nature genetics

Technical Report

https://doi.org/10.1038/s41588-024-01748-0

Direct transposition of native DNA for sensitive multimodal single-molecule sequencing

Received: 9 August 2022

Accepted: 8 April 2024

Published online: 9 May 2024

Check for updates

Arjun S. Nanda[1,2,10], Ke Wu[1,10], Iryna Irkliyenko[1,10], Brian Woo[2,3], Megan S. Ostrowski[1], Andrew S. Clugston[3,4], Leanne C. Sayles[3,4], Lingru Xu[3], Ansuman T. Satpathy[5,6,7], Hao G. Nguyen[3], E. Alejandro Sweet-Cordero[3,4], Hani Goodarzi[2,3,6,8], Sivakanthan Kasinathan[7,9,10] & Vijay Ramani[1,2,3,8]

Concurrent readout of sequence and base modifications from long unamplified DNA templates by Pacific Biosciences of California (PacBio) single-molecule sequencing requires large amounts of input material. Here we adapt Tn5 transposition to introduce hairpin oligonucleotides and fragment (tagment) limiting quantities of DNA for generating PacBio-compatible circular molecules. We developed two methods that implement tagmentation and use 90–99% less input than current protocols: (1) single-molecule real-time sequencing by tagmentation (SMRT-Tag), which allows detection of genetic variation and CpG methylation; and (2) single-molecule adenine-methylated oligonucleosome sequencing assay by tagmentation (SAMOSA-Tag), which uses exogenous adenine methylation to add a third channel for probing chromatin accessibility. SMRT-Tag of 40 ng or more human DNA (approximately 7,000 cell equivalents) yielded data comparable to gold standard whole-genome and bisulfite sequencing. SAMOSA-Tag of 30,000–50,000 nuclei resolved single-fiber chromatin structure, CTCF binding and DNA methylation in patient-derived prostate cancer xenografts and uncovered metastasis-associated global epigenome disorganization. Tagmentation thus promises to enable sensitive, scalable and multimodal single-molecule genomics for diverse basic and clinical applications.

Third-generation single-molecule sequencing (SMS) technologies deliver accurate, multimodal readouts of genetic sequence and nucleobase modifications on kilobase (kb)-length to megabase-length nucleic acid templates[1]. SMS has facilitated the characterization of previously intractable structural variants and repetitive regions[2,3], assembly of gapless human genomes and high-resolution functional genomics of DNA[4–8] and RNA[9,10]. The intrinsic multimodality of SMS has been exploited by chromatin profiling methods, such as the

[1]Gladstone Institute for Data Science and Biotechnology, Gladstone Institutes, San Francisco, CA, USA. [2]Department of Biochemistry and Biophysics, University of California, San Francisco, San Francisco, CA, USA. [3]Helen-Diller Cancer Center, San Francisco, CA, USA. [4]Department of Pediatrics, University of California, San Francisco, San Francisco, CA, USA. [5]Department of Pathology, Stanford University, Stanford, CA, USA. [6]Parker Institute for Cancer Immunotherapy, San Francisco, CA, USA. [7]Gladstone-University of California, San Francisco Institute for Genomic Immunology, Gladstone Institutes, San Francisco, CA, USA. [8]Bakar Computational Health Sciences Institute, San Francisco, CA, USA. [9]Division of Rheumatology, Department of Pediatrics, Stanford University, Stanford, CA, USA. [10]These authors contributed equally: Arjun S. Nanda, Ke Wu, Iryna Irkliyenko, Sivakanthan Kasinathan. e-mail: skas@stanford.edu; vijay.ramani@gladstone.ucsf.edu

single-molecule adenine-methylated oligonucleosome sequencing assay (SAMOSA)[4,11], Fiber-seq[5], Nanopore sequencing of nucleosome occupancy and methylome[7] and others[6,8,12]. These approaches establish a paradigm for encoding functional genomic information (for example, histone–DNA and transcription factor–DNA interactions) as separate SMS 'channels' concurrently with primary sequence and endogenous epigenetic marks, such as CpG methylation.

Over the past decade, improvements in cost, data quality, read length and computational tools have driven rapid maturation of the Pacific Biosciences of California (PacBio) and Oxford Nanopore Technologies (ONT) SMS platforms. For example, the cost of PacBio sequencing has decreased from US$2,000 to US$35 per gigabase (Gb), concomitant with increases in yield (100 Mb to 90 Gb per instrument run), read length (from approximately 1.5 kb to 15–20 kb) and accuracy (from approximately 85% to more than 99.95%)[13]. A key limitation of PacBio SMS is the amount of input DNA required for PCR-free library preparation (typically at least 1–5 µg, or 150,000–750,000 human cells; Supplementary Note 1) owing to sample losses during mechanical or enzymatic fragmentation, adapter ligation and serial reaction cleanups. While low-input protocols are available, they typically rely on PCR amplification, which erases modified bases and may introduce biases. This obstacle has limited the primary use of SMS to genome assembly and medical genetics, precluding analyses of rare clinical samples and post-mitotic cell populations, single cells and microorganisms.

Simultaneous transposition of sequencing adapters and template DNA fragmentation (that is, tagmentation) using hyperactive Tn5 transposase poses an attractive solution to this problem[14]. The reduced input requirement and workflow complexity of Tn5-based short-read library preparation has transformed bulk genome, epigenome and transcriptome profiling[15–17] and enabled single-cell and spatial monoplex[18–20] and multiomic sequencing[21–23]. Reasoning that the high efficiency of tagmentation and consolidation of protocol steps would similarly facilitate low-input SMS, we optimized transposition of hairpin adapters to yield long circular molecules for PacBio sequencing[24]. We then applied this principle to develop two PCR-free multimodal methods: (1) single-molecule real-time sequencing by tagmentation (SMRT-Tag) for assaying the genome and epigenome; and (2) SAMOSA by tagmentation (SAMOSA-Tag), which adds a concurrent channel for mapping chromatin structure. SMRT-Tag accurately detected genetic and epigenetic variants from as little as 40 ng of DNA. SAMOSA-Tag maps of single-fiber CTCF and nucleosome occupancy and CpG methylation uncovered metastasis-associated global chromatin deregulation in technically challenging patient-derived xenografts (PDXs) from a patient with prostate cancer. These results extend tagmentation to PacBio library preparation and have the potential to enable sensitive, scalable and cellularly resolved single-molecule genomics.

Results

Tn5 transposition produces PacBio-compatible molecules

Two technical factors need to be addressed to efficiently generate long (>1 kb) molecules for PacBio SMS via transposition of hairpin adapters into genomic DNA (gDNA) (illustrated with the SMRT-Tag workflow; Fig. 1a). First, the conventional Tn5 enzyme used in many short-read sequencing methods optimally produces 100–500 bp fragments. Therefore, we selected a triple-mutant Tn5 enzyme (hereafter referred to as Tn5), which permits concentration-dependent control of fragment size[25]. We loaded Tn5 with custom oligonucleotides consisting of the hairpin PacBio adapter and mosaic end sequences needed to assemble transposomes. Analytical electrophoresis of gDNA tagmented with adapter-loaded Tn5 at varying reaction conditions confirmed generation of fragments more than 1-kb long, which are favored at low transposome concentrations and temperature (Fig. 1b). Additional considerations for controlling library size are detailed below and in Supplementary Note 2.

Second, Tn5 transposition introduces 9-nt gaps into template molecules[26] (Fig. 1a), which must be sealed for productive SMS. While hairpin transposition has been reported for short-read, single-cell genomics[18], and Tn5 is used in some ONT protocols, efficient gap repair to create closed, circular molecules has, to our knowledge, not been reported. We thus tested 62 conditions (Supplementary Table 1) to optimize gap filling. Two enzyme combinations proved to be the most robust based on yield (Supplementary Fig. 1) and electrophoretic fragment lengths (Supplementary Fig. 2) of gDNA subjected to tagmentation, repair and exonuclease digestion to select for closed circles: Phusion polymerase and Taq DNA ligase ('Phusion/Taq') and T4 DNA polymerase and Ampligase ('T4/Ampligase'). These produced exonuclease-resistant libraries from as little as 50 ng gDNA, with typical yields of more than 20% of input mass (Supplementary Table 2). In all subsequent experiments, we used Phusion/Taq because it provided significantly higher yields on gDNA than T4/Ampligase ($P = 0.0093$, two-sided t-test).

SMRT-Tag produces tunable libraries for multiplexed SMS

We applied direct transposition in SMRT-Tag, a simple method for whole-genome analysis, and explored library and sequencing characteristics. To evaluate the sequencing efficiency of SMRT-Tag, we tagmented 120 ng of HG002 gDNA (equivalent to approximately 20,000 human cells) in eight separate reactions and used solid-phase reversible immobilization (SPRI) beads to fractionate the resulting libraries for sequencing using PacBio's proprietary 2.1 and 2.2 polymerases optimized for short and long templates, respectively (Supplementary Note 2). Circular consensus sequencing (CCS) read length distributions of the 3,524,301 molecules (14.3 Gb total) sequenced over two runs were concordant with size selection and polymerase choice (Fig. 1c; 2,081 ± 935.8 bp versus 5,940 ± 3,097 bp for polymerases 2.1 and 2.2, respectively; mean ± s.d.). The per-read quality scores (Q-scores; Fig. 1d) and number of CCS passes (Fig. 1e) were sufficient for PacBio high-fidelity ('HiFi') sequencing with more than 99% (>Q20) base accuracy, which typically requires 5 or more redundant passes per molecule.

To assess demultiplexing using the 8-nt barcode included in the SMRT-Tag hairpin adapter (Fig. 1a), we first performed low-pass sequencing of libraries pooled after tagmentation, gap repair and exonuclease digestion of gDNA from the extensively genotyped HG002, HG003 and HG004 human trio (in total, seven 80-ng reactions sequenced to 0.75× HG002, 1.39× HG003 and 1.30× HG004 depths; Supplementary Fig. 3a). We inspected the left and right barcodes of molecules, which were identical (99.9% concordance; Supplementary Fig. 3b). Taking advantage of the pedigree to query genotype mixing of multiplexed libraries, we confirmed that HG003 and HG004 (unrelated parents) shared few private single-nucleotide variants (SNVs) (0.60% HG003 versus HG004; 0.67% HG004 versus HG003), while HG002 (child) was a mixture of parental genotypes (33.1% overlap; Supplementary Fig. 3c). Second, to determine if samples could be multiplexed immediately after tagmentation, we sequenced gDNA libraries from four separate reactions pooled before gap repair and exonuclease digestion (Supplementary Fig. 3d). Barcode concordance (99.9%; Supplementary Fig. 3e) and Smith–Waterman barcode alignment scores reported by the lima demultiplexer (mean = 97.9, s.d. = 6.78, normalized scale 0–100; Supplementary Fig. 3f) were excellent. This confirmed that there was no tagging of previously transposed molecules during gap repair, exonuclease cleanup and pooling, and was consistent with the zero turnover activity of Tn5.

