Advancing Locus-specific Chromatin Isolation Methods for Multi-omic Discoveries

Yuzhen Liu

A dissertation

submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee

Brian J Beliveau, Chair

Celeste Berg,

Hao-Yuan Kueh

Program Authorized to Offer Degree:

Molecular & Cellular Biology Graduate Program

University of Washington

Abstract

Advancing Locus-specific Chromatin Isolation Methods for Multi-omic Discoveries

Yuzhen Liu

Chair of the Supervisory Committee:
Brian Beliveau
Department of Genome Sciences

Chromatin functions, such as gene expression, DNA replication and repair, and 3D genome architecture, are mediated by the concerted action of the unique molecular repertoire of interacting proteins and DNAs at the gene locus of interest. Thus, a comprehensive functional understanding of a locus requires the thorough characterization of its local chromatin composition. As a prerequisite to unbiased identification of locus-bound factors, the purification of specific gene loci has attracted considerable scientific attention as an important technical challenge that if solved, promises crucial biological discoveries. During my dissertation work, I have contributed to the development of locus-specific chromatin isolation methods easily adaptable to large, repetitive loci as well as small, single-copy ones. The work I describe here includes the development of “DNA O-MAP”, a labeling method that biotinylates proteins around a gene locus using peroxidase-conjugated oligonucleotides. In collaboration with colleagues, I demonstrate that DNA O-MAP, coupled with sample multiplexed

quantitative proteomics and next-generation sequencing, can generate well-annotated DNA-protein networks at specific gene loci.

ACKNOWLEDGEMENTS

First and foremost, I would like to extend my deepest gratitude to my advisor, Dr. Brian Beliveau, for his empowering guidance for my scientific pursuits, unwavering belief in my abilities, invaluable feedback for my intellectual and personal growth. I am profoundly grateful for the privilege of pursuing my PhD in an environment that champions scientific rigor, technological creativity, and work-life well-being. Over the past five years, I attained transformative growth in my critical thinking, mental tenacity, and physical fitness. None of this would have been possible if Brian did not welcome me into his lab. Thank you immensely, Brian.

I am also deeply indebted to mentors in my thesis committee: Dr. Devin Schweppe, for unveiling those infinitely minute local proteomes with his expertise and for running a remarkably friendly and supportive lab for their collaborators; Dr. Celeste Berg, whose encouragement and appreciation of my work always feels like a ray of warm morning sun; and Drs. Hao-Yuan Kueh and Gavin Ha, who were both infinitely patient and attentive as I navigate through the challenges in both graduate school and the life beyond.

I also had the pleasure of working alongside many kind and brilliant colleagues and friends. A special thanks to Chris McGann, who never ceased to fill our collaboration with his genuine camaraderie and uplifting humor. I cannot thank my lab enough, especially Eva Nichols, Robin Aguilar, Lily Deng, Sahar Attar, Mary Krebs, Valentino Browning, David Nwizugbo, Qiaoyi Lin, Lidan Li, and Conor Kelly, for selflessly providing me their time, allyship, and insights for my scientific work and personal growth. I am also incredibly fortunate to have many dear friends in Genome

Sciences, especially Leah Anderson, Sayeh Gorjifard, Sophie Moggridge, Ran Zhang, Pengyao Jiang, and Valentina Grillo. Those countless great times we spent together were the most enriching and fulfilling parts of my PhD life.

Chapter I: Introduction

1.1. Chromatin proteins underpin eukaryotic DNA functions.

1.1.1 Chromatin architecture

The genome carries a set of DNA molecules that contain the information to give rise to the cell's structural and functional machinery. In a eukaryotic cell, DNA is packaged tightly in the form of chromatin, with proteins about twice as much as DNA in mass. The human genome is no exception. The 46 linear DNA molecules wrap around cores of histone proteins to form nucleosomes, which coil into chromatin fibers, followed by further compaction by genome architectural proteins of the fibrous secondary structure into higher-order chromatin structures(P. Chen et al., 2021; Maeshima et al., 2021), such as the dynamic chromatin loops organized by CCCTC-binding factor (CTCF) and the cohesin complex(Hansen et al., 2017), as well as lamina-associated domains organized by nuclear pore complexes and nuclear envelope proteins(Briand & Collas, 2020). DNA-binding proteins also protect and stabilize single-stranded DNA that arises temporarily or in specialized regions, such as the replication protein A on newly synthesized DNA during repair and replication(R. Chen & Wold, 2014), as well as the shelterin complex on the single-stranded telomeric region(de Lange, 2005).

1.1.2 Transcriptional activation

Although every cell in the organism contains largely the same nucleotide sequence, a differentiated cell executes highly specialized functions and exhibits adaptable cell states by its unique gene expression in response to changes in physiological signals. The cell's distinct identity is the result of the dynamic protein residents on the genome, the resulting three-dimensional (3D) genome structures, and the gene expression pattern they facilitate.

With nearly two meters of DNA packed into a tiny 10-micron nucleus, the condense nature of chromatin presents a topological challenge to the cell: the timely access of a gene locus for its specialized functions. Beyond providing the structural basis for chromatin, DNA-binding proteins orchestrate gene activation and repression by specifying its precise location, timing, and magnitude of the response. Gene activation is facilitated by a multitude of proteins, histone modification depositions, and long-range DNA interactions. The complex array of molecules includes chromatin modifying/remodeling complexes to alter the nucleosome position and allow access to the locus, gene-specific transcription factors and activators recruited to promoters and enhancers to engage their contact, and general transcription factors including the preinitiation complex and RNA polymerase II holoenzyme to synthesize mRNA. This series of dynamic DNA-centered protein interactions ensures faithful transmission of genetic information(T. I. Lee & Young, 2000).

Specifically, transcriptional activation is initiated by the binding of transcriptional activator proteins with a domain that binds DNA and one that recruits the transcription apparatus to upstream sequences of the gene. The enhancer sequence acts as the integration point of multiple regulatory inputs by forming an incredibly stable nucleoprotein complex of regulator proteins - the enhanceosome. The synergy of these proteins allows for a large dynamic range of transcription activation at a low concentration(Merika & Thanos, 2001). A well-understood example is the virus-inducible enhanceosome of the interferon-β (IFN-β), a normally silent gene. Its enhancer, located between -110 and -45 bases relative to the transcription start site, contains four binding sites for the high mobility group protein HMG I(Y), a protein bearing multiple DNA-binding domains and protein-protein interaction surfaces. Upon viral infection, HMG I(Y) binding unbends the DNA sequence and lowers the free energy required for the recruitment of NF-κB, ATF-2/c-Jun heterodimer, and the interferon regulator proteins into the enhanceosome(Yie et al., 1999). The cooperative binding between the architectural protein

HMG I(Y) and the activators forms a highly stable nucleoprotein complex that permits IFN-β transcription for multiple rounds in response to a viral infection.

1.1.3 DNA replication

In eukaryotes, DNA replication initiates at multiple origins of replication distributed across the genome by the replisome protein assemblies. The highly regulated DNA replication begins with the assembly of a pre-replication complex at the origins of replication in the human genome. The 6-subunit origin recognition complex (ORC) is a conserved complex essential for DNA replication in eukaryotes. Throughout the cell cycle, the binding of ORC to the origins of replication is dynamic in human cells. As the first ORC subunit to appear on mitotic DNA, ORC1 recruits other ORC subunits. The complete, chromatin-bound ORC complex becomes the platform for recruiting other proteins in the pre-replication complex, Cdc6 and Cdt1. The binding of these proteins loads the minichromosome maintenance (MCM) replicative helicase to wrap around the duplex DNA. The hexameric MCM ring, together with Cdc45 and the GINS heterotetramer consisting of Sld5, Psf1, Psf2, and Psf3, forms the active CMG replicative helicase complex. Once activated, the MCM helicase unwinds DNA within its central channel, and reconfigures to encircle each strand of DNA to establish the replication fork to allow other replisome proteins to function(Chou et al., 2021; Costa & Diffley, 2022; Coster & Diffley, 2017). Using the separated DNA strands as templates, DNA polymerases Pol ε and Pol δ synthesize new complementary DNA on the leading and the lagging strands respectively. The PCNA sliding clamp encircles DNA, binds the DNA polymerase and promotes processivity in DNA replication. Many replisome factors, including the fork protection complex, Claspin, And1, and replication factor C, coordinate DNA synthesis with DNA unwinding by regulating the functions of the polymerases and the CMG complex.

Although the 3.055 billion nucleotides of the human genome have been assembled without gaps(Nurk et al., 2022), the above examples only represent a fragmentary

understanding of the essential protein context for genome function. To this day, except several well-known loci, the proteins assembled on most genomic loci remain largely unknown. Mapping how the nucleotides in the locus are packaged with chromatin proteins *in vivo* will serve as the starting point of mechanistic understanding of its regulation and function. Since the second half of last century, the chromatin biology field has seen significant efforts to develop technologies to depict a comprehensive DNA-protein network from a DNA- or protein-centric lens at an increasing resolution.

1.2. *In vitro* DNA-protein binding analysis methods

1.2.1 Electrophoretic mobility shift assays (EMSAs)

Since the 1980s, EMSA has been used to characterize DNA-protein interactions in fresh cell extracts or *in vitro* incubation of purified protein and oligonucleotide pairs in physiological pH, salt, and necessary factor concentrations. Following separation of the bound and free fractions, the bound fraction can be identified by the migration retardation on native electrophoresis and the protein can be identified by immunoblotting, whereas the DNA can be identified by biotin or radioisotope labels, or PCR amplification using selected oligonucleotides(Fried & Crothers, 1981; Garner & Revzin, 1981). One of the earliest complex of DNA and protein interactions evaluated by EMSA is a prokaryotic transcriptional control complex, consisting of cyclic AMP (cAMP), the *E. coli* cyclic AMP receptor protein (CAP), RNA polymerase, and a 214-bp fragment from the *E. coli lac* promoter-operator region. Garner and Revzin revealed that in the presence of cAMP, CAP-promoter binding stabilized RNA polymerase-DNA binding as a complex, which becomes destabilized if CAP is removed, or if the promoter fragment is a CAP-insensitive mutant.

1.2.2 Surface plasmon resonance (SPR) sensing

Although sensitive and versatile for DNA and proteins of various sizes, EMSAs and related variations presented several drawbacks. One of them is throughput – EMSAs are cumbersome to perform and are only able to assess a highly limited set of candidates at a time. Yet eukaryotic genome control processes often involve a sequence of recognition and binding events of multi-unit molecular complexes. Furthermore, the binding between DNA and protein partners are not stable. Electrophoresis itself or even shearing force from vortexing is likely to induce dissociation of the complex(Fried & Crothers, 1981). Electrophoresis may also promote stable complexing of the DNA protein candidates. Therefore, a short electrophoresis is recommended to mitigate this issue(Hellman & Fried, 2007). Besides electrophoresis, other factors that contribute to mobility shift, such as nucleic acid structure formation, may also yield misleading results and/or less reliable quantitation.

Since the late 1990s, SPR sensing, an optical biosensing technique, offers a high-throughput quantitative solution for studying *in vitro* biomolecular interactions including DNA-protein binding. When a light beam is shined to a thin metal film, electrons in the metal become excited and a surface plasmon wave is generated. The wave propagates in parallel to the boundary between the metal film and the external medium and is sensitive to irregularities on the boundary. With any molecular adsorption on the surface, the plasmon wave cannot be formed. This phenomenon can be used to detect binding events by having probe molecules immobilized on the biosensor surface. When the analyte containing the target molecule flows onto the surface, if the target molecule binds to the probe, the refractive index at the surface increases sharply and consequently, the intensity of the reflected light decreases. This change can be detected in real time by capturing the reflected light via a charge-coupled device (CCD) camera. SPR sensing is quick, quantitative, and allows simultaneous processing of hundreds of

samples, facilitating its use for high throughput screening of protein binding activity in drug development(Nguyen et al., 2015; Stockley & Persson, 2009).

Using the commercial SPR biosensor BIAcore system, relatively simple DNA-protein interactions with a small, well-defined set of players have been measured in real time. Fisher *et al.* had measured the different kinetics of the p42 and p51 isoforms of the human ETS1 oncoprotein binding to the wild-type oligonucleotides containing the binding site as well as 9 single-nucleotide mutants(Fisher et al., 1994). Parsons *et al.* traced the prokaryotic transcriptional control complex consisting of the *E. coli* transcriptional repressor protein MetJ, the co-repressor S-adenosylmethionine, RNA polymerase, and a fragment of the 5' operator region of *E. coli* genes involved in methionine synthesis and verified that the removal of the co-repressor leads to the rapid release of MetJ from DNA(Parsons et al., 1995).

In general, *in vitro* techniques like EMSAs and SPRs examine DNA-protein interactions in simplified model systems. They do not intend to enumerate all DNA-protein interactions or to recreate the native cellular environment in which the complex is formed. They also require significant prior knowledge about the sequences and proteins of interest. The emergence of genome-wide and proteome-wide detection approaches such as next generation sequencing (NGS) and mass spectrometry (MS) allow DNA and protein molecules to be identified and quantified in an unbiased manner. Combined with *in vitro* or *in vivo* capture techniques, constructing protein-centric or DNA-centric interactomes becomes aspirations within reach.

1.2.3 Affinity purification followed by MS

Modern MS is a powerful technological advancement that enables the determination of components of complex protein mixtures. Combined with the Stable Isotope Labeling by Amino acids in Cell culture (SILAC) method, two purifications of DNA-binding proteins can be identified, quantified, and compared to yield a sequence-specific proteome. In 2009, Mittler *et al*. introduced an *in vitro* DNA protein interaction screen that aims to identify sequence-specific or

modification-specific protein factors with a one-step affinity purification from metabolically encoded nuclear extracts(Mittler et al., 2009). Synthetic biotinylated double-stranded DNAs containing either the wild-type functional element or the mutant control were used as enrichment handles. In this study, the DNA elements were binding sites for the transcription factors AP2 and ESRRA, as well as control sequences with point mutations. After incubation of HeLa-S3 nuclear extracts from 250 million cells and the biotinylated DNA baits, the protein-DNA complexes formed *in vitro* were purified with streptavidin beads. The washed streptavidin beads were combined and subjected to gel electrophoresis, in-gel digestion, and liquid chromatography(LC)-MS. This screen was able to recover the transcription factors that the DNA motifs were designed to capture, and several other proteins with moderate fold changes. The small number of hits is likely due to interactions between proteins and DNA-baits being significantly weaker than the one between streptavidin and biotin, leading to either loss of the DNA-bound proteins during purification washes, and/or the insufficient wash stringency. More importantly, limited by the *in vitro* experiment design, this method studies DNA-protein interactions on short synthetic DNAs that are only 26 base pairs. Considering that excluding the critical context of 3D genome architecture in the cell may lead to false positive interactors or failure to identify bonafide interactors, this method could serve validation purposes and complement an *in vivo* DNA-protein interaction screen.

1.3. *In vivo* chromatin immunoprecipitation (ChIP) followed by sequencing

Gilmour and Lis introduced the ChIP technique for the first time in 1985 to examine the DNAs bound by RNA polymerase II (Pol II) in Schneider's Drosophila line 2 cells with and without heat shock(Gilmour & Lis, 1985). Specifically, they found that Pol II occupied the entire *hsp70* gene upon heat shock due to heat-induced gene expression, whereas in the absence of

heat shock, Pol II was confined to the 5' end of the gene near the transcription start site, priming the gene for transcriptional activation. This initial study started with UV-crosslinking 300 million or more Schneider line 2 cells to preserve the protein-DNA interactions. Crosslinked nuclei are then extracted in a low-salt, hypotonic buffer. Using CsCl ultracentrifugation, chromatin was extracted from the middle fraction of the CsCl gradient with a needle, removing free DNA in the pellet and free protein in the top layer. Subsequently, chromatin was fragmented by restriction enzyme digestion or random cleavage by sonication, followed by immediate immunoprecipitation (IP) using the antiserum against the protein. Total DNA and immunoprecipitant DNA were ethanol precipitated and treated with RNase A and proteinase K to remove the proteins and RNAs in the chromatin. Gilmour and Lis detected DNAs without amplification using Southern blotting and hybridization assays with radiolabeled cloned DNAs, since this study predates the widespread adoption of PCR technique, which was first described in Saiki *et al*. at the end of 1985(Saiki et al., 1985).

In the decades following its introduction, ChIP became a common method for detecting the interaction between a protein and a DNA sequence *in vivo*. Owing to the technological advances in high-throughput nucleic acid identification including DNA microarrays(Kim & Ren, 2006; van Steensel, 2005) and subsequently high-throughput DNA sequencing(Fields, 2007; Johnson et al., 2007), ChIP realized its full potential by having a robust readout. In 2007, Johnson *et al*. first combined ChIP and high throughput sequencing (ChIP-seq) to examine the distribution of the transcriptional repressor, neuron-restrictive silencer factor, across the human genome in Jurkat cells, a T-cell line. Since its introduction, ChIP-seq facilitated the binding site mapping for various chromatin proteins including transcription factors, chromatin remodelers/modifiers, modified histones, and genome architectural proteins. Nowadays, ChIP-seq makes up a fundamental class of data in the Encyclopedia of DNA Elements (ENCODE) database, a National Human Genome Research Institute-funded research consortium that aims to systematically assign biochemical functions to the genomes of human

and common model organisms(ENCODE Project Consortium, 2004; Luo et al., 2020)[20]. As of now, the ENCODE portal hosts a total of 800 ChIP-seq studies on modified histones and 2989 studies on transcription factors for human cell lines, providing the basis for the annotation of chromatin states and invaluable insights of the distribution of transcription factors, transcriptional machinery, and chromatin structure proteins.

ChIP-seq is a powerful and widely adopted technique that has revolutionized our understanding of gene regulation, genome architecture, and the epigenetic landscape. Despite its immense power, the technique aims to scrutinize DNA-protein interactions through a protein-centric lens. Evaluating one chromatin protein at a time relies on the *a priori* knowledge of the protein's DNA occupancy and the availability of a high signal-to-noise antibody for its immunoprecipitation. Integrative analysis of multiple ChIP-seq studies of co-localizing proteins may reconstruct a partial protein landscape at a specific locus, but the genomics-focused technique does not intend to enumerate the protein repertoire at a specific locus. A complete interactome of a locus can only be achieved when DNA-interacting proteins are detected in an unbiased manner from a locus of interest.

1.4. *In vivo* chromatin enrichment followed by proteomics

Mass spectrometry analysis has emerged as a pivotal method for the unbiased detection of proteins associated with the genome. Purifying chromatin followed by proteomic profiling has proven powerful compared to holistic whole nuclear proteome analysis, as enrichment allows the identification of typically low abundance chromatin proteins without extensive sample fractionation prior to LC-MS(van Mierlo & Vermeulen, 2021). So far, efforts of chromatin enrichment range from the isolation of total chromatin, chromatin subtypes, repetitive genomic regions, to low-copy genomic elements. As we attempt to capture the DNA-interacting proteome at an increasing kilobase-resolution, a larger amount of input material is required to ensure

sufficient peptide amount for detection. In the context of large-scale input preparation and enrichment, the process efficiency becomes particularly important as it directly affects assay throughput.

1.4.1 Isolation of total chromatin

The chromatin is a dynamic structure that undergoes significant proteomic reorganization during embryonic development as well as every cell division. Tracing the chromatin proteome as the cell progresses through the biological processes provides a useful starting point for understanding the dynamic protein network driving the establishment of cellular identity and normal cell proliferation. The biochemical and proteomic studies of chromatin depend on the ability to efficiently isolate chromatin in high yield and purity for downstream analysis. In 2014, Kustatscher *et al*. introduced Chromatin Enrichment for Proteomics (ChEP), a method for the enrichment of whole chromatin that starts by *in vivo* cross-linking of cells with 1% formaldehyde as in standard ChIP procedures(Kustatscher, Hégarat, et al., 2014; Kustatscher, Wills, et al., 2014). After cell lysis in 0.1% Triton X-100, nuclei are digested with RNase to remove RNA-protein complexes, followed by lysis in 4% (wt/vol) SDS. Soluble nuclear proteins that are not crosslinked to DNA were washed away in an 8M urea buffer. These denaturing but salt-free buffer conditions leave the chromatin to aggregate as a transparent, gelatinous pellet after centrifugation, which can be solubilized by sonication and analyzed by MS (**Figure 1**). Using SILAC, Kustatscher *et al*. compared interphase ChEP chromatin with other biochemical fractions such as nuclear and whole-cell lysates, as well as ChEP fractions with and without genome function perturbations including transcription activation/inhibition, DNA damage, replication inhibition in three human cell lines, MCF-7, HeLa, HepG2 cells. By integrative analysis of these chromatin proteomics datasets with a random forest-based algorithm, Kustatscher *et al*. established a probability score for 7635 proteins to describe how likely the protein belongs to the chromatin proteome.

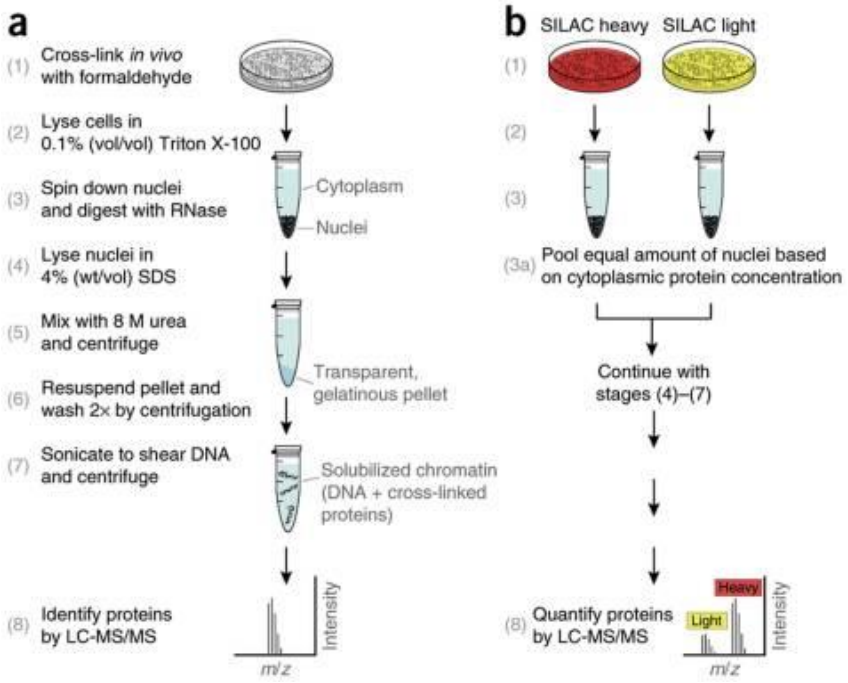


Figure 1. Outline of the ChEP procedure. (A) The key stages (numbers in parentheses) for chromatin enrichment of a single sample. (B) Scheme for a quantitative version of ChEP, comparing chromatin from different samples based on SILAC. Used with permission from Kustatscher et al. 2014.

Slightly optimizing ChEP, Ugur *et al*. characterized the chromatin reorganization during embryonic development through the naïve, formative, and primed states of pluripotent stem cells with Chromatin Aggregation Capture (ChAC) coupled with Data-Independent Acquisition-based proteomics(Ugur et al., 2023). Compared to ChEP, ChAC shifted the nucleus isolation to pre-fixation doing hypotonic extraction from whole cells, likely to better minimize contamination of cytoplasmic proteins crosslinked to chromatin. In addition, after chromatin enrichment with ChEP, ChAC cleans up liberated chromatin proteins with magnetic beads for enzymatic digestion and proteomics sample prep. Cell number requirement of ChEP is lenient (15 million cells per replicates) since total chromatin isolation has a relatively low fold of enrichment.

Although easy to perform in basic molecular biology labs, the ChEP procedure does not completely remove cytosolic contamination. Kustatscher *et al*. concluded that covariation with well-known reference chromatin proteins predicts chromatin components more accurately than

biochemical enrichment by ChEP(Kustatscher et al., 2016). The traditional density-based CsCl gradient ultracentrifugation remains a reliable method for the isolation of total chromatin(Gilmour & Lis, 1985; Ginno et al., 2018; van Mierlo & Vermeulen, 2021). These total chromatin enrichment methods are valuable tools for establishing expectations of protein abundances in the chromatin, so that functionally relevant proteins can be distinguished from the biochemical background in the analysis of locus-specific proteomes.

1.4.2 Isolation of subtypes of chromatin

1.4.2.1 Isolation of nascent chromatin

Beyond separating total chromatin from other cell fractions, several approaches had aimed to further fractionate chromatin by distinctive states or subtypes of interest. In 2014, Alabert *et al*. studied nascent chromatin assembled at or behind the replication fork during cell division(Alabert et al., 2014). After release from thymidine block, cells were incubated with biotinylated dUTPs to label newly synthesized DNA, and crosslinked either 20 minutes or 2 hours after labeling for the comparison between nascent and mature chromatin (**Figure 2**). Using 500 million HeLa S3 cells as input for each condition, biotinylated chromatin fragments were purified from sonicated soluble chromatin and combined for SILAC proteomics.

From the 3995 proteins quantified, enriched proteins included the majority of core replication fork components, accessory components of the fork, known chromatin components, and transcriptional regulatory proteins, indicating a successful enrichment of nascent chromatin overall. However, 878 proteins were functionally uncharacterized, or proteins without an expected chromatin function. Importantly, Alabert *et al*. incorporated the chromatin protein probability score established by Kustatscher *et al.* to further purify the 3995 proteins to distinguish functionally relevant proteins from the background of biochemical enrichment(Kustatscher, Wills, et al., 2014). This *in silico* purification removed a large number of

uncharacterized proteins and proteins with no expected chromatin function, ultimately preserving 93 novel proteins with a high probability to have a function at the replication fork or in the nascent chromatin behind it. Among the three novel proteins experimentally followed up, all three of them showed nuclear localization, two showed colocalization with the replication sliding clamp PCNA, while only one of them was verified by PCNA-fused GST pull-down. This study suggests that the importance of background estimation, *in silico* removal, and experimental validation of hits should not be undermined even in the case of a successful biochemical purification.

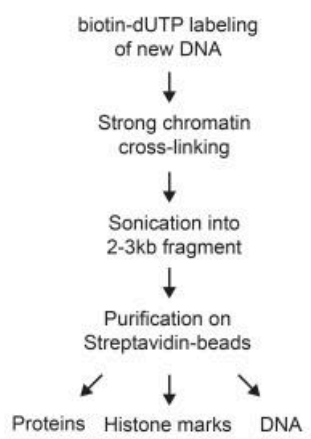


Figure 2. Outline of the Nascent Chromatin Capture protocol. Used with permission from Alabert et al. 2014.

1.4.2.2 Isolation of euchromatin and heterochromatin

Outside of DNA replication, interphase chromosomes exist as two structurally and functionally distinct subtypes: euchromatin and heterochromatin. Euchromatin is relatively less compact, allowing the dynamic access to chromatin remodelers and transcription machinery, and associated with active cell type-specific gene expression. Heterochromatin is highly condensed and transcriptionally silent, suppressing recombination in repeat-rich regions and turns off expression of cell type-inappropriate genes(Maeshima et al., 2021; Morrison & Thakur, 2021). Euchromatin and heterochromatin can be separated by partial micrococcal nuclease

(MNase) digest combined with salt fractionation(Henikoff et al., 2009; Herrmann et al., 2017). Due to its preference for nucleosome-free regions, using MNase at increasing amounts or treatment time can release proteins at euchromatin first, followed by those associated with condensed heterochromatin. In addition, since protein-DNA binding is based on electrostatic interactions, chromatin-associated proteins become more soluble in increasing salt concentrations. Weakly bound proteins such as transcription factors will elute in low salt buffers, whereas tightly bound proteins will become soluble only in high salt buffers. To summarize, mononucleosomal chromatin and euchromatin can be released from MNase-treated cells in low salt concentrations up to 150 mM NaCl, whereas in 600 mM NaCl, nearly 90% of the chromatin is soluble(Henikoff et al., 2009; Herrmann et al., 2017; van Mierlo et al., 2019).

In addition to MNase digest combined with salt fractionation, heterochromatin is more densely packed with high levels of protein occupancy and therefore can be separated by its sedimentation rate on a sucrose gradient(Becker et al., 2017). The “gradient top” contains both euchromatin and soluble proteins and the middle gradient fraction contains sonication-resistant heterochromatin. Since resistance to sonication is a reliable indicator of heterochromatin state, despite that density-based separation methods require a large number of cells as input (10^8), the biophysical enrichment of heterochromatin provides a useful alternative to ChIP methods based on sonication followed by antibody pull-down of heterochromatin marks.

1.4.3 Isolation of local proteome around a chromatin protein of interest

1.4.3.1 ChIP-MS, an IP-MS variant

To understand how a DNA-binding protein functions, it is important to depict the protein-protein interaction network where it is active. The need for a complete molecular context has motivated the development of a wide range of MS-based methods to study chromatin protein-proximal proteomes. ChIP-MS combines chromatin purification with proteomics to

enable the unbiased identification of chromatin-associated proteins directly or indirectly interacting a specific DNA-binding protein or modified histone(Ji et al., 2015; Mohammed et al., 2016; Soldi & Bonaldi, 2014). Using ChIP-MS, Ji *et al*. examined the euchromatin containing the histone modifications H3K27ac, H3K4me3, H3K79me2, and H3K36me3, and heterochromatin containing H3K9me3 and H4K20me3 across the genome of mouse embryonic stem cells. They identified 332 proteins known to associate with chromatin, 46 proteins implicated to associate with chromatin, as well as 114 novel candidates. Specifically, Ji *et al*. confirmed the presence of pluripotency transcription factors such as OCT4, SOX2, ESRRB, and SALL4, as well as transcription initiation apparatus such as Pol II and TFIID in the euchromatin prep.

Although ChIP-MS identified many well studied proteins in the expected chromatin preps, using ChIP for MS analysis comes with its unique limitations. Native ChIP, which relies on MNase digestion to solubilize chromatin, is limited to the isolation of highly abundant stable interactors of DNA such as modified histones, since the loss of rare proteins like transcription factors would be high during IP washes. For sparse proteins of interest, similar to ChIP-seq, the isolation of chromatin relies on chemical crosslinking to covalently preserve chromatin-associated proteins. Yet simultaneously, antibody-based enrichment is sensitive to crosslinking, which could cause epitope occlusion that impairs antibody recognition and eventually the pull-down efficiency. Thus, the fixation used here is light, making it challenging to strike a balance between ensuring epitope availability and performing stringent washes to reduce background. Furthermore, compared to ChIP-seq, the protein of interest needs to have more abundant expression levels in a cell type that can be obtained in sufficient quantities for proteome analysis post enrichment.

Instead of IP, an alternative approach is to affinity-tag the protein of interest in the cell type in study(Lambert et al., 2009; Vermeulen et al., 2010). This approach bypasses the shortcomings of antibodies, which may be unavailable, or have non-specific binding distinct to individual antibodies, complicating the comparison between different chromatin proteins. Affinity

tags also allow for stringent wash regimes that can reduce background noise. Although successfully applied in yeast and easy-to-manipulate mammalian cell lines such as HeLa S3, this strategy may be challenging in cell types that are difficult to transfect or cells that are already fixed. Furthermore, overexpression of a tagged protein may also introduce biological artifacts. Finally, as ChIP-based methods are protein-centric, they do not discriminate between the protein interactions occurring on the DNA from those in the nucleoplasm.

1.4.3.2 The development of enzyme-based proximity labeling (PL) technologies

In the past decade, PL has become an indispensable addition to the toolbox for studying protein interactomes due to their advantage of capturing weak and transient interactors compared to IP. The general principle of PL technologies involves fusing an engineered labeling enzyme, such as peroxidase or biotin ligase, to a specific protein of interest, such as a signaling spatially-restricted protein. This fusion protein targets the enzyme to a protein complex of interest or an organelle, where a small molecule substrate (E.g. biotin) can be covalently conjugated to the endogenous protein within nanometers of its vicinity, leaving a chemical trace of its protein interactions. Subsequently, this biotinylated proteome can be purified under stringent conditions and identified by MS. PL rely on two main classes of enzymes: biotin ligases, which include BioID, TurboID, and miniTurbo, etc., as well as peroxidases, which include horseradish peroxidase (HRP), ascorbate peroxidase (APEX), and APEX2(Guo et al., 2023; Kalocsay, 2019; Qin et al., 2021).

1.4.3.2.1 PL by BioID - an engineered biotin ligase with slow labeling kinetics

BioID, a biotin ligase-based PL technique, was first introduced in 2012(Roux et al., 2012, 2018). It deploys a mutant, promiscuous form of the biotin ligase BirA from *E coli*, denoted as BirA*, Wild-type BirA first catalyzes biotin and adenosine triphosphate (ATP) to form a reactive biotinoyl-5’-adenosine monophosphate (AMP), then retains this activated biotin within its active

site until it reacts with a primary amine on lysine residues on its specific substrate, the AviTag peptide. Since the mutant BirA* has an affinity for biotinoyl-5'-AMP at two orders of magnitude lower than its wild-type form, it prematurely releases the biotinoyl-5'-AMP and its extreme AviTag specificity is lost. Thus, PL is achieved with the promiscuous biotinylation of any adjacent primary amines. In its first introduction, BirA* was fused to a well-studied nuclear lamina component, laminA, an insoluble protein that would be difficult to analyze by affinity purification. Since its introduction, BioID has been successfully applied to elucidate the spatial proteomes of many protein complexes and organelles(Gupta et al., 2015; Youn et al., 2018), and even enabled the a BioID-based map of 192 subcellular proteins, revealing the subcellular locations of 4145 proteins in 2021(Go et al., 2021). However, due to its reduced affinity for biotin, biotinylation by BirA* requires a long labeling time of 18-24 hours in an excess concentration of biotin (50 μM). As a result, BioID has a low temporal resolution which would be ineffective at the capture of rapid signaling cascades or transient interactions.

1.4.3.2.2 PL by peroxidases with rapid labeling kinetics (HRP, APEX/APEX2)

In contrast to BioID, peroxidase-based approaches, such as HRP and APEX, convert the substrate biotin phenol into reactive phenoxyl radicals using hydrogen peroxide and covalently tags nearby proteins within a minute. The fast labeling kinetics enables it to capture snapshots of the changing interactomes in dynamic cellular processes. Although HRP is one of the most sensitive reporter enzymes known, the structure of HRP, maintained by four disulfide bonds, is disrupted by the reducing environment of the cytosol. The reducible HRP structure has limited its use to only extracellular environments such as the cell surface, and oxidizing environments such as the endoplasmic reticulum (ER) and Golgi lumen(Connolly et al., 1994; Martell et al., 2012).

Re-engineering of HRP has failed to overcome its limitation and instead led to the search for a better prototype enzyme. An ascorbate peroxidase of soybean origin naturally active in the

reducing cytosolic environment was engineered into the labeling enzyme APEX, which is active in all cellular compartments(Martell et al., 2012; Rhee et al., 2013). Shortly after BioID introduction, Rhee *et al*. introduced the use of APEX for proteomic labeling of the mitochondrial matrix by fusing APEX to the mitochondrial calcium uniporter, a channel protein on the inner mitochondrial membrane with both N- and C-termini facing the matrix. After normalization using two-state SILAC labeling, the background of endogenously biotinylated proteins was removed, leaving 495 proteins as putative mitochondrial matrix proteins. Owing to the high-quality reference mitochondrial proteome, 94% of the 495 proteins had previous mitochondrial annotation, demonstrating the high specificity of APEX labeling and high confidence of the newly discovered mitochondrial proteins (6%). In addition to membrane-bound organelles, APEX has also been used for proteomic profiling of non-enclosed organelles such as primary cilia(Mick et al., 2015). As an electron microscopy (EM) tag, APEX is able to catalyze the polymerization and precipitation of 3,3'-diaminobenzidine (DAB) to enhance the contrast of the mitochondrial matrix and ER lumen(Martell et al., 2012). However, a major limitation of APEX is that it needs to be expressed at a high level for detectable biotinylation and its overexpression would sometimes perturb the biology in question. This low sensitivity has motivated the improvement of APEX. After yeast display screening of 10^6 APEX variants and the directed evolution for more efficient heme incorporation, Lam *et al*. developed APEX2 with one additional mutation that confers improvements in thermal stability, heme uptake, hydrogen peroxide tolerance, and overall higher activity at a low expression level(Hung et al., 2016; Lam et al., 2015). In its first introduction, APEX2 was fused to the calcium uptake regulatory protein MICU1 to resolve its location in the mitochondrial intermembrane space by EM imaging. The superior performance of the second generation enzyme has enabled the profiling of many protein interactomes and subcellular compartment proteomes, such as the mitochondrial nucleoid, autophagosome, ER lumen in mammalian cells(Han et al., 2017; S.-Y. Lee et al., 2016; Le Guerroué et al., 2017).

Despite many impactful applications in spatially-resolved proteomics in mammalian culture systems, APEX approach is limited by the poor membrane permeability of its substrate biotin phenol when compared to biotin ligase-based approaches that use the easily deliverable biotin. Delivery of the biotin phenol probe can be ameliorated by a 30-minute to one hour of preincubation prior to hydrogen peroxide exposure, but still may be a concern for use in the nucleus, tissues, and animals. Furthermore, APEX labeling requires hydrogen peroxide which is highly toxic to living animals, limiting its use to *in vitro* cell cultures and *ex vivo* tissue slices(Hung et al., 2016). APEX2 *ex vivo* labeling has been demonstrated in acutely prepared mouse tissues, such as isolated perfused hearts and 250-300 um thick brain slices(Dumrongprechachan et al., 2021; Hobson et al., 2022; G. Liu et al., 2020).

