domains[1]. DNA is then tightly wrapped around this octamer in superhelical turns and held in place through multiple dynamic contacts between exposed positively charged lysine and arginine residues and the negatively charged phosphate backbone[1,4,5]. On a single molecule of DNA, multiple nucleosomes are arranged in regular offsets. The intermediary sequence, termed "linker DNA", varies between 20 – 80 bp[1] depending on species and genomic context, and may be similarly bound by the lysine-rich "linker histone" H1[6] to form a chromatosome[7]. H1 binding near the nucleosome exit points deforms DNA and neutralizes the negatively charged linker backbone, further occluding ~160 bp of DNA[8,9]. Serine, lysine and arginine residues located on N-terminal histone tails can also be modified post-translationally (PTMs) with reversible marks such as methylation, acetylation, phosphorylation and ubiquitinoylation, with different marks playing regulatory and structural roles in altering nucleosome spacing[10]. Local compaction of DNA then defines chromatin fibers, which are generally described as euchromatic or heterochromatic to denote the degree of nucleosome-mediated accessibility or inaccessibility to the underlying primary sequence. Fiber organization in turn influences higher order DNA structure, contributing towards defining the 3D genome and eventually chromosomes.

Nucleosomes are predominantly assembled onto DNA in a tightly controlled replication-coupled manner[11]. During S-phase, a set of "canonical" histones (H2A, H2B, H3, H4) are expressed from gene clusters that lack introns[12], then rapidly translated and transported to replication forks by histone chaperones that prevent non-specific DNA binding. Deposition of the $(H3–H4)_2$ tetramer onto the newly synthesized daughter strand is facilitated by the chaperone chromatin assembly factor 1 (CAF1), which interacts with the replication clamp PCNA[13]. Shortly afterwards, two H2A-H2B dimers are installed to reconstitute local chromatin architecture[14]. In contrast, Replication-independent assembly largely proceeds through the exchange of canonical histones for, or novel deposition of, a multitude of non-canonical histone variants of H2A, H2B, and H3. These variants have distinct functions in transcription, replication and DNA repair through altering nucleosome stability[15,16] or providing alternative structural motifs for interaction with chromatin regulatory machinery[13]. Histone variants may also be strongly associated with genomic domains – the principal example being the H3 variant CENP-A, which is installed by the chaperone HJURP (Holliday junction recognition protein)[17] at centromere arms in a cell-cycle dependent manner and essential for cell survival[18,19].

Though a key property of chromatin is structural, compressing the ~2 meters of genomic DNA (gDNA) into the nucleus[1], the primary functional role is to mediate gene expression by regulating the accessibility of underlying sequence to DNA binding proteins such as transcription factors (TFs). As such, the position of nucleosomes on single DNA molecules reflects a combination of intrinsic sequence preferences, proximity to barrier elements[20–22], and ATP-driven processes such as remodeling activity[23] that translates or evicts nucleosomes to generate inaccessible or accessible regions. Seminal work to understand the relationship between sequence and nucleosome positioning used both *in vitro* and *in vivo* techniques to define motifs that strongly enhance nucleosome assembly and binding[24,25]. These include the regular spacing of AA or TT dinucleotides in ~10 bp intervals to introduce the conformational flexibility required to fold DNA, with dinucleotides located on the minor groove to interact with the histone core[26,27]. Similarly, stiff polynucleotide repeat tracts (poly-dA or poly-dT), which are highly abundant in gene promoters and throughout intergenic regions in the mammalian genome[28], are highly

disfavored. Predicting nucleosome positioning *in vivo* using these motifs works reasonably well, suggesting that core DNA sequence has evolved in eukaryotes to promote nucleosome assembly[24]. In contrast, DNA from bacteria such as does not support nucleosome formation[25].

Nucleosome spacing and regularity is also associated with various histone PTMs. Acetylation, particularly of lysine 27 on histone H3 (H3K27ac), by a set specific histone acetyltransferases (HATs) is thought to neutralize charge interactions and promote the spreading of chromatin fibers[29,30]. Conversely, though not exclusively, removal of acetylation marks by histone deacetylases (HDACs) and methylation of H3K27, and of other residues such as H3K9 by histone methyltransferases (HMTs), contribute to chromatin repression[10]. This proceeds largely through recruitment of secondary factors – H3K9me3 marked histones interact directly with heterochromatin protein 1 (HP1), which directly forms constitutive heterochromatin and impedes transcription factor binding[31,32], while H3K27me3 can recruit the Polycomb repressive complex (PRC2), facilitating both facultative chromatin compaction and further H3K27 trimethylation[33]. Critically, H3K27me3, and as well as histone PTMs that mark active transcription and enhancers such as H3K4me3, are inheritable[34], and may be reestablished through diverse mechanisms after replication-coupled nucleosome assembly – providing a "memory" of chromatin landscapes in the course of cell differentiation and growth.

Epigenetic marks influence gene regulation through direct and indirect interactions with chromatin.

Genomes are also populated by reversible epigenetic modifications to nucleotide bases that have regulatory functions. In eukaryotes, these modifications largely converge on the addition of a methyl group to different positions on the heterocyclic ring of cytosine. In humans, the most common addition produces 5-methylcytosine (m^5dC) at cytosine / guanine (CpG) contexts and is mediated by a set cytosine DNA methyltransferases – DNMT3a and DNMT3b[35], for the *de novo* deposition m^5dCpG, and DNMT1[36], for maintaining methylation on newly synthesized DNA. Greater than 70% of CpGs in the mammalian genome are methylated[37], and *de novo* methylation is particularly enriched in embryonic cells[38].

The functional role of m^5dCpG is varied across vertebrates, but in mammals serves primarily as a repressive mark[39]. Regions of high CpG density, termed CpG islands, are organized preferentially at promoters and enhancers and methylation of these CpGs acts to repress transcriptional activation. Similarly, endogenous retroviral elements[40], transposases, and imprinted genes[41] are heavily methylated to suppress their reactivation, and ablating DNMT activity *in vivo* results in genome-wide upregulation of transcription[42]. The mechanism by which repression occurs is varied, but a few models have been proposed. In one, methylation sensitive binding proteins such as MeCP2, and MBD2 facilitate repression by recruitment of HDACs and removal of acetylation marks[43,44]. In a second, m^5dCpG methylation can also disfavor binding of TFs by directly perturbing interacting bases or DNA shape features that TFs require[45,46]. The generality of this second model is less clear, as recent studies using Systematic Evolution of Ligands by Exponential enrichment (SELEX) to probe preferred TF binding motifs have established that a sizable group of TFs can prefer m^5dCpG presence in non-canonical binding motifs *in vitro* [47].

Since DNMTs are sequence agnostic, it is also unclear how CpGs are initially established at specific regions. Chromatin may play a direct regulatory role. This is supported by the observation that m^5dCpG and its oxidation product 5-formylcytosine (f^5dC)[48] destabilize or rigidify[49] nucleosome complexes by altering DNA conformational flexibility[50,51]. Indeed, studies have established a link between histone PTMs and DNMT recruitment – unmethylated H3K4 can interact with inactive regulatory factor DNMT3L to promote DNMT3A-mediated methylation[52], and at intergenic regions, H3K36me2 is required for the recruitment of DNMT3A [53]. However, *de novo* methylation may also result from simple diffusion processes targeting accessible DNA, though it is important none of these mechanisms are mutually exclusive.

Chromatin regulates transcription

In healthy cells, chromatin structure and epigenetic modifications converge to regulate gene expression both distal and proximal to gene bodies. At protein-coding genes, regulation occurs predominantly through a complex set of interactions between histone PTMs and transcription factors that

facilitate the assembly of RNA polymerase II (RNAPII) at promoters and its subsequent translocation through gene bodies. Central to this action is mitigating nucleosome occlusion of primary sequence.

At promoters in mammalian genomes, this is achieved by various chromatin remodeling complexes (CRCs) evicting or sliding nucleosomes to facilitate TF binding and assembly of the RNAPII pre-imitation complex[54]. This region of accessibility, termed the nucleosome depleted region (NDR), is a hallmark of active eukaryotic promoters[24] and is coupled with the depletion of m^5dCpG to permit TF binding. Adjacent to the NDR, the first nucleosome downstream of the TSS is strongly positioned ("+1 nucleosome") and marked with H3K4me3 by the SET1/COMPASS methyltransferase complex[55]. Interestingly, while this mark on its own is not needed for transcription at all genes[56], it is sufficient to induce transcription at Polycomb-repressed genes in specific contexts[57] and its absence is associated with decreased rates of RNAPII elongation[58].

When RNAPII is successfully assembled, release from the promoter is facilitated by the exchange of histone H2 for variant H2A.Z at the +1 nucleosome via a histone chaperone[59]. The exact way this exchange facilitates release is unclear, but studies have suggested the intrinsically less stable nature of the H2A.Z – H2B dimer maybe facilitate RNAPII bypass[15,16,60]. Subsequent progress through gene-body chromatin is facilitated by the histone chaperone FACT, which works to remove H2A-H2B dimers from nucleosomes in front of the transiting RNAPII complex and reassemble them behind[61]. The resulting hexasome intermediate can be bypassed by RNAPII, although the exact dynamics by which this occurs are a topic of current study[62] and likely involve RNAPII pausing and backtracking. Co-transcriptional modification of nucleosomes also occurs through RNAPII C-terminal domain (CTD) -mediated recruitment of the HMT SETD2, which produces H3K36me3 marks across the gene body[63,64]. The role of this mark with respect to elongation is unclear, but broadly serves to actively transcribed regions and can regulate alternative splicing[65]. Similarly, gene bodies are strongly enriched for histone H3 variant H3.3 which is catalyzed by histone chaperone HIRA[66]. H3.3 promotes replication-independent nucleosome assembly and in D. melanogaster, as well as other eukaryotes[11,13], and the degree of H3.3 levels in gene

bodies has been linked to transcription rates[67,68] – suggesting H3.3 exchange is important for RNAPII elongation.

Finally, as elongation completes and RNAPII disassembles from DNA, local chromatin architecture must be reestablished. In yeast, this is achieved by the recruitment of CRCs such as ISWI and CHD through interactions with co-transcriptionally deposited H3K36me3 and RNAPII[61]. The yeast ISWI complex Isw1b in particular directly associates with H3K36me3-marked nucleosomes via a conserved binding domain (PWWP domain), and works to reestablish regular spacing between nucleosomes in the gene body[69]. The degree of reestablishment varies by transcriptional activity. Though singly-transiting RNAPII likely does not evict nucleosomes completely, genes with rapid and continuous RNAPII-mediated transcription are more likely to suffer complete loss of histone octamers[70,71], resulting in chromatin that is heterogeneously and irregularly spaced. Exposed DNA can prime cryptic transcription, and HDACs recruited by H3K36me3 help to suppress this by eliminating acetylation marks that promote chromatin relaxation[72]. Thus, chromatin in transcribed genes is highly dynamic and the complex interplay between histone PTMs, CRCs, and RNAPII must be executed consistently in healthy cells to maintain transcriptional output.

Chromatin remodeler complexes facilitate chromatin accessibility

At sites distal to genes, chromatin also serves to regulate gene expression through the occlusion and uncovering of regulatory DNA. In contrast to gene proximal regulation, distal site accessibility is maintained through interactions involving CRCs and transcription factors.

There are four general families of ATP-dependent CRCs that assist in this process, originally characterized in yeast and highly conserved in humans – INO80, SWI/SNF, ISWI, and CHD/NuRD[73]. All four broadly slide, restructure, or evict nucleosomes through the use of a core ATP-dependent motor that engages nucleosomes via contacts with the DNA and solvent-exposed histones[74,75]. However, complexes assembled around these cores vary significantly in terms of non-catalytic subunit composition and functionality. INO80 slides nucleosomes in *vitro* and is perhaps best known for its role in catalyzing the

exchange of histone H2 with its variant H2A.Z at TSSs[11]. Similarly, ISWI and CHD remodelers generally slide nucleosomes to generate regularly spaced nucleosomal arrays in gene bodies, as described, though the exact mechanism by which precise spacing is established is debated[76]. NuRD remodelers combine CHD family ATPase cores with HDACs and CpG binding proteins, and act primarily as transcriptional co-repressors at enhancers and promoters[23]. In contrast, SWI/SNF remodelers can be recruited to target loci by histone acetylation[77], and evict nucleosomes to generate accessible DNA in opposition to Polycomb-mediated repression[78,79]. In humans, the core ATPase units assemble broadly into three stable complexes – canonical BAF (cBAF), polybromo BAF (PBAF), and non-canonical BAF (ncBAF)[80]. Though the exact regulatory roles of each complex are not well understood, cBAF tends to occupy distal regulatory sites while PBAF is found near promoters and gene bodies. On the other hand, ncBAF, which was most recently discovered and lacks one of the SWI/SNF conserved subunits SMARCB1, localizes to CTCF sequence motifs that play a key role in maintaining 3D genome architecture[81].

Regulatory DNA that is made accessible is overwhelmingly comprised of single or multiple TF binding sites[82]. In the course of cell differentiation the genome-wide accessibility landscape changes by preferentially closing regions bound by pluripotency factors such as OCT4, SOX2 and NANOG, and opening new regions associated with lineage-specific TFs[83]. Multiple models for how these new regions are opened in a cell-type specific manner have been developed. In one, “pioneer” TFs can directly interact with nucleosomal DNA, binding to their occluded sequence motifs or linker DNA and promoting spontaneous nucleosome disassembly[84–86]. Pluripotency factors that can bind nucleosomal DNA *in vitro* are proposed as canonical examples, but the degree to which do so *in vivo* is hotly debated[87].

An alternative model proposes CRCs and TFs indirectly interact to both license new regions and maintain their accessibility. This is supported by experiments showing rapid degradation or inhibition of the core ATPase subunit of SWI/SNF, BRG1, significantly decreases pluripotency factor binding – suggesting TFs alone are insufficient[88,89]. The most distinct example comes from the study of the TF Rap1 binding in yeast. Though Rap1 can engage with nucleosome protected motifs, it ultimately cannot facilitate strong interactions that lead to the establishment of open chromatin without RSC, a chromatin

remodeler with nucleosome eviction capabilities like SWI/SNF[90]. A similar behavior was also observed for the mammalian glucocorticoid receptor, where indirect interactions with Brg1 were required to initiate chromatin remodeling and stable binding at promoters[91]. The exact means by which CRCs are localized to new accessible regions however remains an active area of inquiry. One interesting possibility is that communication occurs via histone PTMs[87] – specific TFs can interact with HDACs, HATs, and HMTs and CRCs are sensitive to PTMs on remodeling substrates *in vitro*[75]. However, further study is required to determine if this coordination occurs *in vivo*.

1.3. The cancer epigenome

Having summarized how the healthy epigenome regulates expression through a range of interactions with TFs, transcriptional machinery, CRCs, and histone PTMs, we now examine how these relationships are altered in tumors.

Dysregulated states are associated with cancer progression

In normal cells, chromatin becomes more restrictive as differentiation proceeds, limiting accessibility to predominantly lineage-specific regulatory DNA and associated trans-acting factors[86,92]. However, a common hallmark of cancer genomes is prevalent hypomethylation at CpGs, with hypomethylated genes displaying significant variability in expression[93]. In colorectal cancers and glioblastomas, this may be accompanied by the hypomethylation of key oncogenes such as LY6K and RBBP6[94]. Further, cancer cells across hundreds of known types share a tendency for de-differentiation, remodeling the accessibility landscape such that enhancers associated with pluripotency are reactivated and novel TF binding sites are licensed[83]. For example, in hormone receptor sensitive prostate and breast cancers, remodeling is strongly associated with oncogenic TFs like ER[95], AR or FOXA1. AR over-expression in late-stage prostate cancer promotes genome-wide relaxation and increased accessibility at AR binding sites[96], while in primary prostate tumors AR is found to bind novel sites together with FOXA1[97]. Gain of function FOXA1 mutants in prostate organoids can have similar effects, with binding

at non-canonical motifs producing recurrently open chromatin at thousands of gene loci associated with pro-luminal transcriptional programs[98]. Chromatin plasticity can also favor cancer cell adaptability, as demonstrated by the emergence of drug resistance subpopulations in lung adenocarcinoma cell lines that are highly sensitive to HDAC inhibition[99].

Conversely, cancer cells may also epigenetically repress tumor suppressors, with the classical example being the m^5dCpG-mediated silencing of *CDK2NA* inhibitor p16. Silencing can also occur at DNA repair genes such as *MLH1* and *MSH2*, promoting a damage tolerant phenotype[100]. At these genes, m^5dCpG may spontaneously deaminate, producing C > T transitions after replication that may be deleterious for gene function[101]. Interestingly, in some tumors silencing may be mutually exclusive with mutational inactivation, as is the case for *BRCA1* in ovarian[102] and breast cancers[103] and *CDKN2A* in squamous cell lung cancers[104]. Drugs targeting hypermethylation, such as 5-aza-2'-decitabine[105], can reverse silencing and are efficacious against solid tumors[94,100]. However, epigenetic repression may also lead to further epigenomic plasticity, as in IDH mutant gliomas where heavily methylated CTCF sites remove insulation between genome compartments and promote aberrant transcriptional activation[106].

Chromatin-mediated changes in expression can also be induced by PRC2, though the effects are highly context specific[107]. PRC2 subunit H3K27 methyltransferase EZH2 can act as tumor suppressor by repressing pro-proliferative genes. This is the case in myelodysplastic syndromes (MDS) and acute myeloid leukemia (AML), where EZH2 or other PRC2 core components are recurrently mutated[108], and loss of EZH2 in hematopoietic cells produces MDS cells with a proliferative phenotype[109]. At a contrast, gain of function alterations that lead to overexpression of EZH2 may also promote cell proliferation *in vivo* and *in vitro*[107]. In B cells, EZH2 overexpression drives lymphomagenesis through transcriptional repression of B cell differentiation factors *IRF4* and *PRDM1*[108,110]. Targeting EZH2 in lymphomas using small molecule HDAC inhibitors has therefore been successful, and clinical trials have been initiated for other malignancies like mesothelioma, sarcomas, and urothelial carcinomas[111].

To what extent do epigenetic alterations initiate oncogenesis? In an extensive study examining hepatocellular carcinomas, a distinct methylation signature was identified prior to neoplastic

transformation that was relevant to later stage disease[112]. This raises the possibility that epigenomic states can act as first "hits", allowing pre-malignant cells to access transcriptional programs that are normally repressed[113]. Permissive states are also associated with the TF-driven epithelial-to-mesenchymal transition (EMT) whereby cells are reprogrammed to facilitate extraversion and colonization of distal niches[114,115]. This is achieved by widespread alteration of histone PTMs including deacetylation at promoters facilitated by NuRD complexes and gene silencing via H3K9me3-mediated formation of constitutive heterochromatin[114]. In particular, the loss of E-cadherin, which is responsible for cell adhesion, is a result of diverse silencing procedures including Polycomb repression and may be stochastically reactivated[116]. Methylation landscapes are also altered, with significant hypomethylation at various enhancers and TF binding sites across cancer types[117]. Together, this highlights the centrality of epigenomic alterations in driving later stage disease.

Mutations in chromatin remodelers facilitate oncogenic epigenomic states

What role, if any, does aberrant CRCs activity play in enabling tumors to attain and maintain these epigenomic states? Pan-cancer surveys have identified SWI/SNF complex components as highly recurrently mutated in ~20% of primary tumors[118]. These mutations are largely inactivating and unevenly distributed across subunits, with the most frequently mutated subunit, ARID1A, being non-catalytic and restricted to cBAF. Mutations are also tissue-specific, with ATPase subunits BRM and BRG1 (*SMARCA2/4*) inactivated in gliomas and lung adenocarcinoma[119]. BAF therefore seems to broadly act as tumor suppressor and loss of function leads to increased Polycomb repression[113]. In mouse models, BRG1 haploinsufficiency leads to a predisposition for mammary tumors[120]. In malignant rhabdoid tumors (MRT), a rare and aggressive childhood cancer with biallelic loss of BAF subunit SNF5 (> 98%), BAF is less stable on chromatin, resulting in a loss of remodeling at enhancers or bivalent promoters[121,122]. EZH2-mediated deposition of H3K27me3 at these loci can subsequently reactivate stem cell-like transcriptional programs that promote tumor proliferation[123].

The inverse relationship is also observed, where tumors are dependent on remodeling activity. In synovial sarcomas and MRT, loss of normal BAF remodeling leads to strong dependency on ncBAF, which can assemble without SMACRB1[80,81]. Eliminating ncBAF assembly by deleting key subunits is therefore synthetic lethal[81]. Similarly, in metastatic prostate cancer models, selective degradation of BRM and BRG1 results in decreased accessibly at AR binding sites and strong antiproliferative effects[124]. Given the near universality of loss of function mutations in SWI/SNF, this raises the exciting possibility that this dependency may be exploited in other cancers. Functional studies to further dissect exactly ncBAF dependencies translate to gene expression changes are ongoing[125].

1.4. Sequencing assays for studying the epigenome

Insights into both healthy and dysregulated chromatin states presented in Chapters 1.1 and 1.2 are largely derived from biological assays that use sequencing as a quantitative readout. This suite of assays can measure nucleosome occupancy, histone modification frequencies, and m^5dCpG methylation levels by fragmenting chromatin to produce short molecules that are compatible with short-read sequencing. However, the result is an aggregate measurement across the input sample, fundamentally ignoring sample heterogeneity and the single molecule nature of chromatin.

In general, we can consider an assay as comprised of two parts – 1) a method for *encoding* a property into a measurable signal and 2) a system for detecting, or *reading out*, the encoded signal with high accuracy. Using this framework, we discuss how current epigenomic assays work and how 2nd generation sequencing can be used as a high information content readout.

Review: 2nd generation sequencing

DNA sequencing describes the process of determining the order and identity of nucleotides in a DNA polymer. The predecessor of the modern sequencing reaction (“Sanger sequencing”) utilizes dideoxynucleotide chain terminators to irreversibly halt DNA replication of template molecules[126]. Nucleotide identity, as determined by the specific terminator, can then be localized to positions in the

template based on the length of the terminated fragment. Though read lengths are short (< 1000 bp) and costly, Sanger sequencing paved the way for modern 2nd generation platforms that ultimately follow a similar strategy.

In the most popular sequencing chemistry "sequencing by synthesis" (SBS), commercialized by Illumina over the past two decades, billions of target molecules are immobilized on the surface of flow cells and replicated by the stepwise incorporation of fluorescently labeled nucleotides using an engineered polymerase[126]. Fluorescent signals associated with each base incorporation are detected optically, reading in parallel up to nearly ~25 billion short (~200 – 400 bp), highly accurate (< 0.01% error rate) DNA fragments ("reads") on current instruments[127]. While the original human reference genome was produced at great cost over multiple years[128,129], Illumina SBS platforms can deliver high-depth genomes for less than $1000 in under 2 days. Whole genome sequencing (WGS) using SBS has therefore found wide utility, from the de novo assembly of reference genomes[130] to identifying disease-associated genetic variation in patient samples[131,132]. The latter use case has become increasingly important for understanding how inherited variation predisposes patients to cancer, as well how somatic variation can drive oncogenesis and metastasis[133]. Many diagnostics also now use sequencing, including tests to detect chromosomal abnormalities in early pregnancy[134] or circulating tumor DNA in blood that reflects residual disease[135–137].

Critical to these applications is the process of preparing DNA fragments for compatibility with SBS ("library preparation"). While a known sequence of DNA is required to prime polymerization in Sanger sequencing, modern platforms require the addition of short, standardized DNA "adapters" to molecule ends. These adapters, as the name suggests, are oligonucleotides that serve as an interface between the target molecule and the sequencing chemistry. For Illumina SBS, adapters are required for both immobilizing library molecules to the flow cell surface and as primer handles for PCR amplification to increase the amount of library available for sequencing[126]. Two general techniques have emerged for preparing input material into libraries with minimal GC content biases and consistent fragment size. In ligation-based library preparation, dsDNA is first sheared to ~200 – 500 bp fragments using enzymatic,

acoustic, or mechanical means, then molecule ends enzymatically blunted and adenines added to the 3' ends by non-templated addition. These adenines are used for the overhang ligation of T-tailed SBS adapters[138], which can then be used for PCR amplification. Another library preparation strategy uses hyperactive transposases to simultaneously fragment and insert sequencing adapters into target DNA ("tagmentation")[139,140], reducing a multistep process to a single convenient reaction. The choice transposase, Tn5, has been engineered to bind DNA with high affinity, permitting library preparation from a range of input sources including cDNA[140] and circularized genomes[141–143].

Both procedures have also been adapted to both to work with extremely small amounts of input material, benefitting clinical sequencing efforts. Optimized buffers containing molecular crowding reagents enable reproducible tagmentation from as little as ~100 picograms (pg) of DNA [140,144]. For ligation-based protocols, application specific enzyme optimizations[145] and on-bead molecule capture can facilitate library preparation from highly fragmented ancient[146,147] or cell free DNA. Ligated adapters can also contain unique molecular identifiers (UMIs)[148] , random degenerate sequence that can uniquely index single molecules. Computational merging of sequenced reads carrying the same UMI can reconstruct the original source molecule with extremely low error rates 10^{-8} (1 event / 100M), facilitating rare variant detection and single cell DNA sequencing[149–151].

Encoding chromatin measurements via enzymatic fragmentation

Beyond WGS, the ability to sequence billions of arbitrary DNA fragments has made 2nd generation sequencing an ideal readout for assays that fragment chromatin. The most prominent set, collectively termed "footprinting assays", have evolved from classical nuclease digestion protocols that degrade accessible chromatin and release fragments protected by DNA binding proteins[152,153]. While historically these fragments were *read out* by gel electrophoresis, their short length makes them perfect for Illumina SBS. Mapping sequenced fragments back to the reference genome can reveal, often with base-pair resolution, the genomic coordinates of the original protein-DNA footprints across the millions of cells processed. Footprint size can inform the exact nature of the interaction, ranging from bound

transcription factors (< 50 bp), to di- and tri- nucleosomes complexes (> 300bp)[82]. After appropriate normalization to account for nuclease sequence biases, accessible DNA is then defined as regions lacking footprinting signal.

Two popular nucleases, micrococcal nuclease (MNase-seq)[154] and DNase I endonuclease (DNase-seq)[155,156], have been adopted by multi-center consortia (*e.g.* ENCODE [157]) to map the landscape of protein-DNA interactions across hundreds of human tissues and cell lines. The resulting accessible regions, often described as DNase I hypersensitivity sites (DHSs), reflect sequences that are predominantly open across cells. Fragmentation assays can also be used to directly sequence accessible DNA. The assay for transposase-accessible chromatin (ATAC-seq)[158,159] uses Tn5 transposase to directly tagment nuclei, resulting in the preferential insertion of sequencing adapters into accessible chromatin. The benefits of this approach, chiefly the direct measurement of accessibility and high sensitivity of transposase-mediated library preparation, have made it exceedingly popular in deciphering changes in chromatin state in rare or precious samples, including human tissues post-mortem[160,161].

Methods have also been developed to localize fragmentation to specific proteins of interest rather than footprinting all protein-DNA interactions genome-wide. For example, to improve measurements of nucleosome positioning specifically, various studies have generated histone H3 and H4 cysteine mutants that bind phenanthroline-Cu^{+} reagents to facilitate peroxide-mediated cleavage at very specific offsets from the nucleosome core particle[162–164]. Sequencing the resulting fragments, which exactly span the accessible linker region between two dyads, has produced some of the highest resolution maps of well positioned nucleosomes in yeast and mice. The generalizability of this approach is low however, as mutants must be generated for each protein target. Thus, a popular alternative is Chromatin Immunoprecipitation and sequencing (ChIP-seq)[165], where sheared DNA associated to an epitope of interest is captured by an antibody and converted into a sequencing library. Advances in antibody engineering have enabled ChIP-seq to map the genomic locations of a range of informative epitopes including histone PTMs, transcription factors, and even non-B DNA structures like R-loops[166] or G-quadruplexes[167,168]. When normalized against a mock-immunoprecipitated control, sequencing reads can

also be used to quantify epitope occupancy and enrichment[169]. Extensions to ChIP-seq have further improved resolution by replacing shearing with MNase digestion (MNase-ChIP)[170] or using exonuclease treatment to further reduce immunoprecipitated DNA to only the protein-protected fragment (ChIP-exo)[171]. Native ChIP, without crosslinking to stabilize protein-DNA interactions, has also been promising in reducing high input requirements[172].

However, the most significant improvement to ChIP-seq are assays that tether Tn5 transposase directly to antibodies (CUT&Tag)[173–176] to localize tagmentation and library preparation to only DNA occupied by the target epitope *in situ*. This avoids off-target immunoprecipitation altogether, and the resulting increase in signal using this encoding strategy has enabled CUT&Tag to map histone marks in as few as ~100 cells[176]. Further, loading transposases with barcoded SBS adapters and using antibodies with different Fc regions facilitates multiplexed mapping of multiple marks in single cells[177]. Though CUT&Tag has not yet to date recapitulated the all known ChIP-seq footprints in standardized consortium collections[178], it has been successfully employed across sample input levels and cell types[179], and will likely surpass both chemical cleavage and ChIP-seq as the fragmentation assay of choice for mapping DNA-interacting proteins.

Encoding base modifications as changes in primary sequence

Instead of fragmenting DNA, a subset of assays use chemical and enzymatic conversion of primary sequence to *encode* the presence of base modifications. Bisulfite sequencing[180] uses harsh sodium bisulfite treatment to convert all cytosines except those modified by methylation (m^5dC and hm^5dC) to uracils. Subsequent PCR amplification recodes uracil as thymine, producing C > T transitions at all non-modified cytosines in sequenced reads. Aggregate methylation level across all cells in a sample is then determined per CpGs motif as the fraction of aligned mutated reads out of the total. Improvements to this method leverage the methylcytosine dioxygenase TET2 to oxidize m^5dC and hm^5dC to their end product 5-carboxylcytosine (ca^5dC), followed by cytosine deamination at unmodified cytosines via APOBEC3A

[181]. Enzymatic treatment is less destructive to primary templates, resulting in superior detection of m^5dC from inputs as low as ~100 pg.

While bisulfite sequencing has become the gold standard for clinical detection of hyper- or hypo-methylation at key tumor suppressor genes, modifications to primary sequence make it difficult to distinguish between endogenous C > T variants or mutations induced by deamination. Two general strategies have emerged to solve this problem. In TET2-assisted pyrimidine borane sequencing[182,183], TET2 is once again used to convert m^5dC to ca^5dC, which is then selectively reduced to uracil using pyrimidine borane. Uracil to thymine recoding still proceeds via PCR, but sequenced C >T events at CpGs directly measure m^5dC presence, allowing for greater confidence in estimating sample-wide methylation levels per CpG. Measuring positive signal (*i.e.* CpGs that are methylated, rather than CpGs that are unmethylated) is also achieved by 5-letter sequencing[184], recently developed by the Balasubramanian group and commercialized by Cambridge Epigenetix. Unmodified cytosines are still deaminated using APOCBEC3A, but primary unmodified sequence is recovered because individual strands of DNA duplexes are barcoded and sequenced independently. Because of this *post hoc* computational error correction, 5-letter sequencing is under consideration to replace bisulfite sequencing as a clinical standard.

Encoding chromatin accessibility using methyltransferases

Finally, a third less popular encoding strategy uses DNA methyltransferases (DNMT) to add base modifications at targets of interest[169]. DNMNTs transfer methyl groups abstracted from cofactors to specific positions on DNA nucleobases[185]. The most common methylated bases include the well-characterized m^5dC, N6-methyladenosine (m^6dA), and N4-methylcytosine (m^4dC). While m^4dC has only been identified in bacteria[186] and m^5dC occurs genome-wide at CpG motifs, m^6dA is infrequently present in vertebrate genomes at GATC motifs, particularly during early embryogenesis[187,188]. Hence, adenine methyltransferases (m^6dAse) have been favored as labelling agents, either as N- or C-terminal fusions to target proteins or through exogenous introduction after nuclei permeabilization to capture chromatin

accessibility. One such method, DamID[189,190], leverages endogenous expression of m^6dAse – protein fusions to track genome-wide transcription factor biding patterns over time, but suffers from poor signal due the use of E. coli DNA m^6dAse, which methylates only GATC contexts. The discovery of a non-specific m^6dAses EcoGII and Hia5[191,192] circumvented these restrictions. An improved labelling method MadID[193] uses EcoGII as a fusion partner, and has been used to map telomere-to-telomere and lamina-associated domain contacts with excellent signal to noise across GATC-rich and -poor domains. GpC methyl-transferases, which introduce m^5dC marks at GpC contexts (Gm^5dCase), have also remained popular for labelling accessible DNA[194] – with recent methods coupling GpC labelling with the detection of m^5dC methylation for a multimodal readouts of accessibility [195]. Generating more exotic base modifications for proximity labelling has also been considered – for example using engineered photoactivatable flavin proteins to rapidly generate 8-oxoguanine lesions for recording transient binding events [196,197].

While m^6dAse and Gm^5dCase labelling is non-destructive and therefore preserves chromatin fiber integrity, directly detecting exogenous epigenetic modifications is not possible with 2nd generation platforms. Hence, enzymatic fragmentation assays have been adapted to recode the position of modified bases into fragments that can be sequenced via SBS. In MadID, this is achieved by shearing DNA and immunoprecipitating only fragments containing m^6dA, producing a distribution of reads centered on the modified base. Similarly, in Gm^5dC labelling of accessible DNA, bisulfite treatment is used to quantify methylation levels as a proxy for accessibility. Recoding procedures therefore ultimately decrease the resolution of methyltransferase labeling methods, making them unpopular for studying chromatin state as compared to ChIP-seq, MNase-seq, or ATAC-seq.

1.5. Non-destructive footprinting using third generation sequencing

We would like to understand how single chromatin fibers are altered in cancer progression. However, as we have described, all methods that use enzymatic fragmentation for *encoding* measurements and 2nd generation sequencing as a *readout* ablate this information. Further, sequencing clinically relevant

samples requires addressing challenges in input restrictions, sample quality, and high accuracy for confident variant calling. We therefore seek new methods that can capture fiber information in *cis* without relying on short reads.

Encoding accessibility as epigenetic modifications using methyltransferases is inherently non-destructive and avoids modifying the primary sequence, presenting an attractive solution. When coupled with 3rd generation sequencing platforms, both primary sequence and epigenetic modifications can be read out on long (upwards of 10 kilobase [kb]) single molecules. Recently, a series of methods have utilized this combination of *encoding* and *readout* to ascertain single-molecule chromatin accessibility across the human[198–200], mouse[201], yeast[202,203], and fly genomes[198]. We first describe the aspects of 3rd generation platforms that have enabled these methods, discuss their commonalties, and then highlight limitations that make it difficult to apply them to primary samples.

Third generation platforms for sequencing single molecules

Within the last decade, novel 3rd generation sequencing technologies have challenged the dominant paradigm of Illumina SBS. Rather than rely on stepwise incorporation of nucleotides, 3rd generation platforms use radically different sequencing chemistries to genotype and epigenotype DNA. Though numerous technologies have been proposed and are under development, the market has largely coalesced around two – Nanopore sequencing from Oxford Nanopore (ONT) and SMRT sequencing from Pacific Biosciences (PacBio). In this dissertation, we primarily rely on SMRT sequencing because it meets our requirements for highly accurate sequencing. Nonetheless, we provide a summary of Nanopore sequencing for completeness.

In Nanopore sequencing, single stranded DNA molecules are incrementally ratcheted through engineered protein pores ("nanopores") embedded in an electrically resistant polymer flow cell. An externally applied voltage simultaneously drives negatively charged DNA through nanopores and induces an ionic current that changes as different nucleobases transit the pore. Measuring this signal using specialized sensors followed by computational deconvolution of the ionic current allows for relatively

accurate base calling (~85 – 94%, R9.4 chemistry)[204]. Though the concept was originally proposed nearly 40 years ago, technical advances in both pore and motor protein engineering were required to deliver the required throughput for commercialization. Presently, ONT's highest throughput device can generate 290Gb of data using a flow cell with ~3000 pores[205], with read lengths recorded as high as 2.3Mb[206] under optimized sequencing conditions. Numerous studies have also demonstrated that ionic currents change in response to base modifications[207–209]. Computational frameworks built on recurrent neural networks can then be used to call modifications in read ensembles as well as at single molecule resolution. The configuration of nanopores also allows for both near-real time readouts of sequenced bases and precise control over pore transit rates. Together, these features have be used to develop adaptive sequencing protocols[210,211], where sequencing coverage can be targeted to fragments of interest by rapidly reversing nanopore transit of undesirable DNA. This adaptability, coupled with high throughout, epigenotyping and ultra-long read lengths have attracted numerous method development efforts to the ONT platform. Independently, multiple academic groups have created rapid response programs using nanopore sequencing to facilitate the accelerated diagnosis of newborns or patients with unknown genetic diseases[212–214]. Further, while overall read accuracy has lagged, new sequencing modes released by ONT utilize repetitive re-reading of single molecules as well as novel chemistries (Q20+) to improve modal accuracies to as high as 99.9% at ~20X coverage[215,216].

At a contrast, Single Molecule Real Time (SMRT) sequencing relies on an ultra-processive DNA polymerase and specialized optics to track polymerase-mediated base addition in real time[217]. Central to this process is the zero-mode waveguide (ZMW), a nanowell structure with a volume of ~20 zeptoliters (~20 x 10^{-12} liters) and a diameter smaller than specific wavelengths of light. Double stranded DNA molecules between 2 – 25 kb in size are first converted into templates for rolling circle amplification by ligating annealed hairpin adapters ("SMRT adapters") to DNA ends (**Figure 1.1** – marker 1). Templates are then annealed with engineered sequencing polymerases (originally derived from bacteriophage polymerase Φ29) and single polymerase / DNA complexes anchored to the bottom of each ZMW. Complexes are illuminated from below by a laser and nucleotides with base-specific fluorescent dyes

conjugated to their terminal phosphate groups are added to initiate polymerization. Base incorporation by the polymerase momentarily holds the fluorescent dye in the laser path, triggering fluorescent emission of photons that are captured within the ZMW and detected before the linked pyrophosphate is cleaved to form the phosphodiester bond. This reaction can then continue for hundreds of thousands of bases (on the order of ~300 kb), producing extremely long polymerase reads that are effectively re-reads ("subreads") of each strand of the original library molecule due to the rolling circle process (**Figure 1.1** – marker 2). Subreads are merged computationally, taking advantage of the randomized nature of incorporation errors, to produce a highly accurate circular consensus read per single molecule ("CCS" read)[218] (**Figure 1.1** – marker 3).

On the latest PacBio instruments, flow cells ("SMRTcells") contain between 8M – 25M ZMWs each, generating multiple millions of CCS reads per run (~2 – 3M on the Sequel II, 4 – 6M on the newer Revio[219]), with nearly all (> 90%) meeting the HiFi criteria (*per-base* accuracy > 99.9%). The high single-molecule accuracy and long read lengths of HiFi sequencing have made it the go-to favorite for producing reference grade genome assemblies. For example, the recently completed telomere-to-telomere human reference genome relied heavily on HiFi reads to close assembly gaps, while using nanopore reads for long-distance scaffolding[220–222]. Further, native sequencing without PCR significantly reduces GC biases, and the SMRT sequencing polymerase is not affected by highly repetitive sequence content as in SBS. These properties have recently been leveraged to both genotype repetitive content in disease-relevant genes such as *FMR1*[223,224], as well as comprehensively phase highly related genes that are difficult to map such as *SMN1* and *SMN2*[225].

Critically, SMRT sequencing is highly sensitive to nucleotide modifications – a property which has been leveraged by methyltransferase footprinting methods for native methylation detection. When the SMRT polymerase cognates against bases with epigenetic modifications, it temporarily pauses – extending the duration between the previous base incorporation and the next[217]. This time interval, called the inter-pulse duration (IPD), along with the width of the subsequent fluorescent pulse (pulse width, PW) are two highly informative kinetic parameters produced per base sequenced that uniquely characterize the

epigenetic modification and the surrounding sequence context[226]. While earlier studies deemed changes in PW and IPD too subtle for detection[227], machine learning models, particularly convolutional and recurrent neural networks, trained on these kinetic parameters using whole genome amplified (unmodified, negative control) and methyltransferase treated (modified, positive control) DNA can accurately detect m^6dA and m^5dC with single base and single molecule resolution[201,228,229]. Single molecule accessibility techniques have therefore benefitted from advances in modification detection to efficiently call exogenous m^6dA marks and resolve stretches of accessible sequence.

Measuring single-molecule chromatin accessibility using 3rd generation sequencing

How exactly is methyltransferase footprinting combined with 3rd generation sequencing to generate single molecule accessibility profiles? Though methods differ in their choice of encoding strategy (m^6dAse vs. Gm^5dCase) and readout (PacBio HiFi vs. ONT Nanopore), they largely follow a common workflow presented as a generalized schematic in **Figure 1.2a**. Chromatin fibers dialyzed from nuclei, or nuclei *in situ*, are treated with excess methyltransferase to deposit methylation marks only in accessible DNA, taking inspiration from the methyltransferase-mediated labelling developed for 2nd generation sequencing (**Figure 1.2a** – marker 1). DNA is then deproteinated, extracted, and subjected to platform-specific library preparation to produce long fragments (3 – 20 kb) for sequencing (**Figure 1.2a** – marker 3, SMRT sequencing shown**)**. After sequencing and basecalling, ionic current (ONT) or kinetic measurements (PacBio) are then used as input for network models to predict modification probabilities for each relevant base (A / T or GpC). These per-base modification probabilities are then integrated, often using hidden Markov models, to define methylase-inaccessible footprints on each molecule that reflect the original footprint of DNA-bound proteins, usually nucleosomes or transcription factors (**Figure 1.2a** – marker 4). The resulting data are contiguous alternating accessible and inaccessible stretches on each sequenced molecule.

Where these methods differ significantly is in their inherent resolution. In MeSMLR-seq[203] and nanoNOMe[200], authors used M.CviPI, a Gm^5dCase and footprint calling required specialized procedures

that accounted for the sparsity of GpC motifs across the genome. At a contrast, SAMOSA[199] / Fiber-seq[198] / SMAC-seq[202] used the non-specific m^{6}dAse EcoGII to label accessible adenines, which are significantly less sparse – the median distance between adjacent adenines in the human genome is 1 bp while the median distance between adjacent GpCs is as high as 14 bp (**Figure 1.2b**). This improved resolution contributes to more accurately determined footprints, with recent studies using m^{6}dAse footprinting specifically to capture the stochastic "breathing" at nucleosome ends[201]. Hence, in this dissertation, we use m^{6}dAse footprinting exclusively.

What insights can single molecule accessibility profiling provide? Foremost, single molecule resolution deconvolutes accessible chromatin, as defined from bulk fragmentation assays, into heterogenous populations. For example, studies using SAMOSA have demonstrated that heterochromatin, classically modeled as static and inaccessible, is surprisingly enriched for heterogeneous, irregularly offset nucleosomes. Similarly, studies using Fiber-seq to study DHSs have found that only a subset of fibers are fully uncovered, and the likelihood of two adjacent DHSs both being accessible on a single molecule was ~53% and strongly dependent on distance. Following the notion of mapping coordinated accessibility changes on single molecules, MeSMLR-seq used nanopore reads to show that a set of promoters classically defined as "open" when using fragmentation assays actually displayed combinatorial accessibility patterns on single molecules. A similar analysis was presented by SMAC-seq for two yeast genes, *TMA10* and *HSP26*, tracking joint changes in accessibility during the integrated stress response that are missed by simply measuring bulk nucleosome occupancy.

Footprinting is also not limited observational studies. Recent work by Abdulhay et al.[201] profiled single molecule accessibility on reconstituted chromatin *in vitro* and *in vivo* after remodeling with SNF2h, the essential ATPase component of ISWI CRCs. The resulting positions of nucleosomes was then used to dissect the mechanism by which ISWI CRCs creates regularly spaced arrays, suggesting it depends strongly on nucleosome density and has heterogenous (regularly and irregularly spaced fibers) outcomes. Thus, single molecule studies can reveal nucleosome positioning relationships in *cis* that have been

inaccessible by the past decade of short-read fragmentation-based assays. It would therefore be valuable to use these techniques to study chromatin in primary samples.

1.6. 3rd generation platforms have high input requirements.

The primary barrier to applying single-molecule chromatin accessibility profiling to clinical samples is the high input requirements of both m^6dAse footprinting and HiFi sequencing. Primary samples are inherently precious and must be carefully apportioned between different diagnostic tests. For HiFi sequencing and single-molecule accessibility profiling to become routine, input amounts should become competitive with existing native Illumina SBS which uses as little as ~25 nanograms (ng) gDNA[230,231]. Surveying existing single molecule accessibility assays reveals input requirements ranging from ~1 µg (MeSMLR-seq) to as high as 6 µg (SMAC-seq) (see Chapter 2.11 – **Supplementary Note** on Input-Reduction). Even PacBio's commercially available library preparation kit suggests at least 1 µg of input material to produce enough library for sequencing a single SMRTcell[232].

To understand the limitations of SMRT sequencing, we can calculate the minimal amount of library material required for one sequencing run. On the most widely available PacBio platform, the Sequel II, a minimum of 115 µL of ~40 picomolar (pM) library, or ~4.6 femtomol (fmol), is needed to saturate one SMRTcell after the associated machine loading process. Assuming standard libraries can range from an average size of 2 kb to 15 kb depending on the specific application, this indicates anywhere between 6 ng to 43 ng of library is required. The amount of input material then depends on the library preparation protocol. Current ligation-based protocols for attaching SMRT sequencing adapters to target molecules have low efficiency even at higher input amounts (on the order of 1 – 10%)[233], and do not scale well to lower amounts. For this reason, commercially available PacBio "low-input" kits targeting as low as ~100 ng rely heavily on PCR, negating the benefits of native sequencing[232]. It is therefore evident that new methods are needed for adding SMRT adapters with higher efficiency.

Transposase-mediated library preparation is highly sensitive & customizable

A possible solution for improving the sensitivity of library preparation is tagmentation. As described in Chapter 1.4, engineered Tn5 transposase can be used to simultaneously fragment and insert Illumina SBS sequencing adapters into dsDNA. Tn5 itself is a monomer that readily dimerizes after binding to a short stretch of high affinity sequence termed the "mosaic end"[234], derived from the transposon Tn5 normally propagates in bacteria. Because only the mosaic end is required to for both dimerization and subsequent transposition, arbitrary sequence can be added at the 3' end. In Illumina library preparation these sequences are SBS adapters, while other methods have incorporated barcode sequences for sample- or cell- level indexing[140,144]. Mosaic ends have even been coupled to fluorescent dyes[235,236], allowing for the preferential integration and visualization of fluorescent markers at accessible chromatin like to ATAC-seq. The transposition reaction is also extremely sensitive. Bead-bound Tn5 can tagment as little ~1 ng of input material, and in modified buffer conditions down to ~100 pg[140,144,237]. Tn5 can even be used to introduce primers into single cells after lysis, providing handles that can be used to amplify DNA the many orders of magnitude for required sequencing[238].

However, transposition by Tn5 also leaves behind a genomic scar – nine bases of single stranded DNA between the inserted and original sequence. This scar arises from the strand-transfer complex that forms after Tn5 engages DNA[234]. The DNA duplex is cleaved 9 bp apart on different strands and the free ends attached to the sequences carried by Tn5 dimer, producing two exposed single strands of DNA centered around a double stranded break. In protocols that utilize PCR after tagmentation, such as Illumina library preparation or ATAC-seq, polymerization after deannealing the DNA duplex replaces scarred DNA with newly synthesized strands. The original source strands, scarred or otherwise, are then diluted out and eventually lost.