Finally, to illustrate the tunability of SMRT-Tag, we tagmented gDNA at varying Tn5 concentrations and reaction temperatures, and multiplexed libraries for sequencing. The resulting read length distributions confirmed that the Tn5:DNA ratio and temperature can be varied to shift library size distributions (Supplementary Fig. 4). The mean and s.d. of fragment lengths were respectively controllable over nearly 11-fold and 18-fold dynamic ranges, offering an important

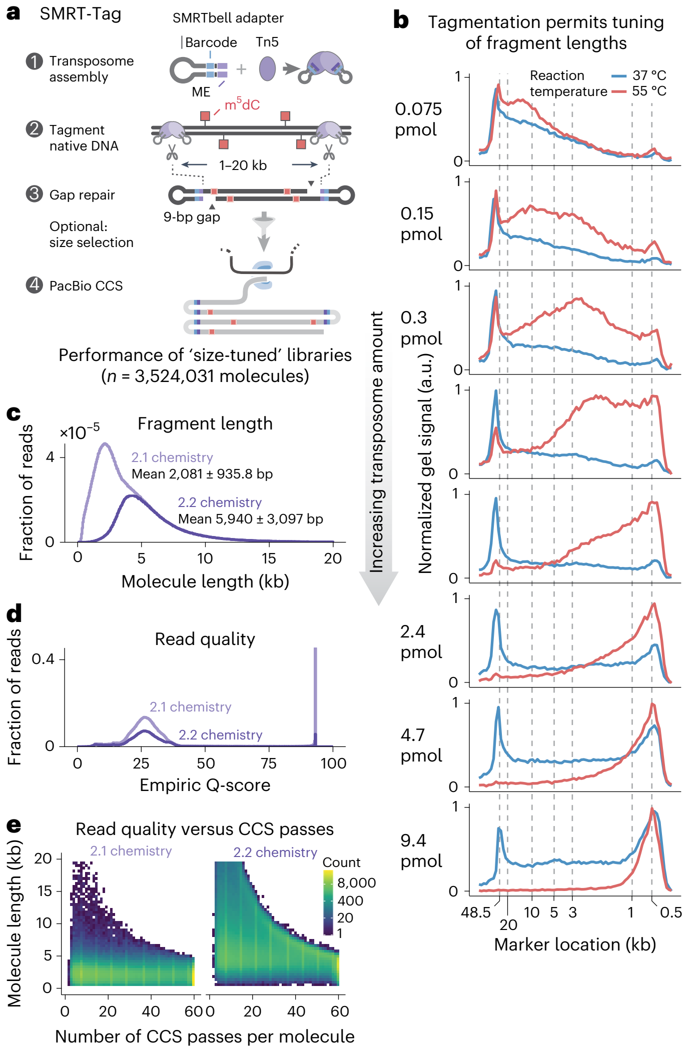

Fig. 1 | Tagmentation enables tunable and sensitive single-molecule real-time sequencing. a, In SMRT-Tag, hairpin adapter-loaded Tn5 transposase was used to fragment DNA into kb-scale fragments. The 9-nt gaps introduced by transposition were closed via optimized gap repair, while exonuclease digestion enriched for the covalently closed templates required for PacBio sequencing. **b**, Varying concentration of hairpin-loaded transposomes and reaction temperature tuned fragmentation of gDNA over a size range of 2–10 kb. **c**, PacBio CCS fragment lengths for SMRT-Tag libraries fractionated into short and long molecules optimal for PacBio polymerases 2.1 (light purple) and 2.2 (dark purple) chemistries, respectively. The distribution for the long-fragment library (2.2 chemistry) has a tail that extends beyond 20 kb. **d**, Empiric Q-score distributions for the 2.1 and 2.2 libraries. **e**, Heatmap of logarithmically scaled counts of CCS length as a function of the number of CCS passes per molecule.

reference point for implementing the approach (Supplementary Fig. 4c). For all experiments, unless otherwise noted, libraries were multiplexed to minimize sequencing cost. Supplementary Note 1 details the rationale for this, and the design choices for library preparation, polymerase binding and flow cell loading. Sequencing and quality metrics for all libraries and pooling strategies for analyses are shown in Supplementary Tables 3 and 4. We conclude that SMRT-Tag generates multiplexable PCR-free PacBio libraries from low-input DNA amounts for multiplex sequencing.

SMRT-Tag accurately detects genetic and epigenetic variation

We next sought to establish the sensitivity and variant calling accuracy of SMRT-Tag. We first determined whether libraries could be generated at the minimum on-plate loading concentration (OPLC) for PacBio Sequel II flow cells of 20–40 pM. We sequenced one SMRT-Tag library generated from 40 ng HG002 gDNA (approximately 7,000 human cell equivalents) achieving 37 pM OPLC (Fig. 2a and Supplementary Note 1). A single flow cell yielded 2,736,674 CCS reads with 2.32-kb median length, equivalent to approximately 2.43× genome coverage (Fig. 2b). While this depth is suboptimal for routine genotyping applications, we next asked whether data quality was sufficient for variant detection. We called SNVs and insertion-deletion (indel) variants using DeepVariant and structural variants (SVs) with pbsv from low-input SMRT-Tag and coverage-matched ligation-based libraries sequenced by the Genome in a Bottle (GIAB) consortium[27]. To evaluate accuracy, we benchmarked detected variants against the gold standard GIAB high-confidence HG002 callset[28] (Fig. 2c–e). Comparing SMRT-Tag and ligation-based libraries, we observed similar recall (0.420 versus 0.527 for SNVs and 0.338 versus 0.408 for indels), precision (0.870 versus 0.898 for SNVs and 0.785 versus 0.797 for indels) and F1 score (0.566 versus 0.664 for SNVs and 0.380 versus 0.539 for indels; Fig. 2c). Performance for SVs was slightly lower (recall 0.129 versus 0.25, precision 0.877 versus 0.879 and F1 score 0.225 versus 0.389; Fig. 2d) probably due to shorter reads affecting the resolution of large indels.

In PacBio SMS, nucleobase modifications are inferred from stereotyped changes in real-time polymerase kinetics during nucleotide addition, offering an opportunity for simultaneous genotyping and epigenotyping[29]. To assess detection of CpG methylation, we predicted the positions of 5-methyl-deoxycytidine (m^5dC) using PacBio's primrose software, which assigns methylation probabilities to CpGs via a convolutional neural network that combines kinetic data from multiple CCS passes. We compared primrose methylation calls from SMRT-Tag and ligation-based PacBio SMS against gold standard bisulfite sequencing data[30]. Per-CpG methylation calls were tightly correlated between the SMRT-Tag and bisulfite m^5dC datasets (Pearson's $r = 0.84$; Fig. 2e). Framing CpG methylation calling as a classification problem (Fig. 2f), we observed excellent performance measured by the area under the curve (AUC) (Fig. 2g), with the SMRT-Tag and ligation-based datasets demonstrating similar AUC (0.935 versus 0.926, respectively).

Finally, to compare performance at higher depths, we sequenced additional HG002 SMRT-Tag libraries to 11.2× median coverage (34.24 Gb on six Sequel II flow cells). We compared SNV, indel and SV calls from SMRT-Tag and coverage-matched ligation-based libraries against the GIAB HG002 benchmark. We found similar recall (0.970 SMRT-Tag versus 0.970 ligation-based PacBio for SNVs and 0.911 versus 0.907 for indels), precision (0.995 versus 0.995 for SNVs and 0.955 versus 0.949 for indels), F1 score (0.983 versus 0.982 for SNVs and 0.932 versus 0.928 for indels) and AUC (0.969 versus 0.968 for SNVs and 0.902 versus 0.897 for indels; Supplementary Fig. 5a–d). CpG methylation detected using high-coverage SMRT-Tag was on par with short-read bisulfite (Supplementary Fig. 5e) and ligation-based PacBio (Supplementary Fig. 5f) data. SMRT-Tag also resolved variants within segmental duplications, repeats, the major histocompatibility complex locus and other challenging regions (Supplementary Fig. 6a; F1 scores 0.977 SMRT-Tag versus 0.967 ligation-based PacBio for SNVs and 0.912 versus 0.905 for indels across all regions with differences probably due to sequencing chemistry) and at varying levels of coverage (Supplementary Fig. 6b). Taken together, these results demonstrate the strong technical concordance between tagmentation and ligation-based libraries and the sensitive detection of genetic and epigenetic variation by SMRT-Tag.

Single-fiber chromatin and methylation profiling with SAMOSA-Tag

Tagmentation is the basis for assay for transposase-accessible chromatin with sequencing (ATAC–seq), a popular method for profiling chromatin accessibility[16]. Reasoning that Tn5 could be used to lower the μg-range input needed for single-molecule chromatin accessibility assays

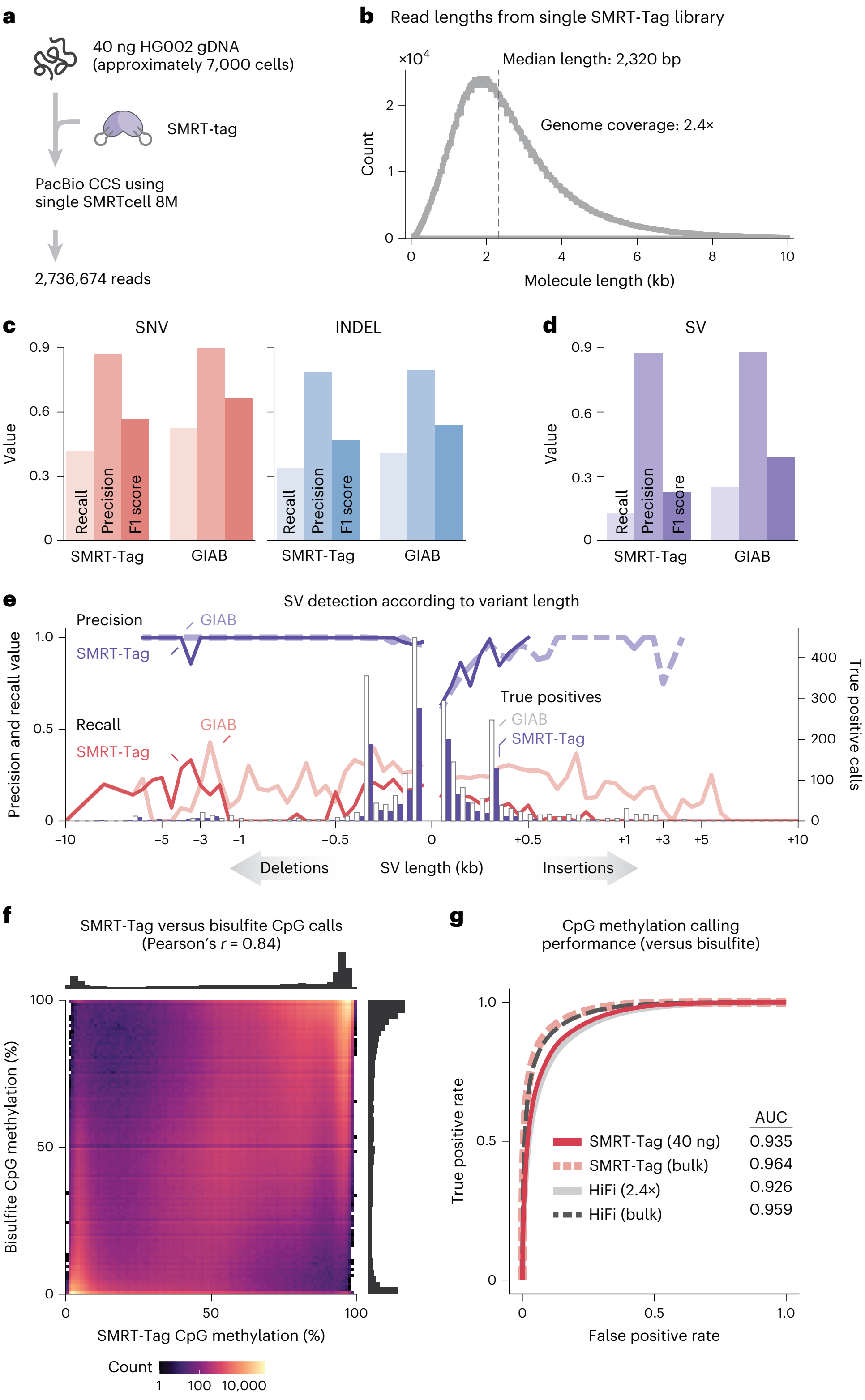

Fig. 2 | SMRT-Tag enables accurate genotyping and epigenotyping of low-input samples. a, To establish whether low-input SMRT-Tag libraries could be sequenced to sufficient depth, we tagmented 40 ng gDNA (equivalent to approximately 7,000 human cells) from GIAB reference individual HG002 and sequenced the resulting library on a single flow cell. **b**, Read length distribution of the 40-ng SMRT-Tag library. **c**,**d**, Precision, recall and F1 scores for DeepVariant SNV and indel calls (**c**) and pbsv SV calls (**d**) from 40-ng SMRT-Tag and coverage-matched ligation-based PacBio data compared with GIAB HG002 variant calling benchmarks. **e**, Precision, recall and number of true positive calls for SVs binned according to size for 40-ng SMRT-Tag and coverage-matched ligation-based data benchmarked against GIAB HG002 SV calls. **f**, Comparison of SMRT-Tag primrose and HG002 bisulfite CpG methylation. **g**, AUCs for CpG methylation detected using 40-ng SMRT-Tag, pooled SMRT-Tag (not coverage-matched) and ligation-based PacBio compared with bisulfite sequencing.