1.4.3.2.3 PL by TurboID, miniTurbo, and LOV-Turbo - engineered biotin ligases with nontoxic, rapid, *in vivo* labeling

TurboID and miniTurbo were developed with the aim to overcome both the slow labeling kinetics of BioID and high toxicity of APEX2 labeling(Branon et al., 2018). Branon et al. performed directed evolution using the template enzyme *E coli* biotin ligase BirA-R118S mutant, a variant 2-fold more active than BioID, by the generation of a library of 10^7 mutants with error-prone PCR, followed by yeast display of this protein library on the surface for proximity biotinylation, detection of biotinylation by streptavidin and tyramide signal amplification, and fluorescence-activated cell sorting (FACS) to enrich for highest biotinylated cells. Two mutant biotin ligases identified through this evolution effort, TurboID and miniTurbo, biotinylated within 10 minutes at similar levels as 18-hour labeling by BioID. Likely due to its 1.5-2-fold lower activity than TurboID, miniTurbo exhibits less background biotinylation using cellular endogenous biotin, providing a more precise temporal control of PL.

To minimize the ‘leaky’ background labeling by TurboID in biotin-rich environments such as neurons, Lee *et al.* developed light oxygen voltage domain(LOV)-Turbo, which activates

under low power blue light in a reversible manner(S.-Y. Lee et al., 2023). The light-sensory LOV domain, a 16kD flavin-containing protein originally from oat, forms a ‘clamp’ in the dark but releases it in the 470 nm light. Inserting this domain into a surface-exposed loop of TurboID allosterically coupled to its active site could release or distort the active site with or without illumination. Screening of 31 insertion sites identified that strongest light-gating was achieved with insertion between amino acids 80/81, which are connected by a beta-strand to the biotin-binding pocket. Furthermore, with a 4-amino-acid truncation in the loop N-terminal to the insertion, the +/- light ratio improved by two-fold. Finally, similar to APEX2 and TurboID, the prototype enzyme for LOV-Turbo underwent yeast display screening and directed evolution to improve stability/expression level while maintaining minimal spontaneous labeling when handled in a red light-illuminated dark room. In addition to exogenous blue light, LOV-Turbo can also activate via bioluminescence resonance energy transfer, allowing it to labeling the proteome associated with a complex defined by the interaction between two proteins fused with LOV-Turbo and the 460 nm light-emitting luciferase NanoLuc respectively.

The rapid kinetics of the 2nd/3rd generation biotin ligases combined with the nontoxic, membrane-permeable biotin substrate enabled the *in situ* proteomic characterizations *in vivo* , a context especially important for the study of complex neural circuits and synapse biology(Rayaprolu et al., 2022; Takano et al., 2020). Using genetically engineered mouse models, TurboID can be delivered in a cell type-specific manner by crossing the *Rosa26*$^{\text{TurboID}}$ knock-in mice with cell-type specific promoter-driven cre-recombinase mice. TurboID can also be delivered via retro-orbital injections of adeno-associated virus (AAV). The biotin substrate can be delivered via daily subcutaneous injections or orally in drinking water.

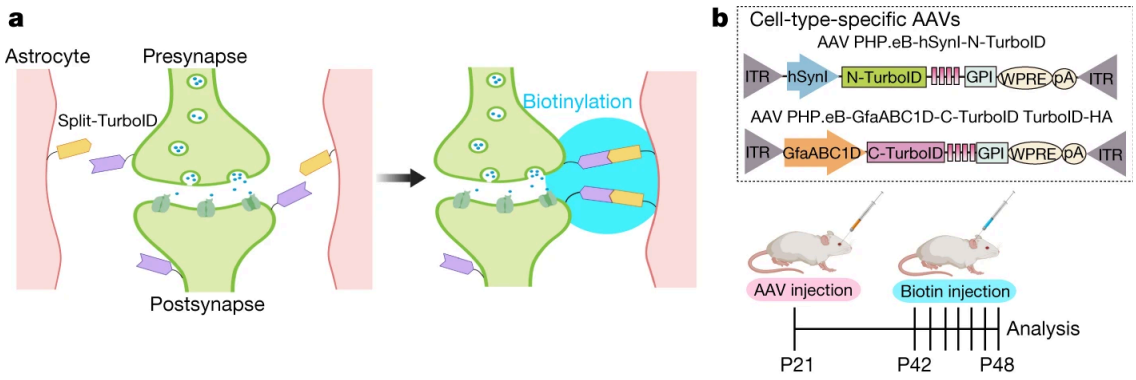


Figure 3. a) Schematic of the Split-surface iBioID approach. b) Outline of Split-TurboID method using cell-type-specific AAVs. ITR, inverted terminal repeats; hSyn1, human synapsin 1 promoter; GPI, glycosylphosphatidylinositol; WPRE, woodchuck hepatitis virus post-transcriptional regulatory element; pA, polyadenylation. Used with permission from Takano et al., 2020.

Takano *et al.* presented a powerful demonstration of *in vivo*, contact-dependent PL application performing perisynaptic proteomic profiling via AAV delivery of N- and C-terminal TurboID fragments driven by cell type-specific promoters to the extracellular surfaces of mouse brain neurons and astrocytes respectively (**Figures 3a, 3b**) (Takano et al., 2020). The two split TurboID fragments rejoined at the astrocyte-neuron junction and biotinylated the local proteome at the synapse, which would otherwise be very challenging to purify and analyze by MS (**Figure 3a**). By comparing protein fractions from the neuron-astrocyte junction and astrocyte surface, normalized with the soluble TurboID control, Takano *et al.* identified a highly confident tripartite synapse proteome consisting of known synaptic proteins such as the channel proteins, excitatory AMPA receptors, and inhibitory $GABA_A$ receptors, among which they focused on NRCAM, a key bridging protein for the adhesion between astrocytes and neurons, and its functions for inhibitory synapse specialization.

1.4.3.2.4 Enzymatic PL applications for chromatin proteins and histone post-translational modifications (PTMs)

The development of engineered biotin ligases provided a large toolbox for the study of local chromatin environments. The fusion of a biotin ligase to a chromatin protein of interest

enables the characterization of the interactomes surrounding chromatin-bound proteins. For example, by fusing BirA* with the type II DNA topoisomerase beta (TOP2B), Uusküla-Reimand *et al*. discovered that TOP2B colocalizes with cohesin and CTCF, suggesting that TOP2 proteins facilitate DNA supercoiling at chromatin loop borders(Uusküla-Reimand et al., 2016). Kochanova *et al*. studied chromocenters, the pericentromeric and centromeric chromatin clusters in *Drosophila* by fusing APEX2 to three centromere-specific proteins, the centromeric-specific variant dCenpA, the hybrid male rescue protein HMR, and heterochromatin protein 1a(Kochanova et al., 2020). In addition to proteins, biotin ligase has also been successfully delivered to histone PTMs that are critical for chromatin biology. Villaseñor *et al*. developed a panel of engineered chromatin readers selective for various methylated histone marks (H3K4me3, H3K9me3, and H3K27me3) and fused them with the biotin ligase BASU(Villaseñor et al., 2020). These constructs were integrated into the mouse genome for stable expression in mouse embryonic stem cells to reveal both the activating and repressing proteins at bivalent sites.

The resulting proteomes from the studies fusing a PL enzyme to a chromatin protein remain protein-centric by virtue of the experimental design. Yet compared to the crosslinking-based ChIP-MS technique, PL enzymes are genetically encodable and thus enables live-cell proteomics that circumvents the drawbacks of crosslinking. Chemical crosslinking with formaldehyde can lead to both cross-links between proteins, as well as cross-links within the same molecule. Furthermore, in addition to modifying lysine side chains, formaldehyde can cause unexpected amino acid modifications at histidine, asparagine, tyrosine, and tryptophan. All of these modifications could lead to a lower sequence coverage than normally expected from a native tryptic digest of the same protein.

1.4.3.3 Photocatalytic proximity labeling and chromatin applications

Enzyme-based PL methods generate phenoxyl radicals or activated AMPs with relatively long half lives of >100 microseconds. With an estimated labeling radius of 300 nanometers (nm), these methods are better suited for characterizing large cellular architectures such as organelles. Besides engineered enzymes, photoactivatable synthetic small molecules provide a new avenue to generate reactive intermediates to label nearby molecules(Fang & Zou, 2023; Knutson et al., 2024). Compared to enzymes, small molecule catalysts are designed to be almost traceless in size so as to minimize the disruption of the native environment. Furthermore, the labeling radius can be modulated using a panel of substrates with increasing half-lives (diazirine < aryl-azide <and phenol probes) and one single photocatalyst(Lin et al., 2024).

Based on the single electron transfer mechanism, light-activated ruthenium derivatives efficiently generate singlet oxygens, followed by triplet oxygens, which are subsequently scavenged by the tyrosine residues of proteins to become phenoxyl radicals. The ruthenium derivative can be directed to important oncogenic receptors by small molecule ligands, and the protein can either be biotinylated or oxidized and inactivated (Sato et al., 2015). Flavin- or xanthene-containing photosensitizers, such as Eosin Y and dibromofluorescein (DBF), can also be excited by visible light to generate singlet-, followed by triplet-oxygen species(Lin et al., 2024; Lynch et al., 2019). For RNA proximity labeling, Engel *et al.* developed Halo-seq, a method that directs DBF as a Halo ligand to nuclear and cytoplasmic Halo-tagged proteins, such that their nearby molecules can be oxidized and alkynylated by propargylamine, a cell-permeable alkyne-containing nucleophile. The alkynylated molecules are efficient click-chemistry substrates and can be linked to an azide-containing tag such as biotin azide, facilitating the purification and high-throughput analysis of the tagged molecules (Engel et al., 2022).

In addition to single electron transfer-based platforms, Geri *et al*. developed a Dexter energy transfer technique named μMap that uses iridium photocatalysts to convert biotin-conjugated diazirines into highly reactive carbenes upon UV light irradiation(Geri et al., 2020). The carbenes readily insert into C-H, N-H, and O-H bonds, biotinylating nearby biomolecules. Compared to singlet-oxygens with half lives of 4 microseconds or more, the reactive carbenes have a half life of 1 nanosecond as they are readily quenched by water. This brief half life limits its diffusion radius to <4 nm and enables a higher resolution mapping of nanoscale protein assemblies beyond organelles or cellular regions. Geri *et al*. demonstrated an immediate application of μMap by conjugating the iridium catalyst to antibodies against cell surface proteins to map the microenvironments on cell membranes.

In addition to labeling extracellular proteomes, iridium photocatalysts can also label chromatin proteins in the nucleus in live cells. Using inteins, which are protein segments that can seamlessly stitch two flanking protein segments into a new protein, the small molecule catalyst can be installed onto chromatin proteins (Seath et al., 2023; Wang et al., 2022). To establish the method, Seath et al. genetically engineered a N-terminal intein to histones H3.1 and centromere protein A, the centromere-specific H3 variant in HEK 293T cells. The C-terminal intein was synthesized via solid-phase peptide synthesis and conjugated to a iridium photocatalyst via click chemistry. The nuclei extracted from the transfected cells were treated with the complementary iridium-conjugated intein to initiate the *in nucleo* protein trans-splicing. After installing the photocatalyst to histones, the nuclei were irradiated in the presence of the substrate biotin-diazirine to initiate proximity labeling (**Figure 4**). This method was applied to map the interactome bound to the oncogenic histone H2A E92K mutant, as well as changes in H2A interactome upon treatment of the bromodomain inhibitor JQ1 and the DOT1-like histone methyltransferase inhibitor pinometostat. In addition to histones, another abundant chromatin resident studied with this technique is RNA Pol II in the presence or absence of the Pol II-stalling small molecule AT7519.

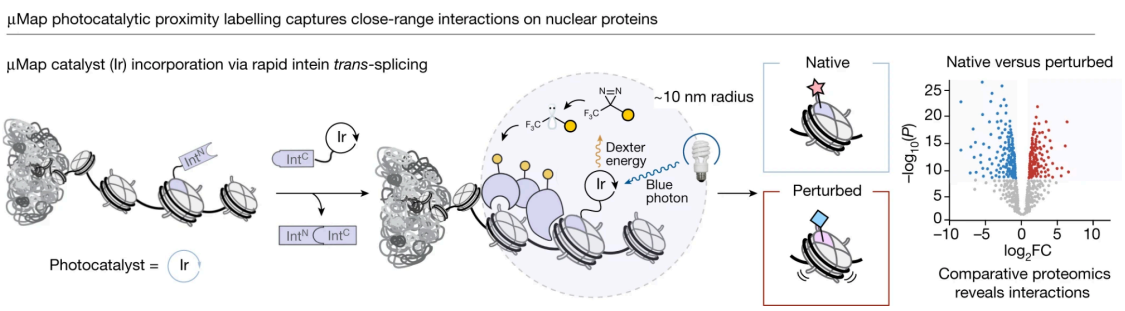


Figure 4. Schematic of the µMap photocatalytic proximity labeling. Used with permission from Seath et al., 2023.

1.4.4 Isolation of chromatin at specific genomic loci

The DNA sequence is the most relevant feature for distinguishing a locus from the genome. Despite intense interest in using the DNA sequence to isolate specific genomic loci, the chromatin biology field has been so far hampered by the lack of methods that are efficient at both sequence recognition and purification((Vermeulen & Déjardin, 2020). The purification of sequence-specific chromatin represents a significant technical challenge and has always been a highly sought-after goal for the field of chromatin biology. The ideal method would need to surmount the combination of several technical difficulties. Firstly, for the unbiased detection of proteins, mass spectrometry platforms are developed later than compared to genomic platforms and proteins/peptides cannot be amplified like DNA to increase analyte amount. Therefore, purification MS studies require much more input biological material to yield eluates at an appropriate quantity and purity for in-depth analysis. Secondly, in comparison to repetitive genomic regions, small, single-copy gene loci are significantly more difficult to purify due to its scarcity. The amount of proteins bound to a small locus is extremely minute compared to those bound to the rest of the genome. The signal from the locus of interest would have to overcome the background noise coming from the rest of the genome. Locus-specific methods begin with an experimenter-defined region of interest. Take a 300 kilobase (kb)-region as an example, the

purification technique would have to achieve a 10,000-fold enrichment in estimate as the chosen region comprises only 0.01% of the human genome. Likewise, studying gene regulatory regions at a 3-kb resolution would require the method to achieve a 1 million-fold enrichment. In part due to the high level of enrichment, and the resulting huge input demand, the locus-specific purification technologies proposed so far have produced proteomic studies on a relatively small scale consisting of only a few samples(Déjardin & Kingston, 2009; Gao et al., 2018; X. Liu et al., 2017; Myers et al., 2018). In addition to the scarcity of a small locus compared to the whole genome, both chromatin structural proteins and gene regulatory proteins are among the proteins truly bound to the locus of interest. Signaling proteins, such as transcription factors, typically bind at much lower frequency than the ubiquitous architecture proteins, making their statistical significant detection even more challenging. The detention limit of current mass spectrometers has been estimated to be 10^{-15} moles. Beyond distinguishing a molecule from noise, the spectrometer's practical quantitation limit for a molecule would be 5 to 10 times higher(Beattie & A H Jones, 2023). Thus, quantifying a singly-bound protein at a unique locus of interest would require 250-500 million diploid cells, assuming zero attrition in the purification and sample preparation process. Owing to these technical challenges to a large extent, no DNA-centric chromatin purification methods have been widely adopted to investigate a given locus.

1.4.4.1 Hybridization-based chromatin capture at specific genomic loci

To address the long standing interest in locus-specific chromatin isolation, the variety of proposed strategies can be summarized into two general categories: nucleic acid-directed and CRISPR-directed. In 2009, Déjardin and Kingston made a notable advancement at locus-specific chromatin isolation by using the unique DNA sequence to discriminate the locus for its purification(Déjardin & Kingston, 2009). The method Proteomics of Isolated Chromatin segments (PICh) uses biotinylated synthetic oligonucleotide probes as the handles for

enrichment of genomic DNA and its associated proteins. Similar to ChIP procedures, PICh begins with formaldehyde fixation of the cells, chromatin solubilization, and lysate pre-clearing (**Figure 5**). Subsequently, the proteinDNA-complexes were briefly denatured to allow hybridization of the biotinylated oligo probes. Specifically, the oligos contained locked nucleic acids by specialized chemical synthesis to increase the melting temperature of the probe. The increase in stability of the oligo-chromatin complexes was critical for maintaining the stringency of the following streptavidin-biotin purification. Déjardin and Kingston validated the methodology using telomeres, a well-characterized test ground for locus-specific chromatin isolation accounting for 0.01-0.07% of the human genome. Telomeres canonically maintained by telomerase and non-canonically maintained by the alternative lengthening of telomeres pathway in two HeLa clones were compared along with scramble probe controls. Using PICh, Déjardin and Kingston revealed 190-210 proteins bound to the two types of telomeres. Among those, 98 proteins were shared between the two. PICh identified 85% of previously identified telomere-associated proteins, validating the technology. Several previously not reported proteins were orthogonally confirmed by immuno-FISH staining.

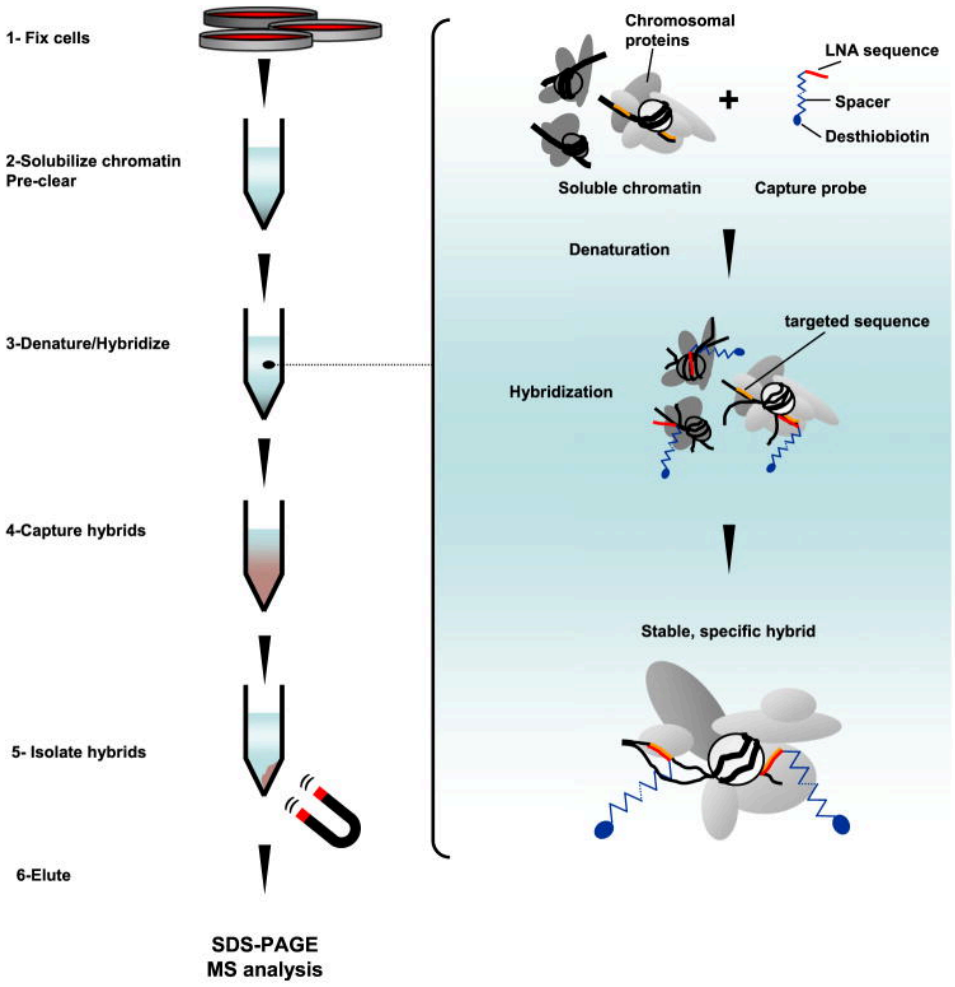


Figure 5. Outline of the PICh Protocol. Used with permission from Déjardin & Kingston, 2009.

Similar to the *in vitro* protein-DNA interaction screen by Mittler et al. in 2009, PICh used biotinylated synthetic DNAs as the enrichment handle. Yet importantly, the DNA-protein complexes in analysis formed endogenously and were captured by fixation in their native context. Studying endogenous protein-DNA complexes significantly increases the number of bonafide hits in comparison to *in vitro* protein-DNA complexing. However, since the biotin handle is on the oligo probe and the oligo-DNA duplex would not sustain stringent biotin purification washes, the chromatin loss in the purification process would have been significant, Therefore, PICh also requires a large cell input to compensate for this loss. Although the HeLa

S3 cells used are hypertriploids with more than 138 telomeres per cell, which significantly reduces the amount of starting material required to obtain a sufficient amount of eluate for analysis, PICh requires 3 x 10^9 cells for one purification of telomeric chromatin. The poor purification efficiency due to the instability of the probe-DNA hybrid limits the use of PICh to repetitive genomic regions(Gauchier et al., 2019; Saksouk et al., 2015; Scelfo et al., 2024).

1.4.4.2 Directing proximity labeling enzymes to specific genomic loci with CRISPR

The combination of DNA-targeting systems and proximity labeling has opened up new avenues for the study of the chromatin composition in a sequence-specific manner. First emerging as a genome-editing platform, the bacteria-derived clustered regularly interspaced short palindromic repeats (CRISPR) system can be repurposed for precise DNA recognition. The CRISPR system only requires a minimal set of two molecules: the Cas9 protein and the engineered small guide RNA (sgRNA). In 2013, Qi et al. engineered a catalytically inactive Cas9 (dCas9) containing two silencing mutations in the nuclease domain, obviating its cleaving activity(Qi et al., 2021). Since then, a variety of strategies have harnessed the RNA-guided sequence recognition of dCas9 to enable the biochemical purification of a specific locus. The general strategy is to endow engineered biotin ligases with locus-specificity using DNA-targeting proteins, followed by biotinylation of the chromatin around the targeted loci and purification-MS analysis(Cenik et al., 2024; Gao et al., 2018; X. Liu et al., 2017; Myers et al., 2018; Qiu et al., 2019; Ugur et al., 2020).

In 2017, Liu *et al*. described the CRISPR affinity purification in situ of regulatory elements (CAPTURE) approach consisting of three essential components: the sgRNA for the locus of interest, the dCas9 fused with a biotin acceptor tandem peptide, and the prototype biotin ligase BirA(X. Liu et al., 2017, X. Liu et al., 2018). Unlike the subsequent CRISPR-proximity labeling methods, only the stably expressed dCas9 protein has a biotin acceptor peptide as a BirA substrate mimic will be biotinylated by the wild type BirA, enabling

streptavidin-purification of dCas9-tethered chromatin. The basic protocol begins with the lentivirus-based establishment and validation of stable, clonal cell lines expressing the dCas9, BirA, and the targeting sgRNAs or non-targeting sgGal4 control (**Figure 6**). The sequence-specificity of the targeting sgRNA is validated by performing Cas9 biotin-ChIP, where the eluate DNA can be evaluated by qPCR or high-throughput sequencing to confirm on-target enrichment. For each purification for proteomic analysis, 250 million to 1 billion validated cells were harvested, fixed in 2% formaldehyde to preserve the dCas9-chromatin association. Liu et al. isolated the nuclei, extracted chromatin with an adaptation of ChEP by Kustatscher et al.(Kustatscher, Wills, et al., 2014), solubilized chromatin with sonication, followed by streptavidin purification and MS detection. For analysis with Western blotting, 100 million cells were cultured. The CAPTURE technology is also equipped with a 3C-sequencing readout to identify long-range DNA interactions. For genomic analysis, 50 million cells were crosslinked and digested with the restriction enzyme DpnII to allow *in situ* proximity ligation of the DNA molecules tethered together by proteins. The ligated chromatin are then sonicated and streptavidin purified for library prep and sequencing. After method validation with telomere isolation,CAPTURE was applied to the locus control region containing the 4 enhancers for the beta-like globin genes: *HS1, HS2, HS3, HS4*, as well as the promoters of beta-globin genes *HBB*, *HBG1,* and *HBG2*.

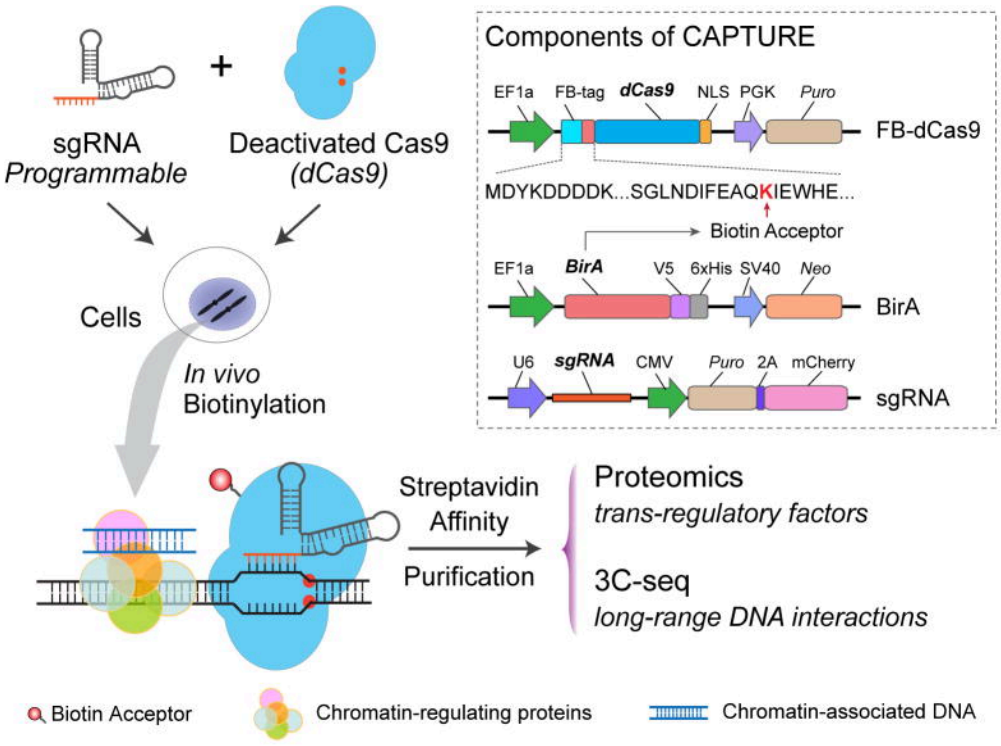


Figure 6. *In Situ* Capture of Locus-Specific Chromatin Interactions by Biotinylated dCas9. Schematic of dCas9-mediated CAPTURE of *cis*-element-associated chromatin interactions. The three components of the CAPTURE system: a FLAG-biotin acceptor-tagged dCas9, a biotin ligase BirA, and a target-specific sgRNA. Used with permission from X. Liu et al., 2018.

The CAPTURE technology introduction exhibited an impressive body of work including gRNA validation by sequencing, comparing on-target enrichment with Cas9 ChIP-seq, the development of quantitative proteomic and 3C readout, and the nuclear isolation and chromatin extraction prior to streptavidin capture which could be helpful for reducing cytoplasmic and nuclear background. However, since the biotin label is only on the dCas9 protein, analogous to ChIP-MS, CAPTURE relies on formaldehyde to retain the chromatin interactions and does not circumvent limitations of chemical crosslinking. Furthermore, each enhancer several hundred bases in length is only sparsely covered by a total of 2-4 sgRNAs, which may explain the huge cell input required for proteomic studies.

Compared to CAPTURE, a significantly more straightforward strategy is to fuse the dCas9 protein to a proximity-labeling enzyme for the deposition of biotin on the nearby

chromatin proteins. Since the nearby proteins are biotinylated *in vivo* prior to cell lysis, this strategy does not rely on chemical fixation and retains proteins in their native state. Several groups have independently proposed methods of this strategy including 1)CasID, a dCas9-BirA* fusion(Ugur et al., 2020), 2)Glo-Pro, a dCas9-APEX fusion (**Figure 7**) (Myers et al., 2018), 3) C-Berst, a dCas9-APEX2 fusion(Gao et al., 2018), 4)CAPLOCUS, also a dCas9-APEX2 fusion(Gao et al., 2018; Myers et al., 2018; Qiu et al., 2019; Ugur et al., 2020), and 5)TurboCas, a dCas9-miniTurbo fusion(Cenik et al., 2024). The development of a dCas9-PL enzyme chromatin isolation technology can be outlined in three stages: 1) the construction of the dCas9-PL enzyme fusion protein in a mammalian expression vector 2) the generation of single-clone cell line stably expressing the fusion enzyme and sgRNAs targeting the locus of interest, as well as characterization for their expression level and on-target enrichment. 3) The large-scale expansion of the characterized cell lines, biotinylation, and harvest in standard cell lysis buffer (Eg. RIPA) for biotin purification and MS prep. Compared to fixation-based methods, these methods enable live-cell proteomics, which holds promise for higher sensitivity and lower cell input requirement for examining small gene elements. For example, Gao et al. reported capturing the telomeric chromatin with C-BERST using only 60 million of the hypertriploid U2-OS cells, a significant decrease compared to the 250 million-3 billion cells reported by PICh and CAPTURE.

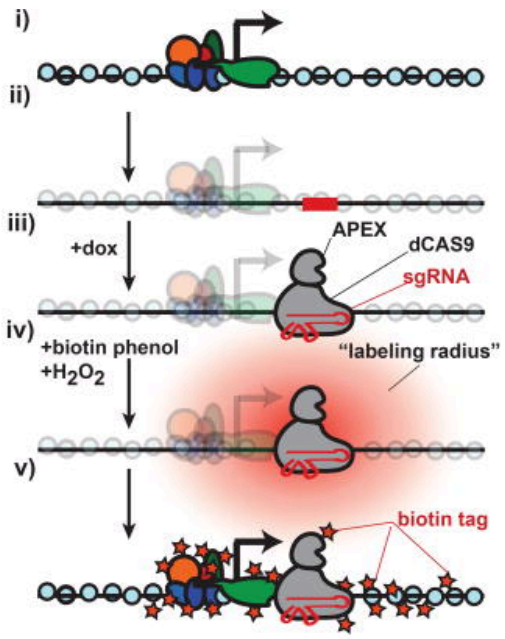


Figure 7. **Illustration of Glo-Pro targeting and affinity labeling reaction. i)** A genomic locus of interest is identified. **ii)** A targeting sequence for the sgRNA is designed (red bar). **iii)** CASPEX expression is induced with doxycycline and, after association with sgRNA, binds region of interest. **iv)** After biotin-phenol incubation, H2O2 induces the CASPEX-mediated labeling of proximal proteins, where the "labeling radius" of the reactive biotin-phenol is represented by the red cloud. **v)** Proteins proximal to CASPEX are labeled with biotin (orange stars) for subsequent enrichment. Used with permission from Myers et al., 2018.

Compared to C-BERST, CAPLOCUS reported even lower cell inputs for readouts including SILAC proteomics, western blotting, and high-throughput sequencing(Qiu et al., 2019). Only two T-75 flasks (estimated to yield <20 million cells in total) of the hypotriploid HEK293T cells were transfected with either the telomere-targeted sgRNA or the non-targeting sgGal4 and later combined for affinity purification and SILAC proteomics, 40 million cells for purification western blotting, and 10 million cells for dCas9 ChIP-seq. However, CAPLOCUS used cells that were only transiently transfected for the subsequent purifications and analyses. For large-scale pull-downs, the expression of the dCas9-PL enzyme protein will need to be similar in all cells within the population and between replicates. Traditional transient transfection is thus ill-advised for more than preliminary functional validation. To avoid repeated large-scale transfections, it will be far more efficient and reproducible to generate a population of cells stably expressing similar

amounts of the dCas9-labeling enzyme fusion protein, followed by the stable introduction of sgRNAs targeting different individual loci.

Since proximity labeling enables live-cell proteomics, this strategy has been explored for the study of much smaller DNA loci such as gene regulatory elements. Using GloPro, Myers et al. attempted to examine *MYC* promoter (500 bases) and the *hTERT* promoter (> 1 kilobases) by generating 5 stable cell lines for each promoter, each expressing one sgRNA respectively to tile across the promoter. They reported using 300 million cells for each cell line and eventually combining the eluates from the 5 cell lines into one MS sample, making the total cell input 1.5 billion. Considering that 5 sgRNA binding sites are only ~ 100 bases, which is thousands-of-fold smaller than telomeres, 1.5 billion constitutes a significant decrease in cell input. However, due to the lack of ground truth knowledge about the chromatin landscape at these promoters, the proteins enriched from this study still call for orthogonal validation study such as ChIP-seq and carefully designed immunofluorescence-FISH, as well as functional validation such as knockdown or knockout of the hit proteins followed by *MYC* transcription. Furthermore, biotin purification followed by high-throughput sequencing would be necessary to assess the genome-wide off-target binding of the dCas9-APEX protein to other locations in addition to the multiple ChIP-qPCRs covering the sgRNA-targeted site reported in the publication.

Similar to dCas9-APEX in GloPro, TurboCas, a dCas9-miniTurbo fusion protein, was directed to the promoter and the first exon of the stress responsive gene *FOS* using three sgRNAs tiled across the 1 kb region. Cenik et al. biotinylated the *FOS* promoter with and without a prior heat-shock treatment to study the local proteome changes needed for heat-shock gene expression. To further validate the method, the *MYC* promoter labeled by GloPro was further studied by TurboCas with 6 sgRNAs. For the TurboCas studies, Cenik et al. performed careful validation of the method and results such as dCas9 ChIP-seq to demonstrate the absence of significant binding outside of the targeted region for each chosen sgRNA, IP-MS analysis of

known proteins involved in the heat shock response, RNA Pol II and cyclin T1, as well as functional validation by shRNA knockdown of selected hit proteins followed by RNA sequencing.

Despite the successful application of GloPro and TurboCas at a few gene promoters, the tiled sgRNA strategy is limited by the challenges accompanying the stable expression of a pool of guide RNAs in the same cells, making it difficult to deliver multiple copies of labeling enzymes to the locus for signal amplification. Both GloPro and TurboCas demonstrated a proof of concept for using a single guide per cell line, and they performed individual pull-downs for each cell line. To efficiently target a non-repetitive region of 1 kb, up to 30 sgRNA may be required, a number that would significantly scale up the pull-down effort. Multiplexed expression of sgRNAs can be achieved by Gibson or Golden Gate assembly of an array of up to 10 sgRNAs linked by *Csy4* cleavage sites, or direct repeats, under a Pol II or Pol III promoter(Kurata et al., 2018; McCarty et al., 2020), with the caveat that internally placed guides will have lower expression levels compared those at the ends. Even when multiplexed sgRNA expression is successful, the sgRNAs would need to compete for a limited pool of dCas9-enzyme fusion proteins, leading to variable targeting efficiencies that are difficult to predict. Furthermore, Cas9:sgRNA complexes can bind genetic loci with as little as 5 nucleotides of homology. These off-target binding effects would scale with the number of simultaneously expressed sgRNA. Coexpression of the *Csy4* endoribonuclease at high concentrations may also lead to cytotoxicity. Due to these intrinsic obstacles of multiplexed CRISPR targeting, the sgRNA tiling for locus-specific chromatin isolation has been very sparse, delivered at the density of a few hundred bases per guide with a simplistic design of a single guide per cell line. Therefore, the sgRNA titling strategy requires a significant amount of cell-line engineering efforts. This requirement puts a limit on its application to easily transfectable systems such as cell lines.

Another limitation of the dCas9-labeling enzyme strategy is related to the bulky fusion protein. With the dCas9 protein at 160 kD and the proximity labeling enzymes at ~28 kD (APEX and miniTurbo), the labeling machinery can be ~188 kD in size. Therefore, excess tiling of

multiple dCas9 fusion proteins may interfere with the local proteome landscape itself. To date, there exists no widely adopted locus-specific chromatin isolation methods that can easily generalize to different loci and biological systems.

1.5. Targeted DNA interaction capture methods

The specific spatial arrangement of higher eukaryotic chromosomes is nonrandom and plays a crucial role in facilitating gene regulation and genome stability. Extensive efforts have been invested into the development of sequencing-based techniques to reconstruct the 3D genome architecture at regional, chromosomal, and whole genome levels.

In 2002, Dekker et al. described the chromosome conformation capture (3C) approach and provided the general strategy that 3C-variant methods rely on(Dekker et al., 2002). 3C begins with a formaldehyde treatment of intact nuclei to crosslink proteins to other proteins and DNA to preserve DNA interactions, restriction enzyme digest to create sticky ends on both DNA fragments bound together by crosslinking, and re-ligation of the sticky ends at a very low DNA concentration to create a chimeric DNA molecule from two fragments in proximity. To detect the interaction frequency of two genomic loci of interest, crosslinking is reversed and individual ligation products are detected with PCR using locus-specific primers and gel electrophoresis. Studying the chromosome organization of the 320-kb yeast chromosome III with 13 primers, Dekker et al. found that, in general, the interaction frequency of two loci decreases as their linear genomic distance increases, except that the two telomeres of the chromosome are closely juxtaposed.

