THE UNIVERSITY OF CHICAGO

QUANTITATIVE STUDIES OF HISTONE MODIFICATIONS: METHODS AND APPLICATIONS OF CHROMATIN IMMUNOPRECIPITATION AND NEXT-GENERATION SEQUENCING

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON GENETICS, GENOMICS, AND SYSTEMS BIOLOGY

BY

ROHAN NISHANT SHAH

CHICAGO, ILLINOIS

JUNE 2022

To my family.

If you cannot — in the long run — tell everyone what you have been doing, your doing has been worthless.

— Erwin Schrödinger, *Science and Humanism*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SUPPLEMENTARY NOTES

ACKNOWLEDGEMENTS

There are many people without whom this work would not have been possible. First, I'd like to thank the Ruthenburg Lab and the entire research community of the 8th floor of Cummings for their frequent help and camaraderie. In particular, I'd like to thank my co-authors with whom I directly worked during graduate school: Jimmy Elias and Bill Richter, Ph.D.'20. I also gratefully acknowledge the helpful feedback and insight that I received from the members of my thesis committee: Mike Rust, Ph.D.; Ed Munro, Ph.D.; and Heng-Chi Lee, Ph.D.

I'd especially like to thank Adrian Grzybowski, Ph.D.'18, who served as my mentor in the lab from my first year of undergrad through my fourth year and, even from afar, has frequently gone out of his way to provide me with any help or support I need in my endeavors. He often took great pains to help me learn to design experiments or perform new techniques – and often at the expense of time spent on his own graduate work. I know that without his help, I would not have been able to progress in the lab to nearly the extent that I have.

And of course, I owe a huge debt of gratitude to my advisor and my mentor of seven years: Alex Ruthenburg, Ph.D. Whether it was in undergrad, medical school, or graduate school, Alex has supported my development as a scientist and the pursuit of my career goals at every opportunity. He's offered me countless hours of advice whenever I needed it while also giving me the independence to pursue my goals and grow as a scientist in my own right. This work would not have been possible without his help, and I am incredibly grateful for his guidance.

And finally, I would like to thank my friends and my family for their consistent and loving support. Their care has been critical at every stage of my training, and it is only with their help that the challenges of the last several years became more surmountable.

ABSTRACT

In the eukaryotic genome, DNA is organized into nucleosomes, which are comprised by 147bp of DNA wrapped around a core of eight histone proteins. These histones can then be post-translationally modified to regulate the function and status of the associated genomic region. The primary method to determine the genomic distribution of these modifications is chromatin immunoprecipitation (ChIP). However, this method, as traditionally practiced, has many problems hampering its interpretability insofar as it is non-quantitative and has indeterminate specificity. Internally calibrated ChIP (ICeChIP) can address some of these issues by employing nucleosomal internal standards, but many open questions remain as to the specificity of commercially available antibodies and many paradigms which are less easily resolved by traditional native ICeChIP. Here, I present my work on methods to extend the use cases of ICeChIP and applications therein. First, I show that many commercially available antibodies against H3K4 methylation states are of low quality and that common methods of antibody validation fail to reflect performance in ChIP, ultimately showing that this low specificity contributed to incorrect biological conclusions in several high-profile studies. Second, I describe our work on the study of bivalency, in which we developed a sequential form of ICeChIP to study nucleosomes bearing both H3K4me3 and H3K27me3, ultimately showing that many paradigms concerning such a combination are incorrect. Third, I describe our development of denaturative ICeChIP and use it to study the role of H3K79me2 in MLL-rearranged leukemias. Finally, I discuss our development of SmartMap, a tool to allocate next-generation sequencing reads that align ambiguously to the genome, demonstrate its ability to improve read depth at regions with low-mappability, and use it to study the role of histone modifications at repetitive elements. Overall, this work shows the power of using specific and quantitative methods in studying histone modifications.

CHAPTER 1: INTRODUCTION

Epigenetics and Chromatin

By the early twentieth century, it was widely accepted that many traits of an organism could be inherited in units now referred to as genes[1,2]. Though initial hypotheses focused on the role of proteins in this process, over the next several decades, researchers accumulated evidence instead pointing to DNA as the carrier of genetic information[3,4]. By the early 1950s, it had become clear that DNA was the primary heritable genetic material and that, in this role, it served to encode the instructions governing the development and activities each cell and the organism as a whole. What remained unclear was how these instructions were carried out in the cellular context.

This question posed a particular challenge for the field of developmental biology. Complex multicellular organisms all have multiple organ systems, each with numerous different cell types with distinct properties, activities, and functions. And yet, these cells all derive from a single fertilized egg, which must then grow, divide, and transform into the myriad cell types of the body. This process, referred to differentiation, was famously analogized by Conrad Waddington as a marble rolling down a contoured terrain which he called the "epigenetic landscape" (Figure 1.1) established by the interactions and effects of different genes on development[5]. As he described in his book *The Strategy of the Genes*:

> Consider a more or less flat, or rather undulating, surface which is tilted so that points representing later states are lower than those representing earlier ones… Then if something, such as a ball, were placed on the surface it would run down towards some final end state at the bottom edge… say, to the eye, and another to the brain, a third to the spinal cord, and so on for each type of tissue or organ. … Since each gene must be regarded as a distinct chemical entity, the path of development as it is observed by the anatomist must be viewed as the resultant of all the very numerous processes in which these genes are involved in the cells concerned.[5]

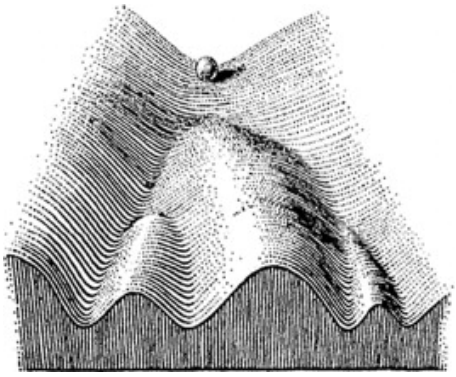


Figure 1.1: Waddington's Epigenetic Landscape.
Conrad H. Waddington's characterization of the pathway of cellular differentiation being akin to a marble rolling down a surface with multiple branch points and final states, representative of different cell types. Adapted from Waddington[5].

As appealing as this model of gene-directed development was, it still left the question: how could the single set of genetic instructions encoded in the DNA of the fertilized egg specifically direct the development and functions of a vast array of different cell types? One hypothesis was that over the course of differentiation, cells would lose portions of the genome that were no longer relevant for that cell type such that the terminally differentiated cells only contained those genes that were necessary for the functions of that cell type[1]. This model was ultimately put to rest in 1970, when Laskey and Gurdon showed that the nucleus of a terminally differentiated somatic cell could drive embryogenesis in an enucleated ovum lacking a genome, thereby showing that even adult cells carried the full genome in their nuclei[6]. This left an alternate hypothesis as the prevailing model: rather than modifying the sequence of the genome, differentiation proceeded by modifying some yet-unspecified regulatory mechanisms that governed the activity and expression of each gene. The study of these regulatory mechanisms would go on to become the central question underlying the field of epigenetics.

Over the course of the last several decades, that term has represented several different concepts. Waddington, for example, defined epigenetics as the study of how genotype and environ-

mental factors impacted developmental phenotypes[5]. The developmental biologist Adrian Bird, by contrast, defined epigenetics much more broadly as the study of the structural changes applied to the genome in order to modulate its activity[7]. As a working definition, we will use the formulation put forth by Riggs and colleagues, which strikes a useful conceptual balance between the restrictiveness of some definitions and the breadth of others. Per their conceptualization, epigenetics refers to the study of heritable traits that are not encoded by modifications to the DNA sequence or the proper pairing of corresponding nucleotides[8]. Rather, epigenetics concerns itself with the other regulatory mechanisms by which genes can be regulated and coordinated. Critically, these traits are not simply transient fluctuations of activity in response to temporary environmental stimuli; rather, this regulation occurs in a manner that can be inherited across cellular divisions or even organismal generations. Over the course of the last several decades, epigenetics has been critical for furthering our understanding of both physiological and pathophysiological processes, including development and differentiation[9–15], regulation of transcription[16–20], and cancer biology[21–25]. The mechanisms implicated in these pathways represent different ways by which cells can interpret a single genome containing all the genes of the organism so as to identify and selectively express the genes that are needed for each of the different cell types in the body.

In the nucleus, DNA is organized in complex with histone proteins into nucleosomes[26–28], which in turn interact with proteins and RNA molecules that associate closely with the genome; these complexes are collectively referred to as chromatin[29], which ultimately associates into flexible and dynamic higher-order strucutres, enabling compaction of the genome[30]. Epigenetic regulatory pathways, in turn, largely fall into one of three categories[31], each primarily involving one of these three components of chromatin: DNA modifications, involving modifications to the DNA that do not affect the DNA sequence; non-coding RNA (ncRNA) regulation, in which RNA molecules that

do not code for proteins regulate the transcription and translation of other genes; and nucleosome variants and modifications, which regulate the most fundamental units of chromatin and DNA organization (Figure 1.2).

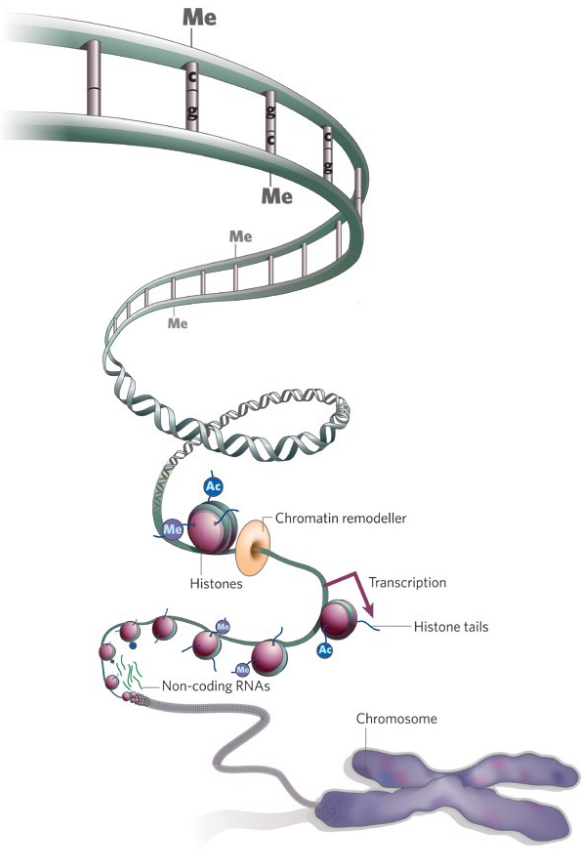
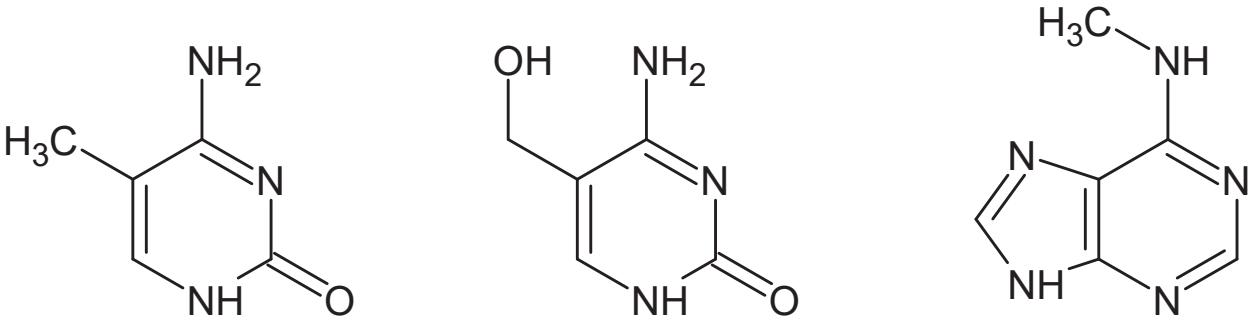


Figure 1.2: Forms of Epigenetic Regulation.
Different forms of epigenetic regulation at every level of chromatin organization, including DNA modifications (depicted as cytosine methylation) to histone modifications and non-coding RNA interactions. Adapted from Jones et al.[31].

The first of these levels of regulation is the language of covalent DNA modifications. In this context, we refer not to modifications that affect the identity of a DNA base or the canonical Watson-Crick base-pairing of a given sequence; we refer to those changes as mutations. Rather, the epigenetic DNA modifications represent covalent modifications to the portions of the DNA bases that do not participate in base pairing (Figure 1.3). The best-characterized and most prominent such

modification is the methylation of carbon 5 of cytosine (5-methylcytosine, abbreviated 5-mC)[32]. This modification, installed in humans by the DNA methyltransferases DNMT1, DNMT3A, and DNMT3B[33], was first described in the 1970s as a repressor of transcription that could be inherited through semiconservative replication[34–37] as part of its role in X chromosome inactivation[37]. Since then, DNA methylation has been found to be a critical repressor of transcription more broadly, with roles including terminal silencing of unnecessary genes over the course of cellular differentiation[38], repression of repetitive regions/endogenous retroviruses to promote genomic stability[1,39], and imprinting of genes in the germ line[40–42]. Indeed, dysregulation of DNA methylation and its effector proteins is now known to be important for the genesis of malignancy[43] or developmental disorders (e.g. Rett syndrome, imprinting disorders)[13]. In addition to the canonical 5-mC form of DNA methylation, several other potentially functional DNA modifications have been described in recent years (Figure 1.3)[32]. These include oxidized forms of 5-mC, such as 5-hydroxymethylcytosine, which has been associated with neurodevelopment and neuronal functions[44–49]; and methylation of nitrogen 6 of adenine (6-methyladenine, abbreviated 6-mA), the function of which is less clear[32,50].



5-methylcytosine (5-mC) | 5-hydroxymethylcytosine (5-hmC) | 6-methyladenine (6-mA)

Figure 1.3: Examples of DNA Modifications.
Structures of 5-methylcytosine, 5-hydroxymethylcytosine, and 6-methyladenine DNA modifications. Note that the hydrogen bond acceptors and donors involved with Watson-Crick base pairing are unaffected by these covalent modifications.

The second category of epigenetic regulation is the modification of transcription and translation by ncRNAs. Indeed, though only 1-3% of the human genome is comprised by protein-coding genes[51,52], it has been shown that roughly 75-90% of the human genome is transcribed into RNA[53,54], with the vast majority of those transcripts existing as non-coding RNAs. The role of some non-coding RNAs have long been clear – for example, ribosomal RNA (rRNA) serves as the catalytic machinery for translation, with transfer RNA (tRNA) serving as the carrier of amino acids and decoder of mRNA in this process. More recently, other ncRNAs have been shown to be critical epigenetic regulators of transcription. One of the first ncRNAs to be so identified was *Xist*, which was identified in the 1990s as essential[55] for X-inactivation (at least for stabilization of the inactivated state[56]) and subsequently found to be sufficient for the same[57]. Since then, many different classes of ncRNAs have been identified as modulators of transcription or transcript stability. These types of ncRNAs include, amongst others: enhancer RNAs (eRNAs), which are short transcripts produced from enhancer regions[58–60] that are thought to specifically regulate transcription of the promoters under enhancement[61]; microRNAs (miRNAs), which are short RNAs that endogenously downregulate transcripts with which they hybridize[62–64]; Piwi-interacting RNAs (piRNAs), which are small RNAs that silence transposable elements and maintain genomic stability, particularly in germ line cells[65–67]; circular RNAs (circRNAs), a variable-length class of RNAs that are thought to absorb excess miRNAs[68]; and chromatin-enriched RNAs (cheRNAs), which are long ncRNAs associated with chromatin that serve as cell-type-specific cis-regulatory activators or repressors of transcription[69–71]. Collectively, these ncRNAs – and others – are able to positively and negatively regulate the transcription and function of the protein-coding segments of the genome through a variety of mechanisms, often in a cell-type specific manner and without modifying the sequence of the genomic DNA itself.

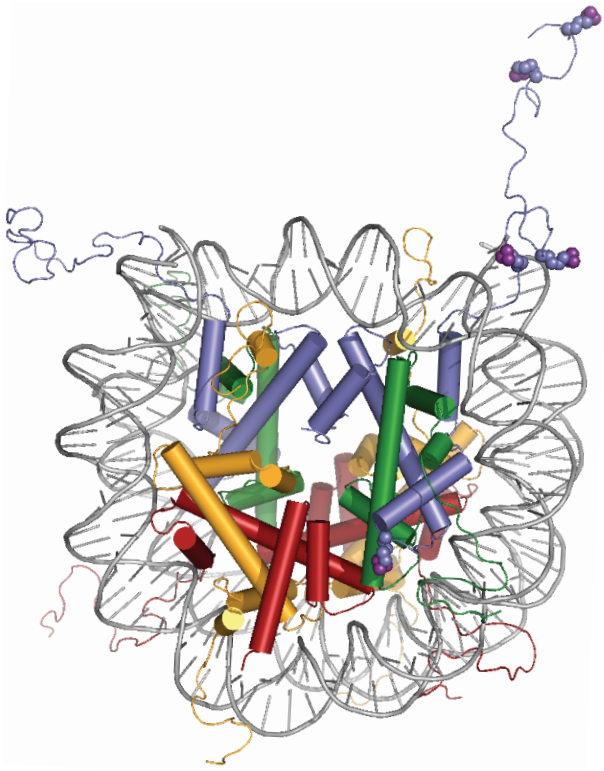


Figure 1.4: Structure of the nucleosome.
Structure of the nucleosome, showing DNA wrapped around an octamer of histone proteins, which can be post-translationally modified. Adapted from Werner and Ruthenburg[75].

The third mechanism of epigenetic regulation is through modification of the nucleosomes and, in particular, the histone proteins. As previously noted, in the eukaryotic genome, DNA is organized into nucleosomes, which are comprised by 147 base pairs of DNA wrapped about a core octamer of histone proteins : two copies each of histones H2A, H2B, H3, and H4 (Figure 1.4)[28]. Nucleosomes, particularly in their higher-order complexes and structures, provide organized compaction to the genome so it can feasibly fit in the nucleus while remaining sufficiently organized to be functional[26,27,31].This compaction must, however, be balanced with the accessibility of the genome; if a region of the genome is highly compacted by nucleosomes (a state of chromatin referred to as heterochromatin), then it will not be easily accessible for transcriptional machinery, thereby repressing gene expression[72,73]. Conversely, if a region of the genome is less compacted

and more accessible (a state of chromatin referred to as euchromatin), then it will be more available for transcription. This provides another method of epigenetic regulation; if the compaction of the genome can be modified at a given location, then that can be used to activate or repress transcription at that locus. This modulation is achieved through post-translational modifications (PTMs) of the histones that comprise the nucleosomes[73,74].

Many different types of histone modifications have been identified with a broad range of roles[76]. These modifications can be largely classified into two categories (which are not mutually exclusive): function through direct impact on structure and function through interaction with binding partners. Members of the first of those classes of histone modifications are thought to act by disruption of the local chromatin structure. For example, some modifications are thought to function by introducing negative charges onto the histone surface and reducing its overall positive charge, weakening the electrostatic interactions between the histone surface and the negatively charged DNA phosphodiester backbone and thus driving transcription. Such modifications include histone phosphorylation, which has been associated with active transcription and DNA damage repair[77], and histone acetylation, which is broadly thought to activate transcription and demarcate enhancer regions[12,78–81]. Other modifications serve to disrupt local chromatin structure through non-electrostatic modulation of steric interactions. These include histone ubiquitylation, which introduces a large modification that sterically disrupts neighboring nucleosomes and activates transcription[82,83], and proteolytic cleavage of the histone tail (also called histone tail clipping), which is through to be important for differentiation[84–86]. These modifications all have the capacity to directly modify the structure of the local chromatin by modulating the favorability of the physical associations within or between nucleosomes and, thereby, to modify the accessibility of the chromatin for transcriptional machinery.

The second class of histone modifications is the set of modifications that function through interactions with specific binding partners. This is not mutually exclusive with the former category; for example, histone lysine acetylation modifications are bound by bromodomain-containing proteins to enact some of their functions[87,88]. However, this class of histone modifications also includes more subtle chemical changes that are unlikely to have significant direct impacts on the structure. One example of such a modification is histone methylation, wherein a proton is replaced with a methyl group. This substitution does not change the charge of a given residue and represents a very subtle change in size relative to the unmodified residue, making significant steric interactions unlikely. Nonetheless, histone methylations are critical for many epigenetic regulatory pathways with a broad range of functions[73,76]. Though recent work has examined the role of histone arginine methylation as an epigenetic regulator[89,90], the best-studied histone methylations are those on lysine residues[31,73,74,76], a class of histone PTMs with highly varied regulatory roles conferred by their recruitment of specific protein binding partners. H3K4 trimethylation (H3K4me3), for example, is bound by PHD fingers domains on proteins such as BPTF[91,92] and TAF3[93,94] to remodel chromatin, make the chromatin more accessible, and activate gene transcription[17,20,94]. H3K27 trimethylation (H3K27me3), by contrast, is bound by Polycomb group proteins[16,95], particularly a subset of chromeobox (Cbx) proteins[96,97], to cause chromatin to be more tightly compacted, less accessible, and transcriptionally repressed. H3K9 di- and trimethylation (H3K9me2 and H3K9me3, respectively) are similarly bound by HP1 family members to similarly cause heterochromatin formation, particularly at repetitive elements and at centromeres or telomeres[98–101]. Many other histone methyllysine PTMs have been described in the literature, with varying degrees of biochemical or functional validation[73,76,102].

Studying the Genomic Distribution of Histone Modifications

Collectively, histone PTMs represent critical regulators of local genomic structure and function with important roles in gene regulation, physiologically cellular differentiation, pathological cellular dysregulation, and oncogenesis[23,103,104]. Accordingly, much work over the last several decades has focused on better understanding the role of histone PTMs in a broad range of cellular, developmental, and/or clinical contexts. But to understand the functions of histone modifications and the genomic features with which they associate, it is first critical to understand the genomic distribution of the same. To that end, the critical first step for most studies of histone modifications is to ask: where in the genome are these histone modifications located, and how prevalent is each modification? Several methods have been developed to answer those questions, each with their own limitations.

In the past, it has been challenging to answer those two questions simultaneously. Broadly, the study of histone modifications coalesced around two classes of methods: those that make quantitative measurements globally and those that make relative measurements locally. The former category describes methods that can quantify histone modifications globally; with these methods, it is possible to measure the absolute abundance of a histone modification (i.e. the proportion of histones or nucleosomes with the modification of interest) across all the histones or nucleosomes in the genome. These methods primarily profile the histone proteins directly to detect the presence and quantity of modifications without concern for the accompanying DNA.

One of the most common methods for this purpose is Western blotting[105], in which a protein sample is separated by SDS-PAGE, transferred to a membrane, and bound by an antibody specific for the target of interest (in this case, a particular histone modification), which can then be detected directly or with a secondary antibody. With appropriate quantitative Western blotting procedures[106,107], including using protein standards on the membrane as calibrants and employing

a reasonable-quality antibody[108], Western blotting can be used to measure the abundance of an individual histone modification quickly and cheaply. However, quantifying more than one histone modification requires separate experiments with separate protein standard calibrants processed alongside the cellular protein samples of interest. Further, this method is critically reliant on the antibody reagent used, which is problematic given that many commercially available antibodies have very low specificity or, on occasion, bind to the wrong target entirely[108–112]. Despite these limitations, for measuring histone modification abundance (and changes therein) genome-wide for a limited set of modifications, Western blotting remains a powerful tool for molecular biologists.

Another method that is often used for global histone modification abundance measurements is mass spectrometry. In most of these workflows, histones are purified, digested, and subjected to liquid chromatography (LC) coupled to tandem mass spectrometry (MS/MS)[113,114]. The post-digestion fragments are separated by the liquid chromatography step, after which they are separated by charge/mass ratio in the first mass spectrometry step[113–116]. Peptides with a given approximate charge/mass ratio are then subjected to the second mass spectrometry step, which fragments the peptide further and measures the mass/charge ratio of those fragments. This fragmentation pattern can then be analyzed for the hallmarks of different histone modifications and, based on the relative contributions of each modification to the fragmentation pattern, the abundance of each such modification[113,114]. Altogether, this method is able to measure the global abundance of a broad range of histone modifications without needing to find new reagents or generate new standards for each such mark. More recent work has extended this process to the purification of entire nucleosomes, followed by a similar LC-MS/MS method to quantify the abundance of different histone modification combinations on a nucleosome, even on different histone proteins[117]. The result there, however, is the same: measurements of a broad range of histone modification global abundances. Though

extremely powerful, mass spectrometry analysis is nowhere near as inexpensive or straightforward as Western blotting; whereas Western blotting can be done in a matter of hours with standard lab equipment and expertise, mass spectrometry experiments can be considerably more involved and require much more expensive and specialized equipment with specialized techniques. As such, it is a useful "gold standard" for quantifying histone modifications but is not as commonly used as Western blotting.