developed by us[4,11] and others[5,8], we optimized a tagmentation-assisted SAMOSA (SAMOSA-Tag; Fig. 3a). In SAMOSA-Tag, nuclei are treated in situ with the non-specific EcoGII methyltransferase, which mediates N6-deoxyadenosine methylation (m^{6}dA), and tagmented using hairpin-loaded Tn5 under conditions optimal for ATAC–seq[31]. DNA is then purified, gap-repaired and sequenced. As proof of concept, we applied SAMOSA-Tag to 50,000 nuclei from *MYC*-amplified OS152 human osteosarcoma cells[32] and used a convolutional neural network hidden Markov model[11] to call inaccessible protein–DNA interaction 'footprints' from m^{6}dA natively detected by PacBio SMS. In total, we sequenced 3,640,652 molecules (7.79 Gb) across eight replicates. Reflecting the transposition of chromatin in nuclei, SAMOSA-Tag CCS read lengths displayed characteristic oligonucleosomal banding (Fig. 3b). When aligned at the 5′ ends, molecules had periodic accessibility signal, which is consistent with transposition adjacent to nucleosomal barriers (Fig. 3c). Individual single-molecule footprint sizes also corresponded to expected mono-nucleosomes, di-nucleosomes, tri-nucleosomes, etc. (Fig. 3d). Finally, single-fiber accessibility visualized in the genomic context, for example, at the amplified *MYC* locus (Fig. 3e) and at copy number loss and neutral loci (Supplementary Figs. 7 and 8), correlated well with ATAC–seq. Importantly, there was only a mild enrichment of SAMOSA-Tag insertions for transcription start sites (TSS) (Supplementary Fig. 9a). However, insertions tended to occur proximal to predicted CCCTC-binding factor (CTCF) binding sites (Supplementary Fig. 9b), which is consistent with blocked Tn5 transposition by strong barrier elements. This subtle preference was also reflected in the fraction of insertions falling near TSS and CTCF sites (Supplementary Fig. 9c; 1.51-fold and 1.58-fold enrichment above background, respectively) and was consistent with propensities reported for Tn5-based shotgun Illumina libraries[33]. Finally, SAMOSA-Tag generalized well to mouse embryonic stem (ES) cells (Supplementary Fig. 10a–c), recovering characteristic 'footprints' around predicted CTCF and REST binding sites, which clustered into distinct accessibility patterns (Supplementary Fig. 10d,e). SAMOSA-Tag can also be performed ex situ wherein DNA is extracted from footprinted nuclei before tagmentation. The barrier effect apparent on aligning 5′-end reads is abrogated in ex situ SAMOSA-Tag (Supplementary Fig. 10b), highlighting the flexibility of the approach for applications requiring more coverage uniformity (Supplementary Note 2).

SAMOSA-Tag permits integrative single-fiber epigenomics

The separability of PacBio polymerase kinetics into m^{6}dA and m^{5}dC channels affords the opportunity to concurrently ascertain DNA sequence, CpG methylation and single-fiber chromatin accessibility to exogenous adenine methyltransferases in a single assay. We first examined m^{6}dA accessibility and CpG methylation at CTCF sites predicted from chromatin immunoprecipitation followed by sequencing (ChIP–seq) in the U2OS osteosarcoma cell line[34]. We recovered hallmarks of CTCF binding, including flanking-positioned nucleosomes, decreased accessibility immediately at the motif (compatible with exclusion of EcoGII by bound CTCF) and depressed CpG methylation within motifs (Fig. 4a). Taking advantage of the single-molecule resolution of SAMOSA-Tag, we deconvolved the differing fiber structures that contribute to the ensemble average chromatin and methylation profiles (Fig. 4a) using Leiden clustering[35] (see the example of four clusters shown in Fig. 4b; cluster sizes are shown in Supplementary Fig. 11). Analysis of pattern-specific average m^{5}dC signal (Fig. 4c) revealed the lowest CpG methylation at CTCF-bound (cluster 1) and unbound and accessible (cluster 2) motif fiber patterns, consistent with previous results[36]. Two additional analyses confirmed minimal confounding of m^{5}dC and m^{6}dA signals. First, the primrose CpG score distributions of EcoGII untreated negative control and footprinted SAMOSA-Tag libraries were concordant (Supplementary Fig. 12a). Second, the average CpG methylation surrounding predicted CTCF sites on fibers with inaccessible motifs compared with those with footprinted motifs was tightly correlated (Supplementary Fig. 12b).

We previously demonstrated that single-fiber chromatin accessibility data can be used to segment the genome by regularity and average spacing of nucleosomes (nucleosome repeat length (NRL))[4,11]. These studies relied on complementary epigenomic assays to ascertain the distribution of 'fiber types' (that is, clusters of molecules with unique regularity or NRL) in euchromatic and heterochromatic domains. We sought to improve on these analyses by directly assessing fiber structure variation with jointly resolved single-molecule CpG content and methylation. To do so, we grouped SAMOSA-Tag molecules into four bins (Fig. 4d) gated on CpG density (>10 CpG dinucleotides per kb) and primrose score (average score greater than 0.5). We then defined fiber types by clustering the m^{6}dA accessibility autocorrelation for each molecule 1 kb or longer in length[4,11]. After removing artifactual molecules, we obtained seven distinct clusters (Fig. 4e; cluster sizes are shown in Supplementary Fig. 13) effectively stratifying the OS152 genome according to NRL (clusters NRL178–NRL208) and regularity (irregularity cluster = irregular spacing). Finally, we carried out a series of enrichment tests to assess domain-specific fiber composition across the four CpG content and methylation bins (Fig. 4f; reproducibility shown in Supplementary Fig. 14). We highlight two findings relevant to chromatin regulation: first, putative hypomethylated CpG islands (high CpG content, low CpG methylation) were enriched for fibers that were irregular (odds ratio (OR) for the irregularity cluster = 1.42, $P < 2.2 \times 10^{-308}$) or have long NRLs (NRL208 OR = 1.09, $P = 4.43 \times 10^{-64}$; NRL197 OR = 1.11, $P = 1.49 \times 10^{-58}$); second, probably hypermethylated, CpG-rich repeats (high CpG content, high CpG methylation) were enriched for fibers that were irregular (irregularity cluster OR = 1.14, $P = 1.33 \times 10^{-139}$) or have short NRLs (NRL172 OR = 1.24; $P < 2.2 \times 10^{-308}$). These results are consistent with our previous in vivo observations of active promoters and heterochromatin in human cells[4] and mouse ES cells[11], pointing to a conserved single-fiber chromatin structure within these domains. Together, these analyses show that SAMOSA-Tag generated multimodal, genome-wide, single-molecule chromatin accessibility data from tens of thousands of cells.

SAMOSA-Tag of PDXs of patients with prostate cancer

One area where SAMOSA-Tag could have immediate utility is in the study of disease models such as cancer PDXs where samples are limited. There are two key challenges with PCR-free PacBio profiling of PDXs propagated in mice: first, after tumor engraftment and growth, cancer cells must be enriched and separated from mouse cells using FACS; second, cells and nuclei from metabolically active or necrotic tumors are often fragile and have damaged native DNA, which impedes sequencing. We thus sought to apply SAMOSA-Tag to generate the first single-fiber chromatin accessibility data from PDX models. We generated PDXs from matched primary and metastatic tumors resected from a patient with castration-resistant prostate cancer[37], and isolated and footprinted approximately 180,000 nuclei from one mouse each per model (Fig. 5a; the FACS gates are shown in Supplementary Fig. 15). To account for the technical difficulty of working with precious PDX samples while ensuring reproducibility, we opted conservatively to perform six replicate SAMOSA-Tag reactions (approximately 30,000 nuclei per reaction). Primary and metastatic PDX libraries were sequenced to depths of 0.32× (0.95 Gb (22.8%) human alignment) and 0.53× (1.57 Gb (95.9%) human alignment). PDX SAMOSA-Tag had similar technical characteristics to mouse ESCs and the experiments involving OS152 cells (Supplementary Fig. 16). Future optimization of cell enrichment, DNA damage repair and nuclei purification will probably permit higher per-sample coverage using lower input than in the proof of concept presented in this study.

Altered CTCF expression and occupancy have been tied to hyperactive androgen signaling[38] and prostate cancer progression[39]. To examine single-molecule chromatin accessibility and CTCF binding in primary and metastatic tumor cells (Supplementary Fig. 17a),

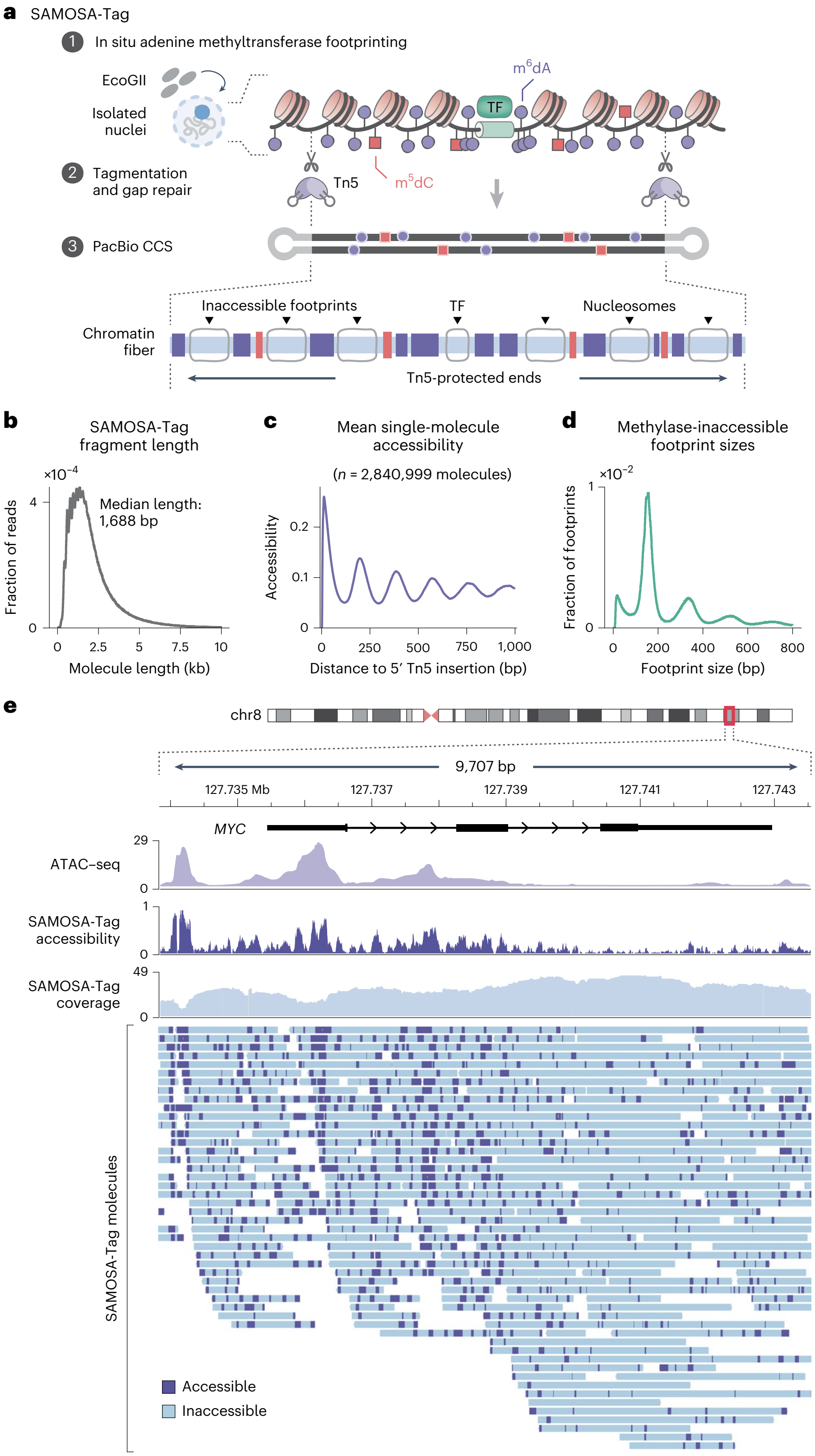

Fig. 3 | SAMOSA-Tag: single-molecule chromatin profiling via tagmentation of adenine-methylated nuclei. a, In SAMOSA-Tag, nuclei were methylated using the non-specific EcoGII m^{6}dAase and tagmented in situ with hairpin-loaded transposomes. DNA was purified, gap-repaired and sequenced, resulting in molecules where the ends resulted from Tn5 transposition, the m^{6}dA marks represented fiber accessibility and computationally defined unmethylated 'footprints' captured protein–DNA interactions. **b**, Length distribution for SAMOSA-Tag molecules from OS152 osteosarcoma cells. **c**,**d**, Average methylation from the first 1-kb of molecules (**c**) and unmethylated footprint size distribution (**d**) for the same data as in **b**. **e**, SAMOSA-Tag fibers at the amplified *MYC* locus. Predicted accessible and inaccessible bases are marked in purple and blue, respectively. Average SAMOSA accessibility is shown in purple; the matched ATAC–seq track is shown in light purple.