1.5.1 3C-based large-scale detection of DNA interactions

Relying on PCR detection, the 3C approach is limited to examining a pre-defined set of loci with pre-designed primers one at a time. The approach easily becomes laborious if the

genomic region in study is larger than several hundred kilobases. Prior to the emergence of high-throughput sequencing, to enable an unbiased genome-wide search of contacting DNAs anchored to one locus of interest, Simonis et al. designed the one-versus-all 4C technology (3C-on-a-chip), which uses a 30-cycle inverse PCR with primers facing outward to encompass any potential chimeric DNA post ligation(Simonis et al., 2006). The PCR products, potentially containing an interacting ligation partner, are then detected on a microarray with each microarray probe designed to be within 100 bp of the restriction enzyme site. Simonis et al. applied 4C to the mouse β-globin locus control region in mouse embryonic day (E) 14.5 liver, where the β-globin genes are actively expressed, and E14.5 brain, where the locus is inactive, and found that the active locus interacts with other genes also in active expression, whereas the same locus in the inactive state preferentially associates with quiescent regions. Around the same time, similar to the microarray-based 4C, 5C (3C-Carbon Copy) also aimed to achieve a large-scale parallel DNA detection with highly multiplexed ligation-mediated amplification, followed by high-throughput sequencing or microarray detection(Dostie et al., 2006). After the generation of a chimeric DNA library with the 3C method, a collection of 5C primers aimed to examine a group of loci can be added to the library to select for interacting DNA pairs of interest. Specifically, when a pair of 5C primers anneal in proximity on the same strand, the primers are able to ligate. The ligation products can be PCR amplified in one reaction with their universal tails as primer binding sites, and the products can subsequently be detected with high-throughput sequencing or DNA microarray. With a pre-selected group of primers, the many-versus-many 5C technology enables the simultaneous detection of DNA interactions among a group of loci.

As the first genome-wide 3C adaptation, Hi-C enriches for ligated DNAs in proximity by filling the restriction-digested sticky ends with biotinylated nucleotides to allow for the purification of the ligation junctions, followed by high throughput sequencing of the chimeric DNAs. Since Hi-C tests possible pairwise interactions in an all-versus-all manner, the resolution of a Hi-C

interaction map relies heavily on high sequencing depth. The first human Hi-C experiment detected genome-wide DNA contacts at 1-Mb resolution using 8.6 million read pairs(Lieberman-Aiden et al., 2009). For the mouse or human genome, several hundred million read pairs can only achieve a Hi-C map of 100-kb, a resolution insufficient for the study of regulatory elements.

1.5.2 Targeted 3C-based DNA interactions

Compared to generating genome-wide DNA interaction maps at higher resolution, it may be more cost-effective to dedicate sequencing depth only to the viewpoints of interest and their interacting domains so as to locally increase resolution. The many-vs-all Capture-C method combines 3C library generation with targeted biotinylated oligonucleotide capture to enrich the selected loci for sequencing. The Micro Capture-C (MCC) method provides improvements to Capture-C by removing strong detergents to maintain nuclear integrity, using MNase over restriction enzymes to provide higher footprinting resolution, and by using deeper targeted sequencing and new bioinformatic approaches to located ligation junctions(Downes et al., 2022; Hamley et al., 2023).

However, Capture-C and MCC belong to the 3C family of techniques for assaying chromatin topology and therefore are limited by the inherent properties of proximity ligation-based techniques: once the ends of a DNA fragment are ligated, they are consumed so one fragment can only be ligated to one or two neighboring ones, greatly limiting the number of interactions that can be detected. Multi-way DNA interactions only become apparent in the aggregation of signals from a large pool of cells. Ideally, to efficiently detect the multivalent interactions with minimal information loss in samples of low cell numbers, the DNA interactions should ideally be preserved via routes other than ligation.

1.6. Objectives of the dissertation

The primary objective of this dissertation is to address the technical challenges in purifying small, non-repetitive gene loci for a comprehensive description of their locus-bound chromatin factors. Aiming for a method achievable without genetic engineering, I utilize the hybridization of a set of genome-binding oligonucleotides to discriminate the locus, followed by the hybridization of secondary, peroxidase-conjugated oligonucleotides to biotinylate proteins in its proximity. I seek to develop a scalable locus-biotinylation method to yield sufficient cells to benchmark the performance of purification, proteomics, and genomics methods. With the established workflow, I aim to quantify DNA-protein and DNA-DNA interactions at specific gene loci and explore the performance of this method at non-repetitive loci of limited abundance.

To achieve these objectives, I initially developed the locus-biotinylation method for cells fixed on glass with a microscopy end goal. The imaging-based method provided proof-of-concept of the design and formed the basis of quality-control assays of locus-specificity in subsequent development. To accommodate the fold enrichment and the amount of input material required for the successful purification of small, non-repetitive target loci, I adapted the on-glass biotinylation method to a scaled-up, liquid-phase one that can process hundreds of million cells in solution in one sample. I constructed the two-day protocol with two intermediate quality control assays for a small amount of sampled loose cells to assess the quality of hybridization of genome-binding oligonucleotides, the locus-specificity of biotinylation, as well as general nuclear integrity and structural features as context for the target loci. The technology of in-solution locus-specific biotinylation and streptavidin purification for genomics and proteomics discovery is termed “DNA O-MAP”.

The in-solution workflow I developed consistently generates 16-18 locus-biotinylated samples in parallel with tens-of-million cells per sample for streptavidin purification. Using the in-solution workflow, I routinely performed biotinylation of telomeres and single-copy gene loci,

protein extraction, and affinity purification to produce eluate peptides for our collaborators. My purification effort enabled downstream LC-MS optimizations for single-locus chromatin proteomics by Christopher McGann or Devin Schweppe in the Schweppe Lab. I also optimized the streptavidin purification process at 1000-fold enrichment level, followed by tandem mass tag labeling of the digested peptides from the biotinylated proteins. My optimization efforts enabled the integration of quantitative proteomics for the comparison of locus-specific proteomes at multiple target loci.

Using the in-solution workflow, I biotinylated the human alpha-satellite repeats, telomeres, and mitochondrial genome along with control samples omitting primary oligonucleotides. I performed the whole-cell lysate extraction, streptavidin purification, biotinylated protein on-bead digest, and tandem mass tag labeling of the digested peptides. I carried out this experiment in quadruplicates. Christopher McGann in the Schweppe Lab performed mass spectrometry acquisition and data analysis. I demonstrated that DNA O-MAP can differentiate the proteomes of alpha-satellite repeats, telomeres, and mitochondrial DNA with four highly consistent technical replicates.

To quantify DNA-protein and DNA-DNA interactions at specific genomic loci, I coupled the locus-specific chromatin isolation method with next-generation sequencing. For proof of concept, I performed the locus-biotinylation of a pair of interacting DNA loci and a non-looping locus along with control samples omitting primary oligonucleotides. To demonstrate reproducibility, I performed the locus-biotinylation of three chromatin loop anchors in duplicates along with control samples omitting primary oligonucleotides. For all genomics experiments, I performed chromatin solubilization, streptavidin purification, DNA extraction, NGS library preparation, and data analysis. I recovered pairwise and multiway DNA interactions of cohesin-mediated chromatin loops in a ligation-independent manner using 2-5 million biotinylated cells.

Chapter II: DNA O-MAP uncovers the molecular neighborhoods associated with specific genomic loci

Yuzhen Liu[1-3,6], Christopher D. McGann[1,2,5,6], Mary Krebs[1,2], Thomas A. Perkins, Jr.[1], Rose Fields[1], Conor K. Camplisson[1,2], Chris Hsu[1,2], Shayan Avanessian[1-3], Ashley F. Tsue[2,4,5], Evan E. Kania[2,4,5], David M. Shechner[2,4,5], Brian J. Beliveau[1,2,5*], Devin K. Schweppe[1,2,5*]

[1]Department of Genome Sciences, University of Washington, Seattle, WA, USA

[2]Brotman Baty Institute for Precision Medicine, Seattle, WA, USA

[3]Molecular and Cellular Biology Program, University of Washington, Seattle, WA, USA

[4]Department of Pharmacology, University of Washington, Seattle, WA, USA

[5]Institute of Stem Cell and Regenerative Medicine, University of Washington, USA

[6]These authors contributed equally: Yuzhen Liu, Christopher D. McGann.

*Correspondence e-mails: beliveau@uw.edu; dkschwep@uw.edu

2.1 Abstract

The accuracy of crucial nuclear processes such as transcription, replication, and repair, depends on the local composition of chromatin and the regulatory proteins that reside there. Understanding these DNA-protein interactions at the level of specific genomic loci has remained challenging due to technical limitations. Here, we introduce a method termed “DNA O-MAP”, which uses programmable peroxidase-conjugated oligonucleotide probes to biotinylate nearby proteins. We show that DNA O-MAP can be coupled with sample multiplexed quantitative proteomics and next-generation sequencing to quantify DNA-protein and DNA-DNA interactions at specific genomic loci.

2.2 Introduction

Eukaryotic cells store their genetic material in the form of chromatin, a DNA-protein complex. The function of a eukaryotic DNA locus is executed through the cooperation between its nucleotide sequence and the hundreds of protein factors assembled around it. DNA-protein interactions thus play a fundamental role in regulating both the genome's structure and message storing functions(Bickmore & van Steensel, 2013). Therefore, developing methods to decipher DNA-protein interactions in cells has been a focus of technology development efforts for decades(Jerkovic & Cavalli, 2021). For instance, chromatin immunoprecipitation followed by sequencing (ChIP-seq(Johnson et al., 2007)), which has emerged as a core technology for epigenomics(Ho et al., 2012), surveys the genome-wide binding profile of a target DNA-associated protein. ChIP-seq and related technologies (e.g., DamID(van Steensel & Henikoff, 2000), CUT&Tag(Kaya-Okur et al., 2019)) have produced an abundance of high-quality datasets that enabled the establishment of database consortia such as ENCODE(ENCODE Project Consortium, 2004; Luo et al., 2020) and IHEC(Bujold et al., 2016), and significantly accelerated chromatin state annotation efforts(Ernst & Kellis, 2012; Hoffman et al., 2009). Such methods, which profile DNA-protein interactions through a protein-centric lens, require the *a priori* knowledge of which protein(s) to target and rely on the availability of suitable reagents such as antibodies or genetically engineered cell lines. By targeting a single protein at a time, these methods also inherently ignore the context of protein complexes or transient interactions that may be present at a given locus.

In addition to methods that profile the DNA bound by specific proteins, efforts have been dedicated to addressing the inverse problem—identifying the full collection of proteins assembled on a given DNA locus(Gao et al., 2018; Myers et al., 2018; Qiu et al., 2019; Ugur et al., 2020). Such methods include the foundational proteomics of isolated chromatin segment (PICh) technology, which uses a biotinylated oligonucleotide (oligo) probe to affinity label

specific genomic DNA intervals via *in situ* hybridization (ISH)(Déjardin & Kingston, 2009). To enhance the stability of probe-chromatin interactions throughout the purification workflow, PICh utilizes oligos containing locked nucleic acid residues(Silahtaroglu et al., 2003), which are highly efficient as hybridization probes against repetitive DNA targets but cost-prohibitive to use to target non-repetitive intervals that require dozens to hundreds of probes to produce visible signal(Beliveau et al., 2012). As noted in follow-up work, PICh was effective for repeat sequences but would require significant additional work to extend to more complex genomic sequences(Ide & Dejardin, 2015). Additionally, even with the increased stability gained from the use of locked nucleic acid probes, the probe-chromatin hybrids can be difficult to maintain when coupled with stringent purification washes(Ide & Dejardin, 2015), limiting detection sensitivity. As a consequence, an input of one trillion cells was required for one purification and successful identification of proteins interacting with telomeres(Déjardin & Kingston, 2009).

To reach a higher degree of enrichment, which is critical for lower abundance DNA targets, an alternative strategy is to directly biotinylate the proteins that occupy a target DNA locus. This biotinylation can be achieved via targeted proximity labeling using promiscuous biotin ligases(Cho et al., 2020; Roux et al., 2012) or the engineered ascorbate peroxidase (APEX/APEX2) enzymes(Lam et al., 2015; Martell et al., 2012). Since the development of APEX, several methods including C-BERST(Gao et al., 2018) and GLoPro(Myers et al., 2018), have combined APEX with CRISPR genome targeting to endow it with locus specificity. APEX is fused to a catalytically dead RNA-guided nuclease, Cas9 (dCas9) and directing the fusion enzyme to a specific locus of interest by single guide RNAs (sgRNAs). The locus-docked dCas9-APEX biotinylates the neighboring proteins on electrophilic amino acid side chains, such as tyrosine, enabling protein purification and subsequent identification by mass spectrometry (MS). In the case of GLoPro, APEX-based proximity labeling enhanced protein detection sensitivity, reducing the input required for each replicate analysis to ~300 million cells—a 10-fold reduction in cell input compared to PICh which used 3 billion cells. Nevertheless, a notable

limitation of CRISPR-guided proximity labeling is requiring the introduction of the fusion dCas9-APEX enzyme and sgRNAs into a suitable host cell line. Since a successful locus purification canonically requires tens to hundreds of millions of cells, if not more, most current methods aim to create stable cell lines for this purpose. These requirements limit the use of previous locus proteomics methods since efficient and well-tolerated gene delivery remains a major challenge and considerable effort in primary cells(Mangeot et al., 2019). In addition, the labeling reagents necessary for APEX-based proximity labeling—hydrogen peroxide and biotin phenoxyl radicals—are toxic to cells and living organisms, limiting the use of CRISPR-based proximity labeling to cell lines amenable to genetic engineering. Owing to the large numbers of cells required and the need to maximize sensitivity, previous methods often compared only 1–2 biological replicates(Gao et al., 2018; Myers et al., 2018). In some cases this was limited by the use of stable-isotope-based quantification methods that can only multiplex up to three samples per analysis(Gao et al., 2018). Thus, an unmet need exists for methods capable of scaling and extensibly profiling multiple genomic loci. Moreover, these methods would ideally be capable of scaling and multiplexing comparisons between multiple local proteomes or one local proteome in response to multiple stimuli or perturbations.

We address these pressing technical limitations by introducing DNA O-MAP, a locus purification method that uses oligo-based ISH probes to recruit peroxidase activity to specific DNA intervals. DNA O-MAP builds on our previously introduced RNA O-MAP(Tsue et al., 2023) and pSABER(Attar et al., 2023) techniques, which target peroxidase activity to specific RNAs and RNAs/DNA intervals for purification or visualization, respectively. Here, we describe a cost-effective and scalable bulk hybridization and biotinylation workflows capable of processing millions of cells in parallel in just a few days, and demonstrate the recovered material is compatible with sample multiplexed proteomics(Li et al., 2020). We benchmark the specificity of our approach by recovering telomere-specific DNA binding proteins after targeting telomeric DNA. We further showcase the scalability and sample multiplexing capacity of DNA O-MAP by

distinguishing the DNA-associated proteomes around human pericentromeric alpha-satellite repeats, telomeres, and mitochondrial genomes in quadruplicates using tandem mass tags(Li et al., 2020). Finally, we establish that DNA O-MAP can be used to capture functionally relevant DNA-DNA interactions, read out by DNA sequencing, from intervals as small as 20 kilobases. We anticipate that the flexible targeting, scalable protocol, and robust labeling capabilities provided by DNA O-MAP will lead to its adoption as a platform technology for uncovering locus-specific chromatin interactions.

2.3 Results

2.3.1 Design of DNA O-MAP

DNA O-MAP is a molecular profiling methodology that combines the targeting flexibility of oligo-based (ISH) with the ability of horseradish peroxidase (HRP) to catalyze the localized deposition of small biomolecules at sites where it is bound. DNA O-MAP works by recruiting a ‘secondary’ HRP-conjugated oligo to sites where the primary ISH probes are bound. HRP-mediated deposition of biotin at specific genomic sites then enables the pull-down and purification of chromatin associated proteins and DNA from *trans*-interacting genomic loci. As in RNA O-MAP, the specificity of ISH and/or biotinylation can be assessed by microscopy using a small sample of cells immobilized on solid support before the cell pellets enter affinity purification downstream. Importantly, the HRP-conjugated oligo is available via several commercial sources, allowing researchers without the expertise to perform their own conjugations to utilize DNA O-MAP.

2.3.2 DNA O-MAP deploys a scalable in-solution hybridization-biotinylation workflow.

During the development of DNA O-MAP, it became clear that performing *in situ* hybridization on samples adhered to solid substrates such as microscope slides or well plates would create significant scaling challenges, both in terms of reagent costs and sample processing time. We addressed these challenges by developing a suspension-based hybridization workflow for cost-efficient genomic labeling (**Figure 8A**). We began with adherent cells grown on multi-layer flasks, each yielding 90-120 million cells, and subsequently released and fixed (4% PFA) in order to be compatible with DNA ISH. Samples can be processed in parallel, thereby increasing the number of samples that could be handled in parallel by one experimentalist. Critically, this approach reduces reagent costs by ~1,000-fold relative to conventional ISH protocols performed on solid substrates, making the labeling of millions or more cells with oligo-based ISH probe sets, including those targeting non-repetitive DNA, cost-feasible.

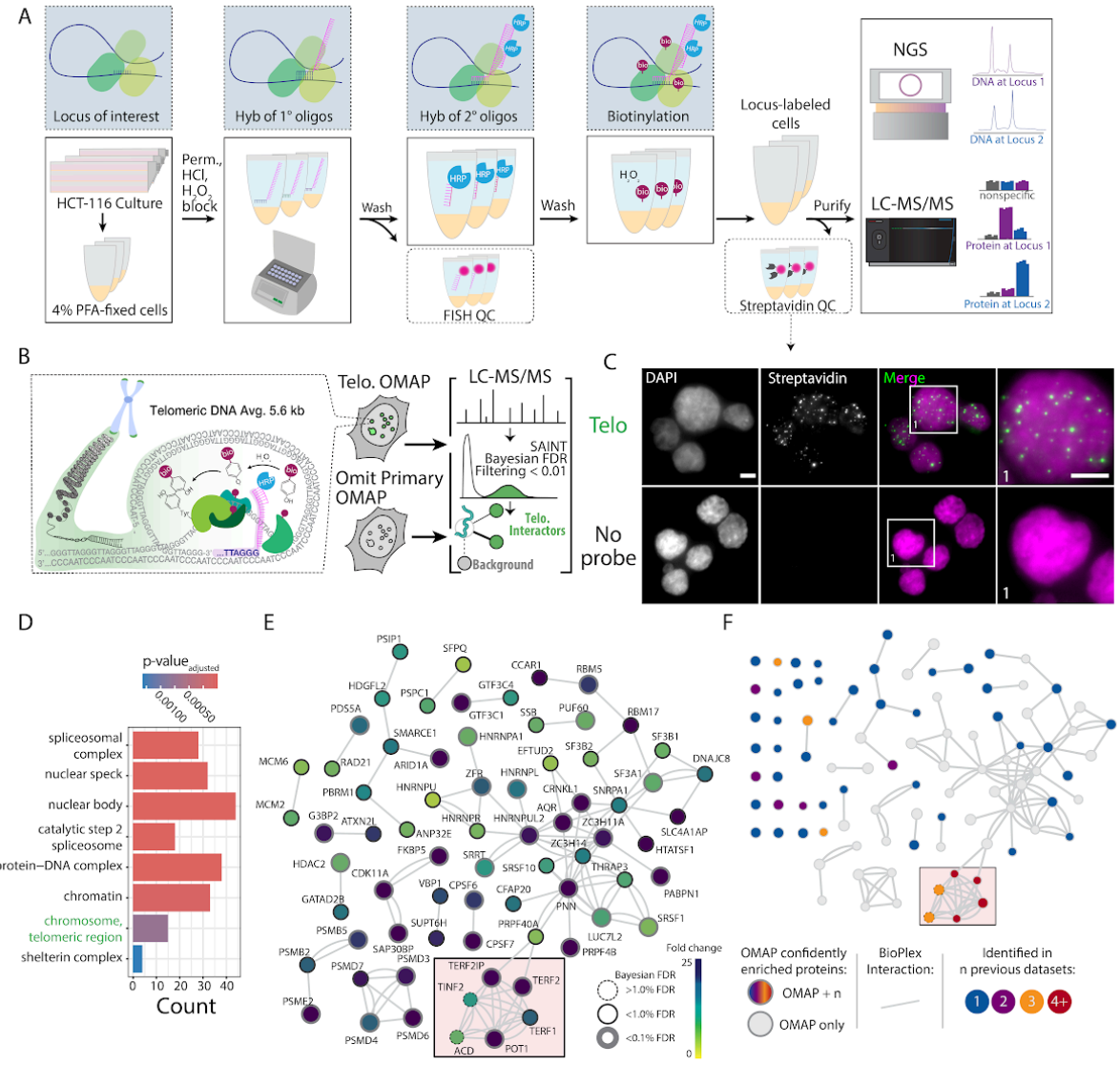


Figure 8. Overview of DNA O-MAP workflow and label-free quantitative proteomics analysis of telomeres. A) Schematic of DNA O-MAP. B) Overview of telomere targeted DNA O-MAP experiment. C) Fluorescent microscopy data showing the observed patterns of DNA (DAPI, left) and *in situ* biotinylation detected by staining with fluorescent streptavidin conjugates (middle, left)l. D) Significant gene sets identified by the Gene Set Enrichment Analysis of the proteins enriched by the telomere probe. E) DNA O-MAP telomeric proteins mapped onto the BioPlex interaction network(Huttlin et al., 2021; Schweppe et al., 2018). The red box highlights shelterin complex proteins. Nodes are colored by the fold-enrichment compared to a no-primary-probe control shown in C, excluding unconnected nodes. F) Telomeric proteins observed in five previous datasets (PICh, C-BERST, CAPLOCUS, CAPTURE, BioID) superimposed onto Figure 8E, colored by the number of prior datasets where the protein was present and including unconnected nodes. Scale bars, 5 μm.

2.3.3 DNA O-MAP reveals the organization of the telomeric proteome.

To demonstrate that O-MAP can successfully purify proteins from small genomic viewpoints, we selected human telomeres for initial testing (**Figure 8B**). Mammalian telomeres are several kilobases of tandemly repeated arrays of 5'-TTAGGG-3' hexamers with terminal 3' single-stranded overhangs at the ends of chromosomes(Chakravarti et al., 2021). Telomeric DNA is specifically bound by a proteinaceous cap that protects the natural chromosome ends from being recognized as damaged DNA—the shelterin complex(de Lange, 2018; Sfeir & de Lange, 2012). Shelterin is a six-subunit complex, which is comprised of the telomeric repeat-binding factor 1 (TERF1), telomeric repeat-binding factor 2 (TERF2), protection of telomeres protein 1 (POT1), adrenocortical dysplasia protein homolog (ACD), TERF2-interacting protein 1 (TERF2IP), and TERF1-interacting nuclear factor 2 (TINF2). Due to the unique telomeric sequence and characteristic DNA structure, the shelterin proteins accumulate exclusively at the ends of the chromosomes. Accordingly, this well-defined set of proteins has been widely accepted as goalposts for a successful locus-specific enrichment experiment(Déjardin & Kingston, 2009; Gao et al., 2018; Myers et al., 2018). In the near-diploid HCT-116 cells, telomeres have an average length of 5.6 kb and their cumulative length approximates 0.017% (~500kb) of the human genome(Myung et al., 2004). Compared to other repetitive elements in the human genome, telomeres are relatively short in HCT-116 cells and thus serve as a rigorous test case for DNA viewpoints of around 500 kb in aggregate across the genome.

We performed a DNA O-MAP experiment in which we either targeted telomeric DNA or omitted the primary hybridization probe (negative control). We purified biotinylated proteins from <60 million cells in three technical replicates followed by imaging of biotinylation and identification of proteins using label-free quantitative proteomics. By streptavidin staining, the punctate fluorescence pattern of biotin-labeled biomolecules closely mimicked telomere FISH,

whereas we did not observe patterns of these puncta in the negative control samples (**Figure 8C**). From our label-free proteomics analysis, we identified 163 proteins as significantly enriched at telomeres. As expected, gene set enrichment analysis(Subramanian et al., 2005) identified significant enrichment of telomeric chromosomal components, chromatin, and protein-DNA complexes (**Figure 8D-E**). Importantly, we identified all six shelterin proteins in the telomere sample and these proteins were completely absent from the control samples. Of the six shelterin proteins, four (TERF1, TERF2, TERF2IP, POT1) passed stringent false-discovery rate control while ACD and TINF2 did not due to low spectral intensity. To benchmark DNA O-MAP, we compared the full set of telomeric proteins to proteins observed in five established telomeric datasets (PICh, C-BERST, CAPLOCUS, CAPTURE, BioID)(Déjardin & Kingston, 2009; Gao et al., 2018; Garcia-Exposito et al., 2016; X. Liu et al., 2017; Qiu et al., 2019) (**Figure 8F**). We then overlaid each called interactor on direct protein interaction data and found that DNA O-MAP enabled greater coverage of known protein interactors, even those not previously identified as enriched at telomeres by other methods. In addition to shelterins, we identified multiple heterogeneous nuclear ribonucleoproteins (hnRNPs) previously annotated as telomere-associated, including HNRNPA1 and HNRNPU. HNRNPA1 has been demonstrated to displace replication protein A (RPA) and directly interact with single-stranded telomeric DNA to regulate telomerase activity(Flynn et al., 2011; LaBranche et al., 1998; Zhang et al., 2006). In addition, HNRNPU belongs to the telomerase-associated proteome(Fu & Collins, 2007) where it binds the telomeric G-quadruplex to prevent RPA from recognizing chromosome ends(Izumi & Funa, 2019). Taken together, this data supports the effectiveness of DNA O-MAP for sensitively and selectively isolating loci-specific proteomes.

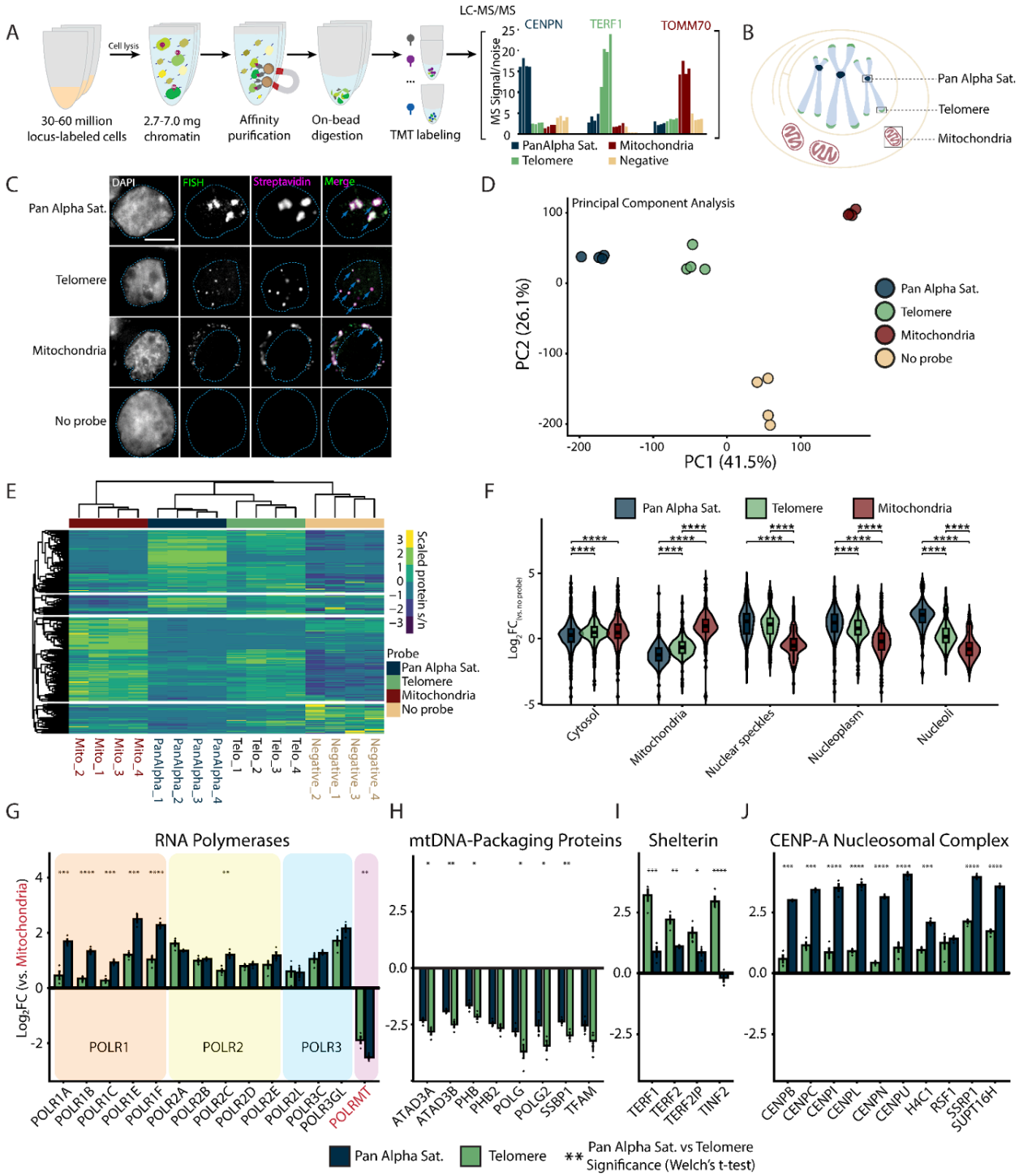


Figure 9. DNA O-MAP reveals distinct features of the sub-proteomes at peri-centromeric alpha satellites, telomeres, and the mitochondrial genome. A) Workflow of DNA O-MAP integrated with sample multiplexing quantitative proteomics B) Schematic of the three DNA loci examined in the TMT16plex experiment: peri-centromeric alpha satellites, telomeres, and mitochondrial genomes. C) Co-localization of DNA FISH and the streptavidin staining of the proteins biotinylated by DNA O-MAP targeting the peri-centromeric alpha satellites, telomeres, and mitochondrial genomes. Scale bar: 5 μm. D) Principal component analysis of scaled intensities of proteins enriched by the pan-alpha probe,

telomere probe, mitochondrial genome oligo pool, and no-primary-probe control. E) Unsupervised hierarchical clustering of scaled intensities of proteins enriched by the pan-alpha probe, telomere probe, mitochondrial genome oligo pool, and no-primary-probe control. F) Log_2 fold change of proteins compared to no-primary-probe control, grouped by HPA subcellular location. Significance calculated based on Welch's t-test for pairwise comparisons (****: p-value <0.0001). G–J) Log_2 fold change of proteins compared to mitochondrial probe enriched proteins for the RNA Polymerases (G), mtDNA nucleoid packaging proteins(Matilainen et al., 2017) (H), Shelterin (I), and CENP-A nucleosomal complexes (J). Significance calculated based on Welch's t-test for pairwise comparisons (p-value: *<0.05, **<0.01, ***<0.001, ****<0.0001).

2.3.4 DNA O-MAP enables multiplexed detection of locus proteomes.

We next evaluated the utility of DNA O-MAP to quantitatively delineate locus-specific proteomes. We integrated sample multiplexing quantitative(Li et al., 2020; Navarrete-Perea et al., 2018; Schweppe et al., 2020) proteomics downstream of DNA O-MAP to enable spectral quantification of all samples simultaneously (**Figure 9A**). In our experimental design, we selected three well-characterized DNA loci with distinct protein occupants in the human genome: 1) telomeres, 2) peri-centromeric alpha satellite repeats; 3) the mitochondrial genome (**Figure 9B**). Centromeres are epigenetically defined chromosomal loci where kinetochore proteins assemble for spindle microtubule attachment to ensure equal chromosome segregation during cell division(McKinley & Cheeseman, 2016; Talbert & Henikoff, 2022). Human centromeres are located within the AT-rich alpha satellite repeats, which are higher-order repeats composed of 171-base-pair monomeric units(Altemose et al., 2022; McNulty & Sullivan, 2018). Due to the sequence independence of centromeres, we utilized a previously described probe(Attar et al., 2023; Deng & Beliveau, 2022) that targets a subset of alpha satellite repeats to represent centromeres, hereafter denoted as the 'Pan Alpha Sat.' probe. The genome-wide binding profile(Aguilar et al., 2024) of the pan-alpha probe closely overlaps with centromeres (**Figure S1**) and covers approximately 35 Mb of the genome according to *in silico* predictions. Mitochondria are intracellular organelles of eukaryotic cells with their own genome (mtDNA). The mtDNA is a circular double-stranded DNA molecule of about 16.6 kb, located in the mitochondrial matrix associated with the inner membrane(Anderson et al., 1981; Rackham & Filipovska, 2022).

To demonstrate the locus-specificity of biotinylation using the new oligo/oligo pools, we performed DNA O-MAP in human HCT-116 cells with a co-hybridization of both fluorescent oligos and HRP oligos in order to observe fluorescent *in situ* hybridization (FISH) and *in situ* biotinylation signals in the same cell. Biotinylation patterns of the pan-alpha, telomere, and

mtDNA probes showed strong concordance with FISH (**Figure 9C**). To quantify the local proteomes corresponding to each of these biotinylated patterns, we prepared replicate (n=4) samples for each probe and control. After *in situ* HRP-mediated labeling, we performed thermal reversal of fixation of cells prior to lysis, enrichment of biotinylated proteins(Paek et al., 2017), tryptic digestion, and labeling with isobaric TMTpro barcodes(Li et al., 2020). We note that artificial lysine alkylation due to cellular fixation with PFA may affect TMTpro labeling of protein, thus we tracked artificial lysine modifications during mass spectrometric analysis to ensure minimal effects of alkylation on protein quantification (1.38% of lysines were alkylated).

In total we quantified 3,055 proteins across all four conditions (**Figure 9D–E**). We observed consistent proteome enrichment by principal component analysis and correlation analyses, with tight clustering of replicates (**Figure 9D–E, S2**). Based on Human Protein Atlas annotations(Thul et al., 2017), we observed significant enrichment of mitochondrial proteins with the mtDNA-probe proteomes and proteins from nuclear locations such as nuclear speckles, nucleoplasm, and nucleoli enriched by the telomere and pan-alpha probes (**Figure 9F, S3**). Notably, the pan-alpha probe enriched proteins from the nucleoli, consistent with the known nucleoli-centromere associations(Bersaglieri et al., 2022); chromosomal passenger complex member AURKB, consistent with the centromeric localization of AURKB in early mitosis to ensure faithful chromosome segregation(Broad et al., 2020; Liang et al., 2020) and the localization of chromosomal passenger complex members to pericentromeric heterochromatin(Ono et al., 2004; Rangasamy et al., 2003). We also observed pericentromeric enrichment of spindle and chromosomal segregation associated proteins TPX2(Kufer et al., 2002) and KIF20A(Khongkow et al., 2016) (**Figure S3, S4**).

Next, we explored the enrichment of several multi-unit protein complexes across the examined loci. To dissect the differences between enriched proteomes for each probe, we chose a subset of proteins of interest and measured the fold change of the two nuclear targets compared to mitochondria. RNA Polymerase I,II,III subunits were all higher in the nuclear

probes than mitochondria, however in contrast to RNA Polymerase II and III, POLR1 proteins are significantly enriched in pan-alpha compared to telomere (**Figure 9G**). This enrichment is likely due to clustering of centromeres around nucleoli(Politz et al., 2013; Rodrigues et al., 2023), the location of ribosomal RNA synthesis by RNA Polymerase I. Conversely, mitochondrial RNA Polymerase POLRMT abundance was significantly lower in the nuclear probe proteomes compared to the mitochondrial probe proteome ($\log_{2\ \text{Pan-Alpha Sat./Mito.}}$= -2.51; $\log_{2\ \text{Telomere/Mito.}}$= -1.88). Similarly, we observed enrichment of mtDNA-packaging nucleoid components(Matilainen et al., 2017) with the mtDNA probes (TFAM, SSBP1, POLG, POLRMT, Lon, ATAD3A/B, and PHB/PHB2; **Figure 9G–H**). As above, we observed consistent enrichment of shelterin components at telomeres (**Figure 9I**). We also observed CENP-A nucleosomal complexes enriched in the pan-alpha proteomes (**Figure 9J**). Histones were enriched with our nuclear probes and a subset (H2A1C, H2AX, and H4C1) were significantly enriched by the pan-alpha probe compared to the telomere probe (**Figure S4**). We also observed enrichment of catenins CTNNB1 and CTNND1 at telomeres (**Figure S3**). The transcription factor CTNNB1 has been observed at the transcriptional start site of h*TERT* where it regulates h*TERT* expression(Hoffmeyer et al., 2012). The h*TERT* gene is located in the subtelomeric region of chromosome 5 (chr5:1,253,167-1,295,068) and expressed in HCT-116 cells(Tsherniak et al., 2017). Collectively, these results demonstrate the sensitivity and subcompartment specificity of DNA O-MAP and highlight how coupling quantitative proteomics with DNA O-MAP can distinguish differential compartment components even for ubiquitous chromatin constituents like histones.

