ChIP-DIP: A MULTIPLEXED METHOD FOR MAPPING PROTEINS TO DNA UNCOVERS COMBINATORICS CONTROLLING GENE EXPRESSION

Thesis by

Andrew Alexander Perez

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2024

(Defended May 8th, 2024)

Andrew Alexander Perez
ORCID: 0000-0002-8723-4859

ACKNOWLEDGEMENTS

Looking back on my graduate career there are many people I would like to thank, not only for making my experience wonderful and exciting, but also for making it possible. From my advisor and fellow graduate students making the experience full of scientific discovery and innovation, to my family and friends who helped the parent of two young children (ages two and four at the time of defending this thesis) survive the challenges of graduate school.

The first person I would like to thank by name is Mitch Guttman. He has been more supportive throughout the past four and half years than I could have possibly imagined. I remember when I first met with him before doing my rotation in his lab, I told him I was expecting a daughter in mid-October, roughly two weeks into my rotation. His first reaction was to congratulate me and told me that I should focus on the upcoming challenges of being a first-time father and that we would figure out an extended rotation so that I could get the full rotation experience when I returned from parental leave. This immediately set the tone for the student–advisor relationship we would have over the years—one of understanding and compassion for even the nonscientific aspects of my life. He's given me toys that his kids have outgrown, and my kids still play with them to this day. That aside, Mitch always made time to discuss new ideas, troubleshoot technical problems, figure out whatthe next logical step in a project would be, or anything in between. He taught me to think about the "important questions" that will move a field forward. For example, the question of "what technology is missing from the field of gene regulation to move exponentially faster?" and how to apply it properly. He gave me every tool I needed to be successful and put a tremendous amount of trust in my judgment on how to move my science forward.

Second, I would like to thank Mario Blanco. If Mitch is my "scientific father", Mario is certainly my "scientific godfather". Mario taught me so much that it would be impossible to enumerate everything. Mario is one of the most kind and caring individuals I have ever met. He always had the time and patience to help me though and think about any problem.

I also want to thank many other graduate students in the lab. In particular, I would like to thank my biggest collaborator, Isabel Goronzy. She joined the lab just as we figured out the fundamentals of ChIP-DIP and she quickly integrated into the team. She eventually "abandoned" me when she graduated just three years into graduate school, a testament to her work ethic and skill.

I also need to thank my committee; Ellen Rothenberg, Paul Sternberg, and Kata Fejes Toth. Kata was kind enough to be on my committee despite being on a completely

different continent in Hungary. She has been selfless with her time and quick to respond to anything I wanted to run by her and talk through. Paul, the chair of my committee, is one of the most efficient people I have ever met. He has also given me some amazing personal advice in the penultimate committee meeting, and I am very grateful for that. Lastly, Ellen has been an incredible resource and has really been a secondary advisor to me. She is selfless with her (very limited) time, and I remember dropping by her office on a couple of occasions to just say hi. These quick interactions both turned into hour-long conversations about science and discussions about my future. She is certainly a person that absolutely loves science without bounds and is an inspiration.

On a more personal note, I would like to thank my parents and my in-laws for their support and love. They would each give up at least one day a week, every week, for more than four years, to care for our two children while my wife and I worked. Not only did they provide childcare while I was in the lab, but they also provided a tremendous amount of encouragement throughout my time as a graduate student. Without their love and support I likely would not be writing this thesis. Not all student parents are as fortunate as I am to have family close by to help. Based on my experience, this is certainly a barrier that parents face that needs to be addressed by the institute.

Last, and certainly not least, I would like to thank my wonderful and loving wife, Allyson Perez. On more occasions than I care to admit, she would handle both kidson her own while I finished an experiment well into the night. She supported and encouraged me even when I was a ball of stress and not a fun person to be around. She sat with me on the couch as I worked on my computer late at night to plan my next experiment, she has been a pillar of strength throughout this journey, and I couldn’t be more thankful for everything that she does.

ABSTRACT

Gene regulation is governed by the complex interplay between thousands of regulatory proteins and chromatin states; understanding how these dynamics give rise to precisely controlled, cell type-specific gene expression has been a central goal of molecular biology. Yet, addressing this goal remains challenging because current methods for mapping proteins to DNA are labor-intensive, resource-demanding, and limited to studying a single or a small number of proteins at a time. To overcome this, we developed ChIP-DIP (ChIP Done In Parallel), a novel split-pool-based method that enables simultaneous, genome-wide mapping of hundreds of diverse regulatory proteins in a single experiment. We demonstrate that ChIP-DIP generates highly accurate maps equivalent to traditional approaches, with data quality unaffected by the number of distinct proteins or the composition of proteins measured within a single experiment. We show that, because of this multiplexed capability, ChIP-DIP enables generation of highly accurate maps using several orders of magnitude fewer cells per protein compared to traditional approaches (~30,000 fold), making it a powerful tool for studying a diverse array of proteins–DNA interactions with limited cellular input. In addition, we show that ChIP-DIP can generate high-quality maps for all classes of DNA-associated proteins, including histone modifications, chromatin regulators, transcription factors, and RNA polymerases. Using these data, we explore quantitative combinations of histone modifications and integrate these signatures with RNA polymerase activity, chromatin regulatory protein binding and transcription factor binding to define distinct classes of regulatory elements (e.g., distinct types of enhancer elements), their functional activity (e.g., transcriptional activity), and their regulatory potential (e.g., poised for activation upon stimulation or differentiation). Together, our results demonstratethat ChIP-DIP enables generation of consortium-level data within a single lab and highlight the importance of this approach for studying mechanisms of gene regulation in a context and cell type-specific manner. Lastly, ChIP-DIP provides a powerful platform to multiplex protein detection and provides a unique opportunity to incorporate other split-pool-based assays such as SPRITE and single-cell SPRITE to detect protein specific nuclear structures and multiplexed single-cell chromatin profiles, respectively. This work represents a transformational framework on how to study biology in a holistic manner.

PUBLISHED CONTENT AND CONTRIBUTIONS

1. Perez, A.A., Goronzy, I.N., Blanco, M.R., Guo, J.K., and Guttman, M. (2023).

   ChIP-DIP: A multiplexed method for mapping hundreds of proteins to DNA uncovers diverse regulatory elements controlling gene expression. bioRxiv, 2023.12.14.571730. 10.1101/2023.12.14.571730.
   A.A.P. conceived of this project with M.R.B. and M.G., led the effort to develop, generate data, analyze data, generate figures, and write the manuscript.

Table of Contents

Chapter 1—Introduction

Why do we need gene regulation?

DNA is the blueprint of all known life. Each organism's genome contains the DNA sequence of every gene needed to grow an organism from a single cell. RNA polymerase transcribe genes in RNAs, which go on to serve other roles, including the production of protein. Multicellular eukaryotic organisms have hundreds to thousands of unique cell types, each containing the same sequence of DNA. However, only a subset of possible genes is expressed in any given cell type; by expressing only the genes needed to carry out a specific role, the organism can use the same blueprint to do thousands of different jobs[1]. For example, a neuron found in the brain and a cell in the lining of the gut they have the identical DNA sequence, but their distinct gene expression patterns lead to very disparate functions (**Figure 1**).

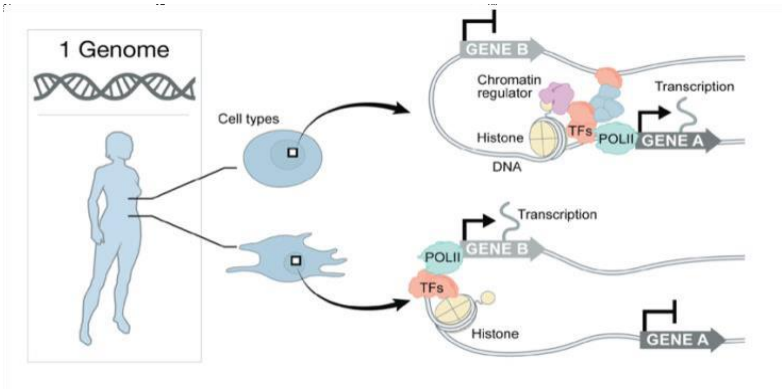

Figure 1. A graphical representation of how an organism can generate many different cell types. The left panel depicts a human with the same DNA sequence in every cell of their body. The middle panel depicts two different cell types with different functional roles in a human. The right panel depicts a simplified example of how the same DNA sequence can produce different transcriptional patterns; In the top cell type Gene A is expressed, but not Gene B. This is reversed in for the bottom cell type.

Defining a chromatin state

Gene expression is a tightly regulated process that requires the coordinated efforts of thousands of different proteins to localize to specific regions of the genome to activate, repress, or quantitatively control levels of transcription. Each genomic region of DNA is a highly organized polymer that is positioned around nucleosomes, comprised of histone proteins that are extensively modified post-translationally[2] (**Figure 2**). The combination of DNA wrapped around histones is called chromatin; the pattern of histone modifications determines the chromatin state (active, repressed, poised, or condensed) of a given region of the genome[3,4]. Chromatin state is dynamically controlled by an

expansive set of proteins known as chromatin regulators that read, write, and erase different histone modifications as well as control nucleosome positioning and accessibility of the chromatin[3,5]. As the nucleosomes change position, the chromatin becomes more or less condensed; this in turn changes the accessibility of the chromatin for sequence-specific transcription factors (TFs)[6], cofactors[7,8] and RNA polymerases[9] to bind and transcribe DNA into RNA. In addition to enabling active transcription, there are also combinations of histone modifications that poise a locus for transcription while maintaining silencing until transcription is required[10]. This is one way that the cell tightly regulates the transcription of genes and maintains, or changes, its cell state and function.

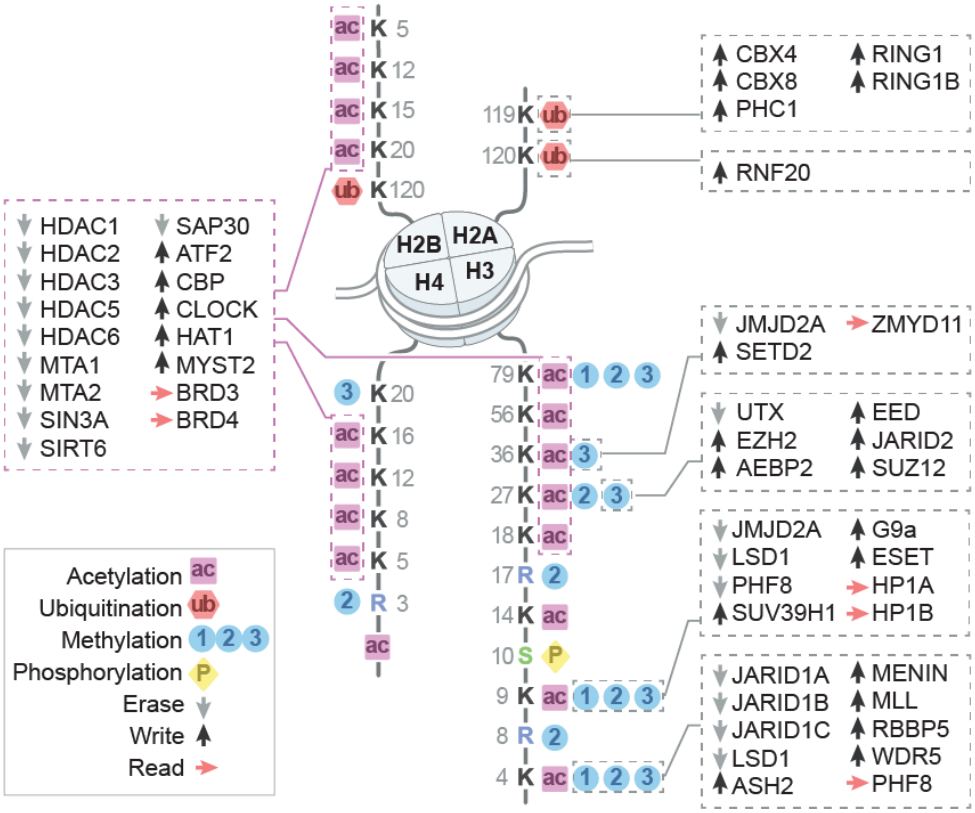

Figure 2: Histone tails are extensively modified. There are over 45 amino acids on histone tails that can be modified by various chromatin readers, writers, and erasers. Each amino acid can only be modified by one chemical group at a time, but there are multiple possible chemical groups possible for most amino acid positions. This allows for an enormous amount of possible combinations of histone combinations at any given locus. Mapping the localization of these modifications is yet to be comprehensively done. It is hypothesized that understanding all combinations of these modifications at every locus will be essential to understand the functional potential of a given genomic locus.

How do we traditionally understand gene regulation?

Understanding which proteins bind where within the genome is key to understanding how each cell generates its genome-wide chromatin state[11]. Because the binding patterns of these proteins to DNA dictate the cell's chromatin state, each cell type has a mostly unique DNA-binding pattern for each of its DNA-interacting proteins. This makes profiling where on the genome each protein is binding for each cell type a daunting challenge. For example, if an organism had 200 different cell types and each of those had 1,000 different transcriptional regulator proteins, 200,000 different protein-DNA profiles would need to be mapped **(Figure 3)**. This challenge is exacerbated by a lack of technologies to quickly map protein-DNA interactions. For example, ChIP-seq[12–15] , CUT&RUN[16], and CUT&tag[17–20] profile the DNA-protein interaction sites of anywhere from one to three different proteins in a single experiment. Given that each experiment takes two to three days, mapping the DNA-protein interactions of thousands of different proteins would take years.

To highlight the importance of this problem, international consortia (ENCODE[6,21–24], PsychENCODE[25], roadmap epigenetics, ImmGen[26], etc.) have been formed to address the problem of the sheer amount of work needed to map thousands of protein-DNA interactomes, one-by-one. ENCODE has directly published ~1000 papers in the space of 15 years and the data it has generated has been used to publish an additional ~2400 paper[27], demonstrating that the data generated by ENCODE is being used to further community knowledge. The fourth and final phase of ENCODE aims to fill a notabledeficit in transcription factor annotations using human fetal tissue, adult tissues, and, importantly, primary cells. While technology has advanced since the pilot phase of ENCODE in 2003, the majority of chromatin profiling assays are still one protein-DNA map per experiment and, as such, mapping protein-DNA interactions is still a relatively slow-going endeavor that relies on dozens of labs to generate data.

In the decade since ENCODE's inception, the focus has been on cell lines and healthy cells **(Figure 3)**. However, if one wanted to move into disease states or perturbations of any sort, the number of experiments that would be required quickly becomes infeasible for a number of reasons. First, acquiring enough cellular material to do one thousand or more protein-DNA profiling experiments in a disease state or in a perturbation system is incredibly challenging. Second, even if one could acquire the cellular material needed to perform over one thousand DNA-protein protein profiling assays, the time it would take to complete all experiments would likely not be worth the effort. There is a clear need for a way to use a reasonable number of cells to generate maps of hundreds of proteins in a reasonable amount of time.

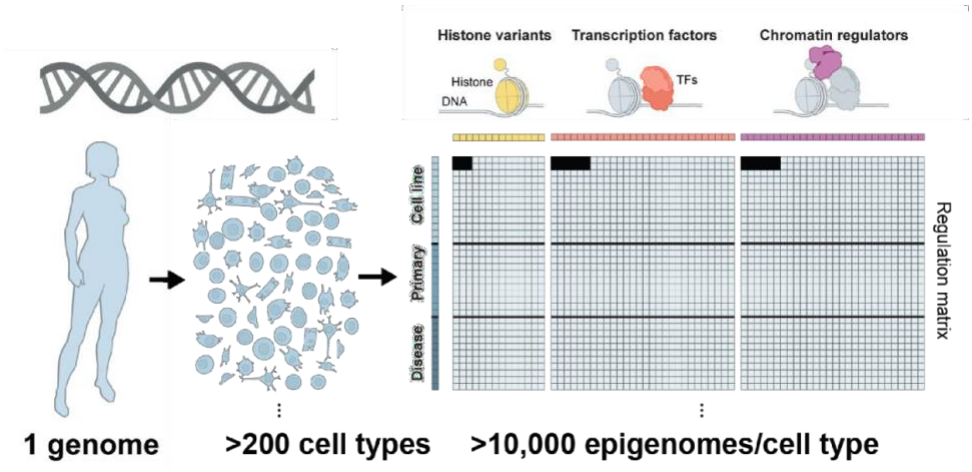

Figure 3: Graphical representation of the staggering expanse of proteins that need to be mapped when considering cell lines, primary cells, and disease states for each regulatory protein type (histone modifications, transcription factors, and chromatin regulators). The filled-in squares represent what consortium efforts have accomplished, while the white squares represent what still needs to be mapped.

Multiplexing the mapping of Protein-DNA interactions

To this end, we wondered if we could leverage our lab's expertise with split-and-pool barcoding to address the need for high-throughput mapping of DNA-protein interactions genome-wide. Specifically, we reasoned that split-and-pool barcoding is used to resolve stable complexes by putting the same string of barcodes on DNA molecules that are stably associated to each other (i.e., when complexes are split and barcoded, pooled together again, and then split out randomly to be barcoded, complexes that are stably associated to each other will always sort together). Our experience with this approach has enabled incredible leaps in technology and biological understanding through split-pool recognition of interactions by tag extension (SPRITE)[28–30] In SPRITE, the same string of barcodes is attached to DNA molecules that are in spatial proximity to each other within the nucleus and held together by crosslinking. Sequencing these barcodes allows us to identify SPRITE clusters containing DNA molecules that may be far apart in primary sequence, but close together in 3D space. We reasoned that we could leverage this for a wider array of applications by building a "synthetic" complex that contains molecules that are stably associated onto a bead (for example, a complex of biotinylated DNA molecules on streptavidin beads) and then resolve the sequence of each DNA molecule that is on each individual bead; the constituent molecules on each bead could then be resolved using split-and-pool barcoding.

Given that ChIP-Seq is used to map proteins to DNA, we sought to adapt its workflow to incorporate split pool. The ChIP-Seq workflow involves binding antibodies to protein Gthat has been covalently immobilized on a magnetic bead. This antibody–bead complex is then added to sheared and crosslinked chromatin to isolate a protein of interest and its associated chromatin. We reasoned that the bead-antibody-protein-chromatin complex is a stably associated complex that can be barcoded and resolved by split-and-pool

barcoding. Further, if multiple sets of beads were bound by unique antibodies and ChIP-Seq was performed on this pool of beads, split-and-pool barcoding could be used to resolve the chromatin associated with each bead (Figure 4A). However, it would be impossible to identify the ***specific*** proteins associated with each piece of chromatin without labeling each bead with a unique identifier. To accomplish this, we developed a novel, simple, and modular method to label each set of beads with a known oligonucleotide sequence using streptavidin-conjugated oligos bound to biotinylated protein G beads (Figure 4B).

This approach was the foundation for ChIP Done in Parallel (ChIP-DIP)[31]. In this thesis, we show that ChIP-DIP can be used generate hundreds of high-quality protein-DNA maps in a single experiment and that those data can be used to stratify the genome into functional states. Not only can this combination of split-pool and bead barcoding be used to multiplex the mapping of protein-DNA interactions genome-wide, it can also be used to multiplex the mapping of protein-RNA interactions genome-wide. For example, we can use a similar approach called SPIDR[32] (Split-Pool Identification of RPB targets), to map dozens of RNA binding proteins (RBPs) in a single experiment. This work represents an important step towards a holistic understanding of biology.

In this thesis, I discuss the method ChIP-DIP, its applications for mapping chromatin states, TFs, chromatin regulators, new insights into combinatorial histone modifications and gene regulations and conclude with a perspective on how this approach and the resulting data will enable a transformation in our understanding of gene regulation by allowing us comprehensively study all molecules within the system simultaneously.

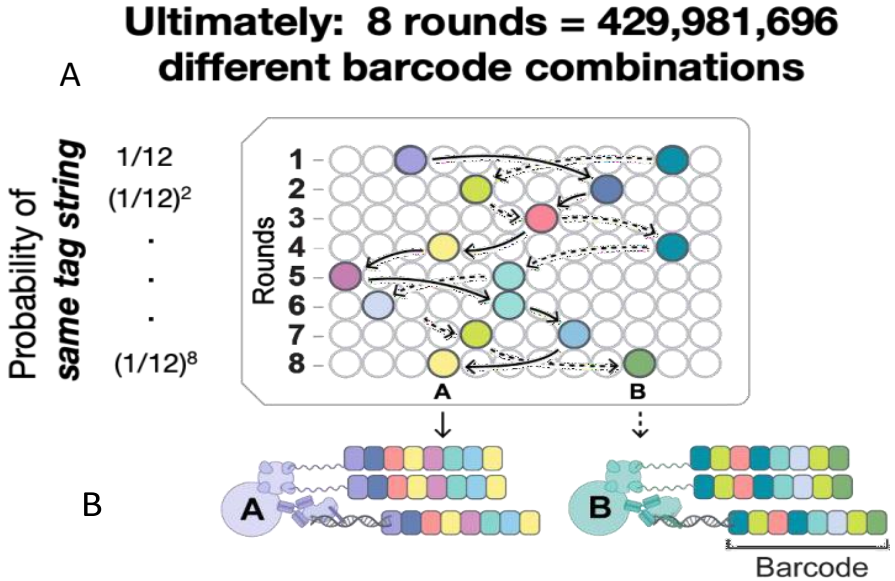

Figure 4: Split-and-pool barcoding is used to put a unique string of barcodes on stably associated DNA molecules. A) Graphical representation of each round of split pool. After beads are put into a well (that contains a unique barcode) ligase is added to ligate the barcode onto the DNA on the beads. After ligation, beads are pooled and then split out into a new row with another set of unique barcodes in each well. In this example, this is performed eight times, yielding ~430 million unique barcode string combinations. The

goal is to have more barcode combinations that total beads to put a unique barcode string on the DNA molecules of each bead. B) graphical representation on what two fully barcodes beads would look like after split-and-pool barcoding is performed. The barcode string on each molecule is unique to the bead.

References

1. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*, 621–628. 10.1038/nmeth.1226.

2. Bednar, J., Horowitz, R.A., Grigoryev, S.A., Carruthers, L.M., Hansen, J.C., Koster, A.J., and Woodcock, C.L. (1998). Nucleosomes, linker DNA, and linker histone form a unique structural motif that directs the higher-order folding and compaction of chromatin. Proc. Natl. Acad. Sci. *95*, 14173–14178. 10.1073/pnas.95.24.14173.

3. Jenuwein, T., and Allis, C.D. (2001). Translating the Histone Code. Science *293*, 1074–1080. 10.1126/science.1063127.

4. Huang, H., Sabari, B.R., Garcia, B.A., Allis, C.D., and Zhao, Y. (2014). SnapShot: Histone Modifications. Cell *159*, 458–458.e1. 10.1016/j.cell.2014.09.037.

5. Tekel, S.J., and Haynes, K.A. (2017). Molecular structures guide the engineering of chromatin. Nucleic Acids Res. *45*, gkx531. 10.1093/nar/gkx531.

6. Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74. 10.1038/nature11247.

7. ROEDER, R.G. (1998). Role of General and Gene-specific Cofactors in the Regulation of Eukaryotic Transcription. Cold Spring Harb. Symp. Quant. Biol. *63*, 201–218. 10.1101/sqb.1998.63.201.

8. Malik, S., and Roeder, R.G. (2023). Regulation of the RNA polymerase II pre-initiation complex by its associated coactivators. Nat. Rev. Genet. *24*, 767–782. 10.1038/s41576-023-00630-9.

9. Barba-Aliaga, M., Alepuz, P., and Pérez-Ortín, J.E. (2021). Eukaryotic RNA Polymerases: The Many Ways to Transcribe a Gene. Frontiers Mol Biosci *8*, 663209. 10.3389/fmolb.2021.663209.

10. Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. Cell *125*, 315–326. 10.1016/j.cell.2006.02.041.

11. Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R.P., et al. (2007). Genome-wide maps of

chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553–560. 10.1038/nature06008.

12. Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science *316*, 1497–1502. 10.1126/science.1141319.

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

14. He, Q., Johnston, J., and Zeitlinger, J. (2015). ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. Nat. Biotechnol. *33*, 395–401. 10.1038/nbt.3121.

15. Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., et al. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat. Methods *4*, 651–657. 10.1038/nmeth1068.

16. Skene, P.J., Henikoff, J.G., and Henikoff, S. (2018). Targeted in situ genome-wide profiling with high efficiency for low cell numbers. Nat Protoc *13*, 1006–1019. 10.1038/nprot.2018.015.

17. Janssens, D.H., Meers, M.P., Wu, S.J., Babaeva, E., Meshinchi, S., Sarthy, J.F., Ahmad, K., and Henikoff, S. (2021). Automated CUT&Tag profiling of chromatin heterogeneity in mixed-lineage leukemia. Nat. Genet. *53*, 1586–1596. 10.1038/s41588-021-00941-9.

18. Kaya-Okur, H.S., Wu, S.J., Codomo, C.A., Pledger, E.S., Bryson, T.D., Henikoff, J.G., Ahmad, K., and Henikoff, S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. Nat Commun *10*, 1930. 10.1038/s41467-019-09982-5.

19. Bartosovic, M., and Castelo-Branco, G. (2023). Multimodal chromatin profiling using nanobody-based single-cell CUT&Tag. Nat. Biotechnol. *41*, 794–805. 10.1038/s41587-022-01535-4.

20. Gopalan, S., Wang, Y., Harper, N.W., Garber, M., and Fazzio, T.G. (2021) Simultaneous profiling of multiple chromatin proteins in the same cells. Mol. Cell 81, 4736–4746.e5. 10.1016/j.molcel.2021.09.019.

21. Vierstra, J., Lazar, J., Sandstrom, R., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Haugen, E., et al. (2020). Global reference mapping of human transcription factor footprints. Nature *583*, 729–736. 10.1038/s41586-020-2528-x.

22. Partridge, E.C., Chhetri, S.B., Prokop, J.W., Ramaker, R.C., Jansen, C.S., Goh, S.-T., Mackiewicz, M., Newberry, K.M., Brandsmeier, L.A., Meadows, S.K., et al.

(2020). Occupancy maps of 208 chromatin-associated proteins in one human cell type. Nature *583*, 720–728. 10.1038/s41586-020-2023-4.

23. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330. 10.1038/nature14248.

24. Abascal, F., Acosta, R., Addleman, N.J., Adrian, J., Afzal, V., Ai, R., Aken, B., Akiyama, J.A., Jammal, O.A., Amrhein, H., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature *583*, 699–710. 10.1038/s41586-020-2493-4.

25. Consortium, P., Akbarian, S., Liu, C., Knowles, J.A., Vaccarino, F.M., Farnham, P.J., Crawford, G.E., Jaffe, A.E., Pinto, D., Dracheva, S., et al. (2015). The PsychENCODE project. Nat Neurosci *18*, 1707–1712. 10.1038/nn.4156.

26. Consortium, T.I.G.P., Heng, T.S.P., Painter, M.W., Elpek, K., Lukacs-Kornek, V., Mauermann, N., Turley, S.J., Koller, D., Kim, F.S., Wagers, A.J., et al. (2008). The Immunological Genome Project: networks of gene expression in immune cells. Nat Immunol *9*, 1091–1094. 10.1038/ni1008-1091.

27. Abascal, F., Acosta, R., Addleman, N.J., Adrian, J., Afzal, V., Aken, B., Ai, R., Akiyama, J.A., Jammal, O.A., Amrhein, H., et al. (2020). Perspectives on ENCODE. Nature *583*, 693–698. 10.1038/s41586-020-2449-8.

28. Quinodoz, S.A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J.M., Detmar, E., Lai, M.M., Shishkin, A.A., Bhat, P., Takei, Y., et al. (2018). Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. Cell *174*, 744–757.e24. 10.1016/j.cell.2018.05.024.

29. Quinodoz, S.A., Jachowicz, J.W., Bhat, P., Ollikainen, N., Banerjee, A.K., Goronzy, I.N., Blanco, M.R., Chovanec, P., Chow, A., Markaki, Y., et al. (2021). RNA promotes the formation of spatial compartments in the nucleus. Cell *184*, 5775–5790.e30. 10.1016/j.cell.2021.10.014.

Quinodoz, S.A., Bhat, P., Chovanec, P., Jachowicz, J.W., Ollikainen, N., Detmar, E., Soehalim, E., and Guttman, M. (2022). SPRITE: a genome-wide method for mappinghigher-order 3D interactions in the nucleus using combinatorial split-and-pool barcoding. Nat Protoc *17*, 36–75. 10.1038/s41596-021-00633-y.

30. Perez, A.A., Goronzy, I.N., Blanco, M.R., Guo, J.K., and Guttman, M. (2023). ChIP-DIP: A multiplexed method for mapping hundreds of proteins to DNA uncovers diverse regulatory elements controlling gene expression. bioRxiv, 2023.12.14.571730. 10.1101/2023.12.14.571730.

31. Wolin, E., Guo, J.K., Blanco, M.R., Perez, A.A., Goronzy, I.N., Abdou, A.A., Gorhe, D., Guttman, M., and Jovanovic, M. (2023). SPIDR: a highly multiplexed method for mapping RNA-protein interactions uncovers a potential mechanism for selective

translational suppression upon cellular stress. bioRxiv, 2023.06.05.543769. 10.1101/2023.06.05.543769.

Chapter 2–Chromatin Immunoprecipitation Done In Parallel (ChIP-DIP)

Introduction

Although every cell in the body contains the same genomic DNA sequence, distinct cell types express different genes to enable cell-type-specific function. Cell-type-specific gene regulation involves the coordinated activity of thousands of regulatory proteins that localize at precise DNA regions to activate, repress, and quantitatively control levels of transcription. Each genomic DNA region is organized around nucleosomes[1], which contain histone proteins that undergo extensive post-translational modifications[2,3] and together define cell-type-specific chromatin states. Chromatin state is controlled by chromatin regulators that directly read, write and erase specific histone modifications[2,4], as well as control nucleosome positioning and DNA accessibility[5,6]. This ultimately determines which genomic regions are accessible for binding by sequence-specific transcription factors[7], the enzymes that transcribe DNA into RNA (RNA polymerases)[8] and other general and specific regulatory proteins that promote or suppress transcriptional initiation[9,10].

Understanding how regulatory protein binding gives rise to cell type-specific gene expression has been a central goal of molecular biology for decades[2]. Over the past 20 years, significant technical advances have enabled genome-wide mapping of regulatory proteins and histone modifications (e.g., ChIP-Seq)[11–14], improved binding site resolution (ChIP-Exo)[15,16], increased sample throughput (e.g., through automation and/or sample pooling)[17,18], and mapping within a limited numbers of cells (e.g.,CUT&RUN/CUT&Tag)[19–21]. Yet, while these innovations have enabled new applications and uncovered critical new insights into gene regulation, all of these approaches still work by studying a single protein at a time. The two exceptions are multiplexed Cut&Tag[22,23] and MAbID[24], which can measure up to three or six histone modifications, but not transcription factors or other regulatory proteins, in a single experiment[23]. Due to the large number of distinct regulatory proteins and histone modifications involved and the cell-type-specific nature of their regulatory interactions, this one-at-a-time mapping approach makes it extremely difficult to construct a comprehensive understanding of gene regulation.

Initial attempts to overcome this challenge led to the formation of various international consortia that aimed to generate reference maps of hundreds of proteins within a small number of cell types (ENCODE[25], PsychENCODE[26], ImmGen[27], etc.). Although these efforts have provided many critical insights[28–30], because protein binding maps and gene expression programs are intrinsically cell-type specific[31,32], it is not possible to study cell type-specific regulation using maps generated from reference cell lines[33]. To date, most mammalian cell types, experimental and disease models, and model organisms remain uncharacterized. Generating additional cell type-specific regulatory maps currently requires consortium-level effort (dozens of labs across the world), time (many years), and resources (>$100 million) for each biological system. Accordingly, there is a clear need for a highly scalable, multiplexed protein profiling method that can increase throughput of protein mapping by orders of magnitude and profile the diverse categories of DNA-

associated proteins, including classes that have been traditionally easier to map (e.g., histone modifications) and those that have been more challenging (e.g., transcription factors)[34]. Such a method would allow any lab to generate comprehensive maps for anycell type of interest in a rapid and cost-effective way and would enable exploration of key questions that is not currently possible.

To address this need, we developed chromatin immunoprecipitation done in parallel (ChIP-DIP), a scalable platform that enables simultaneous, genome-wide mapping of hundreds of diverse regulatory proteins within a single experiment. We utilized ChIP-DIP to generate data from ~180 distinct DNA-associated proteins in human or mouse cells, including a single multiplexed ChIP-DIP experiment containing >225 different antibodies targeting ~160 distinct proteins (**Supplemental Table 1)**. We show that ChIP-DIP generates accurate genome-wide maps, equivalent to those generated by traditional approaches, that are highly robust regardless of the number of antibodies or protein composition contained within a pool and across a range of input cell numbers. We show that ChIP-DIP enables accurate mapping of all classes of DNA-associated proteins, including histone modifications, chromatin regulators, transcription factors and other sequence-specific DNA-binding proteins, and RNA polymerases. Together, our results demonstrate that ChIP-DIP enables the generation of 'consortium-level' comprehensive, context-specific protein-localization maps within any experimental system and within any molecular biology lab and enables the exploration of complex, combinatorial patterns of protein localization that define regulatory activity. Beyond the immediate applications, ChIP-DIP can be directly integrated into a suite of existing split-pool approaches to enable highly multiplexed mapping of protein localization within single cells in combination with measurements of 3D structure[35], ncRNA localization, and nascent transcription[36,37].

Results

To enable highly multiplexed, genome-wide mapping of hundreds of DNA-associated proteins in a single experiment, we developed ChIP-DIP (ChIP Done In Parallel) (**Figure 1A**). ChIP-DIP works by (i) using a rapid, modular, and simple approach to couple individual antibodies to beads containing a unique oligonucleotide tag (**Figure S1A**), (ii) combining sets of different antibody–bead-oligo conjugates to create an antibody–bead pool, (iii) performing ChIP using this pool, (iv) conducting split-and-pool barcoding to match antibody–bead-oligo conjugates to specific genomic DNA regions[36,38,39], and (v) sequencing DNA and computationally matching split-pool barcodes that are shared between genomic DNA and the antibody tag. We refer to all unique reads containing the same split-pool barcode as a cluster. We combine DNA reads from all clusters corresponding to the same antibody to generate a protein-localization map for each individual protein. The output of a ChIP-DIP experiment is analogous to the data generated in a traditional ChIP-Seq experiment, however, instead of a single map, ChIP-DIP provides a set of distinct maps–one for each antibody utilized (**Figure 1B**).

To ensure that chromatin–antibody–bead–oligo conjugates remain intact throughout the ChIP-DIP procedure (rather than dissociating and reforming new complexes), we designed

a series of experiments to measure dissociation between (i) oligo and bead, (ii) antibody and bead, or (iii) antibody and chromatin (**Figure S1B**).

Antibody-ID oligo movement: If antibody-ID oligos were to dissociate from their original bead and bind to other beads during the ChIP-DIP procedure, then multiple antibody-ID oligo types would share a single split-pool barcode, leading to errorsin assigning a chromatin region to the correct protein. If this were the case, we would expect to observe clusters that contain multiple distinct antibody-ID oligos. To explore this, for each split-pool barcode we calculated the proportion of antibody-ID oligo reads corresponding to the maximally represented antibody-ID oligo type. In a representative experiment, we observed that 96% of clusters had only a single type of antibody-ID oligo (100% maximum representation) and 99.4% of barcodes had at least 80% maximum representation (**Figure S1C**). Since most split-pool barcodes have unique representation, this suggests that antibody-ID oligo movement occurs infrequently and should not significantly impact the accuracy of antibody-to-chromatin assignments.