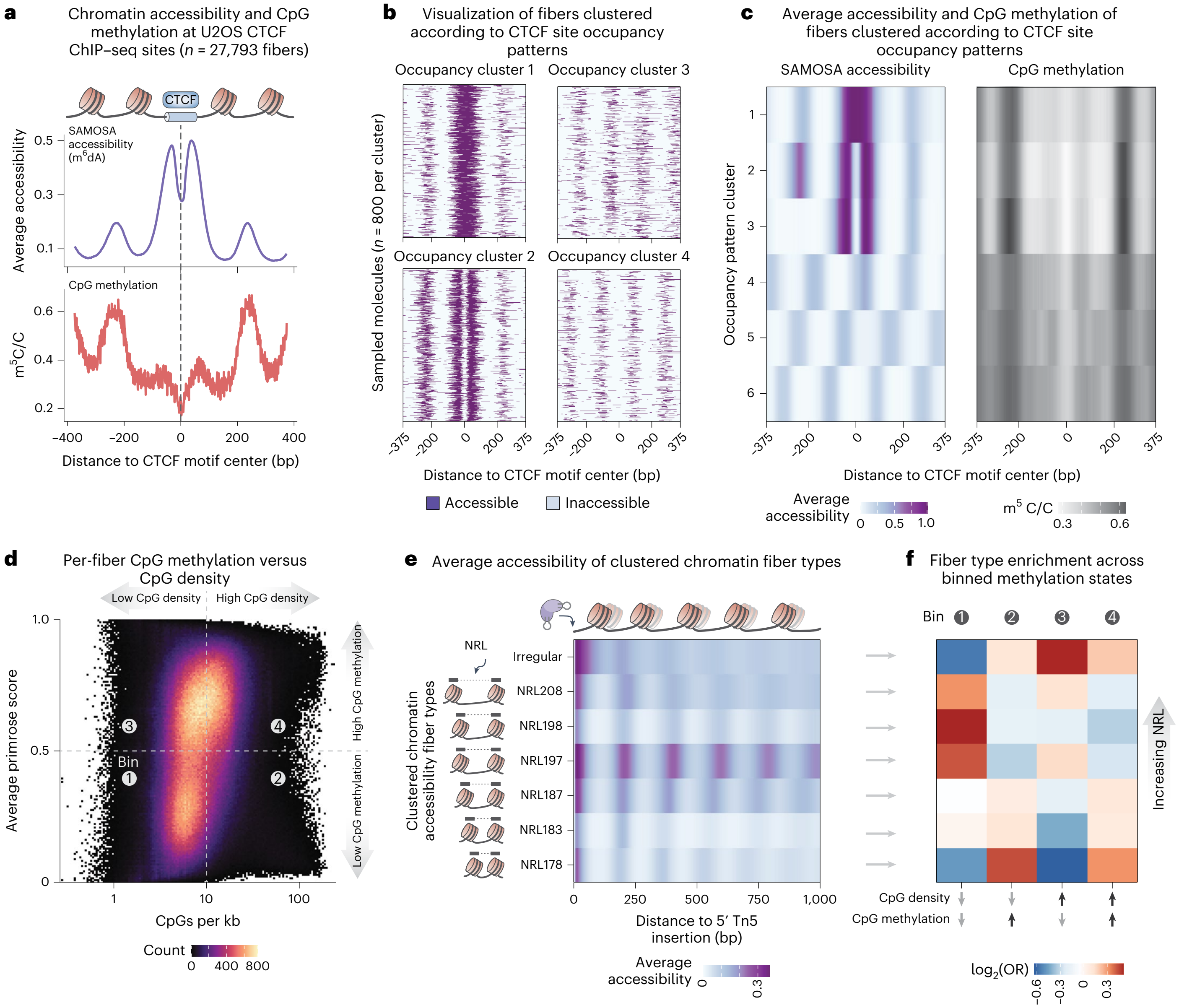

Fig. 4 | SAMOSA-Tag concurrently profiles protein–DNA interactions and CpG methylation on single chromatin fibers. a, Average SAMOSA (m^{6}dA) accessibility and CpG methylation on 27,793 footprinted fibers from OS152 human osteosarcoma cells, centered at the binding sites predicted from published U2OS ChIP–seq data[34]. **b**, Visualization of m^{6}dA signal for individual, clustered fibers centered at the predicted CTCF motifs, reflecting different CTCF-occupied, accessible and inaccessible states (800 molecules per cluster). **c**, Average accessibility (left) and CpG methylation (right) for each of six clustered accessibility states around CTCF motifs. **a**–**c**, The window size was 750 nt. **d**, Average primrose CpG methylation score for individual fibers as a function of density of CpG dinucleotides per kb. We binned molecules into four classes depending on CpG density and average primrose score. **e**, Average accessibility of seven fiber types determined using Leiden clustering of single-fiber m^{6}dA chromatin accessibility autocorrelation. Clusters stratified the entire genome according to NRL (ranging from 178 to 208 nt) or irregularity. **f**, Relative enrichment or depletion of individual fiber types for the same clusters as in **e** in each of four binned states from **d** (one-sided Fisher's exact test, *P* values ranging from $P < 2.2 \times 10^{-308}$ to $P < 2.41 \times 10^{-5}$).

we clustered PDX SAMOSA-Tag reads aligned to CTCF sites predicted using ENCODE ChIP–seq in LnCaP prostate cancer cells. This revealed multiple clusters (Supplementary Fig. 17b) reflecting varying nucleosome occupancy patterns around the CTCF motif (patterns NO1–NO5), direct CTCF occupancy (pattern A) and 'hyperaccessible' fibers devoid of nucleosomes flanking the motif (pattern HA) similar to the OS152 and mouse ESC SAMOSA-Tag (Fig. 4e and Supplementary Fig. 10a). Visualizing differential fiber type usage (Supplementary Fig. 17c) suggested intriguing metastasis-specific shifts in cluster usage, including a decrease in the stereotypic nucleosome phasing at CTCF-bound sites (pattern A) in favor of pattern HA. Analysis of concurrently measured m^{5}dC within these clusters suggested subtle preliminary differences in CpG methylation correlated with single-fiber CTCF motif occupancy patterns (Supplementary Fig. 17d).

Finally, we asked whether single-fiber chromatin architecture differed between matched primary and metastatic tumors (Supplementary Fig. 18a). Unsupervised clustering of autocorrelated single-molecule m^{6}dA signal from primary and metastatic PDXs yielded six fiber types (Fig. 5b): four regular clusters with NRLs ranging from 171 to 208 bp and two irregular clusters (IR1 and IR2). Using published annotations for healthy human prostate as a reference 40, we determined the relative enrichment of fiber types across epigenomic

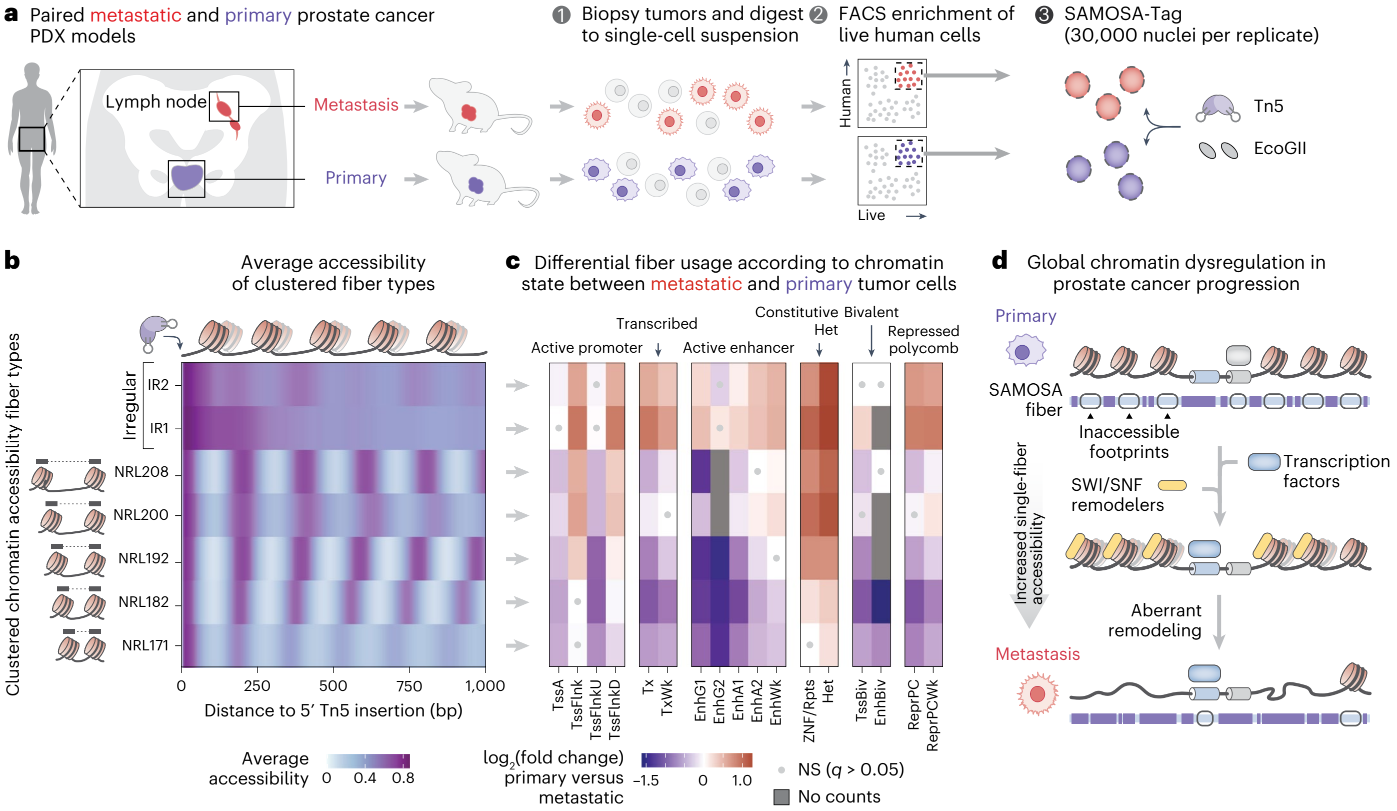

Fig. 5 | SAMOSA-Tag of PDXs nominates global chromatin dysregulation in prostate cancer metastasis. a, Overview of the approach for SAMOSA-Tag of PDXs generated from primary and metastatic castration-resistant prostate tumors sampled from a single patient. Live human cells were enriched from tumors explanted from PDX mice using FACS. Six replicate SAMOSA-Tag reactions were performed using approximately 30,000 nuclei each isolated from primary and metastatic PDXs. **b**, Clustered chromatin fiber types detected in primary and metastatic PDXs falling in one of 17 prostate-specific chromHMM states. Unsupervised Leiden clustering identified seven fiber types—five regular clusters in NRL ranging from 171 to 208 bp and two irregular clusters. **c**, Heatmap of $\log_2$ fold change in fiber type usage across chromHMM states (Δ) in metastatic versus primary PDXs. Effect sizes were the coefficients of case status (primary versus metastatic), which was the predictor variable in a logistic regression model with domain-specific fiber usage as the response variable. Statistically significant differences were identified using a Wald's test of coefficients; two-tailed P values were adjusted for multiple testing using Storey's q[49], with a significance threshold of $q \leq 0.05$. Red indicates fiber types enriched in metastasis, while blue indicates fiber types enriched in primary tumors. The gray dots mark non-significant (NS) results. Chromatin state abbreviations: EnhA1 and EnhA2, active enhancers; EnhBiv, bivalent enhancer; EnhG1 and EnhG2, genic enhancers; EnhWk, weak enhancer; ReprPC, repressed polycomb; ReprPCWk, weak repressed polycomb; TssA, active TSS; TssBiv, bivalent/poised TSS; TssFlnk, flanking TSS; TssFlnkD, downstream flanking TSS; TssFlnkU, upstream flanking TSS; Tx, strong transcription; TxWk, weak transcription; ZNF/Rpts, zinc-finger genes and repeats. **d**, Speculative model of changes in single-molecule chromatin accessibility during prostate cancer progression based on PDX SAMOSA-Tag. Highly accessible, irregular chromatin fibers devoid of phased nucleosomes were enriched in metastatic cells, which was suggestive of deranged activity of BAF remodelers, the prime candidates for generating nucleosome-free, irregular, single-molecule accessibility patterns.

domains (Supplementary Fig. 18b). Applying a logistic regression framework to nominate significant differences in domain-specific fiber usage, we identified several patterns of interest to be followed up in future studies (Fig. 5c). For instance, metastatic tumor cells were significantly enriched for irregular fibers (IR1 and IR2) in heterochromatic domains such as KRAB zinc-finger genes (ZNF/Rpts; IR1 $\log_2$ fold change (Δ) = 0.77, $q = 7.56 \times 10^{-7}$; IR2 Δ = 1.03, $q = 6.15 \times 10^{-15}$) and regions harboring marks of constitutive heterochromatin (Het) IR1 Δ = 1.22, $q = 1.45 \times 10^{-177}$; IR2 Δ = 1.25; $q = 4.46 \times 10^{-125}$). In contrast, distal enhancers were significantly depleted for fibers with specific NRLs (for example, active enhancer 1 (EnhA1); NRL182 Δ = −1.11, $q = 1.07 \times 10^{-71}$). These data hint at the involvement of ATP-dependent chromatin remodelers, such as the Brahma-associated factor (BAF) complex in metastasis-associated nucleosome eviction and chromatin disorganization (Fig. 5d). While BAF has already been implicated as a driver of prostate cancer progression[41], mechanistic studies are needed to evaluate the proposed model. Taken together, these data demonstrate the potential of SAMOSA-Tag to yield biological insights in challenging disease models.