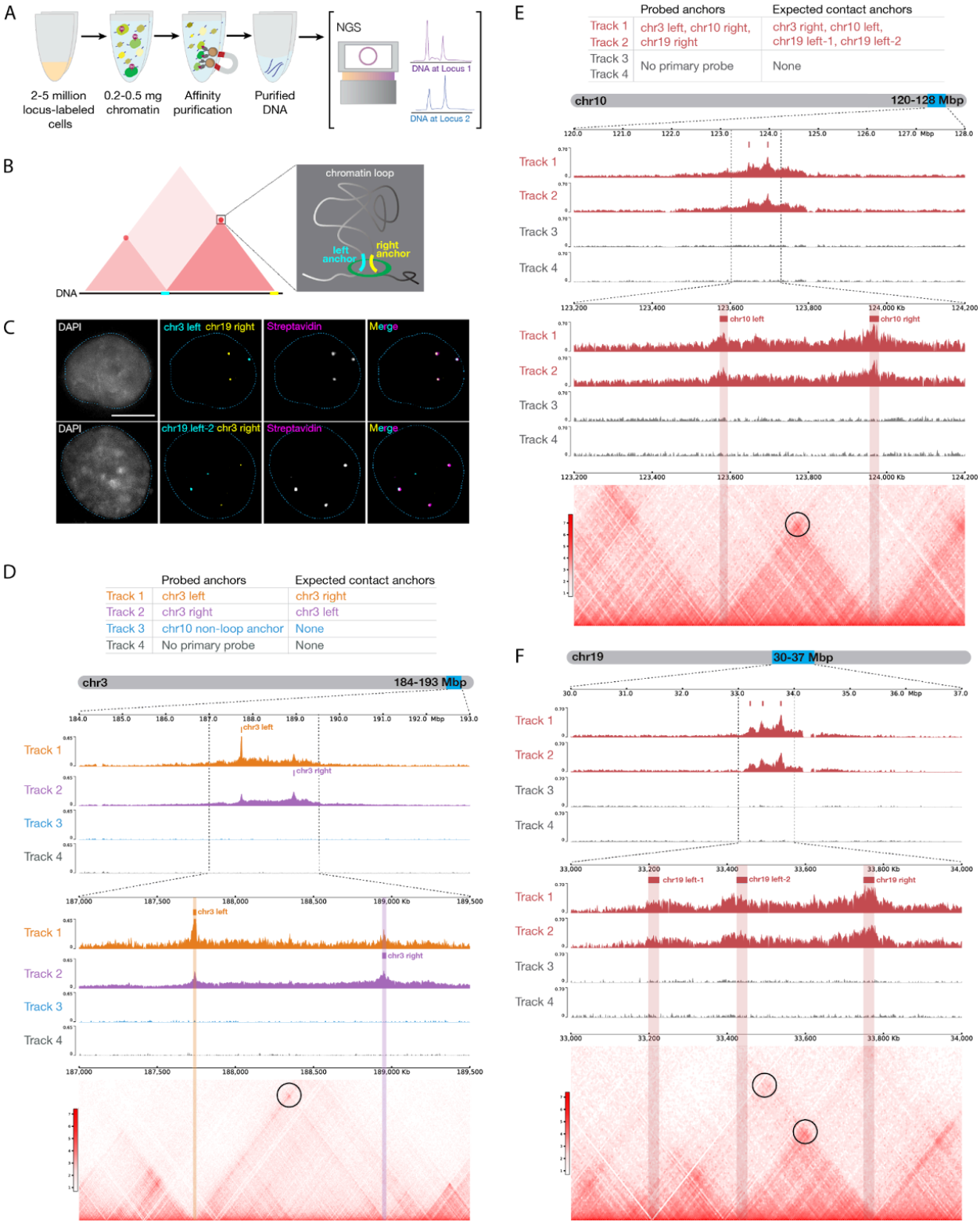


Figure 10. DNA O-MAP efficiently labels single-copy chromatin loop anchors. A) Workflow of DNA O-MAP integrated with biotin purification sequencing B) Schematic of a pair of chromatin loop anchors on

a hypothetical Hi-C map and 3-dimensional space C) DNA FISH and the streptavidin staining of the proteins biotinylated by DNA O-MAP targeting anchors of chromatin loops on chromosome 3 and chromosome 19 D) Table listing the three anchors (Track 1-3) and no-primary-probe control (Track 4) biotinylated by DNA O-MAP and their expected anchors in contact in each track (top). Desthiobiotin purification sequencing signals across the 9-Mb region on chromosome 3 corresponding to the chr3 chromatin loop (middle). Desthiobiotin purification sequencing signals and pairwise contact map at 5-kb resolution across the 2.5-Mb region on chromosome 3 corresponding to the chr3 chromatin loop. Black circle on the contact map indicates the presence of a loop. (bottom). E) Table listing the three chromatin loop anchors (Track 1-2) and no-primary-probe controls (Track 3-4) biotinylated by DNA O-MAP in duplicates and their expected anchors in contact in each track (top). Desthiobiotin purification sequencing signals across the 8-Mb region on chromosome 10 corresponding to the chr10 chromatin loop targeted (middle). Desthiobiotin purification sequencing signals and pairwise contact map at 5-kb resolution across the 1-Mb region on chromosome 10 corresponding to the chr10 chromatin loop. Black circle on the contact map indicates the presence of a loop. (bottom). F) Desthiobiotin purification sequencing signals across the 7-Mb region on chromosome 19 corresponding to the chr19 chromatin loops targeted (top). Desthiobiotin purification sequencing signals and pairwise contact map at 5-kb resolution across the 1-Mb region on chromosome 19 corresponding to the chr19 chromatin loops. Black circles on the contact map indicate the presence of loops (bottom).

2.3.5 DNA O-MAP can uncover DNA-DNA interactions from non-repetitive DNA loci.

Beyond repetitive regions in the human genome, we explored whether DNA O-MAP can recover material from small, single-copy DNA intervals. To this end, we designed an experiment in which we performed *in situ* biotinylation followed by chromatin extraction, affinity purification, and sequencing (**Figure 10A**). The human genome is folded into thousands of chromatin loops where two loci on the same chromosome are tethered to each other (**Figure 10B**). The anchors of the loops are bound by the insulator protein CTCF. The ring-shaped cohesin protein complex is thought to to often stall at CTCF-bound sites while dynamically moving along the genome, creating contact domains of preferential DNA-DNA interaction(Rowley & Corces, 2018). In HCT-116 cells, these contacts between chromatin loop anchors have been captured genome-wide with *in situ* Hi-C(Rao et al., 2014). Normally present in two copies per genome, these 20–25 kb loop anchor intervals are considerably less abundant than telomeres.

We first evaluated whether DNA O-MAP can specifically biotinylate loop anchors with microscopy by a co-hybridization of both fluorescent oligos and HRP oligos at four anchors: chr3 left (chr3:187,729,712-187,749,712), chr3 right (chr3:188,939,711-188,964,711), chr19 left-2 (chr19:33,425,000-33,450,000), and chr19 right (chr19:33,750,000-33,775,000). DNA O-MAP specifically biotinylated the biomolecules proximal to these small DNA intervals, as observed in the co-localizing patterns of FISH and streptavidin staining in the same cells (**Figure 10C**). We next evaluated whether DNA O-MAP could recover the DNA interactions originally discovered by Hi-C. We targeted a pair of intervals with high contact frequency—chr3 left and chr3 right anchors, one non-looping interval (chr10:123,187,984-123,207,984), and no-primary-probe control. We performed DNA O-MAP to biotinylate these DNA intervals, subjected the labeled cells to chromatin solubilization and desthiobiotin purification, and sequenced the eluate DNA. As expected, all three probed DNA intervals were highly enriched

compared with other genomic regions, indicating efficient purification of the loci (**Figures 10D, S5A**). Furthermore, chr3 left and chr3 right anchors reciprocally recovered each other, indicating that DNA O-MAP was able to recover known DNA interactions mediated by proteins. In contrast, the non-looping chr10 anchor did not enrich any other peak except itself (**Figure S5B**). Lastly, in the cells that received no primary oligos, no pronounced enrichment was observed genome wide (**Figure S5B**).

To examine the multiplexability and reproducibility of DNA O-MAP, we simultaneously targeted three chromatin loop anchors: chr3 left, chr10 right (chr10:123,957,984-123,977,984), and chr19 right anchors in duplicates and subjected the cell pellets to purification and DNA sequencing. All three targeted anchors, chr3 left, chr10 right, and chr19 right anchors were successfully enriched (**Figures 10E–F, S6A**), whereas no pronounced enrichment was observed in the no-primary-probe controls genome-wide (**Figure S6B**). Furthermore, chr10 left (contacting chr10 right), chr19 left-1, and chr19 left-2 (both contacting chr19 right) were also efficiently recovered, accurately matching the Hi-C contact maps and the signals from two replicates was consistent (**Figure 10E–F**). These imaging and genomics data demonstrate that DNA O-MAP is capable of labeling small, single-copy DNA intervals with high specificity.

2.4 Discussion

By combining the versatility of hybridization-based genome targeting with robustness of proximity biotinylation, DNA O-MAP offers a scalable approach to study DNA-associated proteomes through a locus specific lens. The liquid-phase hybridization-biotinylation workflow allows for efficient processing of samples and is compatible with both proteomic and genomic readouts. Integration with multiplexed quantitative proteomics enables simultaneous analysis of multiple loci or conditions, increasing data completeness and throughput. Label-free analysis of the telomeres shows strong concordance of labeling with in-situ hybridization and recapitulates previous similar proteomic datasets. Our tri-locus experiment was able to differentiate proteins

with a quantitative profile suggesting general nuclear location from those specifically associated with telomeres and peri-centromeres. DNA O-MAP's ability to target single-copy loci, as evidenced by the chromatin loop anchor experiments, opens up possibilities for studying protein-mediated DNA interactions at a finer resolution than previously possible.

O-MAP has now been shown to be a highly flexible technology for the exploration of biomolecular interactions with RNAs(Tsue et al., 2023) and DNA loci. Using oligos to target the DNA locus, DNA O-MAP can be theoretically adapted for use in any sample types amenable to *in situ* hybridization, including cultured cells, tissue sections, and primary tissue samples(Aguilar et al., 2024; Attar et al., 2023; Hershberg et al., 2021). As the purification tag is decoupled from the probe oligos, labeled chromatin fragments can undergo stringent washes to achieve efficient purification with minimal background. Moreover, without the need to genetically modify the biological system at hand, the probes in this dataset alone could be used to explore telomeric remodeling in cancer cells(Garcia-Exposito et al., 2016), spindle-associated proteome dynamics at the pericentromere(Santos-Barriopedro et al., 2021), and molecular drivers of hetero- or euchromatin formation(Iglesias et al., 2020) at nearly any locus in the human genome (O-MAP probes can feasibly cover >99% of the human genome)(Aguilar et al., 2024; Hershberg et al., 2021).

While this work has laid the foundation for generalized and extensible locus proteomics, further work will be required to achieve the sensitivity required for small, single copy locus proteomics. By taking a comparative quantitative approach, we remove the need to pre-define the local context of probe localization but experimental design is critical and novel interactors likely need further validation to confirm their co-localization at a given locus (e.g., with imaging/FISH). With developments in automation and instrument sensitivity, DNA O-MAP has the potential to expand to locus specific post-translational modifications and be used for large-scale chromatin perturbation screens. We anticipate that DNA-OMAP will have broad

utility for research questions seeking to understand the intricate relationships between DNA sequence, chromatin structure, and cellular function.

2.5 Methods

2.5.1 Cell culture and fixation

Colorectal cancer HCT-116 cells were grown in ATCC-formulated McCoy's 5A Medium Modified (ATCC 30-2007) supplemented with 10% fetal bovine serum and 100 U/ml Penicillin-Streptomycin at 37°C in a humidified atmosphere of 5% CO_2. For each purification, 20 million HCT-116 cells were seeded into one T-500 flask (Thermo Scientific 132867) to culture for 36-48 hours to reach 90–120 million cells. Before collection, cells were briefly rinsed once with Dulbecco's phosphate buffered saline (DPBS) and then incubated with 25 ml of TrypLE Express Enzyme (Gibco 12604-021) at 37°C for two minutes or until loosely attached. The cell suspension was collected into two 50 ml conical tubes and the T-500 flask was rinsed with DPBS. The wash was combined with the cell suspension and centrifuged at 300 G for 5 minutes. After a DPBS wash to remove remaining TrypLE, cells were fixed in 4% paraformaldehyde (wt/vol) (Electron Microscopy Sciences 15710) in phosphate buffered saline (PBS) in suspension at room temperature for 10 minutes with rotation, followed by 125 mM Glycine quenching for 5 minutes at room temperature with rotation and 15 minutes on ice. Fixed cells were collected by centrifugation at 350G for 5 minutes, and stored in fresh DPBS at 4°C until liquid-phase hybridization. Fixed cells were used within 3-5 days.

2.5.2 Primary oligo probes

Primary oligos targeting the human alpha satellite repeat and telomere were purchased as individually column-synthesized DNA oligos from Integrated DNA Technologies. Probe sets targeting mtDNA (chrM:1-16,569), chr3 left anchor (chr3:187,729,712-187,749,712), chr3 right

anchor (chr3:188,939,711-188,964,711), chr10 non-looping anchor (chr10:123,187,984-123,207,984), chr10 right anchor (chr10:123,957,984-123,977,984), and chr19 right anchor (chr19:33,750,000-33,775,000) were designed using PaintSHOP(Hershberg et al., 2021) and ordered in oPool format from Integrated DNA Technologies. More than 300 primary oligos were designed to cover each single-copy DNA interval to ensure a sufficient number of probes at the locus for FISH. The sequences of the oligo and oligo pools used are listed in Supplementary Dataset 1.

2.5.3 Primer exchange reaction (PER)

To extend primary oligos with PER concatemers, reactions were set up as previously described(Kishi et al., 2018) in 100 ul-volume containing 10 mM $MgSO_4$, 300 uM dATP/dCTP/dTTP mix, 100 nM Clean.G hairpin, 80 U/ml Bst DNA Polymerase, Large Fragment (NEB M0275L), 1 uM hairpin, and 1 uM primary oligos in PBS. To verify the length of primary oligos, the reactions were assessed with denaturing polyacrylamide gel electrophoresis. Primary oligos extended to 300-500 nucleotides were used in hybridizations downstream. Unpurified reactions were dehydrated using vacuum concentrators and stored dry at -20°C until hybridization.

2.5.4 In-solution hybridization and biotinylation of cell pellets

Oligo hybridizations were performed on cells in solution for the cost-effectiveness of primary and secondary oligos. Fixed cells were split into 6e7 cell aliquots in 1.5 ml microcentrifuge tubes. All washes and buffer exchanges were performed as follows: centrifuging at 350G for 3.5 minutes or until pelleted, pouring away used buffers from the pellets, adding new buffers, and gentle shaking or low speed vortexing to dislodge cell pellets into tiny clusters or cell suspensions for incubations or washes. Cells in fresh wash buffer were rotated on a low speed nutator for 5 minutes.

Cells were rinsed once with fresh PBS, and permeabilized in PBS-0.5% TritonX-100 (Sigma T8787) for 10 minutes with nutation. After a PBS-0.1% Tween20 (PBS-T) (Sigma T2287) wash, permeabilized cells were incubated in 0.1 N hydrochloric acid (HCl) for 5 minutes. After a PBS-T wash to remove acid, cells were incubated in PBS-T-0.5% hydrogen peroxide to block endogenous peroxidases. After a 2X saline sodium citrate-0.1% Tween20 (2X SSC-T) wash to remove acid, cells were incubated in 2X SSC-T-50% formamide for 20 minutes at 60°C on a Thermomixer C dry block (Eppendorf 2231001005). Cells were exchanged into primary hybridization buffer (Hyb1) comprising 2X SSC-T, 50% (vol/vol) formamide, 10% (wt/vol) dextran sulfate, 0.4 µg/ul RNAse A, and ~1 µM extended primary oligos (resuspended dry, unpurified PER reactions). The cell-Hyb1 mixture was distributed into PCR strip tubes at 1e7-1.5e7 cells in 100 µL volumes. The cells were denatured and primary oligos were hybridized to the genome in the PCR strip tubes in a thermocycler using the cycling protocol: 78°C 3 minutes, 37°C ∞ incubating overnight for more than 18 hours.

The next day, cells were rinsed with 60°C 2X SSC-T into 1.5 ml microcentrifuge tubes, followed by two 2X SSC-T buffer exchanges to remove residual Hyb1. Cell pellets were then washed in 1 ml 2X SSC-T at 60°C, followed by two two-minute washes in 2X SSC-T at room temperature. Fully washed cell pellets were exchanged into 1 ml PBS, and then exchanged into 100 nM secondary HRP oligo that map to the PER concatemer sequence on the primary oligo (custom synthesis by Integrated DNA Technologies or Bio-Synthesis Inc) in PBS. Secondary hybridization was performed at 37°C with nutation for one hour. Cell pellets underwent three 5-minute washes in 1 ml PBS-T at 37°C with nutation. Fully washed cells were incubated in 5 uM desthiobiotin tyramide (Iris Biotech LS-1660) and 1 mM hydrogen peroxide in PBS-T for 5 minutes at room temperature with nutation. To quench the HRP activity, biotinylated cells were washed twice in 10 mM sodium ascorbate and 10 mM sodium azide in PBS-T for 5 minutes at room temperature with nutation. Quenched cells were washed with PBS to remove residual

sodium azide. After sampling cells for quality control, the cell pellets were stored dry in -80°C until chromatin solubilization and affinity purification.

2.5.5 Microscopy-based quality control assays for hybridization and biotinylation

We routinely sample cells along the workflow of preparing AP-MS or NGS samples to monitor the locus specificity of primary oligo hybridization. To assess the quality of primary oligo hybridization, we sampled roughly 5% of fully washed cells from primary hybridization to a new 1.5 ml tube. Cells were incubated with 400 nM fluorescent oligos in PBS at 37°C for an hour with nutation. Hybridized cells underwent three washes in 1 ml PBS-T at 37°C with nutation to remove unbound fluorescent oligos. Washed cells were immobilized on glass slides with Slowfade Gold Antifade Mountant with DAPI (Thermo Fisher S36938) and coverslips for confocal imaging of FISH signal.

We assessed the quality of biotinylation specificity for all samples entering the proteomics or genomics workflow. Roughly 5% of fully quenched cells were sampled into a new 1.5 ml tube and incubated with 0.5-1 μg/ml Alexa Fluor 647-streptavidin (Thermo Fisher S32357) in PBS-T, 1% bovine serum albumin at 37°C for 30 minutes with nutation. Stained cells underwent four washes in 1 ml PBS-T at 37°C with nutation to remove unbound Alexa Fluor 647-streptavidin conjugate. Washed cells were immobilized on glass slides with Slowfade Gold Antifade Mountant with DAPI and coverslips for confocal imaging of Alexa-Fluor 647-streptavidin signals.

2.5.6 Confocal microscopy

Confocal imaging was performed using a Yokogawa CSU-W1 SoRa spinning disc confocal device attached to a Nikon ECLIPSE Ti2 microscope. Excitation light was emitted at 30% of maximal intensity from 405 nm, 488 nm, 561 nm, or 640 nm lasers housed inside of a Nikon LU-NF laser unit. Laser excitation was delivered via a single-mode optical fiber into the

CSU-W1 SoRa unit. Excitation light was directed through a microlens array disk and a SoRa spinning disk containing 50 um pinholes to the rear aperture of a 100x N.A. 1.49 Apo TIRF oil immersion objective lens by a prism in the base of Ti2. Emission light was collected by the same objective and directed by a prism in the base of Ti2 back into the SoRA unit, where it was relayed by a 1x lens (conventional imaging) or 2.8x lens (super-resolution imaging) through the pinhole disk and then directed to the emission path by a quad-band dichroic mirror (Semrock Di01-T405/488/568/647-13X15X0.5). Emission light was then spectrally filtered by one of four single-bandpass filters (DAPI:Chroma ET455/50M; ATTO488: Chroma ET525/36M; ATTO565:Chroma ET605/50M; Alexa Fluor 647: Chroma ET705/72M) and focused by a 1x relay lens onto an Andor Sona 4.2B-11 camera with a physical pixel size of 11 um, resulting in an effective resolution of 110 nm (conventional), or 39.3 nm (super-resolution). The Sona was operated in 16-bit mode with rolling shutter readout and exposure times of 70-300 ms.

2.5.7 FISH-biotinylation co-localization experiment

Fixed cells were split into 5e6 cell aliquots in 1.5 ml microcentrifuge tubes. Primary hybridization and washes were performed similarly to described in the in-solution hybridization and biotinylation of cell pellets with fewer cells. Fully washed cell pellets were exchanged into a secondary co-hybridization buffer containing 30 nM of fluorescent oligos and 100 nM of HRP-oligos in PBS, instead of solely HRP-oligos, for simultaneous hybridization of both species. After washes and biotinylation, the pellets were stained with 0.5-1 µg/ml Alexa-Fluor 647-streptavidin. Cells were immobilized on glass slides with Slowfade Gold Antifade Mountant with DAPI and coverslips for confocal imaging of both FISH and Alexa-Fluor 647-streptavidin signals.

2.5.8 Affinity Purification and sample preparation for proteomics

Biotinylated cell pellets were removed from -80°C to thaw at room temperature. Each cell pellet was resuspended in roughly 0.9 ml of lysis buffer consisting of 1% SDS and 200 mM EPPS with protease inhibitors (Roche 11836170001). The cell mixture was boiled at 95°C for 30 minutes. The boiled cell mixture was sonicated at 4°C using a Covaris LE-220 focused ultrasonicator with the following protocol: 300W peak incident power, 50% duty factor, 200 cycles per burst, with a treatment time of 420 seconds in 1-ml milliTUBEs with AFA fiber (Covaris 520135). The sonicated cell mixture was boiled for a second time at 95°C for 30 minutes. The boiled lysates were cleared by centrifuging at 21130 G for 30 minutes in an Eppendorf 5424 Microcentrifuge at room temperature. The supernatants were transferred to a fresh 1.5-ml tube. To prevent any remnants of cell debris, the supernatants were cleared for a second time by centrifuging at 21130 G for 30 minutes and the supernatants were transferred to a fresh 1.5-ml tube. The supernatants were stored in -80°C until protein quantification.

The cleared cell lysates were quantified using the Pierce BCA Protein Assay Kit (Thermo Fisher 23225). Pierce Streptavidin Magnetic Beads (Thermo Fisher 88817) were washed using 1% SDS, 200 mM EPPS lysis buffer three times before use. From each labeled cell pellet, 2.17 milligrams of protein was used to couple with 500 µg of streptavidin beads in a Protein Lo-Bind tube (Eppendorf EP022431081). The lysates were incubated with the bead slurry for one hour at room temperature with nutation allowing biotinylated proteins to bind. The coupled beads were collected and separated from the flow-through using a magnetic rack (Sergi Lab Supplies 1005a). After the flow-through was removed, the beads underwent the following washes: 2% SDS with 20 mM EPPS twice, 0.1 M Na2CO3, 2 M urea, and 1 M KCl with 20 mM EPPS twice. All washes were performed as follows: after immobilizing the beads on a magnetic rack for 5 minutes, the supernatant was removed and the beads were resuspended in the new wash buffer and incubated for 5 minutes with nutation. Finally, the beads were rinsed once with 20

mM EPPS to remove the excess salt.

The washed streptavidin beads were resuspended in 50 µl of 5 mM TCEP, 200 mM EPPS, pH 8.5 for a 20-minute on-bead protein reduction. The proteins were alkylated on-bead using 10 mM iodoacetamide for one hour in the dark. Then DTT was added to the final concentration of 5 mM to quench the alkylation for 15 minutes. The beads were rinsed twice with 200 mM EPPS for on-bead digest. Assuming 20 µg of eluate protein, 200 ng LysC (Wako) was added to the beads in a 50-ul volume and incubated for 16 hours with vortexing. The next day, 200 ng of trypsin (Promega V5113) was added to the beads and incubated for six hours at 37°C at 200 rpm. After digestion, the peptide-containing supernatant was collected in a fresh 0.5-ml Protein Lo-Bind tube. The beads were rinsed once with 100 µl 50% acetonitrile, 5% formic acid and the wash was combined with the peptides. Peptides were desalted via the stop and go extraction (StageTip)(Rappsilber et al., 2003) method and dried in a vacuum concentrator.

For label free telomere-enriched samples, one sample consisted of HCT-116-Rad21-mAID cells(Natsume et al., 2016) . For samples intended to be multiplexed, dried, desalted peptides were reconstituted in 4 µl of 200 mM EPPS, pH 8.5. The peptides were labeled using 25 µg of TMTpro 16plex Label Reagents (Thermo Fisher A44520) at 33.3% acetonitrile for one hour at room temperature. The labeling reaction was quenched with the addition of 1 µl of 5% hydroxylamine and incubated at room temperature for 15 minutes. The pooled sample was acidified using formic acid and peptides were desalted using a StageTip cartridge. Peptides were eluted in 70% acetonitrile, 1% formic acid and dried by vacuum centrifugation

2.5.9 Mass Spectrometry Data Acquisition Methods and Analysis

Samples were resuspended in 5% acetonitrile/2% formic acid prior to being loaded onto an in-house pulled C18 (Thermo Accucore, 2.6 Å, 150 µm) 30 cm column. Peptides were eluted

over 180 min gradients running from 96% Buffer A (5% acetonitrile, 0.125% formic acid) and 4% buffer B (95% acetonitrile, 0.125% formic acid) to 30% buffer B. Sample eluate was electrosprayed (2700 V) into a Thermo Scientific Orbitrap Eclipse mass spectrometer for analysis. High field asymmetric waveform ion mobility spectrometry (FAIMS) was set at “standard” resolution, 4.6 L/min gas flow, and 3 CVs: −40/–60/–80 were used. MS1 scans were conducted at 120,000 resolving power with a 50 ms max injection time, and the AGC target set to 100%. Peaks from the MS1 scans were filtered by intensity (minimum intensity >5 × 103), charge state (2 ≤ z ≤ 6), and detection of a monoisotopic mass (monoisotopic precursor selection, MIPS). Dynamic exclusion was used, with a duration of 90 s, repeat count of 1, mass tolerance of 10 ppm, and the “exclude isotopes” option checked. For each MS1, 8 data-dependent MS/MS scans were collected. MS/MS scans were conducted in the linear ion trap with the “rapid” scan rate, 50 ms max injection time, AGC target set to 200%, CID collision energy of 35% with 10 ms activation time, and 0.5 m/z isolation window. For TMTPro labelled samples, an MS3 scan was also included in the method. Unless otherwise noted in the methods, the real-time search filter was enabled(Schweppe et al., 2020). Using a human fasta downloaded from Uniprot, fixed modifications for the TMTpro mass (+304.207146) were added to n-terminal residues and lysines. Carbamidomethly (+57.021464) was added for cysteines. Oxidation (+15.9949) was added as a variable modification on methionines. Missed cleavages were set to maximum of 1. “TMT mode” was enabled and thresholds of 1 and 0.05 for Xcorr and dCn respectively were used as minimums to trigger SPS-MS3 scans. SPS ions were set to 10 and MS3 scans were performed at a resolving power of 50,000, with an HCD collision energy of 45%, AGC of 200%, with a maximum injection time of 200 ms.

Label-free mass spectrometry data was analyzed with MSFragger(Kong et al., 2017) search algorithm searched against a full human protein database with forward and reverse protein sequences. Fixed modifications included Carbamidomethyl (+57.021464) on cysteines. Variable modifications included were Oxidation (+15.9949) on methionine and formylation

(+27.994915) on lysines. Peptides up to 2 missed cleavages were included. Peptide spectral matches and proteins were filtered to a 1% false discovery rate using Percolator(Käll et al., 2007).

Multiplexed raw mass spectrometry data was analyzed using the Comet(Eng et al., 2013) search algorithm, searched against a full human protein database with forward and reverse protein sequences (Uniprot 10/2020). Precursor monoisotopic peaks were estimated using the Monocle package. Fixed modifications included TMTpro (+304.207146) on n-terminal residues and lysines and Carbamidomethyl (+57.021464) on cysteines. Variable modifications included were Oxidation (+15.9949) on methionine and formylation (+27.994915) on lysines. Peptides up to 2 missed cleavages were included. Peptide spectral matches and proteins were filtered to a 1% false discovery rate using the rules of parsimony and protein picking. Protein quantification was done using signal-to-noise estimates of reporter ions. Samples were column normalized for total protein concentration. After filtering for contaminants, we performed a two-sided t-test comparing each O-MAP condition using Benjamini-Hochberg adjusted p values (i.e. q-values). Log_2 fold changes of the mean of the biological replicates were also calculated for each biological condition. Human Protein Atlas(Thul et al., 2017) subcellular locations were downloaded and the “main location” was assigned to each protein with a supported or enhanced reliability level. SAINT scores and interaction false discovery rates were calculated with the SAINTexpress software(Choi et al., 2011; Teo et al., 2014). Significant hits were those with a SAINT calculated FDR less than 1%(Choi et al., 2012). BioPlex interaction networks were accessed through the online BioPlex Explorer(Huttlin et al., 2015) (https://bioplex.hms.harvard.edu/). Networks were imaged using Cytoscape 3.10.02(Shannon et al., 2003). Protein complex members were accessed through CORUM(Ruepp et al., 2008). Gene set enrichment analysis was performed with clusterProfiler(Yu et al., 2012) and fgsea(Korotkevich et al., 2021) packages.

2.5.10 Preparation of soluble chromatin for affinity purification followed by next generation sequencing

For confirmation of single-copy O-MAP labeling, loop anchor-biotinylated pellets of 10-20 million cells were removed from -80°C to thaw at room temperature. Each cell pellet was resuspended in an SDS lysis buffer consisting of 1% SDS and 200 mM EPPS with protease inhibitors. The cell mixture was sonicated at 4°C using a Covaris LE-220 focused ultrasonicator with the following protocol: 300W peak incident power, 15% duty factor, 200 cycles per burst, with a treatment time of 20-30 minutes in 130-µl microTUBEs with AFA fiber (Covaris 520077). After the samples had returned to room temperature, the sheared fixed chromatin was transferred to fresh 1.5-ml Protein Lo-Bind tubes and centrifuged at 21130 G for 10 minutes to pellet cellular debris. The supernatants were transferred to a new set of tubes. The cleared chromatin samples were quantified using the Pierce BCA Protein Assay Kit (Thermo Fisher 23225). Next, 50 µl of sheared chromatin was sampled for reverse crosslinking, DNA extraction, and gel electrophoresis to verify that a significant amount of DNA had been sheared to <700 base pairs. A sample of 10 µg sheared chromatin was reserved and stored at -20°C as immunoprecipitation input. 200 µg of chromatin was used to couple with 200 µg of streptavidin beads for one hour in a Protein Lo-Bind tube at room temperature with nutation. The coupled beads were collected and separated from the flow-through using a magnetic rack. After the flow-through was removed, the beads underwent the following washes:

- 2% SDS with 20 mM EPPS
- 2% SDS with 20 mM EPPS
- High Salt Buffer containing 500 mM NaCl, 1 mM EDTA, 50 mM of HEPES pH7.5, 0.1% sodium deoxycholate, and 1% TritonX-100
- LiCl Buffer containing 250 mM LiCl, 1 mM EDTA, 10 mM Tris-HCl pH 8.0, and 0.5% of IGEPAL CA-630

- TE Buffer with 10 mM Tris and 1 mM EDTA
- TE Buffer with 10 mM Tris and 1 mM EDTA

The washes were performed as follows: briefly spin and immobilize the beads on a magnetic rack, pipette out the supernatant as much as possible, resuspend the beads in 0.8 ml of wash buffer, and incubate for 5 minutes with nutation. The washed beads were resuspended in 300 ul of reverse crosslinking buffer containing 300 mM NaCl, 300 mM Tris-HCl pH 8.0, and 1 mM EDTA. Both the eluate beads and the input chromatin were incubated at 65°C for 16 hours for reverse crosslinking. The next day, 4 ul of 20 mg/ml proteinase K (Roche 3115836001) was added to the eluates and inputs and incubated at 50°C for 2 hours to cleave away proteins. The DNA was isolated from the mixture using phenol chloroform extraction followed by ethanol precipitation. Before sequencing library generation, the precipitated DNA was further purified using SPRI beads. The purified DNA was used to generate next-generation sequencing libraries using the NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB E7645S) and NEBNext Multiplex Oligos for Illumina Index Primers Set 1 and 3 (NEB E7335S, E7710S) and PCR-amplified for 15 cycles. The sequencing libraries were quantified using the Qubit 4 fluorometer and library sizes were quantified using the D1000 ScreenTape assay (Agilent 5067-5582) on the TapeStation 4200 automated electrophoresis platform.

2.5.11 DNA sequencing and data analysis

The libraries were mixed and sequenced pair-ended at 50-bp read length on an Illumina NextSeq 2000 sequencer to depths of 14.1-351.8 million reads per eluate sample and 3.14-16.45 millions reads per input sample using the NextSeq 1000/2000 P2 Reagents (100 Cycles) kit (Illumina 20046811). Reads were demultiplexed and adapters were removed using Cutadapt(Martin, 2011). Trimmed reads were mapped to the reference genome (GRCh38) using Bowtie2 version 2.5.3 with the parameter -X 1000 keeping reads with a MAPQ>=30(Langmead & Salzberg, 2012). Duplicate reads were removed using Picard 3.1.1(*Picard*, n.d.). Eluate reads

were normalized to input reads using DeepTools(Ramírez et al., 2016) bamCompare with the following parameters: –binSize 20 –normalizeUsing BPM –smoothLength 60 – extendReads 150. Normalized data were visualized using Coolbox 0.3.9(Xu et al., 2021).

2.6 Data Availability

The mass spectrometry proteomics data have been deposited to the ProteomeXchange(Vizcaíno et al., 2014) Consortium via the MassIVE with the data set identifier PXD054080. Sequencing data will be deposited to Gene Expression Omnibus(*Home - GEO - NCBI*, n.d.) before formal acceptance for publication.

2.7 Author Contributions

Y.L., C.D.M, B.J.B., and D.K.S. conceived and designed the project. Y.L., C.D.M, M.K., T.A.P., R.F., C.H., S.A., A.F.T., and E.K. performed experiments. Y.L., C.D.M., and C.K.C. performed computational analyses. Y.L., C.D.M., B.J.B., and D.K.S. wrote the manuscript. All authors edited and reviewed the manuscript. D.M.S., B.J.B., and D.K.S. supervised the work.

2.8 Competing Interest Statement

D.K.S. is a collaborator with Thermo Fisher Scientific, Genentech, Calico Labs, and AI Proteins. C.K.C., A.F.T., E.K., D.M.S., and B.J.B. have filed a patent application covering aspects of this work. B.J.B. is listed as an inventor on patent applications related to the SABER technology related to this work.

2.9 Acknowledgements

We would like to thank members of the Shechner, Beliveau, and Schweppe labs for constructive feedback and technical assistance in assembling this work. We would also like to

thank Drs. Jay Shendure, Shao-En Ong, Christine Quietsch, Emily Hatch, Gavin Ha, Celeste Berg, Christine Disteche, Andrew Stergachis, and Stanley Fields for helpful discussions of this work. We would like to acknowledge the following sources of support: R35GM137916 (BJB), R35GM150919 (DKS), 5T32HG000035 (CDM), the W.M. Keck Foundation (BJB, DKS), an Andy Hill CARE Distinguished Researcher Award (DKS), a Damon Runyon Dale Frey Award (BJB), a Cancer Consortium New Investigator Award (DKS), and The Pew Charitable Trusts (DKS).

2.10 Supplemental Information

Figure S1. Predicted genome-wide binding profile of the pan-alpha probe.

Figure S2. Replicate analysis of multi-target DNA O-MAP proteomics experiment.

Figure S3. Relative quantitation for the multi-target DNA O-MAP proteomics experiment compared to no-probe control and mtDNA datasets.

Figure S4. Comparison of histone proteins between telomere and pan-alpha probes.

Figure S5. DNA O-MAP biotin purification sequencing of chr3 left, chr3 right, chr10 non-loop anchors, and no-primary-probe control.

Figure S6. DNA O-MAP biotin purification sequencing of multiplexed targeting of chr3 left, chr10 right, chr19 right anchors, and no-primary-probe control in duplicates.

Supplementary Table 1. Oligonucleotide probe sequences used in this work.

Supplementary Table 2. Proteomic data for the enrichment of telomere probe associated proteins.

Supplementary Table 3. Quantitative proteomic data for the multi-target DNA O-MAP proteomics experiment.