Transposase-mediated library preparation enables single molecule sequencing assays

In this dissertation we introduce a novel method built on the principle of extending transposase-mediated library preparation to SMRT sequencing. We primarily achieve this by identifying enzymatic

conditions to efficiently seal the residual genomic scar after transposition, producing a closed molecule similar to the canonical template used in HiFi sequencing.

In Chapter 2, we describe our method, which we term SMRT-Tag, as well as discuss methodological choices and optimizations that enable the preparation of HiFi libraries from as little as ~40 ng of input material – close to the theoretical lower limit described in above section. We extensively benchmark SMRT-Tag in comparison to standard library preparation protocols and reference data, and demonstrate that it enables highly accurate detection of germline variation and m^5dC marks on single molecules. We then combine SMRT-Tag with *in situ* m^6dAse footprinting in a derivative assay we call SAMOSA-Tag, using direct transposition of footprinted nuclei to efficiently prepare HiFi libraries *in situ*. Applying SAMOSA-Tag to cancer cell lines and prostate patient derived xenograft models (PDXs), we obtain single molecule accessibility profiles and capture m^5dC marks on millions of chromatin fibers. Integrating these measurements with existing datasets from ATAC-seq and ChIP-seq for histone modifications reveals a distinct loss of nucleosome fiber regularity associated with metastasis.

Chapter 3 presents further improvements that address losses associated with m^6dAse treatment specifically. We find that through on-bead immobilization of nuclei and cells, sample handling losses can be minimized while maintaining compatibility with both m^6dAse footprinting and tagmentation. Processing nuclei and cells after bead-immobilization produces similarly high-quality single-molecule chromatin accessibility measurements from as few as ~10,000 cells (~60 ng gDNA) – bringing m^6dAse footprinting closer in line with input requirements for 2nd generation assays. We then demonstrate how lower input requirements permit studying chromatin fibers in multiple cell subpopulations sorted from a single heterogenous primary sample.

Finally, in Chapter 4 we present concluding statements, as well as future directions where we believe SMRT-Tag will be of material value to the broader scientific community. Primarily, we believe SMRT-Tag is a general strategy for bringing genomics assays to 3rd generation native sequencing, due to barriers it removes in terms of input. We further discuss our preliminary efforts developing two of these directions – targeted genome enrichment for extremely high resolution accessibility mapping of disease-

associated gene loci, and single cell sequencing for the unbiased characterization of cell-type specific changes in chromatin fiber architecture. In summary, SMRT-Tag is a versatile tool that can enable 3rd generation sequencing as a readout for biological assays.

1.7. Figures

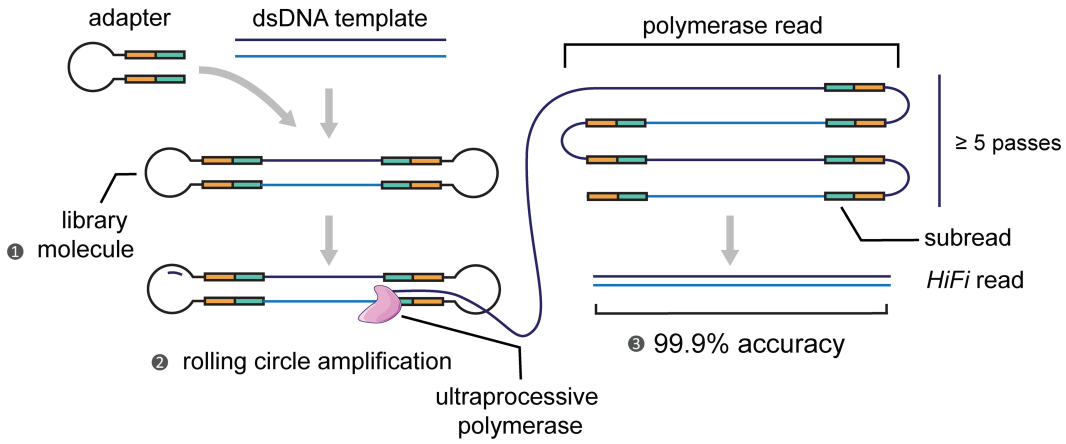


Figure 1.1: Schematic of PacBio circular consensus sequencing. Marker 1 – Hairpin sequencing adapters (SMRT adapters) are ligated to both ends of double stranded DNA templates to produce closed library molecules. Marker 2 – Library molecules are then sequenced using an ultra-processive DNA polymerase, and live base addition recorded via detection of fluorescently labeled nucleotides. Marker 3 – Individual subreads in one long (200 – 300kb) polymerase read are computationally merged and error corrected, producing a circular consensus read. Consensus reads reflect the sequenced single molecule, and are termed HiFi if the per-base accuracy is Q30+ (> 99.9% accuracy).

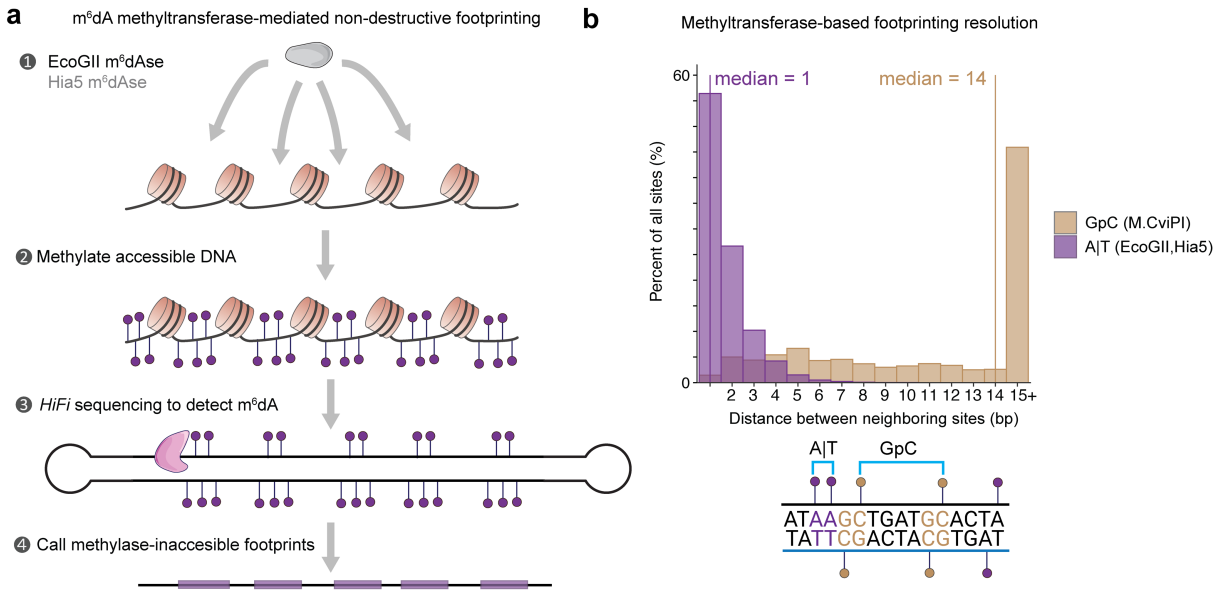


Figure 1.2: Methyltransferase footprinting combined with 3rd generation sequencing for capturing single molecule accessibility. a.) A generalized schematic for single-molecule chromatin accessibility assays that use non-destructive adenine (or cytosine) methyltransferases to label accessible DNA Common m^6dAses include EcoGII and Hia5. Marker 1 – chromatin is treated with excess methyltransferase. Marker 2 – methyltransferases selectively deposit methylation on nucleotides residing in accessible chromatin. Marker 3 – primary sequence and exogenous methylation marks are detected by 3rd generation sequencing, shown here as PacBio HiFi sequencing. Marker 4 – Per-base predictions of methylation marks are integrated to produce methylase-inaccessible and methylase-accessible footprints on single molecules. **b.)** Distribution of distances between adjacent adenines (A or T depending on DNA strand) and adjacent GpC motifs (palindromic on either strand) across all autosomes in the GRCh38.p6 reference assembly. A schematic of this calculation is presented below. GpC motifs are targets for Gm^5dCpG methyltransferases such as M.CviPI (gold) and A or T substates for non-specific adenine methyltransferases such as EcoGII or Hia5 (purple). The median distances between adjacent A or T bases versus GpC motifs is 1 bp vs. 14 bp.

Chapter 2: Direct transposition of native DNA for high-sensitivity multimodal single-molecule sequencing

2.1. Abstract

We present SMRT-Tag: a multiplexable, PCR-free approach for constructing low-input, single-molecule Pacific Biosciences (PacBio) sequencing libraries using Tn5 transposition. SMRT-Tag conservatively reduces the input DNA required for PacBio sequencing by 95 – 99%; libraries prepared from as low as 40 nanograms (ng) human reference DNA (~7,000 human cell equivalents) enable sensitive detection of genetic variation and CpG methylation, with error rates comparable to current state-of-the-art. We further combine SMRT-Tag with *in situ* adenine methyltransferase footprinting of nuclei to develop an approach called SAMOSA-Tag, which facilitates joint analysis of nucleosome repeat length, CTCF occupancy, and CpG methylation on individual chromatin fibers in various cell types. We apply SAMOSA-Tag to perform single-molecule epigenomic measurement of CTCF occupancy, CpG methylation state, and nucleosome positioning in precious matched primary and metastatic human prostate cancer cells from patient-derived xenograft (PDX) prostate cancer models. Together, our novel approaches promise to enable basic and clinical research by offering scalable, sensitive, and multimodal single-molecule genomic and epigenomic analyses in diverse low-input settings.

2.2. Third-generation sequencing platforms are powerful tools for mapping the genome

Third-generation, single-molecule long-read sequencing (SMS) technologies deliver highly accurate genomic and epigenomic readouts of kilobase to megabase-length nucleic acid templates[239] . SMS has facilitated the characterization of previously intractable structural variants and repetitive regions[222,240], assembly of a gapless human genome, and high-resolution functional genomic profiling of both DNA[198–200,202,241] and RNA[242,243]. The multimodality of SMS has also been exploited by single-molecule chromatin profiling methods such as the single-molecule adenine methylated oligonucleosome sequencing assay (SAMOSA)[199,201], Fiber-seq[198], directed methylation long-read sequencing (DiMeLo-

seq)[241], nanopore sequencing of nucleosome occupancy through methylation (NanoNOMe)[200], and others[202,203]. These approaches establish a paradigm for simultaneously measuring functional genomic information (*e.g.,* histone / transcription factor-DNA interactions) as separate SMS "channels" along with primary sequence and endogenous epigenetic marks.

Over the past decade, improvements in cost, data quality, read length, and computational tools have led to the rapid maturation of Pacific Biosciences (PacBio) and Oxford Nanopore (ONT) SMS platforms. For example, the cost of PacBio sequencing has decreased from $2,000 to $35 per gigabase (Gb), concomitant with increases in yield per instrument run (100 Mb to 90 Gb), read length (from ~1.5 kb to 15-20 kb), and accuracy (from ~85% to >99.95%)[244]. A key limitation of SMS, however, remains the amount of input DNA required for PCR-free library preparation (typically ≥1 μg, or ≥ 150,000 human cells). While low-input protocols are available, they often rely on PCR amplification, which erases modified bases, or does not provide adequate coverage of Gb-scale mammalian genomes. This significant obstacle for single-molecule genomic and epigenomic analyses precludes routine analysis of rare and post-mitotic cell types, microorganisms, and clinical samples. As such, the use of SMS has generally been limited to genome assembly and variant detection in clinical and population genetics.

Simultaneous transposition and fragmentation (*i.e.* "tagmentation") using hyperactive Tn5 transposase loaded with sequencing adaptors poses an attractive solution to this problem[139]. Tagmentation serves as the basis for a variety of genomic protocols, including low-input epigenomic profiling[139,158,245], cellularly-resolved monoplex[246] and multiplex[247–249] sequencing, highly accurate duplex sequencing[250], and *in situ* sequencing[251]. We sought to leverage Tn5 for low-input Pacific Biosciences (PacBio) SMS[217] by developing **s**ingle-**m**olecule **r**eal **t**ime sequencing by **tag**mentation (SMRT-Tag) – an amplification-free, low-input library preparation method for simultaneously profiling the genome and epigenome on PacBio sequencers. Here, we detail our optimization process and design choices for the SMRT-Tag protocol, delineate how SMRT-Tag can be used to generate > 7 Gb of high-quality PacBio sequencing data from as low as 40 ng of input DNA, and describe how SMRT-Tag can enable high-sensitivity single-

molecule long-read footprinting of both cell lines and difficult-to-work with patient derived xenograft (PDX) primary tumor cells (an assay we term SAMOSA-Tag).

2.3. Tunable and multiplex Tn5-mediated construction of PacBio libraries

SMRT-Tag (workflow shown in **Figure 2.1a**) relies on tagmentation of high molecular weight genomic DNA (gDNA) by a triple-mutant Tn5 enzyme (hereafter referred to as Tn5), which allows concentration-dependent control of fragment size[144]. We loaded Tn5 with custom oligonucleotides composed of the hairpin PacBio adaptor and mosaic end sequences necessary for transposome assembly and assessed the tunability of gDNA tagmentation at varying transposome concentrations and temperatures by gel electrophoresis (**Figure 2.1b**). This confirmed that hairpin-loaded Tn5 can effectively and tunably tagment DNA, with low temperature and low transposome concentrations favoring generation of fragments >1 kilobase (kb) in length.

We then tested 62 repair conditions (**Supplementary Table 2.1**) to close the 9 base-pair (bp) gap created by tagmentation[234] for productive PacBio sequencing. On the bases of percentage yield of DNA following exonuclease clean-up (**Supplementary Figure 2.1**) and fragment length estimated by analytical gel electrophoresis after tagmentation, repair, and exonuclease clean-up (**Supplementary Figure 2.2**) we found two enzyme combinations to be the most robust: Phusion polymerase and Taq DNA ligase ("Phusion/Taq") and T4 DNA polymerase and Ampligase ("T4/Ampligase"). These combinations yielded exonuclease-resistant libraries using as little as 50 ng of input gDNA, typically producing >20% total DNA yield (**Supplementary Table 2.2**). We used Phusion/Taq for gap repair in all subsequent experiments., as it provided significantly higher yields on high-quality commercial gDNA samples (p = 0.0093, two-sided t-test) in our hands.

To first evaluate the sequencing efficiency of SMRT-Tag libraries, we tagmented 120 ng of reference-grade HG002 gDNA (equivalent to ~20,000 cells) in 8 separate reactions (960 ng total), fractionated the resulting library into two length classes using paramagnetic solid-phase reversible immobilization (SPRI) beads, and sequenced using PacBio's proprietary 2.1 and 2.2 polymerases

optimized for short and long templates, respectively. We generated 3,524,301 molecules over both runs (14.3 Gb total). Fragment length distributions were concordant with size-selection and polymerase choice (**Figure 2.1c**), with shorter mean circular consensus sequence (CCS) lengths observed with 2.1 compared to 2.2 (2,081 ± 935.8 bp vs. 5,940 ± 3,097 bp, mean ± standard deviation [s.d.]). Visualizing the distribution of per-read quality-values (Q-score; **Figure 2.1d**), as well as the read length as a function of the number of individual sequencing passes per molecule (**Figure 2.1e**) demonstrated compatibility of these libraries with PacBio high-fidelity ("HiFi") sequencing, which generally requires at least 5 CCS passes per molecule to achieve greater than 99% (>Q20) base accuracy.

We next tested our ability to multiplex SMRT-Tag reactions. For all sequencing reactions in this study, we used one of eight individually loaded Tn5 transposomes, each harboring a unique 8 nucleotide (nt) barcode. To evaluate barcode demultiplexing, we first carried out a genotype-mixing experiment using high-molecular weight gDNA isolated from previously-genotyped HG002, HG003, or HG004 human samples (**Supplementary Figure 2.3a**). We tagmented samples individually (in total, seven 80 ng reactions), carried out gap-repair and exonuclease cleanup, pooled resulting tagged products, and then sequenced libraries to low-depth (HG002: 0.75X; HG003: 1.39X; HG004: 1.30X). We employed two separate metrics to ascertain barcode fidelity: first, we inspected the "left" and "right" barcodes of all sequenced molecules, which were overwhelmingly identical for all barcoded samples (**Supplementary Figure 2.3b**; 99.9% molecules sequenced with matched barcodes). Second, we assessed sample genotype mixing, with the expectation that HG003 and HG004 should cleanly separate on private genotypes, while HG002 (progeny of HG003 / HG004) should represent a mixture of both genotypes (**Supplementary Figure 2.3c**). As expected, while HG002 shared a moderate level (33.1%) of genotype information with HG003 and HG004, the parental samples had minimal overlap of private SNVs (0.60% HG003 vs. HG004; 0.67% HG004 vs. HG003). This demonstrates that transposomes do not tag previously-transposed templates following gap-repair, exonuclease cleanup, and pooling in our protocol.

We speculated that gap-repair could be performed in a single pool of multiplexed SMRT-Tag reactions (**Supplementary Figure 2.3d**). To test this, we performed four separate tagmentation reactions

on commercially available high-molecular-weight gDNA, pooled tagmentation reactions together, and carried out a single gap-repair and exonuclease treatment prior to sequencing. As in the prior experiment, "left" and "right" barcodes of all sequenced molecules were overwhelmingly identical for all barcoded samples (**Supplementary Figure 2.3e**; 99.9% molecules sequenced with matched barcodes). Furthermore, barcode concordance as measured by PacBio's proprietary demultiplexing software *lima* was very high (mean ± s.d. for *lima* quality scores: 97.9 ± 6.78; **Supplementary Figure 2.3f**). Taken together, these results establish the ability to accurately parallelize and demultiplex SMRT-Tag reactions.

We note that for almost all subsequent described experiments, unless explicitly noted, multiple SMRT-Tag reactions were multiplexed and pooled together on individual flow cells to minimize cost-per-sequenced-base; our rationale for this, and design choices for library prep, polymerase binding, and sequencer loading steps on the PacBio platform are discussed in Chapter 2.11 – **Supplementary Notes**, with detailed information on library quality control for all sequenced and tested libraries incorporated in Chapter 2.13 – **Supplementary File 1.**

Finally, to illustrate the tunability of the SMRT-Tag approach in sequenced libraries, we multiplexed SMRT-Tag reactions on high-molecular weight human gDNA while varying both Tn5 concentration and reaction temperature, and performed sequencing on a single flow cell. Visual examination of resulting read length distributions demonstrated the extent of tunability with the SMRT-Tag approach, as both Tn5-DNA ratio and temperature could be varied to shift library size distributions (**Supplementary Figure 2.4a**). Quantification of these distributions revealed a 1.92-fold dynamic range in mean fragment length, and a 2.30-fold dynamic range in fragment length standard deviation, offering an important reference point for implementing the approach (**Supplementary Figure 2.4b**). Together, these sequencing results demonstrate that SMRT-Tag generates tunable PacBio sequencing libraries from low amounts of input material.

2.4. SMRT-Tag accurately ascertains genomic and epigenomic variation in low-input settings

We next sought to establish the sensitivity and variant-calling accuracy of a single low-input SMRT-Tag reaction. We generated one SMRT-Tag library from 40 ng HG002 gDNA (~7,000 human cell equivalents) and loaded this at the maximum possible on-plate loading concentration (OPLC; **Figure 2.2a**; Chapter 2.10 – **Methods** and Chapter 2.12 – **Supplementary Note**). From a single flow cell, we generated 2.74M circular consensus sequencing (CCS) reads, with a median fragment length of 2.32 kb, equivalent to ~2.43X coverage of the HG002 genome (**Figure 2.2b**). We then evaluated our ability to call variants from this low-input experiment (**Figure 2.2c-e**), using DeepVariant to call single nucleotide variants (SNVs) and small insertions / deletions (indels). Comparing SNV and indel calls from this SMRT-Tag experiment against coverage-matched data from Genome in a Bottle (GIAB), we observed quantitatively similar recall (0.420 vs. for 0.527 for SNVs and 0.338 vs. 0.408 for indels), precision (0.870 vs. 0.898 for SNVs and 0.785 vs. 0.797 for indels), and F1 score (0.566 vs. 0.664 for SNVs and 0.380 vs. 0.539 for indels; **Figure 2.2c**). We observed lower performance on structural variants (SVs; recall 0.129 vs. 0.25, precision 0.877 vs. 0.879, and F1 score 0.225 vs. 0.389; **Figure 2.2d**), but note that this decreased performance derived largely from large insertions, which are inherently more difficult to genotype given the shorter mean fragment lengths of SMRT-Tag libraries. Together, these experiments establish the value of SMRT-Tag in maximizing high-accuracy PacBio sequencing coverage of low-input samples.

Sequencing native DNA on third generation sequencers offers the unique opportunity for simultaneous genotyping and epigenotyping (*i.e.* calling CpG methylation)[252]. To assess whether our low-input SMRT-Tag data effectively captured HG002 CpG methylation, we ran PacBio's *primrose* software, which uses a convolutional neural network to predict CpG modification based on real-time sequencing polymerase kinetics. We then compared genome-wide methylation estimates against publicly available gold-standard bisulfite sequencing data[209] and against GIAB PacBio data. We observed very high correlations between per-CpG methylation calls between our 2.43X SMRT-Tag dataset and benchmark bisulfite-based m^5dC estimates (Pearson's $r = 0.84$; **Figure 2.2e**). Framing CpG methylation detection as

a classification problem (**Figure 2.2f**), we also observed high performance as measured by area-under-curve (AUC), with SMRT-Tag and GIAB data demonstrating similar AUC (0.935 vs. 0.926, respectively).

Finally, to show that SMRT-Tag and standard PacBio sequencing perform similarly at higher coverage, we generated additional SMRT-Tag data from HG002 DNA to achieve a median coverage of 11.2X (34.24 Gb generated over 6 Sequel II flow cells). Comparing SNV, indel, and SV calls from SMRT-Tag against coverage-matched data from GIAB, we observed quantitatively similar recall (0.970 vs. for 0.970 for SNVs and 0.911 vs. 0.907 for indels), precision (0.995 vs. 0.995 for SNVs and 0.955 vs. 0.949 for indels), F1 score (0.983 vs. 0.982 for SNVs and 0.932 vs. 0.928 for indels), and AUC (0.969 vs. 0.968 for SNVs and 0.902 vs. 0.897 for indels; **Supplementary Figure 2.5a-d**). We also observed highly concordant performance for CpG methylation calling for higher-coverage SMRT-Tag data, both compared to bisulfite data (**Supplementary Figure 2.5e**), and against GIAB PacBio HiFi data (**Supplementary Figure 2.5f**). Importantly, SMRT-Tag also performed well in challenging genomic regions (*e.g.,* segmental duplications, tandem repeats, homopolymers, and the MHC locus; **Supplementary Figure 2.6a**), with SMRT-Tag slightly outperforming coverage-matched GIAB in select cases, likely reflecting improvements in sequencing chemistry (F1 scores: 0.977 vs. 0.967 for SNVs and 0.912 vs. 0.905 for indels across all challenging regions). Similarities between SMRT-Tag and GIAB variant-calling performance also did not vary with respect to coverage (**Supplementary Figure 2.6b**). Together, these analyses demonstrate the strong technical concordance between SMRT-Tag and existing PacBio library preparation methods.

2.5. Mapping single-fiber chromatin accessibility and CpG methylation with SAMOSA-Tag

Tn5-tagmentation of intact nuclei is the basis for ATAC-seq, a popular method for quickly and reproducibly profiling bulk chromatin accessibility genome-wide[158]. To adapt SMRT-Tag to analogously assay single-molecule chromatin accessibility[199] (following PacBio-based sequencing methods developed by our group[199,201] and others[198,241]), we developed and optimized a **tag**mentation-assisted **s**ingle-molecule **a**denine **m**ethylated **o**ligonucleosome **s**equencing **a**ssay (SAMOSA-Tag; **Figure 2.3a**). In SAMOSA-Tag,

nuclei are methylated *in situ* using the EcoGII m^6dAse, tagmented using hairpin-loaded Tn5 under conditions optimized for ATAC-seq[253], gap-repaired following DNA purification, and then sequenced on the PacBio Sequel II. As proof-of-concept, we applied SAMOSA-Tag to 50,000 nuclei from *MYC*-amplified OS152 human osteosarcoma cells[254], and used a convolutional neural network hidden Markov model (CNN-HMM)[201] to call inaccessible protein-DNA interaction "footprints" from m^6dA modifications natively detected by the sequencer. In total, across eight replicates, we sequenced 3,640,652 single molecules (7.79 Gb). Consistent with transposition of chromatin in nuclei, SAMOSA-Tag CCS length distributions displayed a characteristic oligonucleosomal banding pattern at shorter lengths (**Figure 2.3b**). When aligned to 5' read ends, SAMOSA-Tag molecules further displayed periodic accessibility signal, consistent with Tn5 transposition adjacent to nucleosomal barriers (**Figure 2.3c**). Sizes of individual footprints corresponded with expected sizes of mono-, di-, tri-, etc. nucleosomes (**Figure 2.3d**). Finally, single-fiber accessibility patterns could be visualized in the context of the genome, for instance at the amplified *MYC* locus (**Figure 2.3e**), and correlated well with ATAC-seq data from the same cell line (**Figure 2.3e**, **Supplementary Figure 2.7;** examples of copy-number loss and copy-number neutral loci in **Supplementary Figure 2.8**).

Importantly, unlike ATAC-seq data, SAMOSA-Tag insertions were only mildly biased toward annotated transcription start sites (TSSs; **Supplementary Figure 2.9a**); insertions did, however, preferentially occur in the vicinity of predicted CCCTC-binding factor (CTCF) binding sites (**Supplementary Figure 2.9b**), consistent with blocked Tn5 transposition by strong barrier elements. This slight insertion preference was also reflected in the overall fraction of insertions falling within TSSs and around CTCF binding sites (**Supplementary Figure 2.9c**; 1.51-fold enrichment above background for TSS; 1.58-fold enrichment above background for CBS), and was consistent with previously reported biases for Tn5-mediated shotgun Illumina sequencing[31]. Finally, SAMOSA-Tag generalized well to mouse embryonic stem cells (mESCs; **Supplementary Figure 2.10**), with SAMOSA-Tag signal demonstrating characteristic "footprint" patterns around predicted Ctcf and Rest binding sites

(**Supplementary Figure 2.10a,b;** left), that could themselves be clustered into distinct accessibility patterns (**Supplementary Figure 2.10a,b**; right).

2.6. Integrative measurement of CpG methylation and single-molecule chromatin accessibility

We speculated that separation of SAMOSA-Tag polymerase kinetics into separate m^6dA and m^5dC channels would enable simultaneous readout of DNA sequence, CpG methylation, and chromatin fiber accessibility. We first examined accessibility and CpG methylation signal surrounding predicted CTCF binding sites derived from ChIP-seq in the U2OS osteosarcoma cell line. Averaged accessibility and CpG methylation signals in 750 nt windows centered at predicted CTCF motifs revealed characteristic hallmarks of CTCF binding, including positioned nucleosomes flanking the motif, decreased fiber accessibility immediately at the motif (consistent with exclusion of EcoGII by fiber-bound CTCF), and depressed CpG methylation within motifs (**Figure 2.4a**). To move past this signal average, we used unbiased Leiden clustering[255] to examine the different fiber structures that make up this pattern (example of 4 clusters shown in **Figure 2.4b;** cluster sizes shown in **Supplementary Figure 2.11**). Analysis of average CpG methylation associated with each fiber structural pattern (**Figure 2.4c)** revealed lowest CpG methylation in clusters displaying direct evidence of CTCF fiber binding (cluster 1; minimum unsmoothed m^5dC/C of 0.14) and motif accessibility without bound CTCF (cluster 2), consistent with prior results[256]. Two additional analyses confirmed minimal confounding of m^5dCpG and m^6dA methylation signals: i.) *primrose* score distributions between negative control (*i.e.* SAMOSA-Tag experiments where EcoGII methylation was omitted) and footprinted samples were concordant (**Supplementary Figure 2.12a**), and ii.) average CpG methylation signal surrounding predicted CTCF sites on fibers without detectable accessibility was tightly correlated with signal from fibers with observed footprints (**Supplementary Figure 2.12b**). These experiments illustrate how the inherent multimodality of SMS can enable joint assessment of protein-DNA interactions, epigenetic modifications, and DNA sequence in a single experiment.

In prior work, we demonstrated that single-fiber accessibility data could be used to cluster the genome based on nucleosome regularity and average distance between regular nucleosomes (nucleosome-repeat length, or NRL)[199,201]. These studies relied on complementary epigenomic datasets to assess how the distribution of so-called "fiber-types" (*i.e.,* collections of fibers with unique regularity or NRL) might differ across euchromatic and heterochromatic domains. We sought to improve on these analyses by directly assessing how fiber structure varies as a function of jointly-measured single-molecule CpG content and CpG methylation. To do so, we assessed the distribution of single-molecule CpG densities, and average *primrose* methylation scores for each sequenced SAMOSA-Tag molecule in our dataset (**Figure 2.4d**). We then sectored these molecules into four different bins, gated on CpG density (> 10 CpG dinucleotides per kilobase), and *primrose* score (average *primrose* score > 0.5). We then (as previously[199,201]; Chapter 2.10 – **Methods**) computed single-molecule autocorrelograms for each sequenced molecule at least 1 kb in length, and clustered autocorrelograms to define fiber types. Following filtering of artifactual molecules, we obtained 7 distinct clusters (**Figure 2.4e**; cluster sizes in **Supplementary Figure 2.13**), which effectively stratified the OS152 genome by NRL (clusters NRL178 – NRL208) and fiber regularity (cluster IR). Finally, using the methylation / CpG content bins, we carried out a series of enrichment tests to assess how these fibers were differentially distributed across high / low CpG content and predicted CpG methylation (**Figure 2.4f;** reproducibility shown in **Supplementary Figure 2.14**). The resulting heatmap relates domain-specific changes in fiber composition as a function of single-molecule CpG state, and we highlight two findings that suggest relevance to chromatin regulation: first, we find that high CpG content / low CpG methylation (*i.e.* likely hypomethylated CpG islands) fibers are enriched for irregular fibers (odds ratio [O.R.] for cluster IR = 1.42; p ~0), as well as fibers with long NRLs (NRL208 O.R. = 1.09 / $p = 4.43 \times 10^{-64}$; NRL197 O.R. = 1.11 / $p = 1.49 \times 10^{-58}$); second, we find that high CpG content / high CpG methylation fibers (*i.e.* likely hypermethylated, CpG rich repetitive sequence) are enriched for irregular fibers (IR O.R. = 1.14 / $p = 1.33 \times 10^{-130}$), as well as short NRL fibers (NRL172 O.R. = 1.24; p ~0). Both results are broadly consistent with our previous *in vivo* SAMOSA observations of active promoters and heterochromatin in human K562 cells[199] and murine embryonic stem

cells (mESCs)[201], pointing to a conserved pattern of single fiber chromosome structure within these domains. Together, these analyses demonstrate that SAMOSA-Tag can easily generate genome-wide, multiomic single-molecule chromatin fiber accessibility data from tens of thousands of cells.

2.7. SAMOSA-Tag applied to patient-derived xenograft prostate tumor cells

One area where SAMOSA-Tag could have significant utility is in the study of clinical / pre-clinical disease models where samples are limited; namely, the study of patient-derived models of cancer progression (*e.g.*, patient-derived xenografted [PDX] mice). There are myriad challenges associated with profiling PDX-derived cells, particularly in a PCR-free setting: first, following tumor cell engraftment and growth, samples must be purified through fluorescence-activated cell sorting (FACS) to enrich for patient-derived tumor cells over host mouse tissue; second, PDX tumor cells derive from highly-necrotic tumors with increased likelihood of damaged native DNA and fragile cells and nuclei. We thus sought to apply SAMOSA-Tag to generate the first single-molecule chromatin accessibility datasets from a pair of primary and metastatic tumors, derived from the same patient diagnosed with castration-resistant prostate cancer (schematic in **Figure 2.5a**). We generated matched primary and metastatic prostate cancer PDX tumor models as previously described[257], and then isolated and methylated ~180,000 nuclei per model (1 mouse per model; FACS gates shown in **Supplementary Figure 2.15**). To account for the significant technical challenges working with mouse-derived primary patient tumor cells, while ensuring reproducibility of our findings, we performed six separate SAMOSA-Tag reactions (estimate 30,000 total input nuclei per reaction) to serve as replicates, which we sequenced on the PacBio Sequel II to a total coverage of 0.32X (0.95 Gb human data; 22.8% human alignment) for the primary PDX model and 0.53X (1.57 Gb human data; 95.9% human alignment) for the metastatic PDX model. We note that while lower-input SAMOSA-Tag reactions are technically feasible, given both preciousness of these samples and aforementioned technical challenges, we opted for a conservative experimental design. Importantly, primary and metastatic PDX SAMOSA-Tag data demonstrated similar technical characteristics to mESC and OS152 SAMOSA-Tag experiments (**Supplementary Figure 2.16**). We also note that future

optimizations (*e.g.* optimized human-mouse separation, DNA damage repair, nuclei purification, etc.) to the proof-of-concept presented here will likely allow for higher coverage (*i.e.* > 1.0X) of individual PDX samples using lower experimental input.

Altered CTCF expression and motif occupancy has been tied to both hyperactive androgen signaling[258] and prostate cancer progression[259]. Thus, we first examined differential single-molecule chromatin accessibility at predicted CTCF sites in primary and metastatic tumor cells (**Supplementary Figure 2.17a**). We aligned SAMOSA-Tag reads from both samples to CTCF sites predicted from ChIP-seq in ENCODE data from the LnCaP model prostate cancer cell line, and then performed clustering as above. As in both OS152 and mESCs, we observed multiple independent clusters (**Supplementary Figure 2.17b**); in these samples, these clusters reflected varying nucleosome occupancy patterns surrounding the core CTCF motif (NO1 – NO5), a cluster with direct evidence of CTCF occupancy (A), and a hyper-accessible cluster (HA) representing fibers devoid of nucleosomes in the vicinity of the CTCF motif. Visualizing the differential usage of these patterns through alluvial plots (**Supplementary Figure 2.17c**) revealed intriguing metastasis-specific shifts in cluster usage, including a decrease in the stereotypic "phased nucleosome / CTCF occupied" A pattern, and an increase in the HA pattern at these sites. Finally, these clusters could be directly associated with concurrently-measured CpG methylation (**Supplementary Figure 2.17d**), providing valuable preliminary insight into differences in CpG methylation state for these single-molecule CTCF motif occupancy states in primary (blue) and metastatic (red) cells.

Finally, we sought to determine whether single-molecule fiber types might differ between primary and metastatic tumor cells (**Supplementary Figure 2.18a**). We again performed unsupervised Leiden clustering of single-molecule autocorrelograms computed on SAMOSA signal. This clustering yielded six different fiber types, four regular clusters ranging in NRLs from 171 to 208 bp, and two irregular clusters (annotated IR1 and IR2; average SAMOSA signal of clusters shown in **Figure 2.5b**). Using previously-published epigenome annotations for healthy human prostate as a reference[260], we next determined the relative enrichment and depletion of fiber types across different human epigenomic domains, for each

sample type (**Supplementary Figure 2.18b**). Finally, we devised a logistic-regression based statistical test to quantify statistically-significant, reproducible differences in domain-specific fiber usage. Our test results reveal many potential patterns-of-interest for future follow-up (**Figure 2.5c**); for instance, metastatic PDX cells were significantly enriched for irregular fiber types IR1 and IR2 in annotated heterochromatic domains such as regions containing KRAB zinc-finger genes / repetitive sequence (label 12, ZNF / Rpt; IR1 $\log_2$ fold-change, or $\Delta = 0.77$, $q = 7.56\text{x}10^{-7}$; IR2 $\Delta = 1.03$, $q = 6.15\text{x}10^{-15}$), and mappable regions harboring marks of constitutive heterochromatin (*e.g.* label 13, Heterochromatin; IR1 $\Delta = 1.22$, $q = 1.45\text{x}10^{-177}$; IR2 $\Delta = 1.25$; $q = 4.46\text{x}10^{-125}$). Furthermore, regions annotated as various types of distal enhancer were significantly depleted for fiber types with specific NRLs (*e.g.* label 9, active enhancer 1; NRL182 $\Delta = -1.11$, $q = 1.07\text{x}10^{-71}$), hinting at potential involvement of ATP-dependent factors such as the Brahma-associated factor (BAF) complex, in evicting nucleosomes and disordering chromatin fibers. BAF has already been implicated as a key driver of prostate cancer progression[124], and while future studies must mechanistically dissect the preliminary model illustrated here (**Figure 2.5d**), our data demonstrate the potential of SAMOSA-Tag to provide mechanistic insight in challenging primary disease models.

2.8. Discussion & Conclusion

Here, we demonstrate direct transposition for sensitively preparing multiplexed, amplification-free PacBio sequencing libraries. We apply this principle to develop two related, single-molecule native DNA sequencing approaches.

Our first technique is SMRT-Tag, which extends the highly-accurate genomic variant and methyl-CpG detection of PacBio HiFi sequencing to very-low native DNA inputs. In this manuscript, we demonstrate that our optimized tagmentation and gap-repair conditions allow for sequencing > 7 Gb of HiFi-quality PacBio data from just 40 ng of input in a monoplex experiment. Pooling samples together to achieve even higher coverage, we demonstrate that SMRT-Tag is virtually indistinguishable in quality from gold-standard PacBio data with respect to genomic variant detection. We further show that SMRT-

Tag detects CpG methylation with performance comparable to both bisulfite sequencing data and previously released gold-standard PacBio data. In summary, combining tagmentation with optimized gap repair allowed the streamlined creation of PacBio libraries from 40 – 100 ng DNA (a minimum of ~7,000 human cell equivalents) compared to current protocols that require > 0.5 – 5 μg DNA (a minimum of ~200,000 human cell equivalents). We anticipate that this reduction in input requirement (conservatively, a 95-99% reduction in required input, see **Supplementary Note** for calculation) will remove a major obstacle to routine PacBio sequencing, and empower basic and translational studies of rare cell populations.

Our second technique is SAMOSA-Tag, which addresses a need for functional genomic methods capable of leveraging the breadth of nucleotide, structural, and epigenomic variation captured by SMS. Inspired by the ATAC-seq assay, SAMOSA-Tag offers a straightforward, scalable method to rapidly profile single-molecule chromatin accessibility without DNA purification. While SAMOSA-Tag does harbor slight insertional biases that ultimately impact genomic coverage uniformity (**Supplementary Figure 2.19**), we note that this can be seen as a desirable feature, particularly if one is already interested in biasing coverage of footprinting experiments to genomic features like CTCF or transcriptional start sites. We successfully constructed SAMOSA-Tag libraries from 30,000 – 50,000 nuclei, which we again note were multiplexed on individual flow cells to maximize sequencing yield. Our proof-of-concept focused on two of many possible SAMOSA-Tag applications: integrative epigenomic analysis of single-molecule CTCF binding, nucleosome architecture, and CpG methylation state in an osteosarcoma cell line, and the first ever single-molecule chromatin accessibility analyses of difficult-to-handle prostate cancer PDX primary and metastatic tumor cells. These applications demonstrate the potential of our approach for driving new epigenomic discoveries. Excitingly, our study also raises the possibility of single-cell resolution SMS approaches for profiling native chromatin and DNA. We envision the further development of SMRT-Tag to include droplet- or combinatorial barcoding-based cellular indexing[19,21,39]; such approaches would extend multimodal long-read analyses to the resolution of hundreds to thousands

of individual cells in parallel, enabling applications ranging from somatic variant detection, to *de novo* assembly, to cell type classification.

While SMRT-Tag and SAMOSA-Tag are powerful tools, both approaches have limitations. SMRT-Tag does not rely on PCR, so in many cases, end-users will likely multiplex SMRT-Tag reactions to maximize OPLC and cost-per-base on PacBio flow cells. Still, we establish here that Sequel II flow cells can be efficiently loaded with as little as 40 ng input. Further, owing to limited input, SMRT-Tag reactions are not readily compatible with pulsed field or other gel-based size-selection procedures; product size distributions in the SMRT-Tag reactions are primarily controlled by transposome concentration and bead-based cleanup protocols, and unsequenced DNA is effectively lost. This also likely sets a ceiling on the maximum amount of input DNA tagmentable by our approach, as the zero-turnover nature of Tn5[234] necessitates that enough Tn5 be used to generate an appropriate number of PacBio-sequenceable fragments. This limitation is particularly important for large-scale structural variant discovery, as the abundance of long, breakpoint-spanning CCS molecules is lower in SMRT-Tag libraries compared to gold-standard HiFi data. While we have partially addressed this by demonstrating tunability of our reactions, future work engineering transposases may enable even more control of library size. Similarly, our SAMOSA-Tag protocol is limited with respect to the minimal amount of nuclei that can be processed. In experiments presented here, we were able to generate high-quality data from 30,000 – 50,000 footprinted nuclei, across multiple pooled replicates. Future optimizations to the SAMOSA-Tag protocol, including light fixation, miniaturized methylation reactions, or immobilization of nuclei on activated beads[261] could further relax this constraint.

More generally, SMRT-Tag and SAMOSA-Tag add to a growing series of technological innovations centered around third-generation sequencing, including Cas9-targeted sequence capture[262], combinatorial-indexing-based plasmid reconstruction[143], and concatenation-based isoform-resolved transcriptomics[263]. The widespread adoption of short-read genomics in basic and clinical applications was catalyzed by the development of tools that democratized Illumina sequencing. Our approaches offer

similar promise for rapidly maturing third-generation sequencing technologies, through scalable, sensitive, and high-fidelity telomere-to-telomere genomics and epigenomics.

2.9. Figures

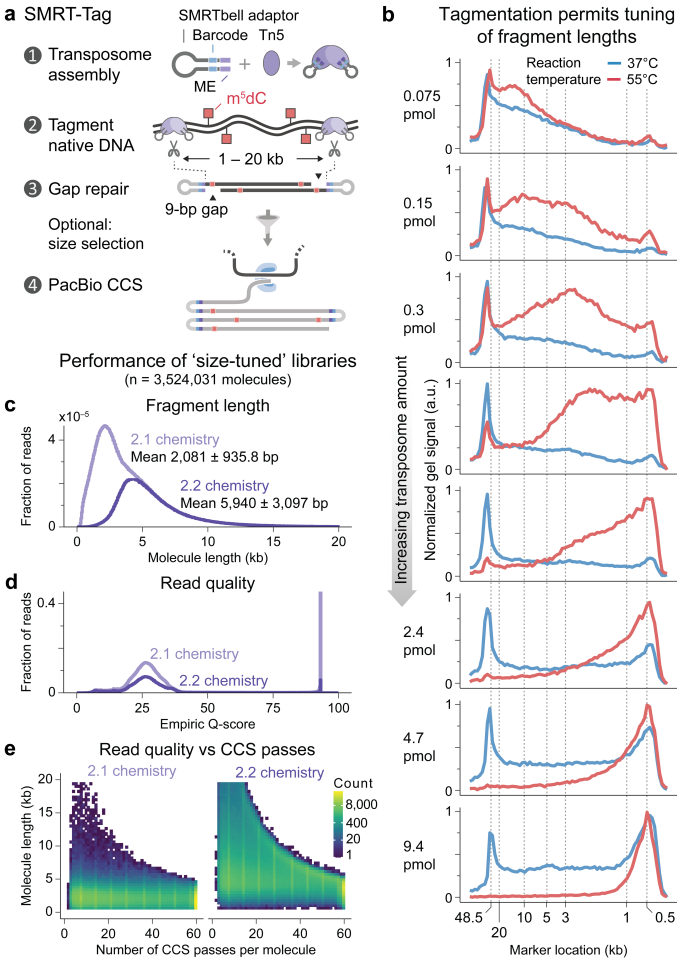


Figure 2.1: SMRT-Tag enables tunable, low-input single-molecule real time sequencing on the PacBio sequencing platform. a.) A schematic of the SMRT-Tag approach. Hairpin adaptor-loaded Tn5 transposase is used to fragment DNA into kilobase-scale fragments. After removing Tn5 transposase, an optimized gap repair reaction is used to fill the 9 bp gaps on either side of the molecule, and an exonuclease treatment is used to purify repaired covalently closed templates. **b.)** Transposomes can tunably fragment high-molecular weight gDNA by tuning reaction temperature and concentration. We targeted conditions that would reliably generate fragments from 2 – 10 kb in size. **c.)** Circular consensus sequencing (CCS) fragment lengths for two size-selected library preps, sequenced using size-appropriate PacBio polymerases (2.1 vs. 2.1). In light purple, shorter libraries, and in dark purple, a longer library. X-axis is capped at 20 kb, though 2.2 libraries exhibit a long-tailed distribution that extends past 20 kb. **d.)** Empirical quality score (Q-score) distributions for 2.1 and 2.2 libraries. **e.)** Heatmap representation of molecule CCS length as a function of number of passes per CCS molecule, with log scaled counts.

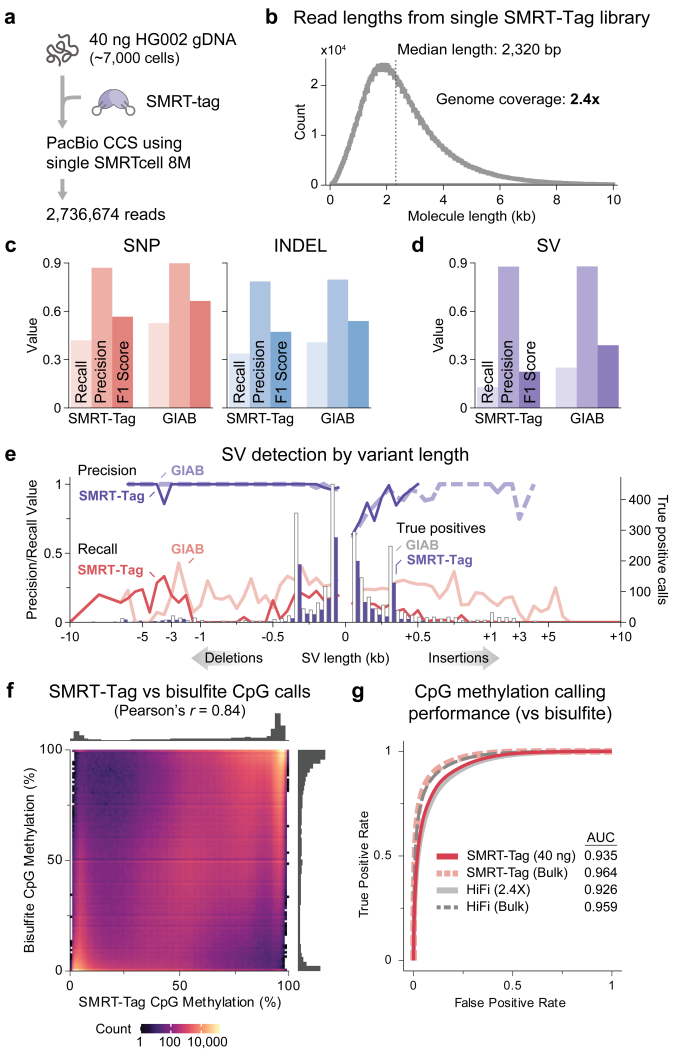


Figure 2.2: SMRT-Tag enables accurate genotyping and epigenotyping of low-input samples. a.) Schematic of monoplex SMRT-Tag experiment to establish ability to sequence low-input samples with maximal coverage on the Sequel II platform. We generated a single SMRT-Tag library using 40 ng gDNA (equivalent to ~7,000 human cells) from Genome in a Bottle (GIAB) reference individual HG002 and sequenced on a single PacBio flow cell **b.)** Read length distribution of monoplex SMRT-Tag library. **c.-d.)** Precision, recall, and F1 scores for DeepVariant single nucleotide polymorphism (SNP) and insertion / deletion (indel) calls (**c**) and *pbsv* structural variant (SV) calls (**d**) from single-plex SMRT-Tag compared to coverage-matched GIAB HG002 PacBio data. **e.)** Precision, recall, and number of true positive variant calls for binned SV sizes for monoplex SMRT-Tag and coverage-matched GIAB HG002 data. **f.)** Monoplex SMRT-Tag *primrose* CpG methylation estimates plotted against bisulfite CpG methylation for HG002. **g.)** Receiver operating characteristic (ROC) curves for HG002 CpG methylation detection using monoplex 40 ng SMRT-Tag, pooled SMRT-Tag from multiplexed libraries (not coverage matched), and GIAB PacBio HiFi data compared to gold-standard bisulfite sequencing.