Discussion

We optimized direct Tn5 transposition of hairpin adapters as a general strategy for preparing amplification-free, multiplexable PacBio libraries from limiting amounts of input DNA. We applied this principle to develop two methods that take advantage of the simultaneous readout of modified and unmodified bases using SMS and highlight the broad potential of Tn5-based PacBio library generation. First, tagmentation coupled with PacBio HiFi sequencing (SMRT-Tag) allowed detection of genetic variation and CpG methylation from as little as 40 ng gDNA (approximately 7,000 human cells) with accuracy comparable to conventional whole-genome and bisulfite sequencing. Second, tagmentation of 30,000–50,000 nuclei after adenine methyltransferase chromatin footprinting (SAMOSA-Tag) concurrently resolved single-fiber DNA sequence, CpG methylation and chromatin accessibility in one assay. Using SAMOSA-Tag libraries multiplexed to maximize sequencing yield, we resolved CTCF binding, nucleosome architecture and CpG methylation in osteosarcoma cells. We also carried out single-molecule epigenome analyses in a preclinical disease model, uncovering global chromatin dysregulation associated with metastatic progression in technically challenging prostate cancer PDX cells.

We anticipate that tagmentation-based protocols will address several obstacles to single-molecule genomics. Simplification of library preparation by combining DNA fragmentation and adapter ligation steps and the high efficiency of Tn5 transposition permitted 90–99% input reduction for SMRT-Tag and SAMOSA-Tag, placing sequencing at the lower limit of the PacBio platform within reach (Supplementary Notes 1 and 2). The ability to profile unamplified DNA has implications for basic and translational analyses of rare cell populations that integrate the breadth of nucleotide, structural and epigenomic variation captured natively by SMS without chemical conversion. Importantly, in situ tagmentation also obviates the need for DNA purification, raising the exciting prospect of multimodal genomics with both single-cell and single-molecule resolution. We envision that developments such as droplet-based or combinatorial barcoding-based cellular indexing[21,23,42] will extend massively parallel PCR-free, single-molecule assays to individual cells for applications ranging from strand-specific somatic variant detection[43] to haplotype-resolved de novo assembly, and cell type classification.

As with any technical advance, while SMRT-Tag and SAMOSA-Tag illustrate the power of Tn5 transposition for PacBio SMS, they have several limitations and areas for improvement. Because these methods do not rely on PCR, libraries may need to be multiplexed to maximize OPLC and reduce per-base cost. Still, we showed that flow cells can be efficiently loaded with as little as 40 ng starting input mass. The length of molecules is lower than the 15–20-kb capability of PacBio SMS and is primarily controlled by transposome concentration and optional bead-based size selection; the limited input amount precludes gel-based-size fractionation. Furthermore, the inverse proportionality between length and molarity for a given input mass implies that more starting material or pooling at higher plexity are needed to yield deep coverage (Supplementary Note 2). This is salient for comprehensively surveying variants and particularly for SV discovery because breakpoint-spanning molecules are less abundant in SMRT-Tag than ligation-based libraries. Limited coverage may also impede resolution of epigenomes in tissues and other heterogeneous samples. Although we have partially addressed this by demonstrating the tunability of tagmentation, adapting engineered[25] and bead-linked[44] transposases may offer finer control of molecule length in the future. SAMOSA-Tag is also limited by the minimum number of nuclei that can be processed. In this study, we generated high-quality data from replicates of 30,000–50,000 nuclei. Optimizations including mild fixation, miniaturized methylation reactions or immobilization of nuclei on beads[45] could further relax this constraint.

More generally, SMRT-Tag and SAMOSA-Tag add to a growing series of innovations centered around third-generation sequencing, including Cas9-targeted sequence capture[46], combinatorial indexing-based plasmid reconstruction[47] and concatenation-based isoform-resolved transcriptomics[48]. The widespread adoption of short-read genomics in basic and clinical applications, and the transition from bulk to single-cell assays was catalyzed by approaches that simplified library preparation and reduced input requirement. Direct transposition offers similar promise for rapidly maturing third-generation sequencing technologies in enabling scalable, sensitive and high-fidelity telomere–telomere genomics and epigenomics.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-024-01748-0.

References

1. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).
2. Aganezov, S. et al. A complete reference genome improves analysis of human genetic variation. *Science* **376**, eabl3533 (2022).
3. Vollger, M. R. et al. Segmental duplications and their variation in a complete human genome. *Science* **376**, eabj6965 (2022).
4. Abdulhay, N. J. et al. Massively multiplex single-molecule oligonucleosome footprinting. *eLife* **9**, e59404 (2020).
5. 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).
6. Shipony, Z. et al. Long-range single-molecule mapping of chromatin accessibility in eukaryotes. *Nat. Methods* **17**, 319–327 (2020).
7. Lee, I. et al. Simultaneous profiling of chromatin accessibility and methylation on human cell lines with nanopore sequencing. *Nat. Methods* **17**, 1191–1199 (2020).
8. 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).
9. Au, K. F. et al. Characterization of the human ESC transcriptome by hybrid sequencing. *Proc. Natl Acad. Sci. USA* **110**, E4821–E4830 (2013).
10. Sharon, D., Tilgner, H., Grubert, F. & Snyder, M. A single-molecule long-read survey of the human transcriptome. *Nat. Biotechnol.* **31**, 1009–1014 (2013).
11. Abdulhay, N. J. et al. Nucleosome density shapes kilobase-scale regulation by a mammalian chromatin remodeler. *Nat. Struct. Mol. Biol.* **30**, 1571–1581 (2023).
12. Wang, Y. et al. Single-molecule long-read sequencing reveals the chromatin basis of gene expression. *Genome Res.* **29**, 1329–1342 (2019).
13. Quail, M. A. et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**, 341 (2012).
14. Adey, A. et al. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol.* **11**, R119 (2010).
15. Adey, A. & Shendure, J. Ultra-low-input, tagmentation-based whole-genome bisulfite sequencing. *Genome Res.* **22**, 1139–1143 (2012).
16. 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).
17. 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).
18. Chen, C. et al. Single-cell whole-genome analyses by linear amplification via transposon insertion (LIANTI). *Science* **356**, 189–194 (2017).
19. Minussi, D. C. et al. Breast tumours maintain a reservoir of subclonal diversity during expansion. *Nature* **592**, 302–308 (2021).
20. Payne, A. C. et al. In situ genome sequencing resolves DNA sequence and structure in intact biological samples. *Science* **371**, eaay3446 (2021).
21. Cusanovich, D. A. et al. Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).

22. Cao, J. et al. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018).
23. Yin, Y. et al. High-throughput single-cell sequencing with linear amplification. *Mol. Cell* **76**, 676–690 (2019).
24. Eid, J. et al. Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
25. Hennig, B. P. et al. Large-scale low-cost NGS library preparation using a robust Tn5 purification and tagmentation protocol. *G3* **8**, 79–89 (2018).
26. Reznikoff, W. S. Tn5 as a model for understanding DNA transposition. *Mol. Microbiol.* **47**, 1199–1206 (2003).
27. Zook, J. M. et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* **3**, 160025 (2016).
28. Krusche, P. et al. Best practices for benchmarking germline small-variant calls in human genomes. *Nat. Biotechnol.* **37**, 555–560 (2019).
29. Flusberg, B. A. et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat. Methods* **7**, 461–465 (2010).
30. Simpson, J. T. et al. Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017).
31. Grandi, F. C., Modi, H., Kampman, L. & Corces, M. R. Chromatin accessibility profiling by ATAC-seq. *Nat. Protoc.* **17**, 1518–1552 (2022).
32. Sayles, L. C. et al. Genome-informed targeted therapy for osteosarcoma. *Cancer Discov.* **9**, 46–63 (2019).
33. Vitak, S. A. et al. Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat. Methods* **14**, 302–308 (2017).
34. Ibarra, A., Benner, C., Tyagi, S., Cool, J. & Hetzer, M. W. Nucleoporin-mediated regulation of cell identity genes. *Genes Dev.* **30**, 2253–2258 (2016).
35. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
36. Wang, H. et al. Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res.* **22**, 1680–1688 (2012).
37. Nguyen, H. G. et al. Development of a stress response therapy targeting aggressive prostate cancer. *Sci. Transl. Med.* **10**, eaar2036 (2018).
38. Alpsoy, A. et al. BRD9 is a critical regulator of androgen receptor signaling and prostate cancer progression. *Cancer Res.* **81**, 820–833 (2021).
39. 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).
40. Wang, T. et al. Integrative epigenome map of the normal human prostate provides insights into prostate cancer predisposition. *Front. Cell Dev. Biol.* **9**, 723676 (2021).
41. Xiao, L. et al. Targeting SWI/SNF ATPases in enhancer-addicted prostate cancer. *Nature* **601**, 434–439 (2022).
42. Ramani, V. et al. Massively multiplex single-cell Hi-C. *Nat. Methods* **14**, 263–266 (2017).
43. Liu, M. H. et al. Single-strand mismatch and damage patterns revealed by single-molecule DNA sequencing. Preprint at *bioRxiv* https://doi.org/10.1101/2023.02.19.526140 (2023).
44. Bruinsma, S. et al. Bead-linked transposomes enable a normalization-free workflow for NGS library preparation. *BMC Genomics* **19**, 722 (2018).
45. Meers, M. P., Bryson, T. D., Henikoff, J. G. & Henikoff, S. Improved CUT&RUN chromatin profiling tools. *eLife* **8**, e46314 (2019).
46. Gilpatrick, T. et al. Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nat. Biotechnol.* **38**, 433–438 (2020).
47. Emiliani, F. E., Hsu, I. & McKenna, A. Multiplexed assembly and annotation of synthetic biology constructs using long-read Nanopore sequencing. *ACS Synth. Biol.* **11**, 2238–2246 (2022).
48. Al'Khafaji, A. M. et al. High-throughput RNA isoform sequencing using programmed cDNA concatenation. *Nat. Biotechnol.* **42**, 582–586 (2023).
49. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100**, 9440–9445 (2003).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Human patient and animal studies

De-identified primary tumor and metastatic lymph node tissue used to generate the PDX models was donated by a patient who provided written informed consent under protocol no. 90911 'Use of marker in cytometric analysis in prostate cancer to predict biological potential' (University of California, San Francisco (UCSF) institutional review board 11-05226). NOD *scid* gamma (NSG) mice (UCSF Breeding Core) were maintained under pathogen-free conditions. This study was performed with assistance from the UCSF Laboratory Animal Resource Center under a protocol approved by the UCSF Institutional Animal Care and Use Committee (no. AN195508).

Cell lines

OS152 cells were regularly tested for authenticity and *Mycoplasma* via CellCheck 9 Plus (IDEXX BioAnalytics) and were cultured in DMEM (catalog no. 10-013-CV, Corning) with 10% Bovine Growth Serum (catalog no. SH30541.03, HyClone) and 1% 100× penicillin-streptomycin-glutamine (catalog no. 30-009-CI, Corning). Embryonic day 14 (E14) mouse ES cells were a gift from E. Nora (UCSF) and were routinely tested for *Mycoplasma* via PCR (NEBNext Ultra II Q5 Master Mix, catalog no. M0544S, New England Biolabs). Feeder-free cultures were passaged at least twice before use and maintained on 0.2% gelatin and in KnockOut DMEM (catalog no. 10829018, Thermo Fisher Scientific) with 10% FCS (catalog no. BW-067C18, Phoenix Scientific), 1% GlutaMAX (catalog no. 35050061, Thermo Fisher Scientific), 1% MEM Non-Essential Amino Acids Solution (catalog no. 1114050, Thermo Fisher Scientific), 0.128 mM 2-mercaptoethanol (catalog no. 1610710XTU, Bio-Rad Laboratories) and purified 1× leukemia inhibitory factor (gifted by B. Panning, UCSF).