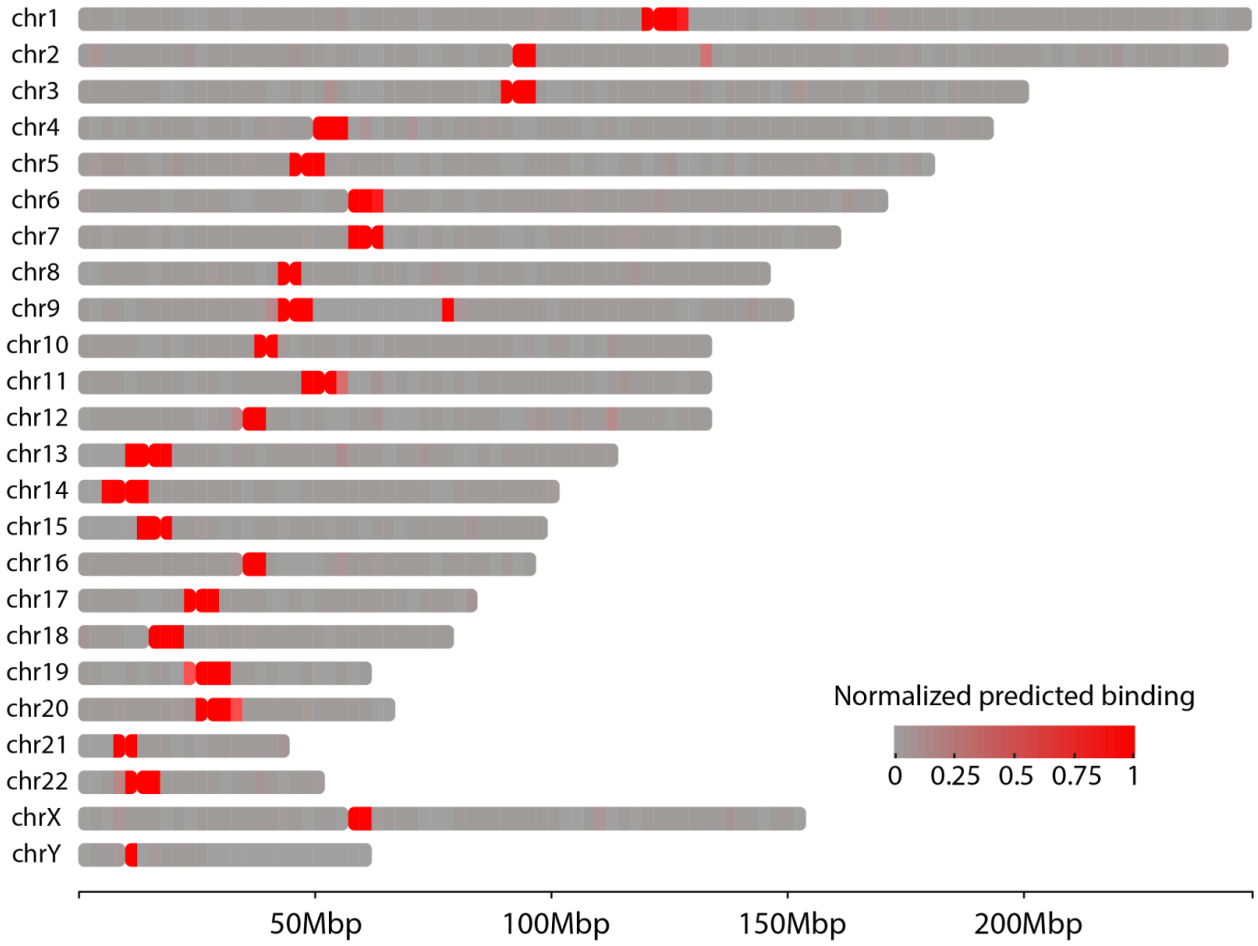


Figure S1. Predicted genome-wide binding profile of the pan-alpha probe. The intensity of red indicates the amount of predicted probe binding.

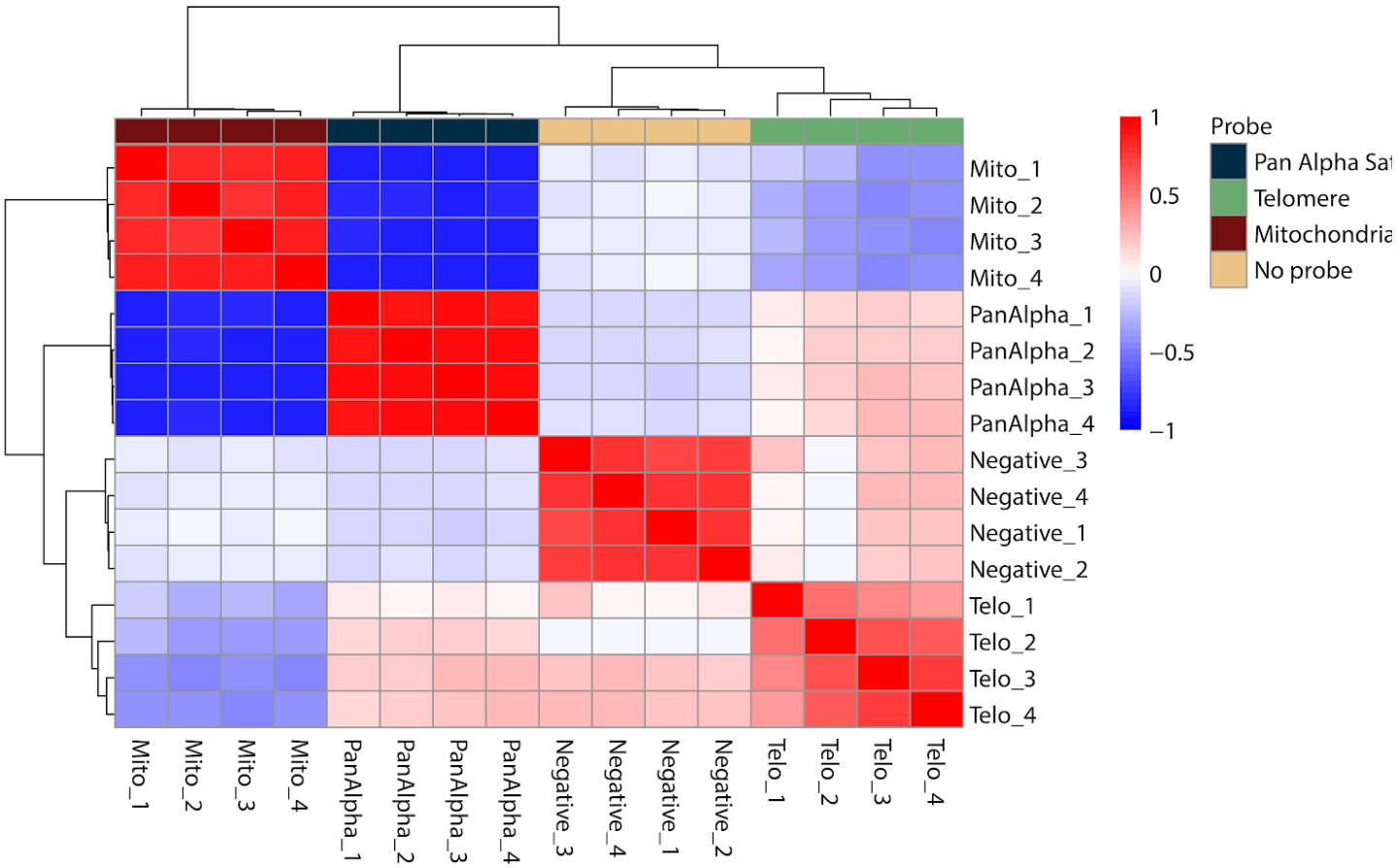


Figure S2. Replicate analysis of multi-target DNA O-MAP proteomics experiment. A) Pearson correlation coefficient of the raw protein intensity values for each replicate of the multiplex with hierarchical clustering on the rows and columns.

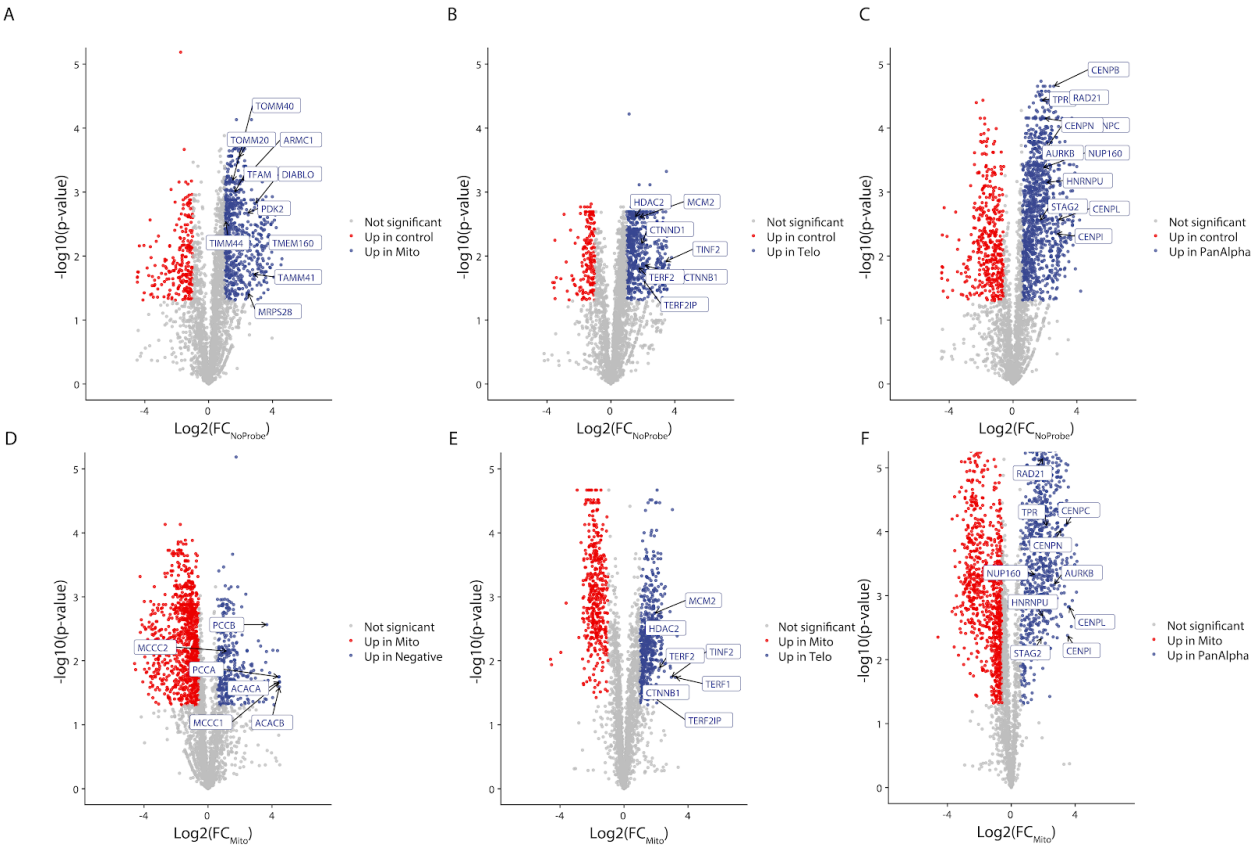


Figure S3. Relative quantitation for the multi-target DNA O-MAP proteomics experiment compared to no-probe control and mtDNA datasets. Volcano plots from multiplexed proteomics experiments with proteins of interest highlighted. A-C) Fold-changes and significance calculated compared to no probe. D-F) Fold-changes and significance calculated compared to mtDNA probe.

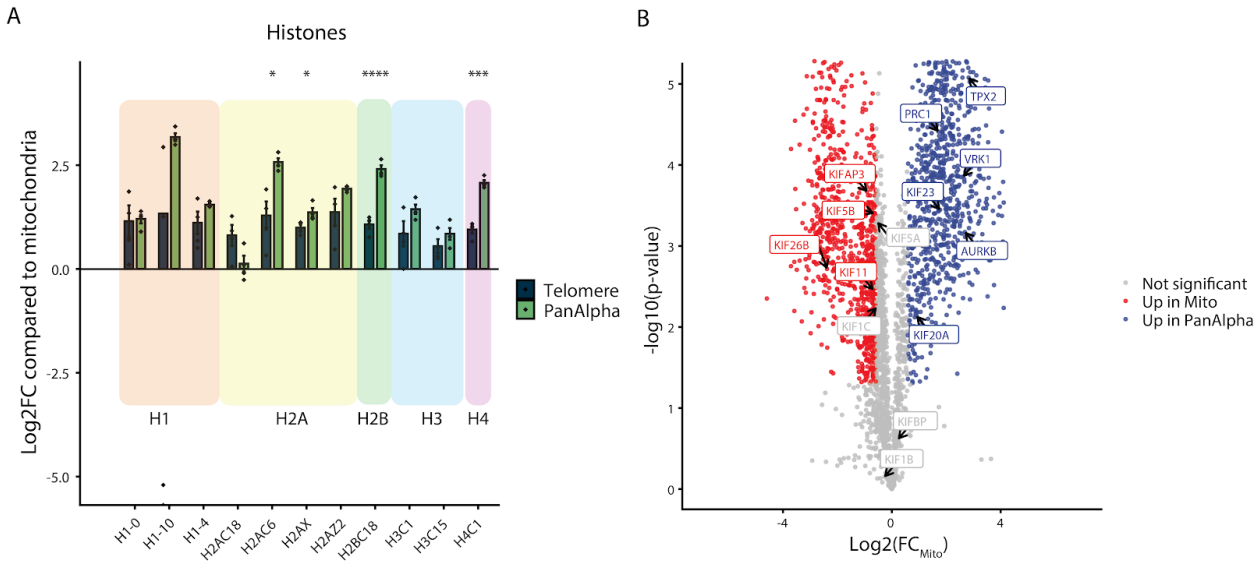
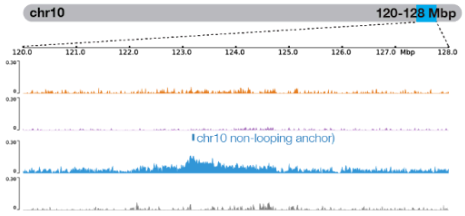


Figure S4. Comparison of histone proteins between telomere and pan-alpha probes. A) Log2 fold change of proteins compared to mitochondrial probe enriched histone complex proteins. Significance calculated based on Welch's t-test for pairwise comparisons (p-value: *<0.05, **<0.01, ***<0.001, ****<0.0001). B) Volcano plot comparing the fold change of pan-alpha to the mtDNA probe with spindle proteins highlighted.

A

	Probed anchors	Expected contact anchors
Track 1	chr3 left	chr3 right
Track 2	chr3 right	chr3 left
Track 3	chr10 non-loop anchor	None
Track 4	No primary probe	None



B

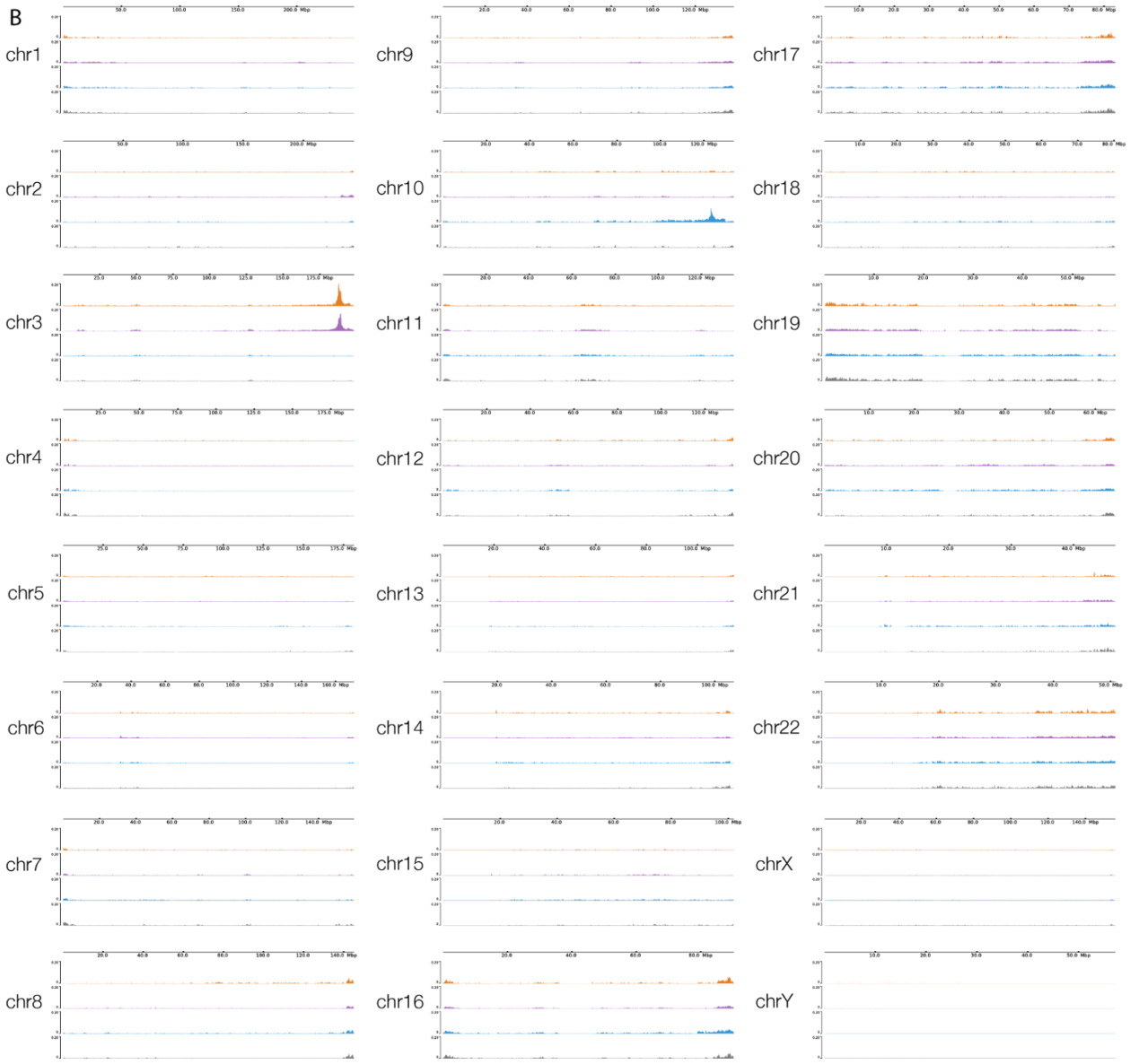
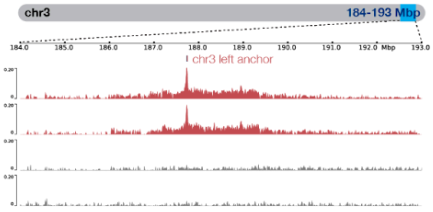


Figure S5. DNA O-MAP biotin purification sequencing of chr3 left, chr3 right, chr10 non-loop anchors, and no-primary-probe control. A) Table listing the three anchors (Track 1-3) and no-primary-probe control (Track 4) biotinylated by DNA O-MAP and their expected contact anchors (left). Biotin purification sequencing signals across the 8-Mb region on chromosome 10 corresponding to the chr10 non-loop anchor targeted (right). B) Biotin purification sequencing signals across every chromosome in the genome for this experiment.

A

	Probed anchors	Expected contact anchors
Track 1 Track 2	chr3 left, chr10 right, chr19 right	chr3 right, chr10 left, chr19 left-1, chr19 left-2
Track 3 Track 4	No primary probe	None



B

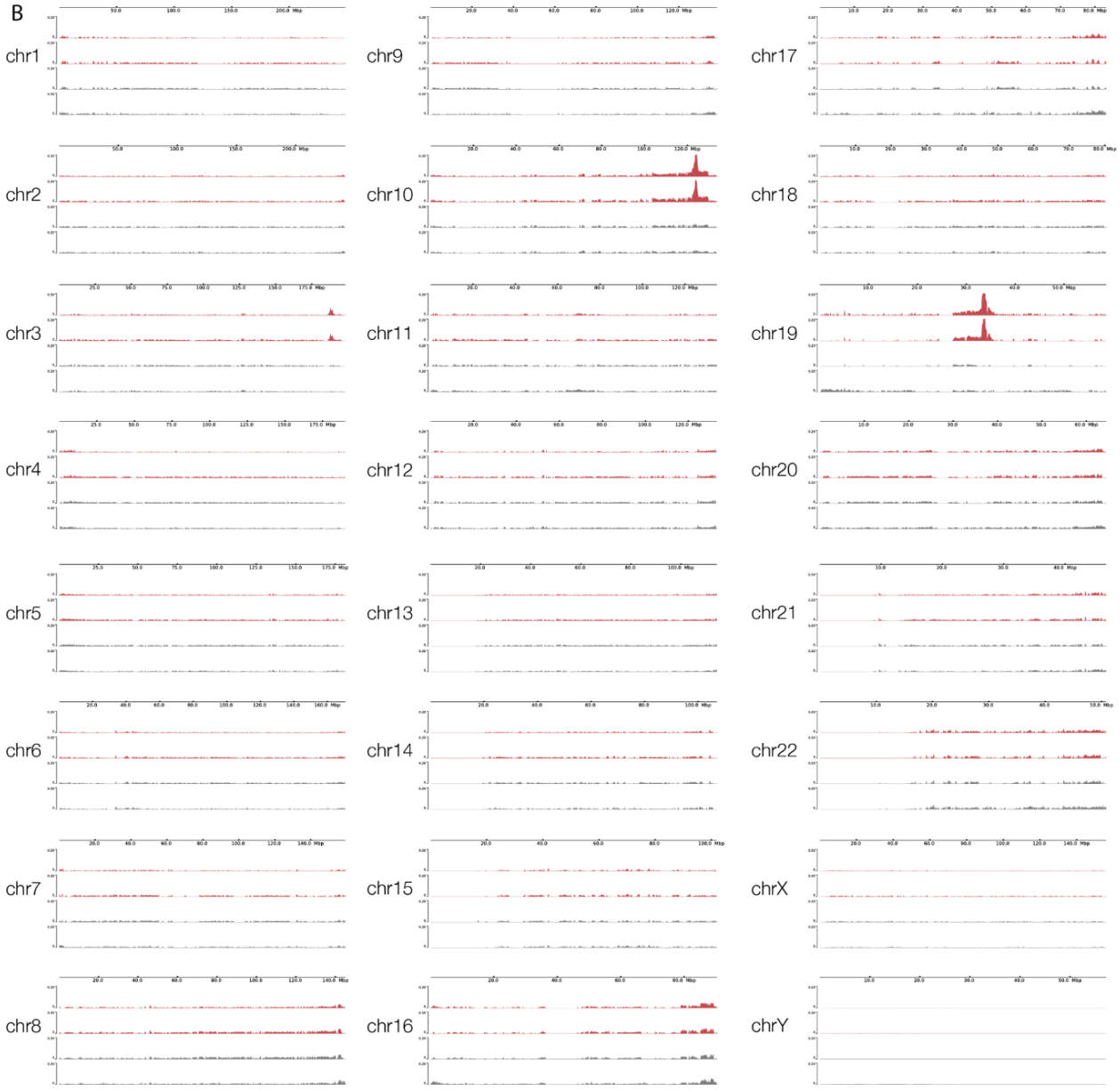


Figure S6. DNA O-MAP biotin purification sequencing of multiplexed targeting of chr3 left, chr10 right, chr19 right anchors, and no-primary-probe control in duplicates. A) Table listing the three anchors (Track 1-2) and no-primary-probe control (Track 3-4) biotinylated by DNA O-MAP and their expected contact anchors (left). Biotin purification sequencing signals across the 9-Mb region on chromosome 3 corresponding chr3 left anchor targeted in Track 1-2 (right). B) Biotin purification sequencing signals across every chromosome in the genome for this experiment.

Chapter III: Perspectives and Future Directions

3.1 Exploring the lower limits of target sizes, cell inputs, and their optimal combinations

3.1.1 Interrogating smaller, single-copy loci for local chromatin proteomes

In Chapter II, I and my colleagues developed a scalable locus-specific biotinylation method that can process hundreds of million cells in solution per sample with up to 18 samples in parallel by one experimenter. We coupled the locus-specific biotinylation method with purification followed by sample multiplexed quantitative proteomics or next-generation sequencing, and termed the technology “DNA O-MAP”. We demonstrated that DNA O-MAP can differentiate the proteomes of alpha-satellite repeats, telomeres, and mitochondrial DNA with four highly consistent technical replicates. We also recovered pairwise and multiway DNA interactions of cohesin-mediated chromatin loops in a ligation-independent manner using 2-5 million biotinylated cells.

We note that the smallest target size in our proteomics dataset remains the telomeres of ~515 kb in aggregate size, which consist of 0.017% of the human genome in HCT-116 cells. Theoretically, our chromatin isolation method offered at least a ~5825-fold enrichment to obtain pure telomeric chromatin. In reality, we purified telomeric chromatin from whole cell lysates, instead of solely genome-bound proteins; thus, the fold enrichment is significantly higher. However, to identify smaller functional features such as 1-kb long gene promoters or enhancers, we would need the method to offer a 3-million-fold enrichment. It remains unclear whether DNA O-MAP can achieve such enrichment levels without further improvements in efficiencies of biotinylation, purification, and MS detection. For instance, to carefully explore the lower limit of

viewpoint size at 10-kb either as an individual locus or a class of loci with similar locus proteomes, we need to identify a proof-of-concept 10-kb viewpoint with well-known, distinctly different proteomes under biological change, or two 10-kb viewpoints with distinctly different proteomes without treatment. Examples of proof-of-concept viewpoint pairs could include the promoter of a gene in the poised state versus the same promoter during gene expression, or the promoters of active genes versus the promoters of repressed ones. Moreover, before moving into 10-kb single-copy viewpoints, several intermediate regions should be studied since the smallest viewpoint we successfully purified for proteome discovery remains the 515-kb repetitive loci using one single genome-binding oligo. As oligonucleotide hybridization forms the basis of locus-specific biotinylation, we would need several thousand oligos to cover a hundred-kb single-copy viewpoint in sufficient density. The pool of genome-binding oligos for a single-copy region may have much less predictable hybridization efficiencies and lead to overall weaker biotinylation.

The *HOX* gene clusters, which are larger single-copy regions at ~130-kb size, are currently under careful examination in the Beliveau and Schweppe Labs. I characterized the biotinylation of *HOXA* and *HOXB* viewpoints with imaging endpoints and streptavidin western blotting. Along with Dr. Conor Herlihy, a postdoctoral fellow in both labs, I biotinylated the *HOXA* and *HOXB* viewpoints in the presence and absence of the EZH2 inhibitor GSK126 to induce the local chromatin proteome shift. It is clear from my preliminary studies that the amount of proteins bound at a smaller genomic region will be lower compared with a larger one. To increase the absolute amount of purified peptides that reach detection, we need to increase the number of cells put into each sample. It is also important to note that K562 cells demonstrate a near-triploid karyotype, which may also increase the absolute amounts of eluates. Considering the viewpoint size, K562 karyotype, and practical parallel sample handling, each *HOXA* and *HOXB* sample contains ~60 million K562 cells, more than twice of the ~27 million cells for the

515-kb telomere locus. This proteomic study holds promise to characterize the performance of DNA O-MAP at 130-kb non-repetitive regions.

3.1.2 Exploring the cell input limits to examine the local chromatin proteomes of a locus of interest

For a more accurate estimation of cell input, we characterized the input for the proteomics dataset by the amount of protein that entered purification instead of cell numbers. We quantified biotinylated whole cell lysates by the bicinchoninic acid assay and used 2.7 mg per sample for the streptavidin purification. Since we consistently obtain ~1 mg of whole-cell lysates from 10 million 4% PFA- fixed HCT-116 cells, we estimated that ~ 27 million cells were used for each sample. This amount happened to be a sufficient input for three viewpoints in the study, especially the smallest and most likely limiting viewpoint, the telomeres. 27 million is likely not the smallest number of cells that can achieve good quality local chromatin proteomes. To explore the upper and lower number of cells required, the locus of interest and the cell line of interest should be selected, as the viewpoint size would determine the fold enrichment and the karyotype of the cell would determine the absolute quantity of eluates possible. The LC-MS method for detection will also present different sensitivity. As TMT quantitative proteomics allows multiple samples to be combined and detected as one, the sample multiplexing may lead to higher sensitivity and better within-experiment comparisons.

To characterize input requirements of telomeric proteome capture in HCT-116 cells, for instance, we can biotinylate the telomeres in 2, 5, 10, 30, 50, 100 million cells respectively in three technical replicates, and detect the 18 telomere proteomes within one TMT experiment. In the lower input experiment, we may see the proteome with a larger percentage of non-specific background such as cytosolic proteins, endogenous biotinylated proteins, and locally bound chromatin architectural proteins compared to the high input samples. As the dynamic range of

locally-bound proteins can be immense, in the higher input experiment, we can expect more rare protein residents or the difficult-to-detect proteins that are small in size or have more sterically buried aromatic residues to be quantified. We would hope to find a cell input in which increasing the number of cells no longer benefits the detection of the rare and biologically meaningful local binders.

3.2 Multi-step purification strategy

3.2.1 Nuclear isolation

Our current purification strategy consists only of a single-step streptavidin purification from whole-cell lysates with extensive stringent washes. For small, non-repetitive loci, a larger amount of input material is needed. Yet simply using the current purification method at a larger scale may also scale up the presence of background proteins. The method would need to be optimized to achieve a much higher fold enrichment. We may implement a multi-step purification strategy to isolate small, less abundant loci to sufficient purity. Prior to streptavidin purification, a nuclear isolation step would remove the majority of cytosolic proteins. Since our method is based on hybridization and requires the cells to be fixed with 4% PFA, the nuclear isolation would need to occur before the cells enter the hybridization-biotinylation workflow. The in-solution hybridization-biotinylation workflow would require careful engineering of the extensive centrifugation-based washes of nuclei instead of whole cells. Since nuclei are smaller in size, they require higher centrifugal speed and longer time in the centrifuge for complete sample recovery after every wash, which in turn may induce more damage to the nuclear structure and more severe clumping.

3.2.2 Total chromatin isolation

Another idea to implement the multi-step purification strategy is to isolate total chromatin from whole cells, which would remove the background of both cytosolic and soluble nuclear proteins. Unlike nuclear isolation, total chromatin isolation can occur after the loci have been biotinylated, preserving the whole cell to aid centrifugal wash. Since the ChEP method introduced by Kustatscher et al. in 2014(Kustatscher, Wills, et al., 2014) and DNA O-MAP protein extraction both begin with PFA-fixed cells and use highly similar buffers and handling steps, ChEP is likely compatible with our downstream processes and could be seamlessly integrated into our protein extraction step. By separating the whole-cell lysate into a chromatin fraction and a remaining fraction containing both cytosolic and soluble nuclear proteins, only the chromatin fraction would enter the streptavidin purification workflow. It would be interesting to assess the number and intensity of signal and noise proteins identified from purifications using whole-cell lysates and ChEP-chromatin fractions using a well-characterized locus, such as the telomeres.

3.3 *In silico* purification to remove cytosolic and nuclear background proteins

To ensure each isolated protein originates from the gene locus, it is important to first define a sizable set of known, locus-specific binders of the target locus from existing knowledge and critically assess whether the isolated proteome at hand encompasses most of these protein residents. If only a few known proteins can be used as positive hits and they abundantly occupy other loci in the genome, it may not be possible for us to distinguish whether the purification produced a complete protein catalog of the locus or only recovered the locus with partial success. In addition to defining an expansive list of positive hits, it is also imperative to establish

a repository of background proteins to ensure that they do not occupy the majority of the proteome at hand. With a good understanding of what we should and should not retrieve, we can interpret the rest of the local proteome as newly discovered local residents and begin to orthogonally validate the identified proteins.

The multi-step fractionation of whole cell lysates may aid in refining our definition of contaminant proteins. Since proteins routinely travel between the nucleus and cytosol, nuclear isolation, although may not easily integrate into the hybridization-biotinylation workflow, can be performed on its own to provide a better understanding of the dynamic cytosolic background. To understand the background from nuclear proteins that are not interacting with DNA, isolated nuclei can undergo PFA fixation and ChEP fractionation to obtain fractions of total chromatin and soluble nuclear proteins. Comparison of the cytosolic, soluble nuclear, and chromatin proteomes should establish probability scores for how likely a protein is seen from each fraction. These probabilities can be used to help us understand the obtained local proteome, in the case of a less successful chromatin isolation.

In addition to contaminants from proteins that did not originate from the locus, highly abundant architectural chromatin proteins could mask the low-abundance regulatory proteins that interact with the locus, a problem not unique to locus-specific proteomic studies. To overcome the interference of abundant proteins that actually did originate from the locus, we are currently using high-field asymmetric waveform ion mobility spectrometry (FAIMS) to induce greater dispersion of ions to increase proteome depth for identification and quantitation. In addition to FAIMS, during MS data acquisition, we use a real-time database search platform(Schweppe et al., 2020) that processes a single spectrum search in less than 10 ms. By filtering out abundant proteins that are already quantified and triggering subsequent scans in an adaptive manner, we increased the data acquisition efficiency for our complex local proteome samples.

References

Aguilar, R., Camplisson, C. K., Lin, Q., Miga, K. H., Noble, W. S., & Beliveau, B. J. (2024). Tigerfish designs oligonucleotide-based in situ hybridization probes targeting intervals of highly repetitive DNA at the scale of genomes. *Nature Communications*, *15*(1), 1027.

Alabert, C., Bukowski-Wills, J.-C., Lee, S.-B., Kustatscher, G., Nakamura, K., de Lima Alves, F., Menard, P., Mejlvang, J., Rappsilber, J., & Groth, A. (2014). Nascent chromatin capture proteomics determines chromatin dynamics during DNA replication and identifies unknown fork components. *Nature Cell Biology*, *16*(3), 281–291.

Altemose, N., Logsdon, G. A., Bzikadze, A. V., Sidhwani, P., Langley, S. A., Caldas, G. V., Hoyt, S. J., Uralsky, L., Ryabov, F. D., Shew, C. J., Sauria, M. E. G., Borchers, M., Gershman, A., Mikheenko, A., Shepelev, V. A., Dvorkina, T., Kunyavskaya, O., Vollger, M. R., Rhie, A., … Miga, K. H. (2022). Complete genomic and epigenetic maps of human centromeres. *Science*, *376*(6588), eabl4178.

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J., Staden, R., & Young, I. G. (1981). Sequence and organization of the human mitochondrial genome. *Nature*, *290*(5806), 457–465.

Attar, S., Browning, V. E., Liu, Y., Nichols, E. K., Tsue, A. F., Shechner, D. M., Shendure, J., Lieberman, J. A., Akilesh, S., & Beliveau, B. J. (2023). Programmable peroxidase-assisted signal amplification enables flexible detection of nucleic acid targets in cellular and histopathological specimens. *bioRxiv : The Preprint Server for Biology*. https://doi.org/10.1101/2023.01.30.526264

Beattie, M., & A H Jones, O. (2023). Rate of advancement of detection limits in mass spectrometry: Is there a Moore's Law of mass spec? *Mass Spectrometry (Tokyo, Japan)*,

12(1), A0118.

Becker, J. S., McCarthy, R. L., Sidoli, S., Donahue, G., Kaeding, K. E., He, Z., Lin, S., Garcia, B. A., & Zaret, K. S. (2017). Genomic and proteomic resolution of heterochromatin and its restriction of alternate fate genes. *Molecular Cell*, *68*(6), 1023–1037.e15.

Beliveau, B. J., Joyce, E. F., Apostolopoulos, N., Yilmaz, F., Fonseka, C. Y., McCole, R. B., Chang, Y., Li, J. B., Senaratne, T. N., Williams, B. R., Rouillard, J.-M., & Wu, C.-T. (2012). Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(52), 21301–21306.

Bersaglieri, C., Kresoja-Rakic, J., Gupta, S., Bär, D., Kuzyakiv, R., Panatta, M., & Santoro, R. (2022). Genome-wide maps of nucleolus interactions reveal distinct layers of repressive chromatin domains. *Nature Communications*, *13*(1), 1483.

Bickmore, W. A., & van Steensel, B. (2013). Genome architecture: domain organization of interphase chromosomes. *Cell*, *152*(6), 1270–1284.

Branon, T. C., Bosch, J. A., Sanchez, A. D., Udeshi, N. D., Svinkina, T., Carr, S. A., Feldman, J. L., Perrimon, N., & Ting, A. Y. (2018). Efficient proximity labeling in living cells and organisms with TurboID. *Nature Biotechnology*, *36*(9), 880–887.

Briand, N., & Collas, P. (2020). Lamina-associated domains: peripheral matters and internal affairs. *Genome Biology*, *21*(1), 85.

Broad, A. J., DeLuca, K. F., & DeLuca, J. G. (2020). Aurora B kinase is recruited to multiple discrete kinetochore and centromere regions in human cells. *The Journal of Cell Biology*, *219*(3). https://doi.org/10.1083/jcb.201905144

Bujold, D., Morais, D. A. de L., Gauthier, C., Côté, C., Caron, M., Kwan, T., Chen, K. C., Laperle, J., Markovits, A. N., Pastinen, T., Caron, B., Veilleux, A., Jacques, P.-É., & Bourque, G. (2016). The International Human Epigenome Consortium Data Portal. *Cell Systems*, *3*(5), 496–499.e2.

Cenik, B. K., Aoi, Y., Iwanaszko, M., Howard, B. C., Morgan, M. A., Andersen, G. D., Bartom, E. T., & Shilatifard, A. (2024). TurboCas: A method for locus-specific labeling of genomic regions and isolating their associated protein interactome. *Molecular Cell*, *84*(24), 4929–4944.e8.

Chakravarti, D., LaBella, K. A., & DePinho, R. A. (2021). Telomeres: history, health, and hallmarks of aging. *Cell*, *184*(2), 306–322.

Chen, P., Li, W., & Li, G. (2021). Structures and Functions of Chromatin Fibers. *Annual Review of Biophysics*, *50*, 95–116.

Chen, R., & Wold, M. S. (2014). Replication protein A: single-stranded DNA's first responder: dynamic DNA-interactions allow replication protein A to direct single-strand DNA intermediates into different pathways for synthesis or repair. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *36*(12), 1156–1161.

Choi, H., Larsen, B., Lin, Z.-Y., Breitkreutz, A., Mellacheruvu, D., Fermin, D., Qin, Z. S., Tyers, M., Gingras, A.-C., & Nesvizhskii, A. I. (2011). SAINT: probabilistic scoring of affinity purification-mass spectrometry data. *Nature Methods*, *8*(1), 70–73.

Choi, H., Liu, G., Mellacheruvu, D., Tyers, M., Gingras, A.-C., & Nesvizhskii, A. I. (2012). Analyzing protein-protein interactions from affinity purification-mass spectrometry data with SAINT. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]*, *Chapter 8*, 8.15.1–8.15.23.