Antibody movement: If antibodies for protein X dissociated from their bead and reassociated with a distinct protein G bead containing a label for an antibody recognizing protein Y, then chromatin captured by that antibody would be improperly assigned to protein Y. To quantify the frequency of such events, we performed a ChIP-DIP experiment where we added in oligo-labeled protein G beads that were not conjugated to an antibody or to an IgG antibody and measured the amount of chromatin that was assigned to these beads. In all cases, we observed minimal detection of chromatin (<0.5%). Specifically, we performed a ChIP-DIP experiment with oligo-labeled beads containing a CTCF antibody, an isotype control IgG antibody, or no antibody (empty). Only beads containing an antibody (i.e., CTCF, IgG) were present during the IP stages and empty beads were added in various post-IP, pre-split- pool processing steps to determine the frequency ofmixing at each step. If antibodies moved between beads during post-IP processing, we would expect to find chromatin associated with empty beads. When we measured the amount of chromatin assigned to each bead, we observed that beads with the IgG control antibody received 0.10% ($1/1000^{th}$) the amount of chromatin compared to the CTCF antibody (**Figure S1E**). Empty beads added during the end-repair reaction and dA-tailing reaction received 0.40% and 0.10% the amount of chromatin of the CTCF antibody beads, respectively. We note that these estimates are likely to be an overestimate of the true mixing rate because this experimental design does not distinguish between chromatin associating due to antibody movement from chromatin that may non-specifically bind to IgG or empty protein G beads. Nonetheless, these results indicate that the impact of antibody movement on chromatin assignment is minimal, representing no more than 0.5% of all chromatin captured.

Chromatin movement: If proteins dissociate from their epitope-specific antibodies post IP, they may specifically bind to other beads containing the same epitope- specific antibodies. If this movement occurs during the split-pool process, then these chromatin fragments would not be assigned because they would lack a paired antibody-ID oligo with the same barcode. We estimated the frequency of chromatin movement using a human-mouse mixing

experiment in which we separately IP'd a human chromatin sample and a mouse chromatin sample using identical pools of labeled beads. After the IP, we mixed the samples and performed the remainder of the standard ChIP-DIP protocol (DNA processing and split pool). If chromatindissociates and rebinds prior to split pool then we would expect that it could rebind to beads from the other species containing the same antibody type as its original bead. We observed that only ~5% of reads were assigned to beads of the other species (4.2% and 5.8%, **Figure S1E**). This suggests that disassociation and reassociation of chromatin-crosslinked proteins from their epitope-specific antibodies during ChIP-DIP processing steps is minimal. Nonetheless, in the cases where this does occur it would impact the assignability of these chromatin fragments (sensitivity of detection) rather than resulting in incorrect assignment (specificity of detection).

We observed minimal dissociation for any of these cases; most beads contain a single oligo type (>95%, **Figure S1C**), beads without a coupled antibody are associated with minimal chromatin (<0.5%, **Figure S1D**) and most purified chromatin originates from the initial capture (> 94%, **Figure S1E**).

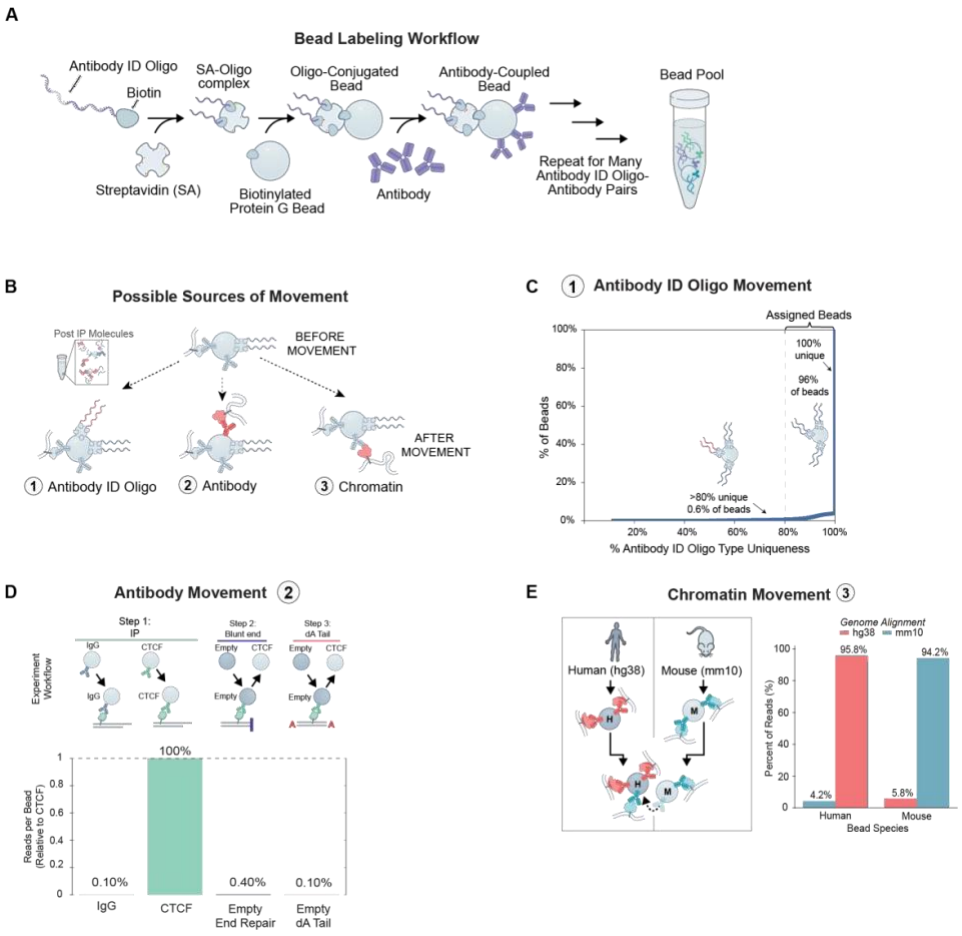

Figure S1: Potential sources of mixing in ChIP-DIP, related to Figure 1. (A) Schematic of labeling strategy to generate Protein G beads coupled with a unique antibody-identifying oligonucleotide and a matched antibody. Protein G beads are covalently modified with a biotin; oligonucleotides containing a 3' biotin are conjugated to streptavidin; oligo–streptavidin complexes are mixed with biotinylated protein G beads and protein G beads are mixed with antibodies. This process is repeated for each unique oligonucleotide–antibody pair and then all bead-antibody conjugates are pooled together. **(B)** Schematic ofthree potential sources of mixing during ChIP-DIP. **(C)** Cumulative distribution plot representing the uniqueness of antibody-ID oligos type (x-axis) within individual clusters. **(D)** Schematic of experimental design to test for antibody movement between beads and quantification of relative reads per bead assigned to true targets (CTCF) or empty beads added during experimental processing steps. **(E)** Schematic of human-mouse experimental design to test for chromatin movement and quantification of species-specific reads assigned to human or mouse beads.

To test whether ChIP-DIP can accurately map genome-wide protein localization, we performed a ChIP-DIP experiment in human K562 cells using four well-studied proteins: 1) the CTCF sequence-specific DNA-binding protein that binds to insulator sequences[40], 2) the H3K4me3 histone modification that localizes at the promoters of active genes[13,41], (3) the RNA Polymerase II enzyme that transcribes RNA[42] and (4) the H3K27me3 histone modification that accumulates over broad genomic regions that are associated with polycomb-mediated transcriptional repression[13,41]. We observed localization patterns that are highly comparable at specific genomic sites (**Figure 1B–C**) and highly correlated genome-wide (r=0.837–0.956, **Figure 1D**) to ChIP-Seq profiles generated by the ENCODE consortium[25,28,43] (**Supplemental Table 2)**.

These genome-wide profiles were highly consistent even when using antibody pools containing different numbers of antibodies and protein composition (**Figure 2A-D, Supplemental Table 3**), including pools containing independent antibodies targeting the same protein (CTCF) or multiple proteins within the same complex (e.g., members of the PRC1/2 complex, **Figure S2**). Because ChIP-DIP enables simultaneous mapping of many proteins within the same experiment, we reasoned that it may dramatically reduce the total number of input cells required per experiment[44] (**Methods**). Indeed, we observed strong genome-wide correlations and peak overlap when measuring these same proteins across a range of input cell numbers (**Figure 2E-H, Figure S3, Supplemental Table 4**). Because ChIP-DIP can generate dozens of individual maps from the same lysate, this further reduces the effective number of cells required for each protein target. In this example, we observed strong correlations with maps generated from ~1,400 cells per protein.

Together, these results demonstrate that ChIP-DIP is highly robust and generates data that are highly comparable to those generated by standard methods.

Figure 1

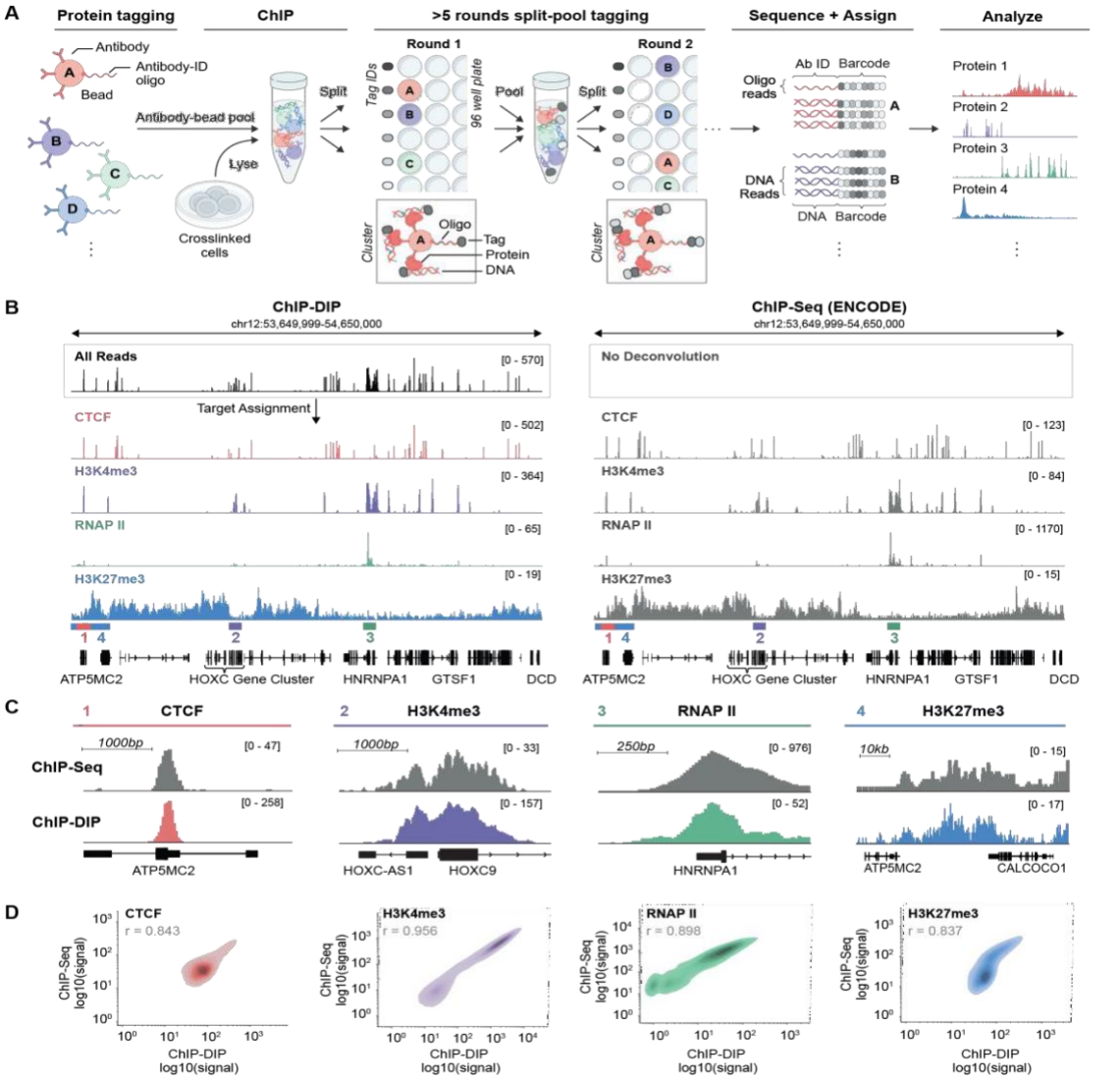

Figure 1: ChIP-DIP: A highly multiplexed method for mapping proteins to genomic DNA. (A) Schematic of the ChIP-DIP method. Beads are coupled with an antibody and associated oligonucleotide (antibody ID). Sets of beads are then mixed (antibody–bead pool) and used to perform ChIP. Multiple rounds of split-and-pool barcoding are performed to identify molecules bound by the same protein G bead. DNA is sequenced and genomic DNA and antibody oligos containing the same split-and-pool barcode are grouped into a cluster, which are used to assign genomic DNA regions to their linked antibodies. All DNA reads corresponding to the same antibody are used to generate protein-localization maps. **(B)** Protein-localization maps over a specific human genomic region (hg38, chr12:53,649,999–54,650,000) for four protein targets—CTCF, H3K4me3, RNAP II and H3K27me3. Left panel: Protein localization generated by ChIP-DIP in K562. Top track shows read coverage prior to protein assignment and bottom four tracks correspond to read coverage after assignment to individual proteins. Right panel: ChIP-Seq data generated by ENCODE within K562 for these same four proteins are shown for the same region. To

enable direct comparison of scales between datasets, we normalized the scale to coverage per million aligned reads. Scale is shown from zero to maximum coverage within each region. **(C)** Comparison of ChIP-DIP and ChIP-Seq maps over specific regions corresponding to zoom-ins of the larger region shown in (B). The locations presented are demarcated by colored bars above the gene track of (B). Scale shown similar to (B). **(D)** Genome-wide comparison (density plots of signal correlation) between the localization of each individual protein measured by ChIP-DIP (x-axis) or ChIP-Seq (y-axis). Points are measured genome-wide across 10kb windows (CTCF, H3K27me3) or all promoter intervals (H3K4me3, RNAP II).

Figure 2

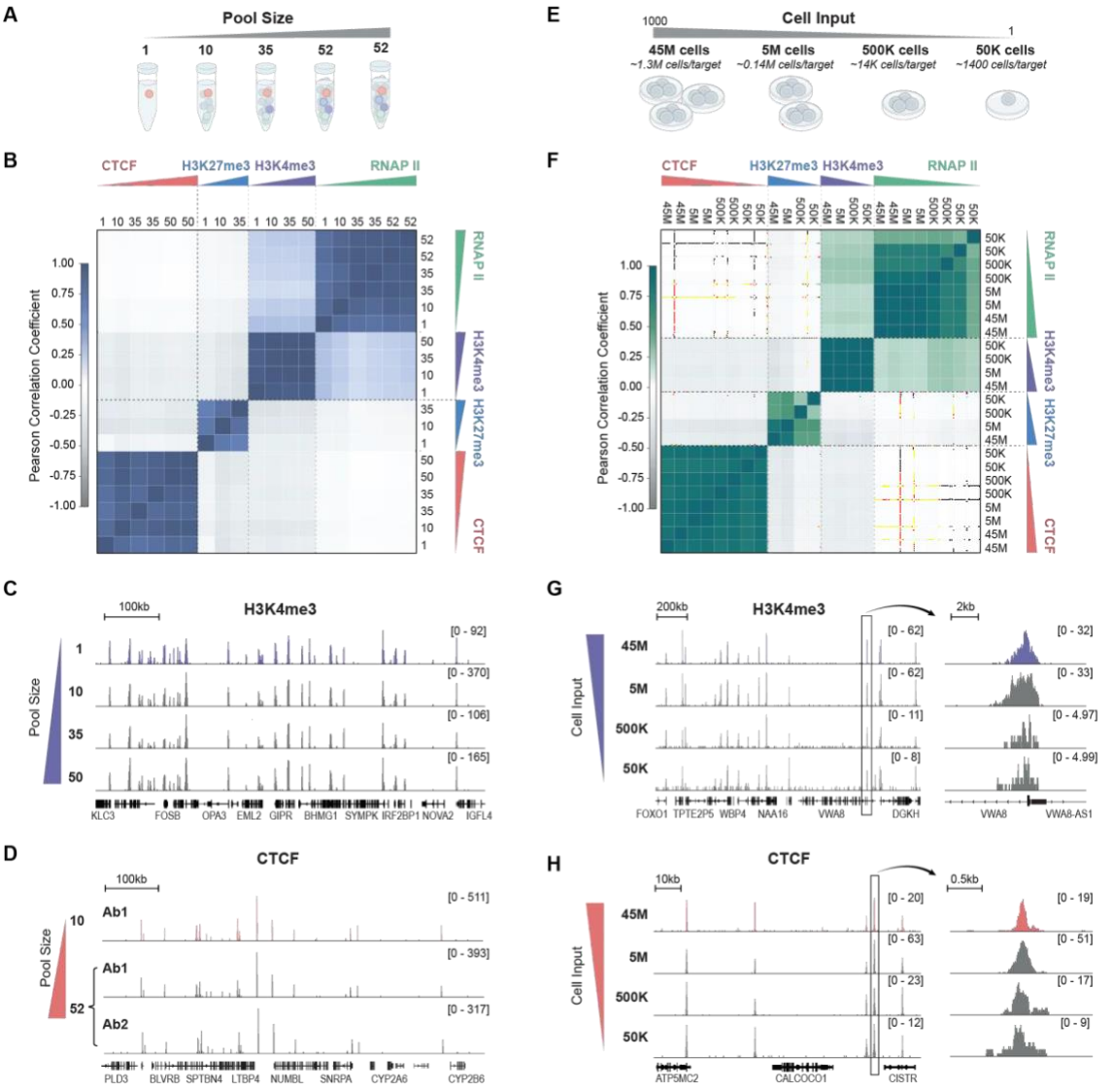

Figure 2: ChIP-DIP accurately maps large sets of proteins using low-levels of cell lysate. (A) Schematic of experimental design to test scalability of antibody–bead pool size and composition. **(B)** Correlation heatmap for protein-localization maps of four proteins–CTCF, H3K4me3, RNAP II and H3K27me3–generated using antibody pools of four different sizes and compositions (see **Methods**). Pool sizes are listed along top and left axis. Replicate proteins in the same pool indicate that a different antibody was used for that protein. Some antibodies were not included in every pool. **(C)** Comparison of H3K4me3 localization over a specific genomic region (hg38, chr19:45,345,500–46,045,500) when measured within various antibody pool sizes and compositions. **(D)** Comparison of CTCF localization over a specific genomic region (hg38, chr19:40,349,999–41,050,000) when measured within a pool of ten antibodies containing a single CTCF-targeting antibody (top)

or two different CTCF-targeting antibodies within a pool of 52 antibodies (bottom). **(E)** Schematic of experimental design to test the amount of cell input required for ChIP-DIP. **(F)** Correlation heatmap for protein-localization maps of four targets–CTCF, H3K4me3, RNAP II and H3K27me3–generated using various amounts of input cell lysate (see **Methods**). Amounts of input cell lysate are listed along top and left axis. **(G)** Comparison of H3K4me3 localization over a specific genomic region (hg38, chr13:40,600,000–42,300,000) when measured using various amounts of input cell lysate. **(H)** Comparison of CTCF localization over a specific genomic region (hg38, chr12:53,664,000–53,764,000) when measured using various amounts of input cell lysate.

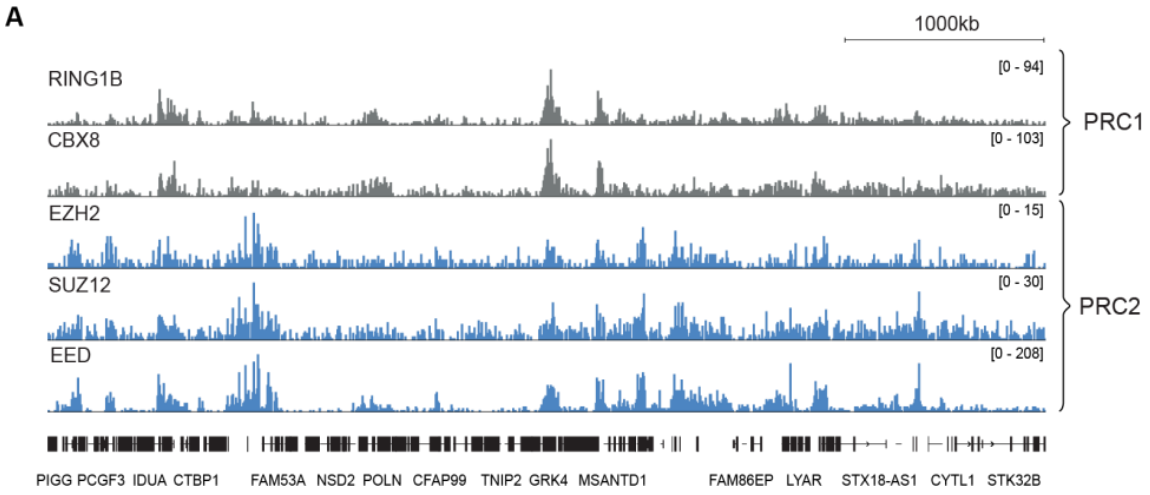

Figure S2: Simultaneous mapping of multiple components of a single protein complex using ChIP-DIP, related to Figure 2. (A) Visualization of various components of the PRC1 (RING1B, CBX8) and PRC2 (EZH2, SUZ12, EED) complexes that were mapped within the same ChIP-DIP pool (K562 52 Antibody Pool) along a genomic region (hg38, chr4:500,000–5,500,000).

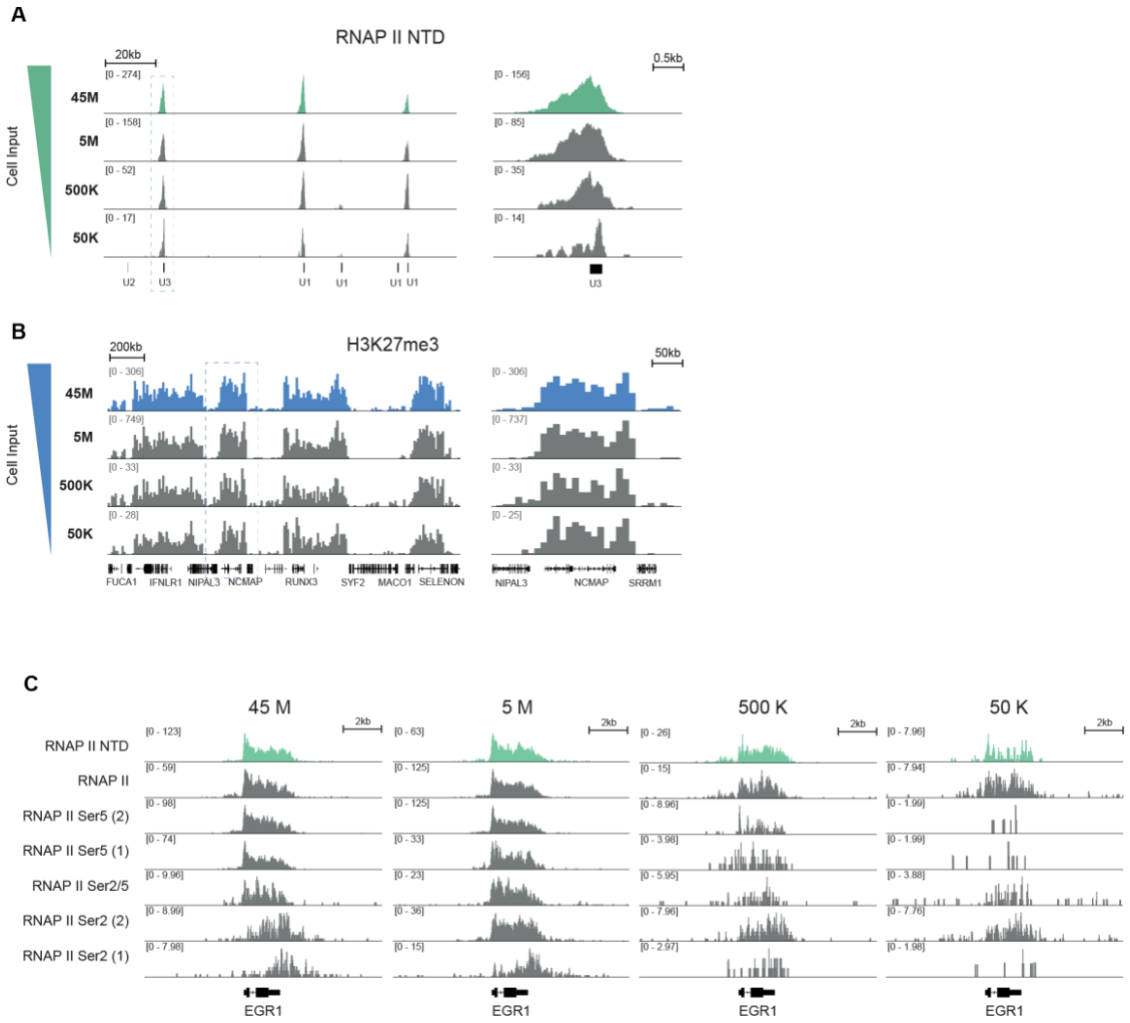

Figure S3. Comparison of protein localization across different amounts of cell lysate, related to Figure 2. (A) Comparison of RNAP II NTD localization across a snRNA gene cluster (hg38, chr17:58,620,000–58,689,000) when generated using various amounts of input K562 cell lysate. **(B)** Comparison of H3K27me3 localization across a genomic region (hg38, chr1:23,850,000–25,850,000) generated using various amounts of input K562 cell lysate. **(C)** Comparison of various isoforms of RNAP II at the EGR1 locus (hg38, chr5:138,455,000–138,480,000) generated using various amounts of input K562 cell lysate.

ChIP-DIP accurately maps hundreds of diverse DNA-associated proteins

We next explored whether ChIP-DIP can simultaneously map proteins from distinct categories, some of which have been traditionally easier to map than others[45,46]. To do this, we performed ChIP-DIP on >60 distinct proteins in human K562 cells and >160 distinct proteins in mouse embryonic stem cells (mESCs) across six experiments (**Supplemental Table 1**). These included 39 histone modifications (HMs), 67 chromatin regulators (CRs), 51 transcription factors (TFs), and all three RNA polymerases (RNAPs) and four of their modified forms.

Histone modifications and chromatin regulators. Histone modifications define cell-type-specific chromatin states and have proven incredibly useful for annotating cell type-specific regulatory elements[47]. We mapped 39 histone modifications–including 18 acetylation, 17 methylation, 3 ubiquitination and 1 phosphorylation marks–in either mESCs or K562s (**Figure 3A**). We confirmed the localization of five histone modifications commonly used to demarcate five functional states[7], as well as additional modifications associated with each state (**Figure S4A-F**): enhancer regions[48] (H3K4me1, H3K4me2, H3K27ac, **Figure 3B**), transcribed regions[13,49,50] (H3K36me3, H3K79me1/2, **Figure 3C**), promoter regions[13,41,51] (H3K4me3, H3K9ac, **Figure 3D**), polycomb-repressed regions[52] (H3K27me3, H2AK119ub, **Figure 3E**), and constitutive heterochromatin regions[53] (H3K9me3, H4K20me3, **Figure 3F)**. These data indicate that ChIP-DIP accurately maps histone modifications with distinct genome-wide patterns (broad and focal localization), that represent distinct activity states (active and repressive), and that localize at distinct functional elements (promoters, enhancers, gene bodies, and intergenic regions).

Chromatin regulators (CRs) are responsible for reading, writing, and erasing specific histone modifications and are critical for the establishment, maintenance, and transition between chromatin states[54,55]. We measured 67 CRs associated with various histone methylation, acetylation, and ubiquitination marks, as well as with DNA methylation, in either mouse ES or human K562 cells (**Figure 3A**). As expected, we observe that an eraser (JARID1A)[56] and a writer (RBBP5-containing complex)[57] of H3K4me3 localize at H3K4me3-modified promoter sites (**Figure 3G, Figure S4G**). Additionally, we observed that components of the PRC1 (RING1B, CBX8)[58] and PRC2 complex (EED, SUZ12, EZH2)[59] co-localize and are enriched over genomic regions containing their respective histone modifications (H2AK119ub and H3K27me3, **Figure 3H, Figure S4H**). Similarly, we observed co-localization of two members of the heterochromatin protein 1 (HP1) family, HP1α and HP1ß, at genomic DNA regions containing their associated heterochromatin marks, H3K9me3 and H4K20me3[60] (**Figure 3I, Figure S4I**). These data indicate that ChIP-DIP accurately maps chromatin regulators from diverse complexes and with distinct functional properties (i.e., modification recognition, enzymatic activity, chromatin packaging).

Figure 3

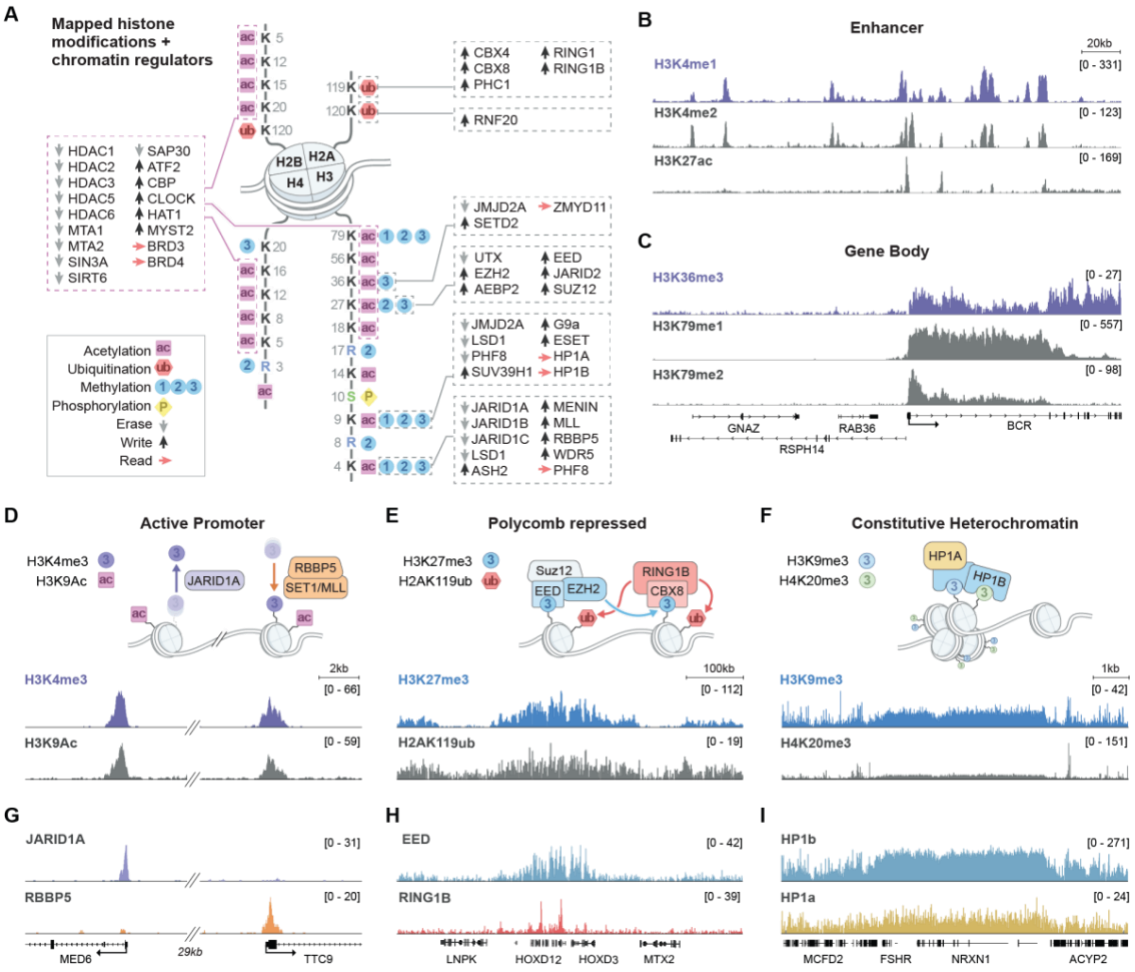

Figure 3: ChIP-DIP accurately maps dozens of functionally diverse histone modifications and chromatin regulators. (A) Illustration of the diverse histone modifications and chromatin regulatory proteins mapped in K562 or mESC using ChIP-DIP. **(B–C)** Visualization of multiple histone modifications across a genomic region (hg38, chr22:23,050,000–23,290,000) in K562 corresponding to multiple histone modifications associated with **(B)** enhancers–H3K4me1, H3K4me2 and H3K27Ac and **(C)** active gene bodies–H3K36me3, H3K79me1 and H3K79me2. **(D)** Top: Schematic of histone modifications and chromatin regulators associated with active promoters. Bottom: Visualization of multiple histone modifications associated with active promoters—H3K4me3 and H3K9Ac—across a genomic region (mm10, chr12:81,590,000–81,636,000) in mouse ESCs. Hashmarks indicate an intervening 29kb region that is not shown. **(E)** Top: Schematic of histone modifications and chromatin regulators associated with polycomb-mediated repression. Bottom: Visualization of multiple histone modifications associated with polycomb-mediated repression—H3K27me3 and H2A119ub—across a genomic region (hg38, chr2:175,846,000–176,446,000) containing the silenced HOXD cluster in K562. **(F)** Top: Schematic of histone modifications and chromatin regulators associated

with constitutive heterochromatin. Bottom: Visualization of multiple histone modifications associated with constitutive heterochromatin—H3K9me3 and H4K20me3—across a genomic region (hg38, chr2:46,200,000–55,700,000) in K562. **(G)** Visualization of an H3K4me3-associated eraser (JARID1A) and writer component (RBBP5) across the same genomic region as (D). **(H)** Visualization of PRC2 (EED) and PRC1 (RING1B) components across the same genomic region as (E). **(I)** Visualization of HP1b and HP1a across the same genomic region as (F).

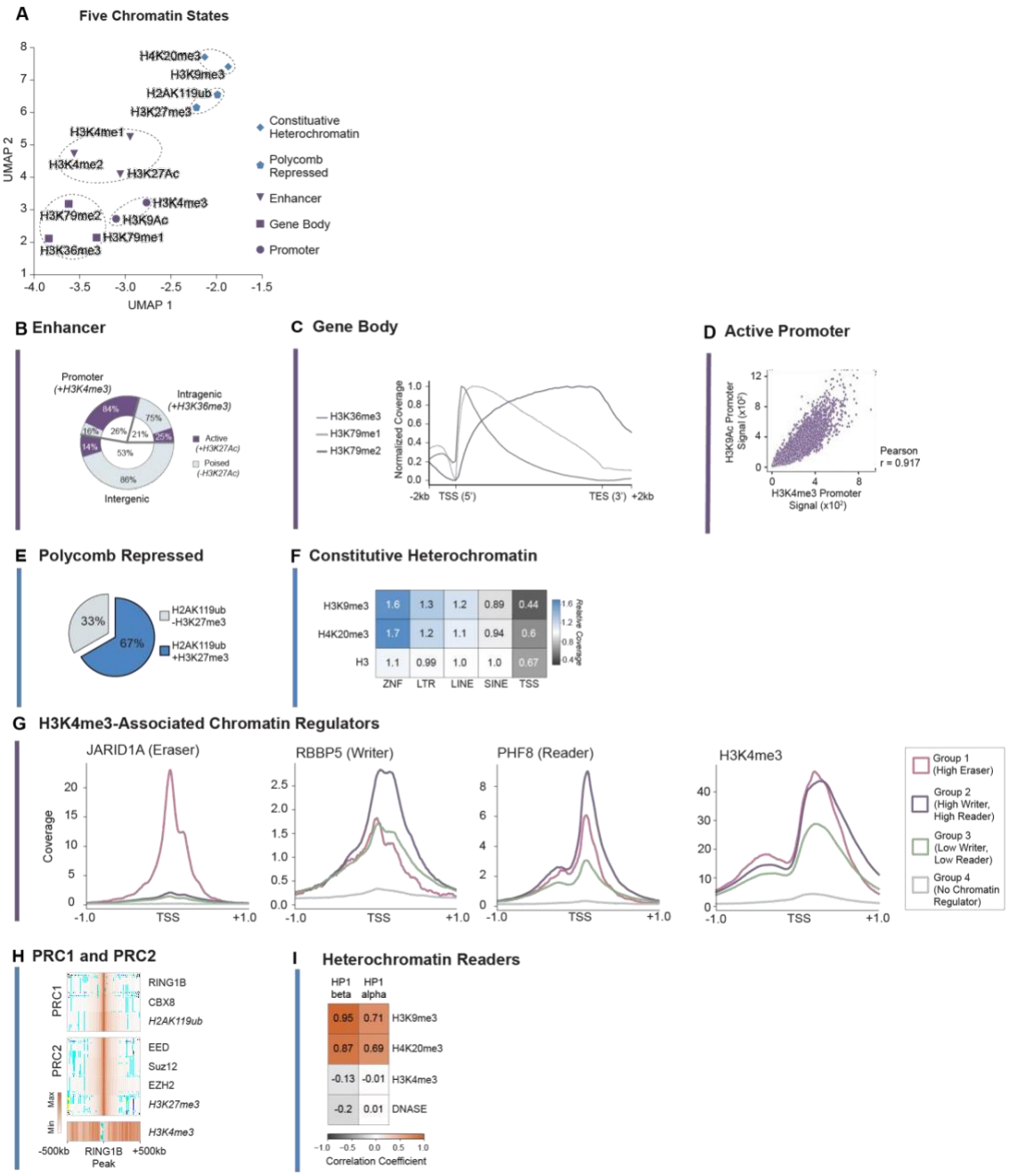

Figure S4: Histone modifications associated with five chromatin states, related to Figure 3. (A) UMAP embedding of 12 histone modifications measured in K562 correspond to five chromatin states. **(B)** Pie chart showing proportion of H3K4me1 peaks in K562 at various genomic position categories—promoters (co-localizing H3K4me3), intragenic (co-localizing H3K36me3) or intergenic—in K562. Proportion of peak regions overlapping with H3K27Ac within each genomic position category are shown in purple. **(C)** Meta plot of signal distribution of H3K36me3, H3K79me1 and H3K79me2 across the gene body of protein coding genes in K562. **(D)** Correlation scatterplot of H3K9Ac and H3K4me3

signals at promoter sites in mESC. **(E)** Pie chart showing overlap of H2AK119ub and H3K27me3 sites in K562. **(F)** Enrichment heatmap of H3K9me3 and H4K20me3 at various associated (ZNF genes, LTRs, LINES) and unassociated (SINES, TSS) genomic elements in K562. H3 is shown as reference. **(G)** Meta plots of read coverage for three H3K4me3-associated chromatin regulators (JARID1A, RBBP5, PHF8) and H3K4me3 at four promoter groups in mESC. Promoter groups were identified using k-means clustering of CR signal (see **Methods**). **(H)** Meta plot showing co-localization of multiple PRC1 and PRC2 members and their respective histone modifications at RING1B sites in K562. **(I)** Genome-wide correlation matrix of multiple HP1 proteins versus heterochromatin and euchromatin markers in K562.