These techniques belong to the class of methods that measure global absolute abundance of the histone modification of interest without any information on its localization. Another class essentially takes the inverse approach, making local measurements of relative modification abundance. These methods primarily focus their readouts on the DNA fragments bound to the nucleosomes with the modification of interest rather than directly probing the protein itself[118]. Though some proof-of-concept work has shown that nucleosomes with a modification of interest can be directly identified and sequenced by microscopy-based methods[119], the vast majority of the methods in this category function by purifying nucleosomes with the target PTM, then recovering and analyzing the associated DNA.

Chromatin immunoprecipitation and its limitations

The most common method operating under that workflow, by a wide margin, is chromatin immunoprecipitation, or ChIP[118,120]. In this method (Figure 1.5A), chromatin is fragmented into mono- or oligonucleosomal fragments and incubated with antibodies that will bind with high affinity to the modification of interest. The Fc stem of the antibody is then captured by a Protein A or Protein G resin, along with anything bound to the variable domain of the antibody (i.e. the nucleosomes of interest). After several washing steps to remove weakly bound off-target nucleosomes, the DNA

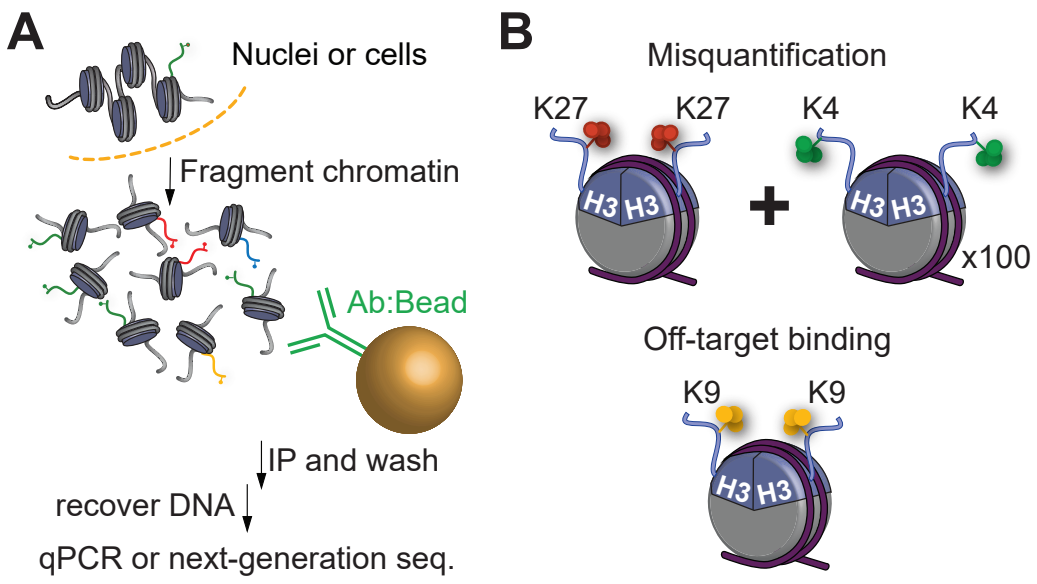


Figure 1.5: Chromatin Immunoprecipitation and Problems.
(A) ChIP workflow. Chromatin is fragmented into mononucleosomal fragments and bound by an antibody targeting a modification of interest. DNA is purified from the recovered nucleosomes and mapped to the genome as a proxy for the histone modification of interest. Adapted from Grzybowski et al.[118] and Shah et al.[124]. **(B)** Problems of ChIP, presented as problems with H3K27me3 ChIP. Conventional ChIP is a relative metric and is thus susceptible to misquantification, making comparison of different marks impossible. Off-target binding of the antibodies also complicates interpretation of a ChIP experiment.

can be recovered from the nucleosomes still bound to the antibody and mapped to the genome using either quantitative polymerase chain reaction (qPCR) or next-generation sequencing (NGS). The interpretation is that the DNA is a proxy for the targeted histone modification; if more DNA is recovered from a given locus than from a control region, then it is assessed that the PTM of interest is enriched at that particular locus[118]. Some similar methods (e.g. CUT&RUN[121], CUT&TAG[122], ChIP-exo[123]) make modifications on this protocol at various stages to serve particular purposes, but the overall interpretation is the same: the amount of DNA from a given location recovered reflects the amount of histone modification at that locus relative to other loci and the rest of the genome.

Since it was first described in 1984 by Gilmour and Lis (for bacterial proteins)[125] and in 1988 by Solomon et al. (for histones in eukaryotic cells)[120], ChIP has become one of the mainstays

of modern molecular biology[17–20,78,126,127]. However, it also has several critical limitations that hamper its interpretability in many applications. As noted above, conventional ChIP measures the enrichment of a histone modification at a given location relative to the rest of the genome, normalized either to control loci (in the case of ChIP-qPCR) or total next-generation sequencing (NGS) read depth (in the case of ChIP coupled to NGS, or ChIP-seq). This can be problematic even for comparisons between loci of a single sample if the loci have different nucleosome occupancies or fragment unevenly[118]. This can be accommodated by sequencing fragmented input chromatin and measuring enrichment as the fold change in the IP over the input, but input sequencing can be expensive due to the high read depths necessary to obtain adequate genomic coverage.

Even with input normalization, standard ChIP-seq experiments cannot be easily compared to each other, making it difficult to compare ChIPs for different modifications or in different cellular contexts. As previously noted, traditional ChIP-seq is a non-absolute quantification method, measuring the relative amount of a given histone modification at a locus as compared to the rest of the genome or a control region[118]. This relative measurement can be adequate for comparing the histone modification levels of different loci within a single cellular context and for a single histone modification. However, if there is a global difference in the abundance of histone modifications, then it is impossible to quantitatively compare ChIP signals from these two different experiments (Figure 1.5B). If a locus has a change in histone modification abundance that is proportional to the change in global abundance, then standard ChIP-seq normalization will not reveal a difference because there is no change in the relative abundance at that locus. Indeed, some work has shown that even a 75% difference in global abundance results in virtually no change to traditional ChIP-seq measurements with global normalization[128]. Because different ChIP-seq experiments are normalized to separate quantities and thus exist on distinct and separate scales, it is impossible to quantitatively compare

two different ChIP-seq tracks – whether they are for two different histone modifications or two different cellular contexts with a global abundance difference[118,128].

Even for a single ChIP experiment, where such questions of quantitative comparison are less relevant, standard ChIP still has one more crucial limitation that hampers its interpretability: antibody specificity. ChIP is critically dependent on the antibody binding the target of interest with high specificity and excluding other species. It has been repeatedly shown, however, that this is not always the case; though the extent of the antibody problem was not fully grasped for some time, it has now been shown in multiple contexts that antibodies often have a propensity to bind histone modifications other than those that they are targeted towards[108,109,111,112,118,124,129] (Figure 1.5B). Indeed, in some cases, commercial antibodies have been highly specific for the wrong target entirely[108]. Without methods to assess the specificity of the antibody, it is difficult to know whether the signal at any given locus represents primarily on- or off-target binding – let alone to know whether signal from two different loci represents equally specific binding.

To address some of these problems, the Ruthenburg Lab developed internally calibrated chromatin immunoprecipitation (ICeChIP)[118]. At the very beginning of this workflow, the sample is spiked with a set of nucleosome standards (Figure 1.6). These standards represent semisynthetic nucleosomes bearing on- or off-target modifications, each with a unique DNA "barcode" for later downstream identification and quantification[118]. These standards are introduced into the workflow prior to chromatin fragmentation, and the remainder of the ICeChIP protocol proceeds as standard for a native ChIP experiment; the pulldown is conducted, DNA is purified, and sequencing reads are mapped to the genome. At this point, the relative pulldown of each of the standards can be computed as a proportion of that present in the input; for example, the uniquely identifiable DNA from the on-target nucleosome standard may have a 35% recovery (or enrichment). This represents the

proportion of nucleosomes that have the modification of interest that would be recovered genome-wide; this number is then used to calibrate the ChIP signal, yielding the absolute proportion of nucleosomes at a given genomic locus with the modification of interest, or the histone modification density (HMD)[118]. Further, the recovery of the off-target standards can be compared to the recovery of the on-target standard; if the on-target standard is recovered with much greater efficiency than the off-target standards, then that is an indication that the IP proceeded with high specificity[118].

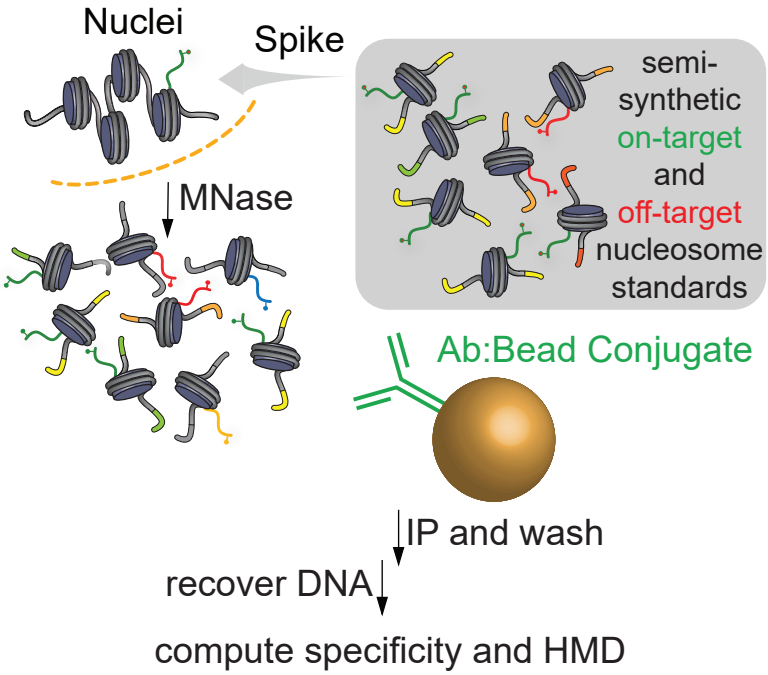


Figure 1.6: Internally Calibrated Chromatin Immunoprecipitation.
Internally calibrated chromatin immunoprecipitation (ICeChIP) workflow. Adapted from Grzybowski et al.[118] and Shah et al.[124].

The result of the ICeChIP experiment is a measurement of absolute histone modification abundance at any given genomic locus, along with a broad assessment of the overall specificity of the pulldown. In this way, ICeChIP addresses both the problem of misquantification of different histone modifications as well as the problem of antibody quality (at least insofar as it becomes possible to measure its specificity).

Open Questions and This Work

Despite its advantages, however, ICeChIP has not solved every problem of chromatin immunoprecipitation. In this work, I will describe my inquiries into four separate knowledge gaps in quantitative chromatin immunoprecipitation.

First is the question of antibody specificity; though it was well-described that antibodies could bind off-target species, it was less clear how prevalent this was in the context of ChIP. This would be of particular concern for pulldowns of different methylation states because of the subtlety of the chemical changes therein. In Chapter 2, I describe my work on this question in the context of H3K4 methylation states, wherein I characterize antibodies targeting the H3K4 methylation states and identify weaknesses in common methods of antibody validation. In the process, I show that low-quality antibodies can drive faulty biological interpretations and, conversely, use high-quality data to develop new quantitative insight into enhancer regulation.

Second is the question of co-occupancy and co-occurrence of different histone modifications on a single nucleosome. ICeChIP remains highly useful for measuring the abundance of an individual modification but, in its published form, is not capable of simultaneously measuring whether a nucleosome has two distinct modifications. At best, it can be determined that two different modifications are enriched at a given genomic locus and that some nonzero quantity must coexist, but the extent of that coexistence (as opposed to the existence of distinct cellular populations with different histone modifications) remained unclear. In Chapter 3, I describe my work on H3K4me3/H3K27me3 bivalent histone modification patterns, in which both H3K4me3 and H3K27me3 modifications exist on a single nucleosome. I describe the development of a sequential form of ICeChIP that can purify nucleosomes with both modifications and use this method to study the role (or lack thereof) of this modification in developmental poising of gene expression.

Third is the question of internal histone modifications. ICeChIP is a native protocol, meaning that it does not massively disrupt the structure of the nucleosome through the immunoprecipitation step. This is perfectly fine for the highly accessible histone tails, which can be easily bound by an antibody even in the native conformation. However, the native structure of the nucleosome presents a challenge for internal modifications on the nucleosome globular domain, such as the H3K79me2 modification, where the antibody is less able to reach and bind the modification of interest resulting in low-specificity pulldowns. In Chapter 4, I describe my work on denaturative ICeChIP, which modifies the ICeChIP procedure to denature the nucleosome in the immunoprecipitation, thereby making internal modifications more highly accessible for antibody capture. With this procedure, I profile H3K79me2 in a variety of cellular contexts to explore its role in leukemogenesis and the maintenance of the leukemic transcriptional profile.

Fourth is the backend of the ICeChIP-seq protocol: alignment and processing of next-generation sequencing reads. The critical first step of NGS read processing is alignment of each read to the reference genome. However, given how highly repetitive most commonly studies genomes are, a given NGS read may map to many distinct loci with acceptable alignment quality. Most analyses simply discard these ambiguously mapped reads, but such a practice leaves many regions of the genome unanalyzed despite the abundance of potentially functional elements in these repetitive regions. In Chapter 5, I describe my work to develop SmartMap, a tool employing a Bayesian reweighting algorithm to allocate ambiguously mapped reads. I then use SmartMap to study the histone modification patterns at several classes of repetitive elements, identifying new classes of repetitive elements with potential functional significance.

Collectively, this document seeks to address knowledge gaps in each of the major variability points of a ChIP-seq experiment: the antibody, the pulldown protocol, and the quantification

backend. In doing so, it is my hope that my work illuminates several of the critical flaws with ChIP-seq as it is traditionally practiced and shows how those flaws can be avoided, thereby driving new insight into the mechanisms of epigenetic regulation.

CHAPTER 2: ON THE STUDY OF H3K4 METHYLATION STATES

Attributions

This chapter has been adapted from: Shah, R. N. *et al.* Examining the Roles of H3K4 Methylation States with Systematically Characterized Antibodies. *Molecular Cell* **72,** 162–177 (2018). Peptide array experiments were conducted by members of the Rothbart Laboratory at the Van Andel Institute, and mouse embryonic stem cell lines were cultured and gifted by members of the Wysocka Laboratory at Stanford University. ICeChIP experiments in Fig. 2.7D-E and 2.8B were conducted by Adrian Grzybowski, PhD'18. The other experiments were conducted by the author.

Abstract

Histone post-translational modifications (PTMs) are important genomic regulators often studied by chromatin immunoprecipitation (ChIP), whereby their locations and relative abundance are inferred by antibody capture of nucleosomes and associated DNA. However, the specificity of antibodies within these experiments have not been systematically studied. Here, we use histone peptide arrays and internally calibrated ChIP (ICeChIP) to characterize 52 commercial antibodies purported to distinguish the H3K4 methylforms (me1, me2, and me3, with each ascribed distinct biological functions). We find that many widely used antibodies poorly distinguish the methylforms and that high- and low-specificity reagents can yield dramatically different biological interpretations, resulting in substantial divergence from the literature for numerous H3K4 methylform paradigms. Using ICeChIP, we also discern quantitative relationships between enhancer H3K4 methylation and promoter transcriptional output and can measure global PTM abundance changes. Our results illustrate how poor antibody specificity contributes to the "reproducibility crisis," demonstrating the need for rigorous, platform-appropriate validation.

Introduction

Over the past several decades, ChIP has contributed many seminal insights into histone PTM regulation and distribution[17–20,126,127,130–132]. However, the interpretation of a ChIP experiment critically relies on the assumption of near-perfect antibody specificity. The validity of this conjecture for the thousands of existing ChIP-seq datasets is uncertain, given that many commercial antibodies display considerable off-target binding in other experimental formats[108,111,112,118,133,134].

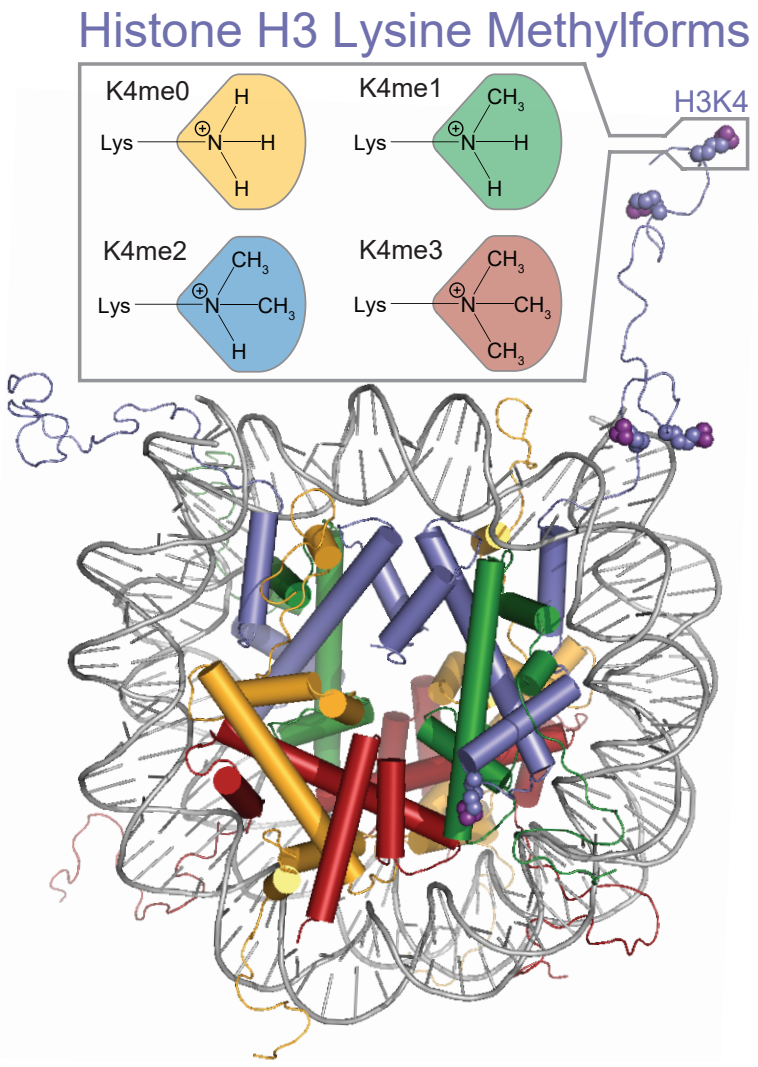


Figure 2.1: Nucleosome with H3K4 methylation states.
Structure of the nucleosome with histone H3 lysine 4 (H3K4) highlighted and schematic of methyl-forms of H3K4. Adapted from Werner and Ruthenburg[75].

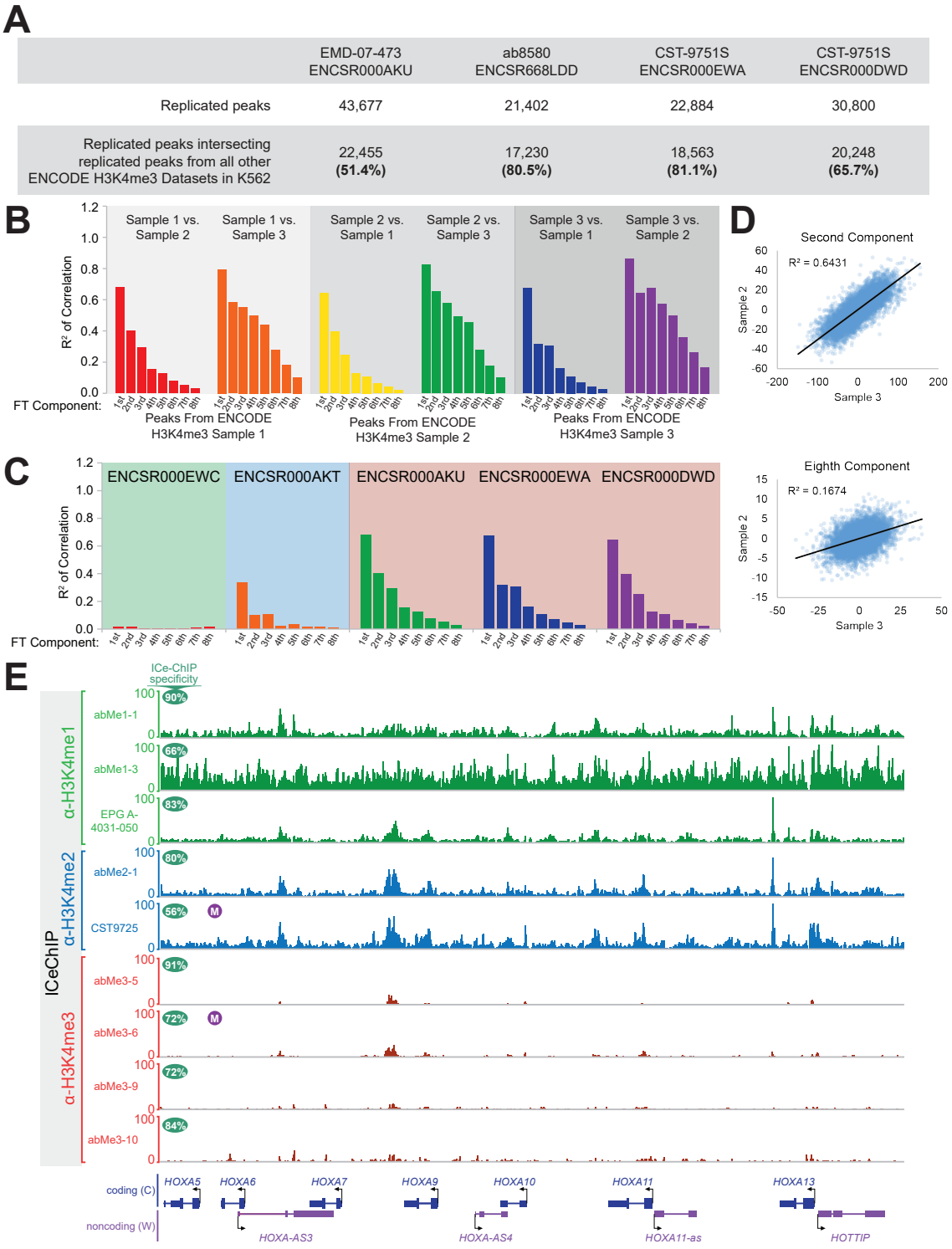


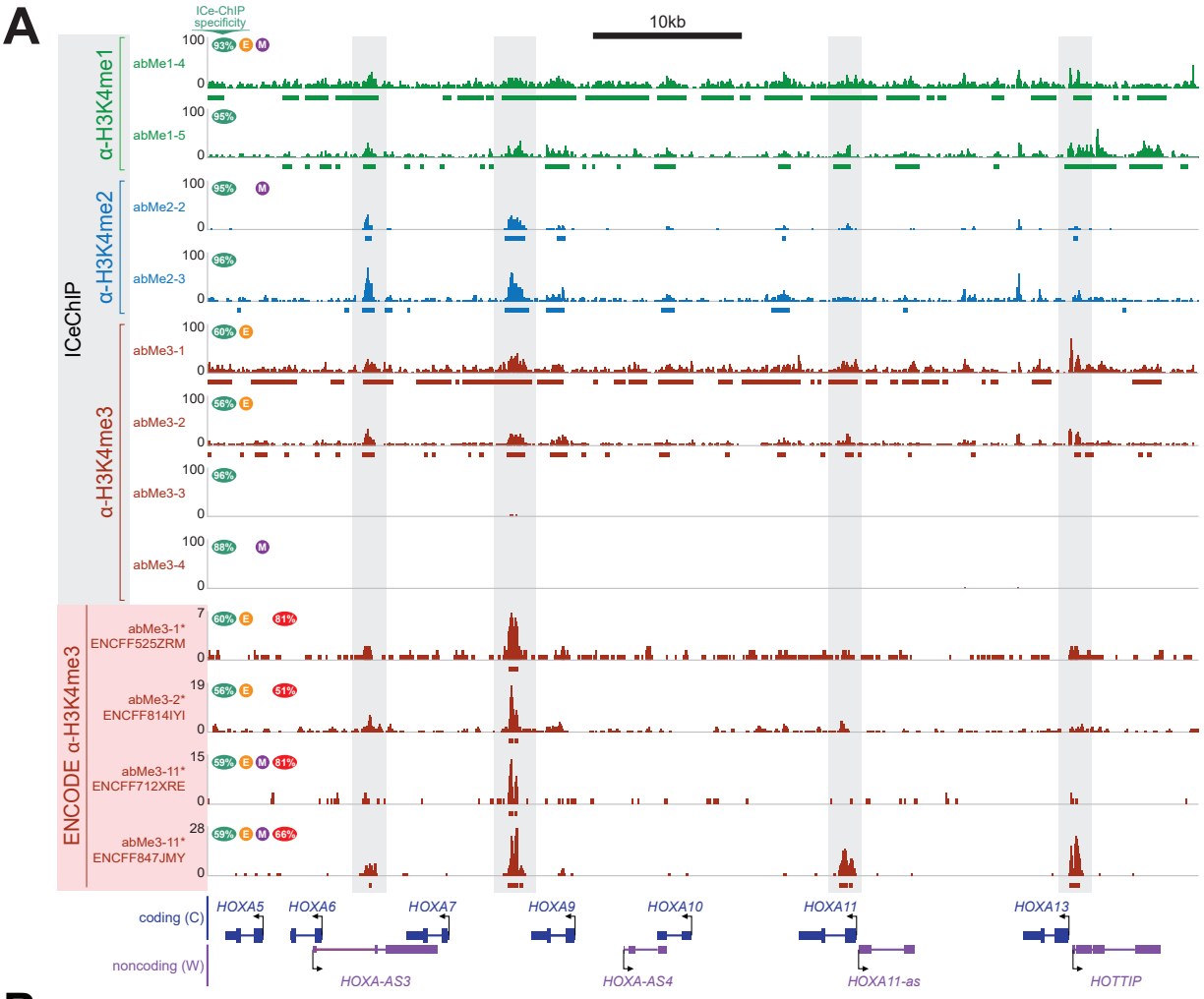
Figure 2.2: ENCODE ChIP-seq datasets display internal inconsistency and incongruity.