Cho, K. F., Branon, T. C., Udeshi, N. D., Myers, S. A., Carr, S. A., & Ting, A. Y. (2020). Proximity labeling in mammalian cells with TurboID and split-TurboID. *Nature Protocols*, *15*(12), 3971–3999.

Chou, H.-C., Bhalla, K., Demerdesh, O. E., Klingbeil, O., Hanington, K., Aganezov, S., Andrews, P., Alsudani, H., Chang, K., Vakoc, C. R., Schatz, M. C., McCombie, W. R., & Stillman, B. (2021). The human origin recognition complex is essential for pre-RC assembly, mitosis, and maintenance of nuclear structure. *eLife*, *10*. https://doi.org/10.7554/eLife.61797

Connolly, C. N., Futter, C. E., Gibson, A., Hopkins, C. R., & Cutler, D. F. (1994). Transport into and out of the Golgi complex studied by transfecting cells with cDNAs encoding horseradish peroxidase. *The Journal of Cell Biology*, *127*(3), 641–652.

Costa, A., & Diffley, J. F. X. (2022). The initiation of eukaryotic DNA replication. *Annual Review of Biochemistry*, *91*(1), 107–131.

Coster, G., & Diffley, J. F. X. (2017). Bidirectional eukaryotic DNA replication is established by quasi-symmetrical helicase loading. *Science (New York, N.Y.)*, *357*(6348), 314–318.

Déjardin, J., & Kingston, R. E. (2009). Purification of proteins associated with specific genomic Loci. *Cell*, *136*(1), 175–186.

Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing chromosome conformation. *Science (New York, N.Y.)*, *295*(5558), 1306–1311.

de Lange, T. (2005). Shelterin: the protein complex that shapes and safeguards human telomeres. *Genes & Development*, *19*(18), 2100–2110.

de Lange, T. (2018). Shelterin-Mediated Telomere Protection. *Annual Review of Genetics*, *52*, 223–247.

Deng, Z., & Beliveau, B. J. (2022). An open source 16-channel fluidics system for automating sequential fluorescent in situ hybridization (FISH)-based imaging. *HardwareX*, *12*, e00343.

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C., Green, R. D., & Dekker, J. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Research*, *16*(10), 1299–1309.

Downes, D. J., Smith, A. L., Karpinska, M. A., Velychko, T., Rue-Albrecht, K., Sims, D., Milne, T. A., Davies, J. O. J., Oudelaar, A. M., & Hughes, J. R. (2022). Capture-C: a modular and flexible approach for high-resolution chromosome conformation capture. *Nature Protocols*, *17*(2), 445–475.

Dumrongprechachan, V., Salisbury, R. B., Soto, G., Kumar, M., MacDonald, M. L., &

Kozorovitskiy, Y. (2021). Cell-type and subcellular compartment-specific APEX2 proximity labeling reveals activity-dependent nuclear proteome dynamics in the striatum. *Nature Communications*, *12*(1), 4855.

ENCODE Project Consortium. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, *306*(5696), 636–640.

Engel, K. L., Lo, H.-Y. G., Goering, R., Li, Y., Spitale, R. C., & Taliaferro, J. M. (2022). Analysis of subcellular transcriptomes by RNA proximity labeling with Halo-seq. *Nucleic Acids Research*, *50*(4), e24.

Eng, J. K., Jahan, T. A., & Hoopmann, M. R. (2013). Comet: an open-source MS/MS sequence database search tool. *Proteomics*, *13*(1), 22–24.

Ernst, J., & Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods*, *9*(3), 215–216.

Fang, Y., & Zou, P. (2023). Photocatalytic proximity labeling for profiling the subcellular organization of biomolecules. *Chembiochem: A European Journal of Chemical Biology*, *24*(8), e202200745.

Fields, S. (2007). Molecular biology. Site-seeing by sequencing. *Science (New York, N.Y.)*, *316*(5830), 1441–1442.

Fisher, R. J., Fivash, M., Casas-Finet, J., Erickson, J. W., Kondoh, A., Bladen, S. V., Fisher, C., Watson, D. K., & Papas, T. (1994). Real-time DNA binding measurements of the ETS1 recombinant oncoproteins reveal significant kinetic differences between the p42 and p51 isoforms. *Protein Science: A Publication of the Protein Society*, *3*(2), 257–266.

Flynn, R. L., Centore, R. C., O'Sullivan, R. J., Rai, R., Tse, A., Songyang, Z., Chang, S., Karlseder, J., & Zou, L. (2011). TERRA and hnRNPA1 orchestrate an RPA-to-POT1 switch on telomeric single-stranded DNA. *Nature*, *471*(7339), 532–536.

Fried, M., & Crothers, D. M. (1981). Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic Acids Research*, *9*(23), 6505–6525.

Fu, D., & Collins, K. (2007). Purification of human telomerase complexes identifies factors involved in telomerase biogenesis and telomere length regulation. *Molecular Cell*, *28*(5), 773–785.

Gao, X. D., Tu, L.-C., Mir, A., Rodriguez, T., Ding, Y., Leszyk, J., Dekker, J., Shaffer, S. A., Zhu, L. J., Wolfe, S. A., & Sontheimer, E. J. (2018). C-BERST: defining subnuclear proteomic landscapes at genomic elements with dCas9–APEX2. *Nature Methods*, *15*(6), 433–436.

Garcia-Exposito, L., Bournique, E., Bergoglio, V., Bose, A., Barroso-Gonzalez, J., Zhang, S., Roncaioli, J. L., Lee, M., Wallace, C. T., Watkins, S. C., Opresko, P. L., Hoffmann, J.-S., & O'Sullivan, R. J. (2016). Proteomic Profiling Reveals a Specific Role for Translesion DNA Polymerase η in the Alternative Lengthening of Telomeres. *Cell Reports*, *17*(7), 1858–1871.

Garner, M. M., & Revzin, A. (1981). A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic Acids Research*, *9*(13), 3047–3060.

Gauchier, M., Kan, S., Barral, A., Sauzet, S., Agirre, E., Bonnell, E., Saksouk, N., Barth, T. K., Ide, S., Urbach, S., Wellinger, R. J., Luco, R. F., Imhof, A., & Déjardin, J. (2019). SETDB1-dependent heterochromatin stimulates alternative lengthening of telomeres. *Science Advances*, *5*(5), eaav3673.

Geri, J. B., Oakley, J. V., Reyes-Robles, T., Wang, T., McCarver, S. J., White, C. H., Rodriguez-Rivera, F. P., Parker, D. L., Jr, Hett, E. C., Fadeyi, O. O., Oslund, R. C., & MacMillan, D. W. C. (2020). Microenvironment mapping via Dexter energy transfer on immune cells. *Science (New York, N.Y.)*, *367*(6482), 1091–1097.

Gilmour, D. S., & Lis, J. T. (1985). In vivo interactions of RNA polymerase II with genes of Drosophila melanogaster. *Molecular and Cellular Biology*, *5*(8), 2009–2018.

Ginno, P. A., Burger, L., Seebacher, J., Iesmantavicius, V., & Schübeler, D. (2018). Cell cycle-resolved chromatin proteomics reveals the extent of mitotic preservation of the genomic regulatory landscape. *Nature Communications*, *9*(1), 4048.

Go, C. D., Knight, J. D. R., Rajasekharan, A., Rathod, B., Hesketh, G. G., Abe, K. T., Youn, J.-Y., Samavarchi-Tehrani, P., Zhang, H., Zhu, L. Y., Popiel, E., Lambert, J.-P., Coyaud, É., Cheung, S. W. T., Rajendran, D., Wong, C. J., Antonicka, H., Pelletier, L., Palazzo, A. F., … Gingras, A.-C. (2021). A proximity-dependent biotinylation map of a human cell. *Nature*, *595*(7865), 120–124.

Guo, J., Guo, S., Lu, S., Gong, J., Wang, L., Ding, L., Chen, Q., & Liu, W. (2023). The development of proximity labeling technology and its applications in mammals, plants, and microorganisms. *Cell Communication and Signaling: CCS*, *21*(1), 269.

Gupta, G. D., Coyaud, É., Gonçalves, J., Mojarad, B. A., Liu, Y., Wu, Q., Gheiratmand, L., Comartin, D., Tkach, J. M., Cheung, S. W. T., Bashkurov, M., Hasegan, M., Knight, J. D., Lin, Z.-Y., Schueler, M., Hildebrandt, F., Moffat, J., Gingras, A.-C., Raught, B., & Pelletier, L. (2015). A dynamic protein interaction landscape of the human centrosome-cilium interface. *Cell*, *163*(6), 1484–1499.

Hamley, J. C., Li, H., Denny, N., Downes, D., & Davies, J. O. J. (2023). Determining chromatin architecture with Micro Capture-C. *Nature Protocols*, *18*(6), 1687–1711.

Hansen, A. S., Pustova, I., Cattoglio, C., Tjian, R., & Darzacq, X. (2017). CTCF and cohesin regulate chromatin loop stability with distinct dynamics. *eLife*, *6*. https://doi.org/10.7554/eLife.25776

Han, S., Udeshi, N. D., Deerinck, T. J., Svinkina, T., Ellisman, M. H., Carr, S. A., & Ting, A. Y. (2017). Proximity biotinylation as a method for mapping proteins associated with mtDNA in living cells. *Cell Chemical Biology*, *24*(3), 404–414.

Hellman, L. M., & Fried, M. G. (2007). Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nature Protocols*, *2*(8), 1849–1861.

Henikoff, S., Henikoff, J. G., Sakai, A., Loeb, G. B., & Ahmad, K. (2009). Genome-wide profiling of salt fractions maps physical properties of chromatin. *Genome Research*, *19*(3), 460–469.

Herrmann, C., Avgousti, D. C., & Weitzman, M. D. (2017). Differential salt fractionation of nuclei

to analyze chromatin-associated proteins from cultured mammalian cells. *Bio-Protocol*, *7*(6). https://doi.org/10.21769/BioProtoc.2175

Hershberg, E. A., Camplisson, C. K., Close, J. L., Attar, S., Chern, R., Liu, Y., Akilesh, S., Nicovich, P. R., & Beliveau, B. J. (2021). PaintSHOP enables the interactive design of transcriptome- and genome-scale oligonucleotide FISH experiments. *Nature Methods*, *18*(8), 937–944.

Hobson, B. D., Choi, S. J., Mosharov, E. V., Soni, R. K., Sulzer, D., & Sims, P. A. (2022). Subcellular proteomics of dopamine neurons in the mouse brain. *eLife*, *11*. https://doi.org/10.7554/eLife.70921

Hoffman, M. M., Buske, O., Bilmes, J. A., & Noble, W. S. (2009). Segway: a dynamic Bayesian network method for segmenting genomic data. *Invertebrate Neuroscience: IN*. http://noble.gs.washington.edu/proj/segway/manuscript/description.pdf

Hoffmeyer, K., Raggioli, A., Rudloff, S., Anton, R., Hierholzer, A., Del Valle, I., Hein, K., Vogt, R., & Kemler, R. (2012). Wnt/β-catenin signaling regulates telomerase in stem cells and cancer cells. *Science*, *336*(6088), 1549–1554.

Ho, J. W. K., Alekseyenko, A. A., Kuroda, M. I., & Park, P. J. (2012). Genome-wide mapping of protein-DNA interactions by ChIP-seq. In *Tag-Based Next Generation Sequencing* (pp. 139–151). Wiley-VCH Verlag GmbH & Co. KGaA.

Home - GEO - NCBI. (n.d.). Retrieved July 24, 2024, from https://www.ncbi.nlm.nih.gov/geo/

Hung, V., Udeshi, N. D., Lam, S. S., Loh, K. H., Cox, K. J., Pedram, K., Carr, S. A., & Ting, A. Y. (2016). Spatially resolved proteomic mapping in living cells with the engineered peroxidase APEX2. *Nature Protocols*, *11*(3), 456–475.

Huttlin, E. L., Bruckner, R. J., Navarrete-Perea, J., Cannon, J. R., Baltier, K., Gebreab, F., Gygi, M. P., Thornock, A., Zarraga, G., Tam, S., Szpyt, J., Gassaway, B. M., Panov, A., Parzen, H., Fu, S., Golbazi, A., Maenpaa, E., Stricker, K., Guha Thakurta, S., … Gygi, S. P. (2021). Dual proteome-scale networks reveal cell-specific remodeling of the human interactome.

Cell, *184*(11), 3022–3040.e28.

Huttlin, E. L., Ting, L., Bruckner, R. J., Gebreab, F., Gygi, M. P., Szpyt, J., Tam, S., Zarraga, G., Colby, G., Baltier, K., Dong, R., Guarani, V., Vaites, L. P., Ordureau, A., Rad, R., Erickson, B. K., Wühr, M., Chick, J., Zhai, B., … Gygi, S. P. (2015). The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell*, *162*(2), 425–440.

Ide, S., & Dejardin, J. (2015). End-targeting proteomics of isolated chromatin segments of a mammalian ribosomal RNA gene promoter. *Nature Communications*, *6*, 6674.

Iglesias, N., Paulo, J. A., Tatarakis, A., Wang, X., Edwards, A. L., Bhanu, N. V., Garcia, B. A., Haas, W., Gygi, S. P., & Moazed, D. (2020). Native Chromatin Proteomics Reveals a Role for Specific Nucleoporins in Heterochromatin Organization and Maintenance. *Molecular Cell*, *77*(1), 51–66.e8.

Izumi, H., & Funa, K. (2019). Telomere Function and the G-Quadruplex Formation are Regulated by hnRNP U. *Cells* , *8*(5). https://doi.org/10.3390/cells8050390

Jerkovic, I., & Cavalli, G. (2021). Understanding 3D genome organization by multidisciplinary methods. *Nature Reviews. Molecular Cell Biology*, *22*(8), 511–528.

Ji, X., Dadon, D. B., Abraham, B. J., Lee, T. I., Jaenisch, R., Bradner, J. E., & Young, R. A. (2015). Chromatin proteomic profiling reveals novel proteins associated with histone-marked genomic regions. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(12), 3841–3846.

Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science*, *316*(5830), 1497–1502.

Käll, L., Canterbury, J. D., Weston, J., Noble, W. S., & MacCoss, M. J. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods*, *4*(11), 923–925.

Kalocsay, M. (2019). APEX peroxidase-catalyzed proximity labeling and multiplexed quantitative proteomics. *Methods in Molecular Biology (Clifton, N.J.)*, *2008*, 41–55.

Kaya-Okur, H. S., Wu, S. J., Codomo, C. A., Pledger, E. S., Bryson, T. D., Henikoff, J. G., Ahmad, K., & Henikoff, S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature Communications*, *10*(1), 1930.

Khongkow, P., Gomes, A. R., Gong, C., Man, E. P. S., Tsang, J. W.-H., Zhao, F., Monteiro, L. J., Coombes, R. C., Medema, R. H., Khoo, U. S., & Lam, E. W.-F. (2016). Paclitaxel targets FOXM1 to regulate KIF20A in mitotic catastrophe and breast cancer paclitaxel resistance. *Oncogene*, *35*(8), 990–1002.

Kim, T. H., & Ren, B. (2006). Genome-wide analysis of protein-DNA interactions. *Annual Review of Genomics and Human Genetics*, *7*(1), 81–102.

Kishi, J. Y., Schaus, T. E., Gopalkrishnan, N., Xuan, F., & Yin, P. (2018). Programmable autonomous synthesis of single-stranded DNA. *Nature Chemistry*, *10*(2), 155–164.

Knutson, S. D., Buksh, B. F., Huth, S. W., Morgan, D. C., & MacMillan, D. W. C. (2024). Current advances in photocatalytic proximity labeling. *Cell Chemical Biology*, *31*(6), 1145–1161.

Kochanova, N. Y., Schauer, T., Mathias, G. P., Lukacs, A., Schmidt, A., Flatley, A., Schepers, A., Thomae, A. W., & Imhof, A. (2020). A multi-layered structure of the interphase chromocenter revealed by proximity-based biotinylation. *Nucleic Acids Research*, *48*(8), 4161–4178.

Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D., & Nesvizhskii, A. I. (2017). MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nature Methods*, *14*(5), 513–520.

Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M. N., & Sergushichev, A. (2021). Fast gene set enrichment analysis. In *bioRxiv* (p. 060012). https://doi.org/10.1101/060012

Kufer, T. A., Silljé, H. H. W., Körner, R., Gruss, O. J., Meraldi, P., & Nigg, E. A. (2002). Human TPX2 is required for targeting Aurora-A kinase to the spindle. *The Journal of Cell Biology*, *158*(4), 617–623.

Kurata, M., Wolf, N. K., Lahr, W. S., Weg, M. T., Kluesner, M. G., Lee, S., Hui, K., Shiraiwa, M.,

Webber, B. R., & Moriarity, B. S. (2018). Highly multiplexed genome engineering using CRISPR/Cas9 gRNA arrays. *PloS One*, *13*(9), e0198714.

Kustatscher, G., Grabowski, P., & Rappsilber, J. (2016). Multiclassifier combinatorial proteomics of organelle shadows at the example of mitochondria in chromatin data. *Proteomics*, *16*(3), 393–401.

Kustatscher, G., Hégarat, N., Wills, K. L. H., Furlan, C., Bukowski-Wills, J.-C., Hochegger, H., & Rappsilber, J. (2014). Proteomics of a fuzzy organelle: interphase chromatin. *The EMBO Journal*, *33*(6), 648–664.

Kustatscher, G., Wills, K. L. H., Furlan, C., & Rappsilber, J. (2014). Chromatin enrichment for proteomics. *Nature Protocols*, *9*(9), 2090–2099.

LaBranche, H., Dupuis, S., Ben-David, Y., Bani, M. R., Wellinger, R. J., & Chabot, B. (1998). Telomere elongation by hnRNP A1 and a derivative that interacts with telomeric repeats and telomerase. *Nature Genetics*, *19*(2), 199–202.

Lambert, J.-P., Mitchell, L., Rudner, A., Baetz, K., & Figeys, D. (2009). A novel proteomics approach for the discovery of chromatin-associated protein networks. *Molecular & Cellular Proteomics: MCP*, *8*(4), 870–882.

Lam, S. S., Martell, J. D., Kamer, K. J., Deerinck, T. J., Ellisman, M. H., Mootha, V. K., & Ting, A. Y. (2015). Directed evolution of APEX2 for electron microscopy and proximity labeling. *Nature Methods*, *12*(1), 51–54.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359.

Lee, S.-Y., Cheah, J. S., Zhao, B., Xu, C., Roh, H., Kim, C. K., Cho, K. F., Udeshi, N. D., Carr, S. A., & Ting, A. Y. (2023). Engineered allostery in light-regulated LOV-Turbo enables precise spatiotemporal control of proximity labeling in living cells. *Nature Methods*, *20*(6), 908–917.

Lee, S.-Y., Kang, M.-G., Park, J.-S., Lee, G., Ting, A. Y., & Rhee, H.-W. (2016). APEX fingerprinting reveals the subcellular localization of proteins of interest. *Cell Reports*, *15*(8),

1837–1847.

Lee, T. I., & Young, R. A. (2000). Transcription of eukaryotic protein-coding genes. *Annual Review of Genetics*, *34*, 77–137.

Le Guerroué, F., Eck, F., Jung, J., Starzetz, T., Mittelbronn, M., Kaulich, M., & Behrends, C. (2017). Autophagosomal content profiling reveals an LC3C-dependent piecemeal mitophagy pathway. *Molecular Cell*, *68*(4), 786–796.e6.

Liang, C., Zhang, Z., Chen, Q., Yan, H., Zhang, M., Zhou, L., Xu, J., Lu, W., & Wang, F. (2020). Centromere-localized Aurora B kinase is required for the fidelity of chromosome segregation. *The Journal of Cell Biology*, *219*(2). https://doi.org/10.1083/jcb.201907092

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., & Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (New York, N.Y.)*, *326*(5950), 289–293.

Li, J., Van Vranken, J. G., Pontano Vaites, L., Schweppe, D. K., Huttlin, E. L., Etienne, C., Nandhikonda, P., Viner, R., Robitaille, A. M., Thompson, A. H., Kuhn, K., Pike, I., Bomgarden, R. D., Rogers, J. C., Gygi, S. P., & Paulo, J. A. (2020). TMTpro reagents: a set of isobaric labeling mass tags enables simultaneous proteome-wide measurements across 16 samples. *Nature Methods*, *17*(4), 399–404.

Lin, Z., Schaefer, K., Lui, I., Yao, Z., Fossati, A., Swaney, D. L., Palar, A., Sali, A., & Wells, J. A. (2024). Multiscale photocatalytic proximity labeling reveals cell surface neighbors on and between cells. *Science (New York, N.Y.)*, *385*(6706), eadl5763.

Liu, G., Papa, A., Katchman, A. N., Zakharov, S. I., Roybal, D., Hennessey, J. A., Kushner, J., Yang, L., Chen, B.-X., Kushnir, A., Dangas, K., Gygi, S. P., Pitt, G. S., Colecraft, H. M., Ben-Johny, M., Kalocsay, M., & Marx, S. O. (2020). Mechanism of adrenergic CaV1.2 stimulation revealed by proximity proteomics. *Nature*, *577*(7792), 695–700.

Liu, X., Zhang, Y., Chen, Y., Li, M., Shao, Z., Zhang, M. Q., & Xu, J. (2018). CAPTURE: In situ analysis of chromatin composition of endogenous genomic loci by biotinylated dCas9. *Et Al [Current Protocols in Molecular Biology]*, *123*(1), e64.

Liu, X., Zhang, Y., Chen, Y., Li, M., Zhou, F., Li, K., Cao, H., Ni, M., Liu, Y., Gu, Z., Dickerson, K. E., Xie, S., Hon, G. C., Xuan, Z., Zhang, M. Q., Shao, Z., & Xu, J. (2017). In Situ Capture of Chromatin Interactions by Biotinylated dCas9. *Cell*, *170*(5), 1028–1043.e19.

Luo, Y., Hitz, B. C., Gabdank, I., Hilton, J. A., Kagda, M. S., Lam, B., Myers, Z., Sud, P., Jou, J., Lin, K., Baymuradov, U. K., Graham, K., Litton, C., Miyasato, S. R., Strattan, J. S., Jolanki, O., Lee, J.-W., Tanaka, F. Y., Adenekan, P., … Cherry, J. M. (2020). New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Research*, *48*(D1), D882–D889.

Lynch, P. G., Richards, H., & Wustholz, K. L. (2019). Unraveling the excited-state dynamics of eosin Y photosensitizers using single-molecule spectroscopy. *The Journal of Physical Chemistry. A*, *123*(13), 2592–2600.

Maeshima, K., Iida, S., & Tamura, S. (2021). Physical Nature of Chromatin in the Nucleus. *Cold Spring Harbor Perspectives in Biology*, *13*(5). https://doi.org/10.1101/cshperspect.a040675

Mangeot, P. E., Risson, V., Fusil, F., Marnef, A., Laurent, E., Blin, J., Mournetas, V., Massouridès, E., Sohier, T. J. M., Corbin, A., Aubé, F., Teixeira, M., Pinset, C., Schaeffer, L., Legube, G., Cosset, F.-L., Verhoeyen, E., Ohlmann, T., & Ricci, E. P. (2019). Genome editing in primary cells and in vivo using viral-derived Nanoblades loaded with Cas9-sgRNA ribonucleoproteins. *Nature Communications*, *10*(1), 45.

Martell, J. D., Deerinck, T. J., Sancak, Y., Poulos, T. L., Mootha, V. K., Sosinsky, G. E., Ellisman, M. H., & Ting, A. Y. (2012). Engineered ascorbate peroxidase as a genetically encoded reporter for electron microscopy. *Nature Biotechnology*, *30*(11), 1143–1148.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, *17*(1), 10–12.

Matilainen, O., Quirós, P. M., & Auwerx, J. (2017). Mitochondria and Epigenetics - Crosstalk in Homeostasis and Stress. *Trends in Cell Biology*, *27*(6), 453–463.

McCarty, N. S., Graham, A. E., Studená, L., & Ledesma-Amaro, R. (2020). Multiplexed CRISPR technologies for gene editing and transcriptional regulation. *Nature Communications*, *11*(1), 1281.

McKinley, K. L., & Cheeseman, I. M. (2016). The molecular basis for centromere identity and function. *Nature Reviews. Molecular Cell Biology*, *17*(1), 16–29.

McNulty, S. M., & Sullivan, B. A. (2018). Alpha satellite DNA biology: finding function in the recesses of the genome. *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology*, *26*(3), 115–138.

Merika, M., & Thanos, D. (2001). Enhanceosomes. *Current Opinion in Genetics & Development*, *11*(2), 205–208.

Mick, D. U., Rodrigues, R. B., Leib, R. D., Adams, C. M., Chien, A. S., Gygi, S. P., & Nachury, M. V. (2015). Proteomics of primary cilia by proximity labeling. *Developmental Cell*, *35*(4), 497–512.

Mittler, G., Butter, F., & Mann, M. (2009). A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements. *Genome Research*, *19*(2), 284–293.

Mohammed, H., Taylor, C., Brown, G. D., Papachristou, E. K., Carroll, J. S., & D'Santos, C. S. (2016). Rapid immunoprecipitation mass spectrometry of endogenous proteins (RIME) for analysis of chromatin complexes. *Nature Protocols*, *11*(2), 316–326.

Morrison, O., & Thakur, J. (2021). Molecular complexes at euchromatin, heterochromatin and centromeric chromatin. *International Journal of Molecular Sciences*, *22*(13), 6922.

Myers, S. A., Wright, J., Peckner, R., Kalish, B. T., Zhang, F., & Carr, S. A. (2018). Discovery of proteins associated with a predefined genomic locus via dCas9-APEX-mediated proximity

labeling. *Nature Methods*, *15*(6), 437–439.

Myung, K., Ghosh, G., Fattah, F. J., Li, G., Kim, H., Dutia, A., Pak, E., Smith, S., & Hendrickson, E. A. (2004). Regulation of telomere length and suppression of genomic instability in human somatic cells by Ku86. *Molecular and Cellular Biology*, *24*(11), 5050–5059.

Natsume, T., Kiyomitsu, T., Saga, Y., & Kanemaki, M. T. (2016). Rapid Protein Depletion in Human Cells by Auxin-Inducible Degron Tagging with Short Homology Donors. *Cell Reports*, *15*(1), 210–218.

Navarrete-Perea, J., Yu, Q., Gygi, S. P., & Paulo, J. A. (2018). Streamlined tandem mass tag (SL-TMT) protocol: An efficient strategy for quantitative (phospho)proteome profiling using tandem mass tag-synchronous precursor selection-MS3. *Journal of Proteome Research*, *17*(6), 2226–2236.

Nguyen, H. H., Park, J., Kang, S., & Kim, M. (2015). Surface plasmon resonance: a versatile technique for biosensor applications. *Sensors (Basel, Switzerland)*, *15*(5), 10481–10510.

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). The complete sequence of a human genome. *Science (New York, N.Y.)*, *376*(6588), 44–53.

Ono, T., Fang, Y., Spector, D. L., & Hirano, T. (2004). Spatial and temporal regulation of Condensins I and II in mitotic chromosome assembly in human cells. *Molecular Biology of the Cell*, *15*(7), 3296–3308.

Paek, J., Kalocsay, M., Staus, D. P., Wingler, L., Pascolutti, R., Paulo, J. A., Gygi, S. P., & Kruse, A. C. (2017). Multidimensional Tracking of GPCR Signaling via Peroxidase-Catalyzed Proximity Labeling. *Cell*, *169*(2), 338–349.e11.

Parsons, I. D., Persson, B., Mekhalfia, A., Blackburn, G. M., & Stockley, P. G. (1995). Probing the molecular mechanism of action of co-repressor in the E. coli methionine

repressor-operator complex using surface plasmon resonance (SPR). *Nucleic Acids Research*, *23*(2), 211–216.

Picard. (n.d.). Retrieved July 15, 2024, from https://broadinstitute.github.io/picard/

Politz, J. C. R., Scalzo, D., & Groudine, M. (2013). Something silent this way forms: the functional organization of the repressive nuclear compartment. *Annual Review of Cell and Developmental Biology*, *29*(1), 241–270.

Qi, L. S., Larson, M. H., Gilbert, L. A., Doudna, J. A., Weissman, J. S., Arkin, A. P., & Lim, W. A. (2021). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell*, *184*(3), 844.

Qin, W., Cho, K. F., Cavanagh, P. E., & Ting, A. Y. (2021). Deciphering molecular interactions by proximity labeling. *Nature Methods*, *18*(2), 133–143.

Qiu, W., Xu, Z., Zhang, M., Zhang, D., Fan, H., Li, T., Wang, Q., Liu, P., Zhu, Z., Du, D., Tan, M., Wen, B., & Liu, Y. (2019). Determination of local chromatin interactions using a combined CRISPR and peroxidase APEX2 system. *Nucleic Acids Research*, *47*(9), e52.

Rackham, O., & Filipovska, A. (2022). Organization and expression of the mammalian mitochondrial genome. *Nature Reviews. Genetics*, *23*(10), 606–623.

Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dündar, F., & Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, *44*(W1), W160–W165.

Rangasamy, D., Berven, L., Ridgway, P., & Tremethick, D. J. (2003). Pericentric heterochromatin becomes enriched with H2A.Z during early mammalian development. *The EMBO Journal*, *22*(7), 1599–1607.

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., & Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, *159*(7), 1665–1680.

Rappsilber, J., Ishihama, Y., & Mann, M. (2003). Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Analytical Chemistry*, *75*(3), 663–670.

Rayaprolu, S., Bitarafan, S., Santiago, J. V., Betarbet, R., Sunna, S., Cheng, L., Xiao, H., Nelson, R. S., Kumar, P., Bagchi, P., Duong, D. M., Goettemoeller, A. M., Oláh, V. J., Rowan, M., Levey, A. I., Wood, L. B., Seyfried, N. T., & Rangaraju, S. (2022). Cell type-specific biotin labeling in vivo resolves regional neuronal and astrocyte proteomic differences in mouse brain. *Nature Communications*, *13*(1), 2927.

Rhee, H.-W., Zou, P., Udeshi, N. D., Martell, J. D., Mootha, V. K., Carr, S. A., & Ting, A. Y. (2013). Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. *Science (New York, N.Y.)*, *339*(6125), 1328–1331.

Rodrigues, A., MacQuarrie, K. L., Freeman, E., Lin, A., Willis, A. B., Xu, Z., Alvarez, A. A., Ma, Y., White, B. E. P., Foltz, D. R., & Huang, S. (2023). Nucleoli and the nucleoli–centromere association are dynamic during normal development and in cancer. *Molecular Biology of the Cell*, *34*(4), br5.

Roux, K. J., Kim, D. I., Burke, B., & May, D. G. (2018). BioID: A screen for protein-protein interactions. *Et Al [Current Protocols in Protein Science]*, *91*(1), 19.23.1–19.23.15.

Roux, K. J., Kim, D. I., Raida, M., & Burke, B. (2012). A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. *The Journal of Cell Biology*, *196*(6), 801–810.

Rowley, M. J., & Corces, V. G. (2018). Organizational principles of 3D genome architecture. *Nature Reviews. Genetics*, *19*(12), 789–800.

Ruepp, A., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Stransky, M., Waegele, B., Schmidt, T., Doudieu, O. N., Stümpflen, V., & Mewes, H. W. (2008). CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Research*, *36*(Database issue), D646–D650.

Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., & Arnheim, N. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science (New York, N.Y.)*, *230*(4732), 1350–1354.

Saksouk, N., Barth, T. K., Ziegler-Birling, C., Olova, N., Nowak, A., Rey, E., Simboeck, E., Mateos-Langerak, J., Urbach, S., Reik, W., Torres-Padilla, M.-E., Imhof, A., & Déjardin, J. (2015). Redundant mechanisms to form silent chromatin at pericentromeric regions rely on BEND3 and DNA methylation. *Molecular Cell*, *57*(1), 202.

Santos-Barriopedro, I., van Mierlo, G., & Vermeulen, M. (2021). Off-the-shelf proximity biotinylation for interaction proteomics. *Nature Communications*, *12*(1), 5015.

Sato, S., Morita, K., & Nakamura, H. (2015). Regulation of target protein knockdown and labeling using ligand-directed Ru(bpy)3 photocatalyst. *Bioconjugate Chemistry*, *26*(2), 250–256.

Scelfo, A., Angrisani, A., Grillo, M., Barnes, B. M., Muyas, F., Sauer, C. M., Leung, C. W. B., Dumont, M., Grison, M., Mazaud, D., Garnier, M., Guintini, L., Nelson, L., Esashi, F., Cortés-Ciriano, I., Taylor, S. S., Déjardin, J., Wilhelm, T., & Fachinetti, D. (2024). Specialized replication mechanisms maintain genome stability at human centromeres. *Molecular Cell*, *84*(6), 1003–1020.e10.

Schweppe, D. K., Eng, J. K., Yu, Q., Bailey, D., Rad, R., Navarrete-Perea, J., Huttlin, E. L., Erickson, B. K., Paulo, J. A., & Gygi, S. P. (2020). Full-Featured, Real-Time Database Searching Platform Enables Fast and Accurate Multiplexed Quantitative Proteomics. *Journal of Proteome Research*, *19*(5), 2026–2034.

Schweppe, D. K., Huttlin, E. L., Harper, J. W., & Gygi, S. P. (2018). BioPlex Display: An Interactive Suite for Large-Scale AP-MS Protein-Protein Interaction Data. *Journal of Proteome Research*, *17*(1), 722–726.

Seath, C. P., Burton, A. J., Sun, X., Lee, G., Kleiner, R. E., MacMillan, D. W. C., & Muir, T. W.

(2023). Tracking chromatin state changes using nanoscale photo-proximity labelling. *Nature*, *616*(7957), 574–580.

Sfeir, A., & de Lange, T. (2012). Removal of shelterin reveals the telomere end-protection problem. *Science*, *336*(6081), 593–597.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, *13*(11), 2498–2504.

Silahtaroglu, A. N., Tommerup, N., & Vissing, H. (2003). FISHing with locked nucleic acids (LNA): evaluation of different LNA/DNA mixmers. *Molecular and Cellular Probes*, *17*(4), 165–169.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., & de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature Genetics*, *38*(11), 1348–1354.

Soldi, M., & Bonaldi, T. (2014). The ChroP approach combines ChIP and mass spectrometry to dissect locus-specific proteomic landscapes of chromatin. *Journal of Visualized Experiments: JoVE*, *86*. https://doi.org/10.3791/51220

Stockley, P. G., & Persson, B. (2009). Surface plasmon resonance assays of DNA-protein interactions. *Methods in Molecular Biology (Clifton, N.J.)*, *543*, 653–669.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–15550.

Takano, T., Wallace, J. T., Baldwin, K. T., Purkey, A. M., Uezu, A., Courtland, J. L., Soderblom, E. J., Shimogori, T., Maness, P. F., Eroglu, C., & Soderling, S. H. (2020). Chemico-genetic

discovery of astrocytic control of inhibition in vivo. *Nature*, *588*(7837), 296–302.

Talbert, P. B., & Henikoff, S. (2022). The genetics and epigenetics of satellite centromeres. *Genome Research*, *32*(4), 608–615.

Teo, G., Liu, G., Zhang, J., Nesvizhskii, A. I., Gingras, A.-C., & Choi, H. (2014). SAINTexpress: improvements and additional features in Significance Analysis of INTeractome software. *Journal of Proteomics*, *100*, 37–43.

Thul, P. J., Åkesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., Alm, T., Asplund, A., Björk, L., Breckels, L. M., Bäckström, A., Danielsson, F., Fagerberg, L., Fall, J., Gatto, L., Gnann, C., Hober, S., Hjelmare, M., Johansson, F., … Lundberg, E. (2017). A subcellular map of the human proteome. *Science*, *356*(6340). https://doi.org/10.1126/science.aal3321

Tsherniak, A., Vazquez, F., Montgomery, P. G., Weir, B. A., Kryukov, G., Cowley, G. S., Gill, S., Harrington, W. F., Pantel, S., Krill-Burger, J. M., Meyers, R. M., Ali, L., Goodale, A., Lee, Y., Jiang, G., Hsiao, J., Gerath, W. F. J., Howell, S., Merkel, E., … Hahn, W. C. (2017). Defining a Cancer Dependency Map. *Cell*, *170*(3), 564–576.e16.

Tsue, A. F., Kania, E. E., Lei, D. Q., Fields, R., McGann, C. D., Hershberg, E., Deng, X., Kihiu, M., Ong, S.-E., Disteche, C. M., Kugel, S., Beliveau, B. J., Schweppe, D. K., & Shechner, D. M. (2023). Oligonucleotide-directed proximity-interactome mapping (O-MAP): A unified method for discovering RNA-interacting proteins, transcripts and genomic loci in situ. *bioRxiv : The Preprint Server for Biology*. https://doi.org/10.1101/2023.01.19.524825

Ugur, E., Bartoschek, M. D., & Leonhardt, H. (2020). Locus-Specific Chromatin Proteome Revealed by Mass Spectrometry-Based CasID. *Methods in Molecular Biology* , *2175*, 109–121.

Ugur, E., de la Porte, A., Qin, W., Bultmann, S., Ivanova, A., Drukker, M., Mann, M., Wierer, M., & Leonhardt, H. (2023). Comprehensive chromatin proteomics resolves functional phases of pluripotency and identifies changes in regulatory components. *Nucleic Acids Research*,

51(6), 2671–2690.