Transcription factors. TFs bind *cis*-regulatory elements in combinatorial patterns to control gene expression. Generating comprehensive maps of TF localization has proven difficult because there are large numbers of distinct TFs, most are cell type-specific, and they are challenging to map by ChIP-Seq because they tend to be lower in abundance and only transiently associated with DNA[61,62]. To explore whether ChIP-DIP can map large sets of TFs, we measured 15 TFs in K562 and 43 TFs in mESC, including constitutive (e.g., SP1 and USF2)[63,64], stimulus-dependent (e.g., p53 and NRF1)[65–67], and developmental/cell type-specific (e.g., Nanog and RFX1)[68,69] DNA- binding proteins[70] **(Figure 4A)**. We obtained high-resolution binding maps for TFs in both cell types, with previously characterized TFs showing localization at their expected genomic DNA targets[63,67,71–74] (**Figure 4A-B, Supplemental Table 5**). Using these genome-wide localization data, we accurately identify DNA-binding motifs for each TF, including the 20bp dimer motif of p53[75] and the 21bp RE-1 consensus sequence of REST[76] (**Figure 4C**). Together, these data indicate that ChIP-DIP generates accurate, high-resolution binding maps of diverse TFs in multiple cell types.

Figure 4

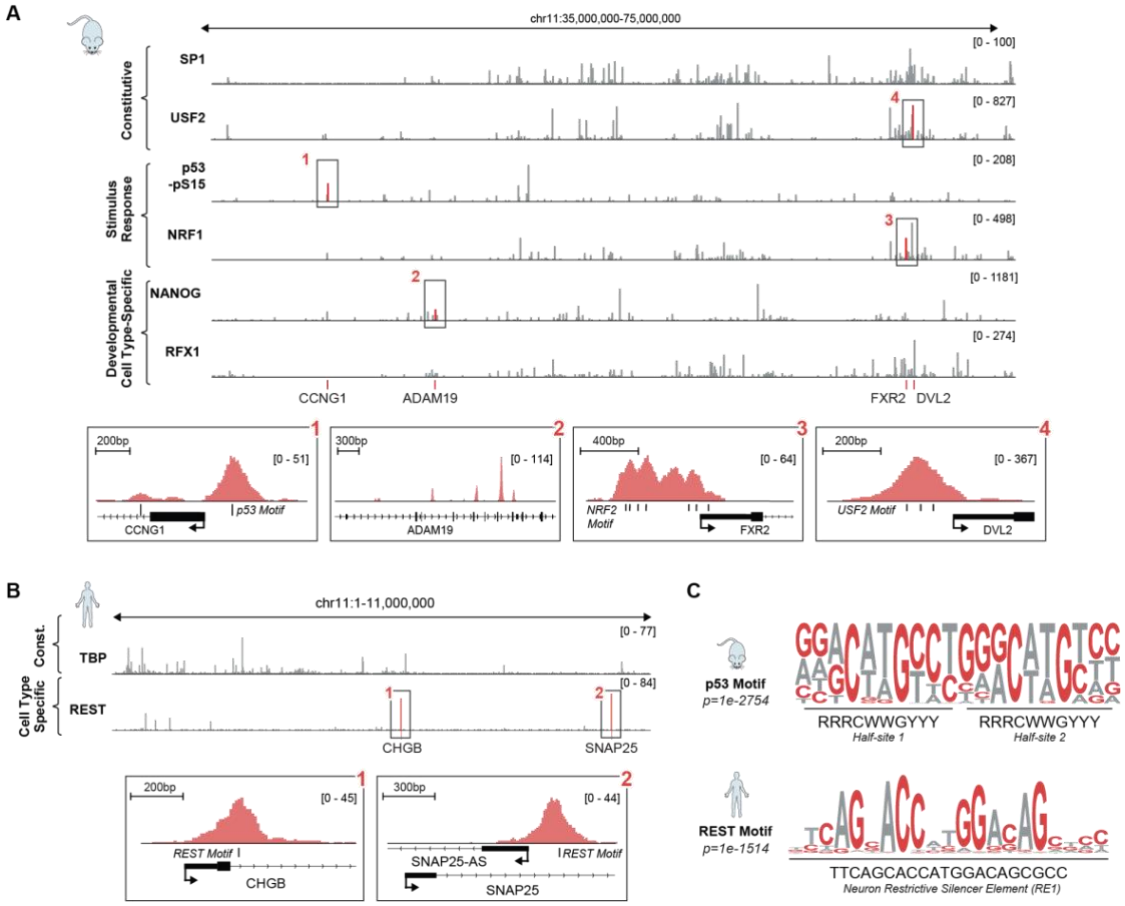

Figure 4: ChIP-DIP maps dozens of transcription factors representing diverse functional classes. (A) Top: Visualization of six transcription factors (SP1, USF2, p53-pSer15, NRF1, NANOG, RFX1) representing three broad functional classes (constitutive, stimulus-response, development/cell type-specific) across a genomic region (mm10, chr11:35,000,000–75,000,000) in mESC. Bottom: Higher-resolution zoom-ins showing individual TF binding patterns at selected targets and motif localization as appropriate. (1) p53 binding the p53 response element on the Cyclin G1 gene promoter. (2) Nanog binding a cluster of sites internal to the developmental gene ADAM19. (3) Nuclear Respiratory Factor 1 (NRF1) binding multiple copies of its motif at the promoter of FXR2.
(4) The constitutively active USF2 binding its triplicate E-box motif. **(B)** Visualization of TBP (constitutive) and REST (cell type-specific) across a genomic region (hg38, chr11:1–11,000,000) in K562 cells. Bottom: Higher-resolution zoom-ins highlight two individual peaks of RE-1 Silencing Transcription Factor/ Neuron-Restrictive Silencer Factor (REST/NRSF) at motif sites near promoters of known neuronal gene targets. **(C)** *de novo* generated motifs for p53 (top) in mESCs and REST (bottom) in K562 cells using binding sites identified using ChIP-DIP.

RNA Polymerases (RNAPs). Different classes of RNA are transcribed by distinct RNA polymerases: RNA Polymerase I (RNAP I) transcribes the 45S ribosomal RNA (rRNA) encoding the 18S, 28S, and 5.8S rRNAs; RNAP II transcribes messenger RNAs and various non-coding RNAs, including snRNAs, snoRNAs and lncRNAs; and RNAP III transcribes diverse small RNAs, including transfer RNAs, 5S rRNA, 7SL, 7SK, and U6 snRNA[8]. We leveraged the power of ChIP-DIP to simultaneously map all three RNAPs and the post-translationally modified forms of RNAP II. We observed that each RNAP localizes with high selectivity to its corresponding classes of genes; RNAP I binds at rDNA, RNAP II at mRNA and snRNA genes, and RNAP III at tRNA genes (**Figure 5A, Figure S5A)**. Moreover, we observed distinct localization patterns of different RNAP II phosphorylation states: serine 5 phosphorylated RNAP II localizes at promoters, while serine 2 phosphorylated RNAP II accumulates over the gene body and past the 3' end of the gene (**Figure S5B**). These data indicate that ChIP-DIP accurately maps the localization of the three RNA polymerases–including multiple functional phosphorylation states of RNAP II–at distinct gene classes and gene features.

Together, these results establish ChIP-DIP as a modular, highly multiplexed method that generates high-quality maps for a wide range of DNA-associated proteins spanning diverse biological functions.

Integrated analysis of protein-localization identifies functionally distinct *cis*-regulatory elements

Previous large-scale analyses have identified histone modifications that demarcate distinct genomic elements (e.g., promoters, enhancers, transcribed regions, etc.)[77], their activity state (active, inactive, repressed), and regulatory potential (poised/primed for activation)[78]. However, because of the large number of histone modifications and regulatory proteins, many efforts have focused on mapping only five histone modifications to identify genomic features and regulatory states (i.e., H3K4me3, H3K4me1, H3K36me3, H3K9me3, and H3K27me3 marking promoters, enhancers, elongated transcripts, heterochromatin, and polycomb-mediated silencing, respectively)[7]. Because ChIP-DIP can map large numbers of diverse proteins, we asked whether additional combinations of histone modifications and regulatory proteins can provide additional information about distinct types, activity states, and regulatory potentials of *cis*-regulatory elements (promoters or enhancers) beyond those captured by the five commonly studied individual histone modifications.

Promoter type and activity state are defined by combinations of histone modifications

H3K4me3 is generally thought to mark the promoters of actively transcribed RNAP II transcripts[13,41,79]. Consistent with this, we find H3K4me3 over the promoters of actively transcribed RNAP II genes, yet we also observe this modification near RNAP I promoters (ribosomal RNA) and many active RNAP III genes (tRNAs) (**Figure 5B**). Similarly, we observe that other histone modifications associated with active RNAP II promoters, including H3K4me2, H3K9Ac, H3K27Ac, and H3K56Ac, are also enriched at RNAP I

and III genes (**Figure 5B-C, Figure S5C**). For example, focusing on a genomic region containing neighboring RNAP II (i.e., histone) and RNAP III (i.e., tRNA) genes, we observe specific binding of the associated polymerase with shared transcription factors and chromatin modification patterns over all genes **(Figure 5C)**.

Although the presence of these histone modifications does not appear to distinguish between genes transcribed by different polymerases, we observed that both their position relative to the transcriptional start site (TSS) and their relative levels differ by polymerase: for RNAP I genes, these modifications localize prior to the TSS; for RNAP II, they flank the promoter and are enriched downstream of the TSS; and for RNAP III, they flank the gene body, localizing both upstream of the TSS and downstream of the transcriptional termination site **(Figure 5B, Figure S5C)**. In addition, the three RNAPs have different relative levels of these histone modifications near their respective gene promoters. For example, we found that RNAP I and II promoters display stronger H3K56Ac enrichment and RNAP I and III display stronger H3K4me2 enrichment relative to H3K4me3 **(Figure 5D)**. In this way, both quantitative combinations of histone modifications and their relative positions define distinct classes of promoters (**Figure 5E**).

Figure 5

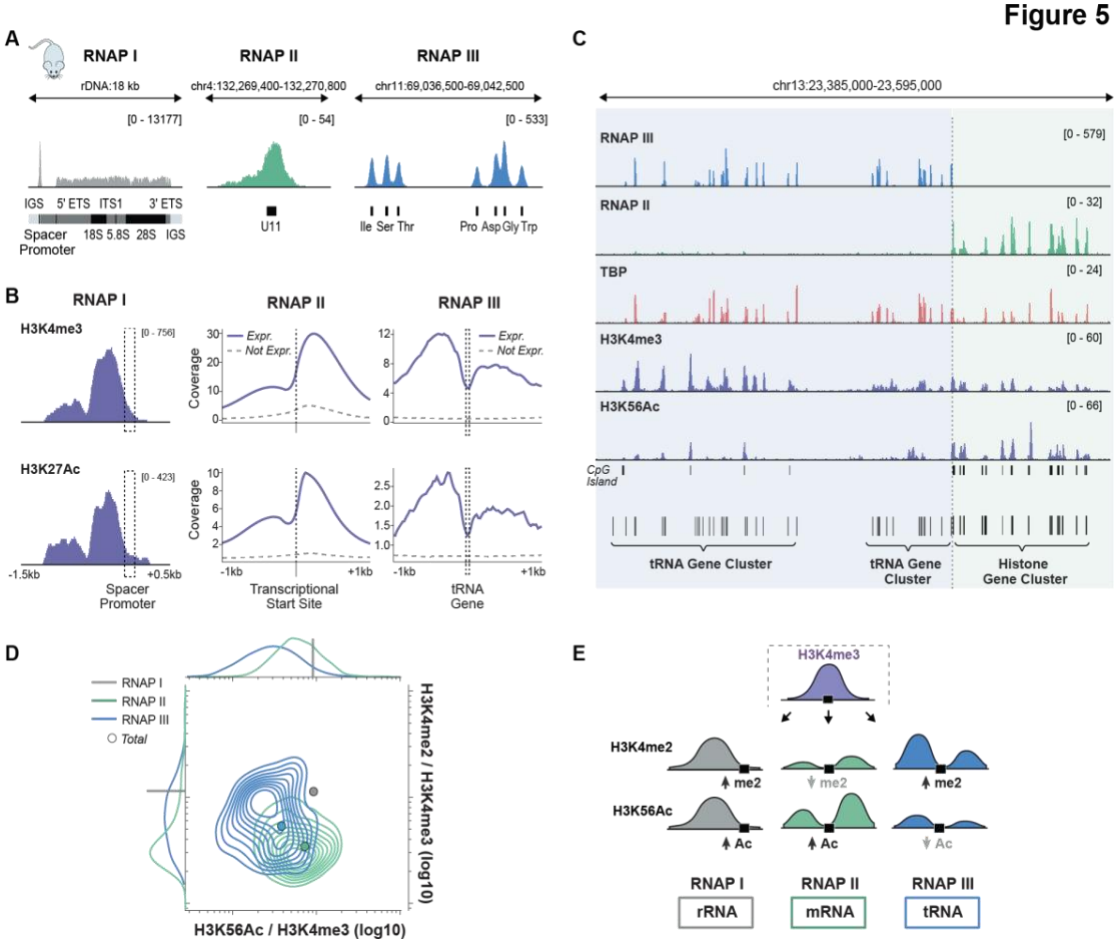

Figure 5: Distinct chromatin signatures define the promoters of each RNA Polymerase. (A) Visualization of RNAP I at the promoter and along the gene body of rDNA (left), RNAP II at a snRNA gene (middle), and RNAP III at a cluster of tRNA genes (right) in mouse ESCs. **(B)** Comparison of H3K4me3 and H3K27Ac profiles at the promoters of RNAP I, II and III genes. RNAP I is displayed over the rDNA spacer promoter (left) while RNAP II and III are displayed as meta plots across active (blue) and inactive (dashed gray) promoters. **(C)** Visualization of RNAP II and RNAP III along with the shared transcription factor TBP, and histone modifications H3K4me3 and H3K56Ac across a genomic region (mm10, chr13:23,385,000–23,595,000) containing a tRNA gene cluster (RNAP III-transcribed genes) adjacent to a histone gene cluster (RNAP II-transcribed genes), separated by dashed line. **(D)** Density distribution of H3K4me2/H3K4me3 versus H3K56Ac/H3K4me3 ratios at RNAP I, active RNAP II and active RNAP III promoters. Points show ratios when computed using the total sum of histone coverage over all promoters. Marginal distributions are shown for RNAP II and III along x and y-axis. Axes are log10 scaled. **(E)** Schematic showing relative levels of histone modifications H3K4me2 and H3K56Ac at H3K4me3 enriched regions and the associated RNAP promoter.

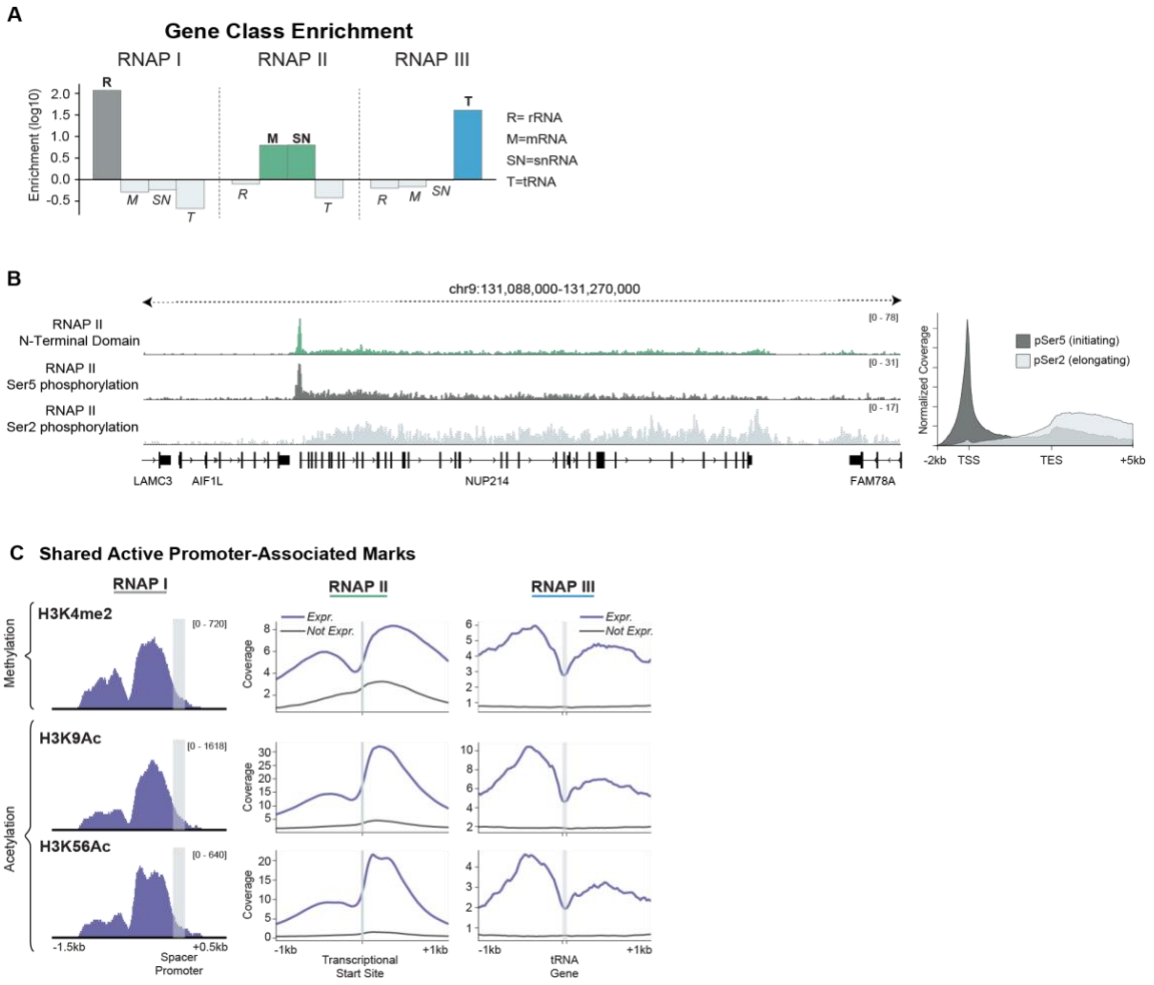

Figure S5: Chromatin states corresponding to distinct RNA polymerases and isoforms, related to Figure 5. (A) Bar graph showing enrichment of gene class coverage (rRNA, mRNA, snRNA or tRNA) for RNAP I, II and III in mESC. For each RNAP, the bar of its associated class (or classes) is highlighted. **(B)** Visualization of RNAP II phosphorylation isoforms across the NUP214 gene in K562 (left). Meta plot of signal distribution of RNAP II phosphorylation isoforms across the gene body of protein coding genes in K562 (right). **(C)** Comparison of histone profiles for H3K4me2, H3K9Ac and H3K56Ac at the promoters of RNAP I, II and III, similar to **Figure 5B**. (Left) Histone modification over the RNAP I-transcribed rDNA spacer promoter. (Middle/Right) Meta plot of histone profiles at active (blue) and inactive (gray) promoters for RNAP II (middle) and RNAP III (right).

Next, we considered whether other histone modifications may distinguish activity states of RNAP II promoters. To explore this, we quantified the levels of ten additional histone modifications at each genomic region containing H3K4me3 and grouped them using hierarchical clustering. We identified five sets of H3K4me3 enriched genomic regions; four are enriched with other histone modifications (sets 1–4) and one is not (set 5). The four co-occurring sets correspond to H3K4me3 along with: H3K27me3/H2AK119ub (set 1), H3K36me3/H3K79me (set 2), H3K9me3/H4K20me3 (set 3), or H3K4me1/H3K27ac (set 4) (**Figure 6A**). These sets correspond to promoters that exhibit distinct transcriptional activity (e.g., high versus low expression) (**Figure 6B**) and are enriched for distinct classes of RNAP II-transcribed genes, such as ribosomal protein and cell cycle genes (set 2), zinc finger protein (set 3) and long intergenic ncRNAs genes (sets 3 and 4) (**Figure 6C-G**)[12,80]. Consistent with the fact that H3K4me3 localization associates with functionally distinct classes of promoters, we found that different combinations of chromatin regulators that read, write, and erase H3K4me3 localize at distinct promoters (**Figure S4G**).

Taken together, these results demonstrate that combinations of histone modifications can distinguish promoter features including polymerase **(Figure 5E),** gene type, and activity level (**Figures 6H**).

Figure 6

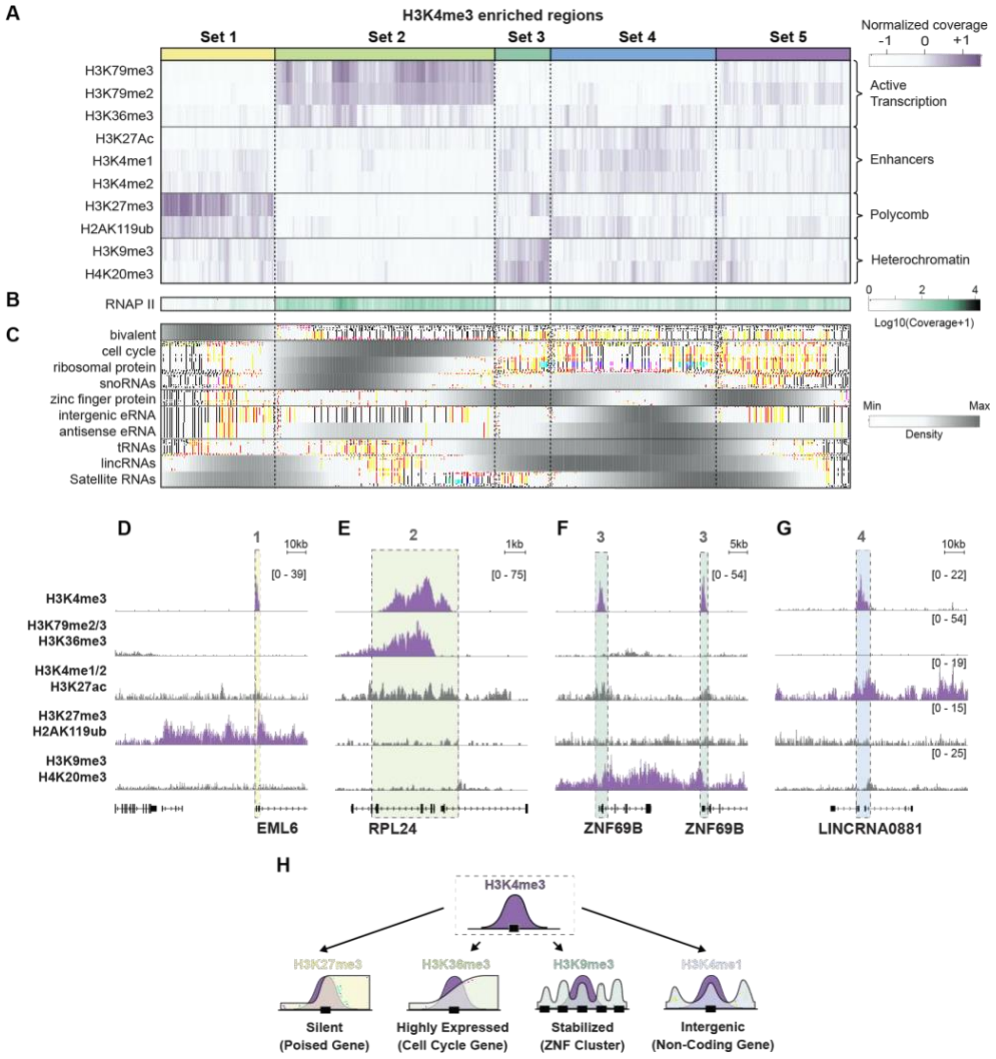

Figure 6: Combinations of histone modifications distinguish RNAP II promoter type, activity and potential. (A) Hierarchically clustered heatmap of coverage levels of ten different histone modifications (y-axis) at individual H3K4me3 enriched genomic regions (x-axis). Five distinct sets of regions are indicated by colored bars along top-axis. **(B)** RNAP II coverage at H3K4me3-enriched regions, as sorted in (A). **(C)** Gene density of ten different gene classes at H3K4me3 enriched regions, as sorted in (A). **(D)** Visualization of H3K4me3 and H3K27me3/H2AK119ub (associated with set 1) across the EML5 gene in K562. **(E)** Visualization of H3K4me3, H3K79me2/H3K79me3/H3K36me3 co-localization (associated with set 2) across the ribosomal protein gene RPL24 in K562. **(F)** Visualization of H3K4me3 and H4K20me3/H3K9me3 co-localization (associated with set 3) across neighboring zinc finger genes in K562. **(G)** Visualization of H3K4me3 and H3K4me1/H3K4me2/H3K27Ac (associated with set 4) across the long intergenic noncoding RNA gene LNCRNA0881. For tracks in (**D**–**G**), the non-H3K4me3 tracks represent the sum of histone tracks associated with each set and are scaled to the maximum value across all panels. H3K4me3 tracks are scaled to the maximum for each panel. **(H)** Illustration

summarizing the co-occurring promoter-associated histone modifications and their associated gene groups.

Enhancer type, activity and potential are defined by combinations of histone modifications

There are >40 different histone acetylation marks[3], many of which have been associated with enhancers and active transcription. We mapped 15 of these in mESCs, including marks on all four core histones and histone variants, and observed that they co-localize at similar sites genome-wide (Pearson r = 0.86–0.97)[81] (**Figure S6**). We considered whether these strong correlations indicate that these marks are redundant or whether there is additional regulatory information encoded by the relative levels of each acetylation mark at specific genomic sites. To explore this, we used a matrix factorization algorithm to define five weighted combinations of acetylation marks at highly acetylated genomic regions (quantitative combinations C1–C5; see **Methods, Figure 7A–B, Supplemental Note 2, Figure S7)**. These quantitative combinations correspond to genomic regions that contain distinct transcription factor and chromatin regulator binding profiles **(Figure 7C–F, Figure S8)**.

Active promoter-proximal elements. The first group of regions (C1) is defined by H3K9Ac and several other H3 acetylation marks (H3K14ac, H3K18ac, H3K36ac, H3K56ac, and H3K79ac) (**Figure 7B)**. Genomic regions containing this signature tend to be localized near the promoter region of transcribed genes and are enriched for RNAP II, general TFs (e.g., TFIIB), and other CpG-island associated factors (e.g., E2F1, CXX1) along with their sequence motifs (e.g., ETS, SP and NRF families) **(Figure 7C, E–F)**.

Poised promoter-proximal elements. The second group of regions (C2) contains high levels of H3K9Ac and acetylation of the histone variant H2AZ (H2AZAc) (**Figure 7B**). Genomic regions containing this signature tend to have lower levels of RNAP II (relative to C1) and are strongly enriched for polycomb (JARID2, SUZ12, RING1B) and other repressive chromatin regulators (KDM2B, HDAC2) **(Figure 7E–F)**.

Stress and signaling response elements. The third group of regions (C3) contains high levels of H2AZAc and H4Ac (**Figure 7B)**. Genomic regions containing this signature are also enriched for RNAP II but are bound by p53 and contain other stress response motifs (e.g., BACH1, NRF2) or signaling response motifs (e.g., CRE) (**Figure 7C, Figure 7E–F**). Consistent with these observations, H2AZ has been proposed as a facilitator of inducible transcription (e.g., signaling pathway responses and p53 regulation)[82–85]. Yet, because

H2AZ is also a component of C2, our results suggest that this association is not solely a property of H2AZAc but of this unique C3 signature.

Active pluripotency distal regulatory elements. The fourth group of regions (C4) is defined by H2BK20Ac and H3K27Ac (**Figure 7B)**. Genomic regions containing this signature tend to be promoter distal (**Figure S8B**) and are associated with actively transcribed embryonic and stem-cell-specific genes (**Figure 7D)**. These regions are enriched for binding of the pluripotency TFs, including Nanog, Oct4, and Sox2, as well as the P300 acetyltransferase and components of mediator **(Figure 7F)**.

Poised differentiation distal regulatory elements. The fifth group of regions (C5) is

defined by high levels of H2BK20Ac (similar to C4) and H3K14Ac (distinct from C4) (**Figure 7B**). Interestingly, these regions displayed similar TF and CR occupancy (e.g., Oct4, Sox2, Nanog, P300 and mediator) to C4 regions **(Figure 7F–G)**[86]. However, in contrast to C4 regions bound by these pluripotency factors, which correspond to enhancers of active genes involved in embryo and stem cell function, C5 regions bound by these factors correspond to enhancers of genes involved in post-embryonic development **(Figure S9)** and are enriched for sequence motifs of TFs involved in lineage specification and morphogenesis (e.g., TEAD family)[87] **(Figure S8C)**. This suggests that C5 enhancers might be important in establishing the gene expression program needed upon differentiation (regulatory potential). Interestingly, we identified a third set of genomic regions that also contain a high density of pluripotency TFs but lack the C4 or C5 acetylation signatures; these are associated with genes involved in later stages of organogenesis (e.g., kidney and sensory systems) (**Figure S9)**.

These analyses indicate that histone acetylation is not a redundant marker of enhancers, but that combinations of acetylation modifications can define unique classes of *cis*-regulatory elements (promoter-proximal versus distal enhancers) that act in distinct ways (stimulus-responsive versus developmentally regulated) and that exhibit different activity (e.g., active gene expression versus poised for activation upon differentiation) (**Figure 7H**).

Overall, these observations highlight the importance of multi-component analyses and demonstrate why ChIP-DIP provides a powerful approach that will be critical for defining unique regulatory features within distinct cell states.

Figure 7

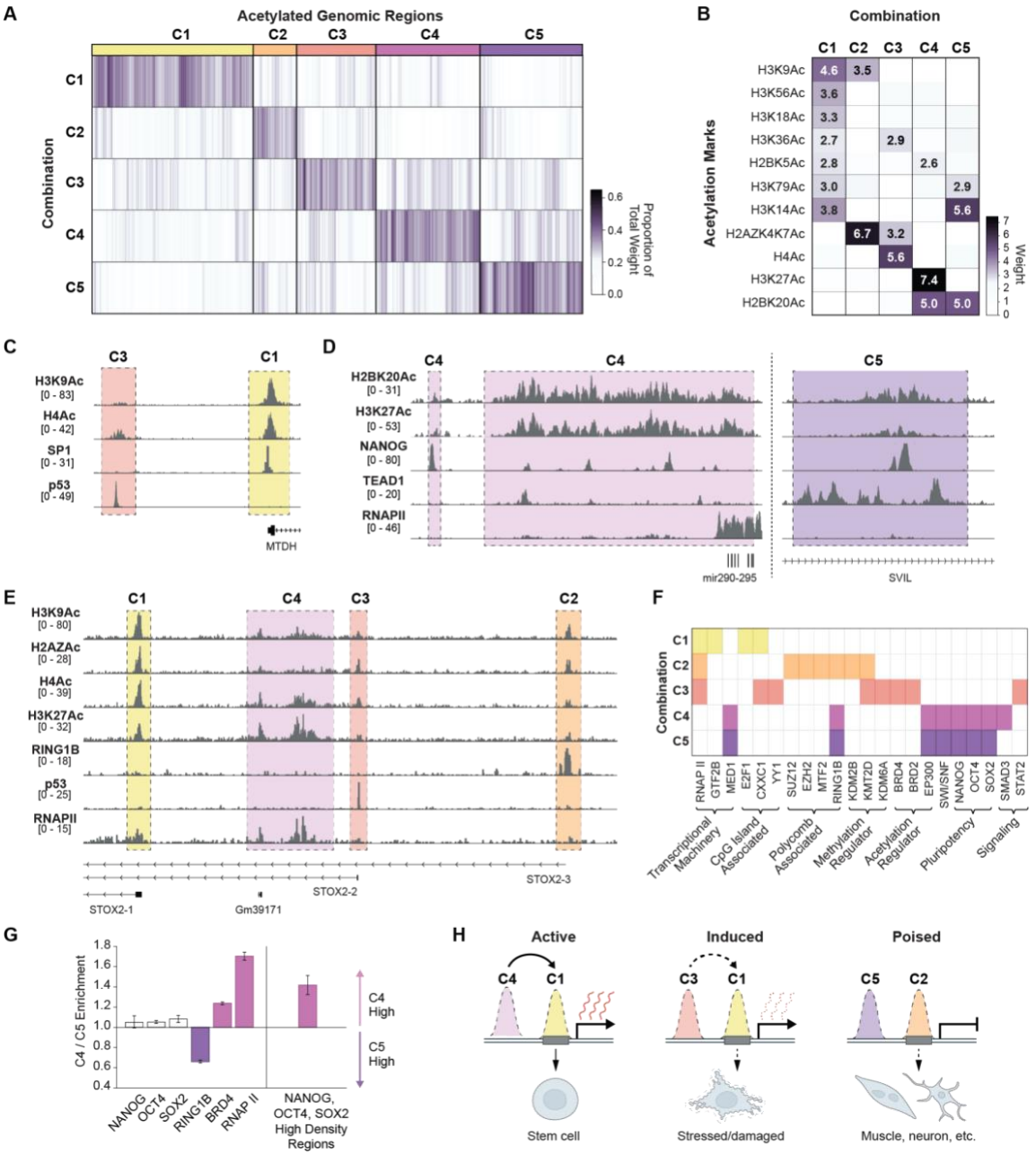

Figure 7: Distinct combinations of histone acetylation marks define unique enhancer types that differ in their activity and developmental potential. (A) The relative weights of five different combinations of histone acetylation marks (C1-C5, y-axis) for each acetylated genomic region (x-axis). Regions are grouped according to the combination that received the greatest weight, as indicated along top-axis. **(B)** The relative weights of each histone acetylation mark (y-axis) within each combination (x-axis). Only weights greater than 2.5 are labeled. **(C)** Visualization of H3K9ac and H4ac along with SP1 and P53 across a genomic region (mm10, chr15:34,065,000–34,086,000) containing enhancers assigned to

the C1 (yellow) and C3 (red) state. **(D)** Visualization of H2BK20Ac and H3K27Ac along with Nanog, Tead1, and RNAP II across a genomic region (left: mm10, chr7:3,191,500–3,221,500, right: mm10, chr18:5,006,500–5,016,500) containing enhancers assigned to C4 (left) and a distinct region assigned to C5 (right). (Scale of the Nanog track is capped to the maximum of the left region; Tead1 data is from published ChIP-Seq data in fetal cardiomyocytes (Akerberg et al., 2019)). **(E)** Visualization of H3K9Ac, H2AZAc, H4Ac along with RING1B, P53, and RNAP II over a genomic region (mm10, chr8:47,272,800–47,427,000) containing enhancers assigned to states C1-C4. **(F)** DNA-associated proteins (x-axis, ordered by function) with significant binding at genomic regions defined by each combination (y-axis) are indicated in color (see **Methods**). (**G**) Enrichment bar graph of selected transcription-associated factors or regions with high density of pluripotency TFs (see **Methods**) in C4 vs. C5 associated regions. Error bars correspond to the enrichment range from bootstrap resampling. **(H)** Schematic of C1-C5 associated regions and their corresponding functions.