(A) ICeChIP-seq and ENCODE ChIP-seq tracks at distal HoxA cluster in K562 cells. Highly specific antibodies reveal absence of H3K4me3; low-specificity antibodies detect appreciable signal from lower methyl forms. ENCODE tracks are reminiscent of ICeChIP tracks but differ from one

Figure 2.2, continued:
another and do not show true H3K4me3 signal. Green oval shows ICeChIP methylform specificity for each antibody , orange circle with E indicates antibody validated to ENCODE standards, purple circle with M indicates monoclonality, and red oval shows percentage of peaks in each ENCODE dataset found in all three of the other ENCODE datasets. Bars below tracks represent peaks. **(B)** Abbreviation codes, specificities in ICeChIP and peptide arrays, and target IP enrichments for antibodies referred to in the main text. Values represent average ± SD.

Here, we have interrogated the specificity of antibodies targeting the three methylation states of lysine 4 on histone H3 (H3K4me1, H3K4me2, and H3K4me3; Fig. 2.1), each ascribed distinct roles in chromatin regulation. H3K4me1 (~5-20% global abundance[135]) is thought to mark enhancers[18,19,131] and flanks promoters[17]. H3K4me2 (~1-4% global abundance[135]) is associated with tissue-specific transcription factor binding sites[136], enhancers[131], and promoter edges[136–138]. H3K4me3 (~1% global abundance[135]) defines active transcription at promoters[17,20,92–94,132], and is also implicated in V(D)J recombination[139], meiotic crossovers[140], and pre-mRNA splicing[141,142]. Critically, many of these conclusions were drawn presuming that ChIP could discriminate between the three methylation states. Concerningly, apparent ChIP-seq replicates with different antibodies for a single such can radically differ, even within a single cell line and when using the highly standardized protocols of the ENCODE consortium (Fig. 2.2, 2.3), raising concerns that antibodies cannot specifically discriminate between these different modifications. As such, we sought to systematically investigate the capacity of antibodies to distinguish different methylation states of H3K4.

To this end, we assessed the specificities of 52 commercial "ChIP grade" antibodies using histone peptide microarrays and ICeChIP (Fig. 2.4, 2.5). In the first approach, antibody is incubated with slide-immobilized peptides, and bound regions identified with a fluorescently labeled secondary antibody (Fig. 2.6A). Peptide microarray measurements allow simultaneous testing of a broad range of different off target, on target, and combinatorial PTMs[108,111,112,133]. The technique



B

Antibody Code	abMe1-1	abMe1-2	abMe1-3	abMe1-4	abMe1-5	abMe1-6	abMe2-1	abMe2-2	abMe2-3	abMe3-1
Manufacturer	Abcam	Active Motif	Active Motif	Cell Signaling	Thermo Fisher	Cell Signaling	Abcam	EMD Millipore	Thermo Fisher	Abcam
Product	ab8895	39297	39635	5326BF	710795	5326	ab7766	05-1338	710796	ab8580
Lot	GR305231-1	01714002	30615011	2	QL230603	1	GR289627-1	2757107	QL230606	GR190229-1
IP Methyl. Specificity (% H3K4 Pulldown)	90 ± 4	76 ± 1	66 ± 17	93 ± 2	94.7 ± 0.5	94.4 ± 0.1	80 ± 4	95.6 ± 0.1	95.3 ± 0.4	60 ± 3
PA Methyl. Specificity (% H3K4 Signal)	96 ± 3	100 ± 0	64 ± 4	99 ± 1	98 ± 2	76 ± 2	57 ± 4	77 ± 5	99.5 ± 0.8	66 ± 2
Target IP Efficiency (% Input)	64 ± 22	0.7 ± 0.4	3 ± 1	33 ± 27	20 ± 5	46 ± 7	55 ± 11	21 ± 5	25 ± 5	63 ± 17

Antibody Code	abMe3-2	abMe3-3	abMe3-4	abMe3-5	abMe3-6	abMe3-7	abMe3-8	abMe3-9	abMe3-10	abMe3-11
Manufacturer	EMD Millipore	Thermo Fisher	Abcam	Abclonal	EMD Millipore	Diagenode	Active Motif	Diagenode	EpiGentek	Cell Signaling
Product	07-473	PA5-40086	ab12209	A2357	05-745R	C15200152	61379	C15410003	A-4033-050	9751
Lot	DAM1623866	RL2301825	GR275790-1	46698	2813867	001-11	24615006	A1052D	606361	9
IP Methyl. Specificity (% H3K4 Pulldown)	56 ± 7	96.5 ± 0.5	88 ± 3	91 ± 2	72 ± 8	73 ± 12	67 ± 1	72 ± 5	84 ± 10	59 ± 7
PA Methyl. Specificity (% H3K4 Signal)	81 ± 5	100 ± 0	88 ± 9	90 ± 9	89 ± 3	86 ± 9	57 ± 3	78 ± 7	81 ± 3	57 ± 2
Target IP Efficiency (% Input)	54 ± 7	17 ± 7	5 ± 1	13 ± 3	55 ± 18	1.4 ± 0.8	0.4 ± 0.3	40 ± 24	48 ± 25	59 ± 22

Figure 2.3: ENCODE ChIP-seq datasets are incongruous with each other and ICeChIP-seq.

(A) Concordance of replicated peaks for each H3K4me3 ENCODE dataset with indicated antibody in K562 cells. Top row shows number of called peaks replicated across the two biological

Figure 2.3, continued:
replicates for each dataset, and bottom row the number and percentage of such peaks that intersect with peaks common to all four of the ENCODE H3K4me3 datasets. **(B)** Peak shape and intensity analysis by pairwise correlation between pre-cosinusoidal factors of eight-component discrete Fourier cosine transform of fold changes over control about peaks from ENCODE H3K4me3 Sample 1 (ENCSR000AKU, left), Sample 2 (ENCSR000EWA, centre), or Sample 3 (ENCSR000DWD, right). If only intensity of signal was different (i.e. same data with different scaling, the R^2 would approach 1, and scalar factor reflecting difference would be apparent from the slope. These comparisons (Movie S1) indicate very limited similarity amongst any of two ENCODE data sets, with modest large-scale similarity (early terms) decaying to negligible fine-scale similarity (late terms). **(C)** R^2 of correlation between pre-cosinusoidal factors of eight-component discrete Fourier cosine transform on antibody-measured peaks of corrected ICeChIP-seq HMD versus antibody-measured fold change over control for H3K4me1, H3K4me2, and H3K4me3 ENCODE ChIP-seq datasets in K562 cells. **(D)** Example scatterplots showing high correlation (top) and low correlation (bottom) of Fourier Transform components on peaks from ENCODE H3K4me3 Sample 3. **(E)** Genome browser view at the HOX locus for antibodies not shown in Figure 1A. Green circle represents ICeChIP aggregate specificity, and purple circle with M indicates monoclonality.

is considered the current gold standard of antibody characterization, but whether it recapitulates antibody performance in ChIP is unclear due to marked differences in experimental format[143]. In contrast, ICeChIP uses DNA-barcoded semisynthetic nucleosome standards encompassing panels of histone PTMs directly spiked into a chromatin sample, allowing the measurement of antibody specificity in situ, and the determination of histone modification density (HMD), the absolute amount of PTM over a genomic interval (Fig. 2.6B)[118]. However, each nucleosome standard must be independently synthesized, which is labor-intensive and technically challenging. Though peptide arrays and ICeChIP have been compared in a very limited way[112], the small scale of such studies precluded broader conclusions. Further, previous studies centered on antibody discrimination between different lysine residues (e.g. H3K4me3 vs. H3K9me3) rather than different methylation states of a single lysine (e.g. H3K4me2 vs. H3K4me3), the latter representing a potentially greater challenge. Integrating peptide array and ICeChIP analyses now enables us to critically evaluate antibodies and determine the extent of data transferability between each format.

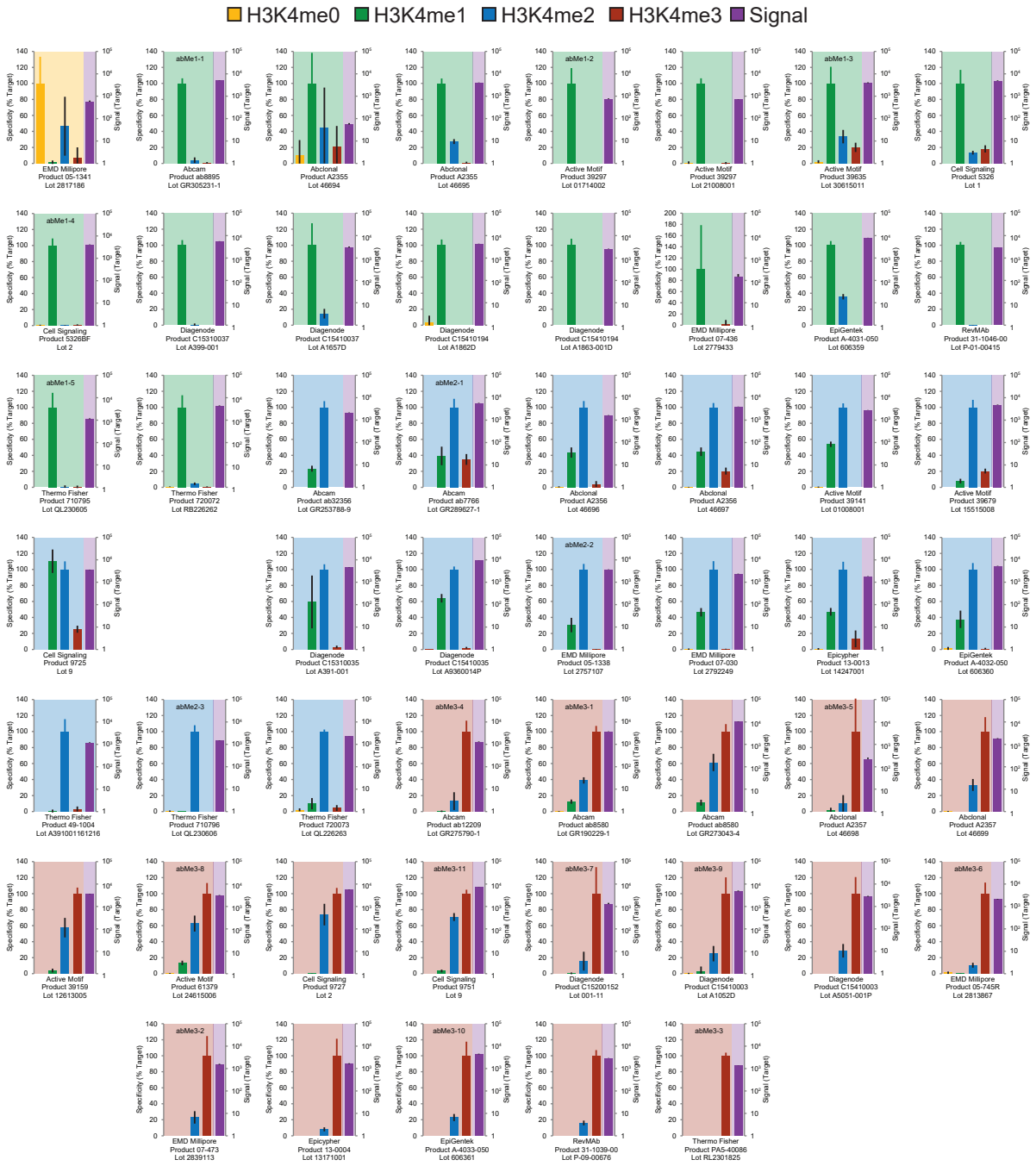


Figure 2.4: Anti-H3K4 methylation antibodies display a broad range of peptide array specificities. The specificity of H3K4 methylform antibody binding on peptide arrays expressed relative to on-target capture. Black error bars represent SD of off-target specificity; colored error bars represent average standard error of on-target signal. Purple bar represents raw fluorescence signal from secondary axis and maps onto secondary axis. Fluorescence measurements for each antibody (n=6), independent at the level of spotting, but simultaneously measured against one antibody dilution.

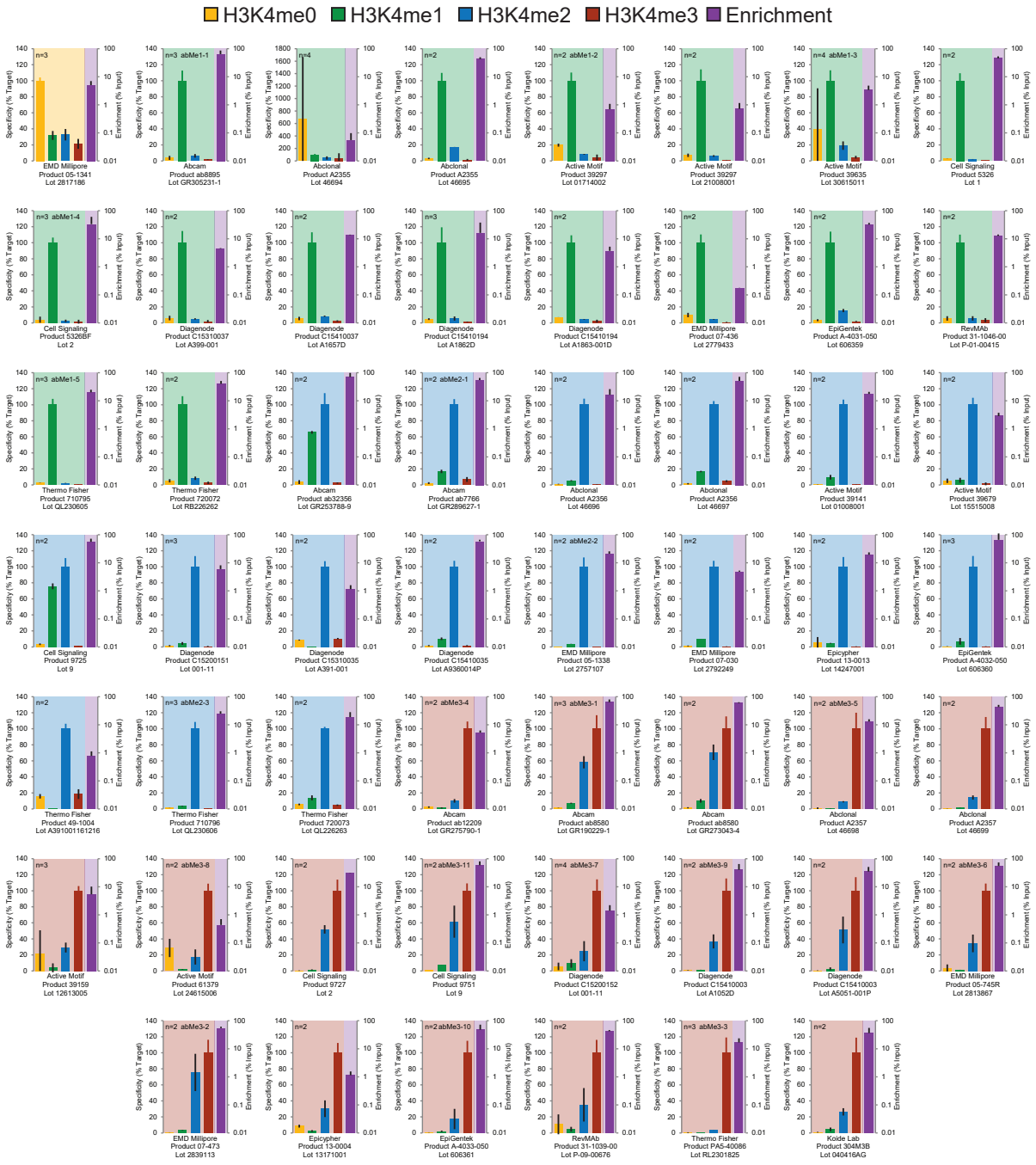


Figure 2.5: Anti-H3K4 methylation antibodies display a broad range of ICeChIP specificities.

The specificity of H3K4 methylform antibody binding in ICeChIP relative to on-target capture. Black error bars represent SD of off-target specificity; colored error bars represent average standard error of on-target signal. Purple bar represents ChIP enrichment and maps onto secondary axis (right). ICeChIP was conducted with 3 µg of mammalian chromatin and 3 µg of each antibody (see Methods). Enrichment of each standard was measured by qPCR; n represents independent ICeChIP experiments averaged for each antibody.

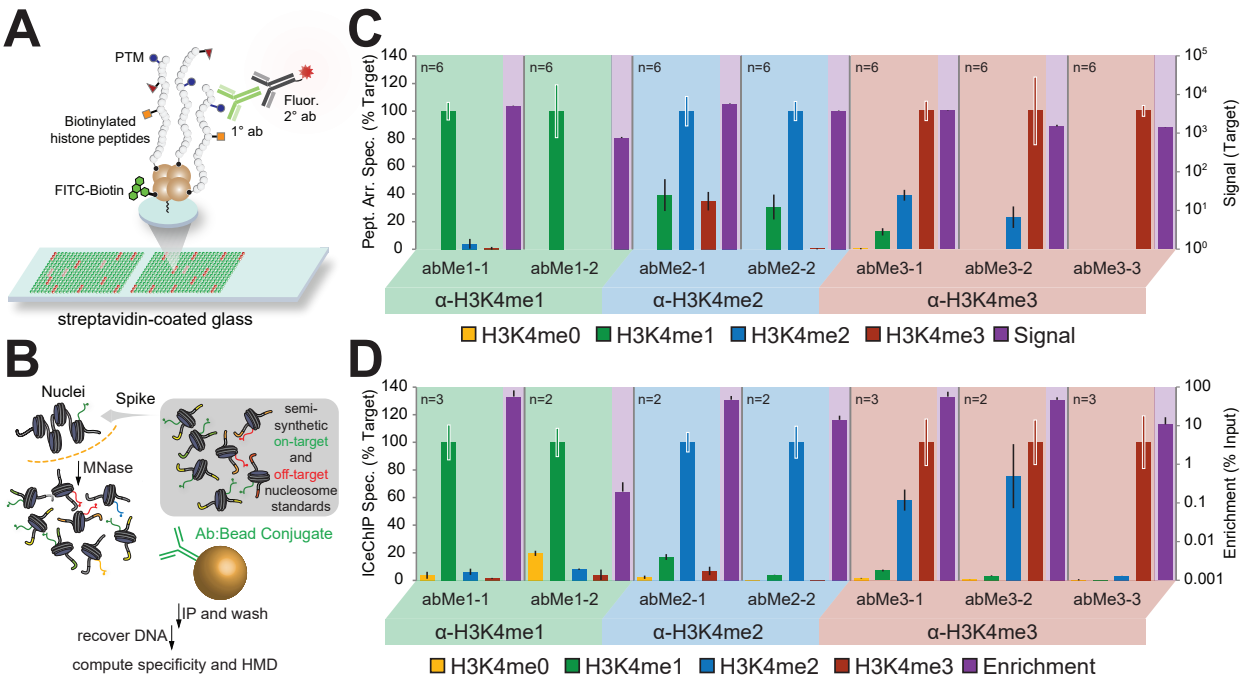


Figure 2.6: Histone 3 lysine 4 (H3K4) antibodies display a range of methylform specificities. **(A, B)** Experimental workflows of (A) peptide arrays and (B) ICeChIP. **(C, D)** A representative selection of methylform binding (target relative to other forms on the left axis) by antibody from (C) peptide arrays and (D) ICeChIP is presented in bar graph form (extracted from the larger set of 52 antibodies: Fig. 2.4, 2.5). Purple bar represents raw fluorescence signal or ChIP enrichment, and maps to right axis ($\log_{10}$ scale). Black error bars represent SD of off-target specificity; colored error bars represent average SD of on-target signal.

Results

Antibody specificities range widely and often diverge across methods

A representative cohort from the 52 antibodies screened with both peptide array and ICeChIP (Fig. 2.4, 2.5) is shown in Fig. 2.6C and 2.6D. High-specificity antibodies, with >90% aggregate methyl-specificity, were identified by both approaches (e.g. abMe1-1 and abMe3-3 in Fig. 2.6C-D; Tables 2.1-2.4), but notably, these reagents are often infrequently used (Tables 2.1-2.4). When present, cross-reactivity most commonly occurred between states differing by a single methyl group (Fig. 2.4, 2.5) and was most severe for the anti-H3K4me3 antibodies.

Table 2.1: Characteristics for antibodies targeting H3K4me0.

Manufact.	Product	Lot	Antibody Code	IP Methyl. Specificity (% H3K4)	PA Methyl. Specificity (% H3K4)	Target IP Efficiency (% Input)	Citat.
EMD Millipore	05-1341	2453179	–	54 ± 5	64 ± 6	5 ± 2	2

Table 2.2: Characteristics for antibodies targeting H3K4me1.

Manufact.	Product	Lot	Antibody Code	IP Methyl. Specificity (% H3K4)	PA Methyl. Specificity (% H3K4)	Target IP Efficiency (% Input)	Citat.
Abcam	ab8895	GR305231-1	abMe1-1	90 ± 4	96 ± 3	64 ± 22	218
Abclonal	A2355	46694	–	37 ± 35	57 ± 19	0.06 ± 0.04	0
Abclonal	A2355	46695	–	82 ± 2	78 ± 2	43 ± 7	0
Active Motif	39297	01714002	abMe1-2	76 ± 1	100 ± 0	0.7 ± 0.4	11
Active Motif	39297	21008001	–	88 ± 2	99 ± 2	0.7 ± 0.4	11
Active Motif	39635	30615011	abMe1-3	66 ± 17	64 ± 4	3 ± 1	1
Cell Signaling	5326	1	abMe1-6	94.4 ± 0.1	76 ± 2	46 ± 7	2
Cell Signaling	5326BF	2	abMe1-4	93 ± 2	99 ± 1	33 ± 27	2
Diagenode	C15310037	A399-001	–	88 ± 4	99 ± 1	4.4 ± 0.2	0
Diagenode	C15410037	A1657D	–	86 ± 3	87 ± 5	14 ± 0.5	2
Diagenode	C15410194	A1862D	–	90 ± 3	97 ± 8	16 ± 22	7
Diagenode	C15410194	A1863-001D	–	88 ± 2	100 ± 0	4 ± 2	7
EMD Millipore	07-436	DAM1687548	–	87 ± 1	97 ± 6	0.179 ± 0.003	16
EpiGentek	A-4031-050	606359	–	83 ± 3	74 ± 2	32 ± 5	0
RevMAb	31-1046-00	P-01-00415	–	87 ± 6	99.9 ± 0.2	13 ± 1	0
Thermo Fisher	710795	QL230603	abMe1-5	94.7 ± 0.5	98 ± 2	20 ± 5	0
Thermo Fisher	720072	RB226262	–	86 ± 4	95 ± 2	40 ± 11	0

Table 2.3: Characteristics for antibodies targeting H3K4me2.