Uusküla-Reimand, L., Hou, H., Samavarchi-Tehrani, P., Rudan, M. V., Liang, M., Medina-Rivera, A., Mohammed, H., Schmidt, D., Schwalie, P., Young, E. J., Reimand, J., Hadjur, S., Gingras, A.-C., & Wilson, M. D. (2016). Topoisomerase II beta interacts with cohesin and CTCF at topological domain borders. *Genome Biology*, *17*(1), 182.

van Mierlo, G., Dirks, R. A. M., De Clerck, L., Brinkman, A. B., Huth, M., Kloet, S. L., Saksouk, N., Kroeze, L. I., Willems, S., Farlik, M., Bock, C., Jansen, J. H., Deforce, D., Vermeulen, M., Déjardin, J., Dhaenens, M., & Marks, H. (2019). Integrative proteomic profiling reveals PRC2-dependent epigenetic crosstalk maintains ground-state pluripotency. *Cell Stem Cell*, *24*(1), 123–137.e8.

van Mierlo, G., & Vermeulen, M. (2021). Chromatin proteomics to study epigenetics - challenges and opportunities. *Molecular & Cellular Proteomics: MCP*, *20*, 100056.

van Steensel, B. (2005). Mapping of genetic and epigenetic regulatory networks using microarrays. *Nature Genetics*, *37 Suppl*(S6), S18–S24.

van Steensel, B., & Henikoff, S. (2000). Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nature Biotechnology*, *18*(4), 424–428.

Vermeulen, M., & Déjardin, J. (2020). Locus-specific chromatin isolation. *Nature Reviews. Molecular Cell Biology*, *21*(5), 249–250.

Vermeulen, M., Eberl, H. C., Matarese, F., Marks, H., Denissov, S., Butter, F., Lee, K. K., Olsen, J. V., Hyman, A. A., Stunnenberg, H. G., & Mann, M. (2010). Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers. *Cell*, *142*(6), 967–980.

Villaseñor, R., Pfaendler, R., Ambrosi, C., Butz, S., Giuliani, S., Bryan, E., Sheahan, T. W., Gable, A. L., Schmolka, N., Manzo, M., Wirz, J., Feller, C., von Mering, C., Aebersold, R., Voigt, P., & Baubec, T. (2020). ChromID identifies the protein interactome at chromatin marks. *Nature Biotechnology*, *38*(6), 728–736.

Vizcaíno, J. A., Deutsch, E. W., Wang, R., Csordas, A., Reisinger, F., Ríos, D., Dianes, J. A., Sun, Z., Farrah, T., Bandeira, N., Binz, P.-A., Xenarios, I., Eisenacher, M., Mayer, G., Gatto, L., Campos, A., Chalkley, R. J., Kraus, H.-J., Albar, J. P., … Hermjakob, H. (2014). ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nature Biotechnology*, *32*(3), 223–226.

Wang, H., Wang, L., Zhong, B., & Dai, Z. (2022). Protein splicing of inteins: A powerful tool in synthetic biology. *Frontiers in Bioengineering and Biotechnology*, *10*, 810180.

Xu, W., Zhong, Q., Lin, D., Zuo, Y., Dai, J., Li, G., & Cao, G. (2021). CoolBox: a flexible toolkit for visual analysis of genomics data. *BMC Bioinformatics*, *22*(1), 489.

Yie, J., Merika, M., Munshi, N., Chen, G., & Thanos, D. (1999). The role of HMG I(Y) in the assembly and function of the IFN-beta enhanceosome. *The EMBO Journal*, *18*(11), 3074–3089.

Youn, J.-Y., Dunham, W. H., Hong, S. J., Knight, J. D. R., Bashkurov, M., Chen, G. I., Bagci, H., Rathod, B., MacLeod, G., Eng, S. W. M., Angers, S., Morris, Q., Fabian, M., Côté, J.-F., & Gingras, A.-C. (2018). High-density proximity mapping reveals the subcellular organization of mRNA-associated granules and bodies. *Molecular Cell*, *69*(3), 517–532.e11.

Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: A Journal of Integrative Biology*, *16*(5), 284–287.

Zhang, Q.-S., Manche, L., Xu, R.-M., & Krainer, A. R. (2006). hnRNP A1 associates with telomere ends and stimulates telomerase activity. *RNA* , *12*(6), 1116–1128.

Appendix

Additional graduate school publication

Published in final edited form as:

Nat Methods. 2021 August ; 18(8): 937–944. doi:10.1038/s41592-021-01187-3.

PaintSHOP enables the interactive design of transcriptome- and genome-scale oligonucleotide FISH experiments

Elliot A. Hershberg[#1], **Conor K. Camplisson**[#1], **Jennie L. Close**[2], **Sahar Attar**[1,3], **Ryan Chern**[1], **Yuzhen Liu**[1,4], **Shreeram Akilesh**[3], **Philip R. Nicovich**[2,6], **Brian J. Beliveau**[1,5,*]

[1]Department of Genome Sciences, University of Washington, Seattle, WA, USA

[2]Allen Institute for Brain Science, Seattle, WA, USA

[3]Department of Pathology, University of Washington, Seattle, WA, USA

[4]Molecular and Cellular Biology Program, University of Washington, Seattle, WA, USA

[5]Brotman Baty Institute for Precision Medicine, Seattle, WA, USA

[6]Present address: Cajal Neuroscience Incorporated, Seattle, WA, USA

[7]These authors contributed equally: Elliot A. Hershberg, Conor K. Camplisson

[#] These authors contributed equally to this work.

Abstract

Fluorescence *in situ* hybridization (FISH) allows researchers to visualize the spatial position and quantity of nucleic acids in fixed samples. Recently, considerable progress has been made in developing oligonucleotide (oligo)-based FISH methods that have enabled researchers to study the three-dimensional organization of the genome at super-resolution and visualize the spatial patterns of gene expression for thousands of genes in individual cells. However, there are few existing computational tools to support the bioinformatics workflows necessary to carry out these experiments utilizing oligo FISH probes. Here, we introduce Paint Server and Homology Optimization Pipeline (PaintSHOP), an interactive platform for the design of oligo FISH experiments. PaintSHOP enables researchers to identify probes for their experimental targets efficiently, to incorporate additional necessary sequences such as primer pairs, and to easily generate files documenting library design. PaintSHOP democratizes and standardizes the process of designing complex probe sets for the oligo FISH community.

Editor's summary

[*]Corresponding Author: beliveau@uw.edu.

Author Contributions

E.A.H., C.K.C., and B.J.B. conceived of the study. E.A.H., C.K.C., R.C., and B.J.B. wrote and optimized software code. C.K.C., J.L.C., S.A., and Y.L. performed validation experiments. E.A.H., C.K.C., J.L.C., Sh.A., P.R.N., and B.J.B. conceptualized features of the web application. E.A.H., C.K.C., and B.J.B. wrote the manuscript. All authors edited and approved the manuscript. Sh.A., P.R.N., and B.J.B. supervised the work.

Competing Interests

The authors declare no competing interests.

Paint Server and Homology Optimization Pipeline (PaintSHOP), an interactive platform for the design of oligo FISH experiments, democratizes and standardizes the process of designing complex probe sets for the oligo FISH community.

Introduction

Fluorescence *in situ* hybridization (FISH) is a powerful technique that allows researchers to visualize the distribution of RNA and DNA at single-cell resolution in fixed samples. Since the introduction of *in situ* hybridization in 1969[1] and the subsequent development of FISH[2–4], the method has continued to be used, updated, and refined as new technologies have become available. Recent advances in DNA sequencing and synthesis technologies have spurred the development of a new generation of advanced FISH techniques that utilize oligonucleotide (oligo) libraries as a source of probe material. Oligo FISH probes offer many advantages compared to conventional probes derived directly from genomic material, as they can be optimized to have specific thermodynamic properties, engineered to bind to precisely defined targets while avoiding repetitive sequences, and programmed to utilize a variety of labeling and detection schemes. Researchers have visualized multicopy targets such as repetitive DNA[5–7] and mRNA[8–10] using 'probe sets' composed of one to a few dozen individually synthesized oligo 'probe' species. Approaches have also been developed to leverage complex oligo libraries created by massively parallel array synthesis[11] to perform oligo FISH experiments targeting single-copy chromosomal regions[12–14]. The use of complex oligo libraries has enabled massively multiplexed DNA FISH experiments[15–20] and spatial transcriptomics approaches targeting hundreds to thousands of individual mRNA molecules[21–23].

While many experimental advances have been made using oligo FISH probes, comparatively little progress has been made in developing computational tools that support the design of these probes and probe sets. Several computational tools exist for various related problems such as designing oligo probes against targets such as bacterial rRNA[24,25], large pools of oligo pairs[26–29], padlock probes[30,31], or for microarrays[32]. Previously, we introduced OligoMiner[33], a bioinformatic pipeline developed to address the bottleneck of computationally designing probe sequences for Oligopaints and other oligo-based FISH approaches. Additionally, a bioinformatic resource called iFISH and the related 'ifpd' Python package[34] were created to design "spotting probe" sets that tile along individual chromosomes or to retrieve probe sets targeting individual regions from a collection of pre-discovered probes. Most recently, a MATLAB-based program called ProbeDealer[35] was introduced to support the *de novo* design of probe sets for MERFISH and chromosome walking experiments with a limited number of configurations. However, when it comes to supporting a wide degree of experimental designs and the necessary steps required to generate complete probe sets, these existing tools either require considerable bioinformatics expertise or lack the scalability and flexibility needed to complete the desired design workflows. To our knowledge, no framework exists to solve common problems such as the freeform appending of additional necessary sequences to probe sequences like primer pairs for PCR amplification to support a broad range of experimental designs or the batched construction of probe sets against multiple targets in parallel.

Here, we introduce Paint Server and Homology Optimization Pipeline (PaintSHOP), a platform that enables the interactive design of oligo-based FISH experiments at transcriptome- and genome-scale. PaintSHOP consists of two components: 1) a bioinformatic pipeline and resulting large-scale collection of over 298 million primary oligo probe sequences targeting the genomes and transcriptomes of 9 different experimental organisms 2) an interactive web application that facilitates the automated creation of ready-to-order probe sets against any target in the genome or transcriptome (paintshop.io) with user-specified patterns. The result is an open source, freely available community resource that bridges the gap between probe set design and experiment.

Results

Interactive probe set construction with PaintSHOP.

The computational design of an oligo-based RNA or DNA FISH experiment consists of two distinct phases. In the first phase, the sequence of a genome assembly is processed by a probe discovery algorithm such as OligoArray[14,32,35], the Perl-based OligoPicker used to generate the iFISH probe database[34], the command-line version of OligoMiner[33], or the web-based OligoMinerApp[36] that wraps the OligoMiner scripts into a graphical user interface to identify probe sequences with specified thermodynamic properties that are predicted to be specific (Fig. 1a and Supplementary Table 1). Such design algorithms typically consider length, GC content, melting temperatures, the presence of undesirable sequence stretches such as homopolymeric runs, and the propensity to form secondary structure[32,33]. In the second phase, these 'primary' probes, which often are designed at the scale of entire genomes, are processed into one or more probe sets that can be ordered from a vendor and processed as needed *in vitro* prior to being deployed in a hybridization reaction (Fig. 1a). While the minimum and optimal number of probes in a set will vary depending on the experimental set-up, sample type, and detection optics used, we would generally recommend using sets of >20 probes spanning a target of >400 nucleotides (nt) for RNA FISH[10,37] and sets of >200 probes spanning a target of >10 kilobases (kb) for DNA FISH[14]. Probe sets can also be designed to consider spacing between individual probes[34] (Supplementary Table 1). PaintSHOP collectively supports both design phases. Primary probes can be discovered *de novo* with the PaintSHOP pipeline (github.com/beliveau-lab/PaintSHOP_pipeline), which uses OligoMiner[33] scripts wrapped in Snakemake[38]. The resulting genome-scale probe collections and pre-existing collections discovered with OligoArray, OligoMiner and iFISH are then placed in a cloud-based database which can be accessed by the PaintSHOP web application (Table 1 and Supplementary Figure 1).

The PaintSHOP web application provides an interactive framework for all facets of the probe set construction process. Users can use the PaintSHOP web application to: 1) retrieve the probes covering their RNA/DNA target(s); 2) ensure the probe sequences have the desired strand orientation; 3) consider trimming or unifying the number of probes per target in their sets; 4) append the necessary primers and barcode sequences for their experimental design (Fig. 1b). The PaintSHOP web application is designed to be modular and flexible, enabling a researcher to use only the features required for their experiment. Use cases can range from simply retrieving probes for a single RNA or DNA target to designing multi-

target experiments with complex codebooks[22]. PaintSHOP is designed for users to be able to retrieve probes for their target from one of the 16 hosted genome-scale probe collections (Table 1), to optimize their set by adding or removing probes if necessary, to append primer and barcode sequences from 13 published and newly introduced sets that are designed to be orthogonal to the genomes of commonly used experimental organisms (Supplementary Table 2), and to generate an order file. For retrieving probes, users have the choice between two approaches: 1) RNA probe design and 2) DNA probe design. The RNA probe design option allows a user to either manually enter a list of RefSeq annotations or upload a file of annotations and returns the probes that cover the inputted targets. The DNA design option accepts BED[39] coordinates (chromosome, start coordinate, stop coordinate) either entered manually or uploaded from a file (Supplementary Figure 2). Each of these set construction operations takes only a few seconds, allowing users to interactively implement both simple and complicated design schemes in a matter of minutes or less using our web-based interface.

Probe set construction options

Once users have retrieved the probes for their target(s), they have the option to use several features to optimize the probe set returned. One important feature that assists with set construction is the ability to tune several probe specificity parameters, enabling precise control over the inherent tradeoff between coverage and specificity (Supplementary Figure 3). To this end, we have created a 'Homology Optimization Pipeline' that employs an updated machine learning model to generate a quantitative prediction of both on-target and off-target binding for every candidate probe in all of the probe sets hosted by our web application (Supplementary Figure 4). This machine learning model builds upon our earlier work using a machine learning classifier[33] to approximate the outcome of performing analytical thermodynamic calculations in NUPACK[8,40–42], which provides in-depth information about the predicted behavior of nucleic acid systems but would be prohibitively slow if deployed for genome-scale probe set design. The machine learning model works by estimating the duplexing probability generated by NUPACK pairwise test-tube simulations based on numerical features computed from the pairwise alignments (Supplementary Figure 4). The underlying predictor is an XGBoost[43] Regressor, which was selected as the highest performing model after an automated evaluation of the >100 supervised learning, unsupervised learning, and dataset transformation models present in the Python scikit-learn library[44] by the TPOT[45,46] genetic search algorithm. The model predictions achieved a root-mean-square error (RMSE) score of 0.0657, and the R^2 score between actual and predicted values on the test set (n = 101,704) was 0.974 (Supplementary Figure 4). Importantly, the ability to accurately predict the pairwise duplexing probability without directly computing NUPACK simulations is what makes the large-scale modeling of off-target binding computationally feasible. For example, as part of this work we computed >140,000,000 duplex predictions in less than one day using a computing cluster, the direct computation of which would have taken more than ~32 days with a similar number of continuously running cluster jobs. The ability to make quantitative binding predictions at scale that can be directly compared thus allows the direct comparison of predicted specificities between different probes; this type of comparison is not possible using our previously reported classifier model. On the PaintSHOP web application, researchers can interactively use these

predictions to tune their probe sets by setting a maximum predicted 'Off-Target Score'—this value is the sum of the predicted duplexing probabilities at all off target sites for a given probe, multiplied by 100. Additionally, users can interactively limit the maximum occurrence of the set of 18-mers contained in each probe sequence using 'Max K-mer Count', as duplexes between *k*-mers on this length scale have been predicted to have thermodynamically relevant binding energies in the FISH assay conditions[33]. Users can also select whether or not to allow the inclusion of repeat-masked[47] sequences that have been annotated as being similar to highly reiterated genomic sequences in their probes. Collectively, these parameters enable users to interactively explore how specificity scores impact the number of probes covering their targets through a dynamically updating interface and potentially make trade-offs between the number of probes in the final probe set and their level of predicted specificity. This trade-off can be important as some probe collections hosted by the PaintSHOP web application have probes with ranges of predicted specificities for users to choose from (Fig. 2); in some cases, users may elect to choose a smaller population of highly specific probes, while in others where probe number is limiting or background is less problematic users may opt for all available probes.

While PaintSHOP enables control of specificity parameters to selectively increase target coverage, another common scenario in probe design is that the user has a desired number of oligos per probe set to help unify signal intensity and/or to facilitate barcoding and detection schemes, as is common in applications such as MERFISH[22] and chromosome walking[15,16,18] experiments. In the best-case scenario, there are an excess of suitable probes at each target site. In this case, users can implement the PaintSHOP "trim" feature that simply rank orders the probes for each target based on their predicted specificity and selects from this ranked list in order until the desired number of probes is reached. In other cases, some or all of the targets may have fewer probes than desired. In this case, users can implement the PaintSHOP "unify number" feature. When using "unify number", targets with a surplus of suitable probes have the desired number chosen using exactly the same logic as used in the "trim" feature, while targets with too few probes have their specificity parameters selectively relaxed until either the desired number is reached or the maximum possible number—being less than the target number—are returned.

A core advantage of oligo-based FISH is the precise control over the composition of the probe sequences that it provides, allowing for the incorporation of primer sequences and barcodes. For example, the Oligopaints[14] technology requires the addition of PCR primer sequences to the 5' and 3' ends of the sequence homologous to the FISH target. The incorporation of primers enables the amplification of ssDNA oligo probes from the oligo library. Additionally, it is possible to incorporate region specific primers, allowing for more advanced imaging experiments such as "chromosome tracing"[48] via sequential hybridization. In similar fashion to DNA FISH, advanced RNA FISH methods require the incorporation of multiple sequences in addition to the region homologous to the target into the final probe sequences. For example, spatial transcriptomics technologies such as MERFISH[22] and seqFISH+[49] require the addition of "barcode" sequences and "readout" sequences in order to perform more complicated experiments with many targets requiring successive hybridization. Similarly, SABER[37], a molecular toolkit for FISH signal

amplification and sequential hybridization, requires a Primer Exchange Reaction[50] (PER) primer to be appended to the 3' end of an oligo probe.

In order to accommodate a wide variety of oligo FISH technologies, PaintSHOP includes a flexible user interface for performing appending operations (Supplementary Figure 5). Through the interface it is possible to append up to three sequences to both the 5' and 3' end of each probe. For each sequence appended, the user can choose from a variety of encoding schemes. A detailed documentation of how to use these appending options and all of the other interactive functions on the PaintSHOP web application can be found at https://paintshop.io/user_guide/ and in Supplementary Note 1. In the simplest case, a selected sequence can be appended to all probes in a given probe set (Supplementary Figure 5). For example, a researcher can add the same 5' primer to each probe in the set. PaintSHOP also allows a user to append a unique sequence to the probes for each target in a set (Supplementary Figure 5). Using the same example, this would mean that a unique 5' primer would be appended to the probes for each target in a set. To add addition flexibility, users can also add multiple sequences to a single position per target (Supplementary Figure 5) or specify an entirely custom configuration. It is possible to quickly use PaintSHOP sequences provided for each position, or to upload a custom set of sequences to append. Collectively, the features in the flexible PaintSHOP appending interface can support a wide variety of oligo FISH technologies and experimental designs.

In addition to the general appending functionality, PaintSHOP provides built-in support for appending bridge sequences according to a MERFISH[22] codebook. With this feature, users can upload a set of MHD4 16-bit barcodes to use with their RNA FISH targets. PaintSHOP automatically generates valid MERFISH probe sets by parsing the barcodes provided and handling the incorporation of the encoded bridge sequences into the probe sequences for each target. To demonstrate this feature, PaintSHOP was used with a set of 90 RNA FISH targets and barcodes (Supplementary Data 1, 2) and a set of 16 readout sequences (Supplementary Data 3) to create an order file for a MERFISH experiment (Supplementary Data 4). Using the hg38 newBalance probe set with default PaintSHOP parameters, the targets had an average of 65.9 probes covering them (Supplementary Figure 6). Additionally, as the targets for MERFISH and other highly multiplexed RNA FISH experiments are often chosen based on single-cell RNA sequencing datasets that do not generally have the ability to resolve the specific isoform(s) that map to a given cell, we have introduced 'isoform flattened' versions of the RefSeq annotations for each of the genome assemblies hosted on the PaintSHOP web application. These 'isoform flattened' annotation sets prioritize shared exonic sequence between isoforms (Methods) in order to maximize the chance of detection and only modestly reduce the coverage of the transcriptome when used for probe intersects (Fig. 3a). Collectively, these new resources will streamline the design and practical implementation of spatial transcriptomic experiments.

The final core feature of the PaintSHOP web application is the download functionality provided. Once a researcher has taken advantage of the features necessary for the design of their FISH experiment, they can freely download all the information necessary for a successful order of their designed library. Additionally, we provide several optional download files that promote reproducibility and clear documentation of design decisions

made and primers used. The generation of these files takes only seconds, and their download time would typically be seconds to tens of seconds based on the user's internet connection speed. By providing features for probe retrieval, set balancing and trimming, sequence appending, and the free download of completed designs, PaintSHOP aims to be the first comprehensive resource for the design of complex oligo FISH experiments.

The PaintSHOP probe collections

The PaintSHOP web application hosts four previously published and experimentally validated genome-scale probe collections for the human hg19 and hg38 genome assemblies[14,33,34] (Table 1). In addition, we have also used the PaintSHOP pipeline (http://github.com/beliveau-lab/PaintSHOP_pipeline) to perform *de novo* probe discovery to augment these pre-existing collections with a group of novel collections (Supplementary Figure 7). Specifically, we first performed a systematic search for probe sequence parameters that increased the total number of candidates identified using thermodynamic settings used in our previously reported "Balance" probe collections (T_m 42–47°C)[33], but expanded the probe length range from 6 nt to 8 nt and used the entirety of the human hg38 to optimize the selection of the length range, whereas only 3 Mb of the hg38 chromosome X were used for length range selection in "Balance"[33]. This search culminated in the creation of the "newBalance" probe sets for all of the genome assemblies hosted on the PaintSHOP web application (Table 1). The newBalance probe sets have a new minimum and maximum probe length window of 30-37, and can include repeat-masked[51] bases. These changes allow users to optionally include these sequences in their design if it is necessary to increase the number of possible probe candidates covering their target of interest, which can be particularly valuable for RNA FISH where repetitive sequences are less of a concern and finding enough quality probes can be challenging. We have used NUPACK[52] to predict the secondary structure formation of all 16 genome-scale probe collections hosted by PaintSHOP, giving users an additional feature that can be used to tune the number of probes returned by the web application. We have validated that newBalance probes behave as expected *in situ* by targeting the *ADAMTS5* mRNA in human kidney mesangial cells (Supplementary Figure 8 and Supplementary Table 3).

While tools exist to efficiently generate genome-scale sets of oligo probes[33,34], to the best of our knowledge no comprehensive database exists to connect the coordinates of the discovered probes with the location of reference annotations such as RefSeq. We set out to create this resource in order to greatly reduce the computational difficulty of retrieving probes for single-molecule FISH[10] (smFISH) experiments with multiple targets. Leveraging the ability to perform fast intersection operations on genomic coordinates with BEDTools[53], we developed a flexible approach to intersect any probe set stored in Browser Extendible Data (BED) format[39] with any annotation set for the assembly. This database enables the retrieval of the probes for an arbitrary number of targets with a simple lookup operation rather than performing a large number of manual intersections to retrieve probes for each target. We have used our approach to intersect the 16 genome-scale probe collections hosted by the PaintSHOP web application (Table 1) with the corresponding RefSeq and isoform-flattened annotations for each collection, producing a two transcriptome-scale subsets. We observe high transcriptome (Fig. 3a) and genome coverage (Fig. 3b) across all 16

collections, giving users a range of options to choose from based on their experimental needs. Importantly, our web application is agnostic to which probe set is used, thus allowing users to harness our newly developed newBalance probes, a number existing of publicly available probe collections, and any additional probe collections that may be released at a future date.

Programming multiplexed FISH experiments with PaintSHOP

In order to demonstrate the effectiveness of PaintSHOP for the design of multiplexed FISH experiments, we designed an oligo library to perform 30-target FISH on the human X chromosome. Specifically, we first selected 30 200 kb windows spaced roughly evenly across the entirety of the X chromosome (range 4.8–6.0 Mb separation, mean 5.2 Mb separation; Supplementary Table 4). We created a BED file containing these regions and uploaded it to PaintSHOP webserver using the 'DNA Probe Design' mode (Supplementary Note 2). We then employed the 'Trim' option to automatically select sets of exactly 1,000 probes that mapped to each of our 30 target regions (Supplementary Note 2). Finally, we used the 'Append Sequences' feature to add a distinct 42 nt 'bridge' sequence[37] to each of the 30 probe sets as well as a universal pair of forward and reverse primers to allow all 30 probe sets to be amplified and processed in parallel (Supplementary Note 2 and Supplementary Data 5).

The unique bridge sequence per-target barcoding scheme designed by PaintSHOP allows for each target to be read out using any available fluorescent channel, which in turn enables the programming of color patterns that span the length of the chromosome (Fig. 4a). In order to showcase this capability, we designed two distinct chromosome-scale targeting patterns, with both being visualized by DNA-SABER[37] via the recruitment of a specified SABER sequence at each site that facilitated the docking of ATTO 488, ATTO 565, or Alexa Fluor 647 labeled imager oligos as in our previous chromosome-scale DNA-SABER experiments[37]. In the first pattern, we programmed a three-color "side-by-side" pattern in which the first 10 target regions starting at on the distal end of the p arm were labeled with ATTO 565 (Fig. 4b, magenta), the middle 10 target regions were labeled with Alexa Fluor 647 (Fig. 4b, yellow), and the last 10 target regions ending at the distal end of the q arm were labeled with ATTO 488 (Fig. 4b, cyan). In the second pattern, we programmed a three-color "repeat" pattern where every third target starting with the first was labeled with labeled with ATTO 565 (Fig. 4c, magenta), every third target starting with the second was labeled with labeled with Alexa Fluor 647 (Fig. 4c, yellow), and every third target starting with the third was labeled with ATTO 488 (Fig. 4c, cyan). We performed DNA FISH using these patterns on XX 46N human metaphase chromosome spreads and in both cases observed specific staining patterns that matched our programmed designs (Fig. 4b,c), demonstrating the ability of PaintSHOP to facilitate the design of multiplexed FISH experiments.

Discussion

PaintSHOP is a freely available computational framework that enables the interactive design of transcriptome- and genome-scale oligo-based FISH experiments. PaintSHOP consists of a large database of genome-scale probe collections that are referenced by a dynamic web

application that facilitates probe retrieval, library design, and the creation of complete order files. Our web application provides substantial control over parameters that impact the coverage of FISH targets, providing the flexibility needed for designing probe sets against targets that have fewer optimal probes to start. In addition to the introduction of a new pipeline and web resource, we have developed newBalance probe sets by optimizing our previously reported approach for genome scale probe mining[33]. The newBalance probe sets for the human, mouse, *C. elegans*, Drosophila, zebrafish, Arabidopsis, *S. cerevisiae*, rat, and chicken genomes are freely available through PaintSHOP along with many other sets created by various technologies[33,34]. Our goal for these technologies and resources is to democratize the ability to design the libraries needed for a wide variety of oligo FISH experiments[14,22,37,49] against any target in the genome or transcriptome. We anticipate that PaintSHOP will enable researchers to perform novel FISH experiments interrogating genome organization and the spatial location of gene expression. Going forward, we expect that the set of organisms and genome assemblies supported by PaintSHOP will continue to expand, particularly as long-read sequencing technologies mature and are applied more broadly.

Methods

Probe sets and Genome Assemblies

OligoMiner hg19 and hg38 'balance' probe sets were downloaded from yin.hms.harvard.edu/oligoMiner. The hg19 probe set from the original Oligopaints study[14] was downloaded from oligopaints.hms.harvard.edu. The hg19 iFISH4U 'full 40 mer' probe set was downloaded from ifish4u.org. The ce11, danRer11, dm6, hg19, hg38, mm9, mm10, sacCer3, rn6, galGal5, and galGal6 genome assemblies were downloaded with soft-masking from genome.ucsc.edu. The tair10 genome assembly was downloaded from arabidopsis.org.

Probe Mining Optimization

OligoMiner[33] was downloaded from github.com/brianbeliveau/OligoMiner. The blockParse script was modified to search for probes in soft-masked genome sequences, and to report candidates in soft-masked regions with a special flag. The modified blockParse script was used to mine for probe candidates for each genome assembly with the parameters "–l 20 –L 60 –t 42 –T 47" to identify all possible probes between 20 and 60 nucleotides in length with a Tm between 42 and 47 degrees. A sliding window was used to identify the 8-nucleotide length window with the highest number of candidates. The candidates with the newly optimized settings and specially flagged candidates in soft-masked regions were termed the 'newBalance' probe sets. All probe mining was performed in a Python 2.7 Anaconda environment[54] with the dependencies required for OligoMiner (Python 2.7, Biopython, scikit-learn) on the Department of Genome Science 'Grid' Cluster at the University of Washington.

PaintSHOP Bridge Set Creation

A set of 1,500 G-depleted 46 nt DNA sequences were generated using Python with the following probabilities for incorporating each base: 0.33 for A, 0.33 for T, 0.33 for C, and 0.0 for G. The following substrings were excluded: "AAAA", "TTTT", and "CCC". Each

sequence had a maximum predicted T_m of 42°C[33,55]. Duplexing probabilities were computed for all pairwise combinations of bridge sequences and their reverse complements. 1,065 sequences with a >0.99 probability of on-target duplexing, a maximum off-target duplexing probability of <=0.015 and an average off-target duplexing probability of <=0.0006 were kept. Pairwise duplexing probabilities were computed for the remaining sequences to screen for potential dimerization between bridge sequences. All simulations were performed with the following FISH conditions: 42° C, 50% formamide, 0.390M sodium. 0.65°C per % (vol/vol) formamide was used to scale temperature values in thermodynamic calculations[33,55]. The 1,065 sequences were aligned to the hg38, mm10, dm6, and ce11 references genomes with Bowtie2[56] using the "--very-sensitive-local" settings. The 818 sequences that aligned 0 times were screened for k-mer sequences against all four reference genomes using the OligoMiner[33] kmerFilter.py utility with the settings "-m 18 -k 10". The 800 remaining sequences were used as the new PaintSHOP bridge set.

Machine Learning Model Development

Model construction was performing using the "probe-target" data set described originally in OligoMiner[33]. Briefly the data set consists of 406,814 pairs of "probe" and "target site" sequences. The "target sites" were generated using a combination of in-silico truncation, insertion, and point mutation of the "probe" sequences. The data set contains a Bowtie2[56] alignment score and the thermodynamic duplexing probability computed using NUPACK 3.0[40–42] for each sequence pair. The following numeric features were engineered to represent the key thermodynamic properties of the "probe" and "target site" sequences: length, GC-content, and dinucleotide counts. These features and the Bowtie2 alignment scores were used to build a machine learning model to predict the duplexing probability of the sequence pairs. The data set was randomly split into a training set and a testing set using scikit-learn[44]. Automatic model selection and hyperparameter optimization was performed using TPOT[45,46]. Negative mean squared error was used as the scoring function. After 10 generations with a population size of 100, TPOT converged on a XGBoost[43] regressor. All model selection and hyperparameter optimization was performed using 5-fold cross-validation. The mean squared error (MSE)

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_n)^2$$

And root mean squared error

$$\frac{1}{n}\sum_{i=1}^{n}\sqrt{(Y_i - \hat{Y}_n)^2}$$

were computed for the test set, where $\hat{Y}$ is the predicted value, Y is the actual value, and n is the total number of samples. Least squares regression was also performed perform to evaluate the correlation between model predictions and actual values on the test set. Feature importance was calculated as the number of times a given feature was used to split the data across all trees.

Pipeline Development

All PaintSHOP probe collections were generated using a pipeline implemented with Snakemake[38]. The pipeline takes as input two files: the FASTA sequence and GTF gene annotations for a given genome assembly. Mining of initial "probe candidate" sequences is performed using steps similar to the OligoMiner workflow[33], and then these sequences are scored using the homology optimization pipeline. Briefly, the pipeline aligns "candidates" to their respective genome using Bowtie2[56] with the settings "--very-sensitive-local -k 100", returning up to 100 alignments. Pairwise alignments are reconstructed from the SAM format[57] alignment results using sam2pairwise[58]. The XGBoost[43] regressor is used to predict the duplexing probability of each pairwise alignment returned for all "probe candidates". The "on-target score" is computed as

$$100 \times P(on)$$

where $P(on)$ is the duplexing probability at the intended "target" site, and the "off-target score" is computed as

$$\sum_{1}^{n} 100 \times P(off)_n$$

where $P(off)$ is any alignment at a site other than the "target", and n is the number of "off-target" alignments. Probabilities are scaled to the 0-100 range for user interpretability in the PaintSHOP web application. In addition to both scores, a k-mer statistic is computed. The number of times each 18-mer for a given candidate occurs within its respective genome is computed using Jellyfish[59]. The pipeline returns the occurrence count of the most frequently occurring 18-mer for each candidate. The scored "probe candidates" constitute the collection of genome-wide DNA-FISH probes designed for the input genome assembly, which is one of three major pipeline outputs. The other two outputs are RNA-FISH probe collections, one with isoform-specific targeting resolution, and another "isoform-flattened" set of probes targeting transcript intervals that are maximally shared across isoforms for a given gene. The isoform-resolved probe collection is generated by intersecting the exon coordinates from the GTF annotations file with the genomic coordinates of the designed probes using BEDTools[53] called via the pybedtools[60] Python wrapper and subsequently reverse-complementing probes when needed to account for (+)/(-)-strand annotations. The "isoform-flattened" set is generated using the same approach but first collapsing the annotated coordinates to intervals shared by the maximal number of isoforms. The pipeline is publicly available at: https://github.com/beliveau-lab/PaintSHOP_pipeline.

User Interface

A web application for interactive probe design was built using the Shiny[61] web framework for the R programming language[62]. The back end of the application consists of two databases and a server. One database consists of the pre-computed set intersection of all probes for a given assembly with all UCSC RefSeq or isoform-flattened annotations in the assembly. The set intersection is computed using BEDTools[53]. The other database consists

of all probes returned from the Homology Optimization Pipeline. The front end of the application enables interactive access to both databases. Users can either retrieve the probes targeting a set of RefSeq annotations or retrieve probes from the full database using any genomic coordinate in their assembly of interest. The front end also dynamically generates an interactive table for the user to view their probes, as well as a visualization of the distribution of probes per target using ggplot2[63]. An additional core feature implemented in the front end is the ability to append the sequences necessary for an oligo library, or a SABER[37] experiment. All designs can be downloaded for use directly from the application.

Chromosome X Library Design

A roughly evenly spaced set of 30 windows was upload to PaintSHOP in BED format. This file was uploaded using the DNA Probe Design feature on the PaintSHOP web interface and 'hg38 newBalance' probes were designed with repeats allowed, an off-target cutoff score of 100, and max k-mer cutoff of 5. The balance set feature was used to trim the probe set to 1,000 probes per target. Using the Append Sequences feature, a 5'-outer primer (ATCCTAGCCCATACGGCAATG) and 3'-outer primer (GTATCGTGCAAGGGTGAATGC), as well as per-target 5' inner primers using the included PaintSHOP 5' Inner Primer Set and 3' per-target bridge sequences using the included Kishi et al. 2019 Bridges[37]. The resulting probes were downloaded from PaintSHOP and ordered as an Oligo Pool from Twist Bioscience. Bridge oligos were designed by reverse-complementing the Kishi et al. 2019 Bridges and appending Primer Exchange Reaction (PER) priming sequences and ordered from Integrated DNA Technologies as a 96-well plate. Also see Supplementary Note 2.

RNA SABER-FISH

The conditionally immortalized human mesangial cell line (K29Mes)[64] was obtained from Dr. Moin Saleem (University of Bristol). Cells were cultured in RPMI-1640 medium supplemented with 10% FBS and ITS+ supplement. For propagation, the cells were grown at 33°C (permissive temperature). For experiments, cells were shifted to 37°C (non-permissive temperature) causing degradation of the temperature sensitive SV40 T-antigen and resulting in growth arrest. K29Mes cells were allowed to adhere to 22 x 22 #1.5 coverslips, then rinsed in 1x PBS, fixed in 4% (wt/vol) paraformaldehyde in 1x PBS for 10 minutes at room temperature, then rinsed in 1x PBS. Samples were then permeabilized in 1x PBS + 0.5% (vol/vol) Triton X-100 for 10 minutes at room temperature, then rinsed in 1x PBS + 0.1% (vol/vol) Tween-20. Samples were then transferred to 2x SSC + 1% (vol/vol) Tween-20 + 40% (vol/vol) formamide and incubated for 30 minutes at 43°C in a benchtop air incubator. Samples were then inverted onto parafilm square containing 80 ¼l of pre-warmed hybridization solution consisting of 2x SSC + 1% (vol/vol) Tween-20 + 40% (vol/vol) formamide + 10% (wt/vol) dextran sulfate and 80 ¼l of lyophilized product from a Primer Exchange Reaction (PER) reaction[37,50] performed on a set of 105 oligo probes targeting the *ADAMTS5* mRNA (Supplementary Table 3). The *ADAMTS5* probe pool was purchased from Integrated DNA Technologies and was PER extended for 90 minutes at 37°C with a probe concentration of 1 μM and a hairpin h25.25[37] concentration of 0.5 μM. Hybridization was allowed to proceed overnight (~16 hours) at 43°C in a humidified chamber placed in a benchtop air incubator. Samples were then washed 2 times for 30

minutes each in 2x SSC + 1% (vol/vol) Tween-20 + 40% (vol/vol) formamide for 30 minutes at 43°C, and then twice for 5 minutes each in 2X SSC + 0.1% (vol/vol) Tween-20 at 43°C, and then twice for 5 minutes each in 1x PBS at room temperature. Samples were then inverted onto parafilm containing 100 μl of a secondary hybridization buffer consisting of 0.16x PBS + 8% (wt/vol) dextran sulfate + 0.04% (vol/vol) Tween-20 and an ATTO565-labeled p25*.25* secondary oligo[37] at 0.4 μM and incubated for 30 minutes at 37°C in a benchtop air incubator. Samples were then washed twice for 5 minutes each in 1x PBS + 0.1% (vol/vol) Tween-20 at 37°C, then stained in 0.1 μg/ml DAPI in 1x PBS for 5 minutes at room temperature. Samples were then washed for 5 minutes in 1x PBS at room temperature, then inverted onto microscope slides containing ProLong Gold Antifade Mountant which cured overnight at room temperature prior to imaging. Images were captured using a Leica SP8X laser scanning confocal microscope using a 63x oil N.A. 1.40 Plan Apo objective lens controlled using Leica LASX Expert software. Images were processed in ImageJ + Fiji[65,66] and Adobe Photoshop.