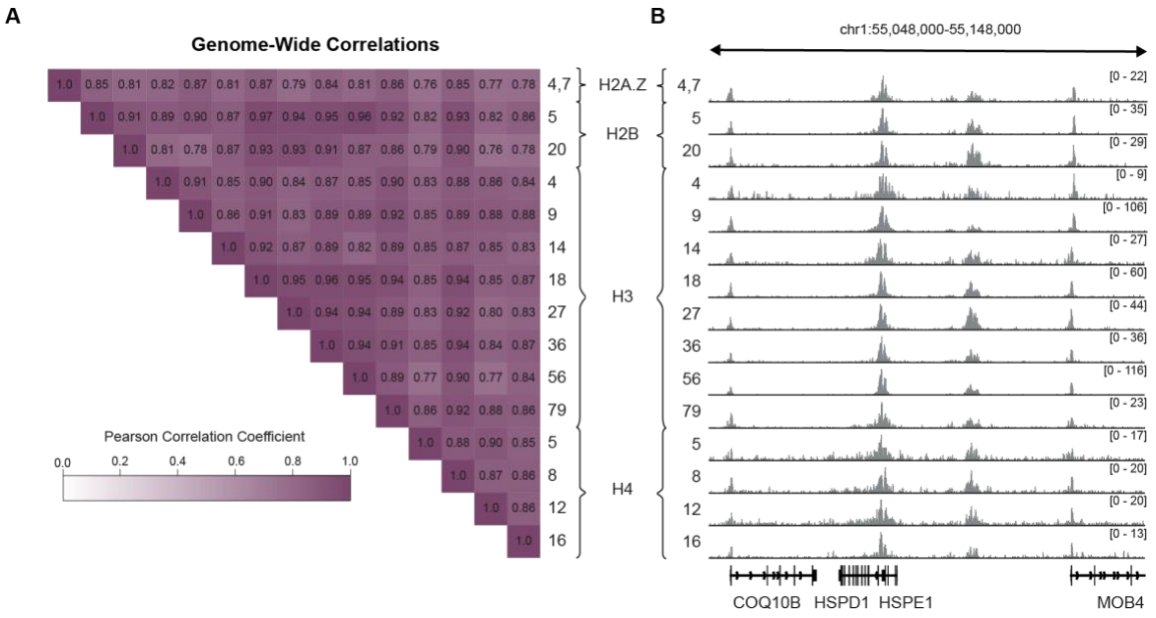

Figure S6: Histone acetylation marks are highly correlated genome-wide, related to Figure 7. (A) Genome-wide Pearson correlation of fifteen different histone acetylation marks in mESC. Correlations are based on coverage computed in 10kB windows. **(B)** Comparison of fifteen different histone acetylation marks across a genomic region (mm10, chr1:55,048,000–55,148,000) in mESC.

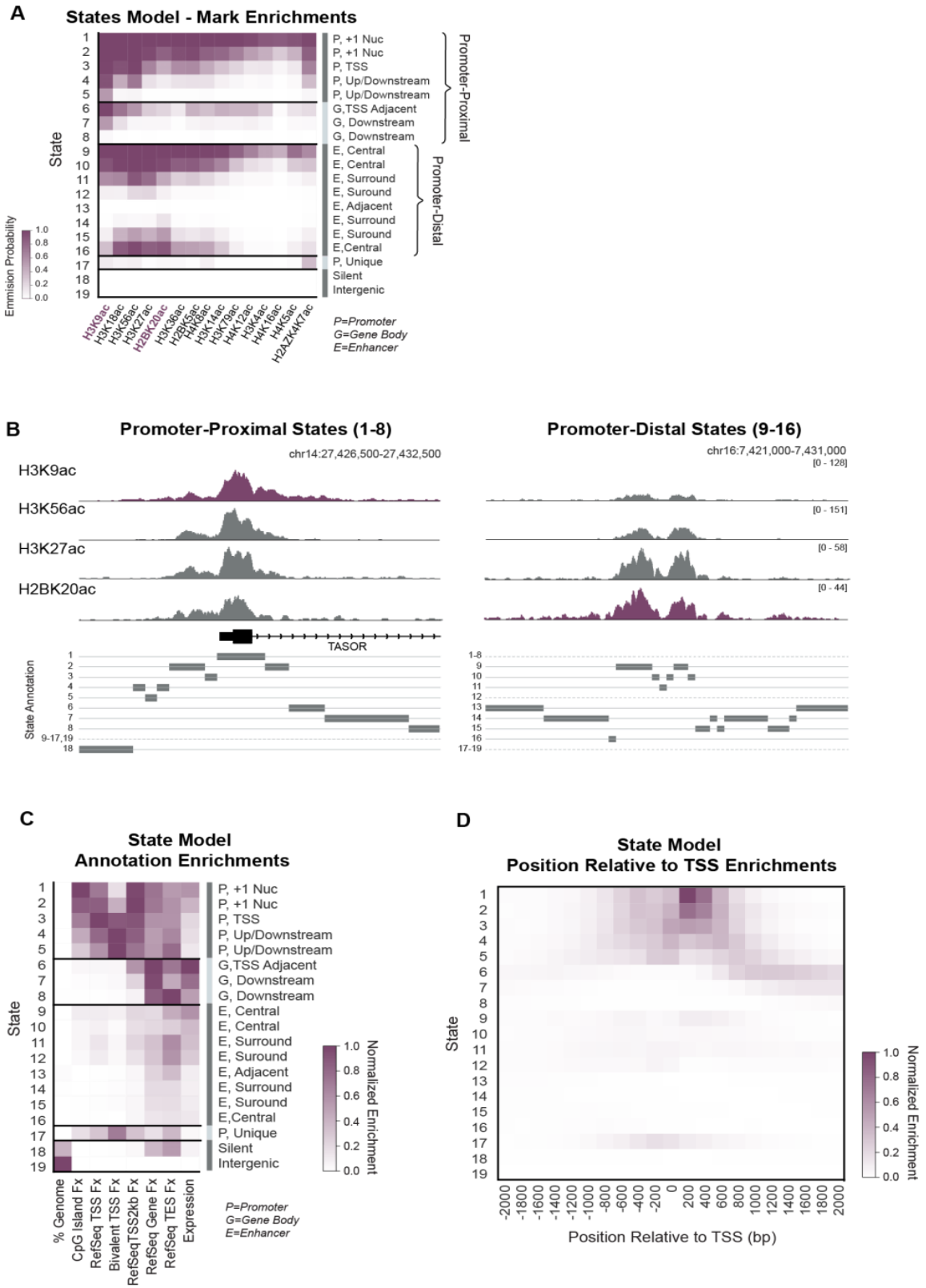

Figure S7: ChromHMM model using histone acetylation marks, related to Figure 7.

(A) Histone acetylation mark emission probability matrix for 19-state ChromHMM model. State annotations (right) were assigned manually based on genomic position enrichments of states. **(B)** Track visualization of histone acetylation marks (top) and chromatin-state annotations (bottom) at example promoter region (left) versus example intergenic region (right). Histone acetylation marks are scaled to the same maximum values at both regions. At each region, the chromatin states that are present are shown with solid lines and a box indicating the exact position; chromatin states that are absent are listed next to dotted lines. **(C)** Heatmap of genome annotation enrichment of chromatin states. Enrichment scores are normalized to the maximum and minimum of each column. **(D)** Heatmap of genomic position enrichment relative to the TSS of chromatin states. Enrichment scores are normalized to the maximum and the minimum of the heatmap.

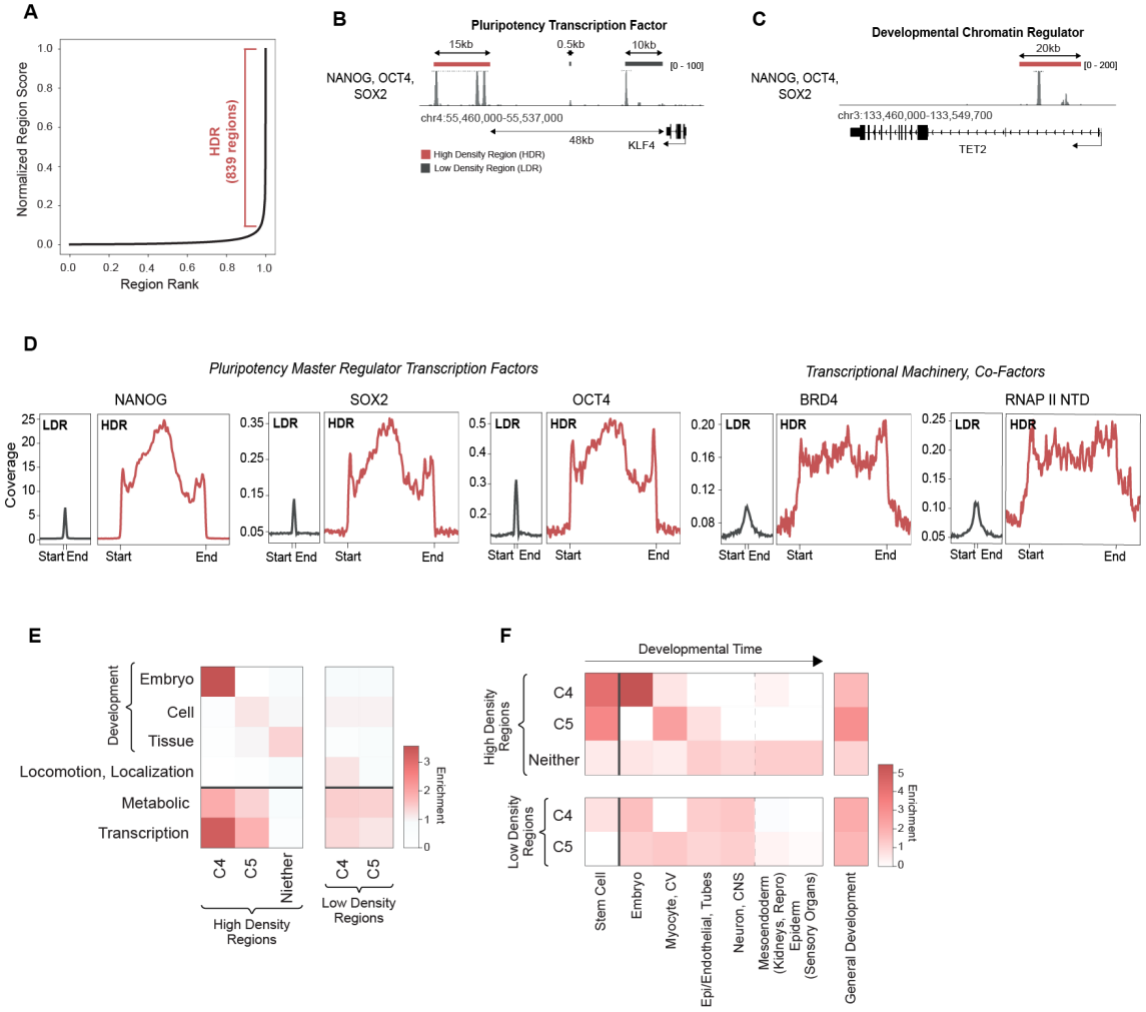

Figure S9: Profiles for high-density regions of NANOG-OCT4-SOX2, related to Figure 7. (A) Plot showing normalized region scores (x-axis) for peak regions of NANOG-OCT4-SOX2, ordered by rank (y-axis). High-density regions are defined as regions past the point where the slope = 1. **(B)** Track visualization of NANOG-OCT4-SOX2 upstream of the gene for KLF4, a pluripotency transcription factor, in mESC. A high-density region is indicated with a red bar; low-density regions are indicated with gray bars. **(C)** Visualization of NANOG-OCT4-SOX2 near the TET2 gene, a developmentally associated chromatin regulator, in mESC. A high-density region internal to the gene is indicated with a red bar. **(D)** Coverage meta plots over low-density regions (LDR) vs. high-density regions (HDR) for pluripotency transcription factors and other transcriptional-related factors. Metagenes are centered on the region and the lengths represent the approximate difference in mean lengths (500bps for LDRs and 14,500bps for HDRs). An additional 4kb surrounding each region is shown. **(E)** Enrichment heatmap for GO terms of genes associated with HDRs or LDRs containing C4, C5 or neither C4/C5 chromatin signatures. **(F)** Enrichment heatmap for development-associated GO terms of genes associated with HDRs or LDRs containing C4, C5 or neither C4/C5 chromatin signatures.

Discussion

We demonstrated that ChIP-DIP enables highly multiplexed mapping of hundreds of regulatory proteins to genomic DNA in a single experiment. Although the largest ChIP-DIP experiment in this study contained >225 distinct antibodies, this number was primarily limited by the availability of high-quality antibodies and we expect that ChIP-DIP could be performed using even larger pools of antibodies. Because this approach employs standard molecular biology techniques, we expect that it will be readily accessible to any lab without the need for specialized training or equipment. As such, we anticipate that ChIP-DIP will enable a fundamental shift from large consortia generating reference maps for a limited number of cell types to individual labs generating cell-type-specific maps within any specific experimental system of interest.

Given the important information encoded within quantitative combinations of histone modifications, chromatin regulators, and transcription factors, comprehensively mapping these factors across cell types will be critical for studying gene regulation and for defining the putative effects of genetic variants associated with human disease. For instance, while specific regulatory states have been shown to be encoded by combinations of histone modifications (e.g., bivalent domains), the details of such states (i.e., number and cell-type specificity) have remained largely unexplored. The large numbers of histone modifications and regulatory proteins have until now necessitated a tradeoff between mapping many marks in a few cell types or a few marks in many cell types. ChIP-DIP overcomes this by mapping hundreds of proteins in a single experiment. Moreover, due to the nature of split-pool barcoding used in ChIP-DIP and because there is negligible antibody–bead–chromatin dissociation during the procedure, ChIP-DIP can also be used to map protein binding within multiple samples simultaneously using distinct sets of antibody-oligo-labeled beads. While we did not directly emphasize this capability in this paper, several ChIP-DIP experiments described here were performed simultaneously by barcoding multiple sample conditions within the same experiment (e.g., crosslinking conditions, IP conditions). In addition to the increase in scale provided by mapping multiple proteins within a single sample, ChIP-DIP also reduces many sources of technical and biological variability associated with processing individual proteins and samples. This ability to measure regulatory proteins at scale, in multiple cell conditions, and with reduced sources of variability will enable large-scale mapping of dynamic protein localization across distinct cell types and timepoints.

Beyond the multiplexing applications highlighted in this work, ChIP-DIP can be directly integrated into multiple existing split-pool approaches to create additional capabilities that are not currently possible. For example, we previously showed that we can map the 3D genome structure surrounding individual protein binding sites (SIP)[88]; integrating this approach with ChIP-DIP will enable mapping of the 3D structures that occur at hundreds of distinct protein binding sites simultaneously. Moreover, we previously developed a method to map 3D genome contacts within thousands of individual single cells using this same split-pool approach[35]. Integrating this single-cell approach with ChIP-DIP will enable comprehensive mapping of hundreds of regulatory protein binding sites within many thousands of individual cells. Finally, we previously showed that split-and-pool barcoding

can be used to simultaneously map the spatial proximity of DNA and RNA to measure ncRNA localization and the levels of nascent RNA transcription at individual DNA sites[37]. Integrating this approach with ChIP-DIP will enable the direct measurements of protein binding and transcriptional activity at individual genomic locations, providing a direct link between binding events and the associated transcription activity within the same cell. For these reasons, we expect that ChIP-DIP will represent a transformative new tool for dissecting gene regulation.

Supplemental note

ChromHMM Model of histone acetylation states. To investigate the spatial relationships between histone acetylation marks, we generated a 19-state genome segmentation model using ChromHMM and fifteen different histone acetylation marks (**Figure S7A**). Based on the transition probabilities, we grouped these nineteen states and found two large sets that differed in their genome positioning: set 1 (States 1–8) was promoter proximal, with the individual states identifying the relationship to the TSS, while set 2 (States 9–16) was promoter distal, with the individual states demarcating the acetylation peaks and surroundings between

them (**Figure S7B–D**). State 17 was also promoter proximal but was grouped separately because of a unique signature with H2AZAc (**Figure S7A, D**). Finally, States 18 and 19 corresponded to silent genic and intergenic regions. Remarkably, this model found that histone acetylation marks were sufficient to define multiple functional elements (e.g., promoters, enhancers, gene bodies, silent, intergenic).

Overall, our state-model found that the acetylation marks cover similar genomic loci; there exist multiple states that have nearly all the marks (i.e., State 1, 2, 9 and 10) and multiple states that appear alike in composition but differ in relationship to genomic annotations (i.e., State 1 vs. State 9) (**Figure S7A, D**). Comparing sets 1 and 2 (i.e., promoter proximal vs. promoter distal), we found that all acetylation marks are present in both sets, however, some marks are more enriched in one set over the other. For example, H3K9Ac was strongest in set 1 (the promoter-proximal set), while H3K18Ac, H3K27Ac and H2BK20Ac were enriched throughout set 2 (the promoter-distal set) (**Figure S7A**). Notably H3K18Ac, H3K27Ac and H2BK20Ac are all targets of the CBP/p300 acetyltransferase[89], which is strongly associated with activity at enhancers. In contrast, the GCN5/PCAF subfamily preferentially acetylates H3K9, H3K14 and H4[90]–all marks we see preferentially in set 1. Comparing all nineteen states individually, we found that a small subset of states had greater selective enrichment for specific histone modifications. For example, State 5 (promoter up/down stream) and State 7 (gene body) both had greater selectivity for H3K9Ac, State 14 was defined by H2BK20Ac and State 17 was defined by H2AZAc. Such states may indicate subtle positional shifts or locations unique to these marks. Correspondingly, by visual comparison, we saw that H3K9Ac appears more enriched downstream of the TSS and into the gene body than other histone acetylation marks.

Our 19-state ChromHMM genome segmentation results corresponded well with the findings of our weighted combinatoric NMF model (**Figure 7**). In our NMF model, we found that TSS associated combinations (C1 and C2) are defined by H3K9Ac; similarly, in our ChromHMM model, we found that H3K9Ac preferred the promoter-associated set of states (set 1). In our NMF model, we predicted a unique role for H2AZAc in defining multiple promoter-associated combinations (C2 and C3); in our ChromHMM model, we found H2AZAc was selectively enriched in State 17, a promoter-associated state. In our NMF model, we found that promoter-distal combinations are defined by H2BK20Ac and H3K27Ac (C4 and C5); in our ChromHMM model, we found that these two marks prefer the promoter-distal set of states (set 2).

Methods

Cell Lines, Cell Culture and Crosslinking

Cell lines used in this study. We used the following cell lines in this study: (i) Female mouse

ES cells (pSM44 mES cell line) derived from a 129 × castaneous F1 mouse cross and (ii) K562, a female human lymphoblastic cell line (ATCC, Cat # CCL-243).

Cell Culture Conditions

(i) pSM44 mES cells were grown at 37°C under 7% CO2 on plates coated with 0.2% gelatin (Sigma, G1393-100ML) and 1.75 mg/mL laminin (Life Technologies Corporation, #23017015) in serum-free 2i/LIF media composed as follows: 1:1 mix of DMEM/F-12 (GIBCO) and Neurobasal (GIBCO) supplemented with 1x N2 (GIBCO), 0.5x B-27 (GIBCO 17504-044), 2 mg/mL bovine insulin (Sigma), 1.37 mg/mL progesterone (Sigma), 5 mg/mL BSA Fraction V (GIBCO), 0.1 mM 2-mercaptoethanol (Sigma), 5 ng/mL murine LIF (GlobalStem), 0.125 mM PD0325901 (SelleckChem) and 0.375 mM CHIR99021 (SelleckChem). 2i inhibitors were added fresh with each medium change. Fresh medium was replaced every 24–48 hours depending on culture density, and cells were passaged every 72 hours using 0.025% Trypsin (Life Technologies) supplemented with 1mM EDTA and chicken serum (1/100 diluted; Sigma), rinsing dissociated cells from the plates with DMEM/F12 containing 0.038% BSA Fraction V.

(ii) K562 cells were purchased from ATCC and cultured in 1x DMEM (Life Technologies, # 11965118), 10% FBS (VWR, #97068-091), 100U/mL Penicillin/Streptomycin (Life Technologies, # 15140122), 1mM Sodium Pyruvate (ThermoFisher, #11360070), 2mM L-Glutamine (Life Technologies # 25030081) at 37°C and 5% CO in 15cm plates (USA Scientific # 5663-9160Q).

Cell Harvest

(i) For harvesting pSM44 mESCs, cells were trypsinized by adding 5 mL of TVP (1 mM EDTA, 0.025% Trypsin, 1% Sigma Chicken Serum; pre-warmed at 37°C) to each 15 cm plate and rocking gently for 3–4 min until cells start to detach. 25 mL of wash solution (DMEM/F-12 supplemented with 0.03% GIBCO BSA Fraction V, pre-warmed at 37°C) was added to each plate to inactivate the trypsin. Detached cells were transferred into 50 mL conical tubes, pelleted at 330 g for 3 min, washed in 4 mL of 1X PBS per 10 million cells and then pelleted in 1X PBS in preparation for crosslinking. (ii) For harvesting K562s, the cell suspension was transferred to 50mL conical tubes, pelleted at 330 g for 3 min, washed with 4 mL of 1X PBS per 10 million cells and then pelleted in 1X PBS in preparation for crosslinking.

Cell Crosslinking

Cells were crosslinked in suspension with 1% Formaldehyde for 10 min at room temperature. For both cell lines, during crosslinking steps and subsequent washes, volumes were maintained at 4 mL of buffer or crosslinking solution per 10 million cells. Pelleted cells were resuspended in 1ml of 1X PBS per 10 million cells and pipetted to disrupt clumps of cells. Next, cells were crosslinked in suspension in a final volume of 4 mL of 1% formaldehyde (FA Ampules, Pierce 28906) diluted in 1X PBS per 10 million cells and rocked gently for 10 min at room temperature. Formaldehyde was immediately quenched with addition of 200 ml of 2.5 M glycine (Sigma G7403-250G) per 1 mL of 1% FA solution and incubated with gentle rocking for 5 min at room temperature. Cells were then washed three times with 0.5% BSA in 1X PBS that was kept at 4°C. Finally, aliquots of 10 million cells were prepared in 1.7 mL tubes; these cell aliquots were pelleted, flash frozen in liquid nitrogen and stored in -80°C until lysis.

Nuclear Isolation and Chromatin Preparation

Nuclear Isolation

Crosslinked cell pellets (10 million cells) were lysed using the following nuclear isolation procedure: cells were incubated in 0.7 mL of Nuclear Isolation Buffer 1 (50 mM HEPES pH 7.4, 1 mM EDTA pH 8.0, 1 mM EGTA pH 8.0, 140 mM NaCl, 0.25% Triton-X, 0.5% NP-40, 10% Glycerol, 1X PIC) for 10 min on ice. Cells were pelleted at 850 g for 10 min at 4C. Supernatant was removed, 0.7 mL of Lysis Buffer 2 (50 mM HEPES pH 7.4, 1.5 mM EDTA, 1.5 mM EGTA, 200 mM NaCl, 1X PIC) was added and the sample was incubated for 10 min on ice. Nuclei were obtained after pelleting and supernatant was removed (as above). Then, 550 uL of Lysis Buffer 3 (50 mM HEPES pH 7.4, 1.5 mM EDTA, 1.5 mM EGTA, 100 mM NaCl, 0.1% sodium deoxycholate, 0.5% NLS, 1X PIC) was added and the sample was incubated for 10 min on ice prior to sonication.

Chromatin fragmentation and size analysis.

Chromatin was fragmented via sonication of the nuclear pellet using a Branson needle-tip sonicator (3 mm diameter (1/8'' Doublestep), Branson Ultrasonics 101-148-063) at 4C for a total of 2.5 min at 4–5 W (pulses of 0.7 s on, followed by 3.3 s off). To check the resulting DNA size distribution, a small aliquot of 20uL of sonicated lysate was then added to 80uL of Proteinase K buffer ((20 mM Tris pH 7.5, 100 mM NaCl, 10 mM EDTA, 10 mM EGTA, 0.5% Triton-X, 0.2% SDS) and reverse crosslinked at 80°C for 30 min. DNA was isolated using Zymo IC DNA Clean and Concentrator columns and eluted in water. 10uL of purified DNA was then run for 10 min on a 1% e-gel (Invitrogen™ E-Gel™ EX Agarose Gels, 1%, Cat.No. G402021). Fragments were found to be 150–700 bp with an average size of roughly 350 bp. The remaining chromatin prep was stored at 4C overnight to be used for the immunoprecipitation the next day.

ChIP-DIP: Bead Preparation

Antibody-ID oligo design

Antibody-ID oligos were designed and ordered from IDT (**Supplemental Table 6**). The sequence is as follows:

/5phos/TGACTTGNNNNNNNNNTATTATGGTAGATCGGAAGAGCGTCGTGTACAC

AGAGTC/3Bio/.

This corresponds to a sticky end that ligates Odd barcodes, UMI, antibody barcode, Illumina primer binding site (i5 primer binding site), spacer sequence. The oligo contains a 5' phosphate to enable ligation and a 3' biotin to enable coupling to beads.

Protein G Bead biotinylation

1 mL of Protein G Dynabeads (ThermoFisher, #10003D) were washed once with 1X PBSt (1X PBS + 0.1% Tween-20), separately keeping the original storage buffer, and resuspended in 1mL PBSt. Beads were then incubated with 20 µL of 5 mM EZ-Link Sulfo-NHS-Biotin (Thermo, #21217) on a HulaMixer for 30 min at room temperature. To quench the NHS reaction, beads were placed on a magnet, 500 µL of buffer was removed and replaced with 500 µL of 1M Tris pH 7.4 and beads were incubated on the HulaMixer for an additional 30 min at room temperature. Beads were then washed twice with 1 mL PBSt and resuspended in their original storage buffer until use.

Preparation of streptavidin-coupled oligonucleotides

Biotinylated antibody-ID oligonucleotides were coupled to streptavidin (BioLegend, #280302) in a 96-well PCR plate. In each well, 20 µL of 10 µM oligo was added to 75 µL 1X PBS and 5 µL 1 mg/mL streptavidin to make a 909 nM (calculated from the molarity of streptavidin molecules) stock. The 96-well plate was incubated with shaking at 1600 rpm on a ThermoMixer for 30 min at room temperature. Each well was diluted 1:4 in 1X PBS for a final concentration of 227 nM.

Preparation of oligonucleotide-coupled Protein G beads

For each antibody in the experiment, 10uL of oligonucleotide-coupled protein G beads were prepared. All biotinylated protein G beads that would be needed for the entire experiment were first pooled into a tube, washed in 1mL of PBSt and resuspended in 200uL of 1x oligo binding buffer (0.5X PBST, 5 mM Tris pH 8.0, 0.5 mM EDTA, 1M NaCl) per

10uL of beads. 200 µL of bead suspension was aliquoted into individual wells of a deep-well 96-well plate (Nunc 96-well deep-well Plates with Shared-Wall Technology, Thermo Scientific, Cat. No. 260251) and 14 µL of 5.675nM (1:40 from 227nM working stock made fresh) of streptavidin-coupled oligo was added to each well. The 96-well plate was then sealed with a Nunc 96-well cap mat (Thermo Scientific, Cat. No. 276000) and shaken at 1200 rpm on a ThermoMixer for 30 min at room temperature. Beads in each well were washed twice with M2 buffer (20 mM Tris 7.5, 50 mM NaCl, 0.2% Triton X-100, 0.2% Na-Deoxycholate, 0.2% NP-40), twice with 1X PBSt, and finally resuspended in 200 µL of 1X PBSt.

Estimating number of oligos per bead

After oligo-coupling, a QC step was performed to estimate the number of oligos bound to each bead. 20% of a representative well for each row of the 96-well plate of oligo-coupled beads was isolated and the "Terminal" tag from split-and-pool barcoding was ligated onto the oligos in these aliquots. Then, half of the ligated product was PCR amplified for ten cycles and purified using 1x SPRI beads. The purified DNA product was run on an Agilent TapeStation using a D1000 tape to estimate molarity and this molarity was used to calculate the total number of molecules post PCR. Using this post-PCR number and the number of cycles of PCR, the number of unique molecules pre-PCR was estimated[39]. Finally, the number of unique molecules was divided by the number of beads put into the PCR reaction (2.7*10^6 beads per 1uL of stock biotinylated protein G beads) to calculate the estimated oligos per bead.

Antibody Coupling

2.5 µg of each antibody was added to each well of the 96-well plate containing oligonucleotide-labeled beads resuspended in 1X PBSt. The plate was incubated on a ThermoMixer overnight at 4C with 30 seconds of shaking at 1200 RPM every 15 min. The following morning, beads were washed twice with 1X PBSt (Sigma, #B4639-5G), resuspended in 200 µL of 1x PBSt + 4mM biotin + 2.5ug Human IgG Fc and left shaking at 1200 rpm for 15 min at room temperature to quench free protein G or streptavidin binding sites.

Preparation of bead pool

All wells containing oligo-labeled, antibody-coupled beads were washed 2X with 200uL 1X PBSt + 2 mM biotin, taking care to remove all supernatant after the final wash. Afterwards, one of two protocols were followed for bead pooling: 1) Equal bead pooling—Beads were pooled using equal amounts of prepared beads for each antibody (10uL of protein G beads per antibody). 2) Titrated bead pooling—Beads were pooled using unequal amounts of prepared beads for each antibody. The relative number of beads for each antibody was determined based on the chromatin pull-down efficiency (chromatin reads per bead) measured in QC experiments. Fewer beads were used for antibodies with higher

pull-down efficiencies and greater beads were used for antibodies with lower pull-down efficiencies or negative controls. This strategy was intended to generate a more uniform distribution for the number of chromatin reads assigned to each antibody in the final experiment.

ChIP-DIP: Immunoprecipitation, Split-and-pool and Library Preparation

Pooled immunoprecipitation

Fragmented lysate was diluted with PBSt +10mM biotin + 1x PIC + 2.5ug of human IgG Fc per 10ul of beads. The pool of labeled beads was added to the lysate and rotated on a HulaMixer for 1 hour at room temperature. Beads were washed 2X with 1mL IP Wash Buffer I (20mM TrispH8.0, 0.05% SDS, 1% Triton X 100, 2mM EDTA, 150mM NaCl in water), 2X with 1mL of IP Wash Buffer II (20mM TrispH8.0, 0.05% SDS, 1% Triton X 100, 2mM EDTA, 500mM NaCl in water) and 2X with 1mL of M2 buffer (20mM Tris pH7.5, 0.2% Triton X100, 0.2% NP-40, 0.2% DOC, and 50mM NaCl).

Chromatin End Repair and dA-Tailing

To blunt end and phosphorylate double stranded DNA, the NEB End Repair Module (E6050L; containing T4 DNA polymerase and T4 PNK) was used. Beads were incubated in 1X NEBNext End Repair Enzyme cocktail + 1X NEBNext End Repair Reaction Buffer + 4mM biotin + 1ug human IgG Fc per 10uL beads at 20°C for 15 min. The reaction was quenched with 3X volume of PBSt + 100uM EDTA and beads were washed 2X with 1mL PBSt. Next, to dA-tail DNA, the NEBNext dA-tailing Module (Klenow fragment (50 -30 exo-, NEBNext dA-tailing Module, E6053L) was used. Beads were incubated in 1X NEBNext dA-tailing Reaction Buffer + 1X Klenow Fragment (exo-) + 4mM biotin + 1ug human IgG Fc per 10uL beads at 37°C for 15 min. The reaction was quenched with 3X volume of PBSt + 100uM EDTA and beads were washed 2X with 1mL PBSt.

Split-and-pool barcoding

Split-and-pool barcoding was performed as previously described[39] with modifications. Specifically, beads were first split-and-pool ligated by DPM to attach a common sticky end to all DNA molecules. Then, beads were split-and-pool ligated for ≥ 6 rounds with sets of "Odd", "Even", and "Terminal" tags. The number of barcoding rounds and number of tags used for each round was determined based on the number of beads that needed to be resolved. These parameters were selected to ensure that virtually all barcode clusters (>95%) represented molecules belonging to unique, individual beads. In most cases, 6 rounds of barcoding with 24–36 tags per round were performed. Each individual tag sequence was used in only a single round of barcoding. All split-and-pool ligation steps were performed for 5 min at room temperature and supplemented with 2mM biotin and 5.4uM ProteinG. After split-and-pool barcoding was complete, beads were resuspended in

1mL of MyRNK buffer [20 mM Tris pH 7.5, 100 mM NaCl, 10 mM EDTA, 10 mM EGTA, 0.5% Triton-X, 0.2% SDS]. Aliquots of various sizes (0.05% to 4% of total beads) were prepared, ensuring that the number of beads within each aliquot was resolvable by the number of possible unique split-and-pool barcodes. Each aliquot was then digested with 8ul of Proteinase K (NEB) for 2 hrs at 55°C, 1200RPM shaking and reverse crosslinking at 65°C, 1600 RPM shaking overnight.

Library Preparation

DNA from each reverse crosslinked aliquot was isolated with a Zymo IC column using a 6X volume of the DNA-binding buffer (Zymo Cat. No. D4014) and eluted in 21ul of H20. Libraries were amplified for 9–12 cycles using Q5 Hot-Start Mastermix (NEB Cat No M0294L) and primers that added the full Illumina adaptor sequences. The following PCR mixture was used: 21uL DNA in H2O, 2uL of i5 primer (12.5uM), 2uL of i7 primer (12.5uM), 25uL 2X Q5 MM. After amplification, libraries were cleaned with 1.2x SPRI (Bulldog Bio CNGS500) and eluted in 20uL. Prior to sequencing, libraries were gel purified to remove unused primers using a 2% agarose gel [Invitrogen Cat No. G401002].

Sequencing

Sequencing was performed on Illumina NovaSeq S4 (300 cycle) and NextSeq (200 cycle or 300 cycle) paired-end runs, Read lengths were asymmetrical in order to capture the full split-and-pool barcode sequence on read 2 (R2) and the chromatin sequence on read 1 (R1). For 300 cycle kit–100 cycles for R1, 200 cycles for R2; For 200 cycle kit–50 cycles for R1 and 150 cycles for R2.

For each experiment, multiple different libraries were generated and sequenced. Each library corresponds to a distinct aliquot which is amplified with a unique pair of primers, providing an additional round of barcoding.

Data Processing Pipeline

Read Processing. Paired-end sequencing reads were trimmed with Trim Galore! V0.6.2 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) to remove adaptor sequences and quality assessed with FastQC v0.11.8. Split-and-pool barcodes were identified from Read 2 using Barcode ID v1.2.0 (https://github.com/GuttmanLab/chipdip-pipeline). Reads missing split-and-pool tags or with tags in the incorrect position given the split-and-pool round they correspond to were discarded. Subsequently, reads were split into two files, one for antibody-ID reads and one for DNA reads, based on the presence of "BPM" (bead tag) or "DPM" (DNA tag), respectively, on Read 1.

For DNA reads, the DPM sequence was trimmed from Read 1 using Cutadapt v3.4[91]. The remaining sequence was aligned to mm10 or hg38 using Bowtie2 (v2.3.5)[92] with default parameters. Only primary alignments with a mapq score of 20 or greater were kept for further analysis. Finally, reads were masked using the repeat genome obtained from ENCODE[93].