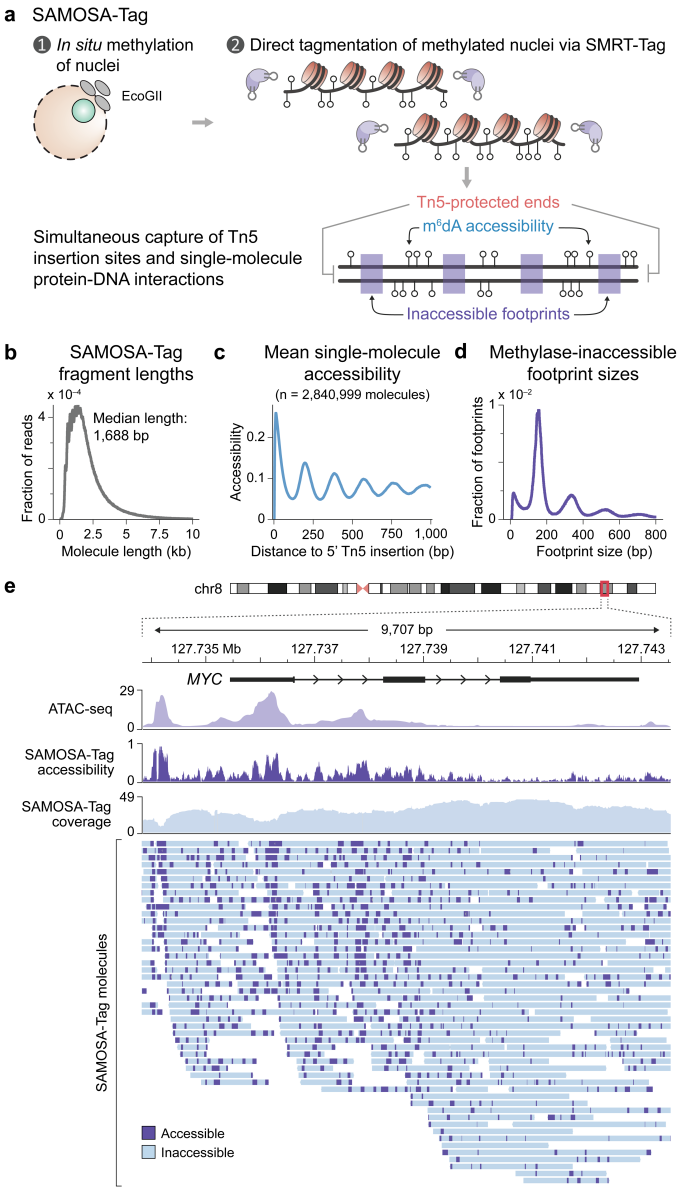


Figure 2.3: SMRT-Tag can be combined with the SAMOSA single-fiber footprinting assay to easily generate single-molecule chromatin accessibility data through direct tagmentation of adenine-methylated nuclei. a.) Schematic overview of the SAMOSA-Tag approach: nuclei are methylated using the nonspecific EcoGII m^{6}dAse and tagmented *in situ* using SMRT-Tag. DNA is purified, gap-repaired, and sequenced on the PacBio Sequel II, resulting in molecules where ends result from Tn5 transposition, m^{6}dA marks represent fiber accessibility, and computationally defined unmethylated footprints capture protein-DNA interaction. **b.)** Fragment length distributions for SAMOSA-Tag data from the OS152 osteosarcoma cell line. **c.)** Average methylation signal from the first 1000 nt of molecules from the same dataset as **b.). d.)** Unmethylated footprint size distribution for the same dataset. **e.)** Genome browser visualization of SAMOSA-Tag data at the amplified *MYC* locus. Purple marks predicted accessible bases, while blue represents predicted inaccessible bases on individual molecules. Average SAMOSA accessibility shown in purple; matched ATAC-seq track shown in blue.

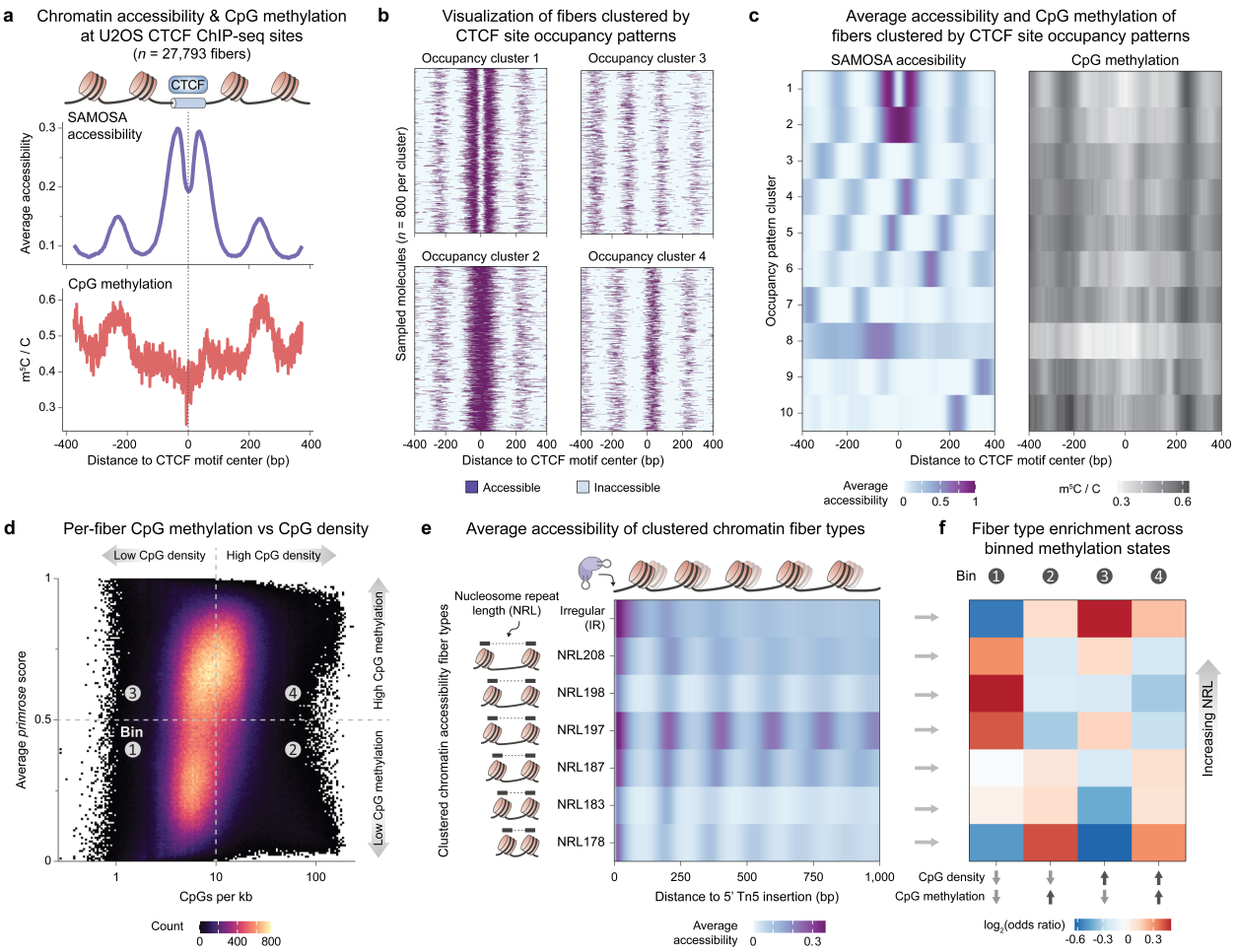


Figure 2.4: SAMOSA-Tag data can simultaneously ascertain CpG methylation state and chromatin accessibility at predicted CTCF binding sites, and can be used to study chromosome fiber structure on differentially CpG methylated fibers. a.) Average SAMOSA accessibility signal and CpG methylation on 27,793 footprinted fibers from OS152 cells, centered at predicted CTCF binding sites taken from published U2OS CTCF ChIP-seq data. **b.)** Molecular visualization of individual, clustered fibers (800 molecules per cluster), reflecting different CTCF-occupied, accessible, and inaccessible fiber states, centered at predicted CTCF binding motifs. **c.)** Simultaneous visualization of average accessibility (left) and CpG methylation (right) for each of 10 clustered accessibility states surrounding CTCF motifs. Window size is 750 nt for **a.) – c.). d.)** Average *primrose* score (methylation prediction) for individual fibers as a function of number of CpG dinucleotides per kilobase on individual fibers. We binned molecules into one of four bins, depending on both CpG density and average *primrose* score. **e.)** Average accessibility of 7 different fiber types determined by performing Leiden clustering on single-molecule autocorrelograms calculated from each footprinted chromatin fiber. Clusters broadly stratify the entire genome on the basis of NRL for regular fibers (ranging from 178 to 208 bp), or irregularity (cluster IR). **f.)** For the same clusters as in **e.)**, relative enrichment or depletion (calculated through Fisher's exact test) of individual fiber types in each of the four binned states from **d.).** All tests shown here are statistically significant (p ranges from ~ 0 to 2.41×10^{-5}).

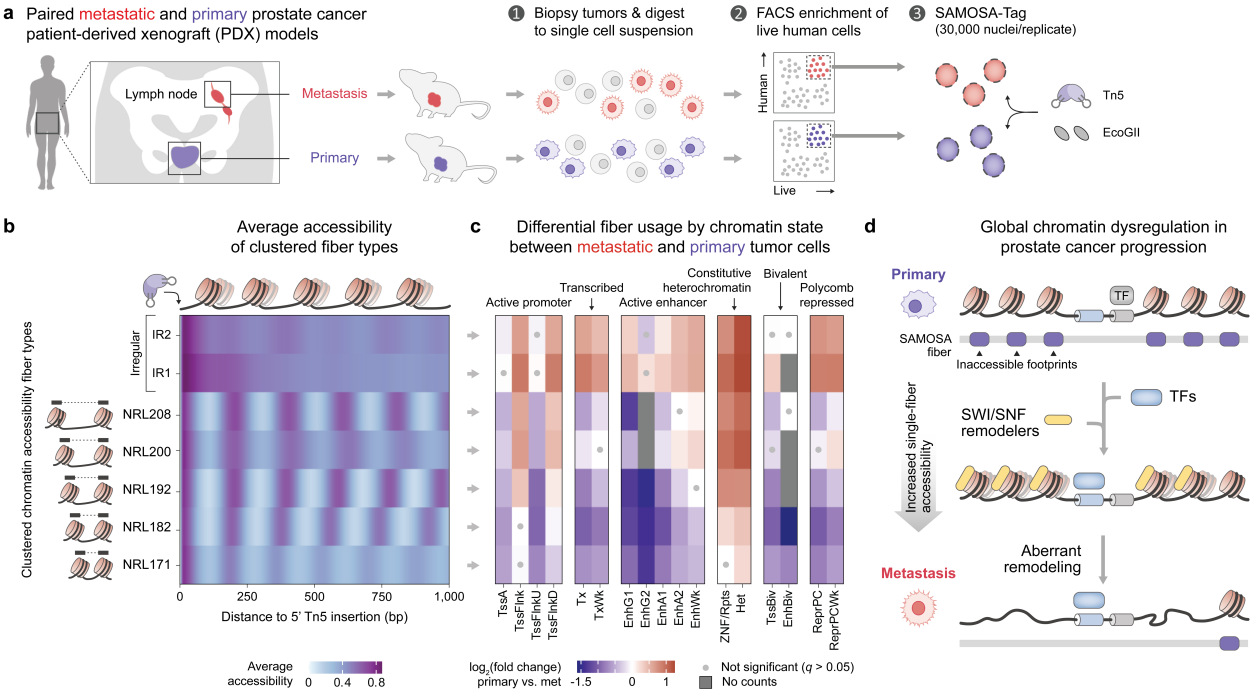
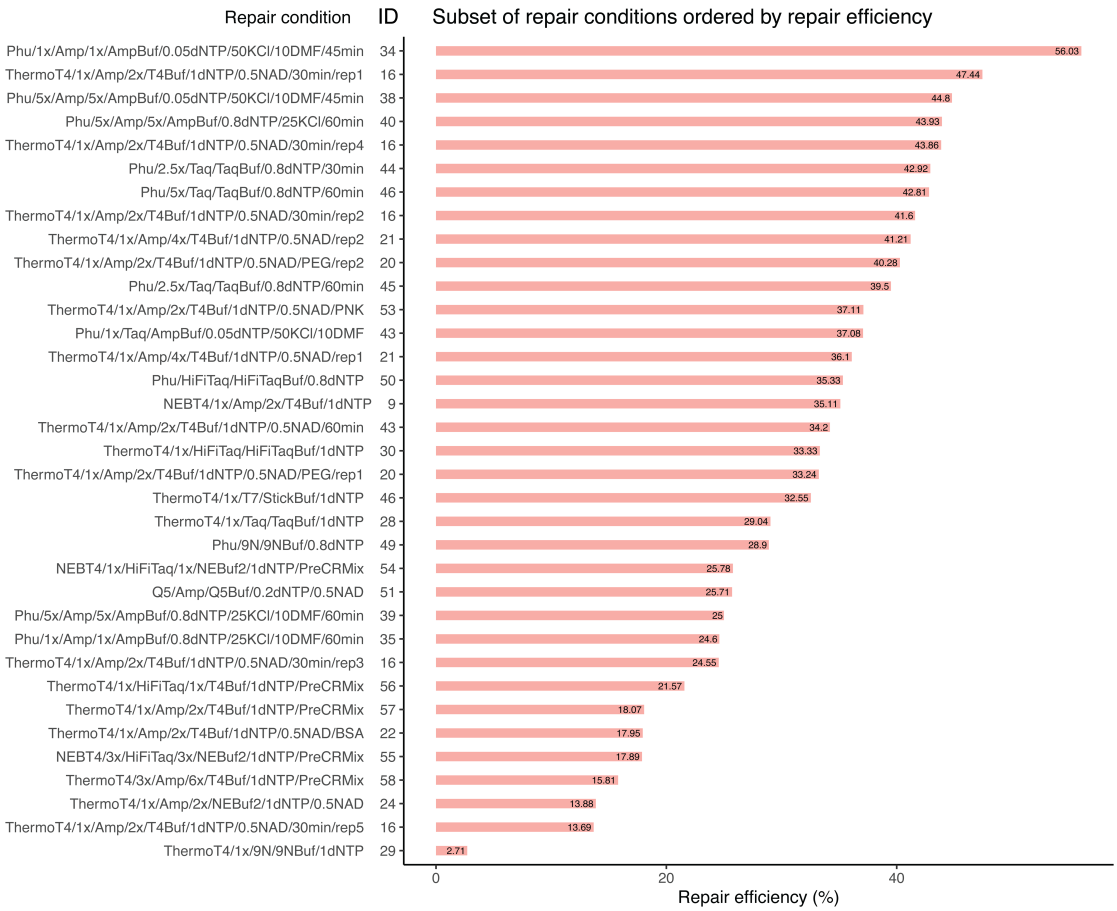
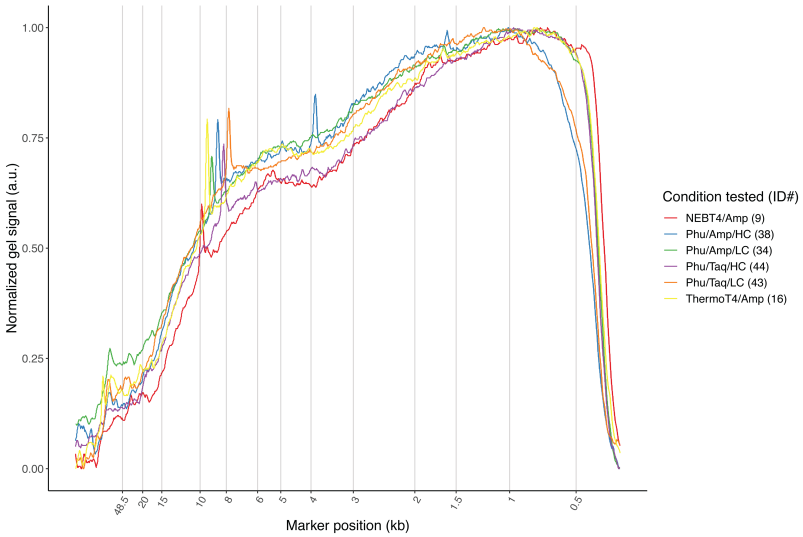


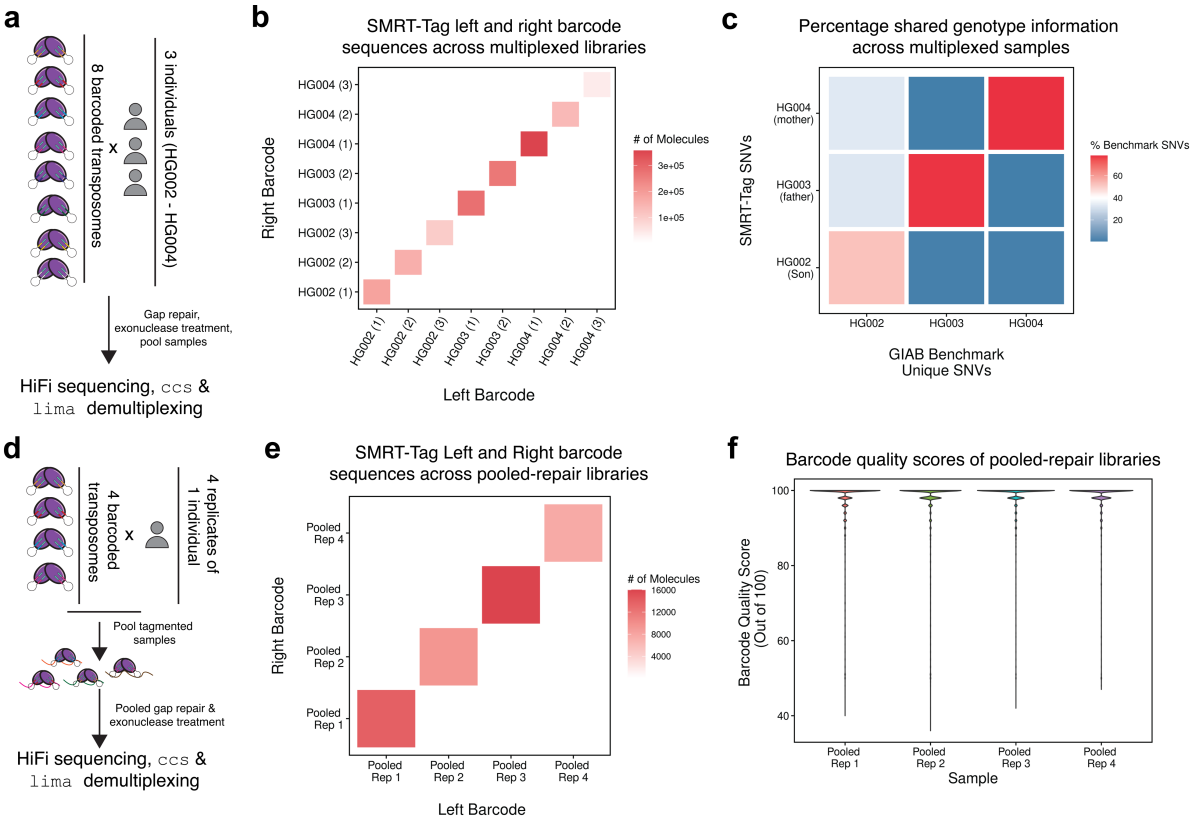
Figure 2.5: SAMOSA-Tag applied to patient-derived xenograft (PDX) models of primary and metastatic prostate cancer. a.) Schematic overview of our approach to performing SAMOSA-Tag on nuclei derived from primary and metastatic PDX mouse models derived from the same patient. PDX was established as previously described, after which tumor cells were biopsied, digested, and then sorted to enrich for live, human cells with FACS. We then performed six parallel SAMOSA-Tag reactions, using an estimated 30,000 nuclei per reaction. **b**.) Clustered fiber types resulting from Leiden clustering of footprinted primary and metastatic chromatin fibers falling in one of 17 different prostate-cancer-specific chromHMM states. Unsupervised Leiden clustering yielded 7 clusters – five regular clusters ranging in NRL from 171 to 208, and two irregular clusters. **c.)** Heatmap result of effect-size estimate from a logistical regression framework designed to call statistically significant differences in fiber type usage across each chromHMM state. Framework considers all six replicates from each of the two different sample types (primary and metastasis). Red indicates fiber types that are enriched in metastatic samples versus primary samples, and blue, vice-versa. Non-significant (N.S.) results marked as grey box or grey dot. **d.)** A preliminary model of single-molecule chromatin accessibility states measured by these SAMOSA-Tag experiments. At both CTCF sites and genome-wide, fibers in metastatic cells are overrepresented for highly accessible, irregular chromatin fibers devoid of phased nucleosomes. We speculate that this might signify deranged activity by SWI/SNF remodelers, which are prime candidates for generating such nucleosome-free / irregular single-molecule accessibility patterns. State legends: 1: TSS, 2: TSS Flank, 3: TSS Flank Upstream, 4: TSS Flank Downstream, 5: Transcribed region, 6: Weakly transcribed region, 7: Genic enhancer 1, 8: Genic enhancer 2, 9: Active enhancer 1, 10: Active enhancer 2, 11: Weak enhancer, 12: KRAB zinc finger / repetitive region, 13: Constitutive heterochromatin, 14: Bivalently-marked TSS, 15: Bivalently-marked enhancer, 16: Polycomb repressed, 17: Weakly polycomb repressed.



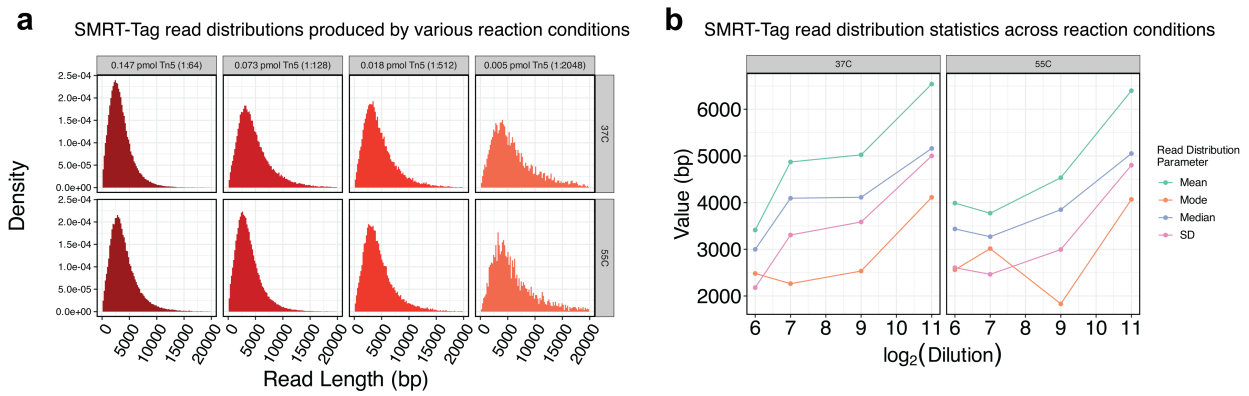
Supplementary Figure 2.1: Tabulated repair efficiency for a subset of the 62 unique conditions tested to optimize gap repair. Repair efficiency (defined as the % yield of final product compared to input DNA by mass following exonuclease treatment) for 35 of the 62 conditions tested. We ultimately selected a mixture of Phusion polymerase and Taq ligase for gap repair as these provided the most consistently high repair efficiency across multiple experiments.



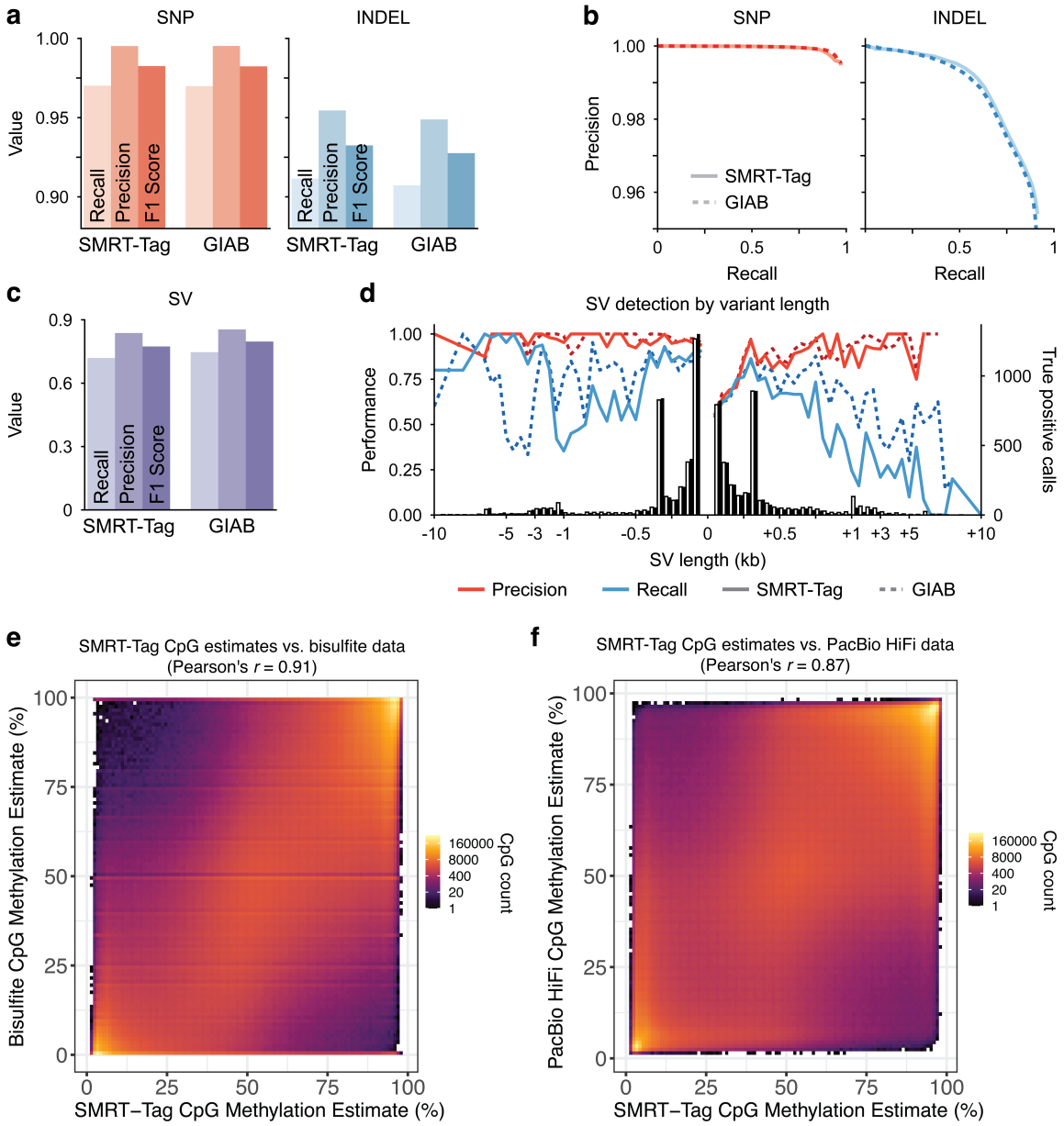
Supplementary Figure 2.2: Example analytical gel trace for validating the size distribution of gap-repaired products for a subset of conditions. In addition to repair efficiency, we also validated that gap repair conditions did not appreciably change the size distribution of resulting libraries by gel electrophoresis. Shown here are analytical gel traces for six specific conditions tested in this study, including Phusion / Taq in multiple buffers.



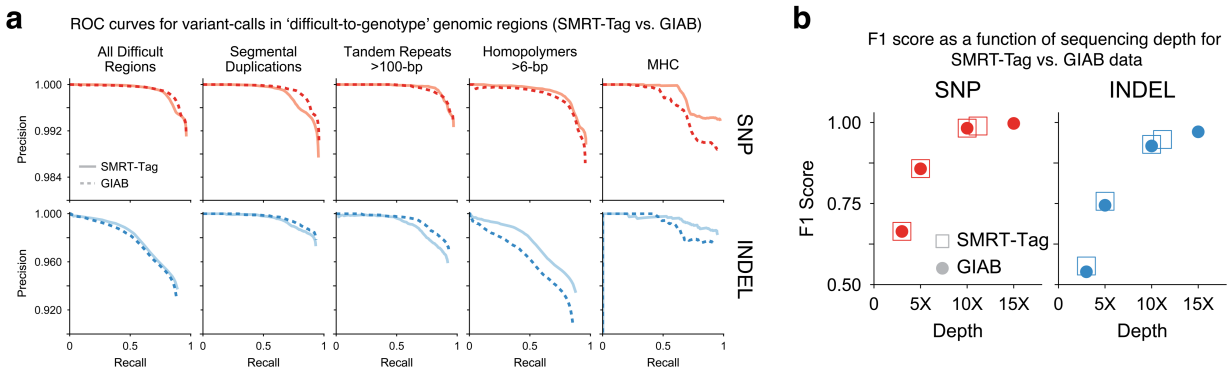
Supplementary Figure 2.3: Control experiments to establish multiplexing with SMRT-Tag. **a.)** Schematic overview of a genotype mixing experiment where gDNA samples from HG003, HG004, and their progeny HG002 are individually barcoded with one of 8 different uniquely-loaded transposomes, gap-repaired, exonuclease-cleaned, pooled, and sequenced on the PacBio Sequel II. **b.)** Heatmap representation of demultiplexing results from PacBio's proprietary *lima* barcode splitting software, which annotates molecules with matching barcodes, versus those with mixed barcodes; on diagonal signal demonstrates minimal cross-contamination across transposome barcodes / samples. **c.)** Percentage shared genotype information across barcoded samples. As expected HG002 harbors shared SNPs with HG003 and HG004, but HG003 and HG004 samples have minimal shared genotype overlap. For this analysis, all private SNVs across HG003 and HG004 were considered. **d.)** Experiment to validate that gap repair can be carried out in a pool without pervasive barcode hopping across molecules. We barcoded gDNA from one individual with one of four different barcoded transposomes, pooled, and then carried out pooled gap repair and exonuclease cleanup. **e.)** As in **b.)** but for pooled experiment. **f.)** Distributions of *lima* quality scores for barcoded molecules in the pool.



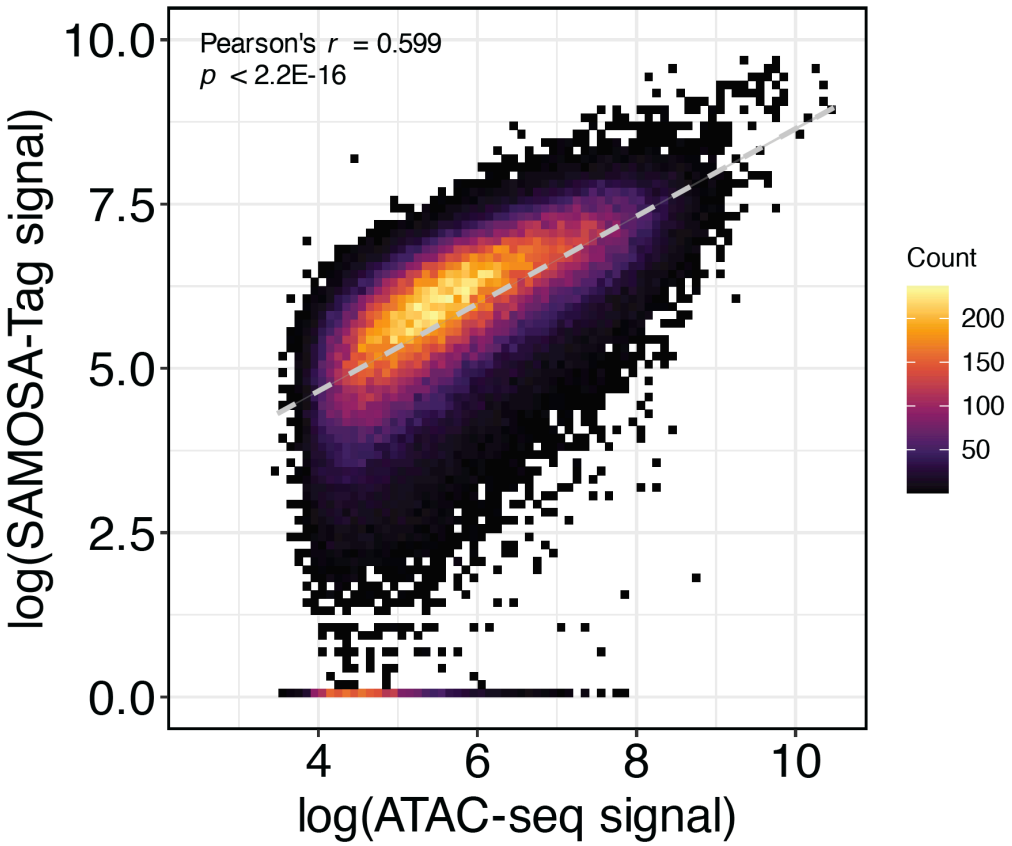
Supplementary Figure 2.4: Establishing the tunability of SMRT-Tag reactions by varying Tn5 concentration and temperature and sequencing resulting libraries. a.) CCS fragment length distributions for various SMRT-Tag libraries constructed by varying Tn5 concentration (columns) or reaction temperature (rows). **b.)** Quantification of mean, mode, median, and standard deviation (SD) for each sequenced library as a function of dilution factor.



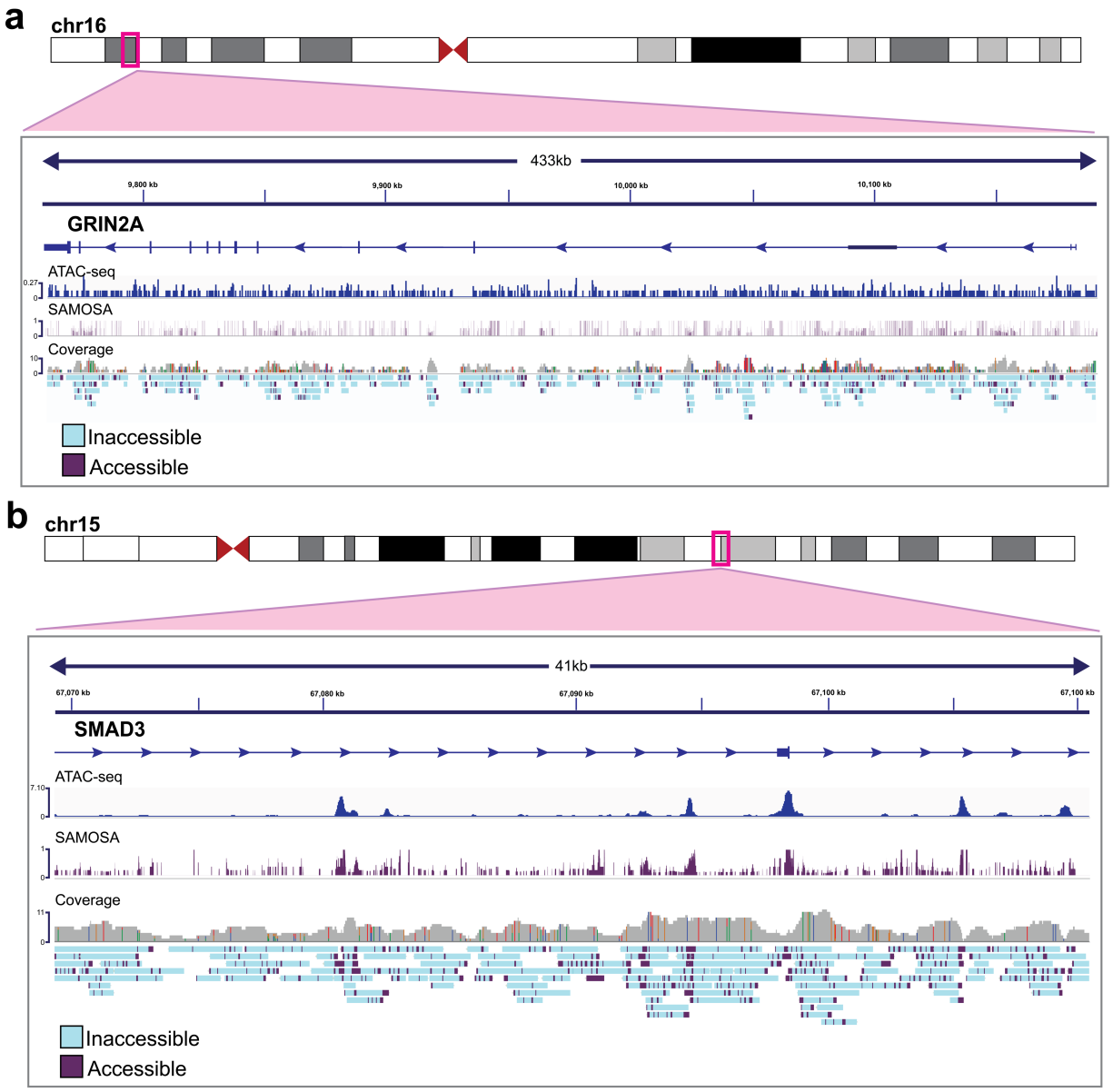
Supplementary Figure 2.5: Benchmarking SMRT-Tag genotype and epigenotype calls at higher coverage. a.) Precision, recall, and F1 scores for DeepVariant single nucleotide polymorphism (SNP) and insertion / deletion (indel) calls from high-coverage SMRT-Tag sequencing of HG002 compared to coverage-matched GIAB data. **b.)** Precision as a function of recall for SNPs and indels for SMRT-Tag versus GIAB data. **c.)** *pbsv* structural variant (SV) calls from high-coverage SMRT-Tag sequencing of HG002 compared to coverage-matched GIAB data. **d.)** Precision, recall, and number of true positive variant calls for binned SV sizes for high-coverage SMRT-Tag and versus GIAB HG002 data. **e.-f.)** Comparisons of *primrose* CpG methylation calls against gold-standard bisulfite data in **e.**) and GIAB HG002 PacBio HiFi data in **f.)** for high-coverage SMRT-Tag sequencing data.



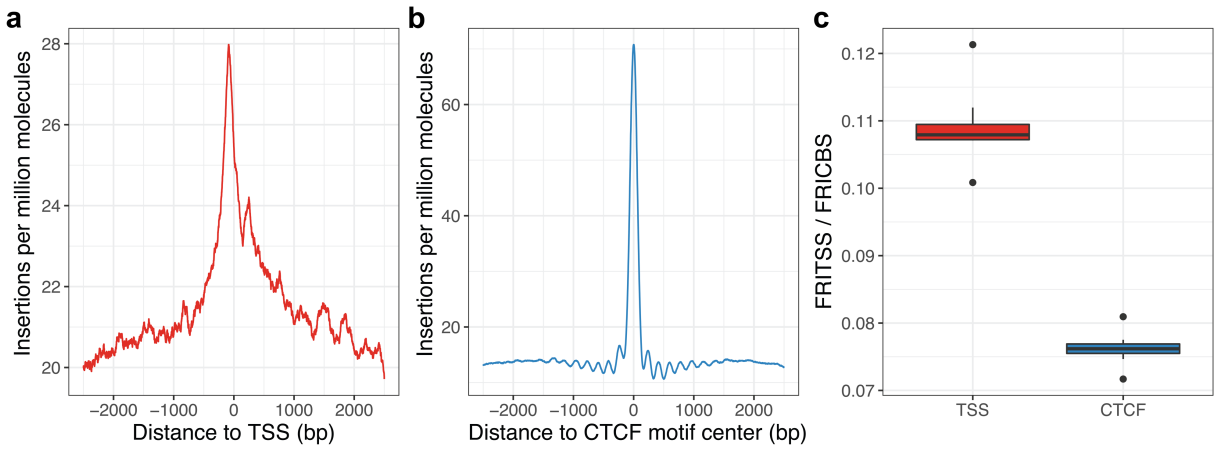
Supplementary Figure 2.6: Genotyping performance of SMRT-Tag data across difficult-to-genotype regions and as a function of sequencing depth. a.) DeepVariant precision / recall curves for SNP (red) and indel (blue) variant calls in challenging genomic regions, including segmental duplications, tandem repeats, homopolymers, and the MHC locus, for high-coverage SMRT-Tag data (solid) versus coverage-matched GIAB data (dashed). **b.)** Composite F1 score for SMRT-Tag (closed circles) versus GIAB data (open square) as a function of sequencing depth, for SNP (red) and indel (blue) variant calls.



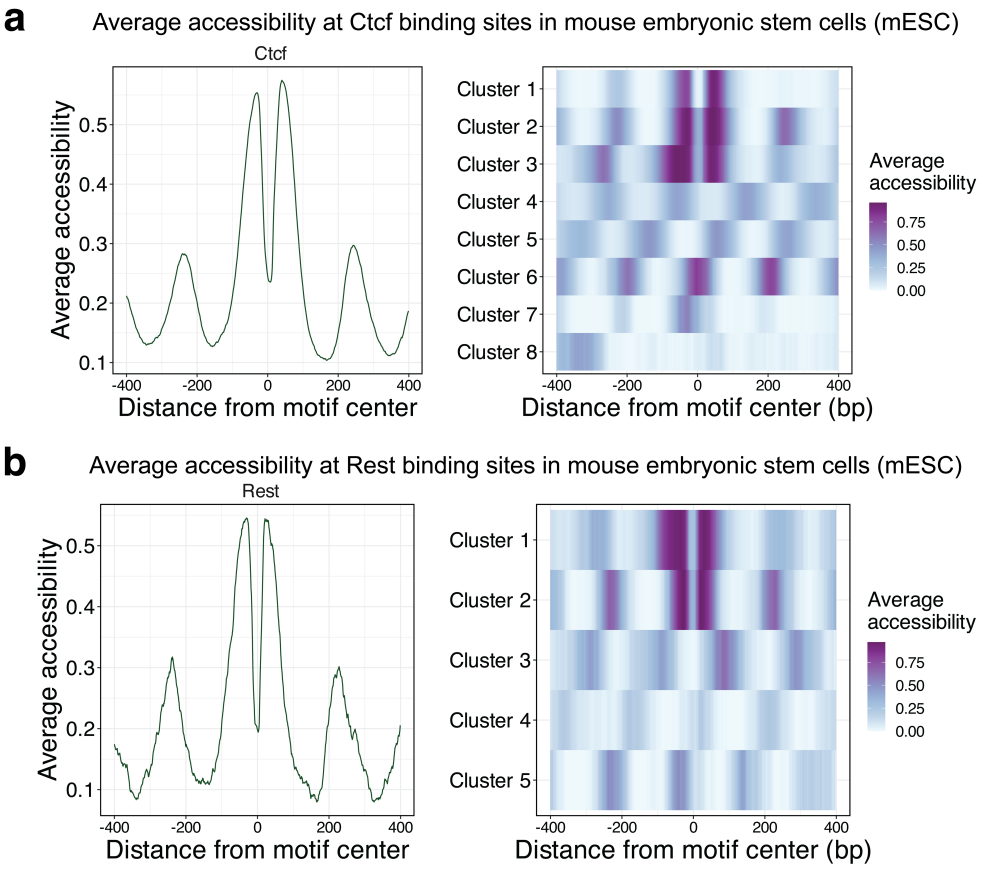
Supplementary Figure 2.7: Genome-wide correlation of OS152 SAMOSA-Tag accessibility measurements with ATAC-seq data. SAMOSA-Tag methyltransferase accessibility signal is significantly and positively correlated with ATAC-seq data.



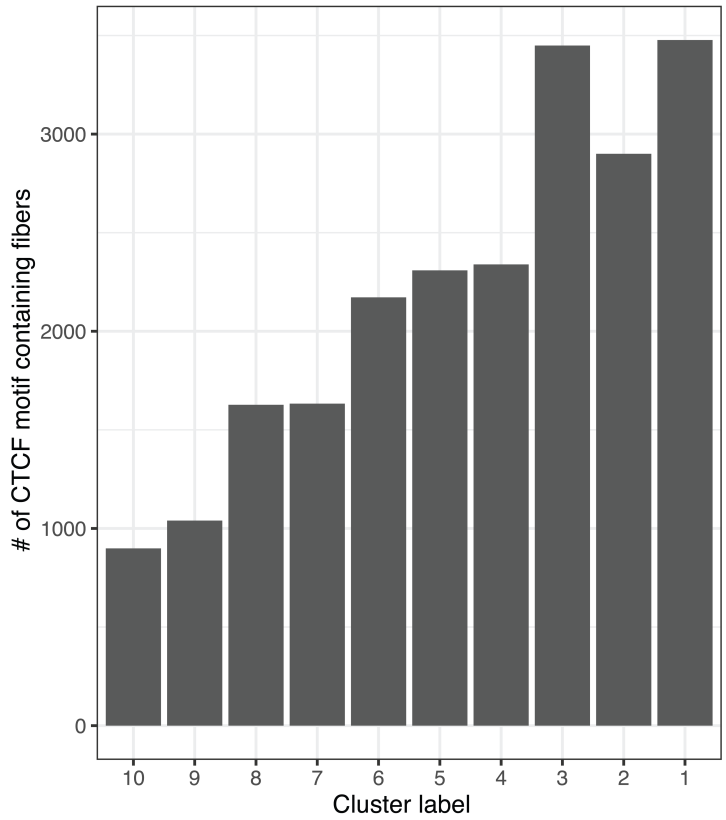
Supplementary Figure 2.8: Examples of SAMOSA-Tag coverage and signal co-plotted with ATAC-seq data for copy-number neutral (SMAD3) and copy-number loss (GRIN2A) genes.



Supplementary Figure 2.9: OS152 SAMOSA-Tag libraries demonstrate slight insertional bias at transcription start sites and CTCF motifs. a.) Metaplot of insertions per million sequenced molecules at hg38 transcriptional start sites (TSSs), in a 5 kb window centered at the TSS for OS152 SAMOSA-Tag libraries. Signal was smoothed using a 100 nt running mean. **b.)** Metaplot of insertions per million sequenced molecules at U2OS ChIP-seq backed CTCF binding sites, in a 5 kb window centered at the center of the CTCF motif. Signal was smoothed using a 100 nt running mean. **c.)** Boxplots of fraction of insertions in TSS (FRITSS) and fraction of insertions in CTCF binding sites (FRICBS) across all eight replicate experiments.