Assembly of hairpin adapter-loaded Tn5 transposomes

Annealing adapters. Uniquely barcoded (Hamming distance ≥4), HPLC-purified hairpin oligonucleotides (Supplementary Table 5) were purchased from Integrated DNA Technologies and normalized to 100 µM in nuclease-free water. Adapters were diluted to 20 µM in annealing buffer (10 mM Tris-HCl, pH 7.5, and 100 mM NaCl), annealed (95 °C for 5 min, 25 °C for 30 min, held at 4 °C) and stored at −20 °C.

Loading Tn5 with SMRT-Tag adapters. Triple-mutant $\mathrm{Tn5}^{\mathrm{R27S,E54K,L372P}}$ enzyme (Tn5) was purified by the QB3 MacroLab (University of California, Berkeley). Aliquots of Tn5 (3.9 mg ml^{-1}) in storage buffer (50 mM Tris-HCl, pH 7.5, 800 mM NaCl, 0.2 mM EDTA, 2 mM dithiothreitol, 10% glycerol) and frozen at −20 °C were thawed at 4 °C, diluted in dilution buffer (50 mM Tris-HCl, pH 7.5, 200 mM NaCl, 0.1 mM EDTA, 2 mM dithiothreitol, 50% glycerol) to approximately 1 mg ml^{-1} Tn5 (18.9-µM monomer) by mixing at 4 °C for 3.5 h until homogenized. Hairpin adapters were loaded by mixing 1.02× volumes of 1 mg ml^{-1} Tn5 with 1× volume of 20 µM annealed adapters, followed by incubation at 23 °C with agitation at 350 r.p.m. for 55 min. Loaded Tn5 stock (9.4-µM monomer) was stored at −20 °C for up to 6 months in glycerol to a final concentration of 50%. Then, 1–2 µl of loaded Tn5 diluted in NativePAGE Loading Buffer was run on a NativePAGE 4–16% Bis-Tris gel (catalog nos. BN2003 and BN1002, Thermo Fisher Scientific) at 150 V for 1 h at 4 °C followed by 180 V for 15 min. Gels were stained with SYBR Gold (catalog no. S11494, Thermo Fisher Scientific) in Tris base, acetic acid and EDTA buffer and then SimplyBlue (catalog no. LC6060, Thermo Fisher Scientific) for 1 h and imaged on an Odyssey XF Imaging System (software v.1.1.0.61, LI-COR).

Assessing the tunability of fragment lengths. Serial dilutions of transposome stock were incubated with 160 ng human gDNA (Promega Corporation) while varying buffers, temperatures and incubation times. Analytical electrophoresis was performed on a 0.4–0.6% Tris base, acetic acid and EDTA agarose gel run at 60–80 °C for 2–3 h. Gels were stained with SYBR Gold and imaged on an Odyssey XF Imaging System.

SMRT-Tag of gDNA

SMRT-Tag library preparation. HG002, HG003 and HG004 gDNAs (Coriell Institute for Medical Research) normalized to 40–160 ng in 9 µl Tagmentation Mix (10 mM [tris(hydroxymethyl)methylamino]propanesulfonic acid-NaOH, pH 8.5, 5 mM $\mathrm{MgCl_2}$, 10% dimethylformamide) were tagmented with 1 µl of barcoded Tn5 at varying concentrations (Supplementary Table 3) at 55 °C for 30 min. Reactions were terminated by adding 0.2% SDS (final concentration 0.04%) before incubation at 25 °C for 5 min, 2× SPRI cleanup and elution in 12 µl elution buffer (EB) (10 mM Tris-HCl, pH 8.5). DNA was gap-repaired at 37 °C for 1 h in repair mix (2 U Phusion-HF, 80 U Taq DNA Ligase with 1× Taq DNA ligase buffer and 0.8 mM deoxynucleotide triphosphate; catalog nos. M0530S, M0208S and N0447S, New England Biolabs), purified with 2× SPRI beads, eluted in 12 µl EB and digested at 37 °C for 1 h in ExoDigest Mix (100 U exonuclease III per 160 ng DNA, 1× NEBuffer 2, catalog nos. M0206S and B7002S, New England Biolabs). Libraries were eluted in 12 µl EB after 2× SPRI cleanup.

Titration of transposome and input amounts at varying temperatures. To characterize the tunability of tagmentation, reactions were carried out as above using hairpin-loaded Tn5 stock (9.4-µM monomer) diluted in nuclease-free water to 0.05-pmol, 0.50-pmol and 5-pmol monomers with 40, 200 and 1,000 ng HG003 gDNA and incubated at 37 °C or 55 °C for 30 min.

Assaying barcode hopping via pooled gap repair. Libraries were prepared as above using barcoded transposomes, but were pooled after tagmentation into a single repair reaction before exonuclease digestion.

Optional size selection of libraries. Size selection was performed to enrich for molecules larger than 5 kb (high-molecular weight molecules) by binding libraries to 3.1× volumes of 35% (v/v) AMPure PB beads (catalog no. 100-265-900, PacBio) diluted in EB at 25 °C for 15 min. Beads were washed twice with 80% ethanol for 1 min and DNA was eluted in 15 µl EB. For some libraries, 0.25× AMPure PB cleanup of the supernatant was used to recover low-molecular weight (<5-kb) DNA, which was eluted in 15 µl EB.

SAMOSA-Tag of cell lines

Nuclei isolation. A total of $1–2 \times 10^6$ OS152 or mouse E14 ES cells were collected by centrifugation (300*g*, 4 °C, 10 min), washed in cold PBS and resuspended using a wide-bore micropipette tip in 1 ml cold nuclei lysis buffer (20 mM HEPES, 10 mM KCl, 1 mM $\mathrm{MgCl_2}$, 0.1% Triton X-100, 20% glycerol, 1× protease inhibitor (catalog no. 04693132001, Roche)). Cells were lysed on ice for 5 min. Nuclei were pelleted (600*g*, 4 °C, 10 min), washed with Buffer M (15 mM Tris-HCl, pH 8.0, 15 mM NaCl, 60 mM KCl, 0.5 mM spermidine) and counted on a Countess 3 instrument (Thermo Fisher Scientific).

SAMOSA footprinting. Permeabilized nuclei were pelleted (600*g*, 4 °C, 10 min) and resuspended in 400 µl Buffer M with 1 mM *S*-adenosyl methionine (catalog no. B9003S, New England Biolabs); 200 µl was reserved as an unmethylated control. Nuclei were treated with 250 U EcoGII (25,000 U ml^{-1}; New England Biolabs) for 30 min at 37 °C with agitation at 300 r.p.m. every 2 min. *S*-adenosyl methionine was replenished to 1.16 mM after 15 min in the methylation and control reactions. Nuclei were pelleted (600*g*, 10 min, 25 °C) and resuspended in 250 µl Omni-ATAC Buffer (10 mM Tris-HCl, pH 7.5, 5 mM $\mathrm{MgCl_2}$, 0.33× PBS, 10% dimethylformamide, 0.01% digitonin, 0.1% Tween 20). Nuclei were filtered through a Flowmi 40-µm strainer (catalog no. BAH136800040, Sigma-Aldrich) and counted and visualized on a Countess 3 to verify dissociation of aggregates.

In situ SAMOSA-Tag. Methylated and unmethylated samples were split into 10,000–50,000 nuclei aliquots and, based on the desired library size and cell type, 9.4–18.8 pmol barcoded transposome was added. Samples were brought up to 50 µl with Omni-ATAC Buffer before tagmentation at 55 °C for 45–60 min. Tagmentation was terminated by treating nuclei with 10 µg RNase A (catalog no. EN0531, Thermo Fisher Scientific) at 37 °C for 15 min with 300 r.p.m. shaking, mixing with 50 µg proteinase K (catalog no. AM2546, Thermo Fisher Scientific), 2.5 µl 10% SDS and 2.5 µl of 0.5 M EDTA, and incubation at 60 °C with 1,000 r.p.m. shaking for 1–2 h. Tagmented DNA was bound to 2× SPRI beads at 23 °C for 30 min with mixing at 350 r.p.m. every 3 min. Beads were collected on a magnet and washed twice in 80% ethanol for 1 min before elution in 20 µl EB at 37 °C for 15 min with mixing at 350 r.p.m. every 3 min. An additional 0.6× SPRI cleanup was used to enrich for fragments larger than 500 bp. DNA was stored at 4 °C overnight, or up to 2 weeks at −20 °C. Libraries were prepared by incubating DNA in Repair Mix at 37 °C for 1 h, followed by 2× SPRI cleanup, treatment with ExoDigest Mix at 37 °C for 1 h, 2× SPRI cleanup and elution in 12 µl EB. For OS152 and mouse E14 ES SAMOSA-Tag cells, eight methylated and unmethylated control replicates were tagmented with barcoded Tn5.

Ex situ SAMOSA-Tag. Permeabilized mouse E14 ES cell nuclei were footprinted as above. After EcoGII methylation, nuclei were digested with 10 µg RNase A at 37 °C for 15 min, mixed with 2.65 µl 10% SDS, approximately 50 µg proteinase K and incubated at 65 °C for 3 h. DNA was extracted in 1× volume of phenol:chloroform:isoamyl alcohol (25:24:1, v/v). Samples were centrifuged (16,000*g*, 2 min, 25 °C) and the aqueous phase was mixed with a 0.1× volume of 3 M NaOAc, 1 µl GlycoBlue Coprecipitant (catalog no. AM9515, Thermo Fisher Scientific) and 3× volumes of cold 100% ethanol. DNA precipitated overnight at −80°C was then pelleted (16,000*g*, 30 min, 4 °C), washed with 500 µl 70% ethanol, air-dried and resuspended in 40 µl EB. Then, 100 ng SAMOSA DNA was tagmented with a 0.046-pmol Tn5 monomer with library preparation as described for SMRT-Tag.

SAMOSA-Tag of PDXs

Generation of prostate cancer PDXs. PDXs were generated[37] using 3–5-mm primary tumor and synchronous lymph node metastasis fragments isolated from a 71-year-old male who presented with high-risk prostate cancer (pretreatment prostate-specific antigen = 19.1 ng ml^{-1}, Gleason 4+5, T3aN1M0) with 6–9-mm bilateral external pelvic lymph node metastases on prostate-specific membrane antigen positron emission tomography scan. To minimize cell death and preserve microenvironment integrity, tumor fragments were taken immediately after prostatic devascularization during robotic prostatectomy and pelvic lymph node dissection, placed in 10 ml of Roswell Park Memorial Institute 1640 and implanted subcutaneously into the flanks of 6–8-week-old male NSG mice to establish the PDX lines. PDXs were cryopreserved after three passages in mice. To ensure that PDXs faithfully recapitulated the original tumor and heterogeneity of prostate cancer, sections were subjected to histopathological comparison after each passage and growth patterns were examined. Passage 10 PDXs were used for SAMOSA-Tag.