For antibody-ID reads, the BPM sequence, which contains the antibody-ID information, was trimmed from Read 1 using Cutadapt v3.4 and the UMI extracted from the remaining sequence.

MultiQC v1.6[94] was used to aggregate metrics from all steps.

Cluster Generation. A "cluster file" was generated by aggregating all reads (i.e., aligned, masked DNA reads and antibody-ID reads) that share the same split-and-pool barcode sequence. During this step, reads in each cluster were deduplicated by alignment position for DNA reads or UMI for antibody-ID reads.

Antibody-ID Oligo Movement Quality Control Check. To assess the frequency of antibody ID oligo movement between beads, the proportion of antibody-ID reads corresponding to the maximum representation in each cluster was calculated. Only clusters with >1 antibody-ID read were considered. For each experiment, these values were plotted as an empirical cumulative distribution function (ECDF) using the Python plotting package seaborn[95].

Cluster Filtering and Assignment. Individual clusters in the "cluster file" were assigned to a specific antibody based on antibody-ID reads within the cluster. First, clusters in the "cluster file" without antibody-ID reads or clusters with >10,000 genomic DNA reads (which likely represent undersonicated material or clumps of beads) were filtered out. Next, each remaining cluster was assigned to the antibody ID that had maximum representation within the cluster if 1) there were greater than two unique reads corresponding to the antibody ID and 2) the antibody ID represented >80% of all antibody-ID reads within the cluster. These criteria were selected empirically to ensure high confidence assignments of antibody IDs to each cluster. Clusters that did not meet these criteria were removed from further analysis.

Antibody-specific protein maps. Genomic DNA alignments were split into separate bam files such that each file corresponded to all alignments associated with an individual antibody based on the antibody-ID assignments within each cluster. DNA reads from clusters that did not have antibody-ID reads, were too large or could not be uniquely assigned to a single antibody ID were filtered out. DNA reads were deduplicated such that only one read per alignment position per cluster was retained.

Visualization and Peak Calling

Bigwig Generation. Bigwigs were generated from each antibody-specific BAM file using the 'bamCoverage' function from deepTools v3.1.3[96] and were visualized with IGV[97]. Track visualizations are scaled to the maximum over the region and scales indicate reads per bin, unless indicated otherwise.

Background Normalization. Because beads from many antibodies are processed together, ChIP-DIP has sources of potential background that are distinct from a traditional ChIP-Seq experiment. In any ChIP experiment, the antibody used will immunoprecipitate its specific protein (and the associated chromatin) but will also non-specifically purify other proteins (and their associated chromatin) at some lower frequency. This non-specific chromatin (background) is generally proportional to the overall distribution of genomic DNA present in the starting material ("input"). In ChIP-DIP, the same is true during the IP; any given antibody will preferentially capture its specific protein but will, at some lower frequency, non-specifically capture other proteins. However, because ChIP-DIP entails purification with many antibodies, the source of proteins and chromatin for this non-specific binding is no longer the entire cellular input but rather the material present within the pooled IP (e. g. the proteins and chromatin that were pulled down by the pool of antibodies). Indeed, we observed that some antibodies displayed background signal that was distinct from the input library. For example, antibodies targeting CTCF displayed higher background over promoter regions, likely reflecting the presence of various promoter-enriched histone modifications present in the same experimental pool. To account for this in our analysis, we used the pool of all genomic DNA reads captured in a ChIP-DIP experiment as the background control. We found that normalization using this empirically-defined background led to a more conservative enrichment calculation for ChIP-DIP data. For example, in the CTCF example noted above, this normalization approach successfully removed the background promoter-associated signal while retaining signal at known CTCF binding sites.

Specifically, a background model was generated for each individual antibody using the total pool of assigned sequencing reads. The background for an antibody contained all reads except those assigned to it, or other antibodies targeting the same or related proteins. For example, for an antibody targeting RNAPII-NTD, reads from all antibodies targeting RNAP II were excluded from this background set. To calculate a scaling factor for this background: 1) the total experiment coverage was calculated in 10kB bins, 2) the high coverage bins (80%) were selected, 3) a per-bin enrichment quotient of the target compared to the background coverage was calculated, 4) a kernel density plot of the enrichment quotient was generated, 5) a threshold was calculated based on the position of the smallest peak and 6) the ratio of total coverage in target versus background bins below the threshold was determined. The goal of this procedure was to locate regions that represented background noise in the target and calculate the target-to-background ratio using only those regions. The kernel density plot was frequently bimodal or with a long tail, with the higher peak or tail representing signal bins and the lower peak representing background bins. Background normalized peaks were called using the scaled background as a substitute for

input. Background normalized bigwigs were generated using the 'bamCompare' function from deepTools v3.1.3 by subtracting the scaled background and, subsequently, removing negative value bins.

Peak Calling. Peaks were called using the HOMER v4.11[98] program 'findPeaks' on tag

directories generated for target datasets using '-style histone' for histone modification targets and '-style factor' for other targets. Background normalized peaks were generated using the scaled background distribution (described above). Specific parameter settings, such as '-minDist' (distance between adjacent peaks), '-size' (width of peaks) or filtering thresholds were tuned according to the nature of the target. For instance, peaks for focal histone modification H3K4me3 were generated using '-F 2 -P 0.001 -L 0' while enriched regions for broad histone modification H3K36me3 were calculated using '-F 2 -P 0.001 -L 0 -size 1000 -minDist 7500 -region'.

Motif Prediction. Transcription factor motifs were predicted using the HOMER program

'findMotifsGenome' on peaks generated using HOMER, as described above. For all transcription factors, motifs were generated using the settings '-s 200 -mask -l 10' and results are reported in **Supplemental Table 5**. For individual examples in **Figure 4**, longer motifs were also predicted.

Ribosomal DNA Alignments

To analyze reads aligning to genomic DNA encoding ribosomal RNA (rDNA), we aligned reads directly to an rDNA reference. We generated a modified reference of the mouse rDNA sequence (NCBI Genbank BK000964.3)[99]. Because the original mouse BK000964.3 sequence begins with the TSS and ends with the Pol I promoter, we transposed a portion at the end of the rDNA reference to the beginning, as previously described[100], to enable a continuous visualization of the promoter-TSS region. Specifically, the rDNA sequence was cut at the 36,000 nt position and the sequence downstream of the cut site were moved upstream of the TSS, such that the resulting rDNA sequence begins with ~10kb of IGS, then the promoters and then transcribed regions. Processing steps prior to sequence alignment followed the standard ChIP-DIP pipeline. After barcode identification, DNA sequence was aligned to the custom rDNA genome using Bowtie2 (v2.3.5) with default parameters. Only primary alignments with a mapq score of 20 or greater were kept for final analysis. The subsequent cluster generation and read assignment steps followed the standard ChIP-DIP pipeline.

ChIP-DIP Experiments

We performed nine ChIP-DIP experiments in this paper, each of which, along with the associated antibodies, proteins, and statistics are described in **Supplemental Table 1**. Briefly, these experiments were:

1. *Chromatin Movement Experiment*: A quality control human and mouse mixing

experiment used to quantify possible chromatin movement during the procedure.

2. *Antibody Movement Experiment:* A quality control human and mouse mixing

experiment used to quantify possible antibody movement during the procedure.

3. *K562 10 Antibody Pool Experiment:* An initial data-generation experiment

performed in human K562 to measure a small number of well-defined targets.

4. *K562 50 Antibody Pool Experiment:* A data-generation experiment performed in

human K562 measuring 50 antibodies.

5. *K562 52 Antibody Pool Experiment:* A data-generation experiment performed in

human K562 measuring 52 antibodies.

6. *K562 35 Antibody Pool Experiment:* A data-generation experiment in human K562

measuring 35 antibodies as a function of different cell input amounts.

7. *mESC 228 Antibody Pool Experiment:* An antibody-screening experiment

performed in mouse ES cells using 228 antibodies to identify good antibodies and the antibody amounts required for deeper characterization in subsequent ChIP-DIP experiments.

8. *mESC 67 Antibody Pool Experiment:* A data-generation experiment performed in

mouse ES cells measuring 67 antibodies.

9. *mESC 165 Antibody Pool Experiment:* A data-generation experiment performed in

mouse ES cells measuring 165 antibodies.

All ChIP-DIP experiments were performed using the same general protocol with the following experiment-specific modifications:

1. Chromatin Movement Experiment: To test whether chromatin dissociates during the

ChIP-DIP procedure and binds to other beads, we designed a human-mouse mixing experiment. Cell lysate from 20M mESC cells, cell lysate from 10M K562 cells and two sets of antibody-coupled, oligonucleotide-labeled beads were prepared according to standard protocol. Prior to IP, lysate yields were quantified using TapeStation and equal amounts of mouse and human chromatin preparations were used for the subsequent, separate IPs. One set of antibody-ID-labeled beads was used for the human IP and the other set of antibody-ID labeled beads was used for the mouse IP. After IP, the two species-specific IPs were mixed and split into three conditions using different quenchers: (i) 10% BSA quencher, (ii) 1X Blocking Buffer quencher and (iii) No quencher. For the 10% BSA quencher condition, end-repair, dA-tailing and DPM reactions were performed in buffer supplemented with 10% BSA. For the 1X blocking buffer quencher condition, end-repair, dA-tailing and DPM reactions were performed in buffer supplemented with 1X protein blocking buffer (Abcam ab126587). The three conditions were combined for split-and-pool barcoding.

For alignment of human-mouse mixing experiments, DNA reads were aligned to a custom combination genome including both mm10 and hg38 genomes using Bowtie2 (v2.3.5) with default parameters. Only primary alignments with a mapq score of 20 or greater were kept for further analysis. Reads were then masked using a merged version of mm10 and hg38 blacklist regions defined by ENCODE. Reads were then uniquely assigned to human beads (beads present only in the human IP condition) or mouse beads (beads present only in the mouse IP condition) using the standard assignment pipeline. Total reads aligned to mm10 or hg38 for each bead set were quantified and the relative proportions were plotted as a bar graph.

2. Antibody Movement Experiment: To test whether antibodies dissociate from their

labeled beads during the ChIP-DIP procedure and bind to other beads, we designed an experiment that involved the addition of labeled antibody-free beads at various steps. Following a similar set-up as the chromatin mixing experiment, cell lysate from 20M mESC cells, cell lysate from 10M K562 cells and two sets of antibody-coupled, oligonucleotide-labeled beads were prepared using the standard protocol. One set of beads was used for the human IP and the other set of beads was used for the mouse IP. After IP, half of each species-specific IP was mixed together, and half was left separate. For this mixed condition only, oligonucleotide-labeled beads without a coupled antibody were added prior to the end-repair and the dA-tailing reactions. These empty beads were added to capture antibodies that dissociated from other IP'd beads. Finally, the three conditions (mouse only, human only, mixed) were ligated with unique sets of DPM adaptors and combined for split-and-pool barcoding. To calculate the frequency of antibody movement,

total reads and total beads assigned to human CTCF beads, human IgG beads, empty beads added prior to end repair and empty beads added prior to dA-tailing were quantified. Reads per bead for each group were normalized to the mean value for human CTCF beads. These normalized values were plotted as a bar graph with 99% CI using the Python plotting package seaborn.

3. K562 10 Antibody Pool Experiment: We performed an initial small scale proof-of-

concept (POC) experiment in K562 using 10 different antibodies. The POC experiment was performed using lysate from 50M K562 cells per IP. Standard protocol with equal bead pooling was used with the exception of IP conditions. Two identical sets of antibody-coupled beads were prepared using different biotinylated oligonucleotides; one set was used for an overnight immunoprecipitation at 4C and one set was used for 1-hr immunoprecipitation at room temperature. DNA processing steps and DPM ligation reactions were performed separately for the two IP conditions and then the two samples were pooled for the remaining rounds of split-and-pool barcoding. See **Supplemental Table 1** for full list of antibodies under the "K562 10 Antibody Pool" tab. For data processing, the standard pipeline generated individual clusters corresponding to antibody–IP condition pairs and individual bam files for each target in each IP condition. Data from both IP conditions were merged for each target, resulting in a single file per antibody.

4. K562 50 Antibody Pool Experiment: The K562 50 Antibody Pool Experiment was

performed using lysate from 50M K562 cells. The standard protocol with equal bead pooling was used. See **Supplemental Table 1** for full list of antibodies under the "K562 50 Antibody Pool" tab.

5. K562 52 Antibody Pool Experiment: The K562 52 Antibody Pool Experiment was

performed using lysate from K562 cells. To test the efficiency of different crosslinking strategies, two parallel IPs were performed using the same pool of prepared beads. One IP utilized 60M K562 cells crosslinked with 1% FA and the other IP utilized 60M K562 cells crosslinked with 1% FA + DSG. Cells for the 1% FA condition were prepared as described above. Cells for the 1% FA + DSG condition were prepared as follows: After harvest and pelleting, K562 cells were crosslinked in 4 mL of 2 mM disuccinimidyl glutarate (DSG, Pierce) dissolved in 1X PBS per 10 million cells for 45 min at room temperature. Cells were then pelleted, washed with 1X PBS and crosslinked with 1% FA, as described above.

For antibody-ID oligonucleotide-labeling of beads, beads were labeled in two sequential rounds. First, beads were labeled according to the standard protocol. Then, beads were labeled again using another 2.5uL of 5.67nM streptavidin-coupled oligo in 200uL of 1x oligo binding buffer for 30 min at room temperature. During the first round of labeling, all wells received a unique streptavidin-coupled oligonucleotide. During the second round of

labeling, most wells received the same streptavidin-coupled oligonucleotide as the first round, with the exception of eleven wells. Eleven pairs of wells received the same streptavidin-coupled oligonucleotide in the second round–one well of each pair was labeled with the same oligonucleotide in both rounds while the other well was labeled with different oligonucleotides. The result was that most beads were labeled with a single, unique oligonucleotide label, eleven beads were labeled with a pair of oligonucleotide labels, and eleven beads were labeled with a single oligonucleotide label that can also be found on other beads. This labeling strategy was designed to test combinatorial labeling of beads. After antibody coupling, beads were pooled in equal amounts and half of the bead pool was used for IP of each crosslinking condition. Following IP, each condition was processed separately and DPM-ligated with unique, condition-identifying sets of adaptors. Conditions were kept separate for the first round of split-and-pool barcoding and then combined for the remaining rounds of split-and-pool. See **Supplemental Table 1** for full list of antibodies under the "K562 52 Antibody Pool" tab.

Sequenced data was processed using the standard ChIP-DIP pipeline up until the clustering assignment step. To account for the dual oligo labeling of selected antibodies, prior to assignment of unique antibodies to each cluster, clusters with multiple labels (clusters containing both oligo types from a known co-occurring pair) were isolated and antibody-ID oligos in these clusters corresponded to the second round of labeling were reassigned to the matched antibody-ID oligo from the first round of labeling. The result is that all antibodies now corresponded to a unique antibody-ID oligo; for the eleven combinatorial pairs, this is the first round of labeling. Afterwards, the remaining steps in the standard ChIP-DIP pipeline (cluster assignment, etc.) were performed as described above.

6. K562 35 Antibody Pool Experiment for input cell number titration: One of the major

challenges with mapping DNA-binding proteins in primary cell types, disease models, and other rare cell populations is the large number of cells required for traditional ChIP-Seq experiments. Because ChIP-DIP enables simultaneous mapping of many proteins within the same experiment, we reasoned that it may dramatically reduce the total number of cells required in two ways: (i) the number of cells required to map any individual protein is instead distributed across all protein targets in a pool, and (ii) the total chromatin purified from multiple proteins may enable purification of lower DNA concentrations associated with a single/low abundance proteins that might otherwise be lost due to experimental handling.

The K562 35 Antibody Pool Experiment was designed to measure the amount of cell input material required for ChIP-DIP. To do this, we performed a series of ChIP-DIP experiments using the same antibody pool and differing amounts of cell lysate. Specifically, we crosslinked >100M cells in a single batch and then performed four ChIP-DIP experiments from this same crosslinked lysate. This experiment involved four separate ChIP-DIP experiments, performed in pairs of two. For the first pair, the 45M and 5M conditions, a 50M cell aliquot was lysed and sonicated and then split into 45M and 5M cell equivalents of lysate. For the second pair, the 500K and 50K conditions, a 1M cell aliquot was lysed

and sonicated and then split into 500k and 50k cell equivalents of lysate. Each pair of experiments used a single preparation of antibody-coupled, antibody-ID oligonucleotide labeled beads that was split in half. See **Supplemental Table 1** for full list of antibodies under the "K562 35 Antibody Pool" tab.

Read coverage profiles of four targets—H3K4me3, H3K27me3, CTCF and RNAP II—were compared. For both RNAP II and CTCF, two different antibodies were included (RNAP II: CST 91151 and 14958S; CTCF: CST 3418S and ABCAM ab128873). Coverage of normalized bigwig files across the set of all peak regions from the 10 Antibody Pool experiment, the same set of regions used for the pool size comparison correlations, was calculated using the 'multiBigwigSummary' function of the Python package deepTools v.3.1.3. Pearson correlation coefficients for all pairs were calculated using the 'plotCorrelation' function of deepTools v.3.1.3 and the plotted as a heatmap, manually ordering the rows/columns from lowest to highest amount of input lysate for each target.

Peak overlaps were calculated for each antibody between pairs of experiments as the (number of peaks in experiment 1 intersecting peaks in experiment 2) / (total number of peaks in experiment 1). The number of intersecting peaks were calculated using the BEDtools v2.29.2[101] and results are reported in **Supplemental Table 4**.

7. mESC 228 Antibody Pool Screen: The mESC 228 Antibody Pool experiment was

performed using 40M mESC cells. The standard protocol was used with the following modifications. Because the number of antibodies exceeded the number of unique antibody-ID oligonucleotides, three plates of antibody-coupled, oligonucleotide-labeled beads were prepared separately, the beads from each plate were pooled into three separate antibody–bead pools, and each pool was used to IP one third (~13M cells) of the prepared cell lysate. Post IP, each sample was processed separately until split pool. During split-pool barcoding, a unique set of "ODD" barcodes was ligated to each sample during the first round and then all samples were pooled for rounds 2–6.

Sequencing data was processed through the standard pipeline, using a concatenated string of three antibody names (i.e., DNMT3B-CST_POLR3E-Bethyl_H3K36Ac-CST), one name for the antibody corresponding to each sample, to match individual antibody-ID sequences during barcode identification. After cluster generation and prior to cluster assignment, each antibody-ID read was assigned to only one antibody based on its first-round 'ODD' split- and-pool tag. Cluster assignment then proceeded using the standard pipeline. Chromatin reads were assigned to each antibody based on cluster assignments and total chromatin for each antibody was quantified (see **Supplemental Table 1**, "mESC 228 antibody pool" tab for counts). These relative chromatin yields per antibody were used to inform the ideal amount of antibody needed and we used these to titrate the amount of beads for each antibody pooled together in ChIP-DIP experiments 8 and 9 below. This experiment was sequenced at low depth to allow for rapid and low-cost antibody screening.

8. mESC 67 Antibody Pool Experiment: The mESC 67 Antibody Pool experiment

was

performed using lysate from 80M mESC cells. The standard protocol with titrated bead pooling was used. This experiment contained 67 different antibodies. See **Supplemental Table 1** for full list of antibodies under the "mESC 67 Antibody Pool" tab.

9. mESC 165 Antibody Pool Experiment: The mESC 165 Antibody pool experiment was

performed using lysate from 60M mESC cells. Similar to the mESC 228 Antibody Pool experiment, because the number of antibodies exceeded the number of unique antibody-ID oligonucleotides, a multi-plate strategy was used. Specifically, two plates of antibody-coupled, oligonucleotide-labeled beads were prepared separately, pooled using the titrated bead pooling strategy and used to IP half of the prepared cell lysate. After IP, the two samples were processed separately up until the third round of split-and-pool barcoding and then combined for the remaining rounds of split-and-pool. This experiment contained 165 different antibodies. See **Supplemental Table 1** for full list of antibodies under the "mESC 165 Antibody Pool" tab. Sequencing data was processed through the standard pipeline, using a concatenated string of antibody names (i.e., LSD1-CST_SAP30-Bethyl) to match individual antibody-ID sequences during barcode identification. After cluster generation and prior to cluster assignment, each read antibody-ID read was assigned to only one antibody based on its first-round split-and-pool tag. Cluster assignment and BAM file generation then proceeded using the standard pipeline.

Protein Target Classification

Antibody targets were assigned to one of five categories: histone modification (HM), transcription factor (TF), chromatin regulator (CR), RNAP and other DNA-associated protein. Transcription factors were defined as proteins with a DNA- binding domain and were manually subclassified into constitutive, stimulus response or cell-type-specific/developmental manually curated these transcription factors based on functional descriptions from GeneCards. Chromatin regulators contained proteins or members of complexes that read, write, or erase histone modifications or DNA methylation. Proteins that were part of chromatin regulator complexes and contained a DNA-binding domain were considered part

of the chromatin regulatory category. Proteins involved in chromatin remodeling (e. g., BRG1) or other structural proteins that interact with chromatin (e.g., LaminA) were also considered chromatin regulators. Dual function proteins (e.g., transcription factors with intrinsic acetyltransferase capabilities) were assigned to a single category (e. g.

transcription factor) but were included in chromatin regulator schematics. Other DNA-associated proteins included a mixture of targets, such as RNAP elongation factors (e.g. ELL), RNA binding proteins (e.g., NONO) and antibodies that detected DNA methylation.

Comparison to ENCODE data

ChIP-DIP comparisons to ENCODE-generated ChIP-Seq data in **Figure 1** were performed using the 10 pool experiment in K562. Visual comparisons were performed using IGV and the raw ENCODE datasets: ENCFF656DMV (H3K4me3), ENCFF785OCU (POLR2A), ENCFF800GVR (CTCF) and ENCFF508LLH (H3K27me3). Genome-wide coverage comparisons were calculated across all RefSeq TSS for H3K4me3 and POLR2A or across 10kB bins for CTCF and H3K27me3. Calculations were performed using the 'multiBamSummary' function of the Python package deepTools v3.1.3 and plotted as 2-D kernel density plots using the Python library seaborn.

Systematic comparisons between ChIP-DIP K562 datasets (10 Antibody Pool, 50 Antibody Pool, 52 Antibody Pool and 35 Antibody Pool) and ENCODE-generated ChIP-Seq were performed for all targets for which ENCODE datasets were available. Genome-wide coverage comparisons were calculated at 1000bp using the 'multiBigwigSummary' function of the Python package deepTools v3.1.3 and Pearson correlation coefficients are reported in **Supplemental Table 2**. The number of overlapping peaks between ChIP-DIP and ENCODE datasets were calculated using as described above. Both the fraction of ENCODE peaks detected by ChIP-DIP and the fraction of ChIP-DIP peaks detected by ENCODE are reported in **Supplemental Table 2.**

Accession numbers used for coverage comparisons include: ENCFF121RHF, ENCFF508LLH, ENCFF035SOZ, ENCFF272JVI, ENCFF816ECC, ENCFF880HKV, ENCFF352HXD, ENCFF446FUS, ENCFF656DMV, ENCFF465UWC, ENCFF149MXA, ENCFF155UQU, ENCFF187HIQ, ENCFF702HIC, ENCFF800GVR, ENCFF982AFE, ENCFF178ARN, ENCFF096FWU, ENCFF844WTT, ENCFF801TEZ, ENCFF174BEG, ENCFF014HSG, ENCFF656FDC, ENCFF457PGP, ENCFF108EMO, ENCFF617EYG, ENCFF986KSB, ENCFF572IXE, ENCFF893LSE, ENCFF785OCU, ENCFF816VGU, ENCFF179XDZ

Accession numbers used for peak comparisons include: ENCFF863MYY, ENCRR698RKX, ENCRR736TRL, ENCFF222OPH, ENCFF731WDM, ENCFF479VOH, ENCFF514SHW, ENCFF643TQX, ENCFF586UHP, ENCFF456SZC, ENCFF108JNI, ENCFF632MQY, ENCFF924JXT, ENCFF834RTA, ENCFF660HBV, ENCFF567BOX, ENCFF563CWK, ENCFF394JNY, ENCFF307DMW, ENCFF181AVH, ENCFF169ZQQ, ENCFF040TWS, ENCFF990EZP, ENCFF755KWM, ENCFF269JZL, ENCFF106GOL, ENCFF827OVS, ENCFF789NDS, ENCFF567HEH

Pool Size Comparison Analysis

To measure the influence of the number of antibodies contained within an individual pool, read coverage profiles of four targets–H3K4me3, H3K27me3, CTCF and RNAP II–generated in four different ChIP-DIP experiments in K562 cells were compared. ChIP-DIP experiments included the 10 Antibody Pool, the 45M condition from the 35 Antibody Pool, the 50 Antibody Pool and the 52 Antibody Pool in K562. For both RNAP II and CTCF, two different antibodies were included (RNAP II: CST 91151 and 14958S; CTCF: CST 3418S and ABCAM ab128873). Coverage of normalized bigwig files across the set of all peak regions from the 10 Antibody Pool experiment was calculated using the 'multiBigwigSummary' function of the Python package deepTools v.3.1.3. Pearson correlation coefficients for all pairs were calculated using the 'plotCorrelation' function of deepTools v.3.1.3 and plotted as a heatmap, manually ordering the rows/columns from smallest to largest pool size for each target.

Peak overlaps were calculated for each target between experiments of different pool sizes as described above and are reported in **Supplemental Table 3**.

Histone Modification Diversity Analysis

Chromatin-State: Genome-wide coverage for 10kb windows for 12 histone marks (H3K27me3, H2AK119ub, H3K9me3, H4K20me3 and H3K9me3 from the 5M condition in the 35 Antibody Pool Experiment in K562; H3K79me2, H3K79me1, H3K4me3, H3K4me2, H3K4me1, H3K9Ac and H3K27Ac from the histone panel in K562) was calculated using the 'multiBamCoverage' function from deepTools v3.1.3. These values were standardized for each mark by transforming into z-score values. The UMAP reduction was generated using the UMAP[102] Python package and parameters n_components=2 and n_neighbors=3.

Polycomb-Associated Histone Modifications: Validation of polycomb-associated histone modifications used the 5M condition in the K562 35 Antibody Pool Experiment. H3K27me3 and H2AK119ub bam alignment files were converted into binary signal files using the 'BinarizeBam' script from the ChromHMM[103] package with standard settings. The number of bins with only H2AK119ub signal or with both H2AK119ub and H3K27me3 signal were computed and plotted as a pie chart.

Heterochromatin-Associated Histone Modifications: Validation of heterochromatin-associated histone modifications used the 5M condition in the K562 35 Antibody Pool Experiment. Read coverage of H3K9me3, H4K20me3 and H3 were computed over annotation groups (ZNFs, LTRs, LINES, SINES, TSS+/-2kb) using the 'depth' function from SAMtools v1.9[104]. An enrichment score was calculated by normalizing for feature and target abundance. Specifically, let a = total base pairs within an annotation group, b =

effective genome size, c = read coverage of a target over the annotation group and d = total reads of the target. The enrichment score would be (c/d) / (a/b).

Promoter-Associated Histone Modifications: Validation of promoter-associated histone modifications used the ChIP-DIP histone dataset in mESC. Promoter coverage correlations were calculated across promoters from EPDNew[105], a database of non-redundant eukaryotic RNAP II promoters, +/- 500bp using the 'multiBamSummary' and 'plotCorrelations' functions of the Python package deepTools v.3.1.3.

Gene Body-Associated Histone Modifications: Validation of gene body-associated histone modifications used the 5M condition in the K562 35 Antibody Pool Experiment and the K562 50 Antibody Pool Experiment. Coverage meta plots over the gene bodies of all protein coding genes from GENCODE v38 basic annotation were calculated using 'computeMatrix' function of the Python package deepTools v.3.1.3 and normalized to the maximum and minimum for each target.

Enhancer-Associated Histone Modifications: Validation of enhancer-associated histone modifications used the 5M condition in the K562 35 Antibody Pool Experiment and the K562 50 Antibody Pool Experiment. H3K4me1 peaks were assigned to three categories (promoter, gene or intergenic) based on overlap with H3K4me3 (promoter), H3K79me1 (gene) or H3K36me3 (gene). These categories were further sub-divided based on the co-occurrence of H3K27Ac peaks. The proportion of peaks in each category was computed and plotted as a pie chart.

Chromatin Regulator Diversity Analysis

Polycomb-Associated Chromatin Regulators: Validation of polycomb-associated chromatin regulators used the K562 50 Antibody Pool Experiment. Meta plots respective to RING1B peak sites were calculated using 'computeMatrix' function of the Python package deepTools v.3.1.3 with the following settings: 'reference-point -bs 10000 -a 500000 -b 500000'. The resulting read coverage profiles were normalized to the maximum and minimum for each target and plotted as a heatmap.

Heterochromatin-Associated Chromatin Regulators: Validation of heterochromatin-associated chromatin regulators used the K562 50 Antibody Pool Experiment. Genome-wide coverage for 10kB windows and Pearson correlation coefficients were calculated using the 'multiBigwigSummary' function and 'plotCorrelation' function, respectively, of the Python package deepTools v3.1.3.

H3K4me3-Associated Chromatin Regulators: Analysis of H3K4me3-associated chromatin

regulator used the mESC 165 Antibody Pool Experiments. Binding profiles of JARID1A, RBBP5 and PHF8 were measured +/- 1kB around the TSS of all representative promoters from EPDNew and were clustered using k-means clustering with k=4 by the 'plotCoverage' function of the Python package deepTools v.3.1.3. H3K4me3 binding profiles from the mESC 67 Antibody Pool Experiment were measured over the same four promoter groups.

Polymerase Diversity Analysis

RNAP I, II and III Comparison: Validation of the various RNA polymerases used the mESC

165 Antibody Pool Experiment. First, read coverage within a +/- 100bp window surrounding the promoters/TSS of various gene groups were calculated. For tRNAs, the TSS of repeatmasker[106] tRNAs were used. For snRNAs, the TSS of RepeatMasker snRNAs (excluding U6 which is transcribed by RNAP III) were used. For mRNAs, EPDNew TSS annotations were used. For rDNA, the spacer promoter was used. Next, for each polymerase, coverage was normalized to the total reads aligned with any gene group. Finally, an enrichment score of the relative coverage compared to IgG was calculated and plotted as a bar graph.

RNAP II Phosphorylation State Comparison: Validation of the various RNA polymerases

used the K562 52 Antibody Pool Experiment. Meta plots over the gene bodies of all protein

coding genes from GENCODE v38 basic annotation were calculated using 'computeMatrix' function of the Python package deepTools v.3.1.3.

Histone Combinatorial Analyses

Polymerase-Associated Histone Profiles

For RNAP I, track coverage profiles of various histone modifications 1.5kB upstream to 0.5kB downstream of the spacer promoter were visualized using IGV.

For RNAP II, meta plots of coverage profiles for various histone modifications were generated around active and inactive RNAP II promoters using the deepTools v.3.1.3 'computeMatrix' (reference-point -a 1000 -b 1000) and 'plotProfile' functions. Promoters were defined as the TSS of all representative promoters from EPDNew and were grouped into active or inactive based on the read coverage of RNAP II in the surrounding +/-1kB window.

For RNAP III, meta plots of coverage profiles for various histone modifications were generated around active and inactive tRNA genes using the deepTools v.3.1.3 'computeMatrix' (scale-regions -a 1000 -b 1000 -m 75 -bs 25) and 'plotProfile' functions.

tRNA genes were grouped into active or inactive based on the read coverage of RNAP III.

For comparison of relative histone levels, total coverage for each histone mark was calculated in the -1.5kB to +0.5kB window surround the spacer promoter for rDNA, -0.5kB to +0.5kB window around active RNAP II promoters and -0.5kB to +0.5kB window around active RNAP III tRNA gene promoters. To account for differences in window size, the coverage of H3K56Ac and H3K4me2 was normalized to the level of H3K4me3. The density profiles of these ratios were plotted using the seaborn 'jointplot' function with the following kde parameters: "common_norm=False, thresh=0.2, log_scale=True, levels=10, cut=True". For comparison to RNAP I, the total sum ratios (e. g. total H3K4me2 coverage across all active RNAP II promoter intervals divided by total H3K4me3 coverage across all active RNAP II promoter intervals) were also calculated and plotted for RNAP II and RNAP III.

H3K4me3 Enriched Regions Clustering

Combinatorial histone modification analysis for H3K4me3 regions used the 5M condition of the K562 35 Antibody Pool Experiment. Read coverage of ten histone targets (H3K79me3, H3K79me2, H3K36me3, H3K4me1, H3K4me2, H3K27Ac, H3K27me3, H2AK119ub, H3K9me3 and H4K20me3) was calculated over all H3K4me3 peak regions using the 'multicov' function of BEDtools. The resulting region vs. histone data matrix (A) was normalized using log normalization[107]: 1) The log of the data matrix was computed $L = log\ log\ (A)$. 2) The column mean ($\underline{L_{i.}}$), row mean ($\underline{L_{.j}}$), and overall mean ($\underline{L_{..}}$) of the log matrix were computed. 3) All individual cells of the final matrix were computed according to $K_{ij} = L_{ij} - (\underline{L_{i.}}) - (\underline{L_{.j}}) + (\underline{L_{..}})$. This method of normalization is intended to capture the "extra" coverage of histone modification *j* in region *i* that is not explained simply by the overall difference between region *i* and other regions or between histone modification *j* and other histone modifications. Instead, it is special to the combination of region *i* (a region with H3K4me3 enrichment) and coverage of histone modification *j*. The regions of the normalized data matrix were clustered using cluster.hierarchy.linkage function from scipy v.1.6.2[108] with a Euclidean distance metric and complete linkage method. The clustered matrix was visualized using the 'clustermap' function of Python package seaborn.

Gene annotation of H3K4me3 regions was performed using the 'annotatePeaks.pl' function from HOMER v4.11. ZNF genes, RP genes, and lincRNA genes were defined as regions whose annotation gene description contained the terms 'zinc finger protein', 'ribosomal protein' and 'long intergenic', respectively, and had the nearest TSS within 2000bp. snoRNA genes were defined as all regions whose annotation gene type was snoRNA. Satellite RNA genes were defined as regions whose detailed annotation contained the term 'Satellite'. tRNA genes were defined as all regions that intersected with tRNA gene bodies or upstream by 500bp of the tRNA TSS from repeat masker. Cell cycle genes were defined as regions whose gene annotation belonged to the Kegg Cell Cycle Pathway[109]. Bivalent genes were defined as regions whose gene annotation belong to those identified by Court and Arnaud in human H1 cells[110]. Enhancer RNA regions (both antisense and intergenic)

were defined as regions that intersected those identified by Lidschreiber et al.[111] and had the nearest TSS greater than 2000bp away. To visualize enrichments of gene annotations in sets and subsets of the hierarchically clustered heatmap, the kernel density estimate (KDE) was calculated for each annotation group based on their clustering-defined order.

RNAP II levels of individual H3K4me3 regions were measured as the summed coverage over each region from four antibodies targeting RNAP II (RNAP II, RNAP II NTD, RNAP II Ser5, RNAP II Ser2) from the K562 52 Antibody Pool Experiment.

ChromHMM Model of Acetylation

The ChromHMM genome segmentation model was built using 15 different histone acetylation modifications measured in the mESC 67 Antibody Pool Experiment. Bam files were binarized using the BinarizeBam function from ChromHMM with a Poisson threshold of 0.000001 and other default parameters. The signal threshold was increased from default to remove spurious noise. State models with 5–20 states were built using the LearnModel function with default parameters. States were manually reordered and grouped based on transition probabilities between states. 19 states were selected for the final model to retain State 17, a state with a distinctive enrichment and transition profile.