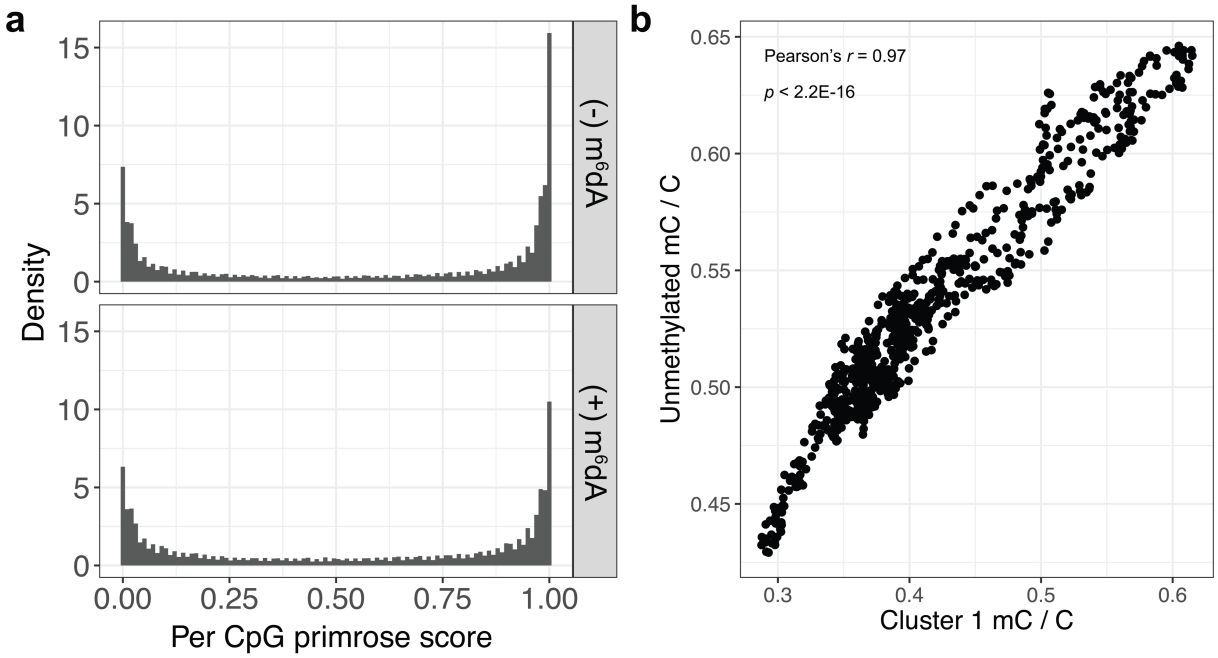


Supplementary Figure 2.10: SAMOSA-Tag generalizes to different cell types and can footprint TFs outside of CTCF / Ctcf. a.) SAMOSA-Tag was performed in mouse embryonic stem cells (mESCs), signal was extracted from fibers containing predicted Ctcf binding sites by ChIP-seq, and resulting molecules were clustered to yield 8 different single-molecule accessibility patterns around predicted Ctcf sites. **b.)** As in **(a)** but for the TF Nrsf / Rest (ChIP-seq data from Yu et al.[264]).

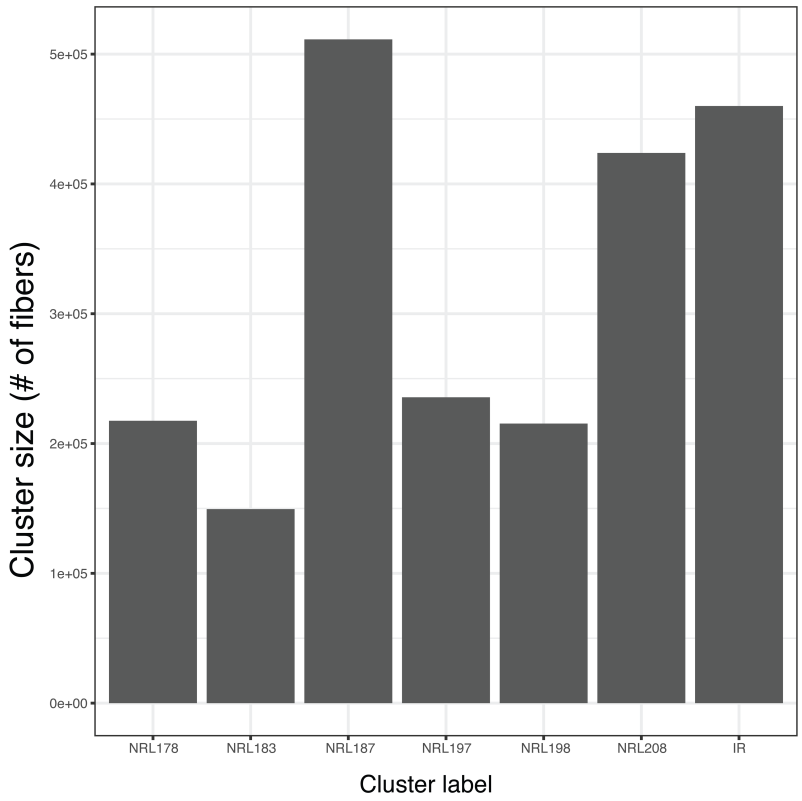


Supplementary Figure 2.11: Cluster sizes resulting from Leiden clustering of single-molecule accessibility patterns surrounding predicted CTCF sites. Cluster labels match the labels used in **Figure 2.4b,c**

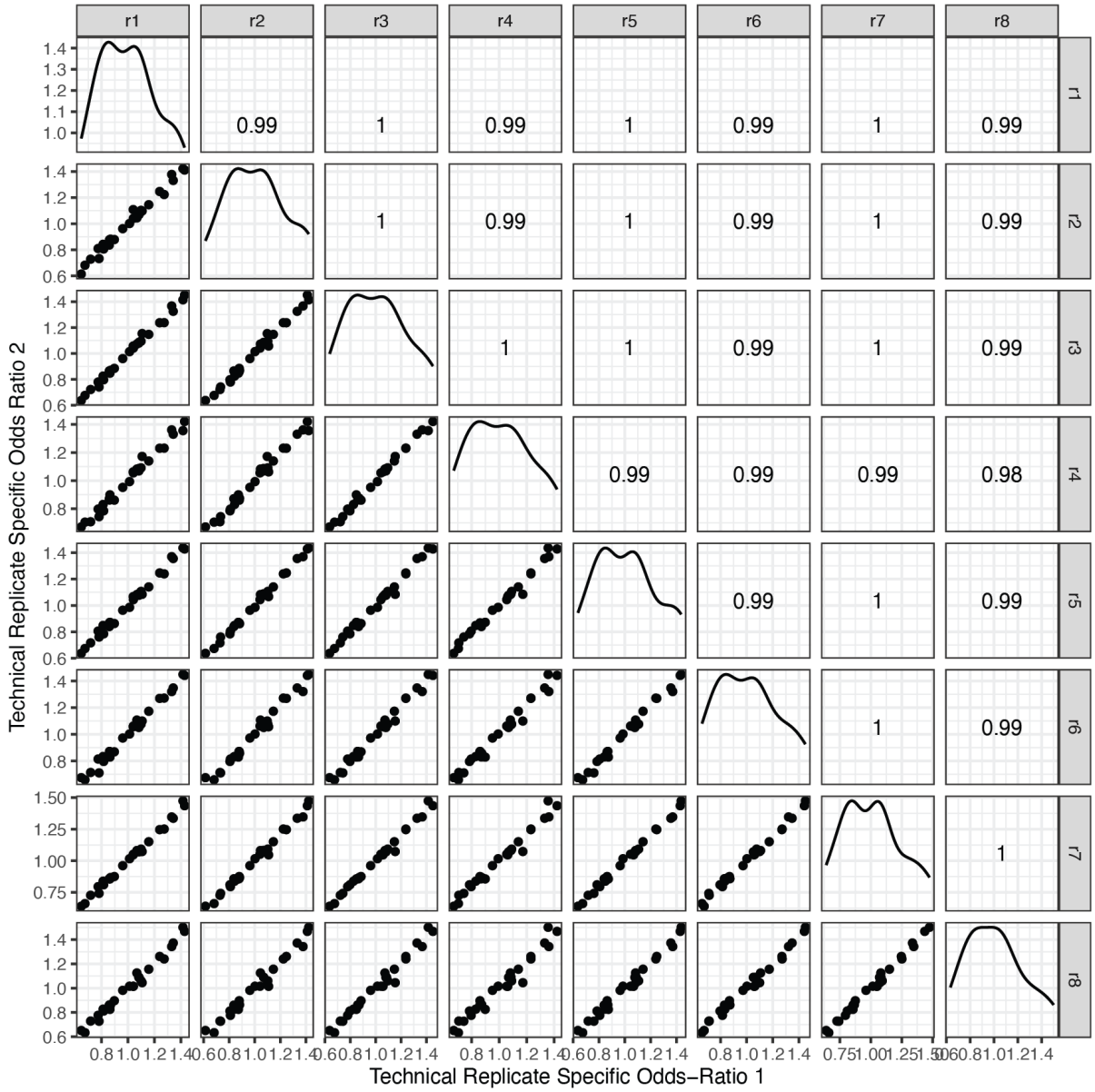
.



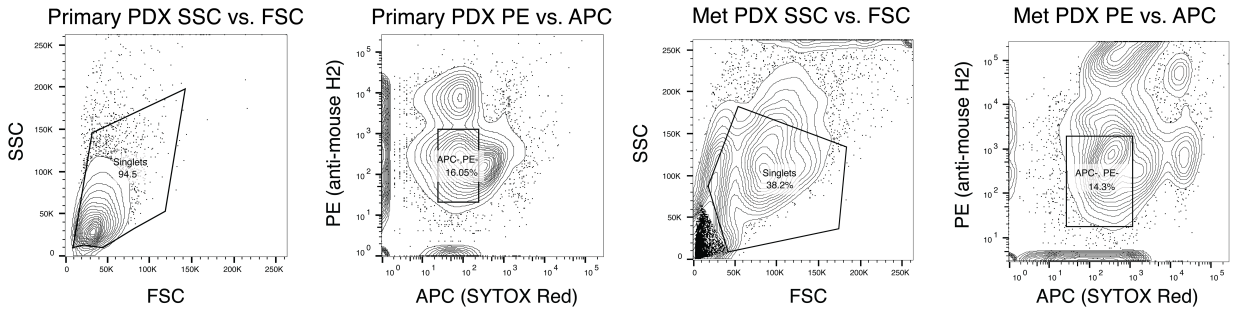
Supplementary Figure 2.12: m6dA footprinting does not appreciably impact primrose CpG methylation predictions. a.) Distribution of per CpG *primrose* scores (50,000 sampled CpGs per experiment) for SAMOSA-Tag control experiments where EcoGII was omitted (top) and SAMOSA-Tag experiments (bottom). **b**.) Correlation of averaged CpG methylation signal from SAMOSA-Tag molecules without any detectable m^6dA methylation surrounding predicted CTCF sites, versus CpG methylation from SAMOSA-Tag molecules from occupancy cluster 1. Signals were correlated with Pearson's *r* of 0.97 ($p < 2.2x10^{-16}$).



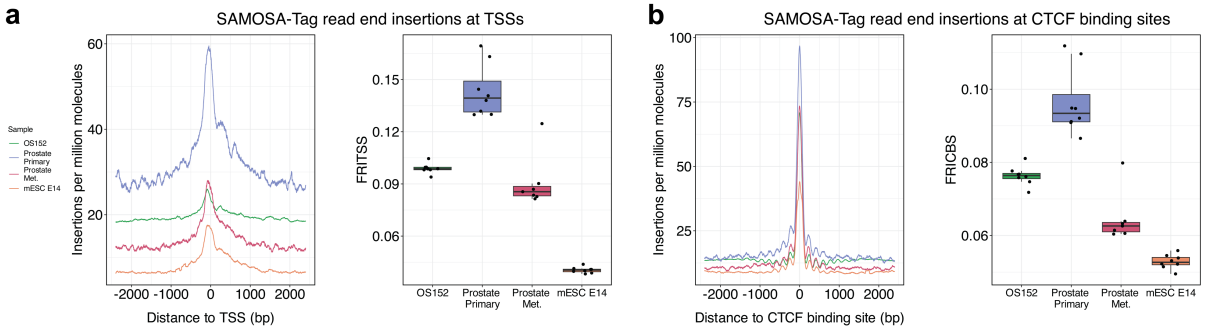
Supplementary Figure 2.13: Cluster sizes resulting from Leiden clustering of single-molecule autocorrelograms. Cluster labels match labels used in **Figure 2.4e,f**.



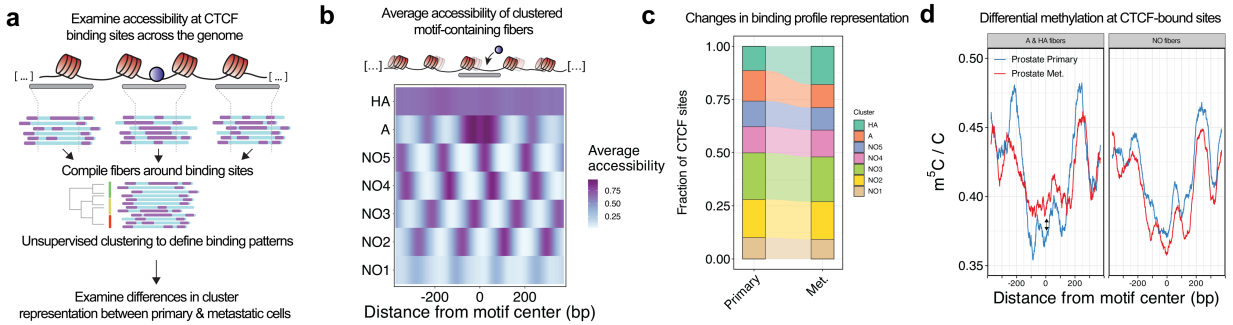
Supplementary Figure 2.14: SAMOSA-Tag fiber enrichments in different CpG content / CpG methylation bins are technically reproducible. Matrix of scatter plots plus Pearson's *r* correlation values across each of eight different replicate OS152 SAMOSA-Tag experiments.



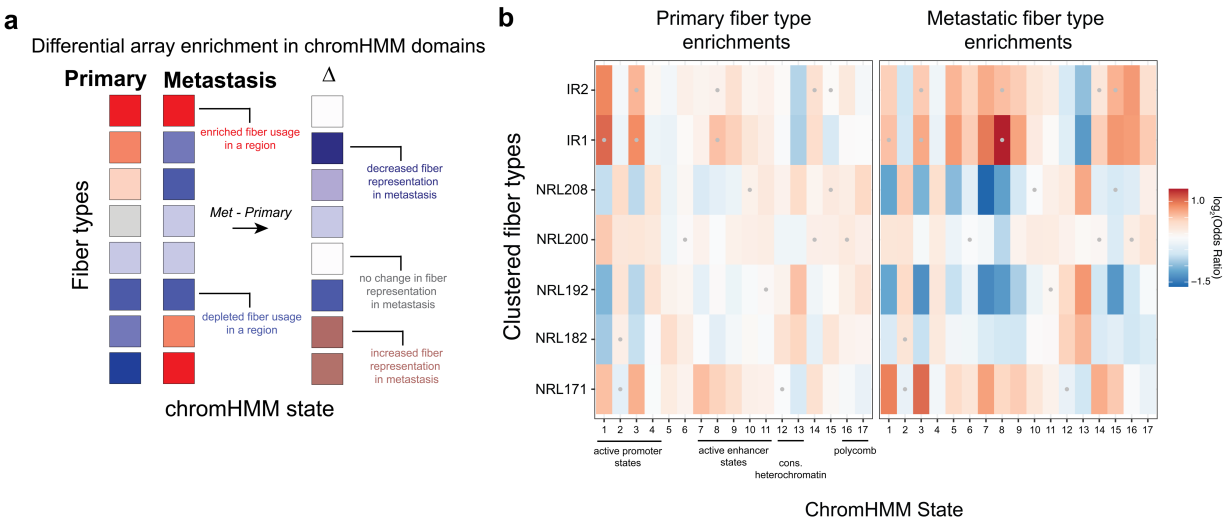
Supplementary Figure 2.15: Raw FACS data for PDX live-dead / human-mouse sorts and associated gating strategies.



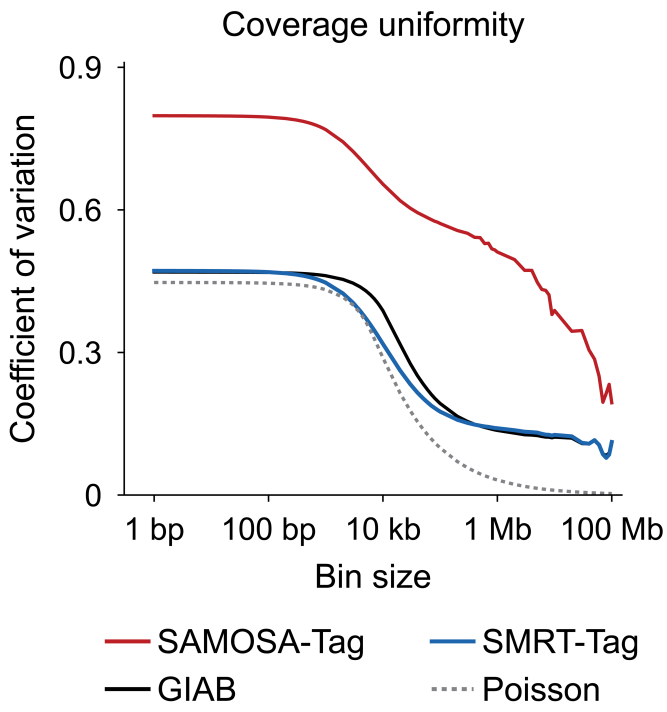
Supplementary Figure 2.16: Comparison of SAMOSA-Tag PDX insertion biases versus cell-line SAMOSA-Tag experiments. a.) TSS insertion bias (left) and FRITSS scores (right) for cell line (OS152 and mESCs) versus PDX SAMOSA-Tag data. **b.)** As in **a.)** but at ChIP-backed CTCF binding sites.



Supplementary Figure 2.17: Analysis of differential single-molecule chromatin accessibility at CTCF sites in primary and metastatic PDX prostate cancer cells. a.) Overview of analytical framework for examining CTCF motif accessibility on individual footprinted chromatin fibers from SAMOSA-Tag primary and metastatic prostate tumor PDX data. **b.)** Unsupervised Leiden clustering of single-molecule chromatin accessibility patterns centered at CTCF motifs reveals 7 different occupancy states, colored here by average accessibility: 5 nucleosome occupied (NO) states with nucleosomes in varying registers around the CTCF binding motif (NO1 – NO5), and two accessible states—accessible (A), which demonstrates the characteristic phasing of nucleosomes surrounding an occupied CTCF binding motif, and hyper-accessible (HA), a state where the entire 750 bp window plotted is accessible to the EcoGII methyltransferase. **c.)** Alluvial plot demonstrating shifts in occupancy state distribution between primary and metastatic samples. Specifically, cluster HA is increased, while cluster A decreases, in metastatic cells compared to primary cells. **d.)** Demonstration of co-measurement of m^6dA accessibility and m^5dCpG methylation with these clustered states. While CTCF motifs that are accessible or hyperaccessible appear to be slightly hypermethylated in metastatic cells compared to primary, motif-containing fibers in the NO state have this effect reversed (slight hypomethylation in the metastatic cells compared to primary).



Supplementary Figure 2.18: Overview of statistical approach for computing differential fiber enrichment and per-sample fiber-type enrichments of SAMOSA-Tag PDX data. a.) Schematic overview of the approach for computing a statistic "delta," which aims to quantify differential representation of fiber types in specific chromHMM domains across the human epigenome in a statistically rigorous manner. Beginning with computed per-domain enrichments in each sample and associated counts, we compute an estimated effect-size (delta) and associated *q* values using a customized logistic regression analysis, and visualize these data in heatmap form with different color scales. **b.)** Fisher's exact test results for each sample (primary vs. met) for clustered fiber types (signal averages shown in **Figure 2.5b**). Red indicates an over-representation of that particular fiber type (y-axis) within the domain (x-axis); blue indicates a depletion of a fiber type within a domain. Grey dots designate tests that are not significant (N.S.). State legends: 1: TSS, 2: TSS Flank, 3: TSS Flank Upstream, 4:TSS Flank Downstream, 5: Transcribed region, 6: Weakly transcribed region, 7: Genic enhancer 1, 8: Genic enhancer 2, 9: Active enhancer 1, 10: Active enhancer 2, 11: Weak enhancer, 12: KRAB zinc finger / repetitive region, 13: Constitutive heterochromatin, 14: Bivalently-marked TSS, 15: Bivalently-marked enhancer, 16: Polycomb repressed, 17: Weakly polycomb repressed.



Supplementary Figure 2.19: Summary plot of differences in coverage uniformity between SAMOSA-Tag, SMRT-Tag, and GIAB samples. Rarefaction curves demonstrating differences in coverage uniformity across multiple window sizes across the genome for SAMOSA-Tag (red), SMRT-Tag (blue), GIAB control data (black), and compared against a random control based on Poisson sampling of reads from the human genome.

2.10. Methods

Data availability

Sequencing data generated from SMRT-Tag libraries derived from HG002, HG003, HG004 and Promega genomic DNA are deposited in the NCBI Sequence Read Archive under accession number PRJNA863422. SAMOSA-Tag data, including subreads and kinetic parameters for OS152 and mESC E14 cell lines are deposited in the Gene Expression Omnibus under accession number GSE225314. SAMOSA-Tag data derived from PDX models are under controlled access to maintain patient privacy.

Code availability

Scripts used to perform analyses are available via GitHub at https://github.com/RamaniLab/SMRT-Tag.

Cell lines and cell culture

OS152 cells were obtained from Alejandro Sweet-Cordero Lab at UCSF, and were routinely tested for cell line authenticity and mycoplasma via CellCheck 9 Plus (IDEXX BioAnalytics). Cells were cultured in standard 1X DMEM (Gibco) supplemented with 10% Bovine Growth Serum (HyClone) and 1% 100X Penicillin-Streptomycin-Glutamine (Corning).

E14 mouse embryonic stem cells were gifted from Elphege Nora Lab at UCSF, and were routinely tested for mycoplasma via PCR (NEBNext® Q5 2X Master Mix). Feeder-free cultures were maintained on 0.2% gelatin, in KnockOut DMEM 1X (Gibco) supplemented with 10% Fetal Bovine Serum (Phoenix Scientific), 1% 100X GlutaMAX (Gibco), 1% 100X MEM Non-Essential Amino Acids (Gibco), 0.128 mM 2-mercaptoethanol (BioRad), and 1X Leukemia Inhibitory Factor (purified and gifted by Barbara Panning Lab at UCSF). Cultures were passaged at least twice before use.

Assembly of SMRT-Tag transposome complexes

Annealing SMRT-Tag Adaptors

HPLC-purified unique SMRT-Tag adaptors were purchased from IDT (Coralville, IA) and normalized to 100 µM in RNase-free water. Barcode sequences were designed with a minimum hamming distance of 4 (**Supplementary Table 2.3**). Adaptors were subsequently diluted to 20 µM in 1X Annealing Buffer (10 mM Tris-HCl pH 7.5 and 100 mM NaCl), annealed via thermocycler (95ºC 5min, RT 30mins, 4ºC hold), and rapidly cooled to -20ºC for long-term storage.

Loading Tn5 transposases with SMRT-Tag adaptors

Purified $Tn5^{R27S,E54K,L372P}$ enzyme was obtained from the Berkeley QB3 MacroLab. Frozen aliquots of $Tn5^{R27S,E54K,L372P}$ enzyme stock (3.9 mg/mL) suspended in Storage Buffer (50 mM Tris-HCl pH 7.5, 800 mM NaCl, 0.2 mM EDTA, 2 mM DTT, 10% glycerol) were thawed at 4ºC, then diluted in Tn5 Dilution Buffer (50 mM Tris-HCl pH 7.5, 200 mM NaCl, 0.1 mM EDTA, 2 mM DTT, and 50% glycerol) to ~1mg/mL Tn5 (18.9 µM monomer) by rotational mixing at 4ºC for 3.5h until fully homogenized. Tn5 was loaded with SMRT-Tag adaptors by gentle mixing of 1.02X volumes of 1mg/mL Tn5 with 1X volume of 20 µM annealed SMRT-Tag adaptors using a wide-bore pipette, followed by an incubation at 23ºC with continuous shaking at 350rpm for 55min. Loaded Tn5 (9.4 µM monomer, "SMRT-Tn5") can be supplemented with glycerol up to a final concentration of 50% and stored at -20° for up to 6 months.

Assays for transposase activity

Confirming Tn5 Loading

Effective adaptor loading was confirmed by blue native PAGE gel-electrophoresis. Briefly, 1-2 µL of Tn5 stock (9.4 µM monomer) diluted in Native Gel Loading Buffer (Invitrogen) was loaded per well on a NativePAGE 4-16% Bis-Tris Gel (Invitrogen) running at 150V for 1 hour at 4ºC, followed by 180V for 15min. Gels were stained with 1X SYBR Gold Solution (Invitrogen) in TAE, followed by 1X Coomassie

Blue (Invitrogen) for 1 hour at room temperature, and imaged on an Odyssey XF imaging system (LI-COR, software version 1.1.0.61).

Assaying Tagmentation Size Tunability

Tagmentation optimization was carried out in parallel using a serial dilution of SMRT-Tn5 stock (9.4 μM monomer) in RNase-free water. Diluted SMRT-Tn5 was incubated with 160 ng of genomic DNA (Promega) using a range of buffers, temperatures and incubation times. Tagmentation reactions were terminated by addition of 0.2% SDS (final concentration 0.04%) and visualized via 0.4-0.6% 1X-TAE-agarose gel. Electrophoresis run time was increased to 2-3h, and voltage decreased to 60-80V to maximize band resolution. Gels were stained with 1X SYBR Gold, and imaged on an Odyssey XF imaging system.

SMRT-Tag on genomic DNA

Preparation of SMRT-Tag libraries

Purified High Molecular Weight genomic DNA (HG002-4, Coriell Institute) was normalized to 40 – 160 ng per sample as input for SMRT-Tag library preparation, which included tagmentation, gap repair, exonuclease cleanup and library validation (Chapter 2.13 – **Supplementary File 1**). For Tn5 tagmentation, reactions were prepared by diluting each sample up to 9 μL in 1X Tagmentation Mix (10 mM TAPS-NaOH pH 7.5, 5 mM MgCl2, and 10% DMF) and adding 1 μL of barcoded Tn5 (varying dilutions from stock, Chapter 2.13 – **Supplementary File 1**). Reactions were incubated at 55ºC for 30min and terminated by addition of 0.2% SDS (final concentration 0.04%) at RT for 5min, followed by a 2X SPRI cleanup and elution in 12 μL of 1X elution buffer (EB, 10 mM Tris-HCl pH 8.5). For gap repair, tagmented samples were incubated in Repair Mix (2U Phusion-HF (New England Biolabs), 80U Taq DNA Ligase (New England Biolabs), 1X Taq DNA Ligase Reaction Buffer, 0.8 mM dNTPs) at 37ºC for 1 hour, followed by a 2X SPRI cleanup and elution in 12 μL of 1X EB. For exonuclease cleanup, reactions were incubated in ExoDigest Mix (100U Exonuclease III (New England Biolabs) per 160 ng, 1X

NEBuffer 2) at 37ºC for 1 hour, followed by a 2X SPRI cleanup and elution in 12 µL of 1X EB. Libraries prepared in the course of method optimization were multiplexed and pooled equimolarly based on the sample concentration measured by Qubit 1X High Sensitivity DNA Assay (Thermo Fisher Scientific).

SMRT-Tag library quality control

To assess repair efficiency, 1 µL of eluted library before and after exonuclease cleanup was measured by Qubit 1X High Sensitivity DNA Assay. To validate library quality, 1 µL of eluted library was assayed via Qubit 1X High Sensitivity DNA Assay and Agilent 2100 Bioanalyzer High Sensitivity DNA Assay (Agilent) to measure sample concentration and library size distribution respectively.

Assaying barcode hopping via pooled gap repair

To assess whether the gap repair reaction affected sample barcoding, we prepared SMRT-Tag libraries as described using barcoded Tn5, but combined samples together after tagmentation into a single gap repair reaction. After gap repair, the pooled sample was treated with ExoDigest mix, as described, to produce a single pooled library.

Optional size selection of SMRT-Tag libraries

For a subset of libraries, an optional size selection step using 35% (v/v) AMPure PB beads diluted in 1X EB was performed to enrich for molecules >5000 bp (HMW). 3.1X of 35% AMPure PB beads was added to a library, incubated at room temperature for 15min, washed twice with 80% ethanol, and the size-selected HMW fraction eluted in 15 µL of 1X EB. Additionally, for some libraries, an additional 0.25X of AMPure PB beads was added to the supernatant and the low molecular weight fraction <5000 bp (LMW) was recovered and eluted in 15 µL of 1X EB.

Sequencing SMRT-Tag libraries

All SMRT-Tag libraries were sequenced on a PacBio Sequel II in house using 8M SMRTcells in both multiplex and monoplex formats. For each SMRTcell, movies were collected for 30 hours, with a 2 hour pre-extension time and a 4 hour immobilization time. Both 2.1 and 2.2 polymerases were used, with polymerase choice dependent on average library size (*i.e.,* HMW fractions were sequenced with 2.2 polymerase, LMW fractions and libraries without size selection with 2.1 polymerase).

SAMOSA-Tag on cell lines

Nuclei isolation

1-2 million OS152 or mESC cells were harvested by centrifugation (300xg, 4ºC, 10min), washed in ice cold 1X PBS, and resuspended in 1 mL cold Nuclear Lysis Buffer (20 mM HEPES, 10 mM KCl, 1 mM MgCl2, 0.1% Triton X-100, 20% Glycerol, 1X Protease Inhibitor (Roche)) by gentle mixing with a wide-bore pipette. The suspension was incubated on ice for 5min, then nuclei were pelleted (600xg, 4ºC, 10min), washed with Buffer M (15 mM Tris-HCl pH 8.0, 15 mM NaCl, 60 mM KCl, 0.5 mM Spermidine), and counted via a Countess III cell counter (Thermo Fisher Scientific).

In situ SAMOSA footprinting

Permeabilized nuclei were pelleted (600xg, 4ºC, 10min) and resuspended in 400 µL Buffer M supplemented with 1 mM S-adenosyl-methionine (SAM, New England Biolabs) and 200 µL aliquoted as an unmethylated control. Nonspecific adenine methyltransferase EcoGII (250U, 10 µL of 25,000U/mL stock, New England Biolabs) was added to the reaction and incubated at 37ºC for 30min with 300rpm shaking every 2min. SAM was replenished to 1.16 mM after 15min in both the reaction and unmethylated control.

Tagmentation of footprinted nuclei

Methylated nuclei along with unmethylated controls were pelleted by centrifugation (600xg, 10min) and gently resuspended in 250 µL 1X Omni-ATAC Buffer (10 mM Tris-HCl pH 7.5, 5 mM $MgCl_2$, 0.33X PBS, 10% DMF, 0.01% Digitonin (Thermo Fisher Scientific), 0.1% Tween-20). The nuclei suspension was then filtered through a 40 µm cell strainer (Scienceware FlowMi), and aggregate dissociation was verified by counting and visualization via Countess III. Both methylated and unmethylated reactions were further distributed into 10,000 – 50,000 nuclei aliquots, and based on the desired library size and cell type, 9.4 – 18.8 pmol of uniquely barcoded Tn5 was added per reaction (Chapter 2.13 – **Supplementary File 1**). Tagmentation reaction volumes were brought up to 50 µL in 1X Omni-ATAC Buffer, then incubated at 55ºC for 45 – 60min.

Tagmentation termination and purification

To terminate tagmentation, reactions were pre-treated with 10 µL of 10mg/mL RNase A (Thermo Fisher Scientific) at 37ºC for 15min with 300rpm shaking. Termination Lysis Buffer (2.5 µL of 20 mg/mL Proteinase K (Ambion), 2.5 µL of 10% SDS and 2.5 µL of 0.5M EDTA) prepared at room temperature was added to the reaction, followed by an incubation at 60ºC with 1000rpm continuous shaking for at least 1 hour, up to 2 hours for improved lysis. To extract tagmented fragments, 2X SPRI beads were added to the reaction, mixed until homogenous, and incubated at 23ºC for 30min with mixing at 350rpm every 3min to keep the beads resuspended. Beads were pelleted via magnet, washed twice in 80% ethanol, then eluted in 20 µL of 1x EB at 37ºC for 15min with interval mixing at 350rpm every 3min to maximize sample recovery. Samples were subjected to an additional 0.6X SPRI cleanup to enrich for fragments > 500bp, and stored at 4ºC overnight, or up to two weeks at -20ºC.

Preparation of SAMOSA-Tag libraries

Purified, tagmented DNA extracted from methylated nuclei and unmethylated controls were normalized up to 160 ng per sample as input for SAMOSA-Tag library preparation. For both OS152 and mESC cells,

a total of 8 methylated replicates along with unmethylated controls, each tagmented with a different set of barcoded SMRT-Tag adaptors, were processed in subsequent steps, including gap repair, exonuclease cleanup and library validation. For gap repair, tagmented samples were incubated in Repair Mix (2U Phusion-HF, 80U Taq DNA Ligase, 1X Taq DNA Ligase Reaction Buffer, 0.8 mM dNTP mix) at 37ºC for 1 hour, followed by a 2X SPRI cleanup and an elution in 12 µL of 1X EB. For exonuclease cleanup, reactions were incubated in ExoDigest Mix (100U Exonuclease III per 160 ng, 1X NEBuffer 2) at 37ºC for 1 hour, followed by a 2X SPRI cleanup and an elution in 12 µL of 1X EB. Repair efficiency and library quality were assessed as for SMRT-Tag.

Sequencing SAMOSA-Tag libraries

SAMOSA-Tag libraries were multiplexed and sequenced on PacBio Sequel II 8M SMRTcells in-house using 2.1 or 2.2 polymerase chemistry depending on the sample (**Supplementary Table 2.2**). For each SMRTcell, movies were collected for 30 hours with a 2 hour pre-extension time and a 4 hour immobilization time.

SAMOSA-Tag on prostate cancer patient derived xenografts (PDX)

Prostate cancer PDX generation and characterization.

Patient derived xenograft (PDX) models were generated as previously described[34]. Briefly, 3-5 mm tumor fragments were isolated from a primary prostate (Gleason 9) tumor and a synchronous metastatic lymph node from the same patient. Tumor fragments were taken immediately after prostatic devascularization during surgery to minimize cell death while preserving the integrity of the tumor microenvironment, placed in 10ml of RPMI 1650 medium for a short transport to the lab from the operating room, and implanted into NSG mice subcutaneously via the flank of NSG mice to establish the PDX lines. After three passages of each PDX tumor in NSG mice, tumors were cryopreserved for future experiments. To ensure that these PDXs faithfully capture the heterogeneity of prostate cancer, tumor sections were subjected to histopathological comparison after each passage. To confirm the passaged PDXs maintained

the integrity of the original PDX, growth patterns were examined. Passage 10 PDXs were processed via SAMOSA-Tag.

PDX sample collection and processing

On the day of collection, PDX tumor samples were surgically removed from mice, aiming to minimize residual mouse tissue, and immediately placed into sterile collection buffer (RPMI-1640) on ice. For each sample, the tumor mass was manually cut to aid dissociation using surgical blades (Fisher Scientific). Each sample was placed into digestion buffer (amount per sample: 5mL of F-12K (Fisher Scientific); 5mL of DMEM (Fisher Scientific); 10 µL DNase I (Worthington Biochemical); 10mg of Liberase-TL (Sigma-Aldrich); 65mg of Collagenase Type III (Worthington Biochemical); 100 µL of 100X Penicillin-Streptomycin (Thermo Fisher Scientific); 40 µl of 0.25 mg/mL Amphotericin B (Fisher Scientific)) and shaken at 750rpm, 37ºC for 1 hour until clumps were visibly dissociated. The resulting single-cell suspensions were spun at 4ºC for 5min at 800xg and the pellets resuspended in cold 1mL PBS (Sigma-Aldrich). The cell suspensions were then strained using a wide-bore P1000 filter tip through a Falcon 70 µm cell strainer (Corning). Samples were then washed twice in 1X PBS via centrifugation at 4ºC, 5min at 800xg. The resulting pellet was resuspended in 1mL Cell Staining Buffer (Biolegend) and counted via hemocytometer at ~8 – 12.5 x 10^6 cells/mL.

Antibody staining and FACS

For blocking, 20 µL of Human TruStain FcX (Fc Receptor Blocking Solution, Biolegend) was added to each sample and incubated for 10min at 4ºC in the dark. 1 µg of PE anti-mouse H-2 Antibody (Biolegend, Cat. 125505) was then added (1 µg for 8 – 12.5 x 10^6 cells total) and allowed to incubate 25mins at 4ºC in the dark. Cells were then washed twice in Cell Staining Buffer and pelleted at 4ºC, 350xg. After antibody staining, 1 µL SYTOX Red Dead Cell Stain (Thermo Fisher Scientific) was added to cells for live-dead staining for 15min at 4ºC in the dark. Cells were kept foil-covered on ice until sorting.

FACS to enrich for live, human cells

To remove contaminant mouse and dead human cells, PDX-derived cells were sorted using a BD FACS Aria II (BD Biosciences) running FACS DIVA software (BD Biosciences) at the UCSF Center for Advanced Technology. Visualization and analysis of the associated data was performed in FlowJo (v10.8.2, BD Biosciences). Cell singlets were selected by gating on forward scatter. Live human cells were selected as PE negative and APC negative, calibrated against single-strain controls, and collected into a 15mL Falcon tube containing 1mL of 1X PBS as receptive buffer. Collection tubes were rinsed with an additional 500 µL of 1X PBS to maximize recovery. Cell counting via hemocytometer estimated between 1.20 – 1.75M cells per sorted PDX sample.

PDX SAMOSA-Tag processing

Sorted cells were placed on ice and immediately processed via SAMOSA-Tag as described for OS152 and mESC E14 cells, with cell pelleting speed reduced from 600xg to 400xg. Due to significant cell loss during preparation, only two unmethylated controls were generated for the primary PDX, and one unmethylated control for the metastasis PDX. Resulting SAMOSA-Tag libraries were assayed for quality as previously described. Primary and metastasis PDX-derived libraries were separately pooled and sequenced on 1 SMRTcell 8M each, using 2.1 polymerase chemistry, and the same run parameters as for OS152 and mESC E14 SAMOSA-Tag libraries.

Preparing low-input genomic DNA libraries using SMRTbell express template prep kit 2.0

SMRTbell libraries were prepared from high molecular weight genomic DNA (HG002, Coriell Institute) using Template Prep Kit 2.0 (TPK2.0, Pacific Biosciences) according to manufacturer's instructions. To assess the efficiency of the enzymatic ligation step specifically, 40 ng of sheared genomic DNA was used as input.

Data Analysis

All scripts and jupyter notebooks used for analyses are available at https://github.com/RamaniLab/SMRT-Tag. All plots were made using R (v.4.2.1) and *ggplot2* (https://ggplot2.tidyverse.org/).

Estimating reaction efficiency

Multiple measures of reaction efficiency were calculated. Tagmentation, gap repair, and exonuclease stepwise efficiencies were determined by dividing the output mass of a given step in nanograms by the input mass in nanograms for that same step. We use the term "repair efficiency" to describe the efficiency of the exonuclease cleanup step, as a proxy for effectiveness of gap repair. Additionally, overall reaction efficiency was either estimated by comparing the final amount of library versus input, or, for libraries where per-step efficiencies were calculated, by multiplying the three stepwise efficiencies together.

Data preprocessing

For all experimental data, HiFi reads were generated from raw subreads using *ccs* (v.6.4.0, Pacific Biosciences) with the additional flag *--hifi-kinetics* to annotate reads with kinetic information. *Lima* (v.2.6.0, Pacific Biosciences) with flag *--ccs* was used to demultiplex runs into sample-specific BAM files, and samples sequenced across multiple cells were merged using *pbmerge* (v1.0.0, Pacific Biosciences). Reads were aligned using *pbmm2* (v.1.9.0, Pacific Biosciences) to the relevant reference genome. SMRT-Tag reads were aligned to the hs37d5 GRCh37 reference genome for variant analyses, and the hg38 reference genome for all other analyses. OS152 SAMOSA-Tag reads was aligned to the hg38 reference genome. mESC E14 SAMOSA-Tag reads were aligned to the GRCm38 reference genome. Primary and metastasis PDX SAMOSA-Tag reads were aligned to a joint hg38 / GRCm39 reference genome and only reads uniquely aligning to hg38 retained for downstream analyses. For all reads, read quality was ascertained from the estimated read quality predicted by *ccs*, and empirical per-read quality score (Q-score) calculated as $-log10\ (\ 1 - (n_{matches} / (n_{matches} + n_{mismatches} + n_{del} + n_{ins})\ )\ )$ or the maximal theoretical quality score when the read contains no variants.

SNV-based analysis of SMRT-Tag demultiplexing

The hs37d5 GRCh37 reference genome[39], GIAB v4.2.1 benchmark[40] VCF and BED files for HG002, HG003, and HG004, and GIAB v3.0 GRCh37 genome stratifications[25] were accessed via the following links:

ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/references/GRCh37/hs37d5.fa.gz

ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio/HG002_NA24385_son/NISTv4.2.1/GRCh37/

ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio/HG003_NA24149_father/NISTv4.2.1/GRCh37/

ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio/HG004_NA24143_mother/NISTv4.2.1/GRCh37/

ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/genome-stratifications/v3.0/v3.0-stratifications-GRCh37.tar.gz

Private SNVs for each individual were obtained using *bcftools* (v1.15.1) and regions for variant calling and evaluation comprising the union of the benchmark BED files were generated using *bedtools* (v2.3.0):

```
bcftools isec \
   --threads 4 \
   -n~100 -w 1 \
   -c some \
   -Oz -o unique.HG002_GRCh37_1_22_v4.2.1_benchmark.vcf.gz \
```

```
    HG002_GRCh37_1_22_v4.2.1_benchmark.vcf.gz \
    HG003_GRCh37_1_22_v4.2.1_benchmark.vcf.gz \
    HG004_GRCh37_1_22_v4.2.1_benchmark.vcf.gz

bcftools isec \
    --threads 4 \
    -n~010 -w 2 \
    -c some \
    -Oz -o unique.HG003_GRCh37_1_22_v4.2.1_benchmark.vcf.gz \
    HG002_GRCh37_1_22_v4.2.1_benchmark.vcf.gz \
    HG003_GRCh37_1_22_v4.2.1_benchmark.vcf.gz \
    HG004_GRCh37_1_22_v4.2.1_benchmark.vcf.gz

bcftools isec \
    --threads 4 \
    -n~001 -w 3 \
    -c some \
    -Oz -o unique.HG004_GRCh37_1_22_v4.2.1_benchmark.vcf.gz \
    HG002_GRCh37_1_22_v4.2.1_benchmark.vcf.gz \
    HG003_GRCh37_1_22_v4.2.1_benchmark.vcf.gz \
    HG004_GRCh37_1_22_v4.2.1_benchmark.vcf.gz

cat HG002_GRCh37_1_22_v4.2.1_benchmark_noinconsistent.bed \
      HG003_GRCh37_1_22_v4.2.1_benchmark_noinconsistent.bed \
      HG004_GRCh37_1_22_v4.2.1_benchmark_noinconsistent.bed | \
sort -k1,1 -k2,2n -k3,3n | \
```

```
bedtools merge | \
bgzip | \
> HG002-4.calling_regions.bed.gz
```

Demultiplexed HG002, HG003, and HG004 SMRT-Tag were aligned to hs37d5 using the *minimap2* aligner (v2.15) and *pbmm2* (v1.9.0) and per-base coverage was tabulated using *mosdepth* (v0.3.3):

```
pbmm2 align \
    --log-level INFO \
    --log-file <OUTPUT_LOG> \
    --preset HiFi \
    --sort \
    --num-threads <THREADS> \
    --sample <SAMPLE_NAME > \
    hs37d5.fa \
    <UNALIGNED_BAM> \
    <OUTPUT_BAM>

mosdepth \
    --threads <THREADS> \
    --use-median \
    --by GRCh37_notinalllowmapandsegdupregions.bed.gz \
    <OUTPUT_PREFIX> \
    <ALIGNED_BAM>
```

Given low depth of coverage, we naively called SNVs within regions defined in the GIAB benchmark BED files supported by at least 2 reads and with minimum mapping quality of 15 using *samtools mpileup* (v1.15.1) and a custom script.

```
samtools mpileup \
   --no-BAQ \
   --fasta-ref hs37d5.fa \
   --positions HG002-4.calling_regions.bed.gz \
   <ALIGNED_BAM> | \
bgzip > <OUTPUT_PLP_GZ>

zcat <OUTPUT_PLP_GZ> | \
plp2vcf.py -q <MIN_MAP_Q> -d <MIN_DEPTH> - | \
bgzip > <OUTPUT_VCF>
```

For each of HG002, HG003, and HG004, naive SNV calls were intersected with private benchmark SNVs in regions labeled “not difficult” in the GIAB v3.0 genome stratification and covered by at least 2 SMRT-Tag reads using *bedtools* (v2.30.0), *samtools* (v1.15.1), and *bcftools* (v1.15.1). For example, the analysis for HG002 SMRT-Tag calls were intersected with HG003 benchmark private SNVs:

```
zcat HG002/mosdepth/HG002.per-base.bed.gz | \
awk -v D=2 '{if ($4 >= D) print}' | \
bedtools merge -i - | \
bedtools intersect \
        -u -a - -b GRCh37_notinalldifficultregions.bed.gz | \
        bgzip \
```

```
> HG002.d2.GRCh37_notinalldifficultregions.bed.gz

bcftools isec \
        --threads <THREADS> \
        -n =2 -w 1 \
        -c some \
        --regions-file HG002.d2.GRCh37_notinalldifficultregions.bed.gz \
        -Oz -o HG002.q15.d2_vs_HG003_unique.vcf.gz \
        HG002.q15.d2.vcf.gz \
        unique.HG003_GRCh37_1_22_v4.2.1_benchmark.vcf.gz

bcftools index \
        -t -f \
        --threads <THREADS> \
        HG002.q15.d2_vs_HG003_unique.vcf.gz

bcftools stats \
        --threads <THREADS> \
        HG002.q15.d2_vs_HG003_unique.vcf.gz \
> HG002.q15.d2_vs_HG003_unique.stats

# Determine total number of covered SNVs:
bcftools view \
        --threads <THREADS> \
        --regions-file HG002.d2.GRCh37_notinalldifficultregions.bed.gz \
        -Oz -o HG002.d2_vs_HG003.base.vcf.gz \
```

```
        --types snps \
        unique.HG003_GRCh37_1_22_v4.2.1_benchmark.vcf.gz

    bcftools index \
        -t -f \
        --threads <THREADS> \
        HG002.d2_vs_HG003.base.vcf.gz

    bcftools stats \
        --threads <THREADS> \
        HG002.d2_vs_HG003.base.vcf.gz \
    > HG002.d2_vs_HG003.base.stats
```

HG002 small variant (SNV and indel) calling and benchmarking

In addition to the hs37d5 GRCh37 reference genome, GIAB v4.2.1 benchmark VCF and BED files for HG002, and GIAB GRCh37 v3.0 genome stratifications used in the genotype demultiplexing analysis, we downloaded publicly available HG002 PacBio Sequel II HiFi reads (SRX5527202), which were generated with ~11 kb size selection and Sequel II chemistry 0.9 and SMRTLink 6.1 pre-release, and are available aligned to the same reference genome via GIAB:

ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/AshkenazimTrio/HG002_NA24385_son/PacBio_SequelII_CCS_11kb/HG002.SequelII.pbmm2.hs37d5.whatshap.haplotag.RTG.10x.trio.bam

We used *pbmm2* for alignment of HG002 SMRT-Tag CCS reads to hs37d5 as before. Similarly, median total coverage for SMRT-Tag and GIAB PacBio reads was determined using *mosdepth*. Reads were

subsampled to 3-, 5-, 10-, and 15-fold depths using *samtools* (v1.15.1) based on *mosdepth* median coverage:

```
samtools view \
    --threads <THREADS> \
    --subsample <FRAC>\
    --subsample-seed 0 \
    --bam \
    --with-header \
    --write-index \
    --output <OUTPUT_BAM> \
    <ALIGNED_BAM>
```

Small variants (SNVs and indels) were called using *DeepVariant* (v1.4.0):

```
run_deepvariant \
    --model_type PACBIO \
    --num_shards <THREADS> \
    --verbosity 0 \
    --logging_dir <DIR> \
    --reads <ALIGNED_BAM> \
    --ref hs37d5.fa \
    --output_vcf <OUTPUT_VCF>
```

We then compared variants called from SMRT-Tag and HG002 PacBio Sequel II HiFi data against NIST v4.2.1 benchmarks[2] using *hap.py* (v0.3.12) and GIAB v3.0 GRCh37 genome stratifications:

```
hap.py \
    -r hs37d5.fa \
    -o <OUTPUT_PREFIX> \
    -f HG002_GRCh37_1_22_v4.2.1_benchmark_noinconsistent.bed \
    --threads <THREADS> \
    --pass-only \
    --engine=vcfeval \
    --verbose \
    --logfile <OUTPUT_LOG> \
    --stratification v3.0-GRCh37-v4.2.1-stratifications.tsv \
    HG002_GRCh37_1_22_v4.2.1_benchmark.vcf.gz \
    <DEEPVARIANT_VCF>
```

Structural variant calling and benchmarking

HG002 SMRT-Tag and GIAB Sequel II data were pre-processed as described above for small variant detection. Benchmark NIST Tier 1 SV calls for HG002 (v0.6) and tandem repeats for hs37d5 were obtained from:

https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/AshkenazimTrio/HG002_NA24385_son/NIST_SV_v0.6/HG002_SVs_Tier1_v0.6.bed

https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/AshkenazimTrio/HG002_NA24385_son/NIST_SV_v0.6/HG002_SVs_Tier1_v0.6.vcf.gz

ftp://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.trf.bed.gz

Reads were subsampled as described above for small variant analysis. Structural variants were called using *pbsv* (v2.8.0; https://github.com/PacificBiosciences/pbsv):

```
pbsv discover \
    --hifi \
    --log-level INFO \
    --log-file <LOG_NAME> \
    --tandem-repeats hg19.trf.bed.gz \
    <ALIGNED_BAM> <OUTPUT_PREFIX>.svsig.gz

pbsv call \
    --hifi \
    --log-level INFO \
    --log-file <LOG_NAME> \
    --num-threads <THREADS> \
    hs37d5.fa <OUTPUT_PREFIX>.svsig.gz <OUTPUT_PREFIX>.vcf
```

VCF files output by *pbsv* were compressed and indexed using *samtools*. We then benchmarked variants against the NIST v0.6 Tier 1 structural variant calls for HG002 using *Truvari* (v3.3.0)[265]:

```
truvari bench \
    --comp <PBSV_VCF_GZ> \
    --base HG002_SVs_Tier1_v0.6.vcf.gz \
    --includebed HG002_SVs_Tier1_v0.6.bed \
    --reference hs37d5.fa \
```

```
--output <OUTPUT_PREFIX> / \
--giabreport \
--refdist 1000 \
--pctsim 0 \
--passonly \
--debug
```

Predicting CpG methylation in single molecule reads

HiFi reads produced using both 2.1 and 2.2. polymerase chemistries were first demultiplexed with *lima* (v.2.6.0) to remove barcode sequences, then *primrose* (v.1.3.0, Pacific Biosciences) used to predict m^5dC methylation status at CpG dinucleotides. Methylation probabilities encoded using the BAM tags ML and MM were parsed to continuous values and used for downstream single-molecule methylation predictions. Per-CpG methylation estimates were made using tools available at https://github.com/PacificBiosciences/pb-CpG-tools.