Manufact.	Product	Lot	Antibody Code	IP Methyl. Specificity (% H3K4)	PA Methyl. Specificity (% H3K4)	Target IP Efficiency (% Input)	Citat.
Abcam	ab32356	GR253788-9	–	58 ± 2	81 ± 2	70 ± 29	35
Abcam	ab7766	GR289627-1	abMe2-1	80 ± 4	57 ± 4	55 ± 11	55
Abclonal	A2356	46696	–	94 ± 1	68 ± 3	16 ± 10	0
Abclonal	A2356	46697	–	81.3 ± 0.7	61 ± 3	51 ± 21	0
Active Motif	39141	01008001	–	91 ± 2	65 ± 1	18 ± 3	8
Active Motif	39679	15515008	–	89 ± 1	78 ± 3	3 ± 1	0
Cell Signaling	9725	9	–	56 ± 1	42 ± 3	56 ± 16	4
Diagenode	C15200151	001-11	–	94 ± 2	—-	6 ± 2	0
Diagenode	C15310035	A391-001	–	83.8 ± 0.2	62 ± 13	1.1 ± 0.4	0
Diagenode	C15410035	A9360014P	–	88 ± 2	60 ± 2	58 ± 8	6
EMD Millipore	05-1338	2757107	abMe2-2	95.6 ± 0.1	77 ± 5	21 ± 5	6
EMD Millipore	07-030	DAM1479603	–	89.8 ± 0.2	68 ± 2	4.8 ± 0.4	126
Epicypher	13-0013	14247001	–	90 ± 5	62 ± 4	19 ± 4	1
EpiGentek	A-4032-050	606360	–	92 ± 4	72 ± 6	66 ± 45	0
Thermo Fisher	49-1004	A391001161216	–	74 ± 2	97 ± 4	0.8 ± 0.4	2
Thermo Fisher	710796	QL230606	abMe2-3	95.3 ± 0.4	99.5 ± 0.8	25 ± 5	0
Thermo Fisher	720073	QL226263	–	81 ± 1	86 ± 6	19 ± 9	0

Remarkably, apparent specificity in peptide arrays and ICeChIP is only weakly correlated ($R^2 = 0.2337$: Fig. 2.7A) and is independent of both raw fluorescence in peptide arrays (Fig. 2.7B) and IP enrichment in ICeChIP (Fig. 2.7C), suggesting that antibody specificity trends are not driven by affinity alone. Notably, there was much greater platform disagreement for antibodies to H3K4me2 than for those to H3K4me1 or H3K4me3 (Fig. 2.8A).

Table 2.4: Characteristics for antibodies targeting H3K4me3.

Manufact.	Product	Lot	Antibody Code	IP Methyl. Specificity (% H3K4)	PA Methyl. Specificity (% H3K4)	Target IP Efficiency (% Input)	Citat.
Abcam	ab12209	GR275790-1	abMe3-4	88 ± 3	88 ± 9	5 ± 1	6
Abcam	ab8580	GR190229-1	abMe3-1	60 ± 3	66 ± 2	63 ± 17	418
Abcam	ab8580	GR273043-4	–	55 ± 4	58 ± 4	59 ± 4	418
Abclonal	A2357	46698	abMe3-5	91 ± 2	90 ± 9	13 ± 3	0
Abclonal	A2357	46699	–	86 ± 2	75 ± 4	43 ± 8	0
Active Motif	39159	12613005	–	66 ± 11	62 ± 5	5 ± 5	80
Active Motif	61379	24615006	abMe3-8	67 ± 1	57 ± 3	0.4 ± 0.3	2
Cell Signaling	9727	2	–	65 ± 3	58 ± 4	31 ± 0.7	11
Cell Signaling	9751	9	abMe3-11	59 ± 7	57 ± 2	59 ± 22	24
Diagenode	C15200152	001-11	abMe3-7	73 ± 12	86 ± 9	1.4 ± 0.8	0
Diagenode	C15410003	A1052D	abMe3-9	72 ± 5	78 ± 7	40 ± 24	43
Diagenode	C15410003	A5051-001P	–	65 ± 8	78 ± 5	35 ± 16	43
EMD Millipore	05-745R	2813867	abMe3-6	72 ± 8	89 ± 3	55 ± 18	10
EMD Millipore	07-473	DAM1623866	abMe3-2	56 ± 7	81 ± 5	54 ± 7	189
Epicypher	13-0004	13171001	–	71 ± 5	93 ± 2	1.1 ± 0.4	1
EpiGentek	A-4033-050	606361	abMe3-10	84 ± 10	81 ± 3	48 ± 25	0
RevMAb	31-1039-00	P-09-00676	–	67 ± 5	86 ± 2	42 ± 5	0
Thermo Fisher	PA5-40086	RL2301825	abMe3-3	96.5 ± 0.5	100 ± 0	17 ± 7	0
Koide Lab	304M3B	040416AG	–	76 ± 1	—-	36 ± 19	0

We found that specificity in ICeChIP was not substantially affected by changes in relative methylform abundances for the antibodies screened (Fig. 2.8B), suggesting that different chromatin abundances of the methylforms do not mask true antibody ChIP specificity. Yet, for approximately half of the antibodies screened in peptide arrays, changing the amount of epitope or antibody altered observed specificity (Fig. 2.7D-E and 2.8C-D). We speculate that these differences in antibody

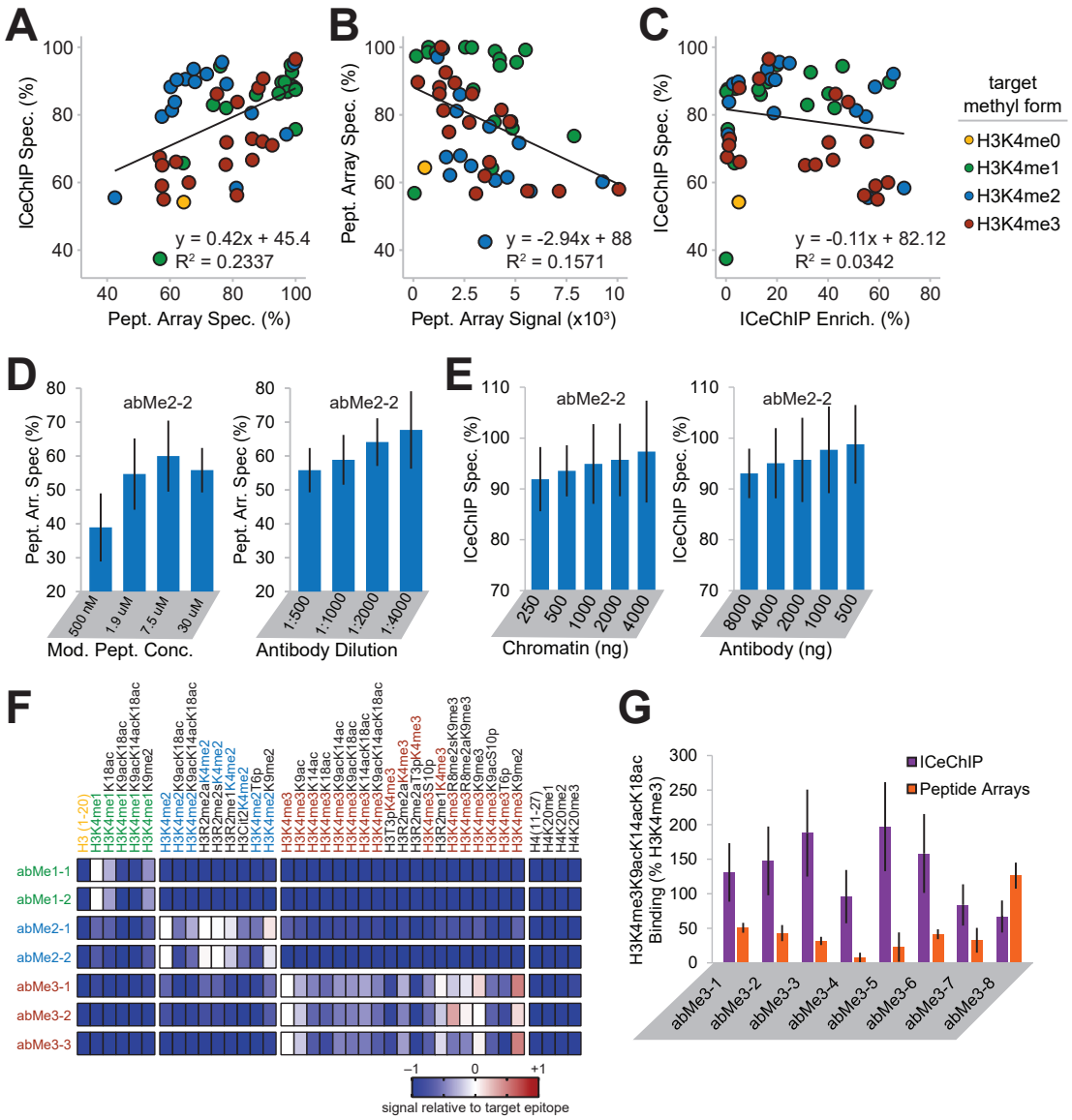


Figure 2.7: Antibodies can display different specificities in peptide arrays and ChIP.

(A) Specificity computed for each antibody (of 52 tested) as target H3K4 methylform (indicated by dot colour) enrichment normalized to the sum of all H3K4 methylform enrichments. **(B)** Methylform specificity versus on-target signal in peptide arrays. **(C)** Methylform specificity versus on-target enrichment in ICeChIP. **(D)** Aggregate specificity in peptide arrays of abMe2-2, varying concentration of modified peptide (left) or antibody dilution (right). **(E)** Aggregate specificity of abMe2-2 in ICeChIP when varying amount of input chromatin (left) or amount of antibody (right). **(F)** Heatmap of peptide array antibody binding normalized to target for select combinatorial modifications (full peptide set detected in Figure S5). **(G)** Binding in ICeChIP and peptide arrays of selected anti-H3K4me3 antibodies to H3K4me3K9acK14acK18ac relative to singly modified H3K4me3. All peptide arrays were conducted with six fluorescence measurements, and all ICeChIPs with one of each pulldown. Error bars represent SD.

A

	ICeChIP More Specific	Approx. Equally Specific	Pept. Arrays More Specific	Total
α-H3K4me0	0	1	0	1
α-H3K4me1	1	11	5	17
α-H3K4me2	12	2	2	16
α-H3K4me3	1	11	6	18
Total	14	25	13	52

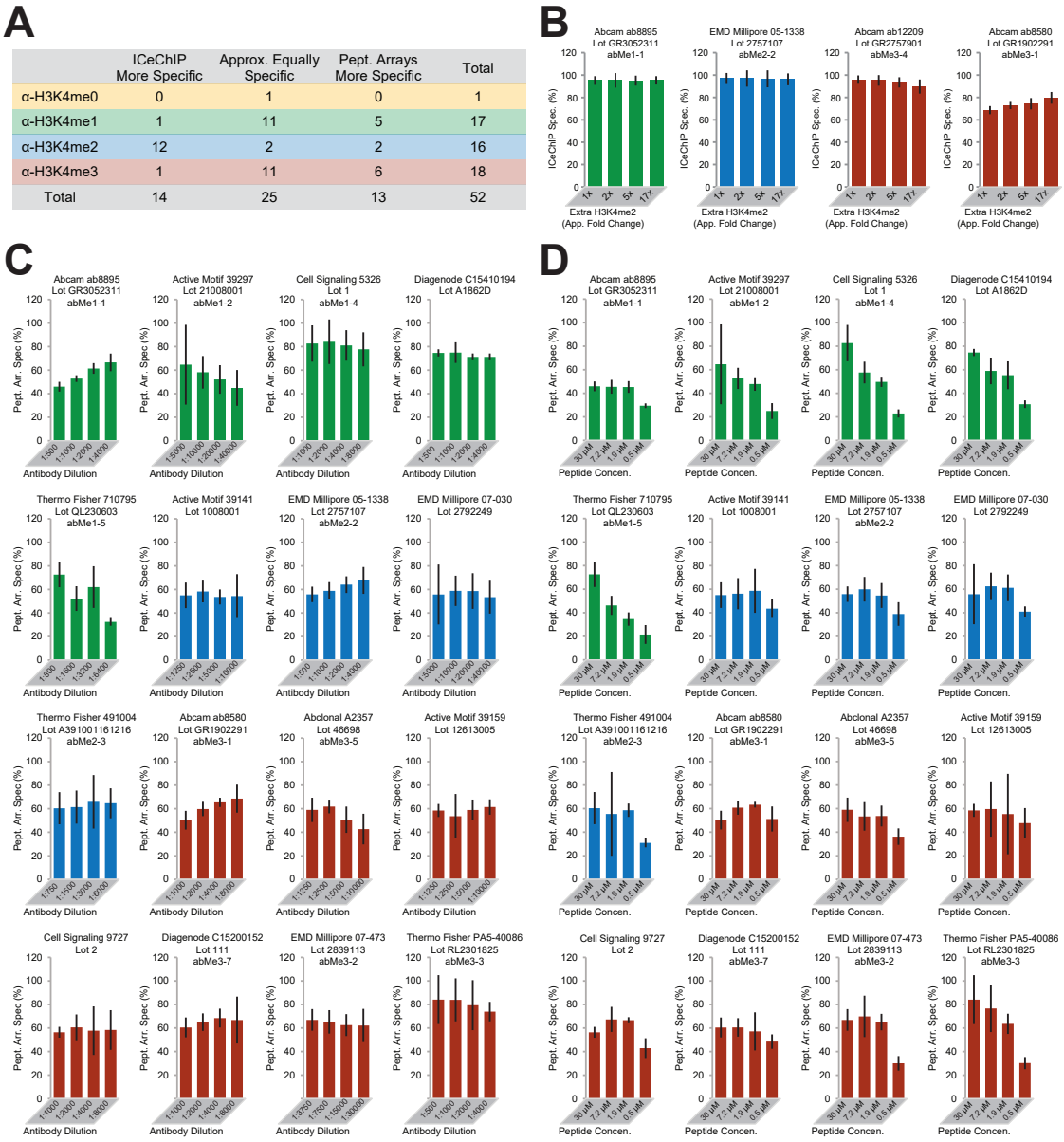


Figure 2.8: ICeChIP and peptide arrays have discrepancies that can be modulated.

(A) Agreement of antibody of methyl-form specificity between ICeChIP and peptide arrays, to within 10 percentage points. **(B)** ICeChIP aggregate specificity for four antibodies when H3K4me2 nucleosomes bearing a different DNA sequence is added in excess of endogenous H3K4me2. One replicate per ICeChIP experiment for a total of four ICeChIPs per antibody. **(C, D)** Antibody specificity on EpiTitrate peptide arrays with varying amounts of (C) antibody and (D) modified peptide. Approximately half the antibodies tested show marginally altered specificity with increasing dilution, albeit not always in the same direction. Most antibodies show decreased specificity at the most dilute modified peptide concentration, and approximately a third show decreasing specificity with increased modified peptide dilution more broadly. Six fluorescence readings per peptide array experiment, with one independent experiment per antibody dilution.

specificities are the result of the different physical interactions underpinning the two methods: in peptide arrays, dilute antibody binds densely packed epitope on a surface, whereas ICeChIP (and ChIP more generally) is the opposite. However, a complete understanding of these differences remains a challenge for future inquiry.

Peptide arrays permit simultaneous querying of combinations of H3K4 methylations with other PTMs[108,111,112,133]. In this context, many antibodies displayed reduced affinity for their target with flanking lysine acetylation (Fig. 2.7F, all except abMe2-1 and abMe3-2; and Fig. 2.9), which are thought to occasionally coexist[144,145]. Yet in ICeChIP, we largely do not observe such reduced binding, with several antibodies displaying the opposite trend (Fig. 2.7G). Although these proximal modifications do impact apparent H3K4me3 capture in both platforms, the effects are subtle and poorly aligned between the two methods.

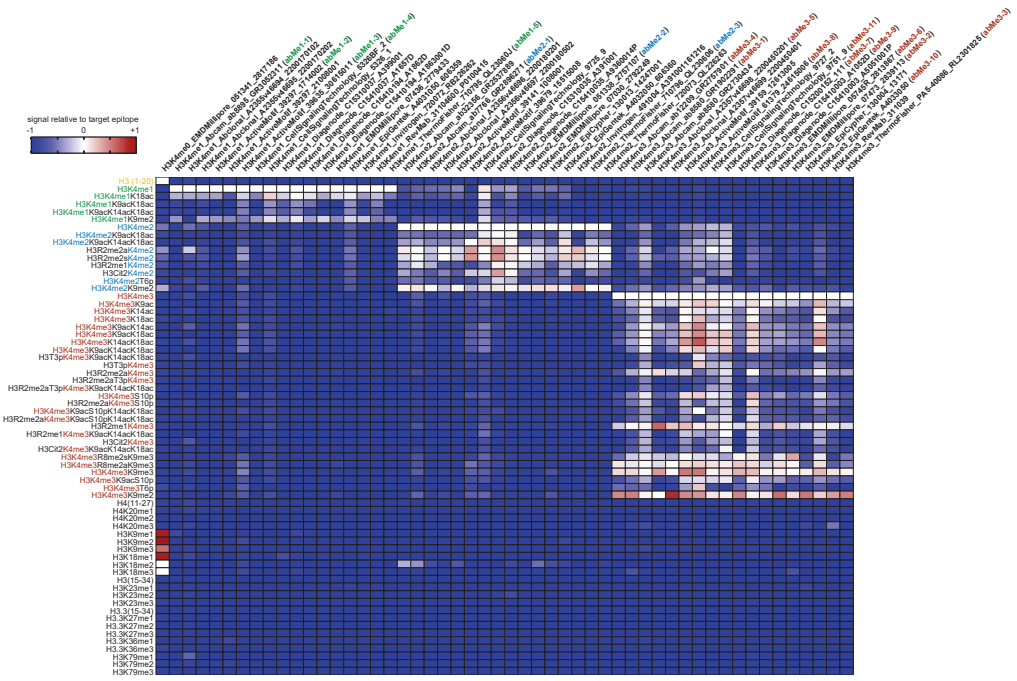


Figure 2.9: Combinatorial modifications can impact antibody binding in peptide arrays.
Heatmap of antibody binding to a wide range of combinatorial and off-target peptides on peptide arrays. Signal is normalized to singly modified target epitope.

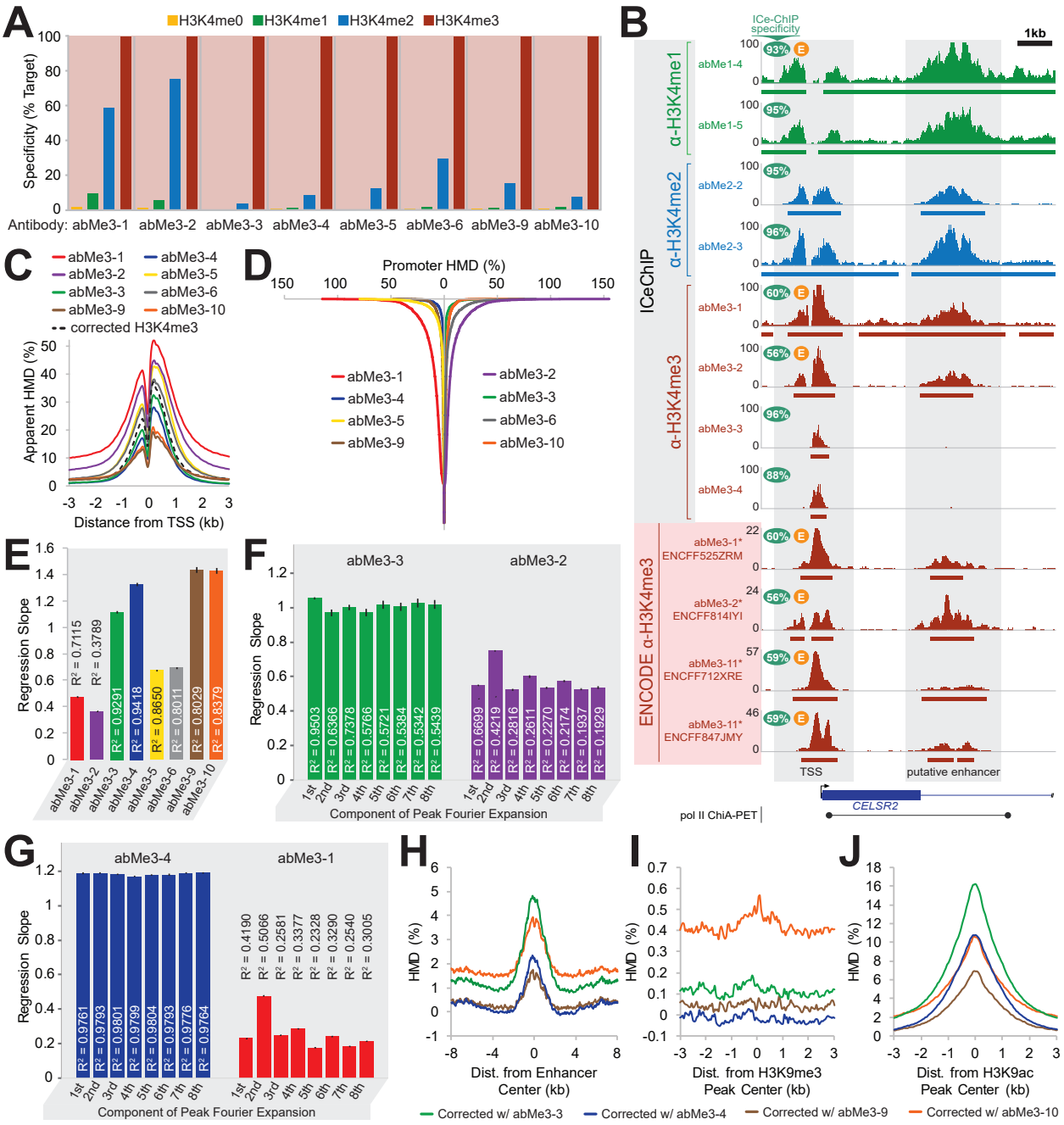


Figure 2.10: Antibodies with different specificities yield markedly different ChIP-seq profiles in K562 cells.

(A) Specificity profiles of anti-H3K4me3 antibodies measured by ICeChIP-seq (full range of standards in Fig. 2.11A). **(B)** A representative chromosomal coordinate view showing several antibody ICeChIP-seq modification profiles and ENCODE project H3K4me3 modification profiles in K562 cells, with a putative promoter-enhancer connection (Li et al., 2012). Bars below tracks represent peaks. **(C)** Anti-H3K4me3 antibodies and signal-corrected H3K4me3 modification profiles contoured over all TSSs for all Refseq genes. **(D)** Average HMD measured by anti-H3K4me3 antibodies

Figure 2.10, continued:
(sorted in descending order) at the +1 and +2 nucleosomes of genes with signal-corrected H3K4me3 HMD ≤ 0 (7,666 Refseq genes). Vertical axis represents position in sorted gene list. **(E)** Correlation between average HMD of signal-corrected H3K4me3 versus antibody-measured HMD at antibody peaks for anti-H3K4me3 antibodies. Error bars represent 99.99% CI of regression slopes. **(F)** Correlation between pre-cosinusoidal factors of eight-component discrete Fourier cosine transform on antibody-measured peaks of signal-corrected H3K4me3 versus antibody-measured HMD for abMe3-3 (left) and abMe3-2 (right). Error bars represent 99.99% CI of regression slopes. **(G)** Correlation between pre-cosinusoidal factors of eight-component discrete Fourier cosine transform of measured HMDs by abMe3-3 versus abMe3-4 (left) or abMe3-1 (right), on peaks from abMe3-4 (left) or abMe3-1 (right). **(H, I, J)** Signal-corrected H3K4me3 modification profiles, generated from abMe1-5, abMe2-3, and the noted H3K4me3 antibody, contoured over (H) stringently defined enhancers, (I) H3K9me3 peaks, and (J) ENCODE H3K9ac peaks.