Non-Negative Matrix Factorization of Acetylated Regions

Non-negative matrix factorization analysis utilized the histone acetylation mark data from the mESC 67 Antibody Pool Experiment. NMF is a matrix factorization technique to reduce dimensionality and explain the observed data using a limited number of combinatorial components[107]. NMF decomposes the original data matrix (dimensions: N x M) into a basis matrix (dimensions: N x k) and a mixture coefficient matrix (dimensions: k x M). In this case, N represents genomic regions of interest, M represents individual histone acetylation marks and k represents the number of combinatorial histone acetylation states. High coverage regions were defined using the results of the ChromHMM Model. Specifically, the 200bp genomic bins corresponding to states with enrichment of multiple histone acetylation marks (states 1,2,3,4,6,9,10,11,12,15,16) were merged to form high coverage regions. Then, to reduce the number of fragmented or spurious regions, bins with 400 base pairs (2 genomic windows) were merged and regions with size less than 400 base pairs (2 genomic windows) were filtered out. A initial normalized data matrix (N x M) was generated by computing the coverage of each histone modification over each region and normalizing for region size and histone abundance. Specifically, to account for differences in region size between regions, the total reads per region was scaled by region size and, to account for differences in total measured histone abundance between marks, sigmoidal scaling was used[112,113]. NMF was then performed using 'Nimfa'[114], a Python library for nonnegative matrix factorization, with the nndsvd initialization method. The rank *k* was selected empirically, taking into account the biological assignability of the resulting states, the complexity of the model and the stability of the factorization (the number of iterations the algorithm required to coverage).

After factorization, the resulting basis matrix (N x k) contained the coefficient of each

combination *i* for each genomic region. A sorted heatmap of the basis matrix was generated by grouping the regions according to the combination that contributed the greatest coefficient for each region. For visualization, this heatmap was normalized by dividing the coefficients for each region by the total coefficient sum of the region.

To profile and assign a biological interpretation to individual combinations, each region was assigned to the combination with the maximum coefficient. Identification of transcription factors with significant binding overlap to regions assigned to a single combination was performed using the Cistrome Data Browser, an interactive database of public ChIPseq[115]. For each combination, the top 100 scores were filtered for targets with at least two hits in any cell type. Motif enrichment was calculated using the HOMER function 'findMotifs' on all genomic regions assigned to each combination. For comparison of enrichment levels in C4 versus C5, enrichments were calculated using bedGraphs from the mESC 165 Antibody Pool Experiment and the ChromHMM program 'OverlapEnrichment' (java -jar ChromHMM.jar OverlapEnrichment -binres 1 -signal). Interval bars for these

enrichments were generated by bootstrap resampling; enrichments were recalculated for 200 independent draws of 75% of the regions assigned to C4 or C5.

High-Density Regions of NANOG-OCT4-SOX2

High-density regions of pluripotency associated transcription factors were calculated using the NANOG, OCT4 and SOX2 data from the mESC 165 Antibody Pool Experiment. Specifically, high-density and low-density regions were defined using the super-enhancer setting of the 'callPeaks' function from HOMER on the merged tag directors of the three transcription factors. To remove nonspecific background peaks, the merged tag directories of the background models for these three factors was used as input. Briefly, the super-enhancer setting with default parameters first identifies peaks, then stiches together individual peaks that are within 12.5kb of each other, calculates a 'super-enhancer score' for each region based on input-normalized read coverage, generates a 'super-enhancer plot' (regions sorted by score vs. number of regions) and identifies the regions where the slope of the plot is greater than 1. These regions are labeled as putative 'super enhancers' while all remaining regions are labeled as 'typical enhancers'. We consider the 'super-enhancer' regions as HDR and the 'typical enhancer' regions as LDR.

TF and CR enrichments over HDRs versus LDRs were calculated using the 'computeMatrix' function with scale-regions setting from deepTools v.3.1.3. To account for the differences in typical region size between LDRs and HDRs, which tended to be much larger, the -m parameter was set to approximately the median region size for each group.

GO terms associated with the intersection of HDRs, LDRs and NMF-based acetylation combinations were calculated using the GO analysis function of 'annotatePeaks' from HOMER. To limit the number of terms under consideration, only terms assigned to the biological process category that received a cutoff $p<0.001$ were used. Terms were then manually grouped into larger categories (e. g. developmental, metabolic). Enrichment scores were calculated by normalizing for the total number of possible unique terms

assigned the category and the total number of terms assigned to the intersection group.

Statistics

Pearson correlation coefficients for coverage comparisons versus ENCODE were calculated using Pearson function of scipy.stats library[108]. Pearson correlation coefficients for heatmaps were generated using the 'plotCorrelation' function from deepTools v.3.1.3[96].

References

1. Bednar, J., Horowitz, R.A., Grigoryev, S.A., Carruthers, L.M., Hansen, J.C., Koster, A.J., and Woodcock, C.L. (1998). Nucleosomes, linker DNA, and linker histone form a unique structural motif that directs the higher-order folding and compaction of chromatin. Proc. Natl. Acad. Sci. *95*, 14173–14178. 10.1073/pnas.95.24.14173.

2. Jenuwein, T., and Allis, C.D. (2001). Translating the Histone Code. Science *293*, 1074–1080. 10.1126/science.1063127.

3. Huang, H., Sabari, B.R., Garcia, B.A., Allis, C.D., and Zhao, Y. (2014). SnapShot: Histone Modifications. Cell *159*, 458–458.e1. 10.1016/j.cell.2014.09.037.

4. Tekel, S.J., and Haynes, K.A. (2017). Molecular structures guide the engineering of chromatin. Nucleic Acids Res. *45*, gkx531. 10.1093/nar/gkx531.

5. Mashtalir, N., Dao, H.T., Sankar, A., Liu, H., Corin, A.J., Bagert, J.D., Ge, E.J., D'Avino, A.R., Filipovski, M., Michel, B.C., et al. (2021). Chromatin landscape signals differentially dictate the activities of mSWI/SNF family complexes. Science *373*, 306–315. 10.1126/science.abf8705.

6. He, S., Wu, Z., Tian, Y., Yu, Z., Yu, J., Wang, X., Li, J., Liu, B., and Xu, Y. (2020). Structure of nucleosome-bound human BAF complex. Science *367*, 875–881. 10.1126/science.aaz9761.

7. Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330. 10.1038/nature14248.

8. Barba-Aliaga, M., Alepuz, P., and Pérez-Ortín, J.E. (2021). Eukaryotic RNA Polymerases: The Many Ways to Transcribe a Gene. Front. Mol. Biosci. *8*, 663209. 10.3389/fmolb.2021.663209.

9. ROEDER, R.G. (1998). Role of General and Gene-specific Cofactors in the Regulation of Eukaryotic Transcription. Cold Spring Harb. Symp. Quant. Biol. *63*, 201–218. 10.1101/sqb.1998.63.201.

10. Malik, S., and Roeder, R.G. (2023). Regulation of the RNA polymerase II pre-initiation complex by its associated coactivators. Nat. Rev. Genet. *24*, 767–782. 10.1038/s41576-023-00630-9.

11. Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science *316*, 1497–1502. 10.1126/science.1141319.

12. Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R.P., et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553–560. 10.1038/nature06008.

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

14. Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., et al. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat. Methods *4*, 651–657. 10.1038/nmeth1068.

15. He, Q., Johnston, J., and Zeitlinger, J. (2015). ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. Nat. Biotechnol. *33*, 395–401. 10.1038/nbt.3121.

16. Serandour, A.A., Brown, G.D., Cohen, J.D., and Carroll, J.S. (2013). Development of an Illumina-based ChIP-exonuclease method provides insight into FoxA1-DNA binding properties. Genome Biol. *14*, R147. 10.1186/gb-2013-14-12-r147.

17. Tehranchi, A.K., Myrthil, M., Martin, T., Hie, B.L., Golan, D., and Fraser, H.B. (2016). Pooled ChIP-Seq Links Variation in Transcription Factor Binding to Complex Disease Risk. Cell *165*, 730–741. 10.1016/j.cell.2016.03.041.

18. Aldridge, S., Watt, S., Quail, M.A., Rayner, T., Lukk, M., Bimson, M.F., Gaffney, D., and Odom, D.T. (2013). AHT-ChIP-Seq: a completely automated robotic protocol for high-throughput chromatin immunoprecipitation. Genome Biol. *14*, R124. 10.1186/gb- 2013-14-11-r124.

19. Janssens, D.H., Meers, M.P., Wu, S.J., Babaeva, E., Meshinchi, S., Sarthy, J.F., Ahmad, K., and Henikoff, S. (2021). Automated CUT&Tag profiling of chromatin heterogeneity in mixed-lineage leukemia. Nat. Genet. *53*, 1586–1596. 10.1038/s41588-021-00941-9.

20. Kaya-Okur, H.S., Wu, S.J., Codomo, C.A., Pledger, E.S., Bryson, T.D., Henikoff, J.G., Ahmad, K., and Henikoff, S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. Nat. Commun. *10*, 1930. 10.1038/s41467-019-09982-5.

21. Skene, P.J., Henikoff, J.G., and Henikoff, S. (2018). Targeted in situ genome-wide profiling with high efficiency for low cell numbers. Nat. Protoc. *13*, 1006–1019. 10.1038/nprot.2018.015.

22. Bartosovic, M., and Castelo-Branco, G. (2023). Multimodal chromatin profiling using nanobody-based single-cell CUT&Tag. Nat. Biotechnol. *41*, 794–805.

10.1038/s41587- 022-01535-4.

23. Gopalan, S., Wang, Y., Harper, N.W., Garber, M., and Fazzio, T.G. (2021). Simultaneous profiling of multiple chromatin proteins in the same cells. Mol. Cell *81*, 4736–4746.e5. 10.1016/j.molcel.2021.09.019.

24. Lochs, S.J.A., Weide, R.H. van der, Luca, K.L. de, Korthout, T., Beek, R.E. van, Kimura, H., and Kind, J. (2023). Combinatorial single-cell profiling of major chromatin types with MAbID. Nat. Methods, 1–11. 10.1038/s41592-023-02090-9.

25. Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74. 10.1038/nature11247.

26. Consortium, P., Akbarian, S., Liu, C., Knowles, J.A., Vaccarino, F.M., Farnham, P.J., Crawford, G.E., Jaffe, A.E., Pinto, D., Dracheva, S., et al. (2015). The PsychENCODE project. Nat. Neurosci. *18*, 1707–1712. 10.1038/nn.4156.

27. Consortium, T.I.G.P., Heng, T.S.P., Painter, M.W., Elpek, K., Lukacs-Kornek, V., Mauermann, N., Turley, S.J., Koller, D., Kim, F.S., Wagers, A.J., et al. (2008). The Immunological Genome Project: networks of gene expression in immune cells. Nat. Immunol. *9*, 1091–1094. 10.1038/ni1008-1091.

28. Partridge, E.C., Chhetri, S.B., Prokop, J.W., Ramaker, R.C., Jansen, C.S., Goh, S.-T., Mackiewicz, M., Newberry, K.M., Brandsmeier, L.A., Meadows, S.K., et al. (2020). Occupancy maps of 208 chromatin-associated proteins in one human cell type. Nature *583*, 720–728. 10.1038/s41586-020-2023-4.

29. He, Y., Hariharan, M., Gorkin, D.U., Dickel, D.E., Luo, C., Castanon, R.G., Nery, J.R., Lee, A.Y., Zhao, Y., Huang, H., et al. (2020). Spatiotemporal DNA methylome dynamics of the developing mouse fetus. Nature *583*, 752–759. 10.1038/s41586-020- 2119-x.

30. Sisu, C., Muir, P., Frankish, A., Fiddes, I., Diekhans, M., Thybert, D., Odom, D.T., Flicek, P., Keane, T.M., Hubbard, T., et al. (2020). Transcriptional activity and strain- specific history of mouse pseudogenes. Nat. Commun. *11*, 3695. 10.1038/s41467-020- 17157-w.

31. Chasman, D., and Roy, S. (2017). Inference of cell type specific regulatory networks on mammalian lineages. Curr. Opin. Syst. Biol. *2*, 130–139. 10.1016/j.coisb.2017.04.001.

32. Ota, M., Nagafuchi, Y., Hatano, H., Ishigaki, K., Terao, C., Takeshima, Y., Yanaoka, H., Kobayashi, S., Okubo, M., Shirai, H., et al. (2021). Dynamic landscape of immune cell-specific gene regulation in immune-mediated diseases. Cell *184*, 3006–3021.e17. 10.1016/j.cell.2021.03.056.

33. Madhani, H.D., Francis, N.J., Kingston, R.E., Kornberg, R.D., Moazed, D.,

Narlikar, G.J., Panning, B., and Struhl, K. (2008). Epigenomics: A Roadmap, But to Where? Science *322*, 43–44. 10.1126/science.322.5898.43b.

34. Kidder, B.L., Hu, G., and Zhao, K. (2011). ChIP-Seq: technical considerations for obtaining high-quality data. Nat. Immunol. *12*, 918–922. 10.1038/ni.2117.

35. Arrastia, M.V., Jachowicz, J.W., Ollikainen, N., Curtis, M.S., Lai, C., Quinodoz, S.A., Selck, D.A., Ismagilov, R.F., and Guttman, M. (2022). Single-cell measurement of higher-order 3D genome organization with scSPRITE. Nat. Biotechnol. *40*, 64–73. 10.1038/s41587-021-00998-1.

36. Quinodoz, S.A., Jachowicz, J.W., Bhat, P., Ollikainen, N., Banerjee, A.K., Goronzy, I.N., Blanco, M.R., Chovanec, P., Chow, A., Markaki, Y., et al. (2021). RNA promotes the formation of spatial compartments in the nucleus. Cell *184*, 5775–5790.e30. 10.1016/j.cell.2021.10.014.

37. Goronzy, I.N., Quinodoz, S.A., Jachowicz, J.W., Ollikainen, N., Bhat, P., and Guttman, M. (2022). Simultaneous mapping of 3D structure and nascent RNAs argues against nuclear compartments that preclude transcription. Cell Rep. *41*, 111730. 10.1016/j.celrep.2022.111730.

38. Quinodoz, S.A., Ollikainen, N., Tabak, B., Palla, A., Schmidt, J.M., Detmar, E., Lai, M.M., Shishkin, A.A., Bhat, P., Takei, Y., et al. (2018). Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. Cell *174*, 744–757.e24. 10.1016/j.cell.2018.05.024.

39. Quinodoz, S.A., Bhat, P., Chovanec, P., Jachowicz, J.W., Ollikainen, N., Detmar, E., Soehalim, E., and Guttman, M. (2022). SPRITE: a genome-wide method for mapping higher-order 3D interactions in the nucleus using combinatorial split-and-pool barcoding. Nat. Protoc. *17*, 36–75. 10.1038/s41596-021-00633-y.

40. Kim, S., Yu, N.-K., and Kaang, B.-K. (2015). CTCF as a multifunctional protein in genome regulation and gene expression. Exp. Mol. Med. *47*, e166–e166. 10.1038/emm.2015.33.

41. Kouzarides, T. (2007). Chromatin Modifications and Their Function. Cell *128*, 693–705. 10.1016/j.cell.2007.02.005.

42. Girbig, M., Misiaszek, A.D., and Müller, C.W. (2022). Structural insights into nuclear transcription by eukaryotic DNA-dependent RNA polymerases. Nat. Rev. Mol. Cell Biol. *23*, 603–622. 10.1038/s41580-022-00476-9.

43. Abascal, F., Acosta, R., Addleman, N.J., Adrian, J., Afzal, V., Ai, R., Aken, B., Akiyama, J.A., Jammal, O.A., Amrhein, H., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature *583*, 699–710. 10.1038/s41586-020-2493-4.

44. Dahl, J.A., and Gilfillan, G.D. (2017). How low can you go? Pushing the limits of low-input ChIP-Seq. Brief. Funct. Genom. *17*, 89–95. 10.1093/bfgp/elx037.

45. Adli, M., Zhu, J., and Bernstein, B.E. (2010). Genome-wide chromatin maps derived from limited numbers of hematopoietic progenitors. Nat. Methods *7*, 615–618. 10.1038/nmeth.1478.

46. Karimzadeh, M., and Hoffman, M.M. (2022). Virtual ChIP-Seq: predicting transcription factor binding by learning from the transcriptome. Genome Biol. *23*, 126. 10.1186/s13059-022-02690-2.

47. Ernst, J., and Kellis, M. (2017). Chromatin-state discovery and genome annotation with ChromHMM. Nat. Protoc. *12*, 2478–2492. 10.1038/nprot.2017.124.

48. Spicuglia, S., and Vanhille, L. (2012). Chromatin signatures of active enhancers. Nucleus *3*, 126–131. 10.4161/nucl.19232.

49. Steger, D.J., Lefterova, M.I., Ying, L., Stonestrom, A.J., Schupp, M., Zhuo, D., Vakoc, A.L., Kim, J.-E., Chen, J., Lazar, M.A., et al. (2008). DOT1L/KMT4 Recruitment and H3K79 Methylation Are Ubiquitously Coupled with Gene Transcription in Mammalian Cells. Mol. Cell. Biol. *28*, 2825–2839. 10.1128/mcb.02076-07.

50. Gates, L.A., Foulds, C.E., and O'Malley, B.W. (2017). Histone Marks in the 'Driver's Seat': Functional Roles in Steering the Transcription Cycle. Trends Biochem. Sci. *42*, 977–989. 10.1016/j.tibs.2017.10.004.

51. Karmodiya, K., Krebs, A.R., Oulad-Abdelghani, M., Kimura, H., and Tora, L. (2012). H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells. BMC Genom. *13*, 424. 10.1186/1471-2164-13-424.

52. Chen, Z., Djekidel, M.N., and Zhang, Y. (2021). Distinct dynamics and functions of H2AK119ub1 and H3K27me3 in mouse preimplantation embryos. Nat. Genet. *53*, 551–
563. 10.1038/s41588-021-00821-2.

53. Saksouk, N., Simboeck, E., and Déjardin, J. (2015). Constitutive heterochromatin formation and transcription in mammals. Epigenetics Chromatin *8*, 3. 10.1186/1756- 8935-8-3.

54. Chen, T., and Dent, S.Y.R. (2014). Chromatin modifiers and remodellers: regulators of cellular differentiation. Nat. Rev. Genet. *15*, 93–106. 10.1038/nrg3607.

55. Ho, L., and Crabtree, G.R. (2010). Chromatin remodelling during development. Nature *463*, 474–484. 10.1038/nature08911.

56. Kirtana, R., Manna, S., and Patra, S.K. (2020). Molecular mechanisms of KDM5A in cellular functions: Facets during development and disease. Exp. Cell Res. *396*, 112314. 10.1016/j.yexcr.2020.112314.

57. Shilatifard, A. (2008). Molecular implementation and physiological roles for histone H3 lysine 4 (H3K4) methylation. Curr. Opin. Cell Biol. *20*, 341–348. 10.1016/j.ceb.2008.03.019.

58. Geng, Z., and Gao, Z. (2020). Mammalian PRC1 Complexes: Compositional Complexity and Diverse Molecular Mechanisms. Int. J. Mol. Sci. *21*, 8594. 10.3390/ijms21228594.

59. Van Mierlo, G., Veenstra, G.J.C., Vermeulen, M., and Marks, H. (2019). The Complexity of PRC2 Subcomplexes. Trends Cell Biol. *29*, 660–671. 10.1016/j.tcb.2019.05.004.

60. Bosch-Presegué, L., Raurell-Vila, H., Thackray, J.K., González, J., Casal, C., Kane- Goldsmith, N., Vizoso, M., Brown, J.P., Gómez, A., Ausió, J., et al. (2017). Mammalian HP1 Isoforms Have Specific Roles in Heterochromatin Structure and Organization. Cell Rep. *21*, 2048–2057. 10.1016/j.celrep.2017.10.092.

61. Mazzocca, M., Colombo, E., Callegari, A., and Mazza, D. (2021). Transcription factor binding kinetics and transcriptional bursting: What do we really know? Curr. Opin. Struct. Biol. *71*, 239–248. 10.1016/j.sbi.2021.08.002.

62. Bartman, C.R., Hamagami, N., Keller, C.A., Giardine, B., Hardison, R.C., Blobel, G.A., and Raj, A. (2019). Transcriptional Burst Initiation and Polymerase Pause Release Are Key Control Points of Transcriptional Regulation. Mol. Cell *73*, 519–532.e4. 10.1016/j.molcel.2018.11.004.

63. Rada-Iglesias, A., Ameur, A., Kapranov, P., Enroth, S., Komorowski, J., Gingeras, T.R., and Wadelius, C. (2008). Whole-genome maps of USF1 and USF2 binding and histone H3 acetylation reveal new aspects of promoter structure and candidate genes for common human disorders. Genome Res. *18*, 380–392. 10.1101/gr.6880908.

64. O'Connor, L., Gilmour, J., and Bonifer, C. (2016). The Role of the Ubiquitously Expressed Transcription Factor Sp1 in Tissue-specific Transcriptional Regulation and in Disease. Yale J. Biol. Med. *89*, 513–525.

65. Li, Z., Cogswell, M., Hixson, K., Brooks-Kayal, A.R., and Russek, S.J. (2018). Nuclear Respiratory Factor 1 (NRF-1) Controls the Activity Dependent Transcription of the GABA-A Receptor Beta 1 Subunit Gene in Neurons. Front. Mol. Neurosci. *11*, 285. 10.3389/fnmol.2018.00285.

66. Horn, H.F., and Vousden, K.H. (2007). Coping with stress: multiple ways to activate p53. Oncogene *26*, 1306–1316. 10.1038/sj.onc.1210263.

67. Fischer, M. (2017). Census and evaluation of p53 target genes. Oncogene *36*, 3943–3956. 10.1038/onc.2016.502.

68. Akberdin, I.R., Omelyanchuk, N.A., Fadeev, S.I., Leskova, N.E., Oschepkova, E.A., Kazantsev, F.V., Matushkin, Y.G., Afonnikov, D.A., and Kolchanov, N.A. (2018). Pluripotency gene network dynamics: System views from parametric analysis. PLoS ONE *13*, e0194464. 10.1371/journal.pone.0194464.

69. Reith, W., Herrero-Sanchez, C., Kobr, M., Silacci, P., Berte, C., Barras, E., Fey, S., and Mach, B. (1990). MHC class II regulatory factor RFX has a novel DNA-binding domain and a functionally independent dimerization domain. Genes Dev. *4*, 1528–1540. 10.1101/gad.4.9.1528.

70. Brivanlou, A.H., and Darnell, Jr., J.E. (2002). Signal Transduction and the Control of Gene Expression. Science *295*, 813–818. 10.1126/science.1066355.

71. Satoh, J., Kawana, N., and Yamamoto, Y. (2013). Pathway Analysis of ChIP-Seq- Based NRF1 Target Genes Suggests a Logical Hypothesis of their Involvement in the Pathogenesis of Neurodegenerative Diseases. Gene Regul. Syst. Biol. *7*, GRSB.S13204. 10.4137/grsb.s13204.

72. Qi, B., Newcomer, R., and Sang, Q.-X. (2009). ADAM19/Adamalysin 19 Structure, Function, and Role as a Putative Target in Tumors and Inflammatory Diseases. Curr. Pharm. Des. *15*, 2336–2348. 10.2174/138161209788682352.

73. Schoch, S., Cibelli, G., and Thiel, G. (1996). Neuron-specific Gene Expression of Synapsin I MAJOR ROLE OF A NEGATIVE REGULATORY MECHANISM (∗). J. Biol. Chem. *271*, 3317–3323. 10.1074/jbc.271.6.3317.

74. Martin, D., and Grapin-Botton, A. (2017). The Importance of REST for Development and Function of Beta Cells. Front. Cell Dev. Biol. *5*, 12. 10.3389/fcell.2017.00012.

75. Bao, F., LoVerso, P.R., Fisk, J.N., Zhurkin, V.B., and Cui, F. (2017). p53 binding sites in normal and cancer cells are characterized by distinct chromatin context. Cell Cycle *16*, 2073–2085. 10.1080/15384101.2017.1361064.

76. Otto, S.J., McCorkle, S.R., Hover, J., Conaco, C., Han, J.-J., Impey, S., Yochum, G.S., Dunn, J.J., Goodman, R.H., and Mandel, G. (2007). A New Binding Motif for the Transcriptional Repressor REST Uncovers Large Gene Networks Devoted to Neuronal Functions. J. Neurosci. *27*, 6729–6739. 10.1523/jneurosci.0091-07.2007.

77. Ernst, J., and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. Nat. Biotechnol. *28*, 817–825. 10.1038/nbt.1662.

78. Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. Cell *125*, 315–326. 10.1016/j.cell.2006.02.041.

79. Wang, H., Fan, Z., Shliaha, P.V., Miele, M., Hendrickson, R.C., Jiang, X., and Helin, K. (2023). H3K4me3 regulates RNA polymerase II promoter-proximal pause-release. Nature *615*, 339–348. 10.1038/s41586-023-05780-8.

80. Bilodeau, S., Kagey, M.H., Frampton, G.M., Rahl, P.B., and Young, R.A. (2009). SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state. Genes Dev. *23*, 2484–2489. 10.1101/gad.1837309.

81. Zentner, G.E., and Henikoff, S. (2013). Regulation of nucleosome dynamics by histone modifications. Nat. Struct. Mol. Biol. *20*, 259–266. 10.1038/nsmb.2470.

82. Giaimo, B.D., Ferrante, F., Vallejo, D.M., Hein, K., Gutierrez-Perez, I., Nist, A., Stiewe, T., Mittler, G., Herold, S., Zimmermann, T., et al. (2018). Histone variant H2A.Z deposition and acetylation directs the canonical Notch signaling response. Nucleic Acids Res. *46*, gky551-. 10.1093/nar/gky551.

83. Giaimo, B.D., Ferrante, F., Herchenröther, A., Hake, S.B., and Borggrefe, T. (2019). The histone variant H2A.Z in gene regulation. Epigenetics Chromatin *12*, 37. 10.1186/s13072-019-0274-9.

84. Gévry, N., Chan, H.M., Laflamme, L., Livingston, D.M., and Gaudreau, L. (2007). p21 transcription is regulated by differential localization of histone H2A.Z. Genes Dev. *21*, 1869–1881. 10.1101/gad.1545707.

85. Gévry, N., Hardy, S., Jacques, P.-É., Laflamme, L., Svotelis, A., Robert, F., and Gaudreau, L. (2009). Histone H2A.Z is essential for estrogen receptor signaling. Genes Dev. *23*, 1522–1533. 10.1101/gad.1787109.

86. Akerberg, B.N., Gu, F., VanDusen, N.J., Zhang, X., Dong, R., Li, K., Zhang, B., Zhou, B., Sethi, I., Ma, Q., et al. (2019). A reference map of murine cardiac transcription factor chromatin occupancy identifies dynamic and conserved enhancers. Nat. Commun. *10*, 4907. 10.1038/s41467-019-12812-3.

87. Currey, L., Thor, S., and Piper, M. (2021). TEAD family transcription factors in development and disease. Development *148*. 10.1242/dev.196675.

88. Vangala, P., Murphy, R., Quinodoz, S.A., Gellatly, K., McDonel, P., Guttman, M., and Garber, M. (2020). High-Resolution Mapping of Multiway Enhancer-Promoter Interactions Regulating Pathogen Detection. Mol. Cell *80*, 359–373.e8.

10.1016/j.molcel.2020.09.005.

89. Weinert, B.T., Narita, T., Satpathy, S., Srinivasan, B., Hansen, B.K., Schölz, C., Hamilton, W.B., Zucconi, B.E., Wang, W.W., Liu, W.R., et al. (2018). Time-Resolved Analysis Reveals Rapid Dynamics and Broad Scope of the CBP/p300 Acetylome. Cell *174*, 231–244.e12. 10.1016/j.cell.2018.04.033.

90. Fang, Z., Wang, X., Sun, X., Hu, W., and Miao, Q.R. (2021). The Role of Histone Protein Acetylation in Regulating Endothelial Function. Front. Cell Dev. Biol. *9*, 672447. 10.3389/fcell.2021.672447.

91. Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetJ. *17*, 10–12. 10.14806/ej.17.1.200.

92. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. 10.1038/nmeth.1923.

93. Amemiya, H.M., Kundaje, A., and Boyle, A.P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci. Rep. *9*, 9354. 10.1038/s41598-019-45839-z.

94. Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics *32*, 3047–3048. 10.1093/bioinformatics/btw354.

95. Waskom, M. (2021). seaborn: statistical data visualization. J. Open Source Softw. *6*, 3021. 10.21105/joss.03021.

96. Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep- sequencing data analysis. Nucleic Acids Res. *44*, W160–W165. 10.1093/nar/gkw257.

97. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26. 10.1038/nbt.1754.

98. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol. Cell *38*, 576–589. 10.1016/j.molcel.2010.05.004.

99. Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Federhen, S., et al. (2011). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. *39*, D38–D51. 10.1093/nar/gkq1172.

100. George, S.S., Pimkin, M., and Paralkar, V.R. (2023). Customized genomes for human and mouse ribosomal DNA mapping. bioRxiv, 2022.11.10.514243. 10.1101/2022.11.10.514243.

101. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. 10.1093/bioinformatics/btq033.

102. McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv. 10.48550/arxiv.1802.03426.

103. Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods *9*, 215–216. 10.1038/nmeth.1906.

104. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, 1000 Genome Project Data Processing (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. 10.1093/bioinformatics/btp352.

105. Dreos, R., Ambrosini, G., Groux, R., Cavin Périer, R., and Bucher, P. (2017). The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. Nucleic Acids Res. *45*, D51–D55. 10.1093/nar/gkw1069.

106. Smit, A., Hubley, R., and Green, P. RepeatMasker Open-4.0.

107. Kluger, Y., Basri, R., Chang, J.T., and Gerstein, M. (2003). Spectral Biclustering of Microarray Data: Coclustering Genes and Conditions. Genome Res. *13*, 703–716. 10.1101/gr.648603.

108. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272. 10.1038/s41592-019-0686-2.

109. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M., and Ishiguro-Watanabe, M. (2022). KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. *51*, D587–D592. 10.1093/nar/gkac963.

110. Court, F., and Arnaud, P. (2016). An annotated list of bivalent chromatin regions in human ES cells: a new tool for cancer epigenetic research. Oncotarget *8*, 4110–4124. 10.18632/oncotarget.13746.

111. Lidschreiber, K., Jung, L.A., Emde, H., Dave, K., Taipale, J., Cramer, P., and Lidschreiber, M. (2021). Transcriptionally active enhancers in human cancer cells. Mol. Syst. Biol. *17*, e9873. 10.15252/msb.20209873.

112. Jain, A., Nandakumar, K., and Ross, A. (2005). Score normalization in multimodal biometric systems. Pattern Recognit. *38*, 2270–2285. 10.1016/j.patcog.2005.01.012.

113. Cieślik, M., and Bekiranov, S. (2014). Combinatorial epigenetic patterns as quantitative predictors of chromatin biology. BMC Genom. *15*, 76. 10.1186/1471-2164-15-76.

114. Zitnik, M., and Zupan, B. (2018). NIMFA: A Python Library for Nonnegative Matrix Factorization. arXiv. 10.48550/arxiv.1808.01743.

115. Zheng, R., Wan, C., Mei, S., Qin, Q., Wu, Q., Sun, H., Chen, C.-H., Brown, M., Zhang, X., Meyer, C.A., et al. (2018). Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. Nucleic Acids Res. *47*, gky1094-. 10.1093/nar/gky1094.

Chapter 3 - ChIP DIP Development

Introduction

Each gene is constantly and dynamically modulated to ensure the survival and stability of each cell and its cellular functions. There are roughly one thousand five hundred proteins that regulate the activity of all genes in a cell through direct or indirect binding to chromatin. While there are many proteins that regulate transcription, studies generally focus on understanding a handful of proteins at a time. Traditionally, ChIP-seq, and more recently CUT&RUN/CUT&tag, has been used to map where on chromatin each protein is bound one protein at a time. Because of the low throughput nature of current chromatin profiling methods, we reasoned that being able to map hundreds of proteins to DNA would dramatically enhance our ability to understand how the cell establishes and maintains the transcriptional levels of each gene. To do this we realized that SPRITE[1] could be used to resolve the identity of each constituent molecule on a bead. Since ChIP-seq utilizes beads for immunoprecipitation, we reasoned that a simple adaptation would enable ChIP-seq to be multiplexed by labeling each set of beads with a unique oligonucleotide sequence (antibody ID oligo) followed by split pool barcoding. This multiplexed method is coined ChIP DIP, or ChIP Done in Parallel. To accomplish this, we considered many facets including the design of the antibody ID oligo, different modalities to attach the antibody ID oligo to the beads, bead attributes, dissociation rates of noncovalent interactions, antibody amounts for each target, antibody pool composition, and the number of rounds and duration of split pool.

Design Considerations

Oligo Design

Antibody ID oligos need to have several features that are required to serve the purpose of a antibody ID oligo as well as several features that enable them to feed into split pool barcoding ligation reactions. The first requirement is a nine-nucleotide sequence that denotes the identity of the barcode. We chose a nine-nucleotide string of barcodes because this leaves enough combinatorial space for 16,384 different barcode combinations assuming there are up to two sequencing errors when reading the barcode on current high throughput sequencing platforms. Given that there are roughly one thousand different direct and indirect chromatin binding proteins in each cell type[2], this would enable us to choose up to ten antibodies per protein if we wanted to profile every protein in a single experiment. Next, we needed to decide where on the oligonucleotide the barcode is. To avoid a monotemplate sequence during the first few cycles of each sequencing run, which would drastically compromise the sequencing platform's ability to differentiate the different molecules on the flow cell, the barcode starts at the first

nucleotide that is sequenced by the sequencer. Because there is always more than one antibody ID oligo in each experiment, the diversity of the first nucleotides to be sequenced is always diverse using this strategy. Then we considered the length of the unique molecular identifying sequence (UMI). The purpose of a UMI is to distinguish the number of antibody ID oligos on each bead. If all bead oligos have the exact same sequence on the bead, it is impossible to distinguish a unique molecule from a PCR duplicate and prevent the number of antibody ID oligos per bead from being determined. The UMI on each barcode is eight nucleotides long, which will distinguish up to 65,536 antibody ID oligos on each bead. Given that we aim to have 10-20 antibody ID oligos per bead, an eight nucleotide UMI will allow us to identify any PCR duplicated antibody ID oligos. Next, we considered a single stranded antibody ID oligo vs a double stranded antibody ID oligo. A single stranded antibody ID oligo has two advantages. The first is the compatibility with downstream End Repair and dA-tailing enzymatic reactions to blunt and a dA-tail the immunoprecipitated chromatin in preparation of DPM ligation reactions. These reactions will fill in the sticky end on the antibody ID oligo if it is double stranded, destroying the sticky end. The second is that there is no annealing step required as there is with a double stranded antibody ID oligo, saving both on costs and time. Lastly, for ease of use, the bead attachment modality should be on the 3' end of the antibody ID oligo to leave a phosphorylated 5' end with the appropriate sticky end available for downstream ligation to split pool tags.

Labeling Modalities

When choosing a labeling modality for barcoding a bead with an antibody ID oligo, there are three key considerations; 1) the number of antibody ID oligos attached to each bead needs to be controllable and titratable, 2) the modality needs to be stable with a relatively low K_{off} rate to avoid movement of the antibody ID oligo from bead to bead, and 3) the modality needs to be modular, inexpensive, simple. To this end, we tested three different bead labeling modalities; 1) an inert oligo-labeled antibody that can bind to protein G beads through antibody-protein G interactions, 2) a DBCO-Azide based click-chemistry modality, and 3) a streptavidin-biotin labeling modality.

The first modality, oligo-labeled antibody, was able to label protein G beads, but had several drawbacks. First, it was very challenging to efficiently label antibodies with antibody ID oligos using the published CITE-seq protocols. This led to less than one oligo per bead in many cases and made calling the identity of each bead challenging. Second, the K_{off} of antibodies to Protein G are in the $4*10^{-4}$-$1*10^{-5}$ range giving them a half-life of 2-19 hours for an unmodified antibody[3]. The lack of stability between the antibody and protein G interaction led to many different antibody ID oligos on the same bead **(Figure 1)**. Because of these drawbacks, we sought to address them with other labeling modalities.