Predicting nucleosome footprints in SAMOSA-Tag data

SAMOSA-Tag data was preprocessed as above, and subsequently analyzed using a computational pipeline previously developed for detecting m^6dA methylation in HiFi reads[31]. In brief, the kinetics of polymerase base addition were extracted per read, and a series of neural networks trained on kinetic measurements from methylated and unmethylated controls were used to predict the probability of m^6dA methylation at all adenines on both the forward and reverse strands. Methylation probabilities were then binarized into accessibility calls using a two-state hidden Markov model. Accessibility information was encoded per read as a 0/1 modification probability using the BAM tags MM and ML for visualization using a modified version of IGV.

Comparing ATAC-seq signal and SAMOSA accessibility in SAMOSA-Tag data

Total signal, either SAMOSA accessibility or ATAC-seq normalized signal, was aggregated at ATAC-seq peaks identified in the OS152 cell line. Values were log-transformed and Pearson's *r* calculated as a measure of correlation.

U2OS CTCF ChIP-seq processing

Processed BED files from GEO accession GSE87831 were lifted over from reference hg19 to hg38, and then analyzed as in Ramani *et al* (2019)[266] to obtain predicted CTCF binding sites.

Prostate CTCF ChIP-seq processing

Processed BED files containing CTCF ChIP-seq peaks in the metastatic prostate adenocarcinoma cell line LNCaP were obtained from ENCODE (Accession ID: ENCFF275GDH), and analyzed as in Ramani et al. 2019[42] to obtain predicted CTCF binding sites.

Insertion bias analyses at TSS and CTCF sites

Read-ends from SAMOSA-Tag data were extracted from BAM files and tabulated in a 5 kilobase window surrounding annotated GENCODEV28 (hg38) or GENCODEM25 (GRCm38) transcriptional start sites (TSSs) or ChIP-seq-backed CTCF binding sites. For visualization, all metaplots were smoothed with a running mean of 100 nucleotides. FRITSS / FRICBS was calculated as the fraction of read-ends falling within the 5 kilobase window.

CTCF CpG and accessibility analyses

m^6dA accessibility signal surrounding predicted CTCF sites was extracted from accessibility pickle files and *leiden* clustered as in Abdulhay et al. 2020[31]. In addition to filtering out clusters that together accounted for less than 10% of all data, we also manually filtered out 1 cluster that corresponded to completely unmethylated fibers. Compared against all analyzed fibers surrounding CTCF sites, this

cluster accounted for 3,627 fibers, or 11.5% of all CTCF-motif containing fibers in OS152 SAMOSA-Tag data, and 245 fibers or 1.5% in PDX SAMOSA-Tag data. For CpG analyses, we used custom python scripts to convert CpG methylation to a similar pickle format as m^6dA accessibility, and then used identical scripts to extract CpG methylation information per molecule, centered at CTCF sites. All data was then converted into text files for easy loading into R and visualization in *ggplot2*.

Classifying fibers by CpG content and CpG methylation

We binned all sequenced fibers by CpG content and CpG methylation to arrive at four bins, which we defined as high CpG content / methylation (*i.e.* > 0.5 average *primrose* score on a fiber; > 10 CpGs per kilobase), low CpG content / methylation (vice-versa), as well as high / low and low / high bins.

Fiber type clustering

We calculated single-molecule autocorrelograms and performed *leiden* clustering as in[31]. In addition to filtering out all clusters that together comprised less than 10% of all fibers, we also manually filtered out unmethylated / lowly methylated fibers, which fell out of the *leiden* clustering analysis and together accounted for 317,768 fibers (12.5% of all clustered fibers) in OS152 SAMOSA-Tag data.

Fiber type enrichment

Fisher's exact tests to determine fiber type enrichment were performed as in Abdulhay et al. 2020[31]. Briefly, for examining enrichment of fiber type A stratified by feature B, a 2x2 contingency table was constructed by counting fibers that fell into four groups, A ∩ B, A ∩ B ', A ' ∩ B, and A ' ∩ B '. The table was used as input for a one-sided Fisher's exact test and resulting *p* values corrected for multiple testing using Storey's *q* value.

Prostate-specific epigenome stratification

Normal prostate tissue-specific chromHMM annotations in BED format were obtained from[37]. Annotations were lifted over from reference hg19 to hg38.

Differential fiber usage

Differential fiber usage per domain was determined using a logistic regression approach. First, coverage of epigenomic domains by different fiber types in each replicate was determined as in Abdulhay et al. 2020[31]. Then, to determine differential usage for fiber type A in domain B, coverage was aggregated by whether individual fibers were of type A *and* mapped to domain B, or not. Counts for each of these two categories (domain A ∩ fiber B vs. (domain A ∩ fiber B)') were determined for each replicate, and then normalized across replicates using a median of medians approach to account for library depth. Normalized counts per replicate were then used as weights for a logistic regression model with the domain / fiber status as the response variable and case status of the library (primary vs. metastasis PDX) as the predictor. The model was fit using the *glm* function in R (v.4.2.1) and the coefficient of case status used as an estimate of log fold change (Δ, "delta") in metastasis vs. primary. This regression was repeated for every domain and fiber combination observed in the data (7 fiber types, and 17 domain annotations), and the associated *p* value for each fold change corrected for multiple testing using Storey's *q* value[267]. A threshold for significance was set at Storey's $q \leq 0.05$.

2.11. Supplementary Notes

On loading the PacBio instrument

Sequencing on the PacBio platform is fundamentally different from both standard Illumina sequencing and sequencing on the Oxford Nanopore platform. Thus, we feel it important to clarify technical considerations on the PacBio platform that motivated our experimental design decisions when optimizing SMRT-Tag and SAMOSA-Tag. Leveraging the maximum potential of PacBio sequencing (namely, direct detection of specific DNA modifications), requires libraries be made without PCR. This leads to an important limitation, as DNA is lost at every step of a sequencing protocol; importantly this includes steps required for loading the PacBio sequencer—specifically, polymerase binding and loading on SMRTCells ("flow cells"). PacBio sequencing performance is influenced by numerous properties: library fragment length distribution, presence of DNA damage, batch-to-batch SMRTCell and polymerase characteristics, and perhaps most importantly, the on-plate loading concentration (OPLC) of libraries. Maximizing the P1 productivity (fraction of zero-mode waveguides with sequencing one and only one molecule) and CCS yield (and thus, minimizing cost-per-base) of a PacBio flow cell requires a high per-run OPLC, and the only way to maximize OPLC is by 1) minimizing DNA loss during clean-up steps and 2) when possible, pooling libraries together. In our paper, we provide all salient technical details for all SMRT-Tag and SAMOSA-Tag sequencing libraries constructed, including OPLC in Chapter 2.13 – **Supplementary File 1**. While achieving high OPLC to minimize cost-per-base was the primary focus of most experiments presented in this paper, we include as a valuable reference point a sequencing run where a single 40 ng pool of human genomic DNA was tagmented and loaded on the sequencer. This experiment illustrates the power of the SMRT-Tag approach for maximizing coverage of low-input samples.

On estimating input reduction of SMRT-Tag versus conventional library preparation protocols

Given the above-mentioned tradeoffs involved in PCR-free sequencing of native DNA, we think it important to outline the math behind our claim that SMRT-Tag enables sequencing of 1 – 5% the input

of traditional library preparation protocols. The PacBio Template Preparation Kit 2.0 recommends a minimum input of 5 µg, whereas the SMRTbell Prep. Kit 3.0 (which became available in mid-2022 after key experiments for our manuscript were completed) requires 1 – 5 µg. Using 1 – 5 µg (~170,000 – 800,000 cells) purified DNA for current methods and 40 ng (~7,000 cells) as a conservative lower-bound for SMRT-Tag based on our monoplex experiment, SMRT-Tag requires 0.8 – 4% as much input DNA as existing protocols. Or, stated slightly differently, using the recommended minimum of 1 µg as the "average use" lower bound based on manufacturer recommendations, the 40 ng SMRT-Tag input is a 96% reduction. SAMOSA-Tag experiments used 30,000 – 50,000 nuclei (~180 – 300 ng DNA, though the direct comparison is not appropriate given that the substrate is *chromatin* not purified gDNA), which is 0.6 – 9% of the amounts reported in the publications describing the following single-molecule chromatin profiling methods: SAMOSA[4,11] / Fiber-seq[5] (2 µg), DiMeLo-seq[8] (6 – 30 µg), SMAC-seq[6] (6 µg), nanoNOMe[7] (2 – 3 µg), and MeSMLR-seq[12] (input quantity not reported, but minimum recommended for the Oxford Nanopore Ligation Sequencing Kit used to construct libraries is 1 µg).

Based on the above, we report the following conservative underestimates throughout our manuscript: SMRT-Tag requires 1 – 5% as much DNA as existing protocols, equating to input reduction of 95 – 99%. SAMOSA-Tag requires 1 – 10% as much input material as existing protocols, corresponding to reduction by 90 – 99%. Therefore, both SMRT-Tag and SAMOSA-Tag enable reductions in the *magnitude* of input DNA required by approximately one or two orders (i.e., 10-fold or 100-fold).

2.12. Supplementary Tables

Supplementary Table 2.1: Gap repair conditions tested in optimizing SMRT-Tag.

ID	Repair condition - description	Repair condition - abbreviated name
1	NEB T4 DNA Polymerase (3U), Ampligase (10U), Ampligase Buffer, 0.1 mM dNTPs, 30min @ 37ºC	NEBT4/1x/Amp/2x/AmpBuf/0.1dNTP
2	NEB T4 DNA Polymerase (3U), Ampligase (10U), Ampligase Buffer, 1 mM dNTPs, 30min @ 37ºC	NEBT4/1x/Amp/2x/AmpBuf/1dNTP
3	NEB T4 DNA Polymerase (3U), Ampligase (10U), Ampligase Buffer, 10 mM dNTPs, 30min @ 37ºC	NEBT4/1x/Amp/2x/AmpBuf/10dNTP
4	NEB T4 DNA Polymerase (3U), Ampligase (10U), Ampligase Buffer, 0.5 mM dNTPs, 30min @ 37ºC	NEBT4/1x/Amp/2x/AmpBuf/0.5dNTP
5	NEB T4 DNA Polymerase (6U), Ampligase (10U), Ampligase Buffer, 10 mM dNTPs, 30min @ 37ºC	NEBT4/2x/Amp/2x/AmpBuf/10dNTP
6	NEB T4 DNA Polymerase (3U), Ampligase (5U), Thermo T4 DNA Polymerase Buffer, 1 mM dNTPs, 0.5 mM NAD+, 30min @ 37ºC	NEBT4/1x/Amp/1x/T4Buf/1dNTP
7	NEB T4 DNA Polymerase (3U), Ampligase (10U), Thermo T4 DNA Polymerase Buffer, 0.1 mM dNTPs, 0.5 mM NAD+, 30min @ 37ºC	NEBT4/1x/Amp/2x/T4Buf/0.1dNTP
8	NEB T4 DNA Polymerase (3U), Ampligase (10U), Thermo T4 DNA Polymerase Buffer, 0.5 mM dNTPs, 0.5 mM NAD+, 30min @ 37ºC	NEBT4/1x/Amp/2x/T4Buf/0.5dNTP
9	NEB T4 DNA Polymerase (3U), Ampligase (10U), Thermo T4 DNA Polymerase Buffer, 1 mM dNTPs, 0.5 mM NAD+, 30min @ 37ºC	NEBT4/1x/Amp/2x/T4Buf/1dNTP
10	NEB T4 DNA Polymerase (3U), Ampligase (10U), Thermo T4 DNA Polymerase Buffer, 10 mM dNTPs, 0.5 mM NAD+, 30min @ 37ºC	NEBT4/1x/Amp/2x/T4Buf/10dNTP
11	NEB T4 DNA Polymerase (7.5U), Ampligase (25U), Thermo T4 DNA Polymerase Buffer, 1 mM dNTPs, 0.5 mM NAD+, 30min @ 37ºC	NEBT4/2.5x/Amp/5x/T4Buf/1dNTP
12	Thermo T4 DNA Polymerase (5U), Ampligase (10U), Thermo T4 DNA Polymerase Buffer, 1 mM dNTPs, 30min @ 37ºC	ThermoT4/1x/Amp/2x/T4Buf/1dNTP
13	Thermo T4 DNA Polymerase (5U), Ampligase (10U), Thermo T4 DNA Polymerase Buffer, 10 mM dNTPs, 2.5 mM NAD+, 30min @ 37ºC	ThermoT4/1x/Amp/2x/T4Buf/10dNTP/2.5NAD
14	Thermo T4 DNA Polymerase (5U), Ampligase (10U), Thermo T4 DNA Polymerase Buffer, 0.1 mM dNTPs, 0.5 mM NAD+, 30min @ 37ºC	ThermoT4/1x/Amp/2x/T4Buf/0.1dNTP/0.5NAD
15	Thermo T4 DNA Polymerase (5U), Ampligase (10U), Thermo T4 DNA Polymerase Buffer, 0.5 mM dNTPs, 0.5 mM NAD+, 30min @ 37ºC	ThermoT4/1x/Amp/2x/T4Buf/0.5dNTP/0.5NAD
16	Thermo T4 DNA Polymerase (5U), Ampligase (10U), Thermo T4 DNA Polymerase Buffer, 1 mM dNTPs, 0.5 mM NAD+, 30min @ 37ºC	ThermoT4/1x/Amp/2x/T4Buf/1dNTP/0.5NAD/30min
17	Thermo T4 DNA Polymerase (5U), Ampligase (10U), Thermo T4 DNA Polymerase Buffer, 10 mM dNTPs, 0.5 mM NAD+, 30min @ 37ºC	ThermoT4/1x/Amp/2x/T4Buf/10dNTP/0.5NAD

ID	Repair condition – description	Repair condition – abbreviated name
18	Thermo T4 DNA Polymerase (5U), Ampligase (10U), Thermo T4 DNA Polymerase Buffer, 1 mM dNTPs, 0.5 mM NAD+, 60min @ 37ºC	ThermoT4/1x/Amp/2x/T4Buf/1dNTP/0.5NAD/60min
19	Thermo T4 DNA Polymerase (5U), Ampligase (5U), Thermo T4 DNA Polymerase Buffer, 1 mM dNTPs, 0.5 mM NAD+, 30min @ 37ºC	ThermoT4/1x/Amp/1x/T4Buf/1dNTP/0.5NAD
20	Thermo T4 DNA Polymerase (5U), Ampligase (10U), Thermo T4 DNA Polymerase Buffer, 1 mM dNTPs, 0.5 mM NAD+, 5% PEG4000, 30min @ 37ºC	ThermoT4/1x/Amp/2x/T4Buf/1dNTP/0.5NAD/PEG
21	Thermo T4 DNA Polymerase (5U), Ampligase (20U), Thermo T4 DNA Polymerase Buffer, 1 mM dNTPs, 0.5 mM NAD+, 30min @ 37ºC	ThermoT4/1x/Amp/4x/T4Buf/1dNTP/0.5NAD
22	Thermo T4 DNA Polymerase (5U), Ampligase (10U), Thermo T4 DNA Polymerase Buffer, 1 mM dNTPs, 0.5 mM NAD+, 100ug/uL BSA, 30min @ 37ºC	ThermoT4/1x/Amp/2x/T4Buf/1dNTP/0.5NAD/BSA
23	Thermo T4 DNA Polymerase (5U), Ampligase (10U), NEB CutSmart Buffer, 1 mM dNTPs, 0.5 mM NAD+, 30min @ 37ºC	ThermoT4/1x/Amp/2x/CutSmartBuf/1dNTP/0.5NAD
24	Thermo T4 DNA Polymerase (5U), Ampligase (10U), NEB Buffer2, 1 mM dNTPs, 0.5 mM NAD+, 30min @ 37ºC	ThermoT4/1x/Amp/2x/NEBuf2/1dNTP/0.5NAD
25	Thermo T4 DNA Polymerase (10U), Ampligase (20U), Thermo T4 DNA Polymerase Buffer, 1 mM dNTPs, 0.5 mM NAD+, 30min @ 37ºC	ThermoT4/2x/Amp/4x/T4Buf/1dNTP/0.5NAD
26	Thermo T4 DNA Polymerase (10U), Ampligase (20U), Thermo T4 DNA Polymerase Buffer, 1 mM dNTPs, 2.5 mM NAD+, 30min @ 37ºC	ThermoT4/2x/Amp/4x/T4Buf/1dNTP/2.5NAD
27	Thermo T4 DNA Polymerase (12.5U), Ampligase (25U), Thermo T4 DNA Polymerase Buffer, 1 mM dNTPs, 0.5 mM NAD+, 30min @ 37ºC	ThermoT4/2.5x/Amp/5x/T4Buf/1dNTP/0.5NAD
28	Thermo T4 DNA Polymerase (5U), NEB Taq DNA Ligase (80U), NEB Taq DNA Ligase Buffer, 1 mM dNTPs, 30min @ 37ºC	ThermoT4/1x/Taq/TaqBuf/1dNTP
29	Thermo T4 DNA Polymerase (5U), NEB T7 DNA Ligase (3000U), NEB StickTogether Ligase Buffer, 1 mM dNTPs, 30min @ 37ºC	ThermoT4/1x/T7/StickBuf/1dNTP
30	Thermo T4 DNA Polymerase (5U), NEB HiFi Taq DNA Ligase (1U), NEB HiFi Taq DNA Ligase Buffer, 1 mM dNTPs, 30min @ 37ºC	ThermoT4/1x/HiFiTaq/HiFiTaqBuf/1dNTP
31	Thermo T4 DNA Polymerase (5U), NEB 9° N Ligase (80U), NEB 9° N Ligase Buffer, 1 mM dNTPs, 30min @ 37ºC	ThermoT4/1x/9N/9NBuf/1dNTP
32	NEB Phusion High-Fidelity DNA Polymerase (0.8U), Ampligase (2U), Ampligase Buffer, 0.05 mM dNTPs, 50 mM KCl, 20% DMF, 30min @ 37ºC	Phu/1x/Amp/1x/AmpBuf/0.05dNTP/50KCl/20DMF/30min

ID	Repair condition - description	Repair condition - abbreviated name
33	NEB Phusion High-Fidelity DNA Polymerase (0.8U), Ampligase (2U), Ampligase Buffer, 0.05 mM dNTPs, 50 mM KCl, 10% DMF, 30min @ 37ºC	Phu/1x/Amp/1x/AmpBuf/0.05dNTP/50KCl/10DMF/30min
34	NEB Phusion High-Fidelity DNA Polymerase (0.8U), Ampligase (2U), Ampligase Buffer, 0.05 mM dNTPs, 50 mM KCl, 10% DMF, 30min @ 37ºC + 15min @ 45ºC	Phu/1x/Amp/1x/AmpBuf/0.05dNTP/50KCl/10DMF/45min
35	NEB Phusion High-Fidelity DNA Polymerase (0.8U), Ampligase (2U), Ampligase Buffer, 0.8 mM dNTPs, 25 mM KCl, 10% DMF, 60min @ 37ºC	Phu/1x/Amp/1x/AmpBuf/0.8dNTP/25KCl/10DMF/60min
36	NEB Phusion High-Fidelity DNA Polymerase (4U), Ampligase (10U), Ampligase Buffer, 0.05 mM dNTPs, 50 mM KCl, 20% DMF, 30min @ 37ºC	Phu/5x/Amp/5x/AmpBuf/0.05dNTP/50KCl/20DMF/30min
37	NEB Phusion High-Fidelity DNA Polymerase (4U), Ampligase (10U), Ampligase Buffer, 0.05 mM dNTPs, 50 mM KCl, 10% DMF, 30min @ 37ºC	Phu/5x/Amp/5x/AmpBuf/0.05dNTP/50KCl/10DMF/30min
38	NEB Phusion High-Fidelity DNA Polymerase (4U), Ampligase (10U), Ampligase Buffer, 0.05 mM dNTPs, 50 mM KCl, 10% DMF, 30min @ 37ºC + 15min @ 45ºC	Phu/5x/Amp/5x/AmpBuf/0.05dNTP/50KCl/10DMF/45min
39	NEB Phusion High-Fidelity DNA Polymerase (4U), Ampligase (10U), Ampligase Buffer, 0.8 mM dNTPs, 25 mM KCl, 10% DMF, 60min @ 37ºC	Phu/5x/Amp/5x/AmpBuf/0.8dNTP/25KCl/10DMF/60min
40	NEB Phusion High-Fidelity DNA Polymerase (4U), Ampligase (10U), Ampligase Buffer, 0.8 mM dNTPs, 25 mM KCl, 60min @ 37ºC	Phu/5x/Amp/5x/AmpBuf/0.8dNTP/25KCl/60min
41	NEB Phusion High-Fidelity DNA Polymerase (0.32U), NEB Taq DNA Ligase (80U), NEB Taq DNA Ligase Buffer, 0.8 mM dNTPs, 30min @ 37ºC	Phu/0.4x/Taq/TaqBuf/0.8dNTP
42	NEB Phusion High-Fidelity DNA Polymerase (0.32U), NEB Taq DNA Ligase (80U), NEB Taq DNA Ligase Buffer, 0.8 mM dNTPs, 10% DMF, 30min @ 37ºC	Phu/0.4x/Taq/TaqBuf/0.8dNTP/10DMF
43	NEB Phusion High-Fidelity DNA Polymerase (0.8U), NEB Taq DNA Ligase (80U), Ampligase Buffer, 0.05 mM dNTPs, 50 mM KCl, 10% DMF, 30min @ 37ºC	Phu/1x/Taq/AmpBuf/0.05dNTP/50KCl/10DMF
44	NEB Phusion High-Fidelity DNA Polymerase (2U), NEB Taq DNA Ligase (80U), NEB Taq DNA Ligase Buffer, 0.8 mM dNTPs, 30min @ 37ºC	Phu/2.5x/Taq/TaqBuf/0.8dNTP/30min
45	NEB Phusion High-Fidelity DNA Polymerase (2U), NEB Taq DNA Ligase (80U), NEB Taq DNA Ligase Buffer, 0.8 mM dNTPs, 60min @ 37ºC	Phu/2.5x/Taq/TaqBuf/0.8dNTP/60min
46	NEB Phusion High-Fidelity DNA Polymerase (4U), NEB Taq DNA Ligase (80U), NEB Taq DNA Ligase Buffer, 0.8 mM dNTPs, 60min @ 37ºC	Phu/5x/Taq/TaqBuf/0.8dNTP/60min
47	NEB PreCR Repair Mix (1U), ThermoPol Reaction Buffer, 0.1 mM dNTPs, 0.5 mM NAD+, 30min @ 37ºC	PreCR/ThermoPolBuf/0.1dNTP/0.5NAD

ID	Repair condition - description	Repair condition - abbreviated name
48	NEB Bst DNA Polymerase, Full Length (0.8U), NEB Taq DNA Ligase (60U), ThermoPol Reaction Buffer, 1 mM dNTPs, 0.5 mM NAD+, 30min @ 37ºC	Bst/Taq/ThermoPolBuf/1dNTP/0.5NAD
49	NEB Phusion High-Fidelity DNA Polymerase (2U), NEB 9° N Ligase (80U), NEB 9° N Ligase Buffer, 0.8 mM dNTPs, 30min @ 37ºC	Phu/9N/9NBuf/0.8dNTP
50	NEB Phusion High-Fidelity DNA Polymerase (2U), NEB HiFi Taq DNA Ligase (1U), NEB HiFi Taq DNA Ligase Buffer, 0.8 mM dNTPs, 60min @ 37ºC	Phu/HiFiTaq/HiFiTaqBuf/0.8dNTP
51	NEB Q5 High-Fidelity DNA Polymerase (0.4U), Ampligase (10U), NEB Q5 Reaction Buffer, 0.2 mM dNTPs, 0.5 mM NAD+, 30min @ 37ºC	Q5/Amp/Q5Buf/0.2dNTP/0.5NAD
52	NEB Phusion High-Fidelity DNA Polymerase (2U), NEB Taq DNA Ligase (80U), NEB Taq DNA Ligase Buffer, 0.8 mM dNTPs, 0.8 mM ATP, T4 PNK (5U), homemade PreCR Repair Mix, 30min @ 37ºC + 60min @ 37ºC	Phu/2.5x/Taq/TaqBuf/0.8dNTP/PreCRMix
53	Thermo T4 DNA Polymerase (5U), Ampligase (10U), Thermo T4 DNA Polymerase Buffer, 1 mM dNTPs, 0.5 mM NAD+, T4 PNK (5U), 30min @ 37ºC	ThermoT4/1x/Amp/2x/T4Buf/1dNTP/0.5NAD/PNK
54	NEB T4 DNA Polymerase (3U), NEB HiFi Taq Ligase (1U), NEB Buffer2, 1 mM dNTPs, 0.8 mM ATP, T4 PNK (5U), 0.5 mM NAD+, homemade PreCR Repair Mix, 30min @ 37ºC + 30min @ 37ºC	NEBT4/1x/HiFiTaq/1x/NEBuf2/1dNTP/PreCRMix
55	NEB T4 DNA Polymerase (9U), NEB HiFi Taq Ligase (3U), NEB Buffer2, 1 mM dNTPs, 0.8 mM ATP, T4 PNK (5U), 0.5 mM NAD+, homemade PreCR Repair Mix, 30min @ 37ºC + 30min @ 37ºC	NEBT4/3x/HiFiTaq/3x/NEBuf2/1dNTP/PreCRMix
56	Thermo T4 DNA Polymerase (5U), NEB HiFi Taq Ligase (1U), Thermo T4 DNA Polymerase Buffer, 1 mM dNTPs, 0.8 mM ATP, T4 PNK (5U), 0.5 mM NAD+, homemade PreCR Repair Mix, 30min @ 37ºC + 30min @ 37ºC	ThermoT4/1x/HiFiTaq/1x/T4Buf/1dNTP/PreCRMix
57	Thermo T4 DNA Polymerase (5U), Ampligase (10U), Thermo T4 DNA Polymerase Buffer, 1 mM dNTPs, 0.8 mM ATP, T4 PNK (5U), 0.5 mM NAD+, homemade PreCR Repair Mix, 30min @ 37ºC + 30min @ 37ºC	ThermoT4/1x/Amp/2x/T4Buf/1dNTP/PreCRMix
58	Thermo T4 DNA Polymerase (15U), Ampligase (30U), Thermo T4 DNA Polymerase Buffer, 1 mM dNTPs, 0.8 mM ATP, T4 PNK (5U), 0.5 mM NAD+, homemade PreCR Repair Mix, 30min @ 37ºC + 30min @ 37ºC	ThermoT4/3x/Amp/6x/T4Buf/1dNTP/PreCRMix
59	Thermo T4 DNA Polymerase (5U), Ampligase (10U), NEB Buffer2, 1 mM dNTPs, 0.8 mM ATP, T4 PNK (5U), 0.5 mM NAD+, homemade PreCR Repair Mix, 30min @ 37ºC + 30min @ 37ºC + 30min @ 37ºC	ThermoT4/1x/Amp/2x/NEBuf2/1dNTP/PreCRMix

ID	Repair condition - description	Repair condition - abbreviated name
60	NEB Phusion High-Fidelity DNA Polymerase (2U), NEB Taq DNA Ligase (80U), NEB Taq DNA Ligase Buffer, 0.8 mM dNTPs, 0.8 mM ATP, T4 PNK (5U), 1 mM NAD+, 50 mM KCl, homemade PreCR Repair Mix, 30min @ 37ºC + 30min @ 37ºC + 30min @ 37ºC	Phu/2.5x/Taq/TaqBuf/0.8dNTP/1NAD/PreCRMix
61	NEB Phusion High-Fidelity DNA Polymerase (4U), Ampligase (10U), Ampligase Buffer, 0.8 mM dNTPs, 0.8 mM ATP, T4 PNK (5U), 0.5 mM NAD+, 50 mM KCl, homemade PreCR Repair Mix, 30min @ 37ºC + 30min @ 37ºC + 30min @ 37ºC	Phu/5x/Amp/5x/AmpBuf/0.8dNTP/PreCRMix

Supplementary Table 2.2: Gap repair condition efficiencies evaluated in optimizing SMRT-Tag.

Repair condition	ID	Efficiency (%)	Reaction Input Mass	Source	Repair condition - abbreviated name	Subgroup mean repair efficiency	Subgroup std. dev. repair efficiency
Phu/Amp	34	56.03	160	Promega	Phu/1x/Amp/1x/AmpBuf/0.05dNTP/50KCl/10DMF/45 min	36.48	27.64
Phu/Amp	34	16.93	160	Promega	Phu/1x/Amp/1x/AmpBuf/0.05dNTP/50KCl/10DMF/45 min		
Phu/Amp	35	24.60	160	Promega	Phu/1x/Amp/1x/AmpBuf/0.8dNTP/25KCl/10DMF/60 min		
Phu/Amp	37	10.17	160	Promega	Phu/5x/Amp/5x/AmpBuf/0.05dNTP/50KCl/10DMF/30 min		
Phu/Amp	38	44.80	160	Promega	Phu/5x/Amp/5x/AmpBuf/0.05dNTP/50KCl/10DMF/45 min		
Phu/Amp	39	25.00	160	Promega	Phu/5x/Amp/5x/AmpBuf/0.8dNTP/25KCl/10DMF/60 min	25.76	1.07
Phu/Amp	39	26.52	160	Promega	Phu/5x/Amp/5x/AmpBuf/0.8dNTP/25KCl/10DMF/60 min		
Phu/Amp	40	43.93	160	Promega	Phu/5x/Amp/5x/AmpBuf/0.8dNTP/25KCl/60min	36.93	9.91
Phu/Amp	40	29.92	160	Promega	Phu/5x/Amp/5x/AmpBuf/0.8dNTP/25KCl/60min		
Phu/Taq	43	37.09	160	Promega	Phu/1x/Taq/AmpBuf/0.05dNTP/50KCl/10DMF		
Phu/Taq	44	42.92	160	Promega	Phu/2.5x/Taq/TaqBuf/0.8dNTP/30min		
Phu/Taq	45	39.50	160	Promega	Phu/2.5x/Taq/TaqBuf/0.8dNTP/60min	40.45	4.83

Repair condition	ID	Efficiency (%)	Reaction Input Mass	Source	Repair condition - abbreviated name	Subgroup mean repair efficiency	Subgroup std. dev. repair efficiency
Phu/Taq	45	36.16	160	Promega	Phu/2.5x/Taq/TaqBuf/0.8dNTP/60min		
Phu/Taq	45	45.68	160	Promega	Phu/2.5x/Taq/TaqBuf/0.8dNTP/60min		
Phu/Taq	46	42.81	160	Promega	Phu/5x/Taq/TaqBuf/0.8dNTP/60min		
T4/Amp	16	47.44	160	Promega	ThermoT4/1x/Amp/2x/T4Buf/1dNTP/0.5NAD/30min	35.09	9.800
T4/Amp	16	28.33	160	Promega	ThermoT4/1x/Amp/2x/T4Buf/1dNTP/0.5NAD/30min		
T4/Amp	16	41.60	160	Promega	ThermoT4/1x/Amp/2x/T4Buf/1dNTP/0.5NAD/30min		
T4/Amp	16	24.55	160	Promega	ThermoT4/1x/Amp/2x/T4Buf/1dNTP/0.5NAD/30min		
T4/Amp	16	43.86	160	Promega	ThermoT4/1x/Amp/2x/T4Buf/1dNTP/0.5NAD/30min		
T4/Amp	16	36.82	160	Promega	ThermoT4/1x/Amp/2x/T4Buf/1dNTP/0.5NAD/30min		
T4/Amp	16	23.06	160	Promega	ThermoT4/1x/Amp/2x/T4Buf/1dNTP/0.5NAD/30min		
T4/Amp	18	34.2	160	Promega	ThermoT4/1x/Amp/2x/T4Buf/1dNTP/0.5NAD/60min		
T4/Amp	20	33.24	160	Promega	ThermoT4/1x/Amp/2x/T4Buf/1dNTP/0.5NAD/PEG	35.73	3.13
T4/Amp	20	40.28	160	Promega	ThermoT4/1x/Amp/2x/T4Buf/1dNTP/0.5NAD/PEG		
T4/Amp	20	33.02	160	Promega	ThermoT4/1x/Amp/2x/T4Buf/1dNTP/0.5NAD/PEG		
T4/Amp	20	34.51	50	Promega	ThermoT4/1x/Amp/2x/T4Buf/1dNTP/0.5NAD/PEG		

Repair condition	ID	Efficiency (%)	Reaction Input Mass	Source	Repair condition - abbreviated name	Subgroup mean repair efficiency	Subgroup std. dev. repair efficiency
T4/Amp	20	37.60	50	Promega	ThermoT4/1x/Amp/2x/T4Buf/1dNTP/0.5NAD/PEG		
T4/Amp	21	36.10	160	Promega	ThermoT4/1x/Amp/4x/T4Buf/1dNTP/0.5NAD	36.07	5.16
T4/Amp	21	41.21	160	Promega	ThermoT4/1x/Amp/4x/T4Buf/1dNTP/0.5NAD		
T4/Amp	21	30.90	160	Promega	ThermoT4/1x/Amp/4x/T4Buf/1dNTP/0.5NAD		
T4/Amp	57	18.07	160	Promega	ThermoT4/1x/Amp/2x/T4Buf/1dNTP/PreCRMix		
T4/Amp	58	15.81	160	Promega	ThermoT4/3x/Amp/6x/T4Buf/1dNTP/PreCRMix		

Supplementary Table 2.3: Customized SMRT-adapter sequences in IDT compatible format.

Barcode Name	Sequence	Barcode Sequence
SMRT-A_bc-none	/5Phos/CTG TCT CTT ATA CAC ATC TAT CTC TCT CTT TTC CTC CTC CTC CGT TGT TGT TGT TGA GAG AGA TAG ATG TGT ATA AGA GAC AG	AGATGTGTATAAGAGACAG
SMRT-A_bc001	/5Phos/CTG TCT CTT ATA CAC ATC TTT CTT CCG ATC TCT CTC TTT TCC TCC TCC TCC GTT GTT GTT GTT GAG AGA GAT CGG AAG AAA GAT GTG TAT AAG AGA CAG	CGGAAGAAAGATGTGTATAAGAGACAG
SMRT-A_bc003	/5Phos/CTG TCT CTT ATA CAC ATC TTT CCA CAC ATC TCT CTC TTT TCC TCC TCC TCC GTT GTT GTT GTT GAG AGA GAT GTG TGG AAA GAT GTG TAT AAG AGA CAG	GTGTGGAAAGATGTGTATAAGAGACAG
SMRT-A_bc006	/5Phos/CTG TCT CTT ATA CAC ATC TTT GTC GCA ATC TCT CTC TTT TCC TCC TCC TCC GTT GTT GTT GTT GAG AGA GAT TGC GAC AAA GAT GTG TAT AAG AGA CAG	TGCGACAAAGATGTGTATAAGAGACAG
SMRT-A_bc010	/5Phos/CTG TCT CTT ATA CAC ATC TTT AGC TGC ATC TCT CTC TTT TCC TCC TCC TCC GTT GTT GTT GTT GAG AGA GAT GCA GCT AAA GAT GTG TAT AAG AGA CAG	GCAGCTAAAGATGTGTATAAGAGACAG
SMRT-A_bc011	/5Phos/CTG TCT CTT ATA CAC ATC TTC CTA AGG ATC TCT CTC TTT TCC TCC TCC TCC GTT GTT GTT GTT GAG AGA GAT CCT TAG GAA GAT GTG TAT AAG AGA CAG	CCTTAGGAAGATGTGTATAAGAGACAG
SMRT-A_bc012	/5Phos/CTG TCT CTT ATA CAC ATC TTC CGT TGT ATC TCT CTC TTT TCC TCC TCC TCC GTT GTT GTT GTT GAG AGA GAT ACA ACG GAA GAT GTG TAT AAG AGA CAG	ACAACGGAAGATGTGTATAAGAGACAG
SMRT-A_bc013	/5Phos/CTG TCT CTT ATA CAC ATC TTC GAA TCG ATC TCT CTC TTT TCC TCC TCC TCC GTT GTT GTT GTT GAG AGA GAT CGA TTC GAA GAT GTG TAT AAG AGA CAG	CGATTCGAAGATGTGTATAAGAGACAG
SMRT-A_bc014	/5Phos/CTG TCT CTT ATA CAC ATC TTC ACT GTG ATC TCT CTC TTT TCC TCC TCC TCC GTT GTT GTT GTT GAG AGA GAT CAC AGT GAA GAT GTG TAT AAG AGA CAG	CACAGTGAAGATGTGTATAAGAGACAG

Barcode Name	Sequence	Barcode Sequence
SMRT-A_bc015	/5Phos/CTG TCT CTT ATA CAC ATC TTG CAG GAT ATC TCT CTC TTT TCC TCC TCC TCC GTT GTT GTT GTT GAG AGA GAT ATC CTG CAA GAT GTG TAT AAG AGA CAG	ATCCTGCAAGATGTGTATAAGAGACAG
SMRT-A_bc016	/5Phos/CTG TCT CTT ATA CAC ATC TTA TGG CGT ATC TCT CTC TTT TCC TCC TCC TCC GTT GTT GTT GTT GAG AGA GAT ACG CCA TAA GAT GTG TAT AAG AGA CAG	ACGCCATAAGATGTGTATAAGAGACAG
SMRT-A_bc017	/5Phos/CTG TCT CTT ATA CAC ATC TTA CCG ACT ATC TCT CTC TTT TCC TCC TCC TCC GTT GTT GTT GTT GAG AGA GAT AGT CGG TAA GAT GTG TAT AAG AGA CAG	AGTCGGTAAGATGTGTATAAGAGACAG
SMRT-A_bc018	/5Phos/CTG TCT CTT ATA CAC ATC TTA CAA GCC ATC TCT CTC TTT TCC TCC TCC TCC GTT GTT GTT GTT GAG AGA GAT GGC TTG TAA GAT GTG TAT AAG AGA CAG	GGCTTGTAAGATGTGTATAAGAGACAG
SMRT-A_bc019	/5Phos/CTG TCT CTT ATA CAC ATC TCT GAC CAA ATC TCT CTC TTT TCC TCC TCC TCC GTT GTT GTT GTT GAG AGA GAT TTG GTC AGA GAT GTG TAT AAG AGA CAG	TTGGTCAGAGATGTGTATAAGAGACAG
SMRT-A_bc020	/5Phos/CTG TCT CTT ATA CAC ATC TCC TCT CTA ATC TCT CTC TTT TCC TCC TCC TCC GTT GTT GTT GTT GAG AGA GAT TAG AGA GGA GAT GTG TAT AAG AGA CAG	TAGAGAGGAGATGTGTATAAGAGACAG
SMRT-A_bc021	/5Phos/CTG TCT CTT ATA CAC ATC TCC TGT AAC ATC TCT CTC TTT TCC TCC TCC TCC GTT GTT GTT GTT GAG AGA GAT GTT ACA GGA GAT GTG TAT AAG AGA CAG	GTTACAGGAGATGTGTATAAGAGACAG
SMRT-A_bc022	/5Phos/CTG TCT CTT ATA CAC ATC TCC GCA TAA ATC TCT CTC TTT TCC TCC TCC TCC GTT GTT GTT GTT GAG AGA GAT TTA TGC GGA GAT GTG TAT AAG AGA CAG	TTATGCGGAGATGTGTATAAGAGACAG
SMRT-A_bc023	/5Phos/CTG TCT CTT ATA CAC ATC TCA AGT GGA ATC TCT CTC TTT TCC TCC TCC TCC GTT GTT GTT GTT GAG AGA GAT TCC ACT TGA GAT GTG TAT AAG AGA CAG	TCCACTTGAGATGTGTATAAGAGACAG

Barcode Name	Sequence	Barcode Sequence
SMRT-A_bc024	/5Phos/CTG TCT CTT ATA CAC ATC TGT GCA TTC ATC TCT CTC TTT TCC TCC TCC TCC GTT GTT GTT GTT GAG AGA GAT GAA TGC ACA GAT GTG TAT AAG AGA CAG	GAATGCACAGATGTGTATAAGAGACAG
SMRT-A_bc025	/5Phos/CTG TCT CTT ATA CAC ATC TGG CTT CAT ATC TCT CTC TTT TCC TCC TCC TCC GTT GTT GTT GTT GAG AGA GAT ATG AAG CCA GAT GTG TAT AAG AGA CAG	ATGAAGCCAGATGTGTATAAGAGACAG
SMRT-A_bc026	/5Phos/CTG TCT CTT ATA CAC ATC TGG AAC TAC ATC TCT CTC TTT TCC TCC TCC TCC GTT GTT GTT GTT GAG AGA GAT GTA GTT CCA GAT GTG TAT AAG AGA CAG	GTAGTTCCAGATGTGTATAAGAGACAG
SMRT-A_bc027	/5Phos/CTG TCT CTT ATA CAC ATC TGA CGT TAG ATC TCT CTC TTT TCC TCC TCC TCC GTT GTT GTT GTT GAG AGA GAT CTA ACG TCA GAT GTG TAT AAG AGA CAG	CTAACGTCAGATGTGTATAAGAGACAG
SMRT-A_bc028	/5Phos/CTG TCT CTT ATA CAC ATC TGA GTG TCT ATC TCT CTC TTT TCC TCC TCC TCC GTT GTT GTT GTT GAG AGA GAT AGA CAC TCA GAT GTG TAT AAG AGA CAG	AGACACTCAGATGTGTATAAGAGACAG
SMRT-A_bc029	/5Phos/CTG TCT CTT ATA CAC ATC TGA AGA AGG ATC TCT CTC TTT TCC TCC TCC TCC GTT GTT GTT GTT GAG AGA GAT CCT TCT TCA GAT GTG TAT AAG AGA CAG	CCTTCTTCAGATGTGTATAAGAGACAG
SMRT-A_bc030	/5Phos/CTG TCT CTT ATA CAC ATC TAA CAC CTC ATC TCT CTC TTT TCC TCC TCC TCC GTT GTT GTT GTT GAG AGA GAT GAG GTG TTA GAT GTG TAT AAG AGA CAG	GAGGTGTTAGATGTGTATAAGAGACAG

2.13. Supplementary File 1 – Library and sequencing statistics

Supplementary file containing library preparation details and sequencing statistics for SMRT-Tag and SAMOSA-Tag datasets included in Chapter 3.

Chapter 3: Native single-molecule chromatin profiling of primary cells via direct library preparation

3.1. Abstract

Recently, we and others have mapped nucleosomes positions on single molecules with base-pair resolution using non-destructive adenine methyltransferase (m^6dAse) footprinting and 3rd generation sequencing[199,233]. However, mapping single-molecule chromatin structure in input-limited clinical samples has been difficult because m^6dAse footprinting has high input requirements. Here, we overcome this constraint by performing m^6dAse footprinting and library preparation directly on bead-immobilized cells or nuclei. Our improved approach, ConA+SAMOSA-Tag can resolve multimodal chromatin fibers from as few as ~10,000 cells from primary human samples without PCR. When applied to patient-derived xenograft models of prostate cancer progression, ConA+SAMOSA-Tag reproduces previously characterized patterns of histone eviction associated with metastasis. We also use ConA+SAMOSA-Tag to address heterogeneity in primary samples – profiling leukocyte subpopulations from peripheral mononuclear blood cells and revealing a concordant decrease in nucleosome repeat lengths and m^5dCpG methylation during B cell maturation. Our results demonstrate the value of single molecule studies in dissecting biological phenomena and increases their accessibility for the broader research community.