ChrX library amplification and ssDNA probe synthesis

Raw library material was resuspended to a concentration of 20 ng / μl using 10 mM Tris, pH 8.0. The first PCR mix contained 34 μl dH_2O, 10 μl 5X Phusion HF Buffer, 1.5 μl 10 mM dNTP Mix, 1.5 μl 10 μM F primer, 1.5 μl 10 μM R primer, 1.0 μl resuspended oligo pool, 0.5 μl Phusion DNA Polymerase (2 U / μl) for a total volume of 50 μl. The thermal cycler program comprised an initial denaturation at 95°C for 3 minutes, followed by 12 cycles of 98°C for 20 seconds, 60°C for 15 seconds, 72°C for 15 seconds, and a final extension at 72°C for 1 minute followed by a 4°C hold. The first PCR product was purified using a Zymo DNA Clean and Concentrator-5 (DCC-5) kit according to the manufacturer's standard protocol. A dilution of the first PCR product at 20 pg / μl was prepared as a template for the second PCR. The second PCR mix contained 27 μl dH_2O, 10 μl 5X Phusion HF Buffer, 1.5 μl 10 mM dNTP Mix, 5.0 μl 10 μM F Primer, 5.0 μl 10 μM R Primer, 1.0 μl diluted DNA template, 0.5 μl Phusion DNA Polymerase (2 U / μl) for a total volume of 50 μl. The thermal cycler program was the same as before but with 18 cycles instead of 12. The second PCR product was purified as before. RNA was synthesized using the NEB HiScribe T7 Quick High Yield RNA Synthesis Kit with a modified reaction mix containing 8 μl dH_2O, 2.5 μl diluted DNA template, 15.0 μl NTP Buffer Mix, 3 μl T7 RNA Polymerase Mix, 1.5 μl RNaseOUT. The reaction was incubated at 37°C overnight. Enzymatic digestion of the DNA template and precipitation of the RNA using the included Lithium chloride solution were both carried out according to the manufacturer's standard protocol. The reverse transcription reaction contained 55 μl synthesized RNA, 30 μl 5X RT Buffer, 48 μl 10 mM dNTP Mix, 10 μl 100 μM RT Primer, 3 μl RNaseOUT, and 4 μl Maxima H Minus Reverse Transcriptase (200 U / μl) for a total volume of 150 μl, which was split into four 37.5 μl reactions. The reactions were incubated at 50°C for 2 hours and then at 80 °C for 5 minutes. RNA templates were degraded enzymatically by adding 1 μl RNase to each reaction and incubating at 37 °C for 1 hour. To precipitate the final ssDNA probes, 0.1 volumes of 5 M Ammonium acetate, 0.02 volumes of 2% (wt/vol) Glycogen, and 3.0 volumes of 100% (vol/vol) Ethanol were added to the reverse transcription reaction mixture. The resulting mixture was incubated for 15 minutes at −20°C, followed by 10 minutes of centrifugation at 10,000 x *g* at 4°C. The pellet was washed using 750 μl of 70% (vol/vol) and centrifuged again as

before. The pellet was dried for 3 minutes at room temperature and resuspended using 250 μl nuclease free water. The concentration was measured using a NanoDrop spectrophotometer and a 10 μM probe stock solution was prepared and stored at −20°C.

DNA SABER-FISH on spread metaphase chromosomes

Bridge oligos were extended using the primer exchange reaction (PER) as previously described[37] with an extension time of two hours. PER-extended bridge oligos (60 pmol total) and amplified ssDNA primary probes (60 pmol total) were dried using a SpeedVac concentrator. The dried oligos were resuspended using 25 μl of an ISH solution containing 12.5 μl Formamide, 5.0 μl 50% Dextran sulfate, 4.0 μl dH2O, 2.5 μl 20X SSC, 1.0 μl RNase A (10 mg / ml). Human metaphase chromosome spreads (XX 46N, Applied Genetics Laboratories) were denatured in 70% (vol/vol) Formamide in 2X SSCT (2X SSC with 1% (vol/vol) Tween-20) at 70°C (90 seconds) and then transferred to ice-cold 70% (vol/vol) ethanol (5 minutes), to 90% (vol/vol) ethanol (5 minutes), and to 100% ethanol (5 minutes). Slides were air dried and the primary hybridization mix was added and sealed underneath a coverslip with rubber cement. Slides were placed in a humidified chamber and incubated in an oven at 37°C overnight. After hybridization, coverslips were removed and slides were washed in 2X SSCT at 60°C (15 minutes) and in 2X SSCT at room temperature (2 × 5 minutes). A 25 μl secondary hybridization solution comprising 11.0 ul dH2O, 5.0 ul 5X PBS-T (5X PBS with 0.5% (vol/vol) Tween-20), 5.0 ul 50% Dextran sulfate, and 4.0 ul 10 uM total fluorescently labeled secondary oligos. The secondary hybridization mix was added and sealed underneath a coverslip with rubber cement. Slides were placed in a humidified chamber and incubated in an oven at 37°C (1 hour). After hybridization, coverslips were removed and slides were washed in 1X PBS-T (1X PBS with 0.1% (vol/vol) Tween-20) at 37°C (3 x 15 minutes). Slides were imaged on a custom microscopy system consisting of a Nikon Eclipse Ti2 body and an attached Yokogawa CSU-W1 SoRa spinning disc confocal unit. 405 nm, 488 nm, 561 nm, or 640 nm laser excitation was emitted at 30–40% of maximal intensity inside of a Nikon LUNF 405/488/561/640NM 1F commercial launch and coupled into a single-mode optical fiber, which delivered the excitation light into the CSU-W1 SoRa unit. Excitation light was then directed through a microlens array disc and a ‘SoRa’ disc containing 50 μm pinholes and directed to the rear aperture of a 100x N.A. 1.49 Apo TIRF oil immersion objective lens by a prism in the base of the Ti2. Emission light was collected by the same objective and passed via a prism in the base of the Ti2 back into the SoRa unit, where it was relayed by a 1x lens (fields of view) or 2.8x lens (spreads) through the pinhole disc and then directed into the emission path by a quad-band dichroic mirror (Semrock Di01-T405/488/568/647-13x15x0.5). Emission light was then spectrally filtered by one of four single-bandpass filters (DAPI: Chroma ET455/50M; ATTO 488: Chroma ET525/36M; ATTO 565: Chroma ET605/50M; Alexa Fluor 647: Chroma ET705/72M) and focused by a 1x relay lens onto an Andor Sona 4.2B-11 camera with a physical pixel size of 11 μm, resulting in an effective pixel size of 110 nm (fields of view) or 39.3 nm (spreads). The Sona was operated in 16-bit mode with rolling shutter readout and a 300 ms exposure time. Acquisition was controlled by Nikon Elements software. Images were processed in ImageJ + Fiji[65,66] and Adobe Photoshop.

Code Availability

The source code for the PaintSHOP web application is available as Supplementary Software 1 and at https://github.com/beliveau-lab/PaintSHOP. The source code for the Homology Optimization Pipeline is available as Supplementary Software 2 and at https://github.com/beliveau-lab/PaintSHOP_pipeline.

Data Availability

The original 'OligoPaints 2012 hg19" genome-scale probe collection was downloaded from https://oligopaints.hms.harvard.edu/sites/oligopaints.hms.harvard.edu/files/complete-genome-files/hg19.tar.gz. The original OligoMiner hg19 'balance' genome-scale probe collection was downloaded from https://yin.hms.harvard.edu/oligoMiner/probe_seqs/hg19/hg19b.tar.gz. The original OligoMiner hg38 'balance' genome-scale probe collection was downloaded from https://yin.hms.harvard.edu/oligoMiner/probe_seqs/hg38/hg38b.tar.gz. The original 'Full 40-mer' iFISH4u hg19 genome-scale probe collection was downloaded from http://ifish4u.org/custom/dbdownload/hg19.gz. The human hg19 genome assembly was downloaded from https://hgdownload.soe.ucsc.edu/goldenPath/hg19/bigZips/hg19.fa.gz. The human hg38 genome assembly was downloaded from https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz. The mouse mm9 genome assembly was downloaded from https://hgdownload.soe.ucsc.edu/goldenPath/mm9/bigZips/mm9.fa.gz. The mouse mm10 genome assembly was downloaded from https://hgdownload.soe.ucsc.edu/goldenPath/mm10/bigZips/mm10.fa.gz. The *C. elegans* ce11 genome assembly was downloaded from https://hgdownload.soe.ucsc.edu/goldenPath/ce11/bigZips/ce11.fa.gz. The *D. melanogaster* dm6 genome assembly was downloaded from https://hgdownload.soe.ucsc.edu/goldenPath/dm6/bigZips/dm6.fa.gz. The zebrafish danRer11 genome assembly was downloaded from https://hgdownload.soe.ucsc.edu/goldenPath/danRer11/bigZips/danRer11.fa.gz. The *A. thaliana* TAIR10 genome assembly was downloaded from https://www.arabidopsis.org/download_files/Genes/TAIR10_genome_release/TAIR10_chromosome_files/TAIR10_chr_all.fas. The *S. cerevisiae* sacCer3 genome assembly was downloaded from https://hgdownload.soe.ucsc.edu/goldenPath/sacCer3/bigZips/sacCer3.fa.gz. The rat rn6 genome assembly was downloaded from https://hgdownload.soe.ucsc.edu/goldenPath/rn6/bigZips/rn6.fa.gz. The chicken galGal5 genome assembly was downloaded from https://hgdownload.soe.ucsc.edu/goldenPath/galGal5/bigZips/galGal5.fa.gz. The chicken galGal6 genome assembly was downloaded from https://hgdownload.soe.ucsc.edu/goldenPath/galGal6/bigZips/galGal6.fa.gz. All genome-scale probe collections, primer sequences, bridge sequences, SABER-associated sequences, and transcriptome intersects hosted on paintshop.io are available to download from https://github.com/beliveau-lab/PaintSHOP_resources repository. All repositories are available under the MIT license. Raw and processed microscopy images are available upon request.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors thank G. Nir, D. Shechner, A. Tsue, H. Nguyen, J.Y. Kishi, and J. Harke for helpful feedback during the beta testing phase of PaintSHOP, N. Peters and D. Fong for assistance with microscopy, the Genome Sciences IT team for technical assistance, and members of the Beliveau lab for feedback on the written manuscript. We also thank S. Lapan and E. West for productive discussions that inspired aspects of this work. This work was supported by a Damon Runyon Dale F. Frey Breakthrough Award (to B.J.B.), the National Institutes of Health (under grant 1R35GM137916 to B.J.B.), and the DiaCOMP consortium (under grant 19AU3987 to S.A. and B.J.B.). Imaging on the University of Washington W.M. Keck Center Lecia SP8X confocal microscopy was enabled by funding from the NIH (S10 OD016240). We would also like to thank the Allen Institute for Brain Science founder, Paul G. Allen, for his vision, encouragement, and support.

References

1. Pardue ML & Gall JG Molecular hybridization of radioactive DNA to the DNA of cytological preparations. Proc. Natl. Acad. Sci. U. S. A (1969) doi:10.1073/pnas.64.2.600.
2. Rudkin GT & Stollar BD High resolution detection of DNA-RNA hybrids in situ by indirect immunofluorescence [29]. Nature (1977) doi:10.1038/265472a0.
3. Bauman JGJ, Wiegant J, Borst P & van Duijn P A new method for fluorescence microscopical localization of specific DNA sequences by in situ hybridization of fluorochrome-labelled RNA. Exp. Cell Res. (1980) doi:10.1016/0014-4827(80)90087-7.
4. Langer-Safer PR, Levine M & Ward DC Immunological methods for mapping genes on Drosophila polytene chromosomes. Proc. Natl. Acad. Sci. U. S. A (1982) doi:10.1073/pnas.79.14.4381.
5. Moyzis RK et al. A highly conserved repetitive DNA sequence, (TTAGGG)(n), present at the telomeres of human chromosomes. Proc. Natl. Acad. Sci. U. S. A (1988) doi:10.1073/pnas.85.18.6622.
6. Matera AG & Ward DC Oligonucleotide probes for the analysis of specific repetitive dna sequences by fluorescence in situ hybridization. Hum. Mol. Genet 1, 535–539 (1992). [PubMed: 1307254]
7. Dernburg AF et al. Perturbation of nuclear architecture by long-distance chromosome interactions. Cell (1996) doi:10.1016/S0092-8674(00)81240-4.
8. Dirks RW et al. Simultaneous detection of different mRNA sequences coding for neuropeptide hormones by double in situ hybridization using FITC- and biotin-labeled oligonucleotides. J. Histochem. Cytochem (1990) doi:10.1177/38.4.2108203.
9. Femino AM, Fay FS, Fogarty K & Singer RH Visualization of single RNA transcripts in situ. Science (1998) doi:10.1126/science.280.5363.585.
10. Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A & Tyagi S Imaging individual mRNA molecules using multiple singly labeled probes. Nat. Methods (2008) doi:10.1038/nmeth.1253.
11. Kosuri S & Church GM Large-scale de novo DNA synthesis: Technologies and applications. Nature Methods (2014) doi:10.1038/nmeth.2918.
12. Yamada NA et al. Visualization of fine-scale genomic structure by oligonucleotide-based high-resolution FISH. Cytogenet. Genome Res. (2011) doi:10.1159/000322717.
13. Boyle S, Rodesch MJ, Halvensleben HA, Jeddeloh JA & Bickmore WA Fluorescence in situ hybridization with high-complexity repeat-free oligonucleotide probes generated by massively parallel synthesis. Chromosome Res. (2011) doi:10.1007/s10577-011-9245-0.
14. Beliveau BJ et al. Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. Proc. Natl. Acad. Sci. U. S. A (2012) doi:10.1073/pnas.1213818110.
15. Wang S et al. Spatial organization of chromatin domains and compartments in single chromosomes. Science (2016) doi:10.1126/science.aaf8084.
16. Bintu B et al. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. Science (2018) doi:10.1126/science.aau1783.
17. Cardozo Gizzi AM et al. Microscopy-Based Chromosome Conformation Capture Enables Simultaneous Visualization of Genome Organization and Transcription in Intact Organisms. Mol. Cell (2019) doi:10.1016/j.molcel.2019.01.011.
18. Mateo LJ et al. Visualizing DNA folding and RNA in embryos at single-cell resolution. Nature (2019) doi:10.1038/s41586-019-1035-4.

19. Su J-H, Zheng P, Kinrot SS, Bintu B & Zhuang X Genome-Scale Imaging of the 3D Organization and Transcriptional Activity of Chromatin. Cell 182, 1641–1659.e26 (2020). [PubMed: 32822575]
20. Takei Y et al. Integrated spatial genomics reveals global architecture of single nuclei. Nature 1–7 (2021) doi:10.1038/s41586-020-03126-2.
21. Levesque MJ & Raj A Single-chromosome transcriptional profiling reveals chromosomal gene expression regulation. Nat. Methods (2013) doi:10.1038/nmeth.2372.
22. Chen KH, Boettiger AN, Moffitt JR, Wang S & Zhuang X Spatially resolved, highly multiplexed RNA profiling in single cells. Science (2015) doi:10.1126/science.aaa6090.
23. Shah S et al. Dynamics and Spatial Genomics of the Nascent Transcriptome by Intron seqFISH. Cell (2018) doi:10.1016/j.cell.2018.05.035.
24. Pernthaler J, Glöckner FO, Schönhuber W & Amann R Fluorescence in situ hybridization (FISH) with rRNA-targeted oligonucleotide probes. Methods in Microbiology (2001) doi:10.1016/s0580-9517(01)30046–6.
25. Yilmaz LS, Parnerkar S & Noguera DR MathFISH, a web tool that uses thermodynamics-based mathematical models for in silico evaluation of oligonucleotide probes for fluorescence in situ hybridization. Appl. Environ. Microbiol (2011) doi:10.1128/AEM.01733-10.
26. Rogan PK, Cazcarro PM & Knoll JHM Sequence-based design of single-copy genomic DNA probes for fluorescence in situ hybridization. Genome Res. (2001) doi:10.1101/gr.171701.
27. Navin N et al. PROBER: Oligonucleotide FISH probe design software. Bioinformatics (2006) doi:10.1093/bioinformatics/btl273.
28. Nedbal J, Hobson PS, Fear DJ, Heintzmann R & Gould HJ Comprehensive FISH Probe Design Tool Applied to Imaging Human Immunoglobulin Class Switch Recombination. PLoS ONE (2012) doi:10.1371/journal.pone.0051675.
29. Bienko M et al. A versatile genome-scale PCR-based pipeline for high-definition DNA FISH. Nat. Methods (2013) doi:10.1038/nmeth.2306.
30. Baner J Parallel gene analysis with allele-specific padlock probes and tag microarrays. Nucleic Acids Res. (2003) doi:10.1093/nar/gng104.
31. Stenberg J, Nilsson M & Landegren U ProbeMaker: An extensible framework for design of sets of oligonucleotide probes. BMC Bioinformatics (2005) doi:10.1186/1471-2105-6-229.
32. Rouillard JM, Zuker M & Gulari E OligoArray 2.0: Design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. Nucleic Acids Res. (2003) doi:10.1093/nar/gkg426.
33. Beliveau BJ et al. OligoMiner provides a rapid, flexible environment for the design of genome-scale oligonucleotide in situ hybridization probes. Proc. Natl. Acad. Sci. U. S. A. (2018) doi:10.1073/pnas.1714530115.
34. Gelali E et al. iFISH is a publically available resource enabling versatile DNA FISH to study genome architecture. Nat. Commun. (2019) doi:10.1038/s41467-019-09616-w.
35. Hu M et al. ProbeDealer is a convenient tool for designing probes for highly multiplexed fluorescence in situ hybridization. Sci. Rep 10, 22031 (2020). [PubMed: 33328483]
36. Passaro M et al. OligoMinerApp: a web-server application for the design of genome-scale oligonucleotide in situ hybridization probes through the flexible OligoMiner environment. Nucleic Acids Res. 48, W332–W339 (2020). [PubMed: 32313927]
37. Kishi JY et al. SABER amplifies FISH: enhanced multiplexed imaging of RNA and DNA in cells and tissues. Nat. Methods (2019) doi:10.1038/s41592-019-0404-0.
38. Köster J & Rahmann S Snakemake-a scalable bioinformatics workflow engine. Bioinformatics (2012) doi:10.1093/bioinformatics/bts480.
39. Casper J et al. The UCSC Genome Browser database: 2018 update. Nucleic Acids Res. (2018) doi:10.1093/nar/gkx1020.
40. Dirks RM & Pierce NA A partition function algorithm for nucleic acid secondary structure including pseudoknots. J. Comput. Chem (2003) doi:10.1002/jcc.10296.
41. Dirks RM & Pierce NA An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots. J. Comput. Chem (2004) doi:10.1002/jcc.20057.
42. Dirks RM, Bois JS, Schaeffer JM, Winfree E & Pierce NA Thermodynamic analysis of interacting nucleic acid strands. SIAM Rev. (2007) doi:10.1137/060651100.

43. Chen T & Guestrin C XGBoost: A scalable tree boosting system. in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016). doi:10.1145/2939672.2939785.
44. Pedregosa F et al. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res (2011).
45. Olson RS et al. Automating biomedical data science through tree-based pipeline optimization. in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2016). doi:10.1007/978-3-319-31204-0_9.
46. Olson RS, Bartley N, Urbanowicz RJ & Moore JH Evaluation of a tree-based pipeline optimization tool for automating data science. in GECCO 2016 - Proceedings of the 2016 Genetic and Evolutionary Computation Conference (2016). doi:10.1145/2908812.2908918.
47. Smit A, Hubley R & Green P RepeatMasker Open-3.0. RepeatMasker Open-3.0 (1996).
48. Nir G et al. Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling. PLoS Genet. (2018) doi:10.1371/journal.pgen.1007872.
49. Eng CHL et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. Nature (2019) doi:10.1038/s41586-019-1049-y.
50. Kishi JY, Schaus TE, Gopalkrishnan N, Xuan F & Yin P Programmable autonomous synthesis of single-stranded DNA. Nat. Chem (2017) doi:10.1038/nchem.2872.
51. Smit A, Hubley R & Green P RepeatMasker Open-4.0. 2013-2015. http://www.repeatmasker.org (2013).
52. Fornace ME, Porubsky NJ & Pierce NA A Unified Dynamic Programming Framework for the Analysis of Interacting Nucleic Acid Strands: Enhanced Models, Scalability, and Speed. ACS Synth. Biol 9, 2665–2678 (2020). [PubMed: 32910644]
53. Quinlan AR & Hall IM BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics (2010) doi:10.1093/bioinformatics/btq033.
54. Anaconda. Anaconda Software Distribution. Computer software (2014).
55. Cock PJA et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25, 1422–1423 (2009). [PubMed: 19304878]
56. Langmead B & Salzberg SL Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012). [PubMed: 22388286]
57. Li H et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics (2009) doi:10.1093/bioinformatics/btp352.
58. LaFave & Burgess. sam2pairwise version 1.0.0. Zenodo doi:10.5281/zenodo.11377.
59. Marçais G & Kingsford C A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics (2011) doi:10.1093/bioinformatics/btr011.
60. Dale RK, Pedersen BS & Quinlan AR Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. Bioinformatics 27, 3423–3424 (2011). [PubMed: 21949271]
61. Chang W, Cheng J, Allaire JJ, Xie Y & McPherson J shiny: Web Application Framework for R. (2019).
62. R Core Team. R: A Language and Environment for Statistical Computing. (2019).
63. Wickham H ggplot2: Elegant Graphics for Data Analysis. (Springer-Verlag New York, 2016).
64. Establishment of conditionally immortalized human glomerular mesangial cells in culture, with unique migratory properties | American Journal of Physiology-Renal Physiology. https://journals.physiology.org/doi/full/10.1152/ajprenal.00589.2010?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub++0pubmed&.
65. Schneider CA, Rasband WS & Eliceiri KW NIH Image to ImageJ: 25 years of image analysis. Nat. Methods 9, 671–675 (2012). [PubMed: 22930834]
66. Schindelin J et al. Fiji: an open-source platform for biological-image analysis. Nat. Methods 9, 676–682 (2012). [PubMed: 22743772]

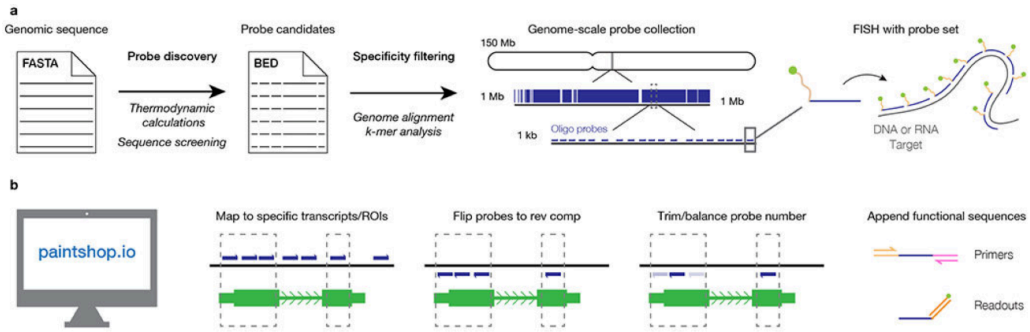


Fig. 1 |. The PaintSHOP workflow.

a, Overview of the genome-scale FISH probe collection design process. **b**, Overview the probe set creation functionality of the PaintSHOP web application.

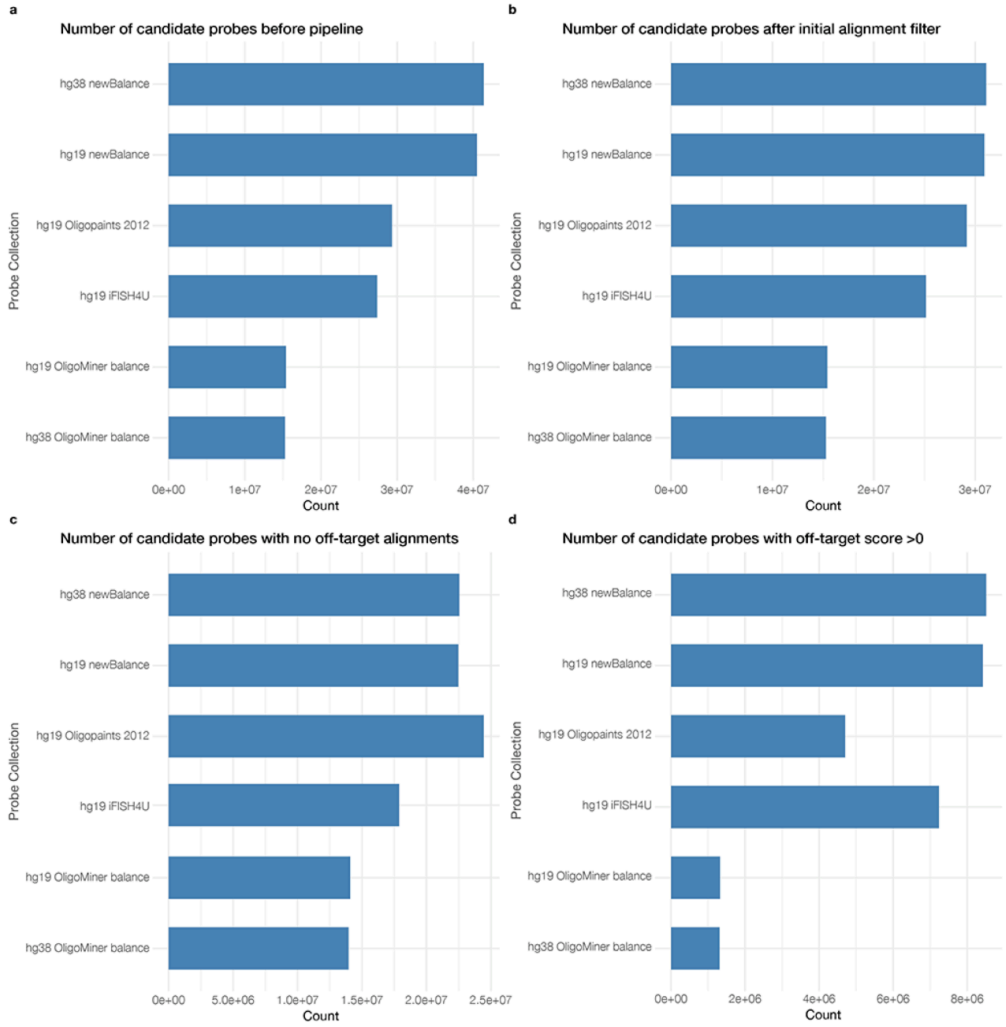


Fig. 2 |. Probe counts for each human probe collection included in PaintSHOP.
a, Counts of the number of candidate probes before any downstream processing was performed. **b**, The number of candidates after filtering for probes with greater than 100 off-target alignments. **c**, The number of remaining probes with no off-target alignments. **d**, The number of remaining probes with between 1 and 100 off-target alignments.

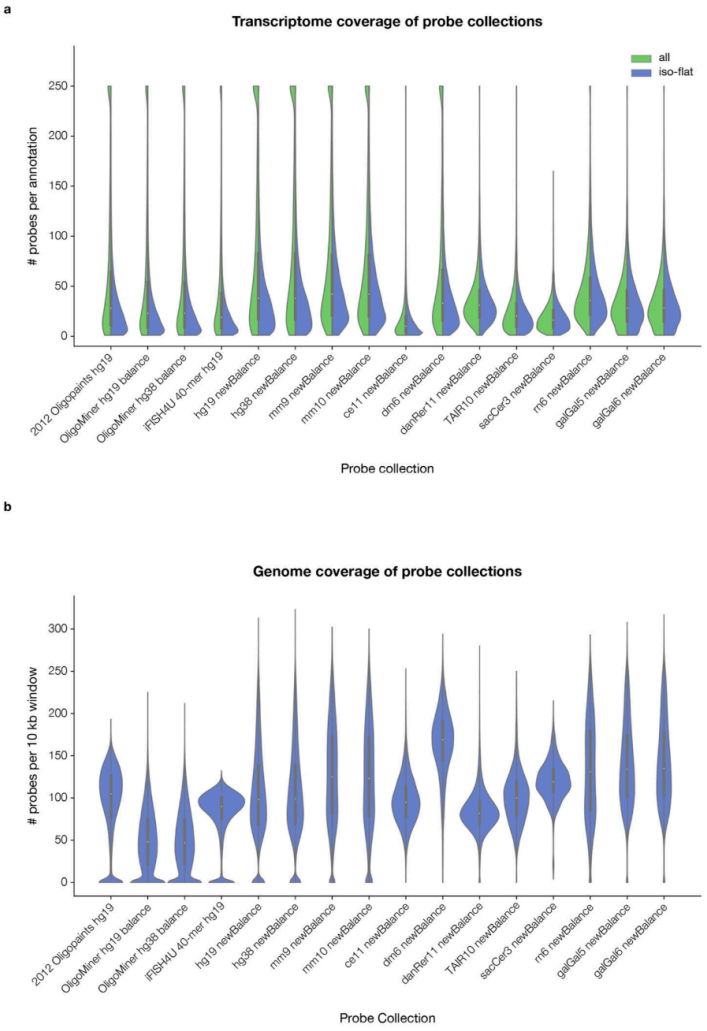


Fig. 3 |. Transcriptome and genome coverage of probe collections.

a, The number of probes per RefSeq transcript annotation ("all"; green, left side of violins) or isoform-flattened RefSeq gene annotation ("iso-flat"; blue, right side of violins) in the genome-scale probe collections hosted by PaintSHOP. Annotations with >250 probes were plotted with a value of 250 to simplify presentation. **b**, The number of probes per 10-kilobase window in the genome-scale probe collections hosted by PaintSHOP. Each genome was split into adjacent (i.e. non-overlapping) windows. Violin plots in a, b show a kernel density estimation (blue, green) along with traditional boxplot elements: median—white dot,

quartiles 2 and 3—thick black box, quartile 1 value – 1.5 * (quartile 3 value – quartile 1 value)—lower bound of black line, quartile 3 value + 1.5 * (quartile 3 value – quartile 1 value)—upper bound of black line. The minimum value of the kernel density estimation was set to 0 for display purposes. In a, annotations with >250 probes mapping to them had their value set to '250' for display purposes.

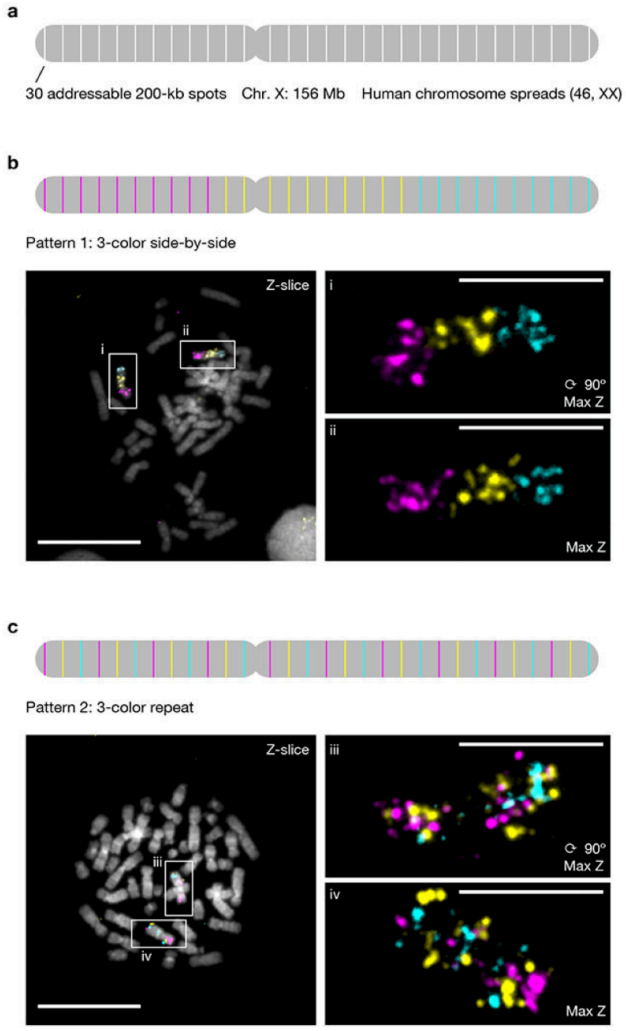


Fig. 4 | Multiplexed DNA FISH programmed by PaintSHOP.
a, Overview of 30-target human X chromosome library design. Each site is individually addressable. **b**, A three-color "side-by-side" pattern imaged on 46 XX human metaphase chromosome spreads. **c**, A three-color "repeat" pattern imaged on 46 XX human metaphase chromosome spreads. Scale bars, 5 μm (spreads) or 20 μm (fields of view). Each pattern was visualized in 3 independent samples and yielded similar results.

Table 1:

Description of the genome-scale probe collections hosted on the PaintSHOP web application.

Collection name	Organism	Genome assembly	Number of probes	Probe length (nt)	Probe T_m range (°C)	Probe %GC (mean ± sd)	Reference
2012 Oligopaints hg19	Human	hg19	29,147,070	32	34.2–49.7	43.6 ± 7.1	Beliveau et al. 2012
OligoMiner hg19 balance	Human	hg19	15,411,378	35–41	42.0–47.0	46.6 ± 6.2	Beliveau et al. 2018
OligoMiner hg38 balance	Human	hg38	15,271,724	35–41	42.0–47.0	46.6 ± 6.2	Beliveau et al. 2018
iFISH4U 40-mer hg19	Human	hg19	25,127,787	40	34.1–59.8	49.6 ± 9.3	Gelali et al. 2019
hg19 newBalance	Human	hg19	32,139,623	30–37	42.0–47.0	50.4 ± 6.8	This study
hg38 newBalance	Human	hg38	32,307,382	30–37	42.0–47.0	50.4 ± 6.8	This study
mm9 newBalance	Mouse	mm9	33,637,090	30–37	42.0–47.0	49.9 ± 6.1	This study
mm10 newBalance	Mouse	mm10	33,811,899	30–37	42.0–47.0	49.9 ± 6.1	This study
ce11 newBalance	*C. elegans*	ce11	972,051	30–37	42.0–47.0	47.5 ± 5.5	This study
dm6 newBalance	Drosophila	dm6	2,265,271	30–37	42.0–47.0	50.5 ± 6.8	This study
danRer11 newBalance	Zebrafish	danRer11	11,331,424	30–37	42.0–47.0	48.1 ± 5.8	This study
TAIR10 newBalance	Arabidopsis	TAIR10	1,197,178	30–37	42.0–47.0	47.3 ± 4.9	This study
sacCer3 newBalance	*S. cerevisiae*	sacCer3	146,574	30–37	42.0–47.0	46.7 ± 4.5	This study
rn6 newBalance	Rat	rn6	36,842,993	30–37	42.0–47.0	50.0 ± 6.2	This study
galGal5 newBalance	Chicken	galGal5	14,209,650	30–37	42.0–47.0	49.6 ± 6.5	This study
galGal6 newBalance	Chicken	galGal6	14,767,514	30–37	42.0–47.0	49.8 ± 6.7	This study

Vita

Yuzhen Liu was born in Pingyuan, China on December 4, 1991. Growing up in Shenzhen, China, she graduated from Shenzhen Foreign Language School in the summer of 2010 and commenced her undergraduate studies at the University of Toronto in Ontario, Canada. During this period, she initiated an academic research career in the labs of Drs. Ali Salahpour (dopamine neurotoxicity), Manuela Neuman (drug hypersensitivity reactions), and Richard Hegele (antiviral activity screening of peptides). She also finished a 1-year internship as an Assistant Researcher in the In Vitro Biology Department at Orion Corporation in Turku, Finland. In the summer of 2015, she graduated from the University of Toronto with an Honours Bachelor of Science in Pharmacology and Biomedical Toxicology. Prior to PhD, she joined the lab of Dr. Andrew Hsieh at Fred Hutchinson Cancer Center and studied the pathogenic molecular mechanisms of mRNA translation in androgen-deficient prostate cancer from 2015-2019. In September 2019, she entered graduate education in the Molecular and Cellular Biology program at the University of Washington. In the lab of Dr. Brian Beliveau, she studied chromatin looping dynamics and developed locus-specific chromatin isolation methods. She defended her dissertation on February 11, 2025.