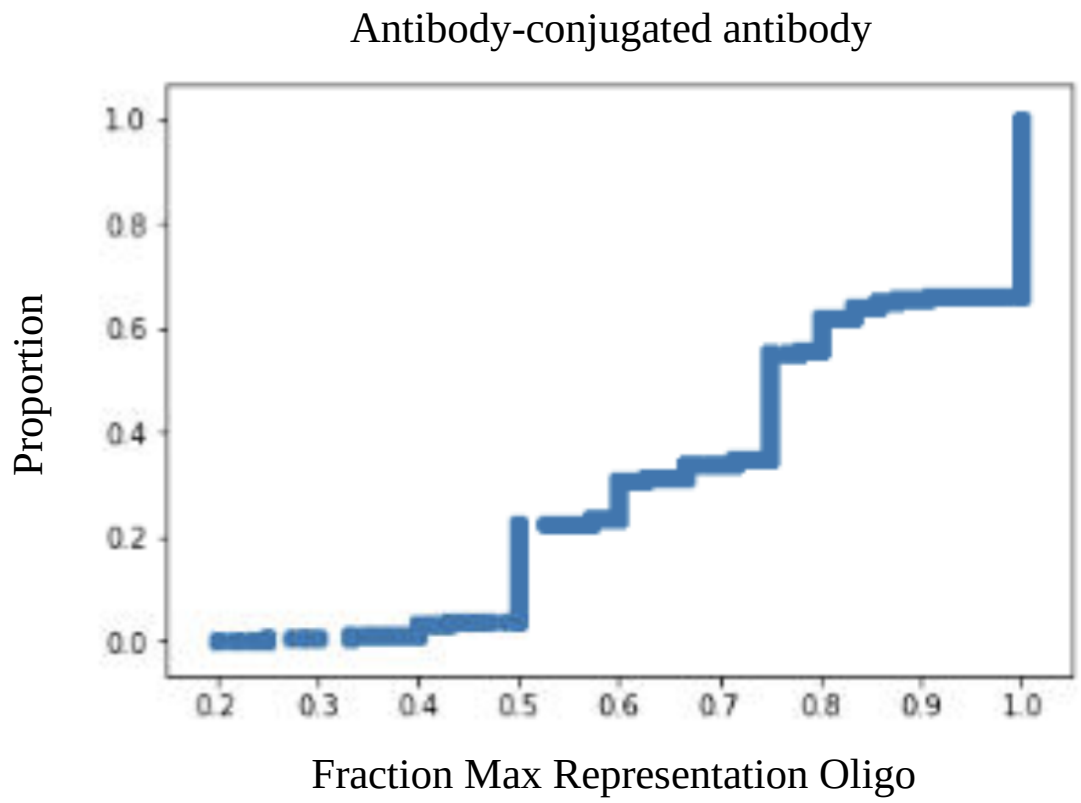

Figure 1: This plot represents the proportion of beads (y-axis) that have a certain fraction of the antibody ID oligo that is maximally represented on each bead (x-axis). This ECDF represents the maximal representation of each antibody ID oligo that bound to protein G beads using the antibody-oligo bead labeling modality.

The second bead labeling modality, DBCO-Azide click chemistry, solved these problems because the chemistry was covalent and there was only one oligo per labeling moiety. This strategy entailed labeling protein G beads with NHS-DBCO and then reacting the DBCO with an antibody ID oligo that had a 3' azide modification (azide-labeled antibody ID oligo). By mixing the DBCO labeled beads with the azide-labeled antibody ID oligo for two hours at room temperature yielded antibody ID oligo labeled beads that are dependent on how much azide-labeled antibody ID oligo is added to the reaction **(Figure 2).** This labeling method suffered from two drawbacks; 1) beads tended to clump after the NHS-DBCO reaction, 2) the use of antibody ID oligos with a 3' azide modification was not scalable due to cost. Due to these considerations, we did not proceed to sequencing with the azide-DBCO click-chemistry labeling modality.

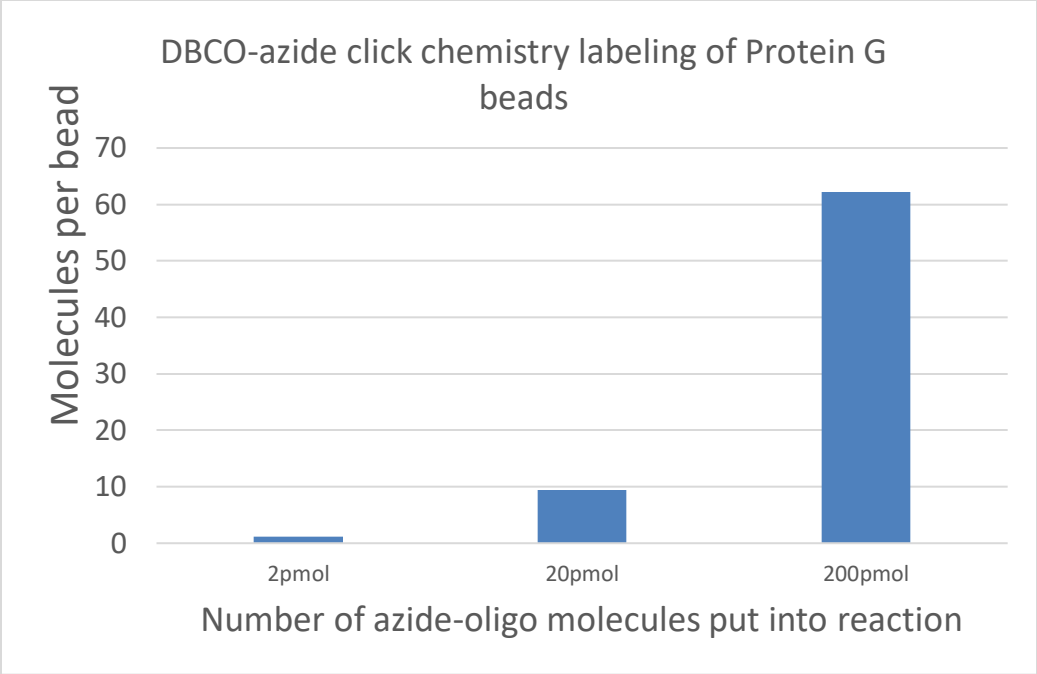

Figure 2: The number of antibody ID oligos per bead detected on each bead using the "bead QC assay" (see protocol chapter for details) after incubating NHS-DBCO labeled Protein G with varying amounts of azide-labeled antibody ID oligo at 25C for 2 hours.

The third bead labeling modality utilized labeling Protein G beads with NHS-biotin followed by binding of the biotinylated beads with subsaturated streptavidin-biotin-oligo to make a "streptavidin (SA) sandwich" to attach the biotinylated antibody ID oligo to the bead. This labeling modality has the most robust results when considering the following metrics; low K_{off} rate for oligo to bead attachment, minimal bead aggregation, reliable oligo labeling densities **(Figure 3)**, and antibody ID oligo loading modularity, costs, and simplicity. Based on these attributes, all experiments utilized this bead labeling modality.

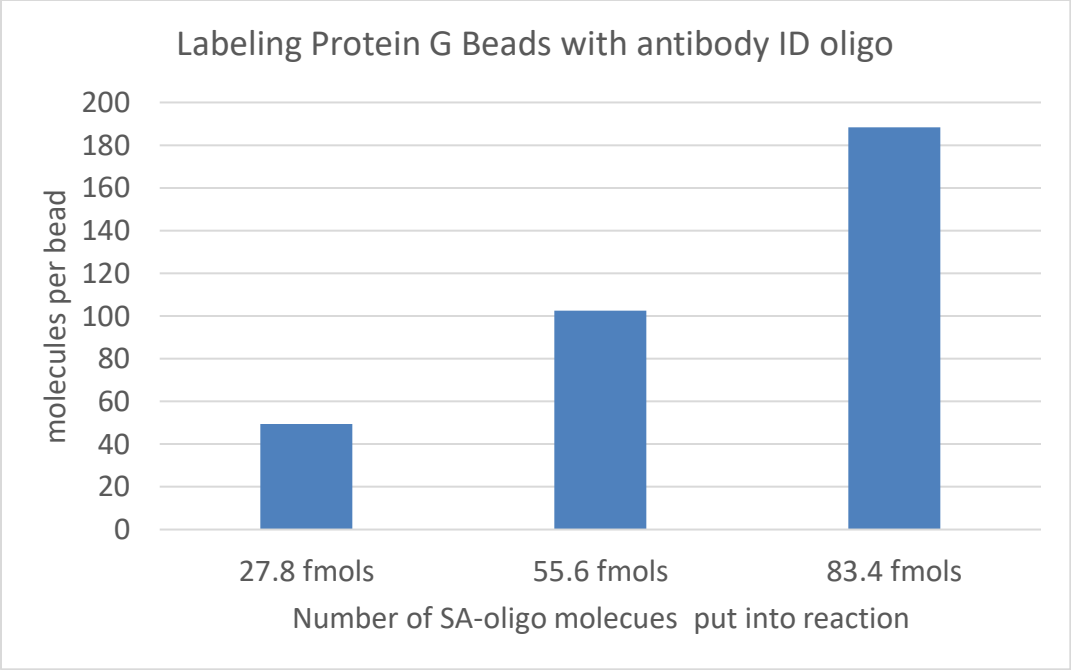

Figure 3: The number of antibody ID oligos per bead detected on each bead using the "bead QC assay" (see protocol chapter for details) after incubating NHS-biotin labeled Protein G with varying amounts of SA-bound antibody ID oligo at 25C for 30 minutes.

Bead Attributes

Next, we considered the bead being used for labeling and capture. For this we considered several attributes; bead size, tendency to clump with other beads, and antibody capacity. We tested two types of beads for this purpose, Protein G Dynabeads (Thermo Fisher Cat. No. 10009D) and Pierce Protein A/G beads (Thermo Scientific Cat. No. 88803). Protein G beads had a bigger diameter (2.8um) compared to protein A/G beads (1um). On the other hand, the antibody binding capacity of protein A/G beads (55-85ug/mg beads) was higher per volume of beads compared to Protein G beads (2.5-3ug/mg beads). Lastly, Protein A/G beads tended to form clumps, making the purity of the oligo on each bead less than desirable **(Figures 4, 5)**. The purity of each antibody ID oligo called on each protein A/G bead made the certainty of correctly assigning chromatin to a protein challenging. For this reason, we selected protein G beads to proceed with.

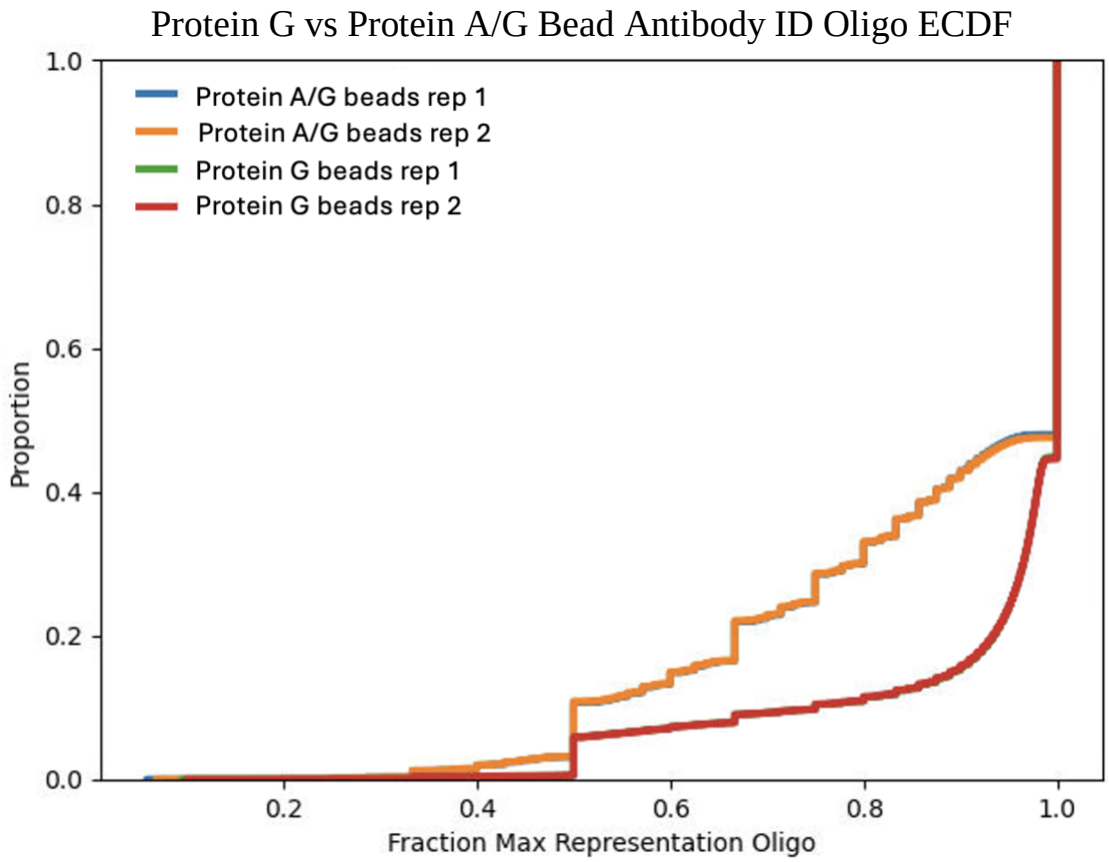

Figure 4: This plot represents the proportion of beads (y-axis) that have a certain fraction of the antibody ID oligo that is maximally represented on each bead (x-axis). We compared two replicates of antibody ID oligo labeled Protein G or Protein A/G beads that were split pooled on separate days. In short, the greater the area under the curve, the more mixed the antibody ID oligos are on each bead. Replicates coincided almost perfectly.

Dissociation rates of noncovalent interactions

Because of the nature of noncovalent interactions that hold the streptavidin to biotin and antibody to protein G, quencher molecules were added to prevent molecules from moving from one bead to bead to another, which would create false signal in a ChIP DIP experiment. In the case of the biotin to streptavidin interactions, this is accomplished by maintaining a concentration of 200uM biotin throughout the experiment. This works in two ways; first the affinity of SA to biotin is cooperative and if all four biotin binding pockets are full then the dissociation constant becomes lower[4], and second, there are no free biotin binding pockets available for antibody ID oligos to move to a different bead than it started on. By adding free biotin, the nonspecific binding of antibody ID oligo is lowered dramatically **(Figure 5)**. In the case of antibody to protein G interactions we tested two different quencher molecules, soluble protein G and soluble rabbit IgG. The rabbit IgG acts similar to the biotin by binding to any bead immobilized protein G that is available to capture antibody that has dissociated from another bead, which prevents the

movement of any antibody that didn't start on a given bead. Soluble protein G works by binding to any antibody that dissociates from its initial bead and keeps any protein G binding sites on the antibody occupied, thus preventing it from binding to any other protein G beads. To test these quenchers, we prepared three sets of antibody ID oligo labeled protein G beads coupled with either; an oligo-conjugated antibody that can be barcoded by split pool, rabbit IgG, or nothing (empty beads) and then pooled all three sets. This pool was treated with either protein G (80uM), IgG (2.5ug/10uL protein G beads), or no quencher control and each condition was split pooled separately. To assess bead to bead antibody movement, that amount oligo-conjugated antibody that ended up on the other two sets of oligo labeled beads, IgG and empty, was assessed. We observed a dramatic reduction of oligo-conjugated antibody that ended up on the empty bead conditions with both protein G and IgG quenchers **(Figure 6)**. We decided to proceed with protein G as a quencher for future experiments because of its low cost and similar efficacy compared to IgG.

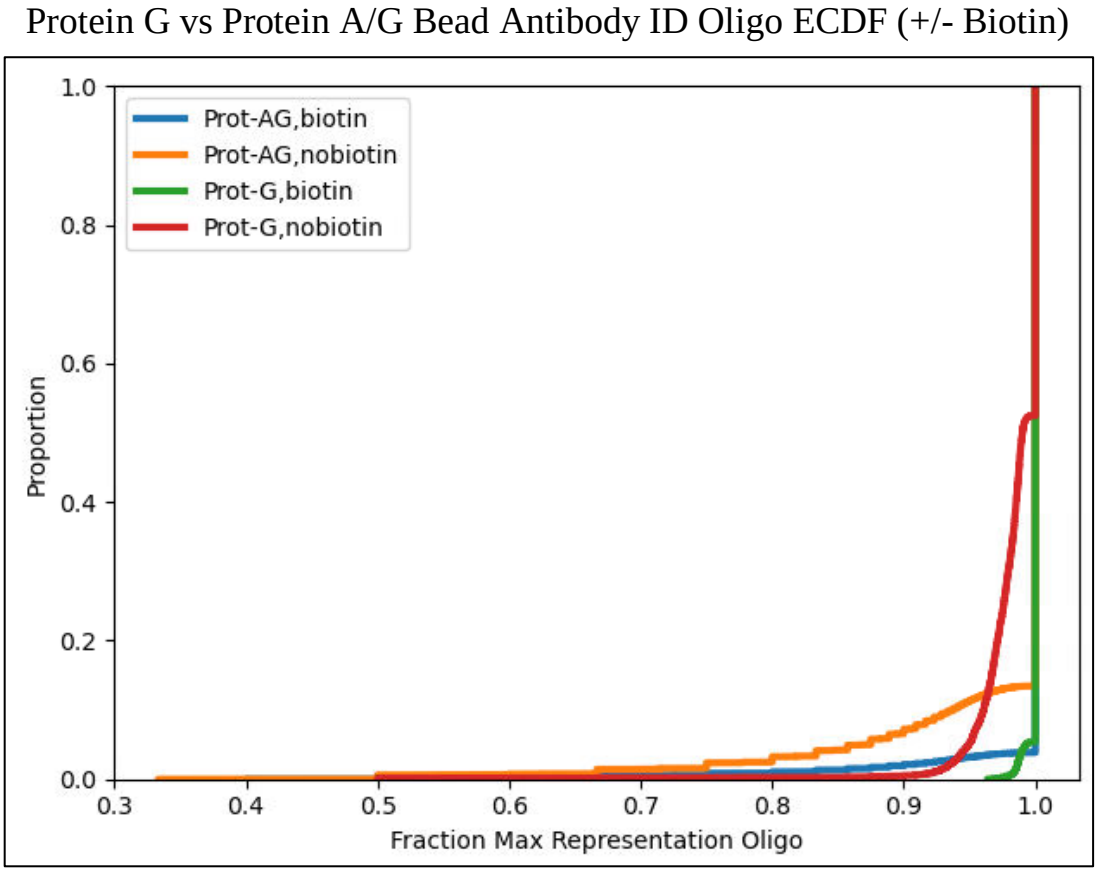

Figure 5: empirical cumulative distribution function (ECDF) plot that represents the proportion of oligo counts (Y-axis) that have a fraction of antibody ID oligo oligos each bead (X-axis). We observed that for both protein G and protein A/G, adding biotin at 200uM final concentration dramatically reduced the number of different antibody ID oligos bound to a given beads. In addition, we observed that protein A/G beads tended to be more mixed compared to their protein G counterpart. Essentially, the greater the area under each curve, the more diverse the antibody ID oligos are on each bead.

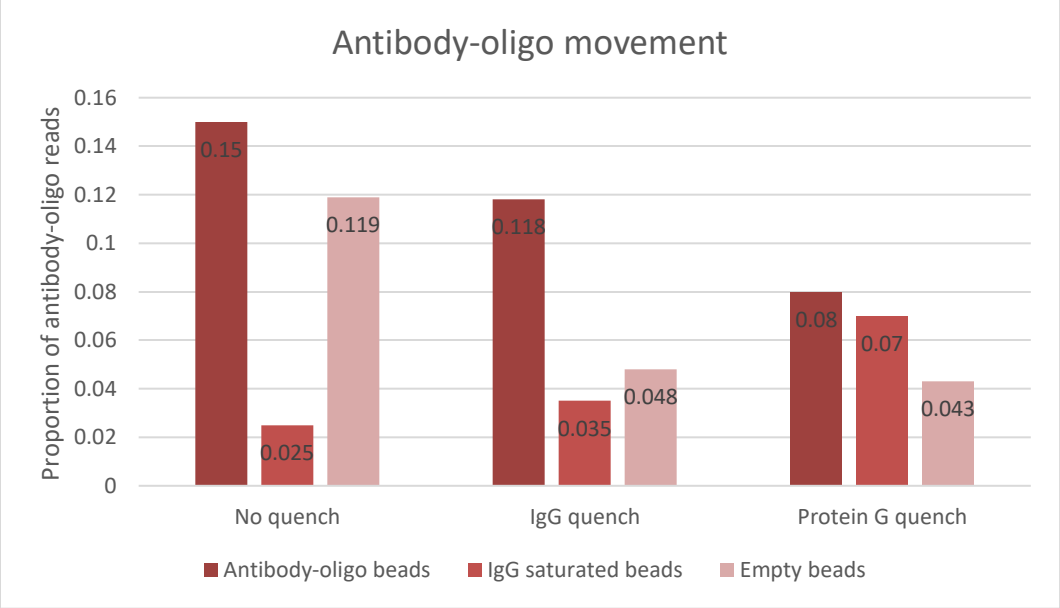

Figure 6: Adding "quencher" molecules, protein G or IgG, reduces the amount of antibody movement from bead to bead. Antibody-oligo beads, IgG saturated beads, and Empty beads were each coupled with unique antibody ID oligos and then pooled, followed by split pooled with either no quench, IgG quench, or protein G quench molecules. The amount of oligo-labeled antibody detected on each set of beads was then read out to quantify the amount of antibody that moved during the experiment in each condition.

Another consideration for the non-covalent interactions is balancing the time it takes to perform a ChIP DIP experiment and the time it takes to complete the various enzymatic reactions throughout the experiment, as well as the buffers used in the various reactions. We found that it is important to follow the recommended reaction times outlined in the protocol chapter to label the maximal number of molecules while minimizing the number of molecules lost during incubations. Specifically, the time required for split pool barcoding should be reasonably minimized by ensuring that the DPM ligation does not exceed one hour, and subsequent split pool ligations should by 8-10 minutes each. This will reduce the amount of chromatin and antibody ID oligo lost throughout the split pool reactions.

Antibody and bead amounts per protein

The amount of chromatin bound to each antibody-bead-oligo combination varies from protein to protein. This means that for an abundant histone modification there could be up to 100 chromatin bound to each bead, while for a lowly expressed transcription factor there could be as little as one chromatin bound to each bead. Because of the pooled nature of ChIP DIP, each antibody-bead-oligo combination cannot be separated after the experiment for sequencing. This was a problem because if the protein that precipitates 100 chromatin per bead gets 10,000,000 reads, then the protein that precipitates one

chromatin per bead would only get 100,000 reads assigned to it. This was addressed in two ways. First, antibodies that precipitate similar amounts of chromatin per bead can be pooled together to prevent large disparities in chromatin per protein. As a rule of thumb, histone antibodies can be pooled together, and transcription factors and chromatin regulator antibodies can be pooled together. Second, to further equalize coverage for each protein target, the amount of antibody can be titrated to precipitate less or more chromatin. For example, if a POLR2A antibody precipitates two chromatin per bead and a Ring1B antibody precipitates 6 chromatin per bead, the amount of Ring1B antibody can be reduced by three-fold to normalize coverage. Alternatively, three-fold more POLR2A antibody (and beads) can be added to the experiment. However, the drawback of adding more beads to an experiment is two-fold; first, more beads need to be resolved with more split pool barcoding and, second, more antibody ID oligo will need to be sequenced for the experiment. For this reason, in general, the amount of antibody was titrated down to normalize coverage between targets. In this way, antibodies that pull down dramatically different amounts of chromatin per bead **(Figure 7)** can give relatively uniform coverage per protein target by normalizing the amount of antibody/bead added to the immunoprecipitation **(Figure 8)**.

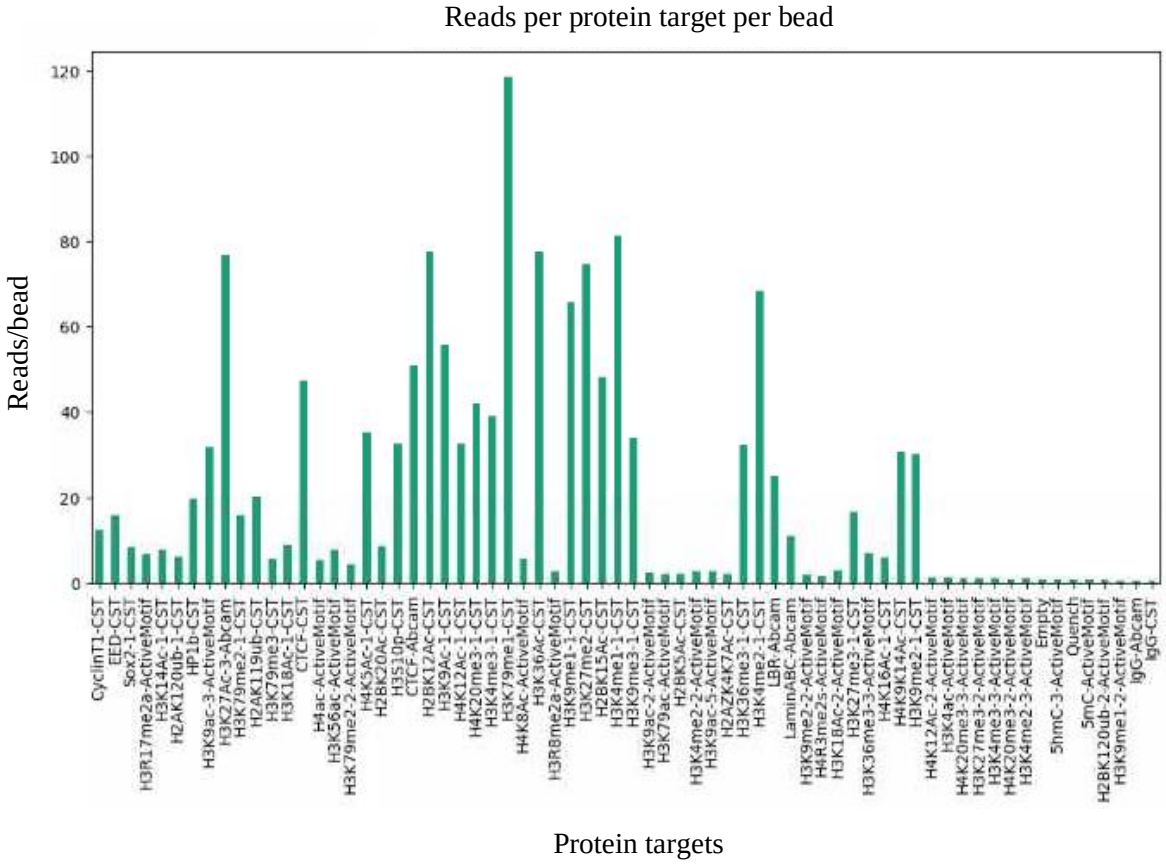

Figure 7: Histogram of the number of reads per bead for each protein target.

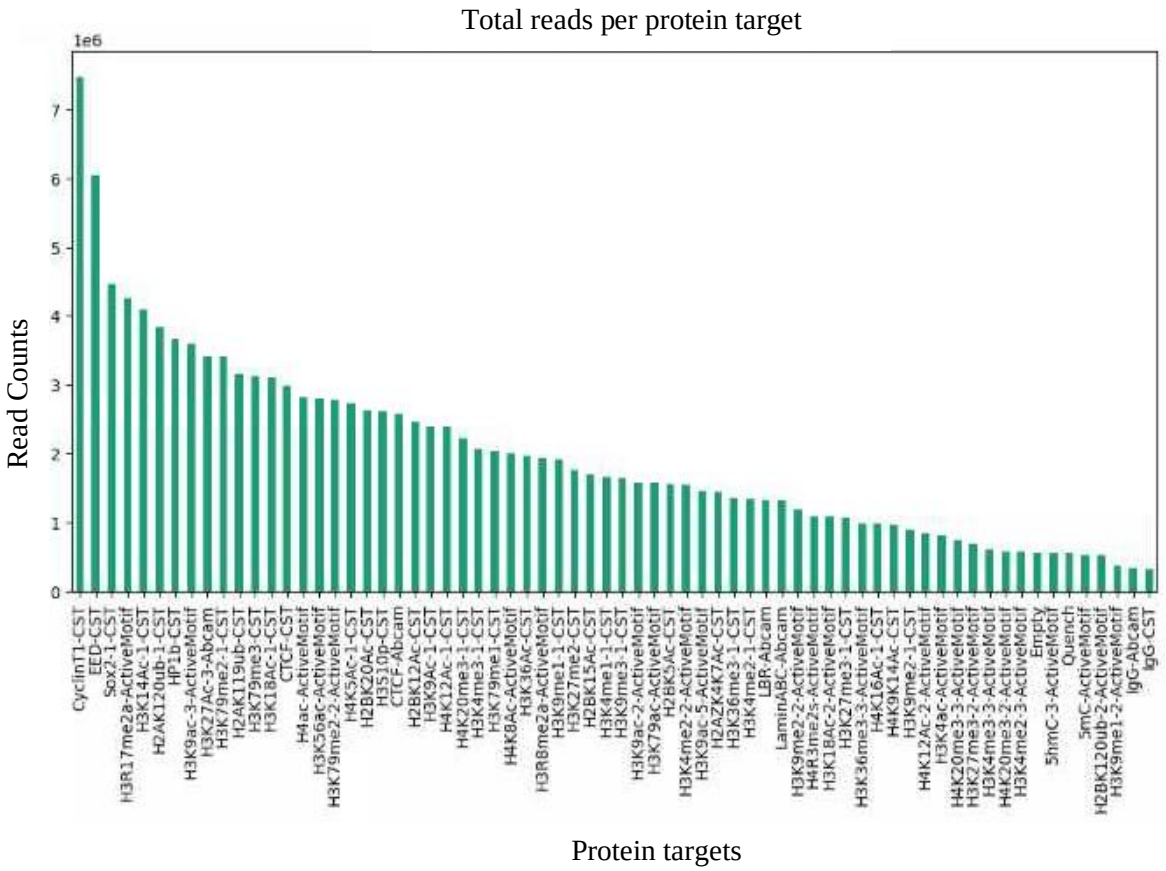

Figure 8: Histogram of the total number of reads for each protein target after adjusting the amount of antibody used to normalize read coverage per protein target.

References

1. Quinodoz, S.A., Bhat, P., Chovanec, P., Jachowicz, J.W., Ollikainen, N., Detmar, E., Soehalim, E., and Guttman, M. (2022). SPRITE: a genome-wide method for mapping higher-order 3D interactions in the nucleus using combinatorial split-and-pool barcoding. Nat Protoc *17*, 36–75. https://doi.org/10.1038/s41596-021-00633-y.

2. Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A., and Luscombe, N.M. (2009). A census of human transcription factors: function, expression and evolution. Nat. Rev. Genet. *10*, 252–263. https://doi.org/10.1038/nrg2538.

3. Corzo, J. (2006). Time, the forgotten dimension of ligand binding teaching. Biochem. Mol. Biol. Educ. *34*, 413–416. https://doi.org/10.1002/bmb.2006.494034062678.

4. Sano, T., and Cantor, C.R. (1990). Cooperative biotin binding by streptavidin. Electrophoretic behavior and subunit association of streptavidin in the presence of 6 M urea. J. Biol. Chem. *265*, 3369–3373. https://doi.org/10.1016/s0021-9258(19)39777-7.

Chapter 4 - ChIP DIP Protocol

Starter Guide to Bead Barcoding

Materials

- EZ-Link Sulfo-NHS-Biotin (Thermo PIA39256)
- DMSO (Sigma-Aldrich D2650-5X5mL)
- Dynabeads Protein G (Life Technologies 10009D) – 2.7*10^6 beads/uL
- Instant Sticky-end Master Mix (NEB M0370L)
- NEBNext Quick Ligation Reaction Buffer (5X) (NEB B6058S)
- 1,2-Propanediol, ACS reagent, ≥99.5% (Sigma-Aldrich 398039-25ML)
- Q5 2x Mastermix (NEB M0494L)
- Streptavidin (BioLegend 280302)
- Terminal Ligates Even Barcode
  - Top: **/5Phos/CAAGTCA**ACCATAATAAGATCGGAAGA
  - Bottom: CAGACGTGTGCTCTTCCGATCTTATTATGGT
- 2Puni primer: 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'
- 2Pbarcoded primer: 5'-CAAGCAGAAGACGGCATACGAGATGCCTAGCCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3'
- Biotinylated and barcoded oligo of bead barcoding
  - Sequence structure
    - /5phos/TGACTTGNNNNNNNNTATTATGGTAGATCGGAAGAGCGTCGTGTACACAGAGTC/3Bio/
- Sticky end that ligates Odd barcodes
- UMI
- Oligo barcode
- Illumina primer binding site (i5 primer binding site)

Buffers:

SPRITE Instant sticky mastermix	
Solution	Volume
NEBNext Quick ligation reaction buffer	1600 uL
Instand Sticky-end Master Mix (2x)	1000 uL
1.2-Propanediol	600 uL
	3200

M2 (SPRITE wash buffer)			
	Stock conc	Final conc	Volume for 50mL solution
H2O			45.5
Tris pH 7.5	1M	20mM	1
Triton-X	10%	0.20%	1
NP-40	10%	0.20%	1
DOC	10%	0.20%	1
NaCl	5M	50mM	0.5
			50

2x SA binding buffer			
	Stock conc	Final conc	Volume for 50mL solution
H20			29.4
Tris pH 7.5	1M	10mM	0.5
NaCl	5M	2M	20
EDTA	.5M	1mM	0.1
			50

PBSt				
reagent	Stock conc,	Final conc	50mL	500mL
PBS	1x	1x	49.5mL	495mL
Tween-20	10%	0.10%	0.5mL	5mL

Dilute M2	
Reagent	1x
M2	7.5mL
Water	7.5mL

Protocol:

Biotinylate Protein G beads

1. Reconstitute NHS-biotin
   a. Allow 1mg tube of sulfo NHS-biotin to equaliberate to RT (10-30 min)
   b. Add 448µL of DMSO to the 1mg vial for a 5mM solution of NHS-biotin
   c. Vortex for 10 seconds
2. Conjugate NHS-biotin to Protein G beads
   a. Put 5mL of beads on magnetic rack

b. Remove storage buffer and put in a separate tube
c. Wash beads with 1mL PBSt
d. Resuspend beads in 1mL PBSt
e. Add 20uL of freshly reconstituted 5mM Sulfo-NHS-Biotin
f. Vortex 10 seconds **immediately**
    i. This is important to ensure all beads are equally biotinylated which is key to an even distribution of antibody ID oligo on all beads.
g. Rotate on hula mixer for 30 minutes to mix end to end
h. Put on magnetic rack, remove 500uL and add 500uL 1M Tris pH 7.4
i. Vortex 5 seconds and rotate on hula mixer for 30 minutes
j. Wash beads 2x with 1mL PBSt
k. Resuspend beads in original buffer for storage
    i. This is to keep the antibody capacity per uL of beads consistent as well as keeping the beads in buffer designed for long term stability at 4C.