3.2. Inherent limitations in single-molecule chromatin profiling

Eukaryotic genomes are wrapped around multicomponent histone complexes, termed nucleosomes, and condensed into chromatin fibers within the cell nucleus[1]. A suite of histone modifying enzymes and ATP-driven remodelers work tirelessly to maintain nucleosomes in regularly spaced arrangements, modulating the accessibility of underlying DNA to gene regulatory machinery such as transcription factors and RNA polymerase II[61]. Human diseases ranging from autoimmune disorders to cancer syndromes are often characterized by fibers with increased accessibility at cis-regulatory elements that drive aberrant transcriptional programs[113,268]. To maintain these irregular fiber architectures,

chromatin remodeler genes are often mutated, epigenetically silenced, or transcriptionally upregulated[80,269].

Understanding how specific fiber architectures are associated with disease progression could provide new therapeutically addressable targets. However, existing assays for profiling chromatin fibers such as micrococcal nuclease digestion (MNase-seq) and the assay for transposase accessible chromatin (ATAC-seq) rely on enzymatic fragmentation of nucleosome-free DNA across millions of cells, fundamentally obscuring nucleosome positioning in *cis*[169]. This ensemble averaging effect is most pronounced in highly heterogeneous patient-derived samples such as tumors with distinct subpopulations responsible for cell proliferation and drug resistance.

One way of avoiding averaging is to study single molecules directly. We and others have coupled non-destructive enzymatic labelling of accessible DNA with native high-throughput 3rd generation sequencing to measure nucleosome footprints, endogenous m^5dCpG methylation, and primary sequence on long (2 – 20 kilobases) single DNA molecules[198–200]. However, large amounts of input material are required for both 1) *in situ* footprinting of nuclei using adenine methyltransferase (m^6dAse) and 2) preparing footprinted molecules for highly accurate HiFi sequencing on the Single-molecule Real-Time (SMRT) platform from Pacific Biosciences (PacBio). To address the second issue, we recently demonstrated that Tn5 transposase loaded with customized SMRT sequencing adapters (SMRT-Tn5) can simultaneously fragment and prepare HiFi-compatible libraries from ~20,000 nuclei directly – a protocol we term SAMOSA-Tag (Chapter 2). Yet the first hurdle remains – typically ~200k– 1M nuclei must be m^6dAse footprinted to guarantee enough DNA is recovered despite sample handling losses. Further, no existing m^6dAse footprinting strategy, including ours, can address sample heterogeneity by assigning single molecules back to their cell of origin, as in other cell-resolved assays like single cell ATAC-seq (scATAC-seq)[270].

We were inspired by recent methods that can assay histone marks in as few as ~100 cells after immobilization on paramagnetic beads coated with concanavalin A (ConA)[173,175,176], a lectin that can bind glycosylated proteins and lipids in eukaryotic membranes[271]. Here, we augmented SAMOSA-Tag with

ConA bead-based immobilization and demonstrated that on-bead m^6dAse footprinting and tagmentation of ~10,000 cells or nuclei can produce high quality single-molecule accessibility measurements concordant with previous findings in mouse embryonic stem cells and primary human samples. We also demonstrate how combining ConA beads and SAMOSA-Tag with multiplexed cell sorting enables characterization of cell-type specific changes in chromatin fiber usage. Finally, we address existing limitations in native sequencing and propose further method optimizations that could improve overall library quality and yield. Overall, our work highlights the key advantages of ConA-immobilized SAMOSA-Tag for single-molecule profiling studies on low-input primary cell populations.

3.3. Concanavalin A beads are compatible with m6dAse footprinting

To enable single-molecule multimodal chromatin profiling directly from low input samples, we modified our original SAMOSA-Tag protocol to utilize nuclei or cells immobilized on ConA beads. Our improved approach, which we term ConA+SAMOSA-Tag (ConA+ST, workflow shown in **Figure 3.1a**), relies on magnet-based precipitation to successively transfer immobilized nuclei between m^6dAse footprinting and *in situ* tagmentation reactions, with minimal loss. Nuclei are subsequently lysed on-bead, tagmented fragments recovered, and HiFi libraries generated using our previously described low-input library preparation strategy.

To confirm that bead-immobilization was compatible with both footprinting and tagmentation, we subjected ~500k freshly extracted nuclei from K562 cells to ConA+ST. We observed that bead-immobilized nuclei incubated at 37ºC in m^6dAse reaction buffer displayed no significant aggregates (**Figure 3.1b**, left), and addition of m^6dAse did not result in partially lysed or evicted nuclei (**Figure 3.1b**, right). Immobilized footprinted nuclei treated with varying concentrations of Tn5 (0.15 – 18.8 picomoles [pmol]) loaded with customized SMRT sequencing adapters (SMRT-Tn5) produced a tunable range of long fragments, allowing for precise control of library size as in our SAMOSA-Tag method (**Figure 3.1c**). However, library yield was significantly reduced after on-bead nuclei lysis and DNA recovery, with lower Tn5 amounts most severely impacted. We speculated that larger molecules were more prone to adsorbing

to ConA beads, as has been observed in other studies[241], and resolved this issue by introducing solid-phase reversible immobilization (SPRI) paramagnetic beads into the post-lysis DNA mixture. Extracting tagmented fragments using both magnetic beads together enabled near-quantitative recovery.

We then confirmed ConA+ST libraires were compatible by HiFi sequencing by lightly sequencing one library as part of a multiplexed SMRTcell. We generated 194,498 molecules that were long (median: 4,351 bp, mean ± standard deviation [s.d]: 5,133 ± 3471 bp, **Figure 3.1d**), and > 99% successfully aligned to the human genome with high accuracy (empirical single-molecule average Q score: Q30+). Further, the library achieved the expected percentage representation in the multiplexed cell, indicating no significant impediments to sequencing efficiency. We therefore concluded our modified ConA+ST protocol could m^6dA footprint and prepare nuclei into HiFi libraries, with no impact on library quality.

3.4. Concanavalin A SAMOSA-Tag enables library preparation from as few as ~5,000 nuclei

We next sought to determine the lower input limit for ConA+ST. We collected E14 mouse embryonic stem cells (mESCs), extracted nuclei, and generated a dilution series of ConA+ST libraries from as few as ~5,000 (equivalent to ~20 ng gDNA) to a maximum of ~500,000 nuclei (~2.5 µg gDNA) per reaction (experiment schematic shown in **Figure 3.2a**). When compared to SAMOSA-Tag control libraries prepared from the same input material, ConA+ST -Tag libraries had overall lower yields (mean ± s.d.: 25.7 ± 5.6% vs. 36.6 ± 2.3%, **Figure 3.1b**). However, as few as ~5,000 input nuclei yielded enough library material to saturate one entire SMRTcell (Chapter 3.13 – **Supplementary File 2**), conservatively a ~20-fold reduction when compared to bulk m^6dAse footprinting requirements of ~200k – 1M nuclei. We multiplexed and sequenced a subset of libraries on one SMRTcell and obtained 30,000 – 120,000 molecules per condition that were on average shorter than equivalent SAMOSA-Tag libraries (median: 1,525 bp vs. 2,786 bp, **Figure 3.2c**, Chapter 3.13 – **Supplementary File 2**). Decreased molecule length may stem from differences in SMRT-Tn5 amount (9.4 pmol vs. 18.8 pmol for ConA+ST vs. SAMOSA-Tag) as well as steric access to nuclei after bead immobilization.

3.5. Concanavalin A SAMOSA-Tag reproduces known fiber enrichment profiles in mouse embryonic stem cells

The chromatin landscape of mouse embryonic stem cells has been extensively characterized through both conventional short-read based assays [272] and single-molecule m^{6}dAse footprinting[201,233]. We therefore performed a series of comparisons between single-molecule chromatin profiling data produced using ConA+ST and reference data produced using SAMOSA-Tag to ascertain which input amounts reproduced expected fiber architectures.

After processing molecules through our established SAMOSA pipeline to call single-molecule accessibility (Chapter , we first visualized the distribution of all inaccessible region sizes, corresponding to the footprints of DNA-bound proteins such as nucleosomes (**Figure 3.2d)**. ConA+ST libraries prepared using lower-input amounts (5k, 10k, 20k nuclei) yielded footprint distributions most similar to SAMOSA-Tag while higher-input libraries (40k, 80k, 500k nuclei) had distinctly shorter footprints (primary peak: ~145 bp vs. ~110 bp). While both were generally concordant with the expected mode murine nucleosomal footprint size of ~125 bp[201], the shorter footprints in higher-input libraries suggested inherent differences in the underlying chromatin fibers. To determine if this was the case, we clustered molecules from all libraires by the regularity of their single-molecule accessibility profiles (**Chapter 3.11 – Methods**). The resulting clusters, termed "fiber types", contained molecules with similar nucleosome repeat length (NRLs) – the average distance between adjacent dyads on a single molecule. We and visualized the average accessibility of each fiber type when molecules were aligned by their 5' ends, (**Figure 3.1e**) and observed 6 were regular, with NRLs ranging from 193 bp to 211 bp, and two irregular (IR1, IR2), with nearly random offsets between nucleosomes.

While the lower-input ConA+ST libraries were comprised of roughly similar amounts of each fiber type as our SAMOSA-Tag control, increasing input amounts favored an overrepresentation of regular fiber types such as NRL193 and NRL199 by as much as 20% (**Figure 3.2f**). We further stratified fibers by whether they mapped to domains marked with different histone posttranslational modifications

("epigenomic domains") and observed fiber types had distinct patterns of enrichment that were strongly correlated with input amount (**Figure 3.2g)**. Surprisingly, while long NRL fibers (NRL211) were consistently depleted (~4 fold) in constitutively heterochromatic H3K9me3-marked domains across all input amounts and our SAMOSA-Tag control, most other fiber type domain-level enrichments were inverted or lost between higher-input libraires and SAMOSA-Tag. A distinct example was the strong and selective (~2-fold) enrichment of IR1 irregular fibers in H3K9me3-marked domains in SAMOSA-Tag and lower-input libraries, a trend which was lost in higher-input libraries. Similarly, NRL193 fibers, which were depleted across all domains including H3K27ac-marked and DNase I hypersensitivity sites (DHS) in SAMOSA-Tag and lower-input libraries, were instead enriched in higher-input libraries. Treating fiber type enrichments as a characteristic library fingerprint revealed the 10k input condition most closely matched SAMOSA-Tag data (Pearson's $r = 0.94$, $p < 2\text{x}10^{-16}$), while the 500k input condition was least similar (Pearson's $r = 0.72$, $p < 2\text{x}10^{-16}$).

To further confirm that fibers captured by lower-input ConA+ST libraries were sampled from the genome similarly to SAMOSA-Tag, we examined the insertional preferences of SMRT-Tn5 at both preferred sites and genome wide. At transcription start sites (TSSs), we found ConA+ST insertion rates were similar to our SAMOSA-Tag control. Only the 500k ConA+ST library had an elevated insertion rate (~62 insertions / 1M molecules vs. SAMOSA-Tag, ~18 insertions / 1M molecules, **Figure 3.2h**) and TSS insertions comprised only slightly more of total molecules sequenced (mean ± s.d.: 4.5 ± 0.4% across ConA+ST vs. 4.1 ± 0.2%, **Figure 3.2i**). However, looking genome-wide we found that IR1 fibers in higher-input conditions were strongly sequence-dissimilar when compared to equivalent fibers from SAMOSA-Tag, as measured by the cosine similarity of normalized 9-mer spectra (cosine distance 0.23 – 0.42, **Figure 3.2j**). Specifically, IR1 fibers were over-enriched (> 7-fold) for 9-mers containing variants of the sequence "AACCCT", a telomere-like hexanucleotide repeat that is relatively infrequent and located predominantly in intronic and intergenic regions of the mammalian genome[273,274] (**Figure 3.2k,** inset). In contrast, lower-input libraries were highly sequence-similar across fiber types (cosine distance < 0.05, **Figure 3.2j**), confirming that ConA+ST applied to ~5k – 20k nuclei can reproduce the expected

chromatin fiber types detected in bulk m^{6}dAse footprinted nuclei at both their expected frequencies and sequence composition.

3.6. Concanavalin A SAMOSA-Tag is compatible with difficult-to-handle clinical samples

We next asked if ConA+ST could reduce sample loss when profiling primary human samples. We generated multiple ConA+ST libraries using ~10,000 nuclei extracted from paired primary and metastatic prostate cancer patient-derived xenografts (PDX)[257] (6 and 8 replicates respectively, Chapter 3.13 – **Supplementary File 2**) that we had previously characterized using bulk SAMOSA-Tag (experiment schematic in F**igure 3.3a**). Because nuclei extraction itself can contribute towards sample loss, we also evaluated whether applying ConA+ST directly to metastatic PDX cells (Chapter 3.11 – **Methods**) could improve library yield. In both cases, individually footprinted ConA+ST libraries displayed similar or marginally reduced performance when compared to bulk SAMOSA-Tag (**Figure 3.3b**), but critically avoided sample loss associated with bulk footprinting (Chapter 3.13 – **Supplementary File 2**). Sequenced molecules were appreciably long (**Figure 3.3c**), had similar rates of human-aligning reads as their bulk SAMOSA-Tag equivalents (~32% and ~92% primary and metastasis respectively, Chapter 3.13 – **Supplementary File 2**), and resolved nucleosome footprints that matched the expected size of ~135 – 147 bp (**Figure 3.3d**).

To evaluate whether ConA+ST libraries could capture metastasis-associated changes in nucleosome regularity, we again defined a set common regular and irregular fiber types via unsupervised clustering of single-molecule accessibility profiles (**Figure 3.3e**). Then, for each method (ConA+ST and SAMOSA-Tag), we independently calculated the differential fiber type usage (DFU) between metastatic and primary PDX cells across a previously annotated set of normal prostate epigenomic domains[260] including promoters, transcriptionally active genes, enhancers and constitutive heterochromatin (schematic of calculation in **Figure 3.3f**). Comparing DFU estimates between methods, we found that ConA+ST largely reproduced key findings from our SAMOSA-Tag dataset including strong differential upregulation of IR1 irregular fibers at enhancers and constitutive heterochromatin and decreased usage of

shorter NRL182 fibers at genic enhancers (IR1, green; NRL182 red; **Figure 3.3g**). ConA+ST DFU estimates for most other fiber and domain combinations had marginal changes compared to SAMOSA-Tag (< 2-fold) and were strongly correlated between methodologies, including cell-direct ConA+ST, in both primary and metastatic PDX samples (Pearson's $R > 0.79$, $p < 1x10^{-16}$, **Figure 3.3h**).

Interestingly, DFU of NRL191 fibers across all epigenomic domains was estimated consistently lower (~4-fold) in ConA+ST libraries as compared to SAMOSA-Tag. We attributed this to the lower representation of NRL191 fibers in metastatic ConA+ST libraries (**Figure 3.3h**, 2nd panel, NRL191 orange). Since our mESC titration experiments suggested increasing amounts of input favored sampling of regular fiber types, we speculated an inverse effect could also occur where decreased input disfavored regular fibers. A slight decrease in regular fiber representation in ~5k and ~10k mESC libraries partially supports this hypothesis (yellow & light green bars, Δ = -0.05 and -0.04 for NRL193 and NRL199 respectively, **Figure 3.2f**). Together, this highlights the importance of maintaining consistent amounts of SMRT-Tn5 across ConA+ST libraries and suggests future comparisons across different input amounts should account for different genome-wide insertional preferences.

3.7. Evaluating differential fiber usage in B cell maturation

Primary human tissues are comprised of phenotypically diverse cell populations with distinct chromatin accessibility and transcriptional states[275,276]. Characterizing the functional role of these subpopulations using cell-resolved approaches has provided invaluable insights into tumor evolution[277], drug resistance[278], and cell differentiation[279].

We therefore asked if our ConA-ST protocol could be used to identify cell-type specific changes in single-molecule chromatin accessibility from a single patient-provided sample. We sampled peripheral blood mononuclear cells (PBMCs) from a healthy donor, enriched for multiple leukocyte subtypes including memory and naïve B, CD4+ and CD8+ T cells, mononuclear cells, and rare bone-derived dendritic cells (DCs) using multiplexed FACS, and prepared 4 x ~20k cell-direct ConA+ST libraries per sorted population (experiment schematic in **Figure 3.4a**). Library yields were on average 23.7 ± 8.8% and

generally consistent across subtypes (one-way ANOVA, $p = 0.225$), apart from memory CD4+ T cells for which two replicate libraries failed (**Figure 3.4b)**.

B cell maturation is associated with well-characterized hypomethylation of lineage-specific enhancers and transcription factor binding sites in precursor and naïve B cells[280]. At germinal centers in lymph organs, somatic hypermutation and terminal fate selection relies on an intricate network of successive transcription factor-mediated gene activation and repression. Functional studies have identified chromatin decompaction globally marked by H3K27ac[281] and nucleosome eviction at key enhancers via Brg1[282,283] as essential for maturation. We therefore speculated our B cell ConA+ST libraries could provide the first map of maturation-associated changes in single-molecule accessibility and m^5dCpG methylation.

We multiplexed and sequenced memory and naïve B cell replicate ConA+ST libraries on one SMRTcell. Though we only obtained between ~2,100 – 18,000 reads per condition, individual molecules were longer than previous ConA+ST libraries (median: 6,205 bp and 6,255 bp, naïve and memory B respectively, **Figure 3.4c**) and could be assigned to distinct regular and irregular fiber types (**Figure 3.4d**. Though a limited number of single-molecule m^6dAse footprinting datasets have been generated in humans, both memory and naïve B cells lacked shorter NRL fibers (*e.g.* NRL176) previously observed in K562 cells[199], osteosarcoma cells, and both primary and metastasis PDX cells. Instead, both B cell subtypes contained irregular fibers (IR1 – IR3 ranging from hyper- to hypoaccessible), that had a marked increase in subnucleosomal footprints (IR1 and IR2 mode subnucleosomal footprint size: 25 bp and 18 bp respectively, **Figure 4.3f**). The two most hyperaccessible irregular fiber types, IR and IR2, (average per base accessibility, 70.7 ± 6.7% and 59.3 ± 6.8%) were also moderately depleted after maturation (1.14- and 1.40-fold, $q < 7.10 \times 10^{-11}$)

We observed that maturation also resulted in a strong reduction (1.92-fold, $q = 9.99 \times 10^{-159}$) in long NRL211 fibers and a comparative increase in shorter NRL192 and NRL190 fibers genome-wide (1.70- and 1.71-fold, $q < 2.35 \times 10^{-60}$). Though we lacked the coverage to further determine differential fiber type usage stratified by genomic domain, we noted that long NRL fibers in osteosarcoma cells were

notably CpG sparse (**Figure 2.4f**). Since nearly 30% of all CpG motifs are differentially methylated, and often demethylated, during B cell maturation[284–286], we speculated fiber type enrichments may be coordinated with demethylation. Thus, we asked if hypomethylation occurred preferentially at the shorter NRL fibers. Surprisingly, we found the opposite trend – while the average m^5dC levels of all regular fibers decreased during the transition from naïve to memory subtypes, the effect was most pronounced at longer NRL fibers including NRL207 and NRL211, independent of CpG density (mean ± s.d.: $\Delta_{\{NRL207,NRL211\}}$ = -0.077 ± 0.019 reduction vs. $\Delta_{\{NRL190,NRL192\}}$ = 0.038 ± 0.013 reduction, Mann-Whitney U $p = 2.08 \times 10^{-2}$, **Figure 3.4g**). A similar effect was also observed for low-CpG density irregular fibers. Together, this suggests maturation-associated hypomethylation is comprised of two parts – 1) moderate global hypomethylation and enrichment of short-NRL fibers, and 2) depletion and strong hypomethylation of long-NRL and CpG-poor irregular fibers. Since changes in accessibility imply changes in transcription factor binding, deeper sequencing could address whether either process influences B-cell lineage-dependent transcription factors. However, our preliminary data alone highlight the power of ConA+ST and multimodal single-molecule analyses in further dissecting epigenomic phenomena previously characterized using bulk assays.

3.8. Addressing sample quality may improve SAMOSA-Tag

We sought to understand why ConA+ST libraries constructed from naïve and memory B cells had reduced sequencing performance (Chapter 3.13 – **Supplementary File 2**, column "CCS") despite reasonable library yields after exonuclease-mediated depletion of non-sequenceable molecules (**Figure 3.4b**). HiFi sequencing is especially sensitive to nicks, breaks and a-basic sites that can prematurely halt the rolling-circle polymerization of the SMRT sequencing polymerase[217]. Thus, we assessed the degree to which sequencing of ConA+ST libraries terminated early by comparing the length of each entire polymerase-sequenced molecule ("polymerase read") to the length of the estimated "subread", the sequence between the two SMRT sequencing adapters. We found that for a significant proportion of library molecules, the SMRT polymerase was unable to complete even one sequencing pass and instead

terminated mid-read (diagonal line, **Supplementary Figure 3.1a**). Combined with the observed low utilization of the SMRTcell (Chapter 3.13 – **Supplementary File 2**, column "P1 (%)"), we hypothesized that endogenous damage was increased in ConA+ST libraries.

To determine if DNA damage stemmed from ConA bead manipulation, we prepared HiFi libraries from genomic DNA (gDNA) extracted directly from the sorted cell populations from our healthy donor using both our low-input library preparation approach, SMRT-Tag, and the commercially available Template Preparation Kit v.2.0 (TPK2.0) from PacBio[232]. In both cases, the resulting libraries failed to sequence well (P1 efficiency, ~1%) regardless of fragment length (**Supplementary Figure 3.2c**), suggesting damage had accrued to a significant degree in the input material itself. We confirmed this was the case by assessing the integrity of our extracted naïve and memory B cell gDNA via automated electrophoresis. Compared to both control high quality genomic DNA and known degraded samples, naïve and B cell gDNA was on average shorter and had a lower DNA integrity number (DIN) (**Supplementary Figure 3.2a,c**). This issue extended to monocyte gDNA as well, suggesting upstream sampling handling in general was at fault. We additionally ruled out our DNA extraction procedure as introducing breaks or nicks, as the commercially available kits from New England Biolabs and Qiagen that we used in this study produced large quantities of high DIN gDNA when applied to ~250k cells from an immortalized B lymphoblast cell line (GM12878, **Supplementary Figure 3.2b**).

One solution implemented by PacBio in their commercially available Template Preparation Kits (version 2.0 and 3.0) is to "repair" gDNA using a cocktail of various nucleases that target damaged DNA bases ("PreCR repair")[287]. Digestion and / or excision of degraded bases leaves behind short gaps that can be efficiently filled in and sealed using a nick-translation DNA polymerase such as *Bst*. Recent efforts capitalizing on this synergistic chemistry have shown repaired patches are < 40 bp long, improving DNA integrity while minimizing erasure of epigenetic modifications[287]. We attempted to rescue our existing ConA+ST B cell libraries using PreCR repair but did not observe an improvement in the DNA-damage induced polymerase termination phenotype (**Supplementary Figure 3.1b**). Libraries prepared using PreCR repaired genomic DNA, using both TPK2.0 or SMRT-Tag, also failed to sequence efficiently.

Hence, we concluded that poor sample handling can irreversibly affect ConA+ST library quality, and, in the absence of amplification to dilute the proportion of molecules containing significant nicks or breaks, resulting in poor sequencing yield. Future studies using ConA+ST should therefore optimize sample handling to produce high quality data, potentially by incorporating changes to buffer conditions developed in the context of nanopore sequencing that help suppress DNases and divalent cations.

3.9. Discussion & Conclusion

We have presented relevant methodological improvements to our SAMOSA-Tag protocol that enable single-molecule m^6dAse footprinting and library preparation directly from low-input cells and nuclei. Our strategy, ConA+ST can produce HiFi-compatible sequencing libraries from as few as ~5,000 – 10,000 nuclei. We validated ConA+ST single-molecule accessibility measurements in mESCs, and showed ConA+ST captures patterns of genome-wide nucleosomal regularity that are highly concordant with SAMOSA-Tag. Critically, applying ConA+ST to precious and difficult-to-handle primary and metastatic PDX samples allowed us to m^6dAse footprint and prepare libraries from ~10,000 nuclei or cells directly, avoiding sample losses from bulk nuclei extraction and methylation. This in turn enabled us probe chromatin fibers in naïve and memory B cells, and deconvolute the well-studied phenotype of maturation-associated global hypomethylation into different trends associated with nucleosome offset distances. Excitingly, this positions ConA+ST to address changes in fiber usage among complex subpopulations from heterogenous tumor biopsies. We envision even rare subpopulations with distinct markers (*i.e.* ~1% or 1000 cells / 1M sorted) could be isolated using FACS, profiled and indexed via ConA+ST, and sequenced effectively in multiplex.

However, ConA+ST is still subject to the common limitations of native sequencing; chiefly, that un-tagemented or unsequenced DNA is effectively lost. Additionally, though we have demonstrated ConA+ST works with both cell models and primary samples, efficient and productive native DNA sequencing requires highly quality damage-free DNA templates. We confirmed that ConA+ST is not likely adding DNA damage to templates, but until a *post hoc* solution exists for recovering damaged

libraries, any experiments using ConA+ST will require significant pre-optimization to confirm isolation, handling and processing of biomaterials does not reduce gDNA integrity. Further while DNA "integrity" and "quality" can be approximately measured by electrophoresis[288], our understanding of how exactly specific damage can be identified and repaired remains lacking. Only a few studies to date have characterized how specific damage sources, including a-basic sites or small patches, interact with 3rd generation sequencing, providing little insight into the correct combination of exo- / endonucleases to remove them[287,289,290]. Thus, we imagine future work in this area will involve the development of chemical and enzymatic methods for sample preservation, as well as new assays for characterizing library quality. A previously explored diagnostic is measuring the fraction of library molecules that support rolling-circle amplification[291–293] – and miniaturized versions of this assay, or others used for assessing DNA quality, could be developed using readily-available microfluidic- or droplet-based readouts[294]. Further methodological improvements including eluting nuclei from ConA beads through competition with other saccharides[295] rather than via lysis could help. These optimizations could also include modulating SMRT-Tn5 titration to maximize the tradeoff between yield and average molecule length.

Nonetheless, we consider ConA+ST a relevant step towards making native 3rd generation sequencing assays competitive with cell-resolved low-input techniques such as CUT&RUN / Tag[176] and scATAC-seq. The use of Tn5 for direct library preparation makes ConA+ST a versatile tool for converting other clinically valuable transposase-mediated techniques, such as spatial methods for mapping clonal composition[277,296], into their single-molecule equivalents. Taken together, we believe ConA+ST will be beneficial to the broader research community by lowering the threshold for single-molecule profiling studies.

3.10. Figures

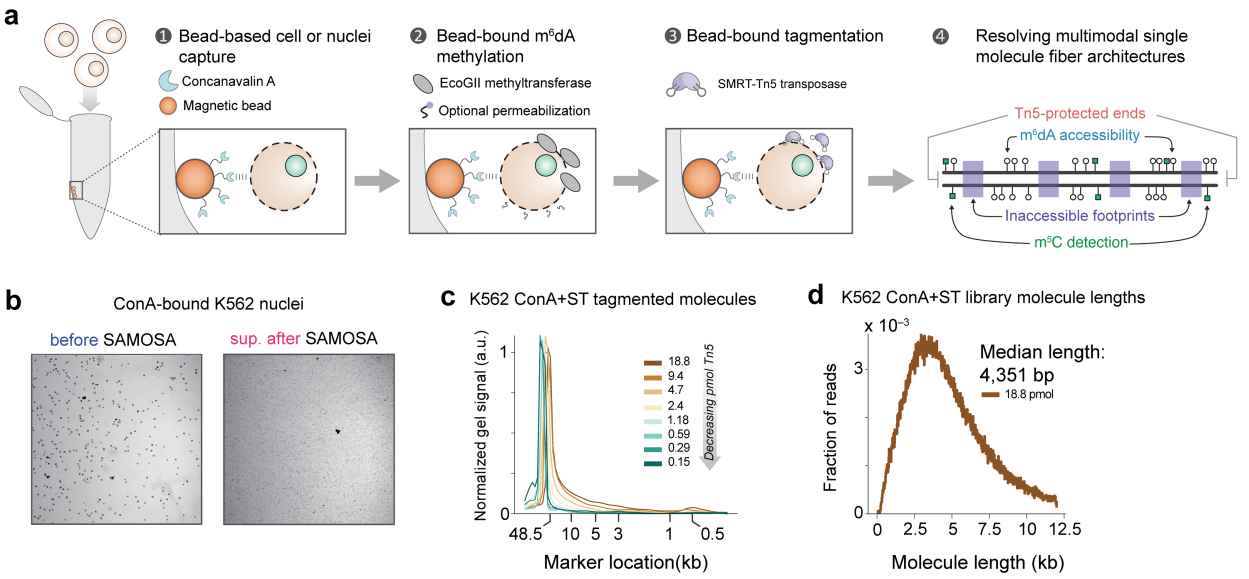


Figure 3.1: Concanavalin A beads are compatible with m^{6}dAse treatment. a.) Schematic overview of the ConA+ST protocol, combining cell or nuclei capture using concanavalin A-coated paramagnetic beads, followed by bead-immobilized EcoGII (m^{6}dAse) treatment and bead-immobilized tagmentation using a custom-loaded Tn5 transposase utilized in both SMRT-Tag and SAMOSA-Tag protocols (Chapter 3.11 – **Methods**). Optional permeabilization steps to improve EcoGII uptake into either nuclei or cells is depicted. Resulting tagmented fragments are repaired using SMRT-Tag gap repair conditions and are compatible with HiFi sequencing. **b.)** Bead-immobilized K562 nuclei visualized after trypan blue staining using bright-field microscopy. Left, bead-immobilized nuclei before m6dAse treatment. Right, supernatant (sup), after treatment. **c.)** Gel electrophoresis traces of ConA+ST libraries indicate tunable tagmentation using SMRT-Tn5. **d.)** Molecule length distribution for a ConA+ST library prepared from K562 nuclei (18.8 pmol Tn5 / 500k nuclei). N = 194,498 molecules sequenced, with median fragment length of 4,351bp.

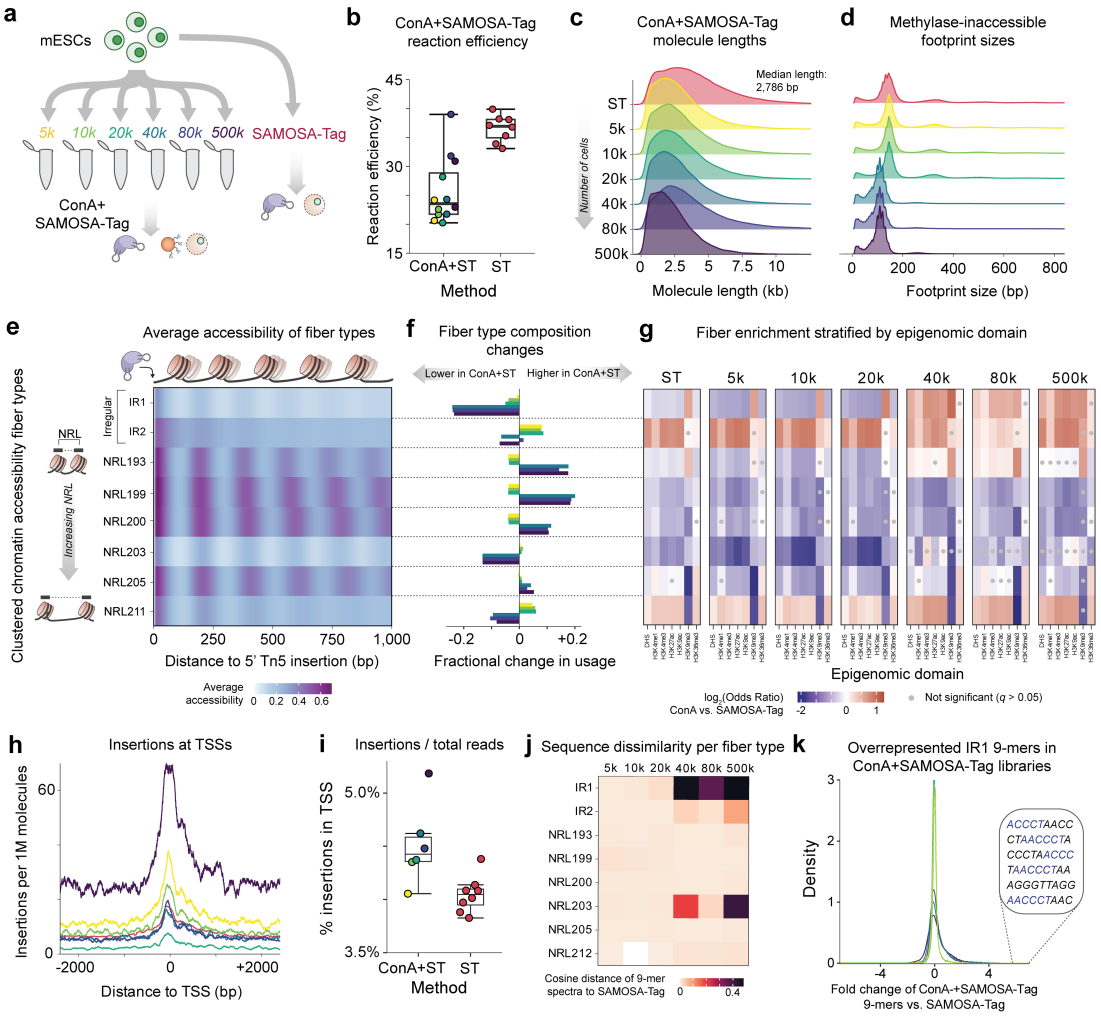


Figure 3.2: ConA+ST resolves multimodal fiber architectures in as few as ~5,000 nuclei and recapitulates expected sampling biases. a.) Schematic overview of ConA+ST pilot experiment. Mouse embryonic stem cell nuclei were subjected to either SAMOSA-Tag in bulk, or serially diluted to a range of input amounts (~5,000, ~10,000, ~20,000, ~40,000, ~80,000, ~500,000) and processed independently via ConA+ST. **b.)** Overall reaction efficiency measured as the percentage of library recovered compared to input material after gap repair and exonuclease cleanup. ST, SAMOSA-Tag. **c.)** Sequenced molecule length and **d.)** methylase-inaccessible footprint distributions for different ConA+ST libraries, as compared to the SAMOSA-Tag control. **e.)** Average accessibility patterns from 5' molecule ends for eight different fiber types, determined by unsupervised clustering. NRL, nucleosome repeat length. Fibers are ordered from shortest to longest NRL. **f.)** Changes in fractional fiber type composition for individual ConA+ST libraries normalized to the SAMOSA-Tag control. **g.)** Per-fiber type enrichment patterns across epigenomic domains. Log2(odds ratio) determined by Fishers Exact shown for each domain (x-axis) and fiber type (y-axis), with red/blue indicating over/under-representation. **h.)** For each ConA+ST library, the rate of Tn5 insertions per million molecules sequenced at transcription start sites (TSSs) and the **i.)** proportion insertions represent of all reads sequenced per library. **j.)** The 9-mer spectrum determined for reads belonging to each fiber type / ConA+ST library compared to the SAMOSA-Tag control using cosine distance. Higher cosine distance (darker) indicates decreased similarity to the SAMOSA-Tag control. **k.)** Fold enrichment of various 9-mers in IR1 fibers from ConA+ST libraries over SAMOSA-Tag background. 9-mers enriched ~7-fold over baseline in ConA+ST libraries with > 40K cells as input are shown, with repetitive sequence highlighted in blue.

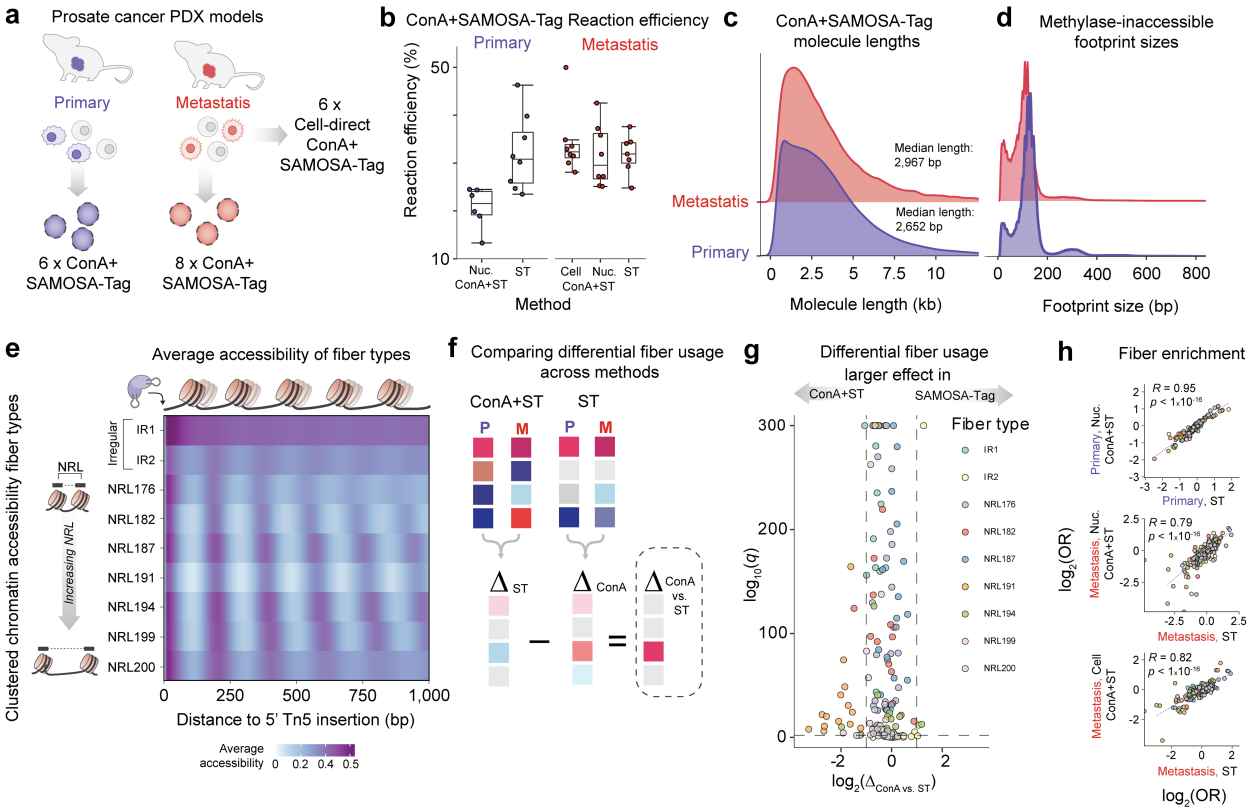


Figure 3.3: ConA+ST applied to patient-derived xenograft models of prostate cancer recapitulate expected fiber usage patterns. a.) Prostate cancer primary and metastasis PDX models previously studied using SAMOSA-Tag were processed using ConA+ST. Cell suspensions from each PDX were sorted to enrich for a live human population. For each model, 6 independent replicates were generated using ~10,000 – 15,000 nuclei per library. For the metastasis model, an additional set of 6 ConA+ST libraries were generated using the sorted cells directly, without nuclei extraction. **b.)** Reaction efficiency for ConA+ST libraries compared to SAMOSA-Tag libraries generated using the same input material. **c.)** Sequenced molecule length and **d.)** methylase-inaccessible footprint distributions for primary and metastasis ConA+ST libraries. **e.)** Average accessibility patterns from molecule ends for nine different fiber types shared across ConA+ST and SAMOSA-Tag primary and metastasis samples, determined by unsupervised clustering, ordered from shortest to longest NRL. **f.)** Schematic overview of comparing fiber usage patterns determined using ConA+ST versus SAMOSA-Tag. For each method, differential fiber usage is calculated (Chapter 3.11 – **Methods**). Differential fiber usage estimates are then compared across methods ($\Delta_{\text{ConA vs. ST}}$). Most estimates should be similar across methods. **g.)** $\Delta_{\text{ConA vs. ST}}$ between ConA+ST versus SAMOSA-Tag across both epigenomic domain and fiber type combinations, colored by fiber type. Values indicate larger effect sizes in ConA+ST (left) or SAMOSA-Tag (right). Dashed vertical lines indicate a method-specific increase or decrease in estimated differential fiber usage of 2-fold. **h.)** Per-method fiber type enrichments ($\log_2$(Odds Ratio)) across epigenomic domains as measured by Fishers Exact test. Enrichments are compared between ConA+ST and SAMOSA-Tag, stratified by PDX model.

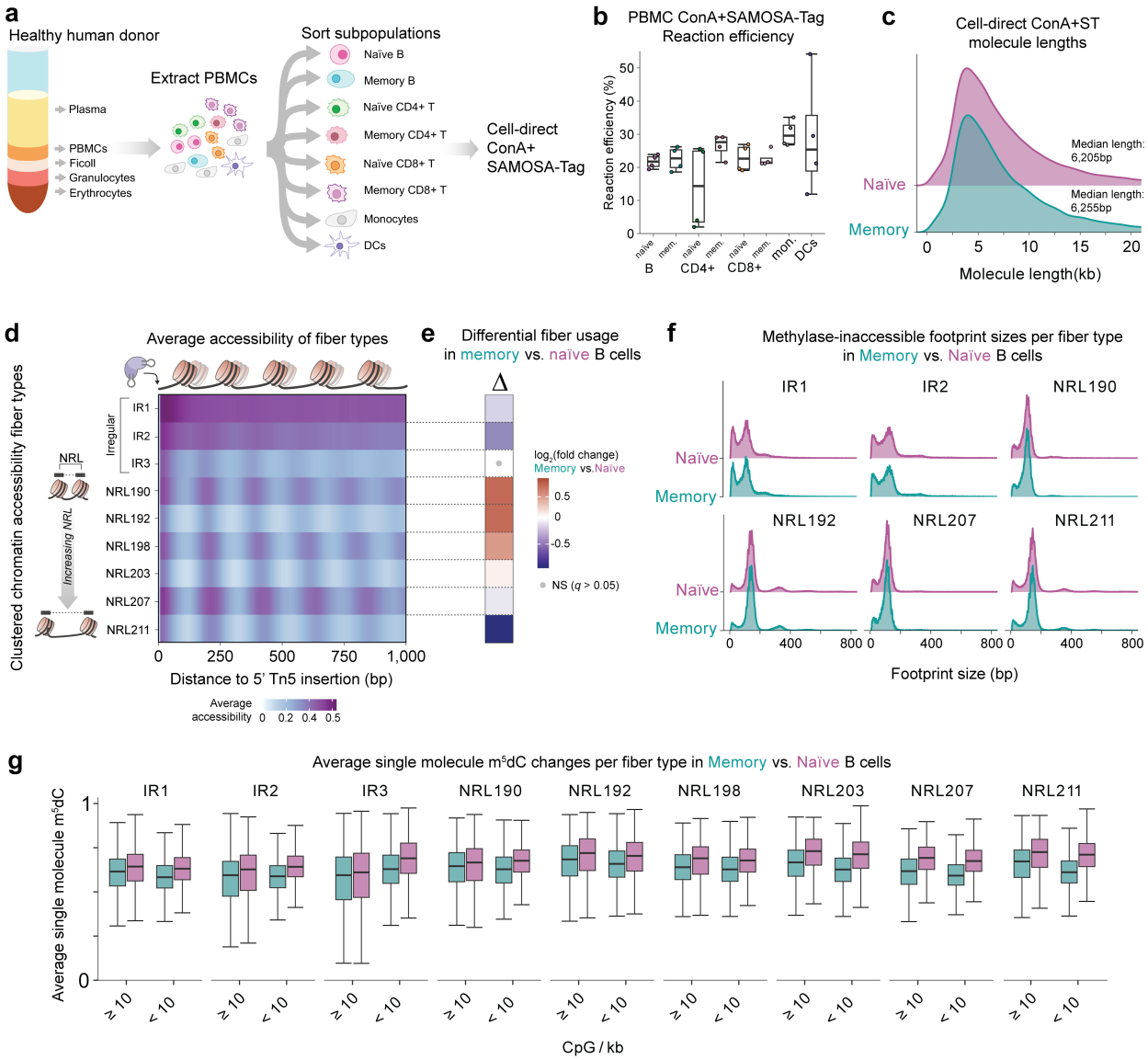
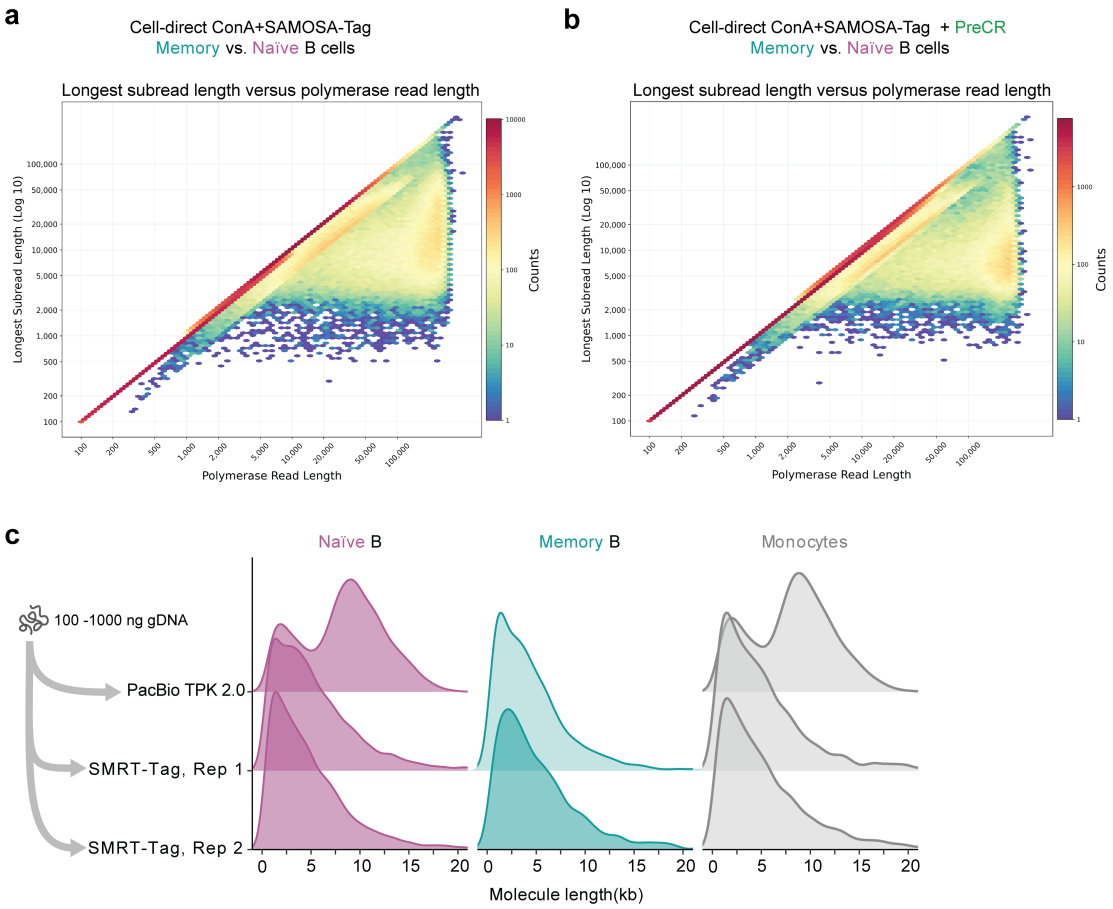
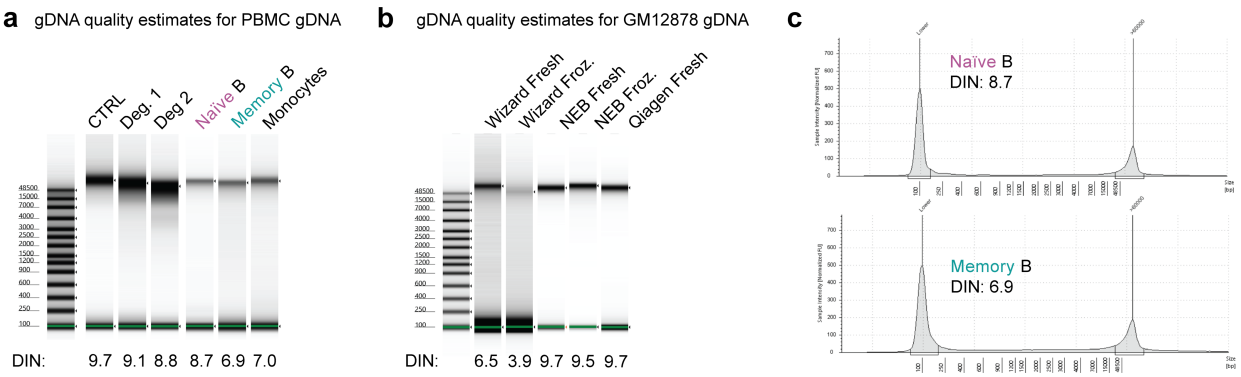


Figure 3.4: ConA+ST reveals concordant increases in nucleosome repeat lengths and m5dC hypomethylation in the course of healthy B cell maturation. a.) Schematic overview of applying ConA+ST to resolve cell-type specific multimodal fibers. Peripheral blood mononuclear cells are extracted from a healthy human donor and sorted into key subpopulations using multiplexed cell sorting. Four ~20,000 cell aliquots are then removed per subpopulation and independently processed using ConA+ST. **b.)** Reaction efficiency for cell-direct ConA+ST libraries prepared from PBMC sub-populations. **c.)** Sequenced molecule lengths for naïve and memory B cell libraries, median molecule lengths 6,205bp and 6,255bp respectively. **d.)** Average accessibility patterns from molecule ends for nine different fiber types shared across cell-direct ConA+ST naïve and memory B cell libraries, determined by unsupervised clustering and ordered from shortest to longest NRL. **e.)** Differential fiber usage between memory B and naïve B cell samples ($\Delta_{\text{memory vs. naïve}}$). Red indicates fiber type usage is significantly increased in memory B cells, blue indicates significantly decreased usage. **f.)** Methylase-inaccessible footprint sizes for naïve and memory B cell samples, stratified by fiber type. Irregular clusters (IR1, IR2) display decreased footprint sizes. **g.)** For each molecule, average m^5dCpG levels. Average levels are then compared between naïve and memory B cells, stratified by both fiber type and single-molecule CpG density. All pairwise comparisons are significant, determined by two-sided Mann-Whitney U test (Chapter 3.11 – **Methods**).