PDX processing and SAMOSA-Tag. Surgically explanted tumors from PDX mice were immediately placed into sterile Roswell Park Memorial Institute 1640 on ice and minced with a scalpel. PDXs were dissociated into single cells by digestion with 10 µg DNase I and 65 mg Collagenase Type 3 (DNase I, RNase & Protease Free, catalog no. LS004206, Worthington Biochemical) and 10 mg Liberase TL (catalog no. 05401020001, Sigma-Aldrich) at 37 °C for 1 h with agitation at 750 r.p.m. in Digestion Buffer: 5 ml DMEM, 5 ml Ham's F-12, 100 µl 100× penicillin-streptomycin and 40 µl 0.25 mg ml^{-1} Amphotericin B (catalog nos. 11965092, 11765054 and 15290018, respectively, Thermo Fisher Scientific). Cells were centrifuged (800*g*, 5 min, 4 °C), resuspended in 1 ml cold PBS, strained through a Falcon 70-µm strainer (catalog no. 352350, Corning) with a wide-bore micropipette tip, washed twice in 1 ml cold PBS and resuspended in 1 ml Cell Staining Buffer (catalog no. 420201, BioLegend) to approximately 8–12.5 × 10^6 cells per ml. Cells were blocked with 20 µl Human TruStain FcX Receptor Blocking Solution (catalog no. 422301, BioLegend) for 10 min at 4 °C, stained with PE anti-mouse H-2 antibody (1 µg antibody per 8–12.5 × 10^6 cells, catalog no. 125505, BioLegend) for 25 min at 4 °C in the dark, washed twice in Cell Staining Buffer and pelleted at 350*g* and 4 °C. Cells were then stained with 1 µl SYTOX Red Dead Cell Stain (catalog no. S34859, Thermo Fisher Scientific) for 15 min at 4 °C. Mouse and dead human cells were depleted using a FACSAria II (FACSDiva software v.9.0.1, BD Biosciences) at the UCSF Center for Advanced Technology. Live human cells were selected as PE$^-$ and APC$^-$ (SYTOX Red) from singlets gated on forward scatter and collected into a conical tube containing 1 ml PBS. The yield per PDX was 1.20–1.75 × 10^6 cells. In situ SAMOSA-Tag was performed using sorted cells with spin speed reduced to 400*g*. Limited unmethylated control replicates for primary ($n = 2$) and metastatic ($n = 1$) PDXs were performed due to sample losses.

Ligation-based library preparation

Preparation of low-input libraries. Replicate libraries were prepared from 40 ng sheared HG002 gDNA using the SMRTbell Express Template Prep Kit 2.0 (TPK2.0, catalog no. 101-696-100, PacBio) protocol, which involves overhang removal, DNA damage repair, end repair, A-tailing, adapter ligation, exonuclease digestion and 1× AMPure PB cleanup. Insufficient mass was obtained for sequencing.

Preparation of high-input libraries. Phenol:chloroform:isoamyl alcohol-extracted mouse E14 ES cell gDNA was fragmented to 6–8 kb using a g-TUBE (catalog no. 520079, Covaris) with an Eppendorf 5424 rotor spun at 7,000 r.p.m. for six passes. A TPK2.0 library was prepared from 2.5 µg sheared DNA and loaded at 44.6 pM to confirm sequencing of ligation-based libraries at low OPLC.

DNA quality control and PacBio sequencing

To assess repair efficiency, 1 µl of library before and after exonuclease digestion was quantified using a Qubit 1× High Sensitivity DNA Assay (catalog no. Q33230, Thermo Fisher Scientific). To validate library concentration and size, 1 µl of library was analyzed using Qubit 1× High Sensitivity DNA and Agilent 2100 Bioanalyzer High Sensitivity DNA assays. Sequencing was performed using 8 M SMRTcells (catalog no. 101-389-001, PacBio) on a PacBio Sequel II running SMRTlink v.11.0.0.146107; movies were collected for 30 h with 2-h pre-extension and 4-h immobilization times. Both 2.1 and 2.2 polymerases were used for SMRT-Tag and OS152, and mouse E14 ES cell SAMOSA-Tag, depending on library size (for example, low-molecular weight and high-molecular weight SMRT-Tag libraries were sequenced with 2.1 and 2.2 polymerases, respectively; Supplementary Note 2). PDX SAMOSA-Tag libraries were multiplexed and sequenced using 2.1 polymerase.

Data analyses

Reaction efficiency. Stepwise tagmentation, gap repair and exonuclease efficiencies were defined as the mass ratio of output to input DNA for a given step. The term 'repair efficiency' refers to exonuclease cleanup efficiency, as a proxy for effectiveness of gap repair and conversion of DNA into sequenceable molecules. Overall efficiency was estimated as the mass ratio of final library to input DNA, or as the product of stepwise efficiencies.

Data preprocessing. CCS/HiFi reads were generated from subreads using ccs v.6.4.0 (PacBio) with the flag --hifi-kinetics. Lima v.2.6.0 (PacBio) with FLAG --ccs was used for demultiplexing and pbmerge v.1.0.0 (PacBio) was used to combine data from libraries sequenced on multiple flow cells. Reads were aligned using pbmm2 v.1.9.0 (PacBio) to the following references: hs37d5 GRCh37 for SMRT-Tag variant analyses; hg38 for OS152 SAMOSA-Tag and all other SMRT-Tag analyses; GRCm38 for mouse E14 ES SAMOSA-Tag cells; and a joint hg38/GRCm39 reference for PDX SAMOSA-Tag with reads uniquely aligned to hg38 retained for subsequent analyses. Read quality was ascertained from ccs and the empiric quality score (Q-score) was calculated as $-\log_{10}(1-(n_{\mathrm{matches}}/(n_{\mathrm{matches}}+n_{\mathrm{mismatches}}+n_{\mathrm{del}}+n_{\mathrm{ins}})))$ or the maximal theoretical quality score if the read contained no variation.

Analysis of SMRT-Tag demultiplexing. Given the low coverage, SNVs were called naively using SAMtools mpileup (v.1.15.1) in GIAB benchmark intervals supported by two or more reads. For each individual, SNV calls were intersected with private SNVs in regions labeled 'not difficult' in the GIAB v.3.0 stratification[25] (Supplementary Methods).

HG002 variant calling and benchmarking. HG002 SMRT-Tag and GIAB PacBio data were subsampled to threefold, fivefold, tenfold and 15-fold depths (Supplementary Methods). SNVs and indels called using DeepVariant (v.1.4.0) were compared against the GIAB/NIST v.4.2.1 benchmarks[27] using hap.py (v.0.3.12) and GIAB GRCh37 stratifications (v.3.0). SVs called using pbsv v.2.8.0 (PacBio) were compared against NIST Tier 1 HG002 SV calls (v.0.6) using Truvari[50] (v.3.3.0).

Predicting CpG methylation in PacBio reads. PacBio primrose v.1.3.0 (now Jasmine) was used to predict CpG methylation. Methylation probabilities encoded in the BAM tags ML and MM were parsed to continuous values for single-molecule methylation prediction. Per-CpG methylation was estimated using the tools available at github.com/PacificBiosciences/pb-CpG-tools.

SAMOSA footprinting. A series of neural networks trained on per-read polymerization kinetics measured during PacBio sequencing of methylated and unmethylated controls were used to predict strand-specific m^6dA methylation probabilities on SAMOSA-Tag CCS reads[4,11]. Probabilities were binarized into accessibility calls using a two-state hidden Markov model and encoded in the MM and ML BAM tags[4,11].

Comparing ATAC–seq and SAMOSA-Tag. SAMOSA accessibility and the normalized ATAC–seq signal were aggregated at OS152 ATAC–seq peaks. Pearson's *r* was used to correlate log-transformed values.

U2OS and LNCaP CTCF ChIP–seq processing. CTCF binding sites were determined[4] from published ChIP–seq peaks from U2OS[34] and LNCaP metastatic prostate adenocarcinoma[51] cells lifted over from hg19 to hg38.

Insertion preference analyses at TSS and CTCF sites. SAMOSA-Tag read ends were tabulated in 5-kb windows around GENCODE v.28 (hg38) or M25 (GRCm38) TSS or ChIP–seq-backed CTCF motifs. Meta-plots were smoothed with a 100-nt running mean. FRITSS and FRICBS were calculated as the fraction of ends falling within the 5-kb window.

CTCF CpG and accessibility analyses. The single-fiber m^6dA accessibility signal around predicted CTCF sites was subjected to Leiden clustering[35]. Clusters comprising less than 10% of data were removed. Unmethylated SAMOSA-Tag fibers were also removed ($n = 3{,}627$ or 11.5% of all CTCF-motif-containing fibers in OS152; and $n = 245$ or 1.5% in PDX).

Classifying fibers according to CpG content and CpG methylation. Fibers were binned according to CpG content (CpGs per kb ≤10 (low) versus >10 (high)) and methylation (average primrose score ≤0.5 (low) versus >0.5 (high)) to define four classes: high CpG content and high methylation; low CpG content and low methylation; high CpG content and low methylation; and low CpG content and high methylation.

Fiber type clustering and enrichment. Unsupervised Leiden clustering[35] of single-molecule accessibility autocorrelation[4,11] was used to identify fiber types. Clusters comprising less than 10% of all fibers were filtered out; unmethylated or lowly methylated molecules were also removed in the OS152 SAMOSA-Tag analyses ($n = 317{,}768$ or 12.5% of fibers). For fiber type A stratified according to feature B, a contingency table of fiber counts in four groups (A ∩ B, A ∩ B, A ' ∩ B and A ' ∩ B ' where the complement of set A is denoted by A') was used as input for a one-sided Fisher's exact test. *P* values were corrected for multiple testing using Storey's *q*[49].

Differential fiber use. chromHMM domains in normal prostate[40] were lifted over from hg19 to hg38. Fiber type-stratified coverage of prostate-specific epigenomic domains was tabulated by aggregating the counts of fiber type A mapping to domain B (A ∩ B) versus the other domains (A ∩ B') per replicate. Counts were normalized across replicates using a median-of-medians approach to account for depth and used as weights for a logistic regression model with domain and fiber status, and case status (primary versus metastasis), as the response and predictor variables, respectively. The glm function (R v.4.2.1) was used to fit the model; the null model was fitted with coefficients set to zero. The case status coefficient was taken to estimate the $\log_2$ fold change (Δ) in metastatic versus primary PDX. This was repeated for all observed combinations of the seven fiber types and 17 domains. A Wald's test was used to evaluate the maximum likelihood-fitted coefficients; two-sided *P* values were adjusted for multiple testing using Storey's *q*[49], with a significance threshold of $q \leq 0.05$.

Visualization. Plots were created using R (v.4.2.1) and ggplot2 (ggplot2.tidyverse.org/). The SAMOSA-Tag data were visualized using a modified version of IGV v.2.17.3 (github.com/RamaniLab/SMRT-Tag/tree/main/igv-vis). The FACS plots were created in FlowJo v.10.8.2 (BD Biosciences).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The SMRT-Tag data are deposited in the NCBI Sequence Read Archive (SRA) under accession no. PRJNA863422. The OS152 and mouse ES cell SAMOSA-Tag data, including the subreads and kinetic parameters, are deposited in the Gene Expression Omnibus (GEO) under accession no. GSE225314. The PDX SAMOSA-Tag data are available via the controlled access database of Genotypes and Phenotypes repository (study accession no. phs003511.v1). The following reference genome assemblies or annotations were used in this study: hs37d5 GRCh37; hg38 GRCm38; a concatenated hg38/GRCm39 reference; GENCODE v.28 and M25; and UCSC hg19 tandem repeats. The NIST/GIAB GRCh37 genome stratifications (v.3.0), small variant benchmarks for HG002, HG003 and HG004 (v.4.2.1), and Tier 1 SV calls for HG002 (v.0.6) were obtained from the NCBI (Supplementary Methods). The following additional publicly accessible datasets were used: GIAB-generated PacBio Sequel II HiFi reads from HG002 (SRA accession no. SRX5527202); CpG methylation calls from bisulfite sequencing of HG002 (ONT Benchmark Datasets; Supplementary Methods); CTCF binding sites determined by ChIP–seq in U2OS (GEO accession no. GSE87831); LNCaP (ENCODE accession no. ENCFF275GDH) cells; and chromHMM annotations for normal prostate (National Genomics Data Center accession no. OMIX237-64-02). Source data are provided with this paper.

Code availability

The source code used to perform the analyses and step-by-step SMRT-Tag and SAMOSA-Tag protocols is publicly accessible via GitHub at github.com/RamaniLab/SMRT-Tag and *Zenodo* (https://doi.org/10.5281/zenodo.10933181)[52].

References

50. 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**, 271 (2022).
51. Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
52. Siva, K., Nanda, A. J., Ramani, V. & Wu, K. SMRT-Tag and SAMOSA-Tag. *Zenodo* https://doi.org/10.5281/zenodo.10933180 (2024).

Acknowledgements

We thank members of the Ramani and Goodarzi laboratories, R. Corces (Gladstone Institutes) and R. Majzner (Dana-Farber Cancer Institute) for critical discussions and the Stanford Research Computing Center for providing the computational resources that contributed to this study. This work was supported by National Institutes of Health grant nos. DP2HG012442 (to V.R.), R01GM123977 (to H.G.), R01CA243555 (to E.A.S.-C.) and T32AR050942 (to S.K.); a Searle Scholars Award (to V.R.); a Pew-Stewart Scholars for Cancer Research Award (to A.T.S.); a Challenge Award from the Prostate Cancer Foundation (to H.G.N.); and a physician-scientist development award from the Stanford Department of Pediatrics (to S.K.). S.K. is an Ernest and Amelia Gallo Endowed Fellow of the Stanford Maternal and Child Health Research Institute.