Generate Streptavidin bound to antibody ID oligo (SA-oligo) complex and optimize bead labeling

1. Couple biotinylated barcoded oligo to Streptavidin at a 2:1 ratio (oligo:Streptavidin)
    a. Prepare the following reaction for each barcoded oligo

Reagent	**molarity (uM)**	**1x**	final concentration of SA (nM)	
Barcoded oligo	10	40		
Streptavidin (1mg/mL)	18.1818	10	909	5.474×10^11 molecules/uL
PBS		150		
	total	200		

    b. **Critical note:** maintaining the 2:1 ratio of oligo:streptavidin is key to effective bead labeling. If bead labeling is not robust, lowering the ratio to 1:1 can help.
    c. Shake 30 minutes at RT at 1200 RPM
    d. Make a 1:4 dilution (227nM final)
        i. Store 227nM stock at 4C. Good for two years (and counting!)
        ii. Store 909nM stock at -80C.
2. Titrate amount of SA-oligo to optimize oligo/bead
    a. **Critical note:** It is important to remove any free biotin from the NHS-biotin labeling step with these washes. Any biotin that was not washed away will negatively impact the bead labeling efficiency with SA-oligo because the biotin will quench the available SA binding pockets for biotin.
    b. Make a fresh dilution from the 227nM to make a working stock at a final concentration of 5.67nM (2:80 dilution)
        i. Wash 10uL of biotinylated Protein G beads per condition 2x with 1mL PBSt (Eg. 5 conditions would be 50uL of protein G beads)
    c. Resuspend beads in 100uL PBSt per 10uL of beads

d. Add 100uL 2x SA binding buffer per 10uL of beads

i. **Critical note:** This buffer is important for efficient binding of SA-oligo to the biotinylated protein G beads. The high salt will neutralize the charge of the single stranded antibody ID oligo and allow the SA-oligo to get close enough to the protein G for binding to the biotin molecules on the protein G.

e. Split into 200uL aliquots (10uL worth of beads) in n number of wells of a deep well plate

f. Add varying amounts of 5.67nM SA-oligo to each well

i. For example: 0uL, 5uL 10uL, 15uL and 20uL of 5.67nM SA-oligo

g. Shake on thermomixer at 1100 RPM at RT for 30 minutes

h. Wash each well 2x with 200uL M2 buffer

i. Resuspend in 200uL Dilute M2 (50:50 of M2:H2O)

3. Quantify oligos/bead

a. Take 2uL worth of beads (40uL of 200uL) of each condition and resuspend beads in 40uL of dilute M2. (Dilute M2 is 50% M2 and 50% water)

b. Prepare the following reaction for each condition

Reagent	**1x (uL)**
Beads in Dilute M2	40
Terminal Ligates Even (4.5uM)	12
Instant Sticky MM	32
total	84

c. Shake at 25C for 15 minutes, 1200RPM

d. Wash 3x with 200uL M2

e. Resuspend in 42uL water

f. Prepare the following reaction with 1/2 of sample (21uL, 1uL of beads)

Reagent	1x (uL)	10x
Beads in water	21	
2pbc (25uM)	2	
2puni (25uM)	2	
2x Q5 MM	25	

g. PCR amplify 10 cycles with the following program

Temp (C)	**Time (s)**	**Cycles**
98	60	1 cycle
98	15	4 cycles
69	15	
72	90	
98	15	6 cycles
72	90	

72	2min	

h. 1.2x SPRI
   i. Add 60uL SPRI beads
   ii. Mix and incubate 5 minutes at room temperature
   iii. Put on magnet and wait until solution clears
   iv. Wash 2x with 200uL 80% **freshly prepared** ethanol
   v. Remove any excess ethanol and allow beads to dry
      1. **Critical note:** Do not over dry, because this will prevent proper elution. Allow to dry enough such that there is no glossy sheen from the beads being wet, but not so dry that the beads begin to show cracks.
   vi. Elute by adding 20uL of water and incubate at 37C

i. Run each sample on tapestation to quantify molarity
   i. The readout will be in picomolar (pM) from the tapestation
   ii. Back calculate the number of unique molecules pre-PCR with the following formula: [(pM)(elution volume)(6.022*10^23)]/[2^((number of PCR cycles)-1)]
   iii. Divide by the number of beads, which in this case in 1uL worth of beads, or 2.7*10^6 beads. This will give you antibody ID oligos per bead
   iv. **Aim for condition that gives 50-80 oligos/bead**

ChIP DIP Protocol

Abbreviations

FA = formaldehyde
SA = streptavidin
QC = Quality control
ProK = Proteinase K
PBS = phosphate buffered solution
MM = mastermix
RT = Room temperature

High level workflow

Small scale test experiment (optional) – This is an optional step in which only a small amount of sequencing is done to understand the IP efficiency (chromatin per bead) for each antibody. There are two goals:

1. Find the amount of chromatin/bead pulled down by each antibody. This information is used to normalize the number of beads used in a large scale experiment so that each target gets roughly the same number of reads.
2. Make sure enough antibody is used to get the library complexity desired. For example, if a TF is very lowly abundant there might only be 5M unique chromatin pulled down with 10uL of antibody-Protein G beads. So for the full experiment you would add 20uL worth of antibody-protein G beads for the large scale experiment.

Workflow

- Generate oligo-labeled beads
- Biotinylate Protein G beads (Day 0)
- Make Oligo-SA complex (Day 0)
- Bind SA-oligo to biotinylated beads (Day 1)
- QC oligo per bead to ensure oligo/bead ratio is in range (80-160 oligo/bead) (Day 1)
- Lyse cells and fragment chromatin using method of choice (Day 1)
- Verify fragments are correct size before binding antibodies
- Store at 4C overnight
- Bind antibodies to oligo-labeled beads (Day 1)
- Bind overnight at 4C
- Quench beads with excess biotin and IgG-Fc (Day 2)
- Wash and pool beads (Day 2)
- Add bead pool to chromatin for IP (Day 2)
- 30 min to 1 hour at RT or 2 hours at 4C
- Wash beads
- End Repair, dA-tailing and DPM to prepare for split pool (Day 2)
- Split pool barcoding (Day 2)
- Take small aliquots and reverse crosslink and ProK overnight (Day 2)

- Column clean aliquots (Day 3)
- Library amplification (Day 3)
- Tapestation or BioAnalyzer to QC libraries (Day 3)
- Illumina sequencing (Day 4)
- Computational demultiplexing (Day 5)
- Assess chromatin/bead for each target for large scale normalization

Large scale experiment for full depth

Workflow

- Generate oligo-labeled beads
- Biotinylate Protein G beads (Day 0)
- Make Oligo-SA complex (Day 0)
- Bind SA-oligo to biotinylated beads (Day 1)
- QC oligo per bead to ensure oligo/bead ratio is in range (80-160 oligo/bead) (Day 1)
- Lyse cells and fragment chromatin using method of choice (Day 1)
- Verify fragments are correct size before binding antibodies
- Store at 4C overnight
- Bind antibodies to oligo-labeled beads (Day1)
- Bind overnight at 4C
- Quench beads with excess biotin and IgG-Fc (Day 2)
- Wash and pool beads (Day 2)
- Add bead pool to chromatin for IP (Day 2)
- 30 min to 1 hour at RT or 2 hours at 4C
- End Repair, dA-tailing and DPM to prepare for split pool (Day 2)
- Split pool barcoding (Day 2)
- Take aliquots and reverse crosslink and ProK overnight (Day 2)
- Column clean aliquots (Day 3)
- Library amplification (Day 3)
- Tapestation or BioAnalyzer to QC libraries (Day 3)
- Illumina sequencing (Day 4)
- Computational demultiplexing and alignment of chromatin from each antibody to the genome (Day 5)
- Downstream ChIP-seq analyses

Equipment

- Branson needle-tip sonicator (3mm diameter (1/8’ doublestep). Branson ultrasonics 101-148-063)
- Agilent TapeStation 4150 (or Agilent Bioanalyzer)
- Multichannel pipet p20 (12 well)
- Multichannel pipet p200 (12 well)

Materials and reagents

- 16% Formaldehyde (w/v), Methanol-Free (Thermo PI28908)
- cOmplete, EDTA-free Protease Inhibitor Cocktail (Millipore Sigma 11873580001)
- EZ-Link Sulfo-NHS-Biotin (Thermo PIA39256)
- DMSO (Sigma-Aldrich D2650-5X5mL)
- Dynabeads Protein G (Life Technologies 10009D) – 2.7*10^6 beads/uL
- Streptavidin (BioLegend 280302)
- Positive control antibodies
- H3K4me3 (CST 9751S)
- CTCF (CST 3418S)
- Q5 2x Mastermix (NEB M0494L)
- SPRI beads (Bulldog Bio CNGS500)
- 96 well PCR plate
- NEBNext End Repair Module (NEB E6050)
- NEBNext dA-tailing Module (NEB E6053)
- IgG Fc (Bio X Cell BE0096)
- Biotin (Sigma Aldrich B4639-5G)
- Proteinase K (NEB P8107S)
- Zymo DNA Clean and Concentrator kit (Zymo 4014)
- Agilent High Sensitivity D1000 ScreenTape (Agilent Technologies 5067-5584)
- Agilent High Sensitivity D1000 Reagent (Agilent Technologies 5067-5603)
- Nunc 96-Well DeepWell Plates with Shared-Wall Technology, Thermo Scientific, Color:Natural, (Thermo Fisher Scientific 260251)
- Instant Sticky-end Ligase Master Mix (ISMM) - 250, rxns (NEB M0370L)
- NEBNext Quick Ligation Reaction Buffer (5X) (NEB B6058S)
- 1,2-Propanediol, ACS reagent, ≥99.5% (Sigma-Aldrich 398039-25ML)
- Bead oligo sample sequence and structure
- /5phos/TGACTTGNNNNNNNNTATTATGGTAGATCGGAAGAGCGTCGTGTACACAGAGTC/3Bio/
  - Sticky end that ligates Odd barcodes
  - UMI
  - Oligo barcode
  - Illumina primer binding site (i5 primer binding site)
  - Spacer sequence

Buffers

Lysis buffer A			
	Stock conc	Final conc	Volume for 10mL solution
HEPES 7.4	1M	50mM	0.5
EDTA	.5M	1mM	0.02

EGTA	.5M	1mM	0.02
NaCl	5M	140mM	0.28
Triton-X	10%	0.25%	0.25
NP-40	10%	0.50%	0.5
Glycerol	50%	10%	2
H2O			6.43
			10mL
Lysis buffer B			
	Stock conc	Final conc	Volume for 10mL solution
HEPES 7.4	1M	50mM	0.5
EDTA	.5M	1.5mM	0.03
EGTA	.5M	1.5mM	0.03
NaCl	5M	200mM	0.4
H2O			9.04
			10mL
Lysis buffer C			
	Stock conc	Final conc	Volume for 10mL solution
HEPES 7.4	1M	50mM	0.5
EDTA	.5M	1.5mM	0.03
EGTA	.5M	1.5mM	0.03
NaCl	5M	100mM	0.2
DOC	10%	0.10%	0.1
NLS	20%	0.50%	0.25
H2O			8.89
			10mL

SPRITE Instant sticky mastermix	
Solution	Volume
NEBNext Quick ligation reaction buffer	1600 uL
Instand Sticky-end Master Mix (2x)	1000 uL
1.2-Propanediol	600 uL
	3200

MyRNK buffer				
	Stock conc	Final conc	Volume for 10mL solution	

Tris pH 7.5	1M	20mM	0.2	1
NaCl	5M	100mM	0.2	1
EDTA	0.5	10mM	0.2	1
EDTA	0.5	10mM	0.2	1
Triton-X	10%	0.50%	0.5	2.5
SDS	10%	0.2%	0.2	1
H2O			8.7	43.5
			10mL	50mL

IP Wash buffer 1 (IPWB1)

	Stock conc,	Final conc	Volume for 30mL solution	50mL		500mL
Tris-HCl pH8.0	1M	20mM	0.6	1	4.77	10
SDS	10%	0.05%	0.15	0.25		2.5
Triton X-100	10%	1%	3	5		50
EDTA	.5M	2mM	0.12	0.2		2
NaCl	5M	150mM	0.9	1.5		15
H2O			25.23	42.05		420.5
			30	50		500

IP Wash buffer 2 (IPWB2)

	Stock conc,	Final conc	Volume for 30mL solution	50mL		500mL
Tris-HCl pH8.0	1M	20mM	0.6	1	6.87	10
SDS	10%	0.05%	0.15	0.25		2.5
Triton X-100	10%	1%	3	5		50
EDTA	.5M	2mM	0.12	0.2		2
NaCl	5M	500mM	3	5		50
H2O			23.13	38.55		385.5
			30	50		500

M2 (SPRITE wash buffer)

	Stock conc	Final conc	Volume for 50mL solution
H2O			45.5
Tris pH 7.5	1M	20mM	1

Triton-X	10%	0.20%	1
NP-40	10%	0.20%	1
DOC	10%	0.20%	1
NaCl	5M	50mM	0.5
			50

2x SA binding buffer			
	Stock conc	Final conc	Volume for 50mL solution
H20			29.4
Tris pH 7.5	1M	10mM	0.5
NaCl	5M	2M	20
EDTA	.5M	1mM	0.1
			50

M2 + 75mM EDTA			
	Stock conc	Final conc	Volume for 50mL solution
H20			38
Tris pH 7.5	1M	20mM	1
Triton-X	10%	0.20%	1
NP-40	10%	0.20%	1
DOC	10%	0.20%	1
NaCl	5M	50mM	0.5
EDTA	.5M	75mM	7.5

PBSt				
reagent	Stock conc,	Final conc	50mL	500mL
PBS	1x	1x	49.5mL	495mL
Tween-20	10%	0.10%	0.5mL	5mL

Dilute M2	
Reagent	1x
M2	7.5mL
Water	7.5mL

Protocol

[Day 0]

Crosslinking

1. Preparation work
    a. Chill one bottle of 1x PBS on ice. Keep one at room temperature.
2. Count and wash cells
    a. Count an aliquot of the cells to determine the total number of cells to crosslink. Prior to counting cells, make sure the cells are in a single cell suspension by pipetting vigorously several times.
        i. total cells:
    b. Pellet cells at room temp for 3 minutes at 330xG. Remove media.
    c. Wash 1x by resuspending in 4mL PBS per 10M cells and pellet cells at room temp for 3 minutes at 330xG. Remove PBS solution.
3. 1% FA crosslinking
    a. NOTE: Right before using, premix a 1% formaldehyde (Thermo cat no. 28908) solution in PBS. Use a **fresh** ampule of 16% formaldehyde. Do not use an ampule opened more than ~20 min before crosslinking the cells.

	1% FA	
	1x (10M cells)	10x (100M cell EXAMPLE)
16% FA	250 uL	2500 uL
1x PBS	3750 uL	37500 uL
	4000 uL	40000 uL

    b. Prepare a 50mL conical centrifuge tube with 3.75 mL PBS for each 10M cells to be crosslinked.
    c. Use 1mL of PBS solution prepared in 3.a. to resuspend cell pellet
    d. Crack open 16% FA ampule and transfer to a labeled 15mL centrifuge tube
    e. Add 250uL 16% FA solution per 4 mL PBS solution from 3.a. to get a ~1% FA solution. Invert to mix
    f. Add resuspended cell pellet to the ~1% FA tube to make a tube of 1% final concentration FA in PBS
    g. Rock gently at room temperature for **exactly** 10 minutes
    h. Immediately add 200uL of 2.5M glycine stop solution per 1mL of 1% formaldehyde solution directly to the tube to quench the crosslinker. Mix well.
        i. Glycine calculation:
    i. Rock gently at room temperature for 5 minutes.
4. Wash and aliquot crosslinked cells
    a. Spin down cells at 4 °C for 4 minutes at 1000xG. Discard FA solution in the FA liquid waste in the fume hood. From here onward, keep cells on ice.

b. Resuspend pellet in cold 1x PBS (4mL PBS per 10M cells)
c. Spin down cells at 4 °C for 4 minutes at 1000xG. Discard solution in FA waste.
d. Repeat the wash step.
e. Resuspend cells in 200uL of PBS per 10M cells.
f. Aliquot cells into microcentrifuge tubes (5M-20M cells each) and spin at 4 °C for 5 minutes at 2000xG. Remove supernatant carefully without disrupting the pellet.
g. **Critical note:** Completely removing the supernatant is important to prevent lysis of cell upon flash freezing for long term storage.

5. Flash freeze in liquid nitrogen and store at -80 or proceed to cell lysis. **(Safe stop if flash freezing and storing at -80C)**

<u>Day 1</u>

Prepare chromatin

6. Cell lysis (for up to a 50M cell aliquot)
    a. Resuspend in 700uL Lysis Buffer A with 14uL 50x Roche C0mplete Protease Inhibitor Cocktail (PIC)
    b. Incubate 10 min on ice
    c. Spin down at 4C for 8 min at 850G
    d. Remove supernatant
    e. resuspend each sample in 700uL Lysis Buffer B with 14uL 50x PIC
    f. Incubate 10 min on ice
    g. Spin down at 4C for 8 min at 850G
    h. Remove supernatant
    i. Resuspend each sample in 550uL Lysis buffer C with 22uL 50x PIC
    j. Incubate 10 min on ice
7. Sonicate
    a. **NOTE: sonication times and power will vary depending on cell type and crosslinking condition and will need to be empirically optimized. These conditions have been optimized for mouse embryonic stem cells, HEK293T cells, and K562 cells.**
    b. Sonicate each aliquot with 20% power (3-4 watts) at 4C for 2.5 min (.7sec on, 3.3sec off)
    c. Check size distribution of chromatin:
        i. Reverse crosslink 10uL-20uL at 80C for 30 minutes. This can be done at a small volume in a thermocycler.
        ii. Clean up with Zymo IC column DNA purification kit
            1. Add 6x DNA binding buffer
            2. Add this to a IC column and spin at 10,000g for 30 seconds
            3. Dump flowthrough

4. Wash 2x with 200uL DNA/RNA wash buffer by spinning at 10,000g for 30 seconds
5. Dump flowthrough
6. Do an empty spin for 1 minute
7. Transfer column to a clean 1.7mL tube
8. Add 20uL H2O to elute
9. Spin at 10,000g for 30 seconds to collect purified DNA

iii. Run half of the purified DNA on a 1% gel to check the chromatin fragment sizes (desired fragment size in 200bp-500bp)

iv. If the size range is 200bp-500bp for most fragments, save this as your input. If not, go back to sonicated sample and sonicate more. Repeat once if needed.

v. **Critical point: Take 20uL input and reverse crosslink overnight**

Bind unique oligos to Protein G beads

8. Aliquot 10uL of biotinylated Protein G beads per IP
9. Wash biotinylated Protein G twice with 1mL PBSt
10. quick spin and put on magnet to remove excess PBSt
11. Resuspend in 100uL PBSt per 10uL of beads
12. Add 100uL of 2x SA binding buffer per 10uL beads. This is the bead master mix (MM).
13. Split bead MM into deep well plate, 200uL per well.
14. Add XuL (optimized volume from bead labeling titration) of 5.67nM SA-PC50 to corresponding wells
15. Shake at RT for 30 minutes
16. Quick spin for 30sec at 100G
17. Wash 2x with 200uL M2
18. wash 2x with 200uL PBSt
19. Keep on ice while prepping chromatin

QC oligos/bead from bead labeling reaction

20. Take 2uL worth of beads (40uL of 200uL) from the last well of each row for representative assessment of coupling for that row
21. Resuspend beads in 40uL of dilute M2. (Dilute M2 is 50% M2 and 50% water)
22. Prepare the following reaction for each condition

Reagent	**1x (uL)**
Beads in Dilute M2	40
Terminal Ligates Even (4.5uM)	12
Instant Sticky MM	32
total	84

23. Shake at 25C for 15 minutes, 1200RPM
24. Wash 3x with 200uL M2

25. Resuspend in 42uL water
26. Take 1/2 of sample (21uL, 1uL of beads) and prepare the following reaction

Reagent	1x (uL)	10x
Beads in water	21	
2pbc (25uM)	2	
2puni (25uM)	2	
2x Q5 MM	25	

27. PCR amplify 10 cycles using the following program:

Temp (C)	**Time (s)**	**Cycles**
98	60	1 cycle
98	15	4 cycles
69	15	
72	90	
98	15	6 cycles
72	90	
72	2min	

28. 1.2x SPRI
    a. Add 60uL SPRI beads
    b. Mix and incubate 5 minutes at room tempuratre
    c. Put on magnet and wait until solution clears
    d. Wash 2x with 200uL **freshly prepared** 80% ethanol
    e. Remove any excess ethanol and allow beads to dry
        i. **Critical note:** Do not over dry, because this will prevent proper elution. Allow to dry enough such that there is no glossy sheen from the beads being wet, but not so dry that the beads begin to show cracks.
    f. Elute by adding 20uL of water and incubate at 37C
29. Run each sample on Tapestation to quantify molarity
    a. The readout will be in picomolar (pM) from the tapestation
    b. Back calculate the number of unique molecules pre-PCR with the following formula: [(pM)(elution volume)(6.022*10^23)]/[2^((number of PCR cycles)-1)]
    c. Divide by the number of beads, which in this case in 1uL worth of beads, or 2.7*10^6 beads. This will give you antibody ID oligos per bead
30. Record the calculated number of antibody ID oligos per bead

Bind Antibodies to Labeled Protein G Beads

31. Resuspend beads in 200uL PBSt
32. Add 2.7ug of antibody to each well
33. Shake at 4C overnight
    a. 10 seconds shaking at 1200 RPM, every 15 minutes

Day 2

34. **Critical note:** It is important to follow the reaction times below. Allowing reactions to proceed for longer than recommended can cause a lower yield in library complexity.
35. 1hr RT IP
36. Shake beads at RT at 1100rpm while preparing reagents and lysate
37. Prepare lysate
    a. Add 500uL PBSt to each tube
    b. 30uL 50x PIC
    c. add XuL of IgG Fc (2.5ug per 10uL beads)
    d. Add XuL 1M biotin (200uM biotin final concentration)
38. Wash beads 2x with 200uL PBSt
39. Resuspend in 200uL PBSt + 200uM biotin + 2.5ug Human IgG Fc per well
    a. This quenches any free protein G and SA to prevent mixing of labels or antibody
40. Shake at RT for 15 minutes

	1X	48X
PBSt	200	9600
Biotin (100mM)	0.8	38.4
Human IgG (8.6ug/uL)	0.3	14.4

41. Pool all beads with volumes outlined in excel sheet to equalize coverage
    a. Normalize number of beads to evenly distribute sequencing coverage for each antibody
        i. **Normalize to the beads with the antibody that pulls down the least amount of chromatin. For example, if antibody A pulls down 100chrom/bead and antibody B pulls down 10chrom/bead then there would be 10x less beads from antibody A compared to antibody B added to the bead pool. This can be skipped if the chromatin per bead for each antibody in unknown or if pooling antibodies of similar IP efficiency.**
42. Wash pooled beads 2x with 1mL PBSt+200uM biotin
43. resuspend beads in 200uL PBSt
44. Add beads to prepared lysate
45. rotate on hula mixer at RT for 1 hour or 2 hours at 4C

Wash beads post IP

46. Wash beads 2x with 1mL IPWB1
    a. Put on magnet and remove supernatant between washes
47. Wash beads 2x with 1mL IPWB2
    a. Put on magnet and remove supernatant between washes
48. Wash beads 2x with 1mL M2
    a. Put on magnet and remove supernatant between washes
49. quick spin, put tube on magnet and remove all excess M2

End Repair and dA-tailing

50. **Critical Note:** It is important to not allow beads to dry out. Prepare mastermixes ahead of time before removing all buffer from beads.

End Repair

51. prepare the following reaction:

Stock Solution	1x
H2O	211.5
100mM biotin	1
End Repair Reaction Buffer (10X)	25
End Repair Enzyme Mix	12.5
Total	250

52. Add 1ug human IgG1 Fc for every 10uL of beads
53. Shake at 20C for 15 min, 1400RPM
54. Quench with 500uL PBSt+100mM EDTA
55. Wash 2x with 1mL M2
    a. Put on magnet and remove supernatant between washes
56. Quick spin and put on magnet, remove all residual solution
57. dA-tailing

dA-tailing

58. Prepare the following reaction

Stock Solution	1x
H2O	214
100mM biotin	1
dA-Tailing Reaction Buffer (10X)	25
Klenow Fragment (exo-)	10
Total	250

59. Add 1ug human IgG1 Fc for every 10uL of beads
60. Shake at 37C for 15 min, 1400RPM
61. Quench with 500uL PBSt+100mM EDTA
62. Wash 2x with 1mL M2
    a. Put on magnet and remove supernatant between washes
63. Quick spin and put on magnet, remove all residual solution

Ligate DPM to chromatin

Reagent	1x (per well)	
Beads in dilute M2+200uM biotin + Protein G	11.2	
DPM (4.5uM)	2.4	
ISMM	6.4	
	20	

64. Resuspend each aliquot of beads in 135uL of dilute M2 + 200uM biotin + Protein G
65. Distribute into 24 wells
66. Add unique DPM to each well
67. Add ISMM with multichannel pipet in volume outlined on reaction table. Pipet up and down to mix

68. Shake at 20C for 20 minutes
69. Quench with 30uL PBSt+100mM EDTA
70. Pool all beads into a single tube
71. Wash 2x with 1mL M2
    a. Quickly vortex to resuspend beads in wash buffer
    b. Put on magnet and remove supernatant
72. Quick spin and put on magnet, remove all residual wash buffer

Split pool

73. 6 rounds of 24 well barcoding (barcode combinations)
    a. Odd (Rows A, B)
    b. Even (Rows A, F)
    c. Odd (Rows C, D)
    d. Even (Rows C, D)
    e. Odd (Rows E, F)
    f. Terminal Ligates Odd (Rows F, G, A)
        i. **Critical note**: It is important to use these rows of terminal ligates Odd/Even because they have the full stagger. This will give better sequencing yield and avoid a monotemplate issue when sequencing.

Stock solution	Volume
beads + dilute M2 + 200uM biotin + Protein G	11.2
Split pool tag (4.5uM)	2.4
Ligation master mix	6.4
total	20

74. Resuspend beads in 270uL dilute M2 + 200uM biotin
75. Distribute into 24 wells
76. Prepare the outlined reaction for each well
    a. Add barcode with multichannel pipet in volume outlined on reaction table. Do not pipet up and down. This will cause beads to stick pipet tip and cause a loss is yield in the final libraries.
    b. Add ISMM with multichannel pipet in volume outlined on reaction table. Pipet up and down to mix
77. incubate at RT for 8-10 minutes
    a. **Critical note:** It is important to minimize the time it takes to perform split pool. This is because the antibody, and by extension the chromatin, is held to the bead via noncovalent interactions. Excess time will lead to shedding of antibody-protein-chromatin off the bead and decrease the yield of chromatin in the final libraries. 8-10 minutes of ligation time is the optimal time for allow for all split pool tags to ligate to the constituent molecules on each bead, but not so long that antibody-protein-chromatin complexes start to excessively shed from the beads.
78. Quench with 30uL PBSt+100mM EDTA
79. Pool into 2mL tube and place on magnet
80. Wash 2x with 1mL M2 using magnet

a. Quickly vortex to resuspend beads in wash buffer
b. Put on magnet and remove supernatant between washes

81. Quick spin and put on magnet, remove all residual solution
82. Repeat 2-9 5 more times
83. wash 2x extra 1mL washes before resuspending in MyRNK
84. resuspend in 1mL MyRNK
85. Take aliquots and reverse crosslink overnight
    a. **Critical note:** Use tubes that are sealed tightly. Some low quality plastics will allow for evaporation. 2mL LoBind Eppendorf tubes tend to have an excellent seal.
    b. 92uL of MyRNK (this includes the aliquot) + 8uL of Proteinase K
    c. Example of a 1% aliquot: 10 uL aliquot in MyRNK + 82 mL MyRNK + 8 uL Proteinase K
    d. **Critical note:** The size of the aliquot is important. When doing split pool based techniques, most molecules in an aliquot must be sampled by sequencing. This is because in order detect to most molecules bound to each bead each molecule need to be sampled by sequencing. If most molecules are not sampled, then you may and up with a cluster (bead) that has only oligo, but no chromatin, making that data useless. To avoid this, we generally take an aliquot with 5 million to 20 million beads. For example, assuming there are 10 molecules per bead, an aliquot with 5M beads would require 50M reads.
86. 50C for 2 hours, 65C overnight

<u>Day 3</u>

Clean up samples

87. Put beads on magnet and transfer supernatant to a new LoBind tube, this contains your IP'd chromatin fragments
88. Add 700uL of DNA binding buffer to sample and add to Zymo IC column
    a. Spin at 10,000g for 30 seconds, dump flow through
    b. Wash 2x with 200uL RNA/DNA wash buffer
        i. 20 sec spin at >10,000 RPM
        ii. dump flowthrough
    c. Empty spin of column in empty collection tube
    d. Put column in a new, labeled 1.5mL LoBind tube
    e. Add 21uL H2O to elute from column, incubate at RT for 3-5 minutes
    f. Spin down 60 sec at > 10,000 RPM
    g. This eluate contains purified DNA fragments

Amplify samples to add sequencing index barcode

89. Prepare the following reaction to amplify **all** of the aliquot

Reagent	1x (uL)
Purified DNA	21
I5 (12.5uM)	3
I7 (12.5uM)	3
2x Q5 MM	27
Total	54

90. Amplify with the following PCR program

Temp (C)	Time (s)	Cycles
98	60	1 cycle
98	15	4 cycles
91. 69	92. 15	
72	90	
98	15	6 cycles
72	90	
72	2min	

93. **Critical note:** Amplify 1-2 aliquots first at 10 cycles. 10 cycles is the recommendation, but if the yield is high there can by overamplification at 10 cycles. Adjust as needed with remaining aliquots. The target molarity after PCR is ~10-16nM
94. Clean up with 1.2x SPRI
    a. Add 60uL SPRI beads and mix with pipet
    b. Incubate at RT for 10 minutes
    c. Wash beads 2x with 200uL **freshly prepared** 80% EtOH
    d. Quick spin and put on magnet to remove residual EtOH
        i. **Critical note:** Do not over dry, because this will prevent proper elution. Allow to dry enough such that there is no glossy sheen from the beads being wet, but not so dry that the beads begin to show cracks.
    e. Add 20uL H2O to elute
    f. Incubate at 37C for 5 minutes
    g. Place beads on a magnet and allow solution the clear.
    h. Transfer supernatant away from SPRI beads into a fresh tube
    i. This contains the amplified library
95. Proceed to Tapestation analysis or freeze for later
    a. Quantify library and number of unique molecules per aliquot with tapestation
        i. Run Tapestation or Bio Analyzer, following manufacturers protocol for each aliquot
        ii. The readout will be in picomolar (pM) from the tapestation
        iii. Back calculate the number of unique molecules pre-PCR with the following formula: [(pM)(elution volume)(6.022*10^23)]/[2^((number of PCR cycles)-1)]
96. Gel extract oligo and chromatin to remove PCR primers.
    a. **Optional:** gel cut chromatin and oligo separately. This is helpful if you want to sequence more chromatin than oligo or vice versa.
97. Sequence on your favorite NGS platform!
98. Analysis pipeline and all details on how to run the pipeline can be found here: https://github.com/GuttmanLab/chipdip-pipeline
    a. Documentation and software is maintain on the Github repository and the link can also be found on the Guttman Lab Webpage.

Future directions

With the core technology of bead barcoding and ChIP-DIP validated and robust, we can now accurately map hundreds of proteins in a single experiment and there is a vast array of questions that can be addressed quickly using this technology. For example, we can quickly profile the protein-DNA and protein-RNA landscape in any disease model, such as cancer or neurogenerative disease, to understand how changes in the protein-DNA/RNA landscape are driving these diseases. Another option would be to perturb healthy cells to understand fundamental biology of gene regulation. For example, one could inhibit various chromatin regulators to understand how chromatin regulators influence to localization of proteins to DNA and RNA. We aim to leverage ChIP-DIP and SPIDR to make landmark observations about gene regulation biology.

Our lab has pioneered the idea that active transcription of RNA is required for nuclear organization and is key to the proper folding and 3D architecture of the genome[1]. Beyond this, we have shown that active transcription is important for the localization of chromatin-associated proteins like HP1-beta[1]. Others have also shown that when transcription is inhibited, large foci of splicing proteins become aberrantly localized when observed with fluorescent microscopy[2]. While it is clear that protein localization can change dramatically upon transcriptional inhibition, it remains unclear how these gross localization changes translate to changes in protein-DNA interactions.

We plan to approach this problem with ChIP-DIP by screening the DNA localization of hundreds of DNA-associated protein upon transcriptional inhibition. This strategy has several advantages to traditional ChIP-Seq. First, and most obviously, dozens to hundreds of protein-DNA interactions can be profiled simultaneously, dramatically reducing the workload associated with inhibiting transcription in 80 million cells per target. Utilizing ChIP-DIP, we can inhibit transcription in 80 million cells to be used to explore the DNA interactions of hundreds of different proteins. We expect one of two different results. First, it is possible that active transcription is required for the **proper** localization of protein; that is, we would observe improper and/or expanded localization of DNA-associated proteins upon inhibition. The second possibility is that active transcription is required for chromatin-associated proteins to find their targets. In this scenario, we would expect to see a dramatic loss of signal compared to cells that did not have transcription inhibited. This would show how RNA guides the localization of different proteins to their chromatin targets.

To inhibit transcription, we could use any of several small molecules that inhibit RNA polymerases to varying specificity. For example, flavopiridol (FVP) is a robust cyclin-dependent kinase 9 (CDK9) inhibitor. CDK9 phosphorylates the largest subunit of RNA polymerase II (RNAPII), POLR2A to initiate the transition from paused RNAPII to actively transcribing RNA RNAPII. By inhibiting CDK9, RNAPII is unable to initiate

transcriptional elongation, effectively inhibiting the production of RNAPII transcripts[3]. In addition to FVP, there are several nonspecific RNA polymerase inhibitors, such as actinomycin D, which form a highly stable complex with DNA, preventing the DNA from unwinding[4]. Since this DNA cannot unwind, RNAPII cannot transcribe. We plan to start our exploration of RNA dependence of nuclear organization by performing ChIP

DIP on cells treated with FVP using a panel of 160 previously validated antibodies to profile ~120 different chromatin-associated proteins to evaluate the role of transcription in the localization of chromatin-associated proteins.

Importantly, we have also demonstrated the ability of bead barcoding to multiplex CLIP to perform high-throughput RNA-protein mapping with SPIDR. We anticipate that bead barcoding will expand the multiplexing capabilities of other IP based assays, such as SPRITE IP[5]. Beyond immunoprecipitation assays, bead barcoding can be used to multiplex other bead-based assays. For example, bead barcoding could be used to multiplex DNA-encoded library (DEL)[6] based drug screening assays by allowing drug discovery of hundreds of proteins, and mutants of proteins, in the same experiment.

References

1. Quinodoz, S.A., Jachowicz, J.W., Bhat, P., Ollikainen, N., Banerjee, A.K., Goronzy, I.N., Blanco, M.R., Chovanec, P., Chow, A., Markaki, Y., et al. (2021). RNA promotes the formation of spatial compartments in the nucleus. Cell *184*, 5775–5790.e30. 10.1016/j.cell.2021.10.014.

2. Yasuhara, T., Xing, Y.-H., Bauer, N.C., Lee, L., Dong, R., Yadav, T., Soberman, R.J., Rivera, M.N., and Zou, L. (2022). Condensates induced by transcription inhibition localize active chromatin to nucleoli. Mol. Cell *82*, 2738–2753.e6. 10.1016/j.molcel.2022.05.010.

3. Balakrishnan, K., Wierda, W.G., Keating, M.J., and Gandhi, V. (2005). Mechanisms of Cell Death of Chronic Lymphocytic Leukemia Lymphocytes by RNA-Directed Agent, 8- NH2-Adenosine. Clin. Cancer Res. *11*, 6745–6752. 10.1158/1078-0432.ccr-05-0553.

4. Lai, W.S., Arvola, R.M., Goldstrohm, A.C., and Blackshear, P.J. (2019). Inhibiting transcription in cultured metazoan cells with actinomycin D to monitor mRNA turnover. Methods *155*, 77–87. 10.1016/j.ymeth.2019.01.003.

5. Vangala, P., Murphy, R., Quinodoz, S.A., Gellatly, K., McDonel, P., Guttman, M., and Garber, M. (2020). High-Resolution Mapping of Multiway Enhancer-Promoter Interactions Regulating Pathogen Detection. Mol. Cell *80*, 359–373.e8. 10.1016/j.molcel.2020.09.005.

6. Satz, A.L., Brunschweiger, A., Flanagan, M.E., Gloger, A., Hansen, N.J.V., Kuai, L., Kunig, V.B.K., Lu, X., Madsen, D., Marcaurelle, L.A., et al. (2022). DNA-encoded chemical libraries. Nat. Rev. Methods Prim. *2*, 3. 10.1038/s43586-021-00084-5.

Limitations

The level of multiplexing is mainly limited by the number of different barcoded oligos and antibodies that are available. However, because the oligo barcoding of beads is so robust, one can use different combinations of bead oligos to denote a unique bead barcode. There is also a start-up cost of doing split-pool-based assays that requires the investment of purchasing the split-pool barcodes, thermomixers, and multichannel pipets. Additionally, we have found that while some antibodies are validated with ChIP-Seq or CLIP, that same antibody may not provide a robust protein-DNA/RNA interaction map with ChIP-DIP or SPIDR. Lastly, because ChIP-DIP and SPIDR rely on the barcode string on the bead oligo to assign chromatin or RNA, respectively, to a specific antibody, there are additional sequencing costs associated with sequencing the bead oligos that provide no additional biological information. We have mitigated this by selecting beads that have a large diameter, increasing the chromatin to oligo ratio required to assign chromatin to a protein. Additionally, the cost of sequencing has dramatically dropped in the past 5 years and continues to drop as newer and more high-throughput sequencing technologies are developed. This makes technologies like ChIP-DIP and SPIDR more and more accessible every day.