Supplementary Figure 3.1: Sequencing quality metrics for naïve and memory ConA+ST libraries. Longest subread length versus polymerase read length for individual ZMWs is visualized for both **a.)** naïve and memory B cell ConA+ST without PreCR and **b.)** with PreCR treatment. **c.)** Sequenced molecule length distributions for HiFi libraries prepared using SMRT-Tag Template Prep Kit 2.0 with 100 ng to 1000 ng (1 µg) of genomic DNA respectively, extracted from naïve and memory B cells, and monocytes.



Supplementary Figure 3.2: Extracted genomic DNA from PBMCs is degraded. a.) Genomic DNA Tapestation traces for control high quality genomic DNA, degraded genomic DNA, and genomic DNA extracted from naïve and memory B cells, and monocytes in the experiment depicted in **Figure 3.4**. **b.)** Genomic DNA Tapestation traces for genomic DNA extracted using three different commercial kits (Promega Wizard, NEB Monarch, Qiagen MagAttract) from ~250k GM12878 cells in short-term culture conditions. **c.)** Tapestation results from **a.)** visualized as traces.

3.11. Methods

Data availability

SAMOSA-Tag sequencing data for mESCs is available from the Gene Expression Omnibus under accession number GSE225314. SAMOSA-Tag data derived from PDX models are under controlled access to maintain patient privacy. Reasonable requests for other ConA+SAMOSA-Tag data in this study will be fulfilled.

Cell lines and cell culture

E14 mouse embryonic stem cells (mESCs) were gifted from Elphege Nora Lab at UCSF and were routinely tested for mycoplasma via PCR (NEBNext® Q5 2X Master Mix). Feeder-free cultures were maintained on 0.2% gelatin, in KnockOut DMEM 1X (Gibco) supplemented with 10% Fetal Bovine Serum (Phoenix Scientific), 1% 100X GlutaMAX (Gibco), 1% 100X MEM Non-Essential Amino Acids (Gibco), 0.128 mM 2-mercaptoethanol (BioRad), and 1X Leukemia Inhibitory Factor (purified and gifted by Barbara Panning Lab at UCSF). Cultures were passaged at least twice before use. Lymphoblastic cell line GM12878 was obtained from the Coriell Institute and cultured in RPMI 1640 (Thermo Scientific), supplemented with 2 mM L-glutamine (Thermo Scientific) and 15% Fetal Bovine Serum (Phoenix Scientific). Cultures were passaged every two days, minimizing cell clumping. K562 cells were similarly cultured in RPMI 1640 supplemented with 2 mM L-glutamine, 10% Fetal Bovine Serum, and penicillin / streptomycin solution (Sigma Aldrich) and passaged every three days.

Nuclei Extraction

Two strategies for nuclei extraction were tested to determine whether overall recovery could be improved after m^6dAse treatment. Extracted nuclei were used directly in downstream processing.

SAMOSA nuclei extraction

1 – 2 million cells were harvested by centrifugation (300xg, 4ºC, 10min), washed in ice cold 1X Phosphate Buffered Saline (PBS, Fisher Scientific), and resuspended in 1 mL cold Nuclear Lysis Buffer (20 mM HEPES, 10 mM KCl, 1 mM MgCl2, 0.1% Triton X-100, 20% Glycerol, 1X Protease Inhibitor (Roche)) by gentle mixing with a wide-bore pipette. The suspension was incubated on ice for 5min, then nuclei were pelleted (600xg, 4ºC, 10min, hard braking disabled) and washed with Buffer M (15 mM Tris-HCl pH 8.0, 15 mM NaCl, 60 mM KCl, 0.5 mM Spermidine).

Alternate nuclei extraction

1 – 2 million cells in culture were rinsed with 1X PBS, then harvested via centrifugation (300xg, 4ºC, 10 – 15min, hard braking disabled) and the cell pellet resuspended in C&R NE1 (20 mM HEPES pH 7.9, 10 mM KCl, 0.1% Triton X-100, 20% glycerol, 1X Protease Inhibitor) by gentle mixing with a wide-bore pipette. 1mL of C&R NE1 was used per 10M cells. Cells were incubated for 10min, nuclei pelleted via centrifugation at 1300xg, 4ºC for 5min, and the nuclei pellet resuspended in C&R Wash Buffer (20mm HEPES pH 7.5, 150 mM NaCl, 0.5M spermidine, 1X Protease Inhibitor)).

Visualizing nuclei bound to Concanavalin A bead

Cells or nuclei were stained with 1X Trypan Blue, and visualized using a Countess III (Thermo Scientific). Brightfield images were exported and color-corrected using Fiji.

Assembly and Preparation of SMRT-Tag transposomes

Custom-loaded hyperactive SMRT-Tn5 transposase was prepared as previously described (Chapter 2.10 – **Methods**). Briefly, 18.9 µM monomer $Tn5^{R27S,E54K,L372P}$ in Tn5 Dilution Buffer (50 mM Tris-HCl pH 7.5, 200 mM NaCl, 0.1 mM EDTA, 2 mM DTT, and 50% glycerol) was combined with 20 µM annealed HPLC purified SMRT-Tag adapter (**Supplementary Table 2.3**), and incubated for 60min at 23ºC with

continuous shaking at 350rpm. Resulting loaded SMRT-Tn5 (9.4 µM monomer) was stored at -20ºC for up to 6 months.

SAMOSA-Tag using Concanavalin A beads

Concanavalin A bead preparation

ConA beads stored at 4ºC were first thawed to room temperature, then gently resuspended from a bead slurry. An *N* x 10 µL aliquot of beads (*N* = number of reactions up to 500k nuclei or cells) was transferred to a 1.5mL protein Lo-bind Eppendorf tube using a wide bore pipette, and 9.6X volumes of ConA Binding Buffer (20 mM HEPES-KOH pH 7.5, 10 mM KCl, 1 mM MnCl2, 1 mM CaCl2) added. Beads were washed twice by pipetting, pelleted on a magnetic rack and supernatant removed. Beads were then resuspended to the same volume as aliquoted and kept on ice until use.

SAMOSA-Tag using Concanavalin Beads (ConA+ST)

A specific number of cells or nuclei (5,000 – 500,0000) in extracted buffer (Buffer M, C&R Wash Buffer) were aliquoted into 250 µL PCR strip tubes (American Scientific), combined with 10 µL of activated ConA beads, and the total reaction volume brought up to 200 µL using C&R Wash Buffer. Binding reactions were incubated for 15min at room temperature to promote binding, with optional agitation at 500rpm using a thermomixer to prevent bead sedimentation.

Optional cell & nuclei permeabilization

ConA bead-immobilized cells or nuclei were pelleted using a magnetic rack for 2min, supernatant removed, and beads resuspended via flicking in 100 µL of ice-cold Dig-Wash Buffer (1X C&R Wash Buffer, supplemented with 0.05% Digitonin (Thermo Fisher Scientific)). Bead-immobilized cells or nuclei were incubated for 30min, up to 2.5h, at 4ºC without agitation to promote membrane permeabilization.

SAMOSA treatment of Concanavalin A bound samples

ConA bead-immobilized cells or nuclei were pelleted using a magnetic rack for 2min, supernatant removed, and resuspended via flicking in 50 µL of Buffer M supplemented with 5-10 µL of high concentration EcoGII m^6dAse (250U, 10 µL of 25,000U/mL stock, New England Biolabs). Methylation reactions were initiated by adding S-adenosyl-methionine (SAM, New England Biolabs) to a final concentration of 1 mM, and incubated at 39ªC with shaking at 350rpm every two min. SAM was replenished to 1.1 6 mM after 15min. Mock-treatment reactions were incubated in the same buffer conditions, without EcoGII. Note, Buffer M was found to promote ConA bead precipitation if incubated for > 30min.

Bead-immobilized tagmentation

Methylated and unmethylated ConA bead-immobilized cells or nuclei were pelleted using a magnetic rack for 2min and supernatant containing residual EcoGII removed. Supernatants were assayed to confirm SAMOSA treatment did not promote sample detachment from ConA beads. Samples were then gently resuspended in Omni-ATAC Buffer (10 mM Tris-HCl pH 7.5, 5 mM $MgCl_2$, 0.33X PBS, 10% DMF, 0.01% Digitonin (Thermo Fisher Scientific), 0.1% Tween-20), supplemented with a varying amount of SMRT-Tn5 transposase (Chapter 3.13 – **Supplementary File 2** for amount used per library). Individual libraires were indexed via tagmentation with uniquely barcoded SMRT-Tag adapters. Tagmentation reactions were incubated for 45min at 55ºC with 300rpm shaking, then terminated by addition of Termination Lysis Buffer (2.5 µL of 20mg/mL Proteinase K (Ambion), 2.5 µL of 10% SDS and 2.5 µL of 0.5 M EDTA), followed by incubation at 60ºC with 1000rpm continuous shaking for 1 hour. To extract tagmented fragments, we determined that joint purification using 2X (115 µL) of SPRI beads and ConA beads already present in each reaction, maximized sample recovery. Homogenous mixtures of both beads and sample fragments were incubated at 23ºC for 30min at 350rpm with shaking every 2min, pelleted via magnet, washed twice with 80% ethanol, and tagmented fragments eluted in 12 µL of Elution Buffer (EB, 10 mM Tris-HCl pH 8.5). Concentration and fragment size of the eluted sample were measured by Qubit

1X High Sensitivity DNA Assay (Thermo Fisher Scientific) and Agilent 2100 Bioanalyzer High Sensitivity DNA Assay (Agilent) respectively.

SMRT-Tag gap repair and library preparation

Gap repair and quality control proceeded as previously described (Chapter 2.10 – **Methods**), as for SAMOSA-Tag. Purified tagmented DNA was combined with Repair Mix (2U Phusion-HF, 80U Taq DNA Ligase, 1X Taq DNA Ligase Reaction Buffer, 0.8 mM dNTP mix (Cat. R0181, Thermo Fisher Scientific), incubated at 37ºC for 1 hour, subjected to a 2X SPRI cleanup and an eluted in 12 µL of 1X EB. Gap-repaired fragments were then incubated in ExoDigest Mix (100U Exonuclease III per 160 ng input DNA, 1X NEBuffer 2) at 37ºC for 1 hour, followed by a 2X SPRI cleanup and an elution in 12 µL of 1X EB.

Library quality control

To assess repair efficiency, 1 µL of eluted library after exonuclease cleanup was measured by Qubit 1X High Sensitivity DNA Assay and compared to input material amount. To validate library quality, 1 µL of eluted library was assayed via Qubit 1X High Sensitivity DNA Assay and Agilent 2100 Bioanalyzer High Sensitivity DNA Assay to measure sample concentration and library size distribution respectively. A subset of prepared libraires were visualized via 0.4 – 0.6% 1X-TAE-agarose gel. Electrophoresis run time was increased to 2-3h, and voltage decreased to 60 – 80V to maximize band resolution. Gels were stained with 1X SYBR Gold (Invitrogen), and imaged on an Odyssey XF imaging system (LI-COR, software version 1.1.0.61).

Optional size selection of SMRT-Tag libraries

For a subset of libraries, an optional size selection step was performed using 0.6X SPRI bead to remove molecules < 500bp, or using 3.1X of 35% (v/v) AMPure PB beads diluted in 1X EB to enrich for molecules > 5000 bp (HMW). In both cases, samples were incubated with beads at room temperature for 15min, washed twice with 80% ethanol, and the size-selected fractions eluted in 15 µL of 1X EB.

Sequencing Concanavalin A SAMOSA-Tag libraries

All ConA+ST libraries were pooled by experiment and sequenced on individual PacBio Sequel II 8M SMRTcells in-house using 2.1 or 2.2 polymerase chemistry depending on the estimated mean fragment distribution of the multiplexed pool. For each SMRTcell, movies were collected for 30 hours with a 2 hour pre-extension time and a 4 hour immobilization time. See Chapter 3.13 – **Supplementary File 2** for information on sequenced libraries.

Concanavalin A SAMOSA-Tag optimization

Concanavalin A SAMOSA-Tag applied to K562 nuclei

K562 cells were cultured as previously described, nuclei extracted following the *SAMOSA nuclei extraction* procedure, and approximately 500,000 per reaction processed via ConA+ST. Post capture, libraries were tagmented using a dilution series of SMRT-Tn5 to ascertain the tunability of bead-immobilized tagmentation (**Figure 3.1c**). A library produced using 18.8 pmol of SMRT-Tn5 was lightly sequenced in a pooled format to determine compatibility with HiFi sequencing (Chapter 3.13 – **Supplementary File 2**).

Concanavalin A SAMOSA-Tag applied to mouse embryonic stem cell nuclei

2 million mouse embryonic stem cells were cultured as described and nuclei extracted following the *alternate nuclei extraction* procedure. Extracted nuclei were first diluted to ~5,000 nuclei / μL as determined by a Countess III cell counter, and then subsequently serially diluted such that the lowest concentration yielded ~5,000 nuclei per ConA binding reaction. Both methylated and unmethylated samples were processed with ConA+ST as described, tagmented using 9.4 pmol of uniquely barcoded SMRT-Tn5, and gap repaired. All methylated libraries and 2 unmethylated libraries were multiplexed for sequencing using 1 SMRTcell (Chapter 3.13 – **Supplementary File 2**). Extracted nuclei from the same

batch were additionally processed via SAMOSA-Tag as a control, following our previously established protocol (Chapter 2.10 – **Methods**).

Concanavalin A SAMOSA-Tag applied to patient-derived xenograft models of prostate cancer

Preparation of PDX samples

PDX samples were prepared as described (Chapter 2.10 – **Methods**). Briefly, paired primary and metastasis patient derived xenograft models of prostate cancer, originally derived from the same treatment-naïve patient, were processed to cell suspensions and contaminant mouse or dead cells depleted via FACS. Nuclei were subsequently extracted following the *SAMOSA nuclei extraction* procedure with additional modifications to minimize loss, and either processed via SAMOSA-Tag, or ConA+ST.

Concanavalin A SAMOSA-Tag applied to mouse-depleted primary and metastasis PDX samples

Approximately ~10k extracted nuclei were aliquoted separately per sample (6 replicates for the primary sample, 8 replicates for the metastasis sample) and processed independently via ConA+ST as described. Samples were tagmented using 9.4 pmol of uniquely barcoded SMRT-Tn5 (Chapter 3.13 – **Supplementary File 2**) and the entire sample used for SMRT-Tag gap repair and library preparation. Libraires were pooled and size selected using 0.6X SPRI beads to deplete < 500bp molecules as described and sequenced in multiplexed format on 1 SMRTcell per sample. Note, no unmethylated libraires were generated, and SAMOSA-Tag unmethylated libraries were used as a reference instead.

Cell-direct Concanavalin A SAMOSA-Tag applied to mouse-depleted metastasis PDX samples

To determine whether ConA+ST was compatible with cell-direct processing, approximately ~15k mouse-depleted metastasis sample cells were aliquoted (8 replicates) and processed via ConA+ST with the following modifications: Cells were bound for 20min instead of 15min and permeabilized for 2.5h in Dig-Wash Buffer as described. SAMOSA treatment was also supplemented with 0.05% digitonin to promote permeabilization during the methylation reaction. Bead-immobilized cells were tagmented using 9.4 pmol

of uniquely barcoded SMRT-Tn5, and three libraries selected for sequencing. Selected libraries were size selected using 0.6X SPRI beads and added to the multiplexed nuclei ConA+ST cell for metastasis samples.

Concanavalin A SAMOSA-Tag applied to healthy peripheral mononuclear cells

Preparation and sorting of leukocyte subsets from a healthy donor

From the blood of a healthy donor termed HC1, a 34 y/o Caucasian female, two vials of leukocytes obtained and frozen for storage (AllCells). On the day of processing, cells aliquots were thawed for 5min at 37ºC in a water bath, transferred to a sterile 50mL conical tube and rinsed with 1mL of pre-warmed 1X PBS. Cell aliquots were combined and brought up to 50mL total volume with pre-warmed 1X PBS, centrifuged for 10min at 400xg, 4ºC, resuspended into 2mL cold PBS, and 25 µL removed for assessing cell count and viability via Trypan Blue staining. Cells were washed once more in 10mL in cold PBS and pelleted at 400xg, 10min, 4ºC. An initial cell count assessed < 5% non-viable cells, and ~22M cells in total.

Enriching for PBMC subpopulations via fluorescence-assisted cell sorting (FACS)

Cell subpopulations were enriched from donor cells via multiplexed FACS. First, the donor cell pellet was resuspended in 100 µL pre-warmed 1X PBS and stained with 100 µL of Zombie NIR (Cat No. 423105, Biolegend) for 10min at 4ºC in the dark, followed by quenching with 5mL FACS buffer. Supernatant was removed via centrifugation at 400xg, 10min and cells resuspended in 75 µL of staining master mix, consisting of 64 µL FACS buffer, 1.2 µL Human TruStain FcX (Cat No. 422302, Biolegend), and 0.8 µL each of Alexa700 CD3 (Cat No. 300423, clone UCHT1, Biolegend), BV650 CD4 (Cat No. 300535, clone RPA-T4, Biolegend), BV510 CD8 (Cat No. 344731, clone SK1, Biolegend) , BV711 CD197/CCR7 (Cat No. 353227, clone G043H7, Biolegend), PerCP/Cy5.5 CD45RA (Cat No. 304121, clone HI100, Biolegend), BV605 CD19 (Cat No. 302243, clone HIB19, Biolegend), PE CD20 (Cat No. 302305, clone 2H7, Biolegend), PE-Cy7 IgD (Cat No. 348209, clone IA6-2, Biolegend), BV421 CD27 (Cat No.

356417, clone M-T271, Biolegend), APC CD38 (Cat No. 303509, clone HIT2, Biolegend), PE-Dazzle594 CD14 (Cat No. 367133, clone 63D3, Biolegend), and FITC HLA-DR (Cat No. 307603, clone L243, Biolegend). Cells were stained for 15min at 4ºC in the dark, then quenched with 430 µL FACS buffer, washed with 5mL of FACS buffer, and sorted with an input concentration of ~5M cells / mL. Fourteen single stain compensation controls were also generated, including no-antibody and APC-Cy7 only controls using 0.8 µL of the relevant antibody, 1 drop of compensation beads (BD Biosciences) and 100 µL of 1X PBS. Cell populations were either collected continuously, or intermittently using the following gating strategy (Supplementary Figure 3.3): T cells (CD3+), both CD4+ naïve (CD3+ CD4+ CCR7+ CD45RA+) and memory (CD3+ CD4+ CD45RA-), CD8+ naïve (CD3+ CD8+ CCR7+ CD45RA+) and memory (CD3+ CD8+ CD45RA-); B cells both naïve (CD3– CD19+ CD20+ CD27- CD38- IgD+) and memory (CD3– CD19+ CD20+ CD27+ CD38- IgD-); monocytes (CD3-, CD14+, HLA-DR+); dendritic cells (CD3-, CD19 -, CD20-, CD14-, and HLA-DR+.). Importantly, CCR7+ was used as a marker for naïve T cell status, but did not produce detectable signal, suggesting issues with antibody staining. As a result, naïve and memory T cell subsets were defined only on CD45RA status. Approximate subpopulation recovery ranged from ~300k (memory B cells) to ~1.5M (monocytes) (Chapter 3.13 – **Supplementary File 2)**

Genomic DNA extraction

Genomic DNA from sorted aliquots was extracted using Qiagen MagAttract (Qiagen), following manufacturers recommendation for “Whole Blood or Tissue”. DNA was eluted in Buffer AE at 25ºC, 1500rpm for 10min, followed by a second elution for 2min to maximize sample recovery. Samples were eluted in 60 µL total volume.

Cell-direct Concanavalin A SAMOSA-Tag applied to PBMC subpopulations

Four replicates of approximately ~20k cells each per subpopulation were processed via ConA+ST as described, with the following modifications: Cells were incubated with ConA beads for 20min to promote

bead-capture and incubated in Dig-Wash Buffer for 30min at 4ºC to promote permeabilization. SAMOSA treatment was additionally supplemented with 0.05% digitonin. Due to the lack of sufficient barcodes, samples were tagmented using 1.47 pmol of SMRT-Tn5 uniquely loaded using all 24 available SMRT-Tag barcodes, with 8 barcodes repeated (SMRT-A_bc001 – SMRT-A_bc0014, Chapter 3.13 – **Supplementary File 2**). All successfully gap-repaired SMRT-Tag libraries derived from naïve and memory B cell subpopulations were then pooled, size selected using 3.1X of 35% Ampure PB beads to enrich for longer fragments, and sequenced in multiplex on one SMRTcell.

Repairing DNA damage

Repairing Concanavalin A SAMOSA-Tag libraries

Existing naïve and memory B cell ConA+ST libraries were pooled and subjected to DNA damage repair using 1 µL of PreCR repair mix (M0309S, New England Biolabs) in 1X ThermoPol Buffer (New England Biolabs), 100 µM dNTP mix, and 1X fresh NAD+. Repaired libraires were cleaned up using 2X SPRI beads, eluted in 12 µL EB, and sequenced in multiplexed format on 1 SMRTcell.

SMRT-Tag and TPK2.0 applied to PreCR-repaired Genomic DNA

Less than 1 µg of genomic DNA extracted from naïve and memory B cells, as well as monocytes, were treated with 1 µL of PreCR repair mix in 1X ThermoPol Buffer (New England Biolabs), 100 µM dNTP mix, and 1X fresh NAD+, cleaned up using 2X SPRI beads and eluted in 12 µL EB. Repaired DNA was prepared via SMRT-Tag as previously described, and where possible prepared using Template Prep. Kit 2.0 from Pacific Biosciences as a control. Resulting libraries were pooled together and lightly sequenced on one multiplexed SMRTcell.

Evaluating DNA quality

Genomic DNA was extracted from ~250k freshly passaged GM12878 cells with high viability and ~250k one month old frozen GM12878 nuclei using three commercially available kits – Wizard Genomic DNA

Purification Kit (Cat. No. A1120, Promega), Monarch Genomic DNA Purification Kit (Cat. No. T3010S, New England Biolabs) and MagAttract HMW DNA Kit (Cat. No. 67563, Qiagen). Extracted genomic DNA, genomic DNA extracted previously from donor HC1 sorted cell populations, known degraded genomic DNA, and an aliquot of high quality genomic DNA (HM24385, Coriell Institute) were also analyzed via Genomic DNA Tapestation (Agilent) and DIN number estimated from software.

Data Analysis

All relevant scripts used for analyses are available at https://github.com/RamaniLab/SMRT-Tag. All plots were made using R (v.4.2.1) and *ggplot2* (https://ggplot2.tidyverse.org/).

Estimating library conversion efficiency and nuclei capture recovery

Reaction efficiency reported here for SAMOSA-Tag and ConA+ST was estimated by dividing the amount of library recovered after SMRT-Tag library preparation by the input amount. Sample recovery from Concanavalin A capture, bead-immobilized m^6dAse methylation, and bead-immobilized tagmentation was estimated using cell counting, as well as amount of DNA recovered after sample lysis and dual-bead purification. Estimated reaction efficiencies are available in Chapter 3.13 – **Supplementary File 2.**

Data preprocessing

Data generated from all SMRTcells was processed using the SAMOSA analysis pipeline[201]. To summarize, subreads were first demultiplexed using *lima* (v.2.6.0, Pacific Biosciences) and HiFi reads generated per sample using *ccs* (v.6.4.0, Pacific Biosciences), with default parameters and mode *--hifi-kinetics* enabled to annotate reads with kinetic sequencing measurements. Sample reads were then processed via *primrose* (v1.3.0, Pacific Biosciences) to annotate reads with m^5dC probabilities at CpG motifs. *Primrose*-processed HiFi reads were used as input for the SAMOSA analysis pipeline, which predicts single-molecule nucleosome footprints from m^6dA modification probabilities through a series of

neural networks and a hidden Markov model (NN-HMM) (see Abdulhay et al.[201] for an in depth description).

Read alignment

Processed HiFi reads were aligned using *pbmm2* (v.1.9.0, Pacific Biosciences) to the relevant reference genome. ConA+ST reads from K562, HC1 naïve B and HC1 memory B libraries were aligned to the GRCh38 reference genome. mESC ConA+ST reads were aligned to the GRCm38 reference genome. Primary and metastasis PDX SAMOSA-Tag reads were aligned to a joint GRCh38 / GRCm39 reference genome and only reads uniquely aligning to GRCh38 retained for downstream analyses.

Insertion bias analyses at TSS and CTCF sites

Read ends from ConA+ST data were extracted from BAM files and tabulated in a 5 kilobase window surrounding annotated GENCODEV28 (GRCh38) transcriptional start sites (TSSs). Counts were normalized per 1M molecule read ends sequenced. Plots were smoothed using a running mean of 100 nucleotides.

Fiber type identification via unsupervised clustering.

We calculated single-molecule autocorrelograms and performed *leiden* clustering as in Nanda et al.[233], with a resolution of 0.4, via the implementation available in Scanpy[297]. Only fibers with at least 1000 bp of sequence were considered. Clusters that comprised less than 5% of all fibers, as well as constitutively unmethylated fibers, were filtered out, and the remaining clusters designated as “fiber types”. Average accessibility profiles per fiber type were determined by averaging binary single-molecule accessibility signal across all fibers classified as a given fiber type, aligned to the prefix of the sequencing read as defined from the HiFi BAM file.

Fractional changes in fiber representation by sample

For each sample, read counts per fiber type were first normalized using a median of medians approach to minimize the effects of library depth, and then normalized to the total normalized depth. Fractional changes were determined by subtracting fractional representation between samples.

Mouse embryonic stem cell E14 domain annotations

The following processed ENCODE 4 annotation files in BED format were obtained from the ENCODE portal [298] and filtered to remove regions annotated on the mitochondrial chromosome: H3K9me3 (ENCFF925BSH), DNaseI-seq (ENCFF048DWN), H3K36me3 (ENCFF362DZS), H3K4me1 (ENCFF158GBZ), H3K4me3 (ENCFF993IIG), H3K27ac (ENCFF519QMV), H3K9ac (ENCFF668UBL).

K-mer spectrum decomposition by fiber type

For a given sample, processed reads were aggregated by fiber type assignment and source method (SAMOSA-Tag, ConA+ST) and sequences written to FASTA files. Canonical k-mer spectra were then generated using *jellyfish*[299], and count files were produced using *jellyfish dump*. For each sample, counts were first normalized using a median of medians approach to account for sequencing depth, and then the cosine similarity of each ConA+ST library to the mean SAMOSA-Tag k-mer spectrum determined. Cosine distance was calculated as 1 – *cosine similarity*.

Determining per-sample fiber type enrichment patterns

Following from previous work (Chapter 2.10 – **Methods**), we determined fiber type overrepresentation in individual domains using a one-sided Fisher's Exact test. Libraries of a given sample type were pooled, and the enrichment of fiber type A in epigenomic domain B was determined by first constructing a 2x2 contingency table with counts A ∩ B, A ∩ B', A' ∩ B, and A' ∩ B '. The table was used as input for a two-sided Fisher's exact test, and resulting *p* values corrected for multiple testing using Storey's *q* value[267]. A

threshold for significance was set at Storey's $q < 0.05$. Estimated odds ratios between methods were then correlated using Pearson's *r* as a measure of technology consistency.

Prostate-specific epigenome stratification

Normal prostate tissue-specific chromHMM annotations in BED format were obtained from Wang et al.[260], as previously (Chapter 2.10 – **Methods**). Annotations were lifted over from reference hg19 to hg38.

Determining differential fiber usage

Differential fiber usage per domain and fiber type (Δ) was calculated using a logistic regression approach defined previously (Chapter 2.10 – **Methods**). To summarize, size-factor normalized counts of a given fiber type and epigenomic domain per replicate library were compared using logistic regression. The regression model was fit using the *glm* function in R (v.4.2.1) and the sample coefficient used as an estimate of $\log_2$ fold change (Δ, "delta"). In cases where domains were not defined, as for B cell populations, differential fiber usage per fiber type alone was determined. Associated *p* values for each Δ were corrected for multiple testing using Storey's *q* value[267]. A threshold for significance was set at Storey's $q < 0.05$.

Method-specific changes in differential fiber usage estimation

Separately, for both ConA+ST and SAMOSA-Tag libraries, we determined differential fiber usage between metastasis and primary samples as described above. The resulting effect sizes per domain and fiber type combination estimate the degree to which a given fiber type is differentially represented in the given domain in metastasis versus primary samples. Effect sizes estimated using each method were log transformed, and the difference in estimates determined ($\Delta_{\text{ConA vs ST}}$)., where a value of 0 indicated the same effect size prediction, and positive or negative deviation indicating a method-specific change in differential fiber usage. A significance value was determined using Fisher's method for combining *p*

values, and corrected for multiple testing using Storey's q value. A threshold for significance was set at Storey's $q < 0.05$.

Comparing average single-molecule m^5dC levels between B cell populations

Reads were first separated by fiber type classification in both memory and naïve B cell libraries. m^5dC modification probability estimates derived from *primrose* were extracted per read, and both the average m^5dC level across the read and number of CpG motifs per read were determined. Distributions of average m^5dC levels per read were then estimated, stratified by both fiber type, sample, and CpG density per read, which we previously determined was significantly associated with both fiber type (Chapter 2.10 – **Methods, Figure 2.4ef**). A Mann-Whitney U test (unpaired) was performed to estimate both effect size and significance, using R, and p values adjusted for multiple testing using the Benjamini-Hochberg procedure, and a threshold for significance set at $q < 0.01$.

3.12. Supplementary Tables

Supplementary Table 3.1: Flow cytometry results when sorting for cell subpopulations from peripheral mononuclear blood cells

Cell Population	Key Markers	Total sorted
Naïve CD4+ T	CD3+ CD4+ CCR7+ CD45RA+	602,427
Mem CD4+ T	CD3+ CD4+ CD45RA-	986,228
Naïve CD8+ T	CD3+ CD8+ CCR7+ CD45RA+	708,979
Memory CD8+ T	CD3+ CD8+ CD45RA-	751,794
Naïve B	CD3– CD19+ CD20+ CD27- CD38- IgD+	805,048
Memory B	CD3- CD19+ CD20+ CD27+ CD38- IgD-	317,768
Monocytes	CD14+ HLA-DR+	1,583,154
Dendritic cells (DCs)	CD3- CD19- CD20- CD14- HLA-DR+	431,002

3.13. Supplementary File 2 – Library and sequencing statistics

Supplementary file containing library preparation parameters and sequencing statistics for ConA+SAMOSA-Tag datasets included in Chapter 3.

Chapter 4: Conclusions & Future Directions

4.1. Summary of findings

In the previous chapters, we have developed novel methods for low input library preparation for HiFi sequencing. In Chapter 2, we demonstrated that our transposase-mediated strategy, SMRT-Tag, can measure primary sequence and epigenetic modifications on single molecules derived from either gDNA or from nuclei directly. In Chapter 3, we developed an extension to the *in situ* version of our assay, SAMOSA-Tag, that enabled direct m^6dAse footprinting of ~10,000 cells or nuclei, significantly reducing sample handling loss. We then generated the first set of single molecule accessibility profiles in primary human samples, including prostate cancer patient-derived xenografts and healthy naïve and mature B cells. Analyses of both datasets revealed distinct changes in nucleosome regularity, including an enrichment for hyper-accessible irregular fibers in metastatic prostate cancer and for hypomethylated short NRL fibers in maturing B cells. Excitingly, across all cases we conservatively lowered input requirements ~20-fold, decreasing the threshold for both 3rd generation sequencing and single-molecule chromatin profiling studies for the broader community.

Beyond the work discussed here, we envision SMRT-Tag as a general tool for developing methods with native single molecule sequencing as a readout. Reducing input requirements allows 2nd generation assays that rely on PCR amplification to be adapted for 3rd generation sequencing. Further, the use of Tn5 enables customizing sequencing adapters for sample barcoding and indexing. There are numerous assays where highly accurate long reads could provide greater resolution. Here, we discuss two – target enrichment and single cell sequencing – that we believe are especially useful.

4.2. Integrating SMRT-Tag with target enrichment

Reagent costs often place practical limits on the scalability of sequencing. As we described in Chapter 2, 3rd generation platforms are only now approaching a competitive price point per base

sequenced when compared to 2nd generation platforms. In any given assay, we would therefore prefer to sequence molecules that provide biological signal ("targets") and avoid wasting sequencing capacity on uninformative ones. Strategies to increase the proportion of targets in a library are called "target enrichment"[300]. In clinical contexts target enrichment is popular for sequencing a subset of key disease-associated regions (*e.g.* genes, enhancers, exons) at a fraction of the cost of whole genome. This is generally achieved by fragmenting gDNA and capturing specific sequences via hybridization to a set of known surface-immobilized probes. However, this process rapidly reduces the amount of available material for library preparation. For example, the Broad Institute's Clinical Research exome probe set, which aims to capture all exons as well as regions associated with rare and inherited cancers, targets only ~1.1% of the genome, reducing material after enrichment by ~99%[301]. Significant PCR amplification is therefore required, and is still considered essential for state of the art methods[302]. We speculated that if we could prepare target-enriched HiFi libraries directly without PCR, then we could profile nucleosome positions and m^5dCpG methylation with high depth at multiple disease relevant loci. Although SMRT-Tag libraries are incompatible with *post hoc* target enrichment, as noted in Chapter 2, SMRT-Tag would be an ideal method for producing native HiFi libraries from low-input gDNA that had already been target-enriched.

To test this, we first designed a set of sgRNA guides targeting eight 200kb regions surrounding genes of interest including *MYC*, *CTCF*, and *INO80* (8 x 200kb targets, ~0.053% of the human genome) and then excised these regions from K562 nuclei using Cas9. Next, we enriched for molecules ~200kb using pulsed field agarose gel electrophoresis and applied SMRT-Tag to the extracted DNA, which contained a mixture of the target loci and contaminant gDNA. Our pilot experiment produced enough library material to sequence on a single SMRT cell, and the resulting reads aligned well to the human genome (~97.8% alignment rate). We examined our target loci and found that coverage of all eight ranged from 50 – 200X, with most reads between 3 – 4kb long. When compared to state of the art PCR-based hybridization panels for long reads, we found our overall enrichment of was competitive (~160X vs. 75 – 190X), but had a significantly reduced on-target rate (~9% vs. ~93 – 99% respectively). Nonetheless,

using *hifiasm*[303], on-target long reads could readily be assembled into contiguous haplotypes and permitted accurate germline genotyping, variant phasing, and detection of low-frequency SNVs.

This exciting result suggests SMRT-Tag is compatible with target enrichment and opens multiple future avenues to pursue, the most immediate being the incorporation of m^{6}dAse footprinting to resolve variation in nucleosome occupancy. Target-enrichment SMRT-Tag could also be useful in determining variants at difficult-to-genotype genes, where HiFi reads have already be found to be highly effective[304,305]. One such locus is the major histocompatibility complex (MHC), where inherited variants in *HLA* genes can predispose individuals to autoimmune and infectious diseases[306]. Using our current benchmark of SMRT-Tag libraries producing ~2.7M HiFi reads from ~40 ng gDNA (Chapter 2.4), we can speculate that scaled up target-enrichment SMRT-Tag could yield up to ~300X coverage per locus – more than enough to assemble and phase multiple MHC haplotypes in a single experiment. Further, although target enrichment reactions are lossy, multiplexing samples from multiple patients by tagmentation with uniquely barcoded SMRT-Tn5 could achieve larger library quantities. Even with a ~9% on-target rate, we estimate 6 patients could be multiplexed with ~50X coverage per locus each. Thus, we envision future work improving on-target rates will make target enrichment SMRT-Tag a highly useful tool for clinical genetics.

4.3. Towards single cell, single molecule profiling using SMRT-Tag

As we noted in Chapter 3, current single-molecule chromatin profiling methods cannot assign single molecules back to their cell of origin, making the study of heterogenous samples exceedingly difficult. While we developed a bead-immobilization approach for physically sorting out subpopulations from primary samples and profiling them separately, prior knowledge is required of both 1) which subpopulations are biologically relevant and 2) the specific markers that enable subpopulation isolation. We therefore sought to develop a version of m^{6}dAse footprinting that could map single cell *and* single-molecule chromatin accessibility.

Single cell techniques measuring RNA (single cell RNA sequencing, scRNA-seq)[307] and chromatin accessibility (single cell ATAC sequencing, scATAC-seq)[270], and copy number variation[238] achieve both molecule and cell resolution by tagging fragments from the same cell with a unique cellular barcode that can be read out during sequencing. Numerous procedures exist to uniquely barcode single cells. These methods include 1) physically isolating cells into microwells and adding sequencing adapters via tagmentation[246,308], 2) encapsulating cells in hydrogel-based droplets and capturing biomolecules with barcoded bead-bound adapters[309], or 3) using successive rounds of combinatorial labelling to attach unique barcodes to molecule ends *in situ*[249,310]. In all cases DNA amplification is required to produce enough input material to overcome sample handling losses, and the resulting libraires are relatively sparse per single cell (*e.g.* 0.00017X – 0.025X genome-wide coverage in scATAC-seq). Both facets make adapting any of these methods difficult. Amplification is incompatible with m^6dAse footprinting because, to our knowledge, epigenetic marks cannot be propagated accurately to new PCR templates. Further, low numbers of molecule per cell would severely underpower any chromatin fiber frequency comparisons.

We observed that tagmenting nuclei with SMRT-Tn5 produced both long (> 1000bp) and short (< 500bp) fragments. Given that SMRT-Tn5 has mild insertional preferences for open chromatin (**Supplementary Figure 2.9a**), we speculated that deep sequencing of short fragments may reveal accessibility information similar to ATAC-seq. If we could resolve ATAC-seq signal per single cell, then we could group together single cells with similar ATAC profiles and aggregate their respective chromatin fibers, allowing for unbiased and well-powered characterization of cell-type specific chromatin fiber accessibility. This would require 1) that SMRT-Tag could produce HiFi libraries from DNA on the order of a single cell (~0.006 ng) and 2) a method for converting these informative short fragments with SMRT adapters into short-read compatible libraries.

To address the first requirement, we tested if SMRT-Tag could prepare HiFi libraries from extremely small amounts of DNA, reasoning that any single cell SMRT-Tag procedure would involve pooled gap repair and exonuclease digestion as we'd demonstrated earlier (Chapter 2.3). We tagmented as little as 0.1 ng to 10 ng of reference gDNA, pooled and gap repaired samples, and then sequenced the

multiplexed pool as part of a larger sequencing run. We recovered 100 – 1,000s of HiFi reads proportional to the frequency of the input samples in the multiplexed cell, indicating no issues with sequencing efficiency. Turning again to our benchmark from Chapter 2, SMRT-Tag library produced from 40 ng of input gDNA yielded ~2.7M HiFi reads in one sequencing run. Following this logic, processing between 7,000 – 10,000 uniquely indexed cells would generate enough material for one SMRTcell, yielding an estimated ~270 – 385 single molecules per cell. While this data would still be sparse (*i.e.,* 0.00025X coverage of the human genome per cell), if ~5 – 10 resolvable populations existed in a sample then aggregated cell types would have between 230,000 – 650,000 fibers each, more than sufficient for fiber-level analyses.

To address the second requirement, we designed a library conversion strategy where a strand displacing polymerase (*Klenow exo-* or *Bst Polymerase, Large Fragment*) first unfolded the SMRT adapter hairpin (**Figure 1.1**) into a linear double stranded molecule. After end repair, linearized molecules were ligated with SBS adapters and PCR amplified to generate an Illumina compatible library. Light sequencing of this library confirmed our conversion approach worked (> 80% clusters passing filter), although with high duplication rates (~40 – 60%). We tested if these fragments had hallmarks of ATAC-seq data including enrichment at transcription start sites but observed only mild enrichment over background (TSS enrichment score 1.5 – 1.8, < 5 failing QC)[311] despite most of the reads being likely subnucleosomal from nucleosome free regions[312]. While short fragments could be readily captured, they likely did not contain enough ATAC signal to produce meaningful clusters.

Nonetheless, our preliminarily results suggest that if ~7,000 – 10,000 individual cells were m^6dAse-treated, deposited in microwells, and tagmented with SMRT-Tn5, we could achieve sparse single molecule coverage per cell. Further, optimizing SMRT-Tn5 quantity and tagmentation conditions can likely shift more accessibility information into short fragments while producing enough longer fragments for HiFi sequencing. An existing method that provides a blueprint for how to proceed is Smart-seq3xpress[313], which probes single cell RNA content by nanoliter-scale reverse transcription and tagmentation in single cells followed by centrifugation to pool and amplify cDNA with minimal loss. One

could easily imagine a similar process with SMRT-Tag, where SMRT-Tn5 is applied to single cells in ~1-5 nanoliter volumes and then tagmented fragments pooled for joint gap-repair and exonuclease digestion. We are therefore optimistic that progress in this direction will turn native single molecule, single cell studies into a reality in the near future.

References

1. Cutter, A. R. & Hayes, J. J. A brief review of nucleosome structure. *FEBS Lett.* **589**, 2914–2922 (2015).
2. Finch, J. T. *et al.* Structure of nucleosome core particles of chromatin. *Nature* **269**, 29–36 (1977).
3. Ou, H. D. *et al.* ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science* **357**, eaag0025 (2017).
4. Arents, G. & Moudrianakis, E. N. The histone fold: a ubiquitous architectural motif utilized in DNA compaction and protein dimerization. *Proc. Natl. Acad. Sci.* **92**, 11170–11174 (1995).
5. Polach, K. J. & Widom, J. Mechanism of Protein Access to Specific DNA Sequences in Chromatin: A Dynamic Equilibrium Model for Gene Regulation. *J. Mol. Biol.* **254**, 130–149 (1995).
6. Noll, M. & Kornberg, R. D. Action of micrococcal nuclease on chromatin and the location of histone H1. *J. Mol. Biol.* **109**, 393–404 (1977).
7. Simpson, R. T. Structure of the chromatosome, a chromatin particle containing 160 base pairs of DNA and all the histones. *Biochemistry* **17**, 5524–5531 (1978).
8. Fyodorov, D. V., Zhou, B.-R., Skoultchi, A. I. & Bai, Y. Emerging roles of linker histones in regulating chromatin structure and function. *Nat. Rev. Mol. Cell Biol.* **19**, 192–206 (2018).
9. Routh, A., Sandin, S. & Rhodes, D. Nucleosome repeat length and linker histone stoichiometry determine chromatin fiber structure. *Proc. Natl. Acad. Sci.* **105**, 8872–8877 (2008).
10. Millán-Zambrano, G., Burton, A., Bannister, A. J. & Schneider, R. Histone post-translational modifications — cause and consequence of genome function. *Nat. Rev. Genet.* **23**, 563–580 (2022).
11. Henikoff, S. & Smith, M. M. Histone Variants and Epigenetics. *Cold Spring Harb. Perspect. Biol.* **7**, a019364 (2015).
12. Marzluff, W. F., Gongidi, P., Woods, K. R., Jin, J. & Maltais, L. J. The Human and Mouse Replication-Dependent Histone Genes. *Genomics* **80**, 487–498 (2002).

13. Buschbeck, M. & Hake, S. B. Variants of core histones and their roles in cell fate decisions, development and cancer. *Nat. Rev. Mol. Cell Biol.* **18**, 299–314 (2017).

14. Serra-Cardona, A. & Zhang, Z. Replication-coupled nucleosome assembly as a passage of epigenetic information and cell identity. *Trends Biochem. Sci.* **43**, 136–148 (2018).

15. Jin, C. *et al.* H3.3/H2A.Z double variant–containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. *Nat. Genet.* **41**, 941–945 (2009).

16. Chen, P., Wang, Y. & Li, G. Dynamics of histone variant H3.3 and its coregulation with H2A.Z at enhancers and promoters. *Nucleus* **5**, 21–27 (2014).

17. Müller, S. *et al.* Phosphorylation and DNA Binding of HJURP Determine Its Centromeric Recruitment and Function in CenH3CENP-A Loading. *Cell Rep.* **8**, 190–203 (2014).

18. Black, B. E. *et al.* Centromere Identity Maintained by Nucleosomes Assembled with Histone H3 Containing the CENP-A Targeting Domain. *Mol. Cell* **25**, 309–322 (2007).

19. Fachinetti, D. *et al.* A two-step mechanism for epigenetic specification of centromere identity and function. *Nat. Cell Biol.* **15**, 1056–1066 (2013).

20. Kornberg, R. D. & Stryer, L. Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. *Nucleic Acids Res.* **16**, 6677–6690 (1988).