Author contributions

S.K. and V.R. conceptualized the study. S.K., V.R. and H.G. designed the study. A.S.N., K.W. and I.I. performed the experiments. M.S.O. assisted with the preparation of the sequencing libraries. A.S.C., L.C.S. and E.A.S.-C. assisted with the osteosarcoma cell experiments. B.W., L.X. and H.G.N. assisted with the PDX experiments. V.R., S.K., A.S.N. and K.W. analyzed the data with support from A.T.S. V.R. and S.K. wrote the manuscript with input from all authors. S.K., V.R. and H.G. supervised the work.

Competing interests

V.R., S.K., A.S.N., H.G. and K.W. are inventors on a provisional patent related to this study. A.T.S. is a scientific founder of Immunai, founder of Cartography Biosciences, founder of Prox Biosciences, advisor to Zafrens, advisor to Pallando Therapeutics, advisor to Wing Venture Capital, and receives research funding from Merck Research Laboratories. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-024-01748-0.

Correspondence and requests for materials should be addressed to Sivakanthan Kasinathan or Vijay Ramani.

Peer review information *Nature Genetics* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

nature portfolio

Corresponding author(s): Vijay Ramani, Sivakanthan Kasinathan

Last updated by author(s): Mar 10, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

□ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

□ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

□ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

□ ☒ A description of all covariates tested

□ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

□ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

□ ☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

☒ □ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☒ □ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

□ ☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Sequencing data was collected using a Pacific Biosciences Sequel II running SMRTlink 11.0.0.146107. Raw data was processed using ccs (Pacific Biosciences, v6.4.0) and demultiplexed using lima (Pacific Biosciences, v2.6.0). Data from libraries sequenced on multiple flow cells was merged using pbmerge (v1.0.0). Images of agarose electrophoresis gels were collected using an Odyssey XF imaging system (LI-COR, software v1.1.0.61). FACS data were collected on a FACS Aria II with FACS Diva v9.0.1 (BD Biosciences).
Data analysis	All scripts required to perform the analyses described in this study and a complete list of required software are available via GitHub at https://github.com/RamaniLab/SMRT-Tag. Scripts utilize Python (v3.8.8) and R (v4.2.1), as well as the following packages and tools: samtools (v1.15.1), bcftools (v1.15.1), bedtools (v2.30.0), mosdepth (v0.3.3), ccs (v6.4.0), pbmm2 (v1.9.0) with minimap2 (v2.15), lima (v2.6.0), primrose (v1.3.0), hap.py (v0.3.12), Truvari (v3.3.0), deepvariant (v1.4.0), pbsv (v2.8.0). FACS data were processed and visualized with FlowJo (v10.8.2, BD Biosciences). SAMOSA-Tag fibers were visualized using a modified version of IGV v2.17.3, available at https://github.com/RamaniLab/SMRT-Tag/tree/main/igv-vis.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

SMRT-Tag data are deposited in the NCBI Sequence Read Archive (SRA; accession number PRJNA863422). OS152 and mESC SAMOSA-Tag data, including subreads and kinetic parameters, are deposited in the Gene Expression Omnibus (GEO; accession number GSE225314). PDX SAMOSA-Tag data are available via the controlled access dbGaP repository (study accession phs003511.v1). The following reference genome assemblies or annotations were used in this study: hs37d5 GRCh37, hg38, GRCm38, a concatenated hg38/GRCm39 reference, GENCODE V28 and M25, and UCSC hg19 tandem repeats. NIST/GIAB GRCh37 genome stratifications (v3.0), small variant benchmarks for HG002, HG003, and HG004 (v4.2.1), and Tier 1 SV calls for HG002 (v0.6) were obtained from NCBI (see below). The following additional publicly accessible datasets were used: GIAB-generated PacBio Sequel II HiFi reads from HG002 (SRA accession SRX5527202), Bismark CpG methylation calls from bisulfite sequencing of HG002 (ONT Benchmark Datasets; see below), CTCF binding sites determined by ChIP-seq in U2OS (GEO accession GSE87831) and LNCaP (ENCODE accession ENCFF275GDH) cells, and chromHMM annotations for normal prostate (NGDC accession OMIX237-64-02; https://ngdc.cncb.ac.cn/omix/release/OMIX237).

hg19 tandem repeats in BED format were downloaded from UCSC:
ftp://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.trf.bed.gz

VCF and BED files for NIST/GIAB small variant benchmarks were obtained from:
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/

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

GIAB-generated HG002 PacBio Sequel II HiFi data were downloaded from:
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

GIAB Tier 1 SV calls for HG002 were downloaded 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

HG002 bisulfite sequencing CpG methylation calls were downloaded from:
https://ont-open-data.s3.amazonaws.com/gm24385_mod_2021.09/bisulphite/cpg/CpG.gz.bismark.zero.cov.gz

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Tissues used to derive prostate cancer xenografts were donated by one participant of male sex. Sex and gender were not considered in study design.
Reporting on race, ethnicity, or other socially relevant groupings	No socially relevant categorization variables were reported or used in the study.
Population characteristics	Tissues used to derive prostate cancer xenografts were donated by a 71-year-old male participant with metastatic castration-resistant prostate cancer.
Recruitment	The participant who donated tissue for derivation of prostate cancer xenografts was recruited from the UCSF Urology service.
Ethics oversight	This study is covered by an active human subjects approval (protocol number 90911 'Use of Marker in Cytometric Analysis in Prostate Cancer to predict biological potential' UCSF IRB 11-05226).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

No specific calculation was performed to establish a required sample size for experiments in this study.

In experiments where specific, relevant conditions were tested that eventually were incorporated into the final SMRT-Tag method (i.e gap repair optimization, top hits) a minimum of two technical replicates were generated as a trade off between reagent usage and accuracy in estimating reaction efficiency.

The number of replicate SAMOSA-Tag experiments was determined to be a minimum of 3 based on previous experiments analyzing m6dA footprinting data (see Abdulhay et al. 2019, eLife, doi: 10.7554/eLife.59404) and was increased to n=8 for our OS152 and mESC SAMOSA-Tag experiments and n=6 for both primary and metastasis PDX SAMOSA-Tag experiments, to improve our analysis of reproducibility.

Data exclusions

For both OS152 SAMOSA-Tag and PDX SAMOSA-Tag analyses, unmethylated or lowly m6dA methylated fibers were excluded on the basis of lack of m6dA signal for determining nucleosome footprints. For OS152 SAMOSA-Tag data, across all n=8 replicates, these excluded fibers accounted for ~12.5% of all examined fibers. For PDX SAMOSA-Tag data (both primary and met.) across all n=6 replicates, these excluded fibers accounted for ~1.5 % of all examined fibers.

Replication

To verify reproducibility for SMRT-Tag gap repair optimization, the top performing gap-repair conditions were tested on multiple input samples, as well as with different amounts of input DNA. For each test, repair efficiency was ascertained by yield as well as fragment size distribution – see Supplementary Figures 1. and 2, and Supplementary Table 2 and Table 3. Selected repair conditions (Phu/Taq) performed adequately across a range of inputs, and library preparation yields were generally consistent when stratified by input amount and type.

For OS152 SAMOSA-Tag experiments, n=8 replicates were generated, and downstream analyses determining fiber type distributions and enrichments performed independently to validate results were reproducible. Supplementary Figure 14 compares fiber type enrichment patterns (odds ratios) across technical replicates, and demonstrates 1) that fiber types discovered via SAMOSA-Tag are reproducible across replicates, and 2) that all 8 replicates are highly consistent with each other. The same analysis performed separately on primary and met. PDX SAMOSA-Tag datasets, n=6 replicates, also indicated a high level of consistency.

Randomization

We did not perform any experiments that required randomization. Our study focused primarily on developing a new method, and demonstrating its utility by profiling relevant samples.

Blinding

We did not perform any experiments that required explicit blinding. For fiber type and binding site cluster determination, clustering analyses on single molecule fibers were performed using an unsupervised (leiden) clustering algorithm. Blinding is not required or routinely performed for this genomic analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
□	☒	Antibodies
□	☒	Eukaryotic cell lines
☒	□	Palaeontology and archaeology
□	☒	Animals and other organisms
☒	□	Clinical data
☒	□	Dual use research of concern
☒	□	Plants

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
□	☒	Flow cytometry
☒	□	MRI-based neuroimaging

Antibodies

Antibodies used

PE anti-mouse H-2 antibody (BioLegend 125505) was used at dilution 1 μg antibody per 8 – 12.5 x 10^6 cells

Validation

Per the manufacturer, "Each lot of this antibody is quality control tested by immunofluorescent staining with flow cytometric analysis." Also per the manufacturer, relevant citations include:
Boyd DF, et al. 2020. Nature. 587:466.
Pyzik M, et al. 2014. J Immunol. 193:6061.
Oyarce C, et al. 2018. Front Immunol. 8:1794.
Bockerstett KA, et al. 2018. Int J Mol Sci. 19:E1096.
Saunderson S and McLellan A. 2017. J Immunol. 10.4049/jimmunol.1601537.

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	Metastatic osteosarcoma cell line OS152 was provided by Alejandro Sweet-Cordero at the University of California, San Francisco. Mouse embryonic stem cells were a gift from Elphege Nora at the University of California, San Francisco.
Authentication	The authenticity of the OS152 cell line was confirmed by genotyping following a protocol previously used (see Sayles et al. 2019, Cancer Discovery, doi: 10.1158/2159-8290.CD-17-1152) using CellCheck 9 Plus (IDEXX BioAnalytics). mESC E14 cells were not authenticated directly, though genotyping data from PacBio sequencing of cells from this line used in various studies appear concordant.
Mycoplasma contamination	The OS152 cell line was tested for mycoplasma contamination, and aliquots used in this study were confirmed to be negative for mycoplasma. The mESC E14 cell line used in this study was also tested for mycoplasma contamination and confirmed to be negative for mycoplasma.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used in this study.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	Patient-derived xenografts were implanted subcutaneously into 6- to 8-week-old male NOD scid gamma (NSG) mice (UCSF Breeding Core) maintained under specific pathogen-free conditions.
Wild animals	n/a
Reporting on sex	Sex was not considered in the study design.
Field-collected samples	The study did not involve field-collected samples.
Ethics oversight	Experiments were performed under a protocol approved by the UCSF Institutional Animal Care and Use Committee (IACUC; number AN195508).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Plants

Seed stocks	n/a
Novel plant genotypes	n/a
Authentication	n/a

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Primary and metastasis PDX models derived from one patient with both primary and lymph-metastatic prostate cancer were used in this study. Creation of these PDX models is described in (Nguyen et al, 2018, Science Translational Medicine). PDX tumors were passaged in NSG mice, with verification of tumor similarity to the original biopsy via histopathological and

growth-rate based comparisons. On the day of processing, samples were first surgically removed from mice, placed into sterile collection buffer, and dissociated manually using sterile surgical blades, as well as using enzymatic treatment (see Methods). Resulting single cell suspensions were washed via centrifugation at 4ºC with PBS, strained to remove aggregates through a 70µm filter, and then stained in Cell Staining Buffer (Biolegend) with 1µg PE anti-mouse H-2 Antibody (Biolegend, Cat# 125505), and 1µL of SYTOX Red Dead Cell Stain (Thermo Fisher).

Instrument

BD FACS ARIA II (BD Biosciences)

Software

Data was acquired with FACS DIVA (v9.0.1, BD Biosciences) and was visualized and analyzed using FlowJo (v.10.8.2, BD Biosciences)

Cell population abundance

The relevant cell population (not mouse, not dead) was determined to be ~ 16.05% of the primary PDX and ~ 14.3% for the metastasis PDX sample via FACS. SAMOSA-Tag libraries prepared from the resulting population were mapped to both human and mouse genomes, and sample purity as estimated by the fraction of human alignments is between 27.5 - 32.1% for the primary PDX and 96.3-96.7% for the metastasis PDX.

Gating strategy

Cell singlets were selected by gating on forward scatter. Subsequently, an APC negative gate corresponding to live cells (SYTOX Red) was defined by calibrating against a single-stain control. A PE negative gate corresponding to "not-mouse" cells was similarly defined by calibrating against a single-stain control. In both cases, the gate was set at the minimum between the negative and positive signal peaks in the single-stain controls. The intersection of the two gates defined the relevant cell population. Supplementary Figure 15 exemplifies the gating strategy.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.