Antibodies with different off-target specificities yield materially different ICeChIP-seq profiles

We next examined 15 antibodies with a range of H3K4 methylform specificities on chromatin from K562 cells, a tier one ENCODE cell line[127]. Using our described method[118], we isolated the on-target ChIP-seq signal for four antibodies to generate signal-corrected tracks (Fig. 2.10A and 2.11A). As anticipated from its performance in both peptide arrays and ICeChIP-qPCR (Fig. 2.6C-D), abMe3-2 captures substantial H3K4me2 (which is more abundant than H3K4me3) in ICeChIP-seq (Fig. 2.10A). Consequentially, its distribution appeared more similar to that of high-specificity H3K4me2 than H3K4me3 antibodies (Fig. 2.10B). Similar off-target capture issues were observed for all other low-specificity antibodies used for ICeChIP-seq (Fig. 2.11).

We then sought to determine if high-specificity and low-specificity antibodies had demonstrably different ChIP-seq profiles genome-wide. High-specificity and corrected H3K4me3 profiles are similar about transcription start sites (TSSs), whereas low-specificity antibodies show inflated apparent HMD, consistent with off-target signal leakage (Fig. 2.10C). Strikingly, at TSSs with no measured H3K4me3 in the corrected profile, the high-specificity anti-H3K4me3 profiles display fewer genes with nonzero apparent HMD than do the low-specificity profiles (Fig. 2.10D). More-

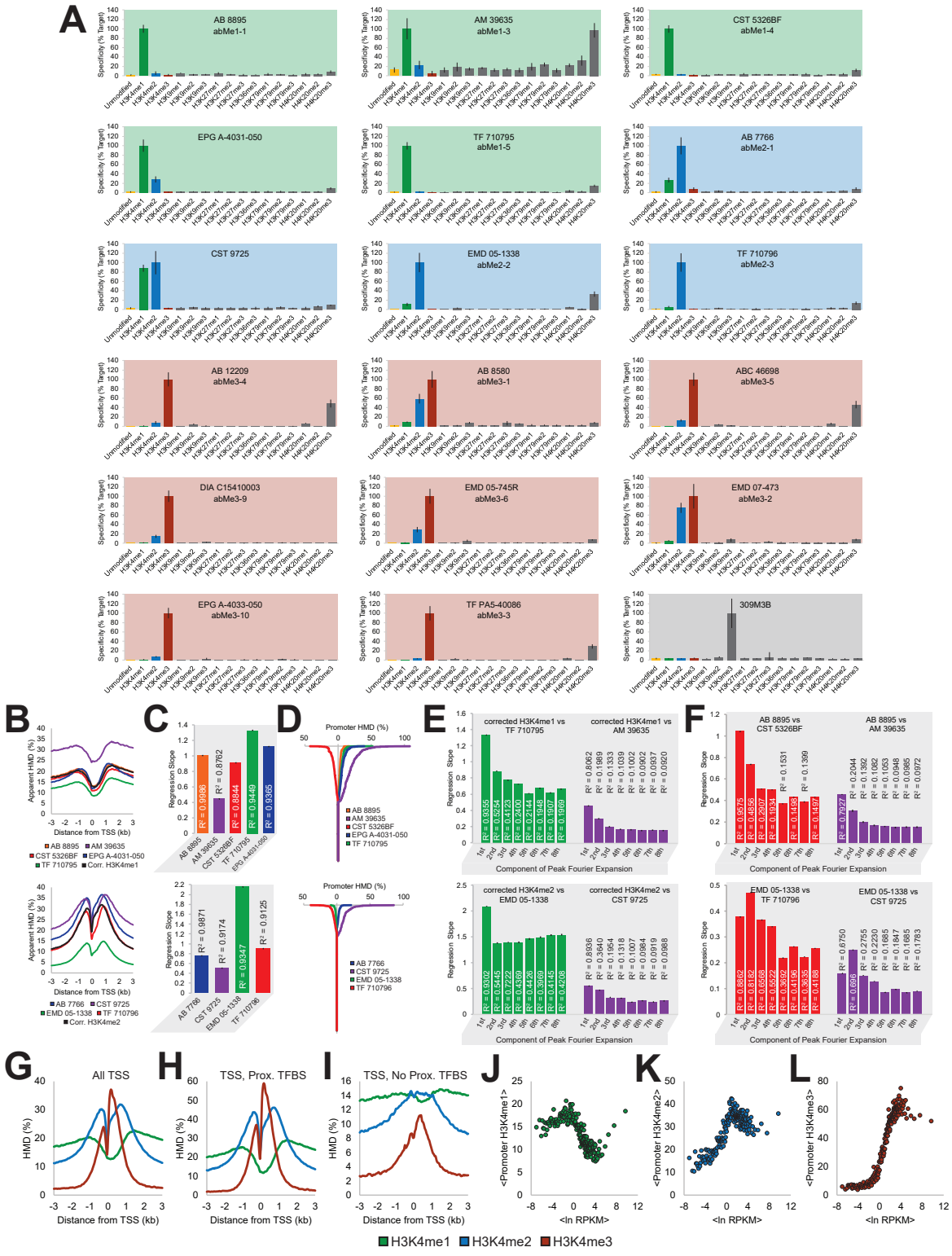


Figure 2.11: Specificity of antibodies is broadly recapitulated in ICeChIP-seq and can impact measured modification profiles.

Figure 2.11, continued:
(A) Specificities of antibodies in ICeChIP-seq experiments are identical within experimental error to those measured by qPCR, and a broader range of off-target internal standards are sampled. Error bars represent SD of estimate based on internal variability of ladder members. **(B)** Metagene contours about Refseq TSSs of anti-H3K4me1 (upper panel) and anti-H3K4me2 (lower panel) antibodies and corresponding corrected profiles. **(C)** Correlation between antibody-measured HMD and corrected HMD at antibody called peaks for anti-H3K4me1 and anti-H3K4me2 antibodies. **(D)** Antibody-measured HMD at +1/+2 nucleosomes of genes with no measured HMD in corrected profile for anti-H3K4me1 and anti-H3K4me2 antibodies. **(E)** Correlation and magnitude analysis of pre-cosinusoidal factors for eight term Fourier series comparing corrected HMD versus antibody-measured HMD for anti-H3K4me1 and anti-H3K4me2 antibodies contoured over called peaks in the corrected. **(F)** Similar analysis of pre-cosinusoidal factors of eight-component discrete Fourier cosine transform of measured HMDs by listed antibodies versus high-specificity reference antibodies abMe1-1 (upper panel) or abMe2-2 (lower panel) on peaks from listed antibodies. **(G, H, I)** Metacontours about TSSs of H3K4me1/2/3 HMD for (G) all TSSs (58,951 TSSs), (H) TSSs with a transcription factor binding site (Wang et al. 2014) within 200bp of the TSS (32,531 TSSs), and (I) TSSs without a transcription factor binding site (Wang et al. 2014) within 200bp of the TSS (26,420 TSSs). **(J, K, L)** Average (J) H3K4me1, (K) H3K4me2, and (L) H3K4me3 corrected HMD of +1/+2 nucleosomes versus ln RPKM of genes. Error bars for all correlations represent 99.99% CI for correlation slope.

over, the HMD of peaks from high-specificity antibodies correlate more closely with the corrected profile than do low-quality antibodies (Fig. 2.10E).

To compare the shapes of the ChIP-seq profiles, we applied a discrete cosine transform to the HMD distributions at called peaks genome-wide for both antibody and corrected profiles. This calculation allowed us to assess concordance of peak shape separately from HMD magnitude. The regression slope for the pre-trigonometric factors indicates concordance of HMD value, whereas the correlation coefficient indicates similarity of distribution shape. For each term, the linear correlation with corrected profile is stronger and the slope closer to unity for high- versus low-specificity antibodies (Fig. 2.10F), demonstrating that the shape and magnitude of high-specificity HMD profiles more closely resemble the signal-corrected profile. Similar comparisons between two additional high- or low-specificity antibodies for each methylform recapitulate these results (Fig. 2.10G and 2.11B-C). Together, these data suggest that the profiles of high- and low-specificity

antibodies are distinct, with different patterns genome-wide. Given that the most widely used ChIP antibodies show poor methylform specificity (Tables 2.1-2.4, Fig. 2.4, 2.5), conclusions drawn from datasets generated with these reagents should be tempered.

Beyond H3K4 methylform analysis, our ICeChIP spike-in pool also contained synthetic barcoded nucleosomes representing H3K9me1/2/3, H3K27me1/2/3, H3K36me3, H3K79me1/2/3, and H4K20me1/2/3 nucleosomes (Fig. 2.11). With the exception of the low-specificity abMe1-3, the tested antibodies did not substantially capture PTMs on other lysines in histone H3, although we note several that showed substantial binding to H4K20me3 in either array testing (Fig. 2.9) or ICeChIP (Fig. 2.11). Off-target recognition of H4K20me3 is surprising given the low primary sequence similarity with H3K4, but such binding has previously been noted in qualitative peptide arrays[133]. As H4K20me3 is relatively rare in rapidly dividing cells[146], this cross-reactivity, though concerning, may be modest in impact.

Several antibodies displayed different sensitivity to flanking additional modifications in peptide arrays, allowing us to test whether those same patterns were apparent in ICeChIP-seq. On peptide arrays, abMe3-3 showed enhanced binding to H3K4me3 paired with H3K9me2 but reduced binding in combination with acetylation marks, whereas the opposite trend was seen for abMe3-9 and abMe3-10 (Fig. 2.9). However, when signal-corrected tracks are generated with these antibodies, at stringently defined enhancers, where H3 acetylation is expected, and H3K9me3 peaks, the differences between the profiles are small and often the opposite of what is predicted by peptide arrays (Fig. 2.10H-I). Similarly, at ENCODE H3K9ac peaks, the profile corrected with abMe3-3 has ~10% higher apparent H3K4me3 HMD over abMe3-10 despite showing reduced capture of acetylated peptides in arrays (Fig. 2.9, 2.10J). Collectively, these results suggest that biases in our ICeChIP analyses due to these combinatorial modifications are modest.

ICeChIP with high-specificity antibodies yields new insights into transcriptional control

Prior studies have relied on ChIP-seq without *in situ* antibody specificity information or calibration, so we next used our robust ICeChIP-seq datasets to critically re-evaluate previous findings and search for new biological insights. In particular, we chose to investigate distal enhancers and the promoters they regulate[147]. H3K4me3 is phenomenologically[17,20,93,94,132] and biochemically[92–94] associated with active promoters[118], where it is flanked by the lower H3K4 methylforms; our present high-quality data recapitulates this general pattern (Fig. 2.10C and 2.11G-L). H3K4me1 and H3K4me2 are canonically thought to be indicative of enhancers, but not of relative enhancer activity[12,18,19,131]. There are scattered reports of H3K4me3 demarcating active enhancers[148], but the accumulated evidence suggests that H3K27ac, rather than H3K4me3, marks active enhancers[12,19,131].

Our data confirm that H3K4me1 and H3K4me2 decorate stringently defined enhancers; however, we detect little evidence for H3K4me3 at these sites (Fig. 2.10H and 2.12). Importantly, though the high-specificity antibodies show little H3K4me3 at a putative enhancer, the low-specificity anti-H3K4me3 antibodies show substantial apparent H3K4me3 at such locations, as do the ENCODE H3K4me3 ChIP-seq tracks (Fig. 2.10B and 2.12F). This artefactual capture, apparent in the low-specificity (but commonly used; see Table S1) anti-H3K4me3 antibodies (abMe3-1 and abMe3-2) and ENCODE data (some of which was performed with the same reagents), is attributable to signal leakage from lower methyl forms, which are abundant at enhancers.

Although there are some differences between datasets using different high-specificity antibodies (Fig. 2.10C-F and 2.10H-J), they all indicate extremely low H3K4me3 levels at enhancers (Fig. 2.10H, 2.12F). If some of the apparent signal inflation of abMe3-3 versus abMe3-4 (Fig. 2.10H-J) was due to enhanced capture of H3K4me3 in the context of flanking acetylation (Fig. 2.7G), these differences are quite modest. While this does not rule out the possibility that other

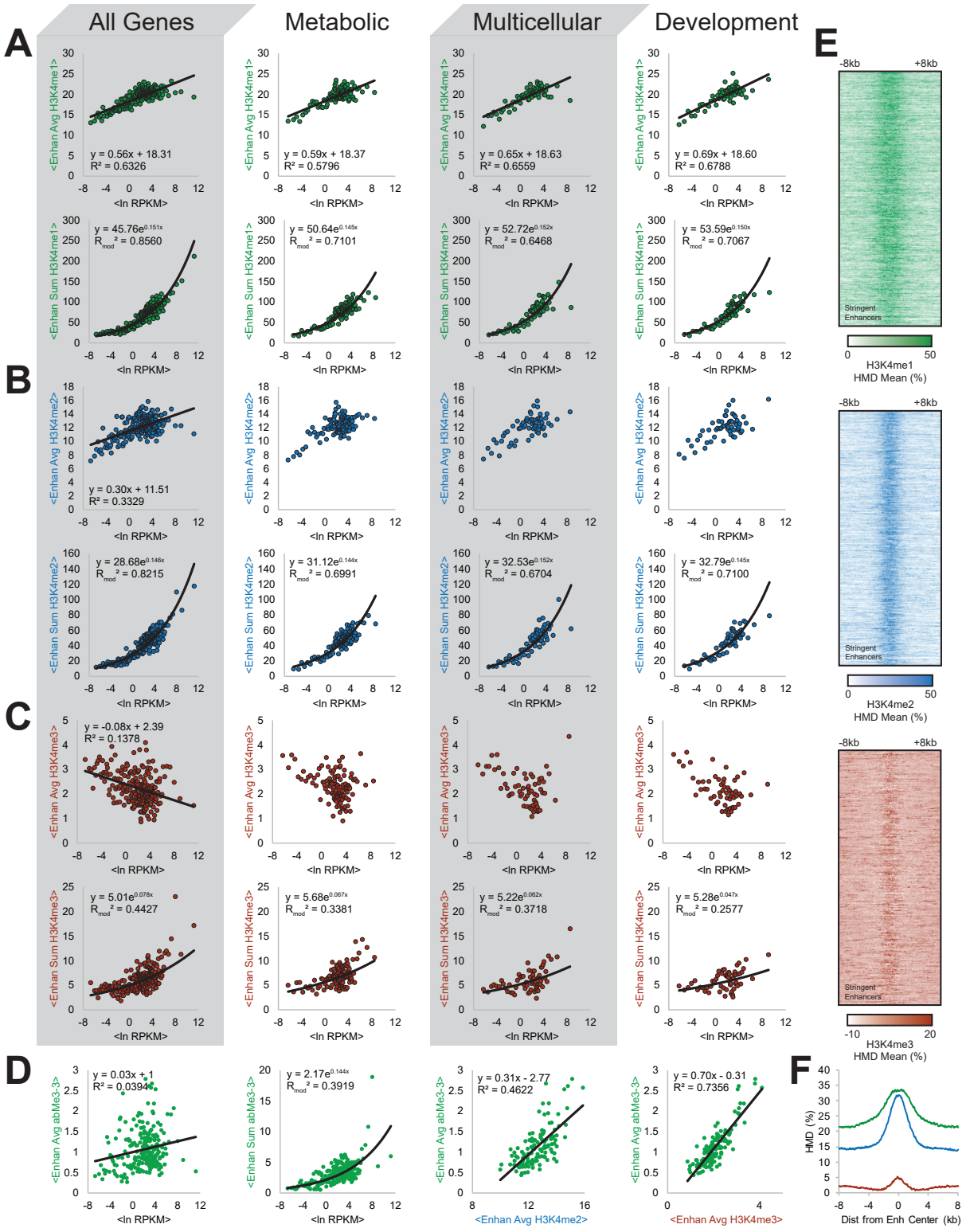


Figure 2.12: High-quality H3K4 methylation HMD datasets reveal quantitative relationships between enhancer H3K4 methylation and promoter activity.

Figure 2.12, continued:
(A, B, C) Average (top) and sum (bottom) of (A) H3K4me1, (B) H3K4me2, and (C) H3K4me3 corrected HMD across all enhancers contacting corresponding promoter regions versus ln RPKM for all classes of genes (left), metabolic genes (centre-left), multicellular system process genes (centre-right), and developmental process genes (right). **(D)** From left to right: abMe3-3 measured HMD sum versus ln RPKM. abMe3-3 measured HMD average across enhancers versus ln RPKM, average corrected H3K4me2 enhancer HMD, and average corrected H3K4me3 enhancer HMD. **(E)** Heatmaps of stringently defined enhancer HMD averages for H3K4me1, H3K4me2, and H3K4me3. All heatmaps sorted by ln RPKM of target genes. R_{mod}^2 represents R^2 of linear correlation between actual and predicted/modeled HMD. **(F)** Signal-corrected H3K4me1, H3K4me2, and H3K4me3 modification profiles contoured over stringently defined enhancers.

proximal modifications could have more severe impacts on capture efficiency, leading to bias in the interpretation agnostic of such effects, for H3K4me and flanking lysine acetylation we observe a less severe dependence than anticipated.

We next investigated the relationship between enhancer H3K4 methylation and target gene expression, as defined by RNA Polymerase II ChIA-PET contacts[147]. Though we find that transcription from a given promoter modestly correlates with the average H3K4me1 HMD across contacting enhancers (Fig. 2.12A; left, top), the sum of H3K4me1 HMD across all contacting enhancers correlates much more strongly (Fig. 2.12A; left, bottom). Similar properties were observed for H3K4me2 (Fig. 2.12B). We interpret these data to mean that the number and collective H3K4me1/me2 density of enhancers predicts promoter activity, suggesting that enhancers may operate *en masse* rather than as isolated elements, and that the lower H3K4 methylforms may play some role in this process. Conversely, neither averages nor sums of enhancer H3K4me3 HMD correlated as well with expression (Fig. 2.12C), nor did the ratio of enhancer H3K4me3 to H3K4me1 (Fig. 2.13A), contrary to prior uncalibrated ChIP studies[148].

H3K4 methylation at enhancers is thought to primarily regulate cell-type specific and developmental genes[18,19,131]. To investigate this, we compared gene expression and enhancer modification levels for metabolic, developmental, and multicellular system process-genes (Fig. 2.12A-C).

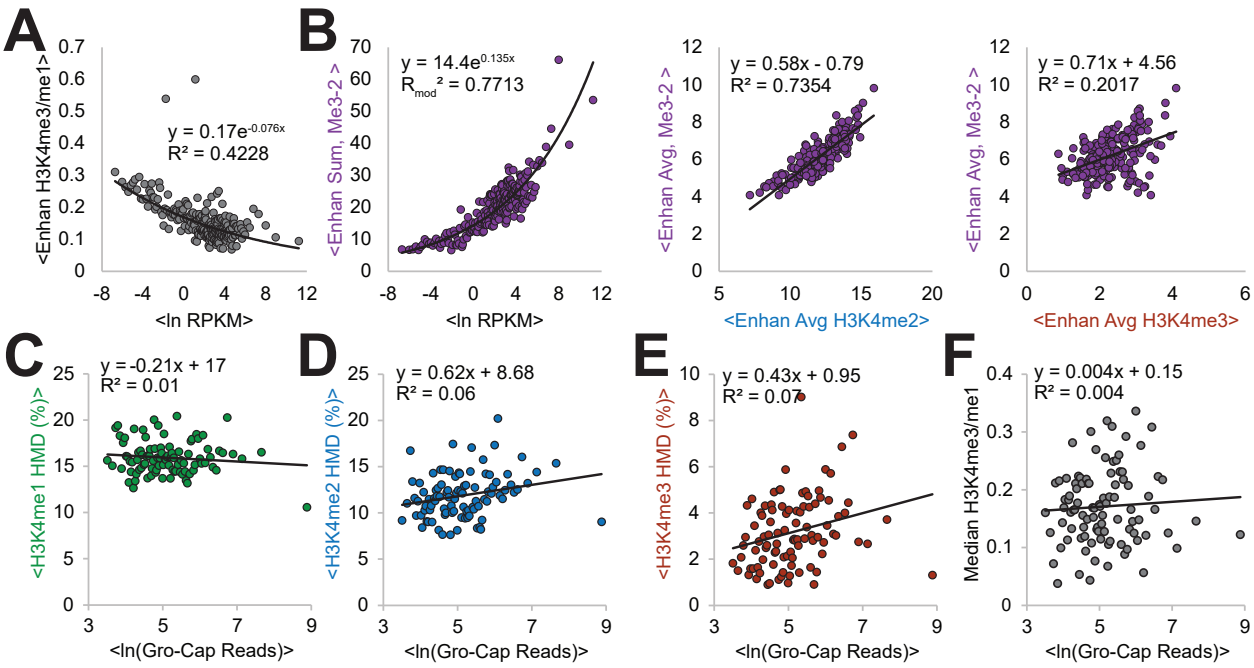


Figure 2.13: H3K4me1 and H3K4me2 HMD across enhancers contacting promoter regions is correlated to gene expression for all, metabolic, and developmental genes.

(A) Ratio of enhancer H3K4me3 HMD to H3K4me1 HMD versus ln (RPKM). **(B)** abMe3-2 measured HMD sum (left) or average (centre and right) across enhancers versus ln RPKM (left), average corrected H3K4me2 enhancer HMD (centre), and average corrected H3K4me3 enhancer HMD. **(C-F)** Average transcript production, measured by average ln GRO-Cap reads, of unstable-unstable classified genes[58] versus (C) average H3K4me1, (D) H3K4me2, (E) H3K4me3 HMD, and (F) median H3K4me3/H3K4me1 ratio. All scatterplots, unless otherwise noted, use corrected H3K4 methylation profiles and show binwise averages; bins contain fifty elements each and were created by sorting on ln (RPKM).

Remarkably, our signal-corrected datasets showed no substantial differences between these gene ontology classes, indicating that enhancer-potentiated transcriptional activation may be more universal in mammalian gene expression than formerly appreciated[18,19,131].

To determine if low-specificity antibodies can materially affect these new observations, we analyzed the HMD sum across enhancers as measured by abMe3-2, which cross-reacts with H3K4me2 (Fig. 2.5, 2.6D, 2.10A, and 2.11A). Here, the apparent H3K4me3 HMD sums at enhancers correlate strongly with gene expression (Fig. 2.13B), unlike corrected or high-specificity H3K4me3 abMe3-3 sums (Fig. 2.12C-D and 2.13B). This apparent HMD at these loci is driven

primarily by H3K4me2 rather than H3K4me3, so the low-specificity abMe3-2 incorrectly attributes this function to the latter PTM (Fig. 2.13B). Importantly, other normalization methods with spike-in chromatin[128], which normalize ChIP experiments but cannot control for specificity, would be similarly susceptible to this misleading artefact, highlighting the importance of internal standard calibration that is sensitive to antibody specificity.

Revisiting literature enhancer mark paradigms with high-specificity antibodies

Multiple reports have implied a role for H3K4me3 at enhancers[58,148], further suggesting that the H3K4me3:H3K4me1 ratio marks active enhancers[148]. Our calibrated data, which enable meaningful ratiometric comparisons, show the opposite trend in K562 cells. Specifically, we find that the ratio of calibrated H3K4me3 to H3K4me1 is inversely related to enhancer activity (Fig. 2.13A), consistent with our observation that enhancers lack substantial H3K4me3 (Fig. 2.10G, 2.12C-F, 2.13C). The prior work relied upon an antibody (abMe3-1) for which two lots performed poorly in our study (Fig. 2.6A, 2.4, 2.5, and 2.10A-B)[148]; the substantial cross-reactivity we observe with H3K4me2, which is abundant at enhancers, may account for the disparity (Fig. 2.12B). The use of crosslinking ChIP, which has been previously noted to reduce specificity[149–151], represents another potential source of the discrepancy. Regardless, several independent lines of evidence (Fig. 2.12, 2.13A) lead us to conclude that the H3K4me3/H3K4me1 ratio is not positively correlated with enhancer activity in K562 cells, and, we suspect this to be more general.

Similarly, based on ENCODE ChIP-seq data, it has been suggested that H3K4me3 levels and the H3K4me3:H3K4me1 ratio at eRNA TSSs are positively correlated with eRNA transcription levels, as measured by GRO-Cap reads in K562 cells[58]. However, there are several potential issues with the ENCODE H3K4 methylation ChIP-seq datasets. Those for H3K4me3 in K562

cells (the only H3K4 methylation state with multiple independent datasets) display substantial divergence from one another (Fig. 2.2, 2.3A-B, 2.10B) and are all very different from our high-specificity ICeChIP-seq datasets (Fig. 2.3C). This could be due to a wide variety of factors, including different antibody quality; sequencing depth; the use of crosslinked ChIP, which leads to greater off-target binding[149–151]; sonication, which can generate a large size distribution of fragments and can damage epitopes[152]; and the effect of single-end sequencing and read extension, which can result in oligonucleosome avidity distortion[118]. Conversely, our ICeChIP-seq datasets were generated with a native procedure, high sequencing depth, and by filtering out fragments with lengths greater than 200bp to avoid oligonucleosome avidity distortion. Whatever the cause, these differences lead to markedly different interpretations when coupled to readouts of eRNA in the same cell line[58]. We find that neither H3K4me1 (Fig. 2.13C), H3K4me2 (Fig. 2.13D), H3K4me3 (Fig. 2.13E), nor the H3K4me3:H3K4me1 ratio (Fig. 2.13F) is substantially correlated to the transcriptional level of eRNAs. This example highlights the need for ChIP-seq procedures that minimize off-target capture and underscores the pitfalls of treating ENCODE datasets as gold standards for these sorts of analyses.

Examining catalytically dead MLL3/4 mutants with high-specificity antibodies and calibration

To further investigate enhancer biology, we conducted ICeChIP-seq in R1 mouse embryonic stem cells (mESCs) with wild-type (WT) and catalytically dead MLL3/4 mutants (dCD MLL3/4) reported to have markedly reduced H3K4me1 global abundance[153]. Sequencing confirms the high specificity of abMe1-6, both relative to H3K4 methylforms and cross-lysine reactivity (Fig. 2.14A). Globally, we observe that WT H3K4me1 abundance is consistent with other global abundance measurements of this PTM in mESCs[145] and we observe roughly three-fold loss of H3K4me1 in dCD mESCs

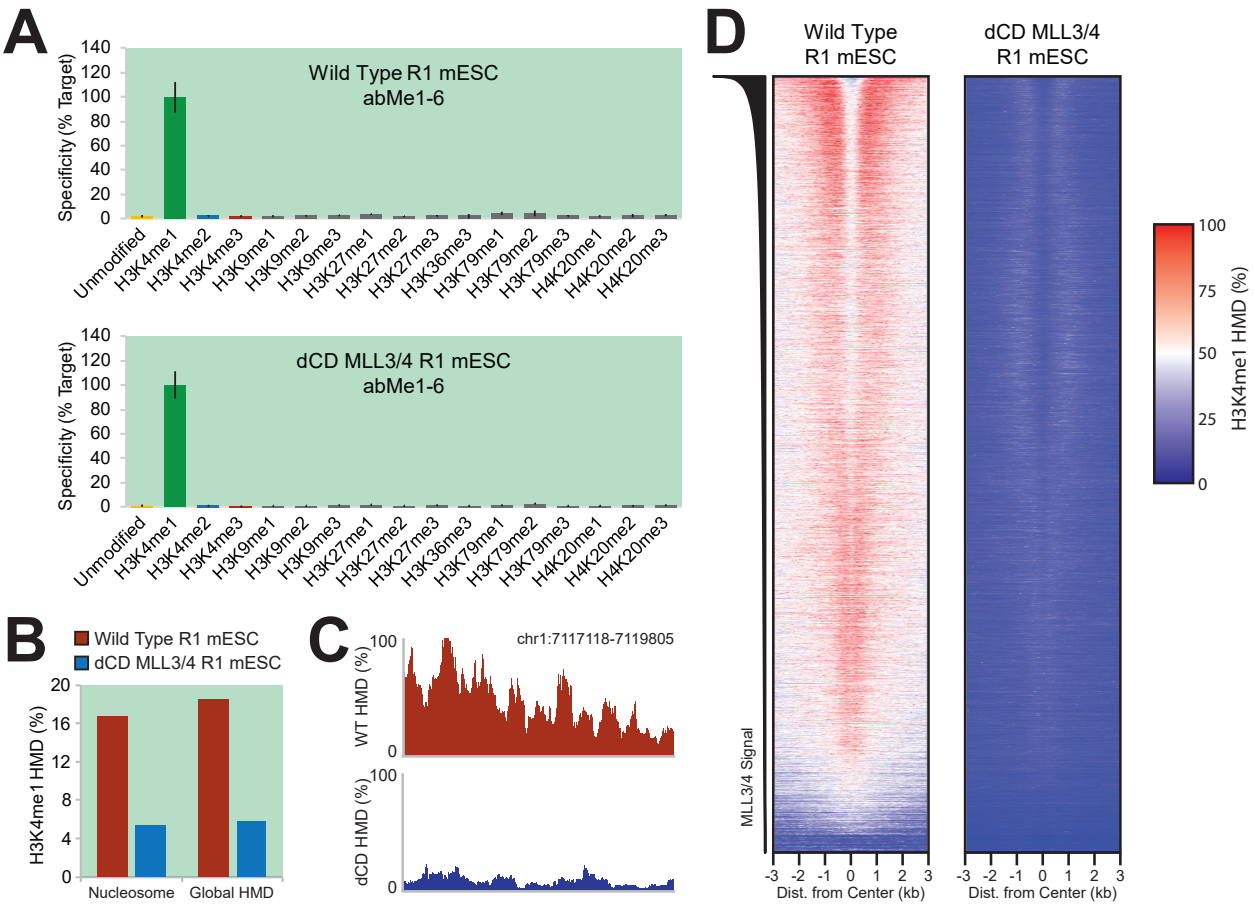


Figure 2.14: Highly specific anti-H3K4me1 ICeChIP-seq can reveal differences between MLL3/4 WT and /catalytically dead cell lines.

(A) Specificity of H3K4me1 ICeChIP-seq in WT and dCD MLL3/4 R1 mESCs. **(B)** Global H3K4me1 abundances, as proportion of nucleosomes (left) and globally integrated HMD (right). **(C)** A representative genome browser view of H3K4me1 HMD in WT and dCD MLL3/4 R1 mESCs near an enhancer[153]. **(D)** Heatmap of H3K4me1 HMD about enhancers in WT and dCD MLL3/4 R1 mESCs, sorted by MLL3/4 ChIP-seq signal[153].

relative to WT, measured either as proportion of nucleosomes or integrated HMD (Fig. 2.14B), confirming that abMe1-6 is specific enough to detect such global abundance differences.

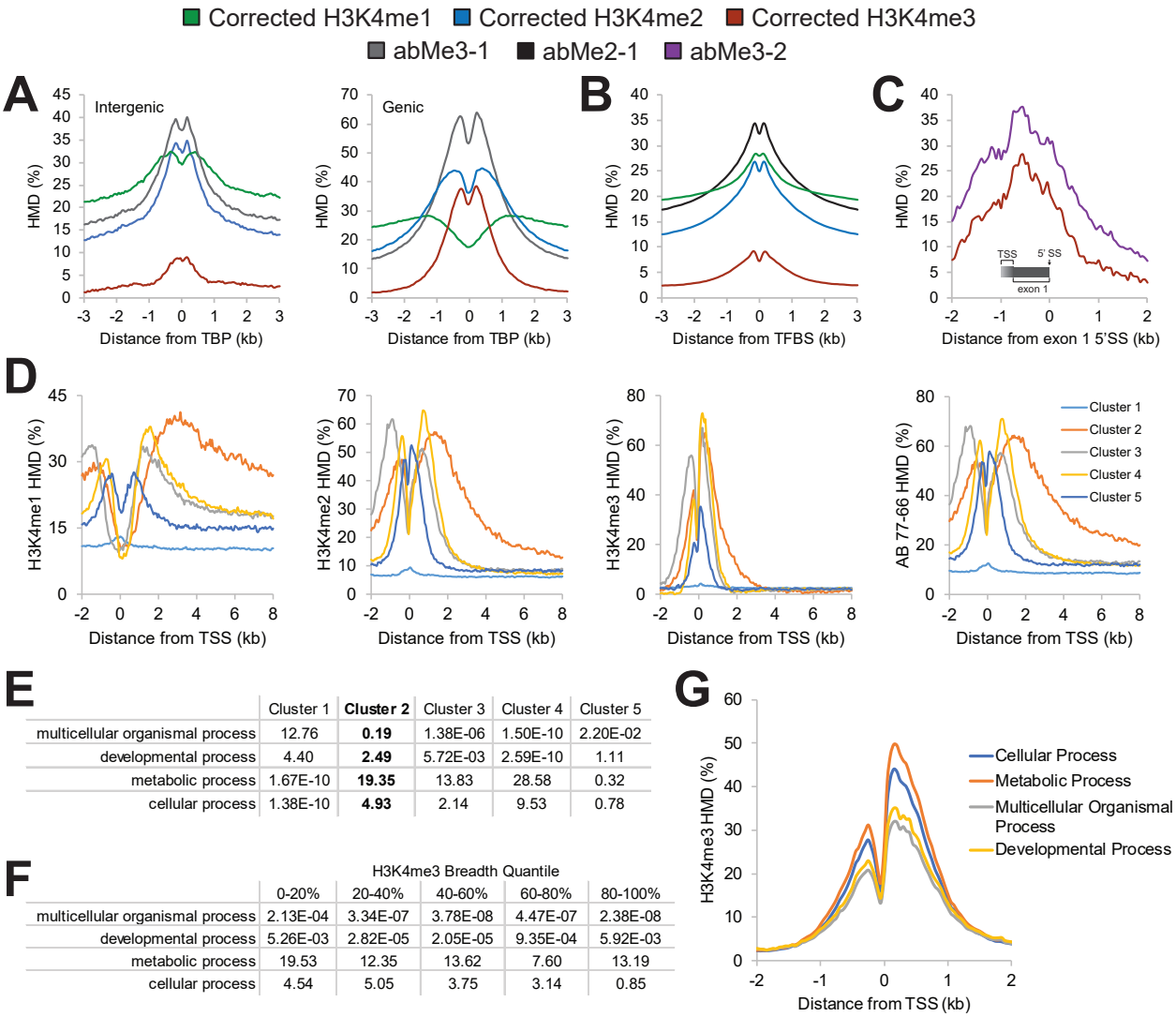
These datasets further serve to highlight the importance of calibration for ChIP-seq. Indeed, ICeChIP-seq genome browser views (Fig. 2.14C) and heatmaps (Fig. 2.14D) of H3K4me1 about enhancer centers for WT and dCD lines show a much more pronounced difference between the two lines than previously reported[153], likely due to inappropriate assumptions inherent in normalization

of uncalibrated data. These discrepancies emphasize the importance both of the absolute quantification offered by ICeChIP and its ability to provide robust quantification amidst to global changes of histone modification abundances, as with these lines.

Reexamining other H3K4 methylform paradigms with high-specificity antibodies

Beyond enhancers, the H3K4 methylforms have been broadly correlated with transcription factor binding. It has been suggested that the H3K4me3 and H3K4me1 profiles are similar in both shape and magnitude between both genic and intergenic TBP sites[154]. Although we cannot confirm the lot of abMe3-1 used in these prior experiments is identical to ours, we recapitulate their results, where the apparent H3K4me3 distribution with abMe3-1 is comparable in shape and magnitude at both genic and intergenic TBP sites (Fig. 2.15A, abMe3-1, grey profile). However, two lots of abMe3-1 show substantial cross-reaction with H3K4me2 (Fig. 2.4, 2.5, 2.10A-B), and its apparent binding profile appears entirely attributable to this methylform. Our H3K4me2 HMD profiles all look quite similar at genic and intergenic TBP sites, whereas the calibrated H3K4me3 distributions are distinct, with little H3K4me3 at the intergenic sites (Fig. 2.15A, red lines). This demonstrates that specificity information within the ChIP experiment is essential for interpretation, as without it, seemingly incorrect conclusions are drawn about the H3K4 methylation state at TBP sites[154].

In addition to enhancers and TFBS, H3K4 methylation is thought to serve biological roles within gene bodies. As an example, there are reportedly two H3K4me3 peaks of comparable magnitude flanking the first exon of genes: the first (canonical) at the TSS[17,20], and the second atop the 5'-spice site that defines the end of the first exon[141]. These observations were based on re-analysis of ENCODE data from K562 cells. Our studies with the same reagent (abMe3-2) indicate its considerable cross-reactivity with H3K4me2 (Fig. 2.4, 2.5, 2.6, 2.11). When we conduct similar



E

	Cluster 1	**Cluster 2**	Cluster 3	Cluster 4	Cluster 5
multicellular organismal process	12.76	**0.19**	1.38E-06	1.50E-10	2.20E-02
developmental process	4.40	**2.49**	5.72E-03	2.59E-10	1.11
metabolic process	1.67E-10	**19.35**	13.83	28.58	0.32
cellular process	1.38E-10	**4.93**	2.14	9.53	0.78

F

	H3K4me3 Breadth Quantile				
	0-20%	20-40%	40-60%	60-80%	80-100%
multicellular organismal process	2.13E-04	3.34E-07	3.78E-08	4.47E-07	2.38E-08
developmental process	5.26E-03	2.82E-05	2.05E-05	9.35E-04	5.92E-03
metabolic process	19.53	12.35	13.62	7.60	13.19
cellular process	4.54	5.05	3.75	3.14	0.85

Figure 2.15: Use of low- vs. high-specificity reagents in the literature may yield demonstrably different biological interpretations for many proposed paradigms.

(A) Apparent HMD profiles of H3K4me1, H3K4me2, H3K4me3 and abMe3-1 about intergenic (left) and genic (right) TATA-binding protein (TBP) sites that have been previously described[154]. **(B)** Apparent HMD profiles of H3K4me1, H3K4me2, H3K4me3, and abMe2-1 about transcription factor binding sites (TFBS) that have been previously described[136]. **(C)** Apparent HMD profiles of H3K4me3 and abMe3-2 about the first exon splice site (SS) for transcripts with a first exon between 750-1000 nucleotides in length. Gradient indicates the region in which the TSS of this set of genes could be. **(D)** Apparent HMD profiles about TSS for H3K4me1, H3K4me2, H3K4me3, and abMe2-1. Clusters were generated using k-means clustering of HMD distribution about TSS[138]. **(E)** -Ln (*p*) of gene ontology enrichment for the clusters profiled in panel (D). **(F)** -Ln (*p*) of gene ontology enrichment for genes by quantile of H3K4me3 peak breadth at said gene. **(G)** Corrected H3K4me3 profiles about TSSs by gene ontology classes in K562 cells.

analyses on genes with the first exon between 750-1000nt (the first length at which the two putative peaks clearly resolve in the original study[141]), we fail to see such a peak at the first exon-intron boundary (Fig. 2.15C), indicative of no H3K4me3 enrichment specific to this splice site. In addition to the above concerns regarding the ENCODE datasets, the previous report used raw H3K4me3 sequencing reads from ENCODE[141], which would not accommodate any differences in nucleosome density at the TSS and the first splice site, whereas ICeChIP (and many conventional ChIP) datasets are normalized to input density and are therefore largely independent of such differences[118]. In this example as well, our high-quality ICeChIP datasets yield different biological interpretations than those proposed in the literature.

Another such example can be found in analysis of the distribution of H3K4me2 over gene bodies. It has been reported that H3K4me2 is highly elevated over the gene body of tissue-specific, immune system process genes in CD4+ T-cells[138]. When we apply the same procedure to identify such genes in K562 cells, we find there is indeed a cluster of genes with somewhat elevated H3K4me2 across the entire gene body (Fig. 2.15D, Cluster 2), though it appears less dramatic and spread-out than may be expected from prior studies. However, we also see that H3K4me1 is more highly spread-out and elevated over this gene class (Fig. 2.15D), reminiscent of their description of the H3K4me2 distribution[138]. We also note that the antibody used by the prior study, abMe2-1, produces results more similar, but not identical results in our analyses (Fig. 2.15D). abMe2-1 displayed some cross-reactivity to H3K4me1 (Fig. 2.4, 2.5, 2.6, and 2.11; abMe2-1), and is likely further compromised by the greater relative abundance (2-10 fold across a variety of cell types) of H3K4me1 over H3K4me2[135]. We also find that this cluster of genes that display the described gene body enrichment profile is, in K562 cells, highly enriched for metabolic processes and not as enriched for cell-type specific processes as previously described (Fig. 2.15E). Thus, while the

differences between our findings and prior reports may be attributable to antibody quality, ChIP procedure, or cell type, the former is likely the most consequential.

Finally, we examined the role of H3K4 methylation domain breadth at gene promoters. It has been proposed that broad H3K4me3 domains mark cell identity genes across a range of cell types, including K562 cells, driving transcriptional constancy[155–157]. To critically assess this phenomenon with our datasets, we analysed the enriched gene ontology classes in K562 cells across different quantiles of H3K4me3 peak breadth. To our surprise, we instead found that metabolic genes were the most enriched class (Fig. 2.15F) and that metabolic processes have, on average, a broader peak structure at TSSs (Fig. 2.15G), suggesting that the proposed role of broad H3K4me3 domains does not apply to K562 cells. As the conclusions in previous publications were largely based on the ENCODE H3K4me3 ChIP-seq tracks, which we have found to be substantially different from our datasets in K562 cells (and indeed, from each other), it is possible that prior interpretations were similarly compromised by antibody quality. All together, these vignettes suggest that in numerous cases, it appears that off-targeting binding by low-specificity antibodies, amongst other factors, has directly led to inaccurate conclusions.

Discussion

Methodological strengths and limitations of peptide arrays and ICeChIP

The largest concern with poor-quality antibodies is that off-target binding will lead to erroneous biological interpretation. In conventional ChIP, with no effective metrics to assess antibody specificity *in situ*, the researcher is effectively blind to this pitfall, potentially compromising their results. Peptide arrays present the only practical way to broadly examine the impact of flanking combinatorial PTMs and have predictive value for other epitope-dense experimental formats, such as

immunoblotting[158]. However, our results suggest that peptide arrays, though commonly used for ChIP antibody validation[108,111,112,133], often fail to accurately reflect antibody performance within ChIP experiments, either for methylform specificity or the impact of combinatorial PTMs. We have begun to examine the physical underpinnings of these differences, but given the distinct experimental formats, they are unlikely reducible to a single concrete principle and in any case, are largely immaterial to the practical matter: that peptide arrays are inappropriate for predicting antibody performance in ChIP.

ICeChIP is not without its limitations. The specificity information afforded by ICeChIP is restricted to the breadth of the semisynthetic nucleosomal standards available. However, these standards, particularly those bearing combinatorial modification patterns, are laborious to construct. If there is a discrepancy between datasets at loci that potentially bear combinatorial modifications, without these additional standards, it is difficult to assess which view is correct. For example, we see modest differences between datasets generated with different highly specific H3K4me3 antibodies (Fig. 2.10H-J) even when the measurement error is reduced by the massive signal averaging implicit in metanalysis. These apparent differences are attributable to several possible sources: differential sensitivity to flanking modifications (either increasing affinity, thereby artifactually inflating the HMD, or the converse); differential off-target nucleosome capture of marks not represented in the panel of nucleosomal standards deployed; and for individual loci, input and IP sampling error can also drive more pronounced peak shape and height differences.

Further, even if a broad range of nucleosome standards bearing combinatorial modifications were constructed, the analysis of histone modifications at co-modified loci would not be straightforward. It is possible, for example, that at a given locus, there are two sub-populations of cells with different PTM states that the two PTMs do not actually coexist on the same nucleosomes. To

evaluate this possibility, a sequential ICeChIP protocol would need to be developed, sequentially selecting for nucleosomes with each PTM. However, to date, the sequential ChIP protocols in the literature[9,130,159–162] have tended towards denaturative, crosslinked protocols with the questionable specificity and IP enrichment inherent in crosslinked ChIP[149–151]. To this end, sequential native ICeChIP remains an active area of study for us, but nonetheless, represents a present limitation of the method.

Beyond combinatorial modifications, ICeChIP is limited in its ability to accurately assess nucleosome-depleted regions. At such regions, input coverage is sparse, leading to low sampling and high uncertainty in HMD values. Though in principle this could be addressed by higher sequencing depth, the relevance of the histone modification density at locations with such low nucleosome occupancy would be questionable. Additionally, ICeChIP assumes that native nucleosomes are stable enough to survive the ChIP protocol, but it has been previously observed certain histone variants and modifications may reduce nucleosome stability[163,164]. If these nucleosomes are unstable during the ChIP experiment, then that may result in artifactually reduced representation in the IP, whereas the DNA will still exist in the input, resulting in deflated apparent HMD.

In this study, we reduced the impact of variability of input preparation, cellular heterogeneity and authentic biological differences between samples by performing the bulk of comparative immunoprecipitations side-by-side from the same pool of input. In other contexts, these factors could become significant contributors to apparent signal.

We often use signal correction in order to more effectively isolate on-target signal from the antibody-measured signal, which is a convolution of on- and off-target binding. Yet such signal correction is not strictly necessary. Indeed, because signal correction uses multiple antibodies to compute a given modification track, the track will often have greater uncertainty than ICeChIP-seq

with a single antibody. We use signal correction for making more nuanced and accurate comparisons in the aggregate, where many loci are being treated and analyzed as one dataset. In these analyses, the error is reduced by averaging. However, when examining individual loci, where the error in a signal-corrected track is more substantial, it may be better to use a single high-quality antibody for the most accurate view of mark distributions.

It is important to note that these limitations also exist with uncalibrated ChIP. However, in that approach, the researcher is completely blind to the questions of specificity and accurate quantification, whereas ICeChIP at least offers some information to that end. Despite its limitations, ICeChIP represents a powerful tool to enable quantitative studies of histone PTMs.

Discrepancies with the literature due to antibody and ChIP quality

Here, we have used our ICeChIP datasets to critically re-examine ENCODE project datasets and other H3K4-methylform paradigms related to transcriptional control. As disagreement between our data and prior literature could reflect cell-type specific differences, we have focused on findings proposed as general features of mammalian chromatin. The examples we have presented here comment on the role of antibodies and ChIP-seq procedures generally in the widely publicized biological "reproducibility crisis"[165]. For a variety of potential reasons, particularly antibody specificity, several of the interpretations currently in the literature are not recapitulated by the high-quality ICeChIP datasets we have produced herein, casting some doubt on the many thousands of existing datasets that currently exist for histone PTMs across a wide range of organisms and cell types, and their use to draw a great many biological conclusions. In several instances, we were able to reproduce the phenomena reported with our K562 ICeChIP datasets using the same antibody catalogue numbers.

However, in each of these cases, the precise interpretation was flawed owing to off-target antibody capture, which the authors could not have known at the time due to inadequate validation criteria.

This set of discrepancies makes a powerful argument for *in situ* metrics of antibody specificity within ChIP experiments as distinct from spike-in normalization for the purposes of comparison[98,118,128]. It is unfortunately commonplace for authors to omit the specific antibody lot numbers used, but if distinctions between our data and the literature arise from lot-to-lot variation[134] this is equally troubling with regard to the scientific reproducibility crisis[165].

Although it is impractical to perform similar analyses of the thousands of papers in the literature that have used the antibodies described here in ChIP experiments, we fear that what we have discovered for a small selection of H3K4 methylation paradigms may represent a larger problem for the field. Furthermore, while we focus here on the specificity problems for antibodies raised to H3K4-methylforms, our ongoing (and comparably extensive) studies of other "PTM-specific" antibodies show similar promiscuity issues (data not shown), and a dose of skepticism for precise conclusions drawn from uncalibrated ChIP with many of these reagents is similarly warranted.

Our results strongly indicate that the field needs to establish and adopt more rigorous quality control standards for ChIP reagents to ensure more robust and reproducible data in the future. Crucially, this includes more careful validation of ChIP antibodies, ideally by direct testing to panels of related internal nucleosome standards that encompass the broadest achievable range of possible cross-reactivities in a ChIP setting[118]. Apart from calibration, we propose that the norms of ChIP-seq data publication should include clear indication of antibody catalogue and lot numbers, sequencing of input, and quantitative analysis rather than use of called peaks, which reduces quantitative data to a mere binary. Different protocols can also affect the specificity of the ChIP-seq experiment, and

though ICeChIP effectively accommodates for this variation[118], we also suggest the use of native ChIP rather than the oftentimes far more noisy, low-efficiency, idiosyncratic and artefact-prone cross-linked ChIP with sonication for accessible histone tails[149–151]. All told, our study demonstrates both the danger of using unvalidated antibodies in ChIP and the power of calibrated ChIP to robustly measure histone PTMs and drive new biological discovery.

Acknowledgements

We wish to thank P. Faber and H. Whitehurst in the University of Chicago Functional Genomics Facility for Illumina sequencing. We also thank the following for donating reagents used in this study: Abclonal (six antibodies), Active Motif (six antibodies), Cell Signaling Technology (three antibodies), and Diagenode (three antibodies). We also thank Dr. Chuck Epstein of the Broad Institute for his generous gift of five antibodies used by the ENCODE Consortium. We further thank Dr. Joanna Wysocka of Stanford University for her generous gift of mESC cell lines.

This study was supported by the National Institutes of Health under award numbers R44-HG008907 to Zu-Wen Sun and Michael-Christopher Keogh (EpiCypher, Inc.), R00-CA181343 and R35-GM124736 to Scott B. Rothbart (Van Andel Institute), and R01-GM115945 to Alexander J. Ruthenburg (University of Chicago).

Methods and Materials

Cell Culture

K562 cell lines were grown at 37°C with 5% CO_2 and 95% humidity in Dulbecco's Modified Eagle Media (DMEM, Gibco; K562 cells only) with 10% (v/v) HyClone FBS Characterized U.S. and

1x Penicillin/Streptomycin (Gibco). Cells were seeded into vented flasks to a density of 200,000 cells/mL of culture and were passaged at 1-2 million cells/mL of culture.

R1 wild-type (WT) and MLL3/4 knockout mESC lines were cultured by the Wysocka Lab as previously described[153] and were generously provided as cell pellets.

Octamer Reconstitution

Symmetric H3K4me3, H3K9me3, H3K27me3, H3K36me3, H3K79me2, and H3K4me3K27me3 octamers were reconstituted from semisynthetic histones as previously described[91,118,166,167]. Recombinant core histones were expressed in BL21 (DE3) with pRARE2 and mixed to equimolarity with the relevant semisynthetic histones in freshly prepared filter sterilized Unfolding Buffer (50 mM Tris-HCl pH 8.0, 6.3 M Guanidine-HCl, 10 mM 2-mercaptoethanol, 4 mM EDTA) to a final concentration of ≥ 1 mg histone per mL. The histone reconstitution was then added to 3500 MWCO SnakeSkin dialysis tubing (Pierce) and dialyzed overnight at 4°C against 500-1000 volumes of filter sterilized Refolding Buffer (20 mM Tris-HCl pH 7.5, 2 M NaCl, 5 mM DTT, 1 mM EDTA).

After dialysis, the histone mixture was centrifuged at 18,000 g for 1 hour at 4°C, and subjected to gel filtration chromatography (Superdex 200 10/300 GL, GE Healthcare, resolved with Refolding Buffer). Each fraction that displayed a peak on the UV chromatogram was analysed by SDS-PAGE (22 mA current in 1x Laemmli Buffer for 70 minutes), stained with SYPRO Ruby (Bio-Rad) per manufacturer instructions, and imaged with a 610BP emission filter at 600V PMT setting. Octamer fractions with equimolar quantities of each core histone were pooled and concentrated (Amicon Ultra-4 Centrifugal Filters, 10,000 MWCO, Millipore) to 5-15 μM octamer, diluted with one volume of Octamer Storage Buffer, and stored at -20°C.

All other octamers were obtained from EpiCypher, Inc.

Nucleosome Reconstitution

Nucleosomes were reconstituted onto 147bp DNAs composed of the core Widom 601 sequence[168] modified with a 22bp barcode on each end, with each barcode composed of two distinct 11bp sequences not found in the human or mouse genomes. The DNA and octamer were mixed to a final concentration of 1µM each in 2 M NaCl, and then dialyzed in dialysis buttons (Hampton Research) and a 10,000 MWCO SnakeSkin dialysis membrane (Pierce) against 200 mL of Refolding buffer for 10 minutes. Dialysis then continued as 2L of Buffer I0 (20 mM Tris-HCl pH 7.5, 1 mM EDTA, 1mM DTT) was added (flow rate 2-2.5 mL per minute).

Dialyzed samples were diluted with an equal volume of Nucleosome Dilution Buffer (20 mM Sodium Cacodylate pH 7.5, 10% v/v glycerol, 1 mM EDTA, 10 mM 2-mercaptoethanol, Filter Sterilized), and 1 µl was analysed by native PAGE (100 V in 1x TBE for 30 minutes), stained with SYBR Gold in 1xTBE for one hour, and visualized with a UV transilluminator gel imager. Fractions containing nucleosomes and minimal free DNA were pooled and diluted to a working concentration of ~ 1 nM with filter sterilized Nucleosome Storage Buffer (10 mM Sodium Cacodylate pH 7.5, 100 mM NaCl, 50% v/v glycerol, 1 mM EDTA, 1x Protease Inhibitor Cocktail [1 mM PMSF, 1mM ABESF, 0.8 µM aprotinin, 20 µM leupeptin, 15 µM pepstatin A, 40 µM bestatin, 15 µM E-64 from a 200x DMSO stock]) and stored at -20°C.

Peptide Microarrays

Peptide microarray experiments were conducted by the Rothbart Lab. Peptide microarrays were fabricated using an Aushon 2470 microarrayer and used as described[110,112]. Briefly, antibodies were diluted according to the manufacturers recommended western blot concentration (unless otherwise indicated) in Array Hybridization Buffer (PBS [137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4,

1.8 mM KH_2PO_4, pH 7.6], 5% BSA, 0.1% Tween-20), and 500 µL (5 µL for 48-well format) was hybridized onto a peptide microarray for 1 hour at 4°C. Slides were washed in PBS and probed with a fluorescently labelled secondary antibody (Life Technologies A-21244 or A-21235). Microarrays were scanned using an Innopsys InnoScan 110AL microarray scanner and analysed using ArrayN-inja[169]. Specificity was calculated as described below for ICeChIP data from the raw fluorescent signal.

ICeChIP

ICeChIP was performed as previously described[118,124,170,171]. Briefly, cell pellets were washed twice with 5 mL of PBS, then washed twice with 5 ml of filter sterilized Buffer N (15 mM Tris-HCl pH 7.5, 15 mM NaCl, 60 mM KCl, 8.5% w/v Sucrose, 5 mM $MgCl_2$, 1 mM $CaCl_2$, 1 mM DTT, 200 µM PMSF, 50 µg/mL BSA, 1x Roche Protease Inhibitor Cocktail), with each wash consisting of complete resuspension of the pellet, centrifugation at 500 g for 5 minutes at 4°C, and removal of supernatant. The washed pellet was then resuspended in at least 2 packed cell volumes (PCV) of Buffer N and mixed with 1 volume of 2x Lysis Buffer (Buffer N supplemented with 0.6% NP-40 Substitute) and incubated on ice for 10 minutes to lyse cells.

The crude nuclei were spun down at 500 g for 5 minutes at 4°C before being resuspended in at least 6 packed nuclear volumes (PNV) of Buffer N and applied to the top of 7.5 mL of filter sterilized Sucrose Cushion N (15 mM Tris-HCl pH 7.5, 15 mM NaCl, 60 mM KCl, 30% w/v Sucrose, 5 mM $MgCl_2$, 1 mM $CaCl_2$, 1 mM DTT, 200 µM PMSF, 50 µg/mL BSA, 1x Roche Protease Inhibitor Cocktail) in a 15 ml centrifuge tube, then spun down at 500 g for 12 minutes at 4°C in a swinging-bucket rotor. The supernatant was discarded, and the pellet resuspended in ~ 2 PNV of Buffer N.

The nucleic acid content of the nuclei per unit volume was quantified by diluting 2 µL of nuclei suspension into 48 µL of 2 M NaCl, water-bath sonicating to solubilize DNA, and spectroscopically measuring nucleic acid concentration by Nanodrop (where one A_{280nm} = 50 ng/µL chromatin). After accounting for the 25-fold dilution of the measurement sample, the concentration of the nuclei was adjusted to 1 µg/µL of chromatin. Nuclei were dispensed to 100 µL aliquots, flash frozen, and stored at -80°C prior to use.

For use, nuclei aliquots were thawed and spiked with ~ 1 µl of each barcoded nucleosome standard per 50 µg of chromatin. This suspension was then mixed by pipette, transferred to a new tube, and warmed to 37°C for 2 minutes. 1 unit of micrococcal nuclease (MNase, Worthington) per 4.375 µg of chromatin was added, and samples incubated at 37°C while shaking at 900 rpm for 12 minutes. Digestions were stopped by adding 1/9 volume of filter sterilized 10x MNase Stop Buffer (100 mM EDTA, 100 mM EGTA) while slowly vortexing, and nuclei lysed by adding 5 M NaCl to a final concentration of 600 mM while slowly vortexing. 66 mg of HAP resin (BioRad, CHT™ Ceramic Hydroxyapatite, Type I, 20 um) per 100 µg of chromatin digested was rehydrated with 200 µl of filter sterilized HAP Buffer 1 (5 mM Sodium Phosphate pH 7.2, 600 mM NaCl, 1 mM EDTA, 200 µM PMSF) per 100 µg of chromatin digested. Lysed nuclei were centrifuged at 18,000 g for 1 minute to pellet insoluble nuclear debris, and the soluble fraction added to the rehydrated HAP resin and incubated for 10 minutes at 4°C with rotation.

After incubation, the HAP resin slurry was added to a centrifugal filter unit (Millipore Ultrafree MC–HV Centrifugal Filter 0.45 µm) and spun at 1000 g for 30 seconds at 4°C. The HAP resin left on the filter unit was then washed 4 times with 200 µL HAP Buffer 1, and 4 times with 200 µl filter sterilized HAP Buffer 2 (5 mM Sodium Phosphate pH 7.2, 100 mM NaCl, 1 mM EDTA, 200 µM PMSF) by spinning at 1000 g for 30 seconds at 4°C. HAP resin was eluted into a clean

tube with three 100 µl solutions of filter sterilized HAP Elution Buffe r(500 mM Sodium Phosphate pH 7.2, 100 mM NaCl, 1 mM EDTA, 200 µM PMSF). The nucleic acid content of the elution was then quantified by Nanodrop.

Antibodies and quantities used for each ICeChIP experiment are shown in Appendix A. With the exception of the 304M3B and 309M3B antibodies, the indicated amount of Protein A Dynabeads (Invitrogen) for each ICeChIP was washed with 50 µL of ChIP ChIP Buffer 1 by use of a magnetic rack, then resuspended in 50 µL of ChIP Buffer 1. In a separate set of tubes, the antibody was diluted to 100 µL with ChIP Buffer 1. The antibody and Protein A Dynabead suspensions were combined and incubated on a rotator at 4°C for at least one hour, then washed with 200 µL of ChIP Buffer 1 by use of a magnetic rack and resuspended in 50 µL of ChIP Buffer 1 (25 mM Tris pH 7.5, 5 mM $MgCl_2$, 100 mM KCl, 10% v/v glycerol, 0.1% v/v NP-40 Substitute, 50 µg/ml of BSA).

The antibodies 304M3B and 309M3B were prepared similarly with Streptavidin M-280 Dynabeads (Invitrogen) rather than Protein A Dynabeads. The beads were washed and antibodies added and incubated as above. After incubation, the beads were washed twice with 200 µL of ChIP Buffer 1 by use of a magnetic rack. They were then washed twice with 200 µL of ChIP Buffer 1 supplemented with 5 µM of biotin by incubating for 10 minutes at 4°C on a rotator, then removing supernatant by use of a magnetic rack.

After antibodies were prepared and washed, the input chromatin concentration adjusted to 20 ng/µl with filter sterilized ChIP Buffer 1, and the amount of chromatin specified in Appendix A was added to each antibody-bead conjugate and incubated for 15 minutes on a rotator at 4°C. Beads were then washed twice with filter sterilized ChIP Buffer 2 (25 mM Tris pH 7.5, 5 mM $MgCl_2$, 300 mM KCl, 10% v/v glycerol, 0.1% v/v NP-40 Substitute) and once with filter sterilized ChIP Buffer 3 (10 mM Tris pH 7.5, 250 mM LiCl, 1 mM EDTA, 0.5% Sodium Deoxycholate, 0.5% v/v NP-40

Substitute), with a wash consisting of removal of the existing supernatant by use of a magnetic rack, resuspension into 150 µl of buffer, transfer to a new siliconized tube, and incubation on the rotator for 10 minutes at 4°C. After these washes, the supernatant was removed, the beads resuspended in ChIP Buffer 1, transferred to a new siliconized tube, rinsed once with 200 µl of TE before being resuspended in 50 µl of ChIP Elution Buffer (50 mM Tris pH 7.5, 1 mM EDTA, 1% w/v SDS, Filter Sterilized) and incubated at 55°C for 5 minutes.

After incubation, the supernatant was transferred to a new set of siliconized tubes, and the beads discarded. To each supernatant was then added 2 µl of 5 M NaCl, 1 µl of 500 mM EDTA, and 1 µl of 10 mg/mL Proteinase K. 15 µl of Input DNA was also diluted to 50 µl with 35 µl of ChIP Elution Buffer and was supplemented with 2 µL of 5 M NaCl, 1 µL of 500 mM EDTA, and 1 µL of 10 mg/mL Proteinase K. The IP elutions and diluted input were then incubated at 55°C for 2 hours for a Proteinase K digestion. After digestion, the DNA was purified by adding 1.5 volumes of Serapure HD (1:50 dilution of Sera-Mag SpeedBeads [Fisher], 20% PEG-8000, 2.5 M NaCl, 10 mM Tris pH 7.5, 1 mM EDTA, 0.05% Tween-20, Filter Sterilized prior to addition of SpeedBeads), incubating at room temperature for 15 minutes, then collecting the beads on a magnetic rack, washing twice with 150 µl of 70% ethanol, and eluting into 50 µl ddH2O, which was then recovered and stored at -20°C.

DNA Quantification and Analysis by Quantitative PCR

To assess local histone modification density and/or antibody specificity, our DNA from the ChIP experiments was quantified by quantitative PCR (qPCR). qPCR was conducted using TaqMan Gene Expression Master Mix (Applied Biosystems) using the primers and hydrolysis probes previously described[118]. These primers and probe for the barcoded sequences were previously qPCR validated

for effectiveness and quality[118]. Primers were used at 900 nM; hydrolysis probe at 250 nM, in the TaqMan Gene Expression Master Mix (Applied Biosystems). The qPCR program was run at 95°C for 10 minutes, followed by 40 cycles, each consisting of 15 seconds at 95°C followed by 1 minute at 60°C and concluding with a plate read.

Cq values were analysed using the ΔΔCq method. Briefly, the Cq values for each target for each sample were averaged together to obtain the mean Cq value. Enrichment for each barcode was then computed as Enrichment $= 2^{Cq_{INPUT} - Cq_{IP}} * 10$, accounting for the 10-fold dilution of Input relative to IP and multiplying by 100% for Enrichment as a percentage of target. Off-target binding to alternate PTMs were computed by normalizing each enrichment to that of the on-target PTM: referred to as "Specificity (% Target)". For H3K4 methylform specificity analyses, overall specificity was computed by dividing the enrichment of the target PTM by the sum of the enrichments for all H3K4 methylforms (i.e. H3K4me0 + H3K4me1 + H3K4me2 + H3K4me3); this is referred to as "Aggregate Specificity."

Illumina Library Preparation and Sequencing

Illumina libraries were prepared as described[118], with minor modifications. Briefly, Serapure purified DNA was quantified using Quant-iT^{TM} PicoGreen (Thermo Fisher) as per manufacturer instructions. Libraries were then generated from up to 10 ng of each DNA sample (input or IP) with the NEBNext Ultra II DNA Library Prep kit (New England Biolabs) per manufacturer instructions. The DNA content of each library was then quantified and pooled for Illumina sequencing. Cluster generation and paired-end sequencing was conducted using standard Illumina protocols by the University of Chicago Genomics Facility on the Illumina NextSeq. One replicate of each antibody was sequenced.

Bioinformatic Analyses

To align reads, a reference genome was first created, which consisted of the either human genome (GRCh38/hg38) or the mouse genome (mm9) appended respectively by the sequences of each of the nucleosome standard barcodes. Reads were then mapped to the appropriate reference genome using Bowtie2 using the sensitive pre-set and end-to-end alignment options[172]. Using SAMTools[173], any reads which were not paired, not mapped in a proper pair, or mapped with a map quality < 20 were discarded to prevent low-quality reads from impacting downstream analyses. Reads were then flattened to create a single mapping from each matched pair of reads by retaining only one fragment per pair, and any mappings with lengths > 200bp were also discarded to ensure only mononucleosomes were being analyzed[118].

Bedgraphs of genome coverage were then generated using BEDTools[174], and IP / input genome coverage bedgraphs were merged using BEDTools[174]. The sum of reads across ladder members for each nucleosomal standard was computed for each sample and HMD bedgraphs were then generated from the merged bedgraphs using awk to apply the following formula:

$$\text{HMD } (\%) = 100\% * \frac{\text{IP}_{\text{locus}}/\text{Input}_{\text{locus}}}{\text{IP}_{\text{barcode}}/\text{Input}_{\text{barcode}}}$$

Error and 95% confidence intervals were computed with Poisson statistics and error propagation from the merged bedgraphs using awk to apply the following formula:

$$\text{95CI Error } (\%) = 1.96 * \text{HMD } (\%) * \sqrt{\frac{1}{\text{IP}_{\text{locus}}} + \frac{1}{\text{Input}_{\text{locus}}}}$$

Bigwig files were generated for visualization using the bedGraphToBigWig tool[175].

Correction was conducted using the antibodies AB 8895 (abMe1-1), AB 7766 (abMe2-1), AB 12209 (abMe3-4), and AB 8580 (abMe3-1), unless otherwise noted. Correction was done using our previously described method[118] against H3K4me1, H3K4me2, H3K4me3, and H4K20me3

off-target binding. Briefly, measured HMD by each antibody can be described by a vector *M*, and the measured specificities by each antibody described by a square matrix *S*. Then, we can state, if other off-target binding is negligible, that the correct HMDs for H3K4me1, H3K4me2, H3K4me3, and H4K20me3 can be expressed by the vector *C* such that $M=CS$. As such, the vector *C* can be computed as $CSS^{-1} = C = MS^{-1}$. The elements of S^{-1} were then used to compute the HMD and Error of the corrected profiles using awk to linearly combine the AB 8895 (abMe1-1), AB 7766 (abMe2-1), AB 12209 (abMe3-4), and AB 8580 (abMe3-1) profiles.

Peak calling was conducted for all H3K4 methylation antibodies using Macs2 using the bdgpeakcall command[176], with the input being the HMD bedgraphs computed for each sample. To compute average HMD across a series of intervals, a "double mapping" procedure was used. First, the HMD bedgraph was mapped onto 1bp windows made for each interval using BEDTools[174]. Then, the mapped windows were mapped onto the original intervals using BEDTools[174]. This procedure ensured that the degree of overlap of the interval with each value of the HMD bedgraph was accounted for in the mapping procedure. Using this double-mapping procedure, the average HMD and average 95% CI Error of each called peak was computed. At this point, those peaks with greater average HMD than average 95% CI Error were selected as "high-confidence" peaks. All subsequent peak analyses were conducted with these "high-confidence" peaks. For the H3K9me3 antibody, peak calling was conducted using Macs2 using the bdgbroadcall command[176], with the input being the HMD bedgraph. These peaks were treated as the H3K9me3 broad peaks. Peak HMD correlations, were conducted by computing average HMD as measured by antibody and corrected profile across antibody-measured peaks and subsequently correlating these computed average HMDs using R, forcing through origin. Stringently defined enhancers were defined as those that are not overlapping with a Refseq promoter and have a transcription factor binding site[136],

GRO-Cap TSS[58], ATAC-seq peak[177], and ENCODE H3K27ac peak, FAIRE-seq peak, DNase HS site, and P300 peak[127]) which make contact with at least one promoter by pol II ChIA-PET[147].

For Fourier analyses, the 1200bp region centered upon each peak centre was sectioned into eight 150bp windows using BEDTools[174]. For each window, the average HMD as measured by the antibodies or corrected profile to be used, depending on the analysis employed, was computed as above. The eight windows were then assembled into eight-element vectors for each peak interval, and the pretrigonometric factors of a Fourier Discrete Cosine Transform computed on these vectors using Mathematica 10.2 with the command FourierDCT. The pretrigonometric factors were then correlated using R for each of the eight components, forcing through origin.

Profiles of HMD distributions about features including transcription start sites, first exons, and TBP sites were generated using HOMER annotatePeaks[178]. Gene ontology was conducted using HOMER findGO[178]. Gene ontology terms were largely classed into the overarching PANTHER GOSlim terms[179].

Integrated genome-wide HMD was computed by computing average of HMD across all base-pairs in genome. Nucleosome global modification abundance was computed as ratio of total genomic IP to input reads divided by ratio of barcode IP to input reads, much like computation of locus-specific HMD. The integrated genome-wide HMD represents the proportion of the genome that has the modification of interest; the nucleosome global modification abundance represents the proportion of nucleosomes bearing the modification of interest. These two would be equivalent if nucleosomes were uniformly distributed about the genome but are otherwise not necessarily equivalent.

Statistical details of experiments can be found in the relevant figure legends. Linear correlations with R were forced through origin for more appropriate slope comparison.

Data and Software Availability

The ICeChIP-seq datasets generated in this study have been deposited in the Gene Expression Omnibus under accession number GSE103543.

CHAPTER 3: RETHINKING THE ROLE OF NUCLEOSOMAL BIVALENCY IN EARLY DIFFERENTIATION

Attributions

This chapter has been adapted from: Shah, R. N. *et al.* Re-evaluating the role of nucleosomal bivalency in early development. Preprint at *bioRxiv*, doi: 10.1101/2021.09.09.458948. (2021). Asymmetric disulfide-linked H3K4me3-H3K27me3 were synthesized and provided by the Fierz Laboratory at École polytechnique fédérale de Lausanne, Switzerland. The 304M3B-1xHRV3C antibody was developed by the Koide Lab at New York University with Adrian Grzybowski, PhD'18. Dr. Grzybowski also developed the reICeChIP method, conducted reICeChIP-seq on naïve mouse embryonic stem cells, and conducted methyltransferase assays. Jimmy Elias cultured primed mouse embryonic stem cells and neuronal precursor cells. The other experiments and analyses were conducted by the author.

Abstract

Nucleosomes, composed of DNA and histone proteins, represent the fundamental repeating unit of the eukaryotic genome; posttranslational modifications of these histone proteins influence the activity of the associated genomic regions to regulate cell identity. Traditionally, trimethylation of histone 3-lysine 4 (H3K4me3) is associated with transcriptional initiation, whereas trimethylation of H3K27 (H3K27me3) is considered transcriptionally repressive. The apparent juxtaposition of these opposing marks, termed "bivalent domains", was proposed to specifically demarcate of small set transcriptionally poised lineage-commitment genes that resolve to one constituent modification through differentiation, thereby determining transcriptional status. Since then, many thousands of studies have canonized the bivalency model as a chromatin hallmark of development in many cell

types. However, these conclusions are largely based on chromatin immunoprecipitations (ChIP) with significant methodological problems hampering their interpretation. Absent direct quantitative measurements, it has been difficult to evaluate the strength of the bivalency model. Here, we present reICeChIP, a calibrated sequential ChIP method to quantitatively measure H3K4me3/H3K27me3 bivalency genome-wide, addressing the limitations of prior measurements. With reICeChIP, we profile bivalency through the differentiation paradigm that first established this model[9,130]: from naïve mouse embryonic stem cells (mESCs) into neuronal progenitor cells (NPCs). Our results cast doubt on every aspect of the bivalency model; in this context, we find that bivalency is widespread, does not resolve with differentiation, and is neither sensitive nor specific for identifying poised developmental genes or gene expression status more broadly. Our findings caution against interpreting bivalent domains as specific markers of developmentally poised genes.

Introduction

H3K4me3 is canonically considered to be a marker of active transcription[17,18,20,92–94], whereas H3K27me3 is thought to be a transcriptional repressor[11,180–183]. In its original conception, the bivalency model posits that the combination of H3K4me3 and H3K27me3 (or a so-called "bivalent domain"[9,130,184]) represents a specific regulatory marker of developmentally staged genes. Specifically, lineage commitment genes are thought to be held in a low-expression, transcriptionally "poised" state by promoter nucleosomes bearing both H3K4me3 and H3K27me3[9,130,185,186]. Upon differentiation, the bivalent domain "resolves" into a monovalent state, and the associated gene is either transcriptionally activated or terminally repressed if H3K27me3 or H3K4me3 is lost, respectively[9,130,185–188]. The elegance of this instructive model inspired a host of follow-on studies that

have suggested that bivalency is important in differentiation[160,189–195], embryogenesis[14,15,184,196,197], genome architecture[185,198–201], and oncogenesis[25,202–205].

In the absence of unambiguous biochemical or functional validation[184,206,207], these studies have largely relied upon ChIP, with the vast majority of studies defining loci with independent ChIP enrichment for H3K4me3 and H3K27me3 as bivalent domains. However, this analysis cannot distinguish whether the two modifications coexist or represent two distinctly marked subpopulations of alleles or cells. Further, because different ChIPs are normalized separately, they exist on separate relative scales and cannot be quantitatively compared without internal calibration[118,124,171]. As such, it is impossible to quantify the extent of bivalency at a given locus or to measure its changes through differentiation.

To address the first problem, several studies have used sequential ChIP[9,145,159,160,162], measuring coexistence by using the eluent of an IP against H3K4me3 as the substrate for an IP against H3K27me3 (or vice versa). However, these experiments were uncalibrated, were often undersampled[119,162], and used antibodies of unknown specificity[9,145,159,160,162], precluding quantification of the extent of modification. Moreover, many used relatively large chromatin fragments in their pulldowns, making it difficult to determine whether modifications coexisted on one nucleosome or discretely marked neighbouring nucleosomes[9,159,160,162]. The limitations of these sequential ChIP studies preclude accurate assessment of key properties of bivalency.

Our previous work introduced internally calibrated ChIP (ICeChIP), in which barcoded nucleosome internal standards are used to measure antibody specificity and as analytical calibrants that enable computation of the histone modification density (HMD), or the proportion of nucleosomes at a given locus with the modification of interest[118,124,171]. By identifying regions with high H3K4me3 and H3K27me3, we indirectly identified many promoters with a nonzero amount of bivalency, in-

cluding those regulating developmental and metabolic genes[118]. However, this analysis was limited; it was not sensitive for bivalency at less extensively modified loci, nor could it quantify the extent of bivalency. Here, we directly quantify this nucleosomal mark pattern by calibration of a modified sequential ChIP approach to critically evaluate the bivalency model in the differentiation system in which the foundational observations were made.

Results

Measuring bivalency with reICeChIP

To directly measure bivalency and evaluate its role in differentiation, we first attempted to deploy our calibrants with published sequential ChIP methods. However, when evaluated with internal standards, these methods[9,145,160] displayed extremely low enrichment and variable specificity (Fig. 3.1A), with common elution methods either failing to release most of the captured material[159] or compromising the specificity of the second IP (Fig. 3.1B-C). With such heavy losses, we became concerned that we would undersample and potentially bias the measurement of bivalent nucleosomes. We sought a method of elution from the primary IP that was both more efficient and would preserve nucleosome integrity for the second IP. To that end, we modified a recombinant biotinylated Fab (304M3-B) specific for H3K4me3[208] with an intervening HRV 3C endoprotease cleavage site to enable quantitative elution by enzymatic cleavage under mild conditions.

We then leveraged this reagent to develop reICeChIP (Fig. 3.2A). The first pulldown was conducted with the cleavable α-H3K4me3 Fab from native mononucleosomes[118,170] spiked with nucleosome internal standards. We then eluted the captured nucleosomes from streptavidin resin by cleaving the antibody with HRV 3C endoprotease[209] and, with this eluent, conducted a second pulldown against H3K27me3 with a conventional antibody. This method eluted material from the

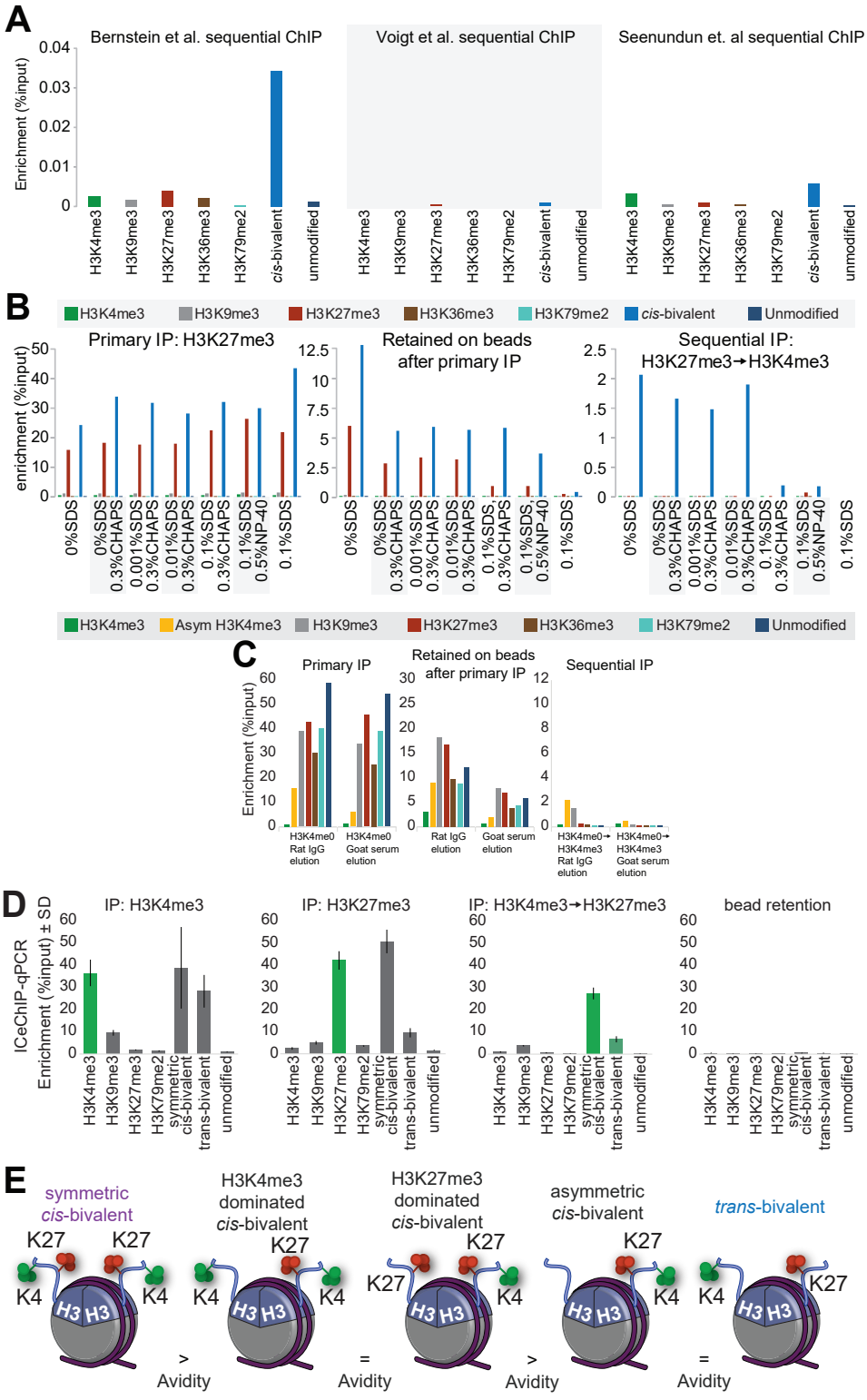


Figure 3.1: Evaluation of sequential ChIP methods.

Figure 3.1, continued:
(A) Enrichment of on- and off-target nucleosome standards under sequential ChIP protocols developed by Bernstein et al.[9], Voigt et al.[145], and Seenundun et al.[160]. **(B-C)** Enrichment at different sequential ICeChIP steps with (B) chemical denaturant elution and (C) immunoglobulin and serum elution. **(D)** Enrichment of different nucleosome standards with ICeChIP-qPCR performed against H3K4me3, H3K27me3, and bivalency, with beads showing very little retention of chromatin (n=3 technical replicates). Error bars represent standard deviation. **(E)** Different configurations of bivalency on a single nucleosome. Of these, only trans-bivalency has been identified by mass spectrometry[116,145].

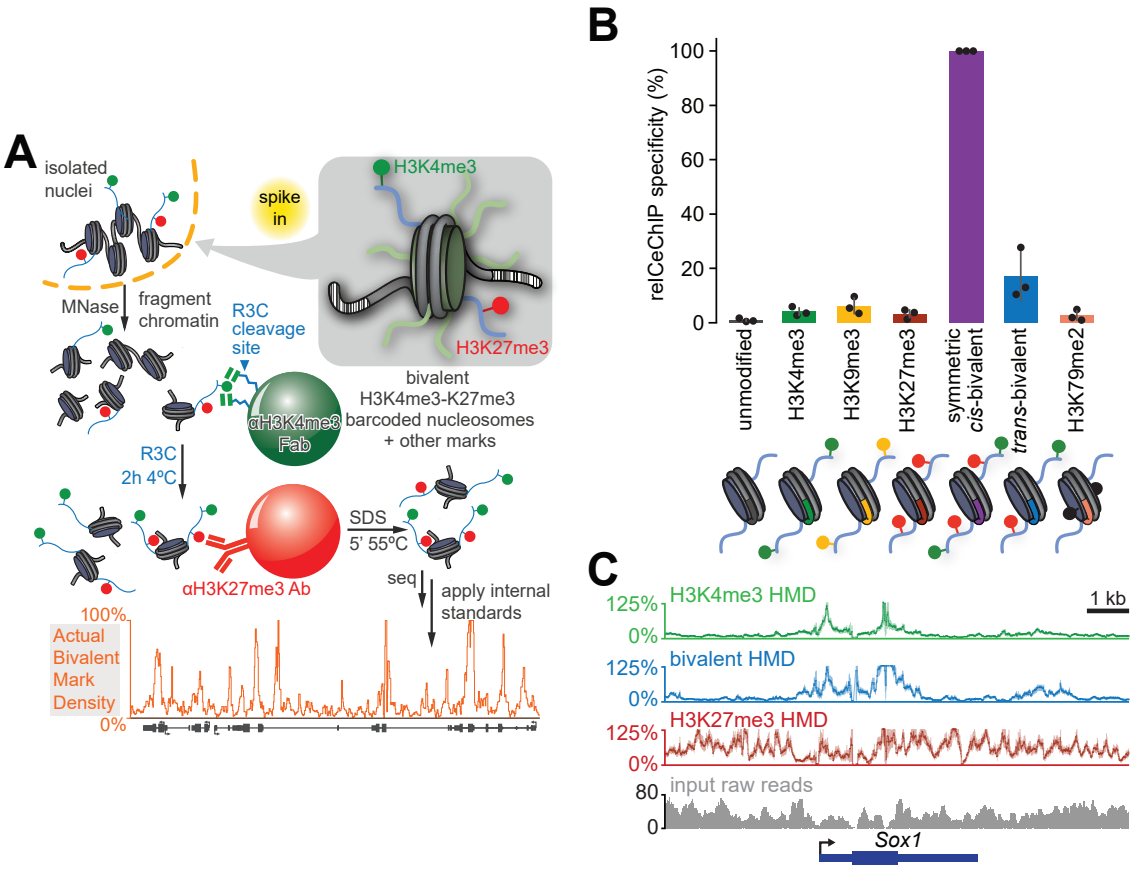


Figure 3.2: Workflow and evaluation of reICeChIP-seq.
(A) Schematic of reICeChIP-seq. The recombinant α-H3K4me3 Fab 304M3-B achieves high affinity by ”clasping” the histone tail between two Fab molecules[208], a binding mode readily achieved by multiple copies of the Fab presented on a bead, but not by the Fab in solution. Thus, protease cleavage not only elutes nucleosomes from the beads but also likely from the Fab complex. **(B)** Enrichment of different barcoded nucleosomes in reICeChIP-seq (n=3 biological replicates). Error bars represent S.D. **(C)** Representative line plot showing histone modification density of H3K4me3, H3K27me3, and bivalency ICeChIP-seq presented with 95% confidence intervals (lighter shade) and input read depth in naïve mESCs. Bivalency is calibrated to the trans-bivalency nucleosome standard and corrected for off-target H3K9me3 pulldown.

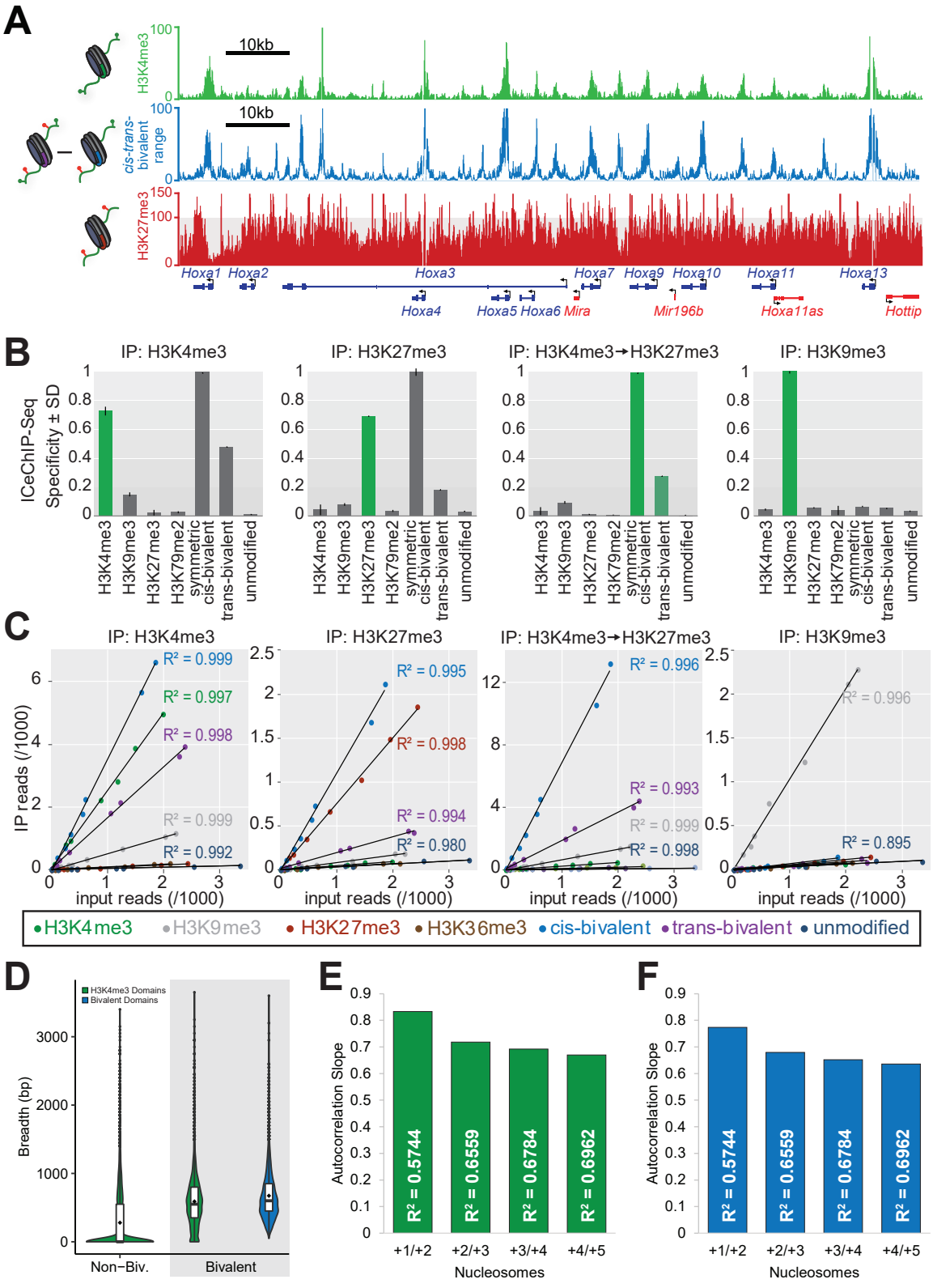


Figure 3.3: Evaluation of reICeChIP specificity and standards.

Figure 3.3, continued:
(A) Representative genome browser view of H3K4me3, H3K27me3, and bivalency, shown as a range of possible values by normalization to trans-bivalent (upper limit) or cis-bivalent (lower limit) nucleosome standards. **(B)** Relative pulldown of different nucleosome standards in ICeChIP-seq, normalized to the most-enriched standard. **(C)** Scatterplots of reads from DNA barcodes applied to nucleosome standards in ICeChIP-seq. **(D)** Violin plots of peak breadth (consecutive segment of 50bp windows overlapping promoter with >25% HMD) for H3K4me3 (green) and bivalency (blue) at non-bivalent and bivalent genes (>25% HMD) in naïve mESCs. **(E-F)** Autocorrelation of (E) H3K4me3 and (F) bivalency HMDs between nucleosomes in naïve mESCs. Nucleosomes are defined as sequential 200bp windows from the TSS.

primary pulldown more efficiently (Fig. 3.1D), resulting in 1000-2500x higher enrichment of the target over the published methods (Fig. 3.1A, 3.2B). This improvement enabled genome-wide measurement of bivalency HMD (Fig. 3.2C), representing the proportion of nucleosomes at a given locus modified with both H3K4me3 and H3K27me3, using the trans-bivalent nucleosome standards[210] as the calibrant (Fig. 3.1E, 3.3; Supplementary Note 3.1).

Bivalency through differentiation

With reICeChIP, we sought to study the role of bivalency in development by tracking its changes across a differentiation pathway that was used in several classic studies of bivalency[9,130,211]: differentiation from naïve mESCs[211] through the primed mESC state[211] to NPCs. In naïve mESCs, we noted that bivalency was far more widespread than previously reported (Fig. 3.4A-B); rather than ~1000 bivalent genes in naïve mESCs[211], we observed at least 10% bivalency HMD at most promoters (25768/42622), with almost 5000 promoters bearing bivalency at more than 50% of their nucleosomes (Fig. 3.4A,C; Supplementary Note 3.2, 3.3). This trend is recapitulated with primed mESCs, with the consensus set of bivalent promoters representing fewer than 2000 genes[130,185,200,212], as compared to more than 25,000 that are >25% bivalent in our analysis.

Even more striking were the changes in bivalency across this differentiation scheme. Previous studies suggested that bivalency largely disappears upon differentiation to NPCs[9,130,187,188].

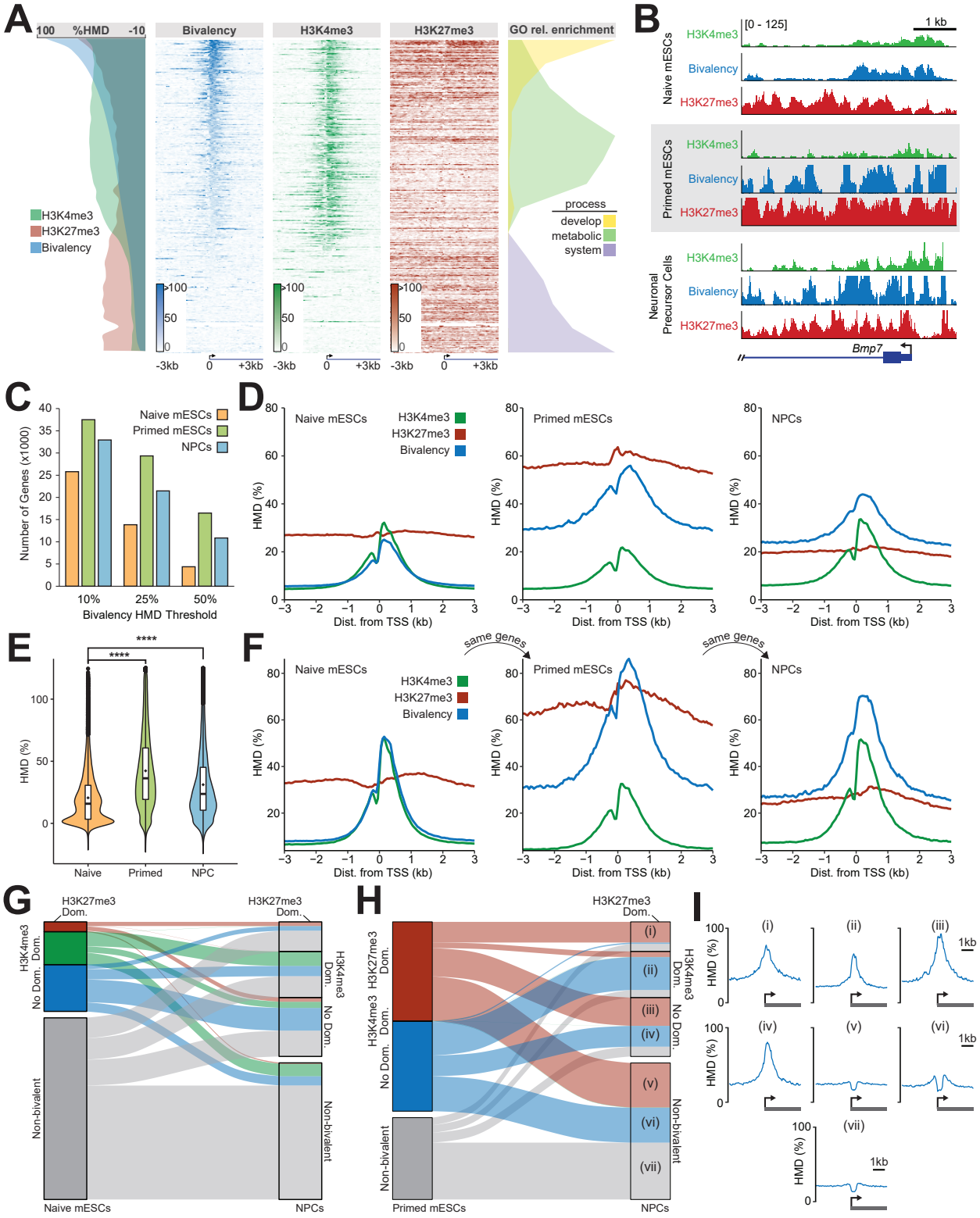


Figure 3.4: Bivalency is widespread and does not resolve over differentiation.

Figure 3.4, continued:
(A) Bivalency, H3K4me3, and H3K27me3 at all Refseq promoters in naïve mESCs, with relative enrichment of GO terms. Genes are rank ordered by bivalency HMD at promoter, defined as the region from 0 to +400 bp relative to the TSS. **(B)** Representative locus view of H3K4me3, H3K27me3, and bivalency at promoters in naïve mESCs (top), primed mESCs (centre), and NPCs (bottom), presented on the same scale of 0-125% HMD. **(C)** Number of promoters with bivalency HMDs above the given thresholds in each cell type out of a total of 42,622 Refseq promoters. **(D)** Metaprofiles of H3K4me3, H3K27me3, and bivalency at all promoters in naïve mESCs, primed mESCs, and NPCs. Heatmaps for primed mESCs and NPCs are presented in Extended Data Fig. 3b. **(E)** Distribution of bivalency HMDs at all Refseq promoters in three cell states, zoomed to below 125% HMD. Overall, 99.5% of naïve promoters, 87.3% of primed promoters, and 91.6% of NPC promoters have an HMD below 100%. Full plot in Extended Data Fig. 3a. **(F)** Metaprofiles of H3K4me3, H3K27me3, and bivalency at promoters identified as bivalent in naïve mESCs (25% HMD threshold), tracked from naïve mESCs to primed mESCs to NPCs. Heatmaps for bivalency are presented in Extended Data Fig. 3f. **(G-H)** Alluvial plots of dominance and bivalency of genes from (G) naïve mESCs to NPCs or (H) primed mESCs to NPCs. Bivalency [>25% HMD] can be subcategorized into dominance classes by independent ICeChIP for the constituent marks, with H3K27me3 in excess (H3K27me3/H3K4me3 > e^{1}), H3K4me3 dominant (H3K27me3/H3K4me3 < e^{-1}), or intermediate ratios (no dominance). **(I)** Bivalency metaprofiles for gene subsets indicated in panel (h) from -3kb to +3kb relative to the TSS. $^{****}p < 2.2x10^{-16}$.

However, we found the opposite; promoter bivalency *increases* upon differentiation (Fig. 3.4D-E, 3.5A-B), with thousands more genes meeting bivalency HMD thresholds relative to naïve mESCs (Fig. 3.4C). Similarly, we find that bivalent domains do not resolve upon differentiation; tracking bivalent genes from naïve mESCs through differentiation, we observe that bivalency is higher at these same promoters in primed mESCs and NPCs (Fig. 3.4F, 3.5C-F). As previously reported, primed mESCs have the most bivalency, likely related to the high level of promoter H3K27me3 in this state[211] (Fig. 3.4D). Accordingly, there are 27% fewer bivalent genes in NPCs than in primed mESCs (Fig. 3.4E). However, this decrease is nowhere near the previously reported decrease of 92%[130], and bivalent genes from primed mESCs remain highly bivalent in NPCs (Fig. 3.5G-H). Collectively, these data suggest that bivalency is far more widespread in this system than previously appreciated and remains elevated through differentiation, rather than resolving to one of the two monovalent states.