21. Mavrich, T. N. *et al.* A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res.* **18**, 1073–1083 (2008).

22. Fedor, M. J., Lue, N. F. & Kornberg, R. D. Statistical positioning of nucleosomes by specific protein-binding to an upstream activating sequence in yeast. *J. Mol. Biol.* **204**, 109–127 (1988).

23. Bracken, A. P., Brien, G. L. & Verrijzer, C. P. Dangerous liaisons: interplay between SWI/SNF, NuRD, and Polycomb in chromatin regulation and cancer. *Genes Dev.* **33**, 936–959 (2019).

24. Kaplan, N. *et al.* The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**, 362–366 (2009).

25. Zhang, Y. *et al.* Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. *Nat. Struct. Mol. Biol.* **16**, 847–852 (2009).

26. Ioshikhes, I. P., Albert, I., Zanton, S. J. & Pugh, B. F. Nucleosome positions predicted through comparative genomics. *Nat. Genet.* **38**, 1210–1215 (2006).

27. Struhl, K. & Segal, E. Determinants of nucleosome positioning. *Nat. Struct. Mol. Biol.* **20**, 267–273 (2013).

28. Iyer, V. & Struhl, K. Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. *EMBO J.* **14**, 2570–2579 (1995).

29. Allfrey, V. G., Faulkner, R. & Mirsky, A. E. Acetylation and methylation of histones and their possible role in the regulation of rna synthesis*. *Proc. Natl. Acad. Sci.* **51**, 786–794 (1964).

30. Cosgrove, M. S., Boeke, J. D. & Wolberger, C. Regulated nucleosome mobility and the histone code. *Nat. Struct. Mol. Biol.* **11**, 1037–1043 (2004).

31. Allshire, R. C. & Madhani, H. D. Ten principles of heterochromatin formation and function. *Nat. Rev. Mol. Cell Biol.* **19**, 229–244 (2018).

32. Soufi, A., Donahue, G. & Zaret, K. S. Facilitators and Impediments of the Pluripotency Reprogramming Factors' Initial Engagement with the Genome. *Cell* **151**, 994–1004 (2012).

33. Oksuz, O. *et al.* Capturing the Onset of PRC2-Mediated Repressive Domain Formation. *Mol. Cell* **70**, 1149-1162.e5 (2018).

34. Escobar, T. M. *et al.* Active and Repressed Chromatin Domains Exhibit Distinct Nucleosome Segregation during DNA Replication. *Cell* **179**, 953-963.e11 (2019).

35. Okano, M., Xie, S. & Li, E. Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nat. Genet.* **19**, 219–220 (1998).

36. Bestor, T., Laudano, A., Mattaliano, R. & Ingram, V. Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells: The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. *J. Mol. Biol.* **203**, 971–983 (1988).

37. Feng, S. *et al.* Conservation and divergence of methylation patterning in plants and animals. *Proc. Natl. Acad. Sci.* **107**, 8689–8694 (2010).

38. Kaneda, M. *et al.* Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. *Nature* **429**, 900–903 (2004).

39. Schmitz, R. J., Lewis, Z. A. & Goll, M. G. DNA Methylation: Shared and Divergent Features across Eukaryotes. *Trends Genet.* **35**, 818–827 (2019).

40. Walsh, C. P., Chaillet, J. R. & Bestor, T. H. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat. Genet.* **20**, 116–117 (1998).

41. Li, E., Beard, C. & Jaenisch, R. Role for DNA methylation in genomic imprinting. *Nature* **366**, 362–365 (1993).

42. Ehrlich, M. DNA hypomethylation in cancer cells. *Epigenomics* **1**, 239–259 (2009).

43. Nan, X. *et al.* Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature* **393**, 386–389 (1998).

44. Ng, H.-H. *et al.* MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex. *Nat. Genet.* **23**, 58–61 (1999).

45. Rao, S. *et al.* Systematic prediction of DNA shape changes due to CpG methylation explains epigenetic effects on protein–DNA binding. *Epigenetics Chromatin* **11**, 6 (2018).

46. Samee, Md. A. H., Bruneau, B. G. & Pollard, K. S. A De Novo Shape Motif Discovery Algorithm Reveals Preferences of Transcription Factors for DNA Shape Beyond Sequence Motifs. *Cell Syst.* **8**, 27-42.e6 (2019).

47. Yin, Y. *et al.* Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**, eaaj2239 (2017).

48. Raiber, E.-A. *et al.* 5-Formylcytosine organizes nucleosomes and forms Schiff base interactions with histones in mouse embryonic stem cells. *Nat. Chem.* **10**, 1258–1266 (2018).

49. Choy, J. S. *et al.* DNA Methylation Increases Nucleosome Compaction and Rigidity. *J. Am. Chem. Soc.* **132**, 1782–1783 (2010).

50. Ngo, T. T. M. *et al.* Effects of cytosine modifications on DNA flexibility and nucleosome mechanical stability. *Nat. Commun.* **7**, 10813 (2016).

51. Jimenez-Useche, I. & Yuan, C. The Effect of DNA CpG Methylation on the Dynamic Conformation of a Nucleosome. *Biophys. J.* **103**, 2502–2512 (2012).

52. Ooi, S. K. T. *et al.* DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature* **448**, 714–717 (2007).

53. Weinberg, D. N. *et al.* The histone mark H3K36me2 recruits DNMT3A and shapes the intergenic DNA methylation landscape. *Nature* **573**, 281–286 (2019).

54. Hartley, P. D. & Madhani, H. D. Mechanisms that Specify Promoter Nucleosome Location and Identity. *Cell* **137**, 445–458 (2009).

55. Chen, K. *et al.* Broad H3K4me3 is associated with increased transcription elongation and enhancer activity at tumor-suppressor genes. *Nat. Genet.* **47**, 1149–1157 (2015).

56. Henikoff, S. & Shilatifard, A. Histone modification: cause or cog? *Trends Genet.* **27**, 389–396 (2011).

57. Cano-Rodriguez, D. *et al.* Writing of H3K4Me3 overcomes epigenetic silencing in a sustained but context-dependent manner. *Nat. Commun.* **7**, 12284 (2016).

58. Wang, H. *et al.* H3K4me3 regulates RNA polymerase II promoter-proximal pause-release. *Nature* **615**, 339–348 (2023).

59. Weber, C. M., Ramachandran, S. & Henikoff, S. Nucleosomes Are Context-Specific, H2A.Z-Modulated Barriers to RNA Polymerase. *Mol. Cell* **53**, 819–830 (2014).

60. Li, S., Wei, T. & Panchenko, A. R. Histone variant H2A.Z modulates nucleosome dynamics to promote DNA accessibility. *Nat. Commun.* **14**, 769 (2023).

61. Venkatesh, S. & Workman, J. L. Histone exchange, chromatin structure and the regulation of transcription. *Nat. Rev. Mol. Cell Biol.* **16**, 178–189 (2015).

62. Bintu, L. *et al.* Nucleosomal Elements that Control the Topography of the Barrier to Transcription. *Cell* **151**, 738–749 (2012).

63. Kizer, K. O. *et al.* A Novel Domain in Set2 Mediates RNA Polymerase II Interaction and Couples Histone H3 K36 Methylation with Transcript Elongation. *Mol. Cell. Biol.* **25**, 3305–3316 (2005).

64. Vakoc, C. R., Sachdeva, M. M., Wang, H. & Blobel, G. A. Profile of Histone Lysine Methylation across Transcribed Mammalian Chromatin. *Mol. Cell. Biol.* **26**, 9185–9195 (2006).

65. Luco, R. F. *et al.* Regulation of Alternative Splicing by Histone Modifications. *Science* **327**, 996–1000 (2010).

66. Scott, W. A. & Campos, E. I. Interactions With Histone H3 & Tools to Study Them. *Front. Cell Dev. Biol.* **8**, (2020).

67. Chow, C.-M. *et al.* Variant histone H3.3 marks promoters of transcriptionally active genes during mammalian cell division. *EMBO Rep.* **6**, 354–360 (2005).

68. Deal, R. B., Henikoff, J. G. & Henikoff, S. Genome-Wide Kinetics of Nucleosome Turnover Determined by Metabolic Labeling of Histones. *Science* **328**, 1161–1164 (2010).

69. Smolle, M. *et al.* Chromatin remodelers Isw1 and Chd1 maintain chromatin structure during transcription by preventing histone exchange. *Nat. Struct. Mol. Biol.* **19**, 884–892 (2012).

70. Kulaeva, O. I., Hsieh, F.-K. & Studitsky, V. M. RNA polymerase complexes cooperate to relieve the nucleosomal barrier and evict histones. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 11325–11330 (2010).

71. Kulaeva, O. I., Hsieh, F.-K., Chang, H.-W., Luse, D. S. & Studitsky, V. M. Mechanism of transcription through a nucleosome by RNA polymerase II. *Biochim. Biophys. Acta BBA - Gene Regul. Mech.* **1829**, 76–83 (2013).

72. Carrozza, M. J. *et al.* Histone H3 Methylation by Set2 Directs Deacetylation of Coding Regions by Rpd3S to Suppress Spurious Intragenic Transcription. *Cell* **123**, 581–592 (2005).

73. Narlikar, G. J., Sundaramoorthy, R. & Owen-Hughes, T. Mechanisms and functions of ATP-dependent chromatin-remodeling enzymes. *Cell* **154**, 490–503 (2013).

74. Valencia, A. M. *et al.* Recurrent SMARCB1 Mutations Reveal a Nucleosome Acidic Patch Interaction Site That Potentiates mSWI/SNF Complex Chromatin Remodeling. *Cell* **179**, 1342-1356.e23 (2019).

75. Dann, G. P. *et al.* ISWI chromatin remodellers sense nucleosome modifications to determine substrate preference. *Nature* **548**, 607–611 (2017).

76. Oberbeckmann, E. *et al.* Ruler elements in chromatin remodelers set nucleosome array spacing and phasing. *Nat. Commun.* **12**, 3232 (2021).

77. Becker, P. B. & Workman, J. L. Nucleosome Remodeling and Epigenetics. *Cold Spring Harb. Perspect. Biol.* **5**, a017905 (2013).

78. Alfert, A., Moreno, N. & Kerl, K. The BAF complex in development and disease. *Epigenetics Chromatin* **12**, 19 (2019).

79. Weber, C. M. *et al.* mSWI/SNF promotes Polycomb repression both directly and through genome-wide redistribution. *Nat. Struct. Mol. Biol.* **28**, 501–511 (2021).

80. Mashtalir, N. *et al.* Chromatin landscape signals differentially dictate the activities of mSWI/SNF family complexes. *Science* **373**, 306–315 (2021).

81. Michel, B. C. *et al.* A non-canonical SWI/SNF complex is a synthetic lethal target in cancers driven by BAF complex perturbation. *Nat. Cell Biol.* **20**, 1410–1420 (2018).

82. Vierstra, J. *et al.* Global reference mapping of human transcription factor footprints. *Nature* **583**, 729–736 (2020).

83. Stergachis, A. B. *et al.* Developmental Fate and Cellular Maturity Encoded in Human Regulatory DNA Landscapes. *Cell* **154**, 888–903 (2013).

84. Li, G. & Widom, J. Nucleosomes facilitate their own invasion. *Nat. Struct. Mol. Biol.* **11**, 763–769 (2004).

85. Soufi, A. *et al.* Pioneer Transcription Factors Target Partial DNA Motifs on Nucleosomes to Initiate Reprogramming. *Cell* **161**, 555–568 (2015).

86. Zaret, K. S. & Mango, S. E. Pioneer transcription factors, chromatin dynamics, and cell fate control. *Curr. Opin. Genet. Dev.* **37**, 76–81 (2016).

87. Isbel, L., Grand, R. S. & Schübeler, D. Generating specificity in genome regulation through transcription factor sensitivity to chromatin. *Nat. Rev. Genet.* 1–13 (2022) doi:10.1038/s41576-022-00512-6.

88. King, H. W. & Klose, R. J. The pioneer factor OCT4 requires the chromatin remodeller BRG1 to support gene regulatory element function in mouse embryonic stem cells. *eLife* **6**, e22631 (2017).

89. Iurlaro, M. *et al.* Mammalian SWI/SNF continuously restores local accessibility to chromatin. *Nat. Genet.* **53**, 279–287 (2021).

90. Mivelaz, M. *et al.* Chromatin Fiber Invasion and Nucleosome Displacement by the Rap1 Transcription Factor. *Mol. Cell* **77**, 488-500.e9 (2020).

91. Swinstead, E. E., Paakinaho, V., Presman, D. M. & Hager, G. L. Pioneer factors and ATP-dependent chromatin remodeling factors interact dynamically: A new perspective. *BioEssays* **38**, 1150–1157 (2016).

92. Perino, M. & Veenstra, G. J. C. Chromatin Control of Developmental Dynamics and Plasticity. *Dev. Cell* **38**, 610–620 (2016).

93. Pujadas, E. & Feinberg, A. P. Regulated Noise in the Epigenetic Landscape of Development and Disease. *Cell* **148**, 1123–1131 (2012).

94. Lu, Y. *et al.* Epigenetic regulation in human cancer: the potential role of epi-drug in cancer therapy. *Mol. Cancer* **19**, 79 (2020).

95. Miranda, T. B. *et al.* Reprogramming the Chromatin Landscape: Interplay of the Estrogen and Glucocorticoid Receptors at the Genomic Level. *Cancer Res.* **73**, 5130–5139 (2013).

96. Urbanucci, A. *et al.* Androgen Receptor Deregulation Drives Bromodomain-Mediated Chromatin Alterations in Prostate Cancer. *Cell Rep.* **19**, 2045–2059 (2017).

97. Pomerantz, M. M. *et al.* The androgen receptor cistrome is extensively reprogrammed in human prostate tumorigenesis. *Nat. Genet.* **47**, 1346–1351 (2015).

98. Adams, E. J. *et al.* FOXA1 mutations alter pioneering activity, differentiation and prostate cancer phenotypes. *Nature* **571**, 408–412 (2019).

99. Sharma, S. V. *et al.* A Chromatin-Mediated Reversible Drug-Tolerant State in Cancer Cell Subpopulations. *Cell* **141**, 69–80 (2010).

100. Baylin, S. B. & Jones, P. A. A decade of exploring the cancer epigenome — biological and translational implications. *Nat. Rev. Cancer* **11**, 726–734 (2011).

101. Rideout, W. M., Coetzee, G. A., Olumi, A. F. & Jones, P. A. 5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes. *Science* **249**, 1288–1290 (1990).

102. Bell, D. *et al.* Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).

103. Glodzik, D. *et al.* Comprehensive molecular comparison of BRCA1 hypermethylated and BRCA1 mutated triple negative breast cancers. *Nat. Commun.* **11**, 3747 (2020).

104. Hammerman, P. S. *et al.* Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).

105. Issa, J.-P. J. *et al.* Phase 1 study of low-dose prolonged exposure schedules of the hypomethylating agent 5-aza-2′-deoxycytidine (decitabine) in hematopoietic malignancies. *Blood* **103**, 1635–1640 (2004).

106. Flavahan, W. A. *et al.* Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**, 110–114 (2016).

107. Comet, I., Riising, E. M., Leblanc, B. & Helin, K. Maintaining cell identity: PRC2-mediated regulation of transcription and cancer. *Nat. Rev. Cancer* **16**, 803–810 (2016).

108. Xu, B., Konze, K. D., Jin, J. & Wang, G. G. Targeting EZH2 and PRC2 dependence as novel anticancer therapy. *Exp. Hematol.* **43**, 698–712 (2015).

109. Sashida, G. *et al.* Ezh2 loss promotes development of myelodysplastic syndrome but attenuates its predisposition to leukaemic transformation. *Nat. Commun.* **5**, 4177 (2014).

110. Velichutina, I. *et al.* EZH2-mediated epigenetic silencing in germinal center B cells contributes to proliferation and lymphomagenesis. *Blood* **116**, 5247–5255 (2010).

111. Cheng, Y. *et al.* Targeting epigenetic regulators for cancer therapy: mechanisms and advances in clinical trials. *Signal Transduct. Target. Ther.* **4**, 1–39 (2019).

112. Ding, X. *et al.* Genomic and Epigenomic Features of Primary and Recurrent Hepatocellular Carcinomas. *Gastroenterology* **157**, 1630-1645.e6 (2019).

113. Flavahan, W. A., Gaskell, E. & Bernstein, B. E. Epigenetic plasticity and the hallmarks of cancer. *Science* **357**, eaal2380 (2017).

114. Tam, W. L. & Weinberg, R. A. The epigenetics of epithelial-mesenchymal plasticity in cancer. *Nat. Med.* **19**, 1438–1449 (2013).

115. Mani, S. A. *et al.* The Epithelial-Mesenchymal Transition Generates Cells with Properties of Stem Cells. *Cell* **133**, 704–715 (2008).

116. Graff, J. R., Gabrielson, E., Fujii, H., Baylin, S. B. & Herman, J. G. Methylation Patterns of the E-cadherin 5′ CpG Island Are Unstable and Reflect the Dynamic, Heterogeneous Loss of E-cadherin Expression during Metastatic Progression*. *J. Biol. Chem.* **275**, 2727–2732 (2000).

117. Bell, R. E. *et al.* Enhancer methylation dynamics contribute to cancer plasticity and patient mortality. *Genome Res.* **26**, 601–611 (2016).

118. Kadoch, C. *et al.* Proteomic and bioinformatic analysis of mammalian SWI/SNF complexes identifies extensive roles in human malignancy. *Nat. Genet.* **45**, 592–601 (2013).

119. Kadoch, C. & Crabtree, G. R. Mammalian SWI/SNF chromatin remodeling complexes and cancer: Mechanistic insights gained from human genomics. *Sci. Adv.* **1**, e1500447 (2015).

120. Bultman, S. *et al.* A Brg1 Null Mutation in the Mouse Reveals Functional Differences among Mammalian SWI/SNF Complexes. *Mol. Cell* **6**, 1287–1295 (2000).

121. Wang, X. *et al.* SMARCB1-mediated SWI/SNF complex function is essential for enhancer regulation. *Nat. Genet.* **49**, 289–295 (2017).

122. Nakayama, R. T. *et al.* SMARCB1 is required for widespread BAF complex–mediated activation of enhancers and bivalent promoters. *Nat. Genet.* **49**, 1613–1623 (2017).

123. Wilson, B. G. *et al.* Epigenetic antagonism between polycomb and SWI/SNF complexes during oncogenic transformation. *Cancer Cell* **18**, 316–328 (2010).

124. Xiao, L. *et al.* Targeting SWI/SNF ATPases in enhancer-addicted prostate cancer. *Nature* **601**, 434–439 (2022).

125. Otto, J. E. *et al.* Structural and functional properties of mSWI/SNF chromatin remodeling complexes revealed through single-cell perturbation screens. *Mol. Cell* **83**, 1350-1367.e7 (2023).

126. Shendure, J. *et al.* DNA sequencing at 40: past, present and future. *Nature* **550**, 345–353 (2017).

127. NovaSeq 6000 Sequencing System.

128. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

129. Venter, J. C. *et al.* The Sequence of the Human Genome. *Science* **291**, 1304–1351 (2001).

130. Zerbino, D. R. Using the Velvet de novo assembler for short-read sequencing technologies. *Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al* **CHAPTER**, Unit-11.5 (2010).

131. Dewey, F. E. *et al.* Clinical Interpretation and Implications of Whole-Genome Sequencing. *JAMA* **311**, 1035–1045 (2014).

132. Austin-Tse, C. A. *et al.* Best practices for the interpretation and reporting of clinical whole genome sequencing. *Npj Genomic Med.* **7**, 1–13 (2022).

133. Aaltonen, L. A. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).

134. Kitzman, J. O. *et al.* Noninvasive Whole-Genome Sequencing of a Human Fetus. *Sci. Transl. Med.* **4**, 137ra76-137ra76 (2012).

135. Markus, H. *et al.* Evaluation of pre-analytical factors affecting plasma DNA analysis. *Sci. Rep.* **8**, 7375 (2018).

136. Sasieni, P. *et al.* Modelled mortality benefits of multi-cancer early detection screening in England. *Br. J. Cancer* 1–9 (2023) doi:10.1038/s41416-023-02243-9.

137. Alexander, G. E. *et al.* Analytical validation of a multi-cancer early detection test with cancer signal origin using a cell-free DNA–based targeted methylation assay. *PLOS ONE* **18**, e0283001 (2023).

138. Hess, J. F. *et al.* Library preparation for next generation sequencing: A review of automation strategies. *Biotechnol. Adv.* **41**, 107537 (2020).

139. Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol.* **11**, 1–17 (2010).

140. Picelli, S. *et al.* Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–2040 (2014).

141. Ludwig, L. S. *et al.* Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. *Cell* **176**, 1325-1339.e22 (2019).

142. Lareau, C. A. *et al.* Massively parallel single-cell mitochondrial DNA genotyping and chromatin profiling. *Nat. Biotechnol.* **39**, 451–461 (2021).

143. Emiliani, F. E., Hsu, I. & McKenna, A. Circuit-seq: Circular reconstruction of cut in vitro transposed plasmids using Nanopore sequencing. *bioRxiv* (2022) doi:10.1101/2022.01.25.477550.

144. Hennig, B. P. *et al.* Large-Scale Low-Cost NGS Library Preparation Using a Robust Tn5 Purification and Tagmentation Protocol. *G3 GenesGenomesGenetics* **8**, 79–89 (2017).

145. Xiong, K. *et al. Duplex-Repair enables highly accurate sequencing, despite DNA damage*. http://biorxiv.org/lookup/doi/10.1101/2021.05.21.445162 (2021) doi:10.1101/2021.05.21.445162.

146. Gansauge, M.-T. & Meyer, M. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nat. Protoc.* **8**, 737–748 (2013).

147. Artegiani, B. *et al.* Fast and efficient generation of knock-in human organoids using homology-independent CRISPR–Cas9 precision genome editing. *Nat. Cell Biol.* **22**, 321–331 (2020).

148. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11**, 163–166 (2014).

149. Abascal, F. *et al.* Somatic mutation landscapes at single-molecule resolution. *Nature* **593**, 405–410 (2021).

150. Bae, J. H. *et al.* Single duplex DNA sequencing with CODEC detects mutations with high sensitivity. *Nat. Genet.* 1–9 (2023) doi:10.1038/s41588-023-01376-0.

151. Kennedy, S. R. *et al.* Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat. Protoc.* **9**, 2586–2606 (2014).

152. Galas, D. J. & Schmitz, A. DNAase footprinting a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res.* **5**, 3157–3170 (1978).

153. Dingwall, C., Lomonossoff, G. P. & Laskey, R. A. High sequence specificity of micrococcal nuclease. *Nucleic Acids Res.* **9**, 2659–2674 (1981).

154. Henikoff, J. G., Belsky, J. A., Krassovsky, K., MacAlpine, D. M. & Henikoff, S. Epigenome characterization at single base-pair resolution. *Proc. Natl. Acad. Sci.* **108**, 18318–18323 (2011).

155. Crawford, G. E. *et al.* Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res.* **16**, 123–131 (2006).

156. Boyle, A. P. *et al.* High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell* **132**, 311–322 (2008).

157. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

158. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).

159. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr. Protoc. Mol. Biol.* **109**, 21.29.1-21.29.9 (2015).

160. Hauberg, M. E. *et al.* Common schizophrenia risk variants are enriched in open chromatin regions of human glutamatergic neurons. *Nat. Commun.* **11**, 5581 (2020).

161. Fullard, J. F. *et al.* An atlas of chromatin accessibility in the adult human brain. *Genome Res.* **28**, 1243–1252 (2018).

162. Chereji, R. V., Ramachandran, S., Bryson, T. D. & Henikoff, S. Precise genome-wide mapping of single nucleosomes and linkers in vivo. *Genome Biol.* **19**, 19 (2018).

163. Voong, L. N. *et al.* Insights into Nucleosome Organization in Mouse Embryonic Stem Cells through Chemical Mapping. *Cell* **167**, 1555-1570.e15 (2016).

164. Brogaard, K. R., Xi, L., Wang, J.-P. & Widom, J. Chapter Fourteen - A Chemical Approach to Mapping Nucleosomes at Base Pair Resolution in Yeast. in *Methods in Enzymology* (eds. Wu, C. & Allis, C. D.) vol. 513 315–334 (Academic Press, 2012).

165. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science* **316**, 1497–1502 (2007).

166. Wahba, L., Costantino, L., Tan, F. J., Zimmer, A. & Koshland, D. S1-DRIP-seq identifies high expression and polyA tracts as major contributors to R-loop formation. *Genes Dev.* **30**, 1327–1338 (2016).

167. Hänsel-Hertsch, R. *et al.* G-quadruplex structures mark human regulatory chromatin. *Nat. Genet.* **48**, 1267–1272 (2016).

168. Hänsel-Hertsch, R., Spiegel, J., Marsico, G., Tannahill, D. & Balasubramanian, S. Genome-wide mapping of endogenous G-quadruplex DNA structures by chromatin immunoprecipitation and high-throughput sequencing. *Nat. Protoc.* **13**, 551–564 (2018).

169. Klein, D. C. & Hainer, S. J. Genomic methods in profiling DNA accessibility and factor localization. *Chromosome Res.* **28**, 69–85 (2020).

170. Skene, P. J. & Henikoff, S. A simple method for generating high-resolution maps of genome-wide protein binding. *eLife* **4**, e09225 (2015).

171. Rhee, H. S. & Pugh, B. F. Comprehensive Genome-wide Protein-DNA Interactions Detected at Single Nucleotide Resolution. *Cell* **147**, 1408–1419 (2011).

172. Kasinathan, S., Orsi, G. A., Zentner, G. E., Ahmad, K. & Henikoff, S. High-resolution mapping of transcription factor binding sites on native chromatin. *Nat. Methods* **11**, 203–209 (2014).

173. Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife* **6**, e21856 (2017).

174. Skene, P. J., Henikoff, J. G. & Henikoff, S. Targeted in situ genome-wide profiling with high efficiency for low cell numbers. *Nat. Protoc.* **13**, 1006–1019 (2018).

175. Kaya-Okur, H. S. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* **10**, 1930 (2019).

176. Kaya-Okur, H. S., Janssens, D. H., Henikoff, J. G., Ahmad, K. & Henikoff, S. Efficient low-cost chromatin profiling with CUT&Tag. *Nat. Protoc.* **15**, 3264–3283 (2020).

177. Meers, M. P., Llagas, G., Janssens, D. H., Codomo, C. A. & Henikoff, S. Multifactorial profiling of epigenetic landscapes at single-cell resolution using MulTI-Tag. *Nat. Biotechnol.* 1–9 (2022) doi:10.1038/s41587-022-01522-9.

178. Hu, D. *et al.* CUT&Tag recovers up to half of ENCODE ChIP-seq peaks in modifications of H3K27. 2022.03.30.486382 Preprint at https://doi.org/10.1101/2022.03.30.486382 (2023).

179. Li, Y. *et al.* Chromatin and transcription factor profiling in rare stem cell populations using CUT&Tag. *STAR Protoc.* **2**, 100751 (2021).

180. Frommer, M. *et al.* A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci.* **89**, 1827–1831 (1992).

181. Vaisvila, R. *et al.* Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. *Genome Res.* **31**, 1280–1289 (2021).

182. Liu, Y. *et al.* Accurate targeted long-read DNA methylation and hydroxymethylation sequencing with TAPS. *Genome Biol.* **21**, 54 (2020).

183. Liu, Y. *et al.* Bisulfite-free direct detection of 5-methylcytosine and 5-hydroxymethylcytosine at base resolution. *Nat. Biotechnol.* **37**, 424–429 (2019).

184. Füllgrabe, J. *et al.* Simultaneous sequencing of genetic and epigenetic bases in DNA. *Nat. Biotechnol.* 1–8 (2023) doi:10.1038/s41587-022-01652-0.

185. Lyko, F. The DNA methyltransferase family: a versatile toolkit for epigenetic regulation. *Nat. Rev. Genet.* **19**, 81–92 (2018).

186. Sánchez-Romero, M. A., Cota, I. & Casadesús, J. DNA methylation in bacteria: from the methyl group to the methylome. *Curr. Opin. Microbiol.* **25**, 9–16 (2015).

187. Wu, T. P. *et al.* DNA methylation on N6-adenine in mammalian embryonic stem cells. *Nature* **532**, 329–333 (2016).

188. Xiao, C.-L. *et al.* N6-Methyladenine DNA Modification in the Human Genome. *Mol. Cell* **71**, 306-318.e7 (2018).

189. Steensel, B. van & Henikoff, S. Identification of in vivo DNA targets of chromatin proteins using tethered Dam methyltransferase. *Nat. Biotechnol.* **18**, 424–428 (2000).

190. van Steensel, B., Delrow, J. & Henikoff, S. Chromatin profiling using targeted DNA adenine methyltransferase. *Nat. Genet.* **27**, 304–308 (2001).

191. Murray, I. A. *et al.* The non-specific adenine DNA methyltransferase M.EcoGII. *Nucleic Acids Res.* **46**, 840–848 (2018).

192. Drozdz, M., Piekarowicz, A., Bujnicki, J. M. & Radlinska, M. Novel non-specific DNA adenine methyltransferases. *Nucleic Acids Res.* **40**, 2119–2130 (2012).

193. Sobecki, M. *et al.* MadID, a Versatile Approach to Map Protein-DNA Interactions, Highlights Telomere-Nuclear Envelope Contact Sites in Human Cells. *Cell Rep.* **25**, 2891-2903.e5 (2018).

194. Aughey, G. N., Estacio Gomez, A., Thomson, J., Yin, H. & Southall, T. D. CATaDa reveals global remodelling of chromatin accessibility during stem cell differentiation in vivo. *eLife* **7**, e32341 (2018).

195. Kelly, T. K. *et al.* Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res.* **22**, 2497–2506 (2012).

196. Shu, X. *et al.* A Genetically Encoded Tag for Correlated Light and Electron Microscopy of Intact Cells, Tissues, and Organisms. *PLOS Biol.* **9**, e1001041 (2011).

197. Ruiz-González, R. *et al.* Singlet Oxygen Generation by the Genetically Encoded Tag miniSOG. *J. Am. Chem. Soc.* **135**, 9564–9567 (2013).

198. Stergachis, A. B., Debo, B. M., Haugen, E., Churchman, L. S. & Stamatoyannopoulos, J. A. Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science* **368**, 1449–1454 (2020).

199. Abdulhay, N. J. *et al.* Massively multiplex single-molecule oligonucleosome footprinting. *eLife* **9**, e59404 (2020).

200. Lee, I. *et al.* Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *Nat. Methods* **17**, 1191–1199 (2020).

201. Abdulhay, N. J. *et al.* Single-fiber nucleosome density shapes the regulatory output of a mammalian chromatin remodeling enzyme. 2021.12.10.472156 Preprint at https://doi.org/10.1101/2021.12.10.472156 (2021).

202. Shipony, Z. *et al.* Long-range single-molecule mapping of chromatin accessibility in eukaryotes. *Nat. Methods* **17**, 319–327 (2020).

203. Wang, Y. *et al.* Single-molecule long-read sequencing reveals the chromatin basis of gene expression. *Genome Res.* **29**, 1329–1342 (2019).

204. Wang, Y., Zhao, Y., Bollas, A., Wang, Y. & Au, K. F. Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* **39**, 1348–1365 (2021).

205. PromethION. *Oxford Nanopore Technologies* https://nanoporetech.com/products/promethion.

206. Payne, A., Holmes, N., Rakyan, V. & Loose, M. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics* **35**, 2193–2198 (2019).

207. Liu, Q. *et al.* Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. *Nat. Commun.* **10**, 2449 (2019).

208. McIntyre, A. B. R. *et al.* Single-molecule sequencing detection of N6-methyladenine in microbial reference materials. *Nat. Commun.* **10**, 579 (2019).

209. Simpson, J. T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017).

210. Loose, M., Malla, S. & Stout, M. Real-time selective sequencing using nanopore technology. *Nat. Methods* **13**, 751–754 (2016).

211. Payne, A. *et al.* Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nat. Biotechnol.* **39**, 442–450 (2021).

212. Goenka, S. D. *et al.* Accelerated identification of disease-causing variants with ultra-rapid nanopore genome sequencing. *Nat. Biotechnol.* **40**, 1035–1041 (2022).

213. Gorzynski, J. E. *et al.* Ultrarapid Nanopore Genome Sequencing in a Critical Care Setting. *N. Engl. J. Med.* **386**, 700–702 (2022).

214. Galey, M. *et al.* 3-hour genome sequencing and targeted analysis to rapidly assess genetic risk. 2022.09.09.22279746 Preprint at https://doi.org/10.1101/2022.09.09.22279746 (2022).

215. Accuracy. *Oxford Nanopore Technologies* https://nanoporetech.com/accuracy.

216. The power of Q20+ chemistry. *Oxford Nanopore Technologies* https://nanoporetech.com/q20plus-chemistry (2021).

217. Eid, J. *et al.* Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* **323**, 133–138 (2009).

218. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).

219. PacBio Revio | Long-read sequencing at scale. https://www.pacb.com/revio/.

220. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).

221. Gershman, A. *et al.* Epigenetic patterns in a complete human genome. *Science* **376**, eabj5089 (2022).

222. Aganezov, S. *et al.* A complete reference genome improves analysis of human genetic variation. *Science* **376**, eabl3533 (2022).

223. Loomis, E. W. *et al.* Sequencing the unsequenceable: Expanded CGG-repeat alleles of the fragile X gene. *Genome Res.* **23**, 121–128 (2013).

224. TRGT: Tandem Repeat Genotyper. (2023).

225. Chen, X. *et al.* Comprehensive SMN1 and SMN2 profiling for spinal muscular atrophy analysis using long-read PacBio HiFi sequencing. *Am. J. Hum. Genet.* **110**, 240–250 (2023).

226. Feng, Z. *et al.* Detecting DNA Modifications from SMRT Sequencing Data by Modeling Sequence Context Dependence of Polymerase Kinetic. *PLOS Comput. Biol.* **9**, e1002935 (2013).

227. Clark, T. A. *et al.* Enhanced 5-methylcytosine detection in single-molecule, real-time sequencing via Tet1 oxidation. *BMC Biol.* **11**, 1–10 (2013).

228. Jha, A. *et al.* Fibertools: fast and accurate DNA-m6A calling using single-molecule long-read sequencing. 2023.04.20.537673 Preprint at https://doi.org/10.1101/2023.04.20.537673 (2023).

229. Ni, P. *et al.* DNA 5-methylcytosine detection and methylation phasing using PacBio circular consensus sequencing. 2022.02.26.482074 Preprint at https://doi.org/10.1101/2022.02.26.482074 (2023).

230. Kozarewa, I. *et al.* Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of GC-biased genomes. *Nat. Methods* **6**, 291–295 (2009).

231. Illumina DNA PCR-Free Prep | For sensitive WGS applications. https://www.illumina.com/products/by-type/sequencing-kits/library-prep-kits/dna-pcr-free-prep.html.

232. Library prep + barcoding. *PacBio* https://www.pacb.com/products-and-services/consumables/library-prep-and-barcoding-kits/.

233. Nanda, A. S. *et al.* Sensitive multimodal profiling of native DNA by transposase-mediated single-molecule sequencing. 2022.08.07.502893 Preprint at https://doi.org/10.1101/2022.08.07.502893 (2022).

234. Reznikoff, W. S. Tn5 as a model for understanding DNA transposition. *Mol. Microbiol.* **47**, 1199–1206 (2003).

235. Chen, X. *et al.* ATAC-see reveals the accessible genome by transposase-mediated imaging and sequencing. *Nat. Methods* **13**, 1013–1020 (2016).

236. Xie, L. *et al.* 3D ATAC-PALM: super-resolution imaging of the accessible genome. *Nat. Methods* **17**, 430–436 (2020).

237. Chen, Z. *et al.* Ultra-low input single tube linked-read library method enables short-read second-generation sequencing systems to generate highly accurate and economical long-range sequencing information routinely. *Genome Res.* gr.260380.119 (2020) doi:10.1101/gr.260380.119.

238. Laks, E. *et al.* Clonal Decomposition and DNA Replication States Defined by Scaled Single-Cell Genome Sequencing. *Cell* **179**, 1207-1221.e22 (2019).

239. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).

240. Vollger, M. R. Segmental duplications and their variation in a complete human genome. *Science* **376**, 6965 (2022).

241. Altemose, N. *et al.* DiMeLo-seq: a long-read, single-molecule method for mapping protein–DNA interactions genome wide. *Nat. Methods* **19**, 711–723 (2022).

242. Au, K. F. Characterization of the human ESC transcriptome by hybrid sequencing. *Proc Natl Acad Sci U A* **110**, 4821–30 (2013).

243. Sharon, D., Tilgner, H., Grubert, F. & Snyder, M. A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol* **31**, 1009–1014 (2013).

244. Quail, M. A. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**, 341 (2012).

245. Schmidl, C., Rendeiro, A. F., Sheffield, N. C. & Bock, C. ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. *Nat Methods* **12**, 963–965 (2015).

246. Chen, C. *et al.* Single-Cell Whole Genome Analyses by Linear Amplification via Transposon Insertion (LIANTI). *Science* **356**, 189–194 (2017).

247. Cusanovich, D. A. Epigenetics. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).

248. Cao, J. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018).

249. Yin, Y. *et al.* High-Throughput Single-Cell Sequencing with Linear Amplification. *Mol. Cell* **76**, 676-690.e10 (2019).

250. Minussi, D. C. Breast tumors maintain a reservoir of subclonal diversity during expansion. *Nature* **592**, 302–308 (2021).

251. Payne, A. C. *et al.* In situ genome sequencing resolves DNA sequence and structure in intact biological samples. *Science* **371**, eaay3446 (2021).

252. Flusberg, B. A. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods* **7**, 461–465 (2010).

253. Grandi, F. C., Modi, H., Kampman, L. & Corces, M. R. Chromatin accessibility profiling by ATAC-seq. *Nat Protoc* **17**, 1518–1552 (2022).

254. Sayles, L. C. *et al.* Genome-Informed Targeted Therapy for Osteosarcoma. *Cancer Discov.* **9**, 46–63 (2019).

255. Traag, V. A., Waltman, L. & Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* **9**, 5233 (2019).

256. Wang, H. Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res* **22**, 1680–1688 (2012).

257. Nguyen, H. G. *et al.* Development of a stress response therapy targeting aggressive prostate cancer. *Sci. Transl. Med.* **10**, eaar2036 (2018).

258. Alpsoy, A. *et al.* BRD9 Is a Critical Regulator of Androgen Receptor Signaling and Prostate Cancer Progression. *Cancer Res.* **81**, 820–833 (2021).

259. Shan, Z. *et al.* CTCF regulates the FoxO signaling pathway to affect the progression of prostate cancer. *J. Cell. Mol. Med.* **23**, 3130–3139 (2019).

260. Wang, T. *et al.* Integrative Epigenome Map of the Normal Human Prostate Provides Insights Into Prostate Cancer Predisposition. *Front. Cell Dev. Biol.* **9**, (2021).

261. Meers, M. P., Bryson, T. D., Henikoff, J. G. & Henikoff, S. Improved CUT&RUN chromatin profiling tools. *eLife* **8**, e46314 (2019).

262. Gilpatrick, T. *et al.* Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nat. Biotechnol.* **38**, 433–438 (2020).

263. Al'Khafaji, A. M. *et al.* High-throughput RNA isoform sequencing using programmable cDNA concatenation. 2021.10.01.462818 Preprint at https://doi.org/10.1101/2021.10.01.462818 (2021).

264. Yu, H.-B., Johnson, R., Kunarso, G. & Stanton, L. W. Coassembly of REST and its cofactors at sites of gene repression in embryonic stem cells. *Genome Res* **21**, 1284–1293 (2011).

265. English, A. C., Menon, V. K., Gibbs, R. A., Metcalf, G. A. & Sedlazeck, F. J. Truvari: refined structural variant comparison preserves allelic diversity. *Genome Biol.* **23**, 1–20 (2022).

266. Ramani, V., Qiu, R. & Shendure, J. High Sensitivity Profiling of Chromatin Structure by MNase-SSP. *Cell Rep.* **26**, 2465-2476.e4 (2019).

267. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci.* **100**, 9440–9445 (2003).

268. Yadav, T., Quivy, J.-P. & Almouzni, G. Chromatin plasticity: A versatile landscape that underlies cell fate and identity. *Science* **361**, 1332–1336 (2018).

269. Valencia, A. M. & Kadoch, C. Chromatin regulatory mechanisms and therapeutic opportunities in cancer. *Nat. Cell Biol.* **21**, 152–161 (2019).

270. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).

271. Saleemuddin, M. & Husain, Q. Concanavalin A: A useful ligand for glycoenzyme immobilization—A review. *Enzyme Microb. Technol.* **13**, 290–295 (1991).

272. Domcke, S. *et al.* Competition between DNA methylation and transcription factors determines binding of NRF1. *Nature* **528**, 575–579 (2015).

273. Tóth, G., Gáspári, Z. & Jurka, J. Microsatellites in Different Eukaryotic Genomes: Survey and Analysis. *Genome Res.* **10**, 967–981 (2000).

274. Srivastava, S., Avvaru, A. K., Sowpati, D. T. & Mishra, R. K. Patterns of microsatellite distribution across eukaryotic genomes. *BMC Genomics* **20**, 153 (2019).

275. Dentro, S. C. *et al.* Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell* **184**, 2239-2254.e39 (2021).

276. THE TABULA SAPIENS CONSORTIUM. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).

277. Lomakin, A. *et al.* Spatial genomics maps the structure, nature and evolution of cancer clones. *Nature* **611**, 594–602 (2022).

278. Lee, M.-C. W. *et al.* Single-cell analyses of transcriptional heterogeneity during drug tolerance transition in cancer cells by RNA sequencing. *Proc. Natl. Acad. Sci.* **111**, E4726–E4735 (2014).

279. Cuomo, A. S. E. *et al.* Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nat. Commun.* **11**, 810 (2020).

280. Akkaya, M., Kwak, K. & Pierce, S. K. B cell memory: building two walls of protection against pathogens. *Nat. Rev. Immunol.* **20**, 229–238 (2020).

281. Kieffer-Kwon, K.-R. *et al.* Myc Regulates Chromatin Decompaction and Nuclear Architecture during B Cell Activation. *Mol. Cell* **67**, 566-578.e10 (2017).

282. Bossen, C. *et al.* The chromatin remodeler Brg1 activates enhancer repertoires to establish B cell identity and modulate cell growth. *Nat. Immunol.* **16**, 775–784 (2015).

283. Schmiedel, D., Hezroni, H., Hamburg, A. & Shulman, Z. Brg1 Supports B Cell Proliferation and Germinal Center Formation Through Enhancer Activation. *Front. Immunol.* **12**, (2021).

284. Lee, S.-T. *et al.* A global DNA methylation and gene expression analysis of early human B-cell development reveals a demethylation signature and transcription factor network. *Nucleic Acids Res.* **40**, 11339–11351 (2012).

285. Caron, G. *et al.* Cell-Cycle-Dependent Reconfiguration of the DNA Methylome during Terminal Differentiation of Human B Cells into Plasma Cells. *Cell Rep.* **13**, 1059–1071 (2015).

286. Kulis, M. *et al.* Whole-genome fingerprint of the DNA methylome during human B cell differentiation. *Nat. Genet.* **47**, 746–756 (2015).

287. Zatopek, K. M. *et al.* RADAR-seq: A RAre DAmage and Repair sequencing method for detecting DNA damage on a genome-wide scale. *DNA Repair* **80**, 36–44 (2019).

288. 'Giron' Koetsier, P. A. & Cantor, E. J. A simple approach for effective shearing and reliable concentration measurement of ultra-high-molecular-weight DNA. *BioTechniques* **71**, 439–444 (2021).

289. Clark, T. A., Spittle, K. E., Turner, S. W. & Korlach, J. Direct Detection and Sequencing of Damaged DNA Bases. *Genome Integr.* **2**, 10 (2011).

290. Guiblet, W. M. *et al.* Long-read sequencing technology indicates genome-wide effects of non-B DNA on polymerization speed and error rate. *Genome Res.* **28**, 1767–1778 (2018).

291. Inoue, J., Shigemori, Y. & Mikawa, T. Improvements of rolling circle amplification (RCA) efficiency and accuracy using Thermus thermophilus SSB mutant protein. *Nucleic Acids Res.* **34**, e69 (2006).

292. Ducani, C., Bernardinelli, G. & Högberg, B. Rolling circle replication requires single-stranded DNA binding protein to avoid termination and production of double-stranded DNA. *Nucleic Acids Res.* **42**, 10596–10604 (2014).

293. Ordóñez, C. D., Lechuga, A., Salas, M. & Redrejo-Rodríguez, M. Engineered viral DNA polymerase with enhanced DNA amplification capacity: a proof-of-concept of isothermal amplification of damaged DNA. *Sci. Rep.* **10**, 15046 (2020).

294. Rhee, M., Light, Y. K., Meagher, R. J. & Singh, A. K. Digital Droplet Multiple Displacement Amplification (ddMDA) for Whole Genome Sequencing of Limited DNA Samples. *PLOS ONE* **11**, e0153699 (2016).

295. Clegg, R. M., Loontiens, F. G., Van Landschoot, A. & Jovin, T. M. Binding kinetics of methyl .alpha.-D-mannopyranoside to concanavalin A: temperature-jump relaxation study with 4-methylumbelliferyl .alpha.-D-mannopyranoside as a fluorescence indicator ligand. *Biochemistry* **20**, 4687–4692 (1981).

296. Zhao, T. *et al.* Spatial genomics enables multi-modal study of clonal heterogeneity in tissues. *Nature* 1–7 (2021) doi:10.1038/s41586-021-04217-4.

297. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 1–5 (2018).

298. Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res.* **48**, D882–D889 (2020).

299. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).

300. Mamanova, L. *et al.* Target-enrichment strategies for next-generation sequencing. *Nat. Methods* **7**, 111–118 (2010).

301. Twist Alliance Clinical Research Exome. https://www.twistbioscience.com/resources/safety-data-sheet/twist-alliance-clinical-research-exome-349-mb-bed-files.

302. Steiert, T. A. *et al.* High-throughput method for the hybridisation-based targeted enrichment of long genomic fragments for PacBio third-generation sequencing. *NAR Genomics Bioinforma.* **4**, lqac051 (2022).

303. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).

304. Zook, J. M. *et al.* A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.* **38**, 1347–1355 (2020).

305. Wagner, J. *et al.* Curated variation benchmarks for challenging medically relevant autosomal genes. *Nat. Biotechnol.* **40**, 672–680 (2022).

306. Matzaraki, V., Kumar, V., Wijmenga, C. & Zhernakova, A. The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome Biol.* **18**, 76 (2017).

307. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med.* **9**, 75 (2017).

308. Zahn, H. *et al.* Scalable whole-genome single-cell library preparation without preamplification. *Nat. Methods* **14**, 167–173 (2017).

309. Lareau, C. A. *et al.* Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat. Biotechnol.* **37**, 916–924 (2019).

310. Mulqueen, R. M. *et al.* High-content single-cell combinatorial indexing. *Nat. Biotechnol.* **39**, 1574–1580 (2021).

311. Hitz, B. C. *et al.* The ENCODE Uniform Analysis Pipelines. 2023.04.04.535623 Preprint at https://doi.org/10.1101/2023.04.04.535623 (2023).

312. Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol.* **21**, 1–16 (2020).

313. Hagemann-Jensen, M., Ziegenhain, C. & Sandberg, R. Scalable single-cell RNA sequencing from full transcripts with Smart-seq3xpress. *Nat. Biotechnol.* **40**, 1452–1457 (2022).

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
Arjun Scott Nanda
E8253238B6D042F...
Author Signature

5/8/2023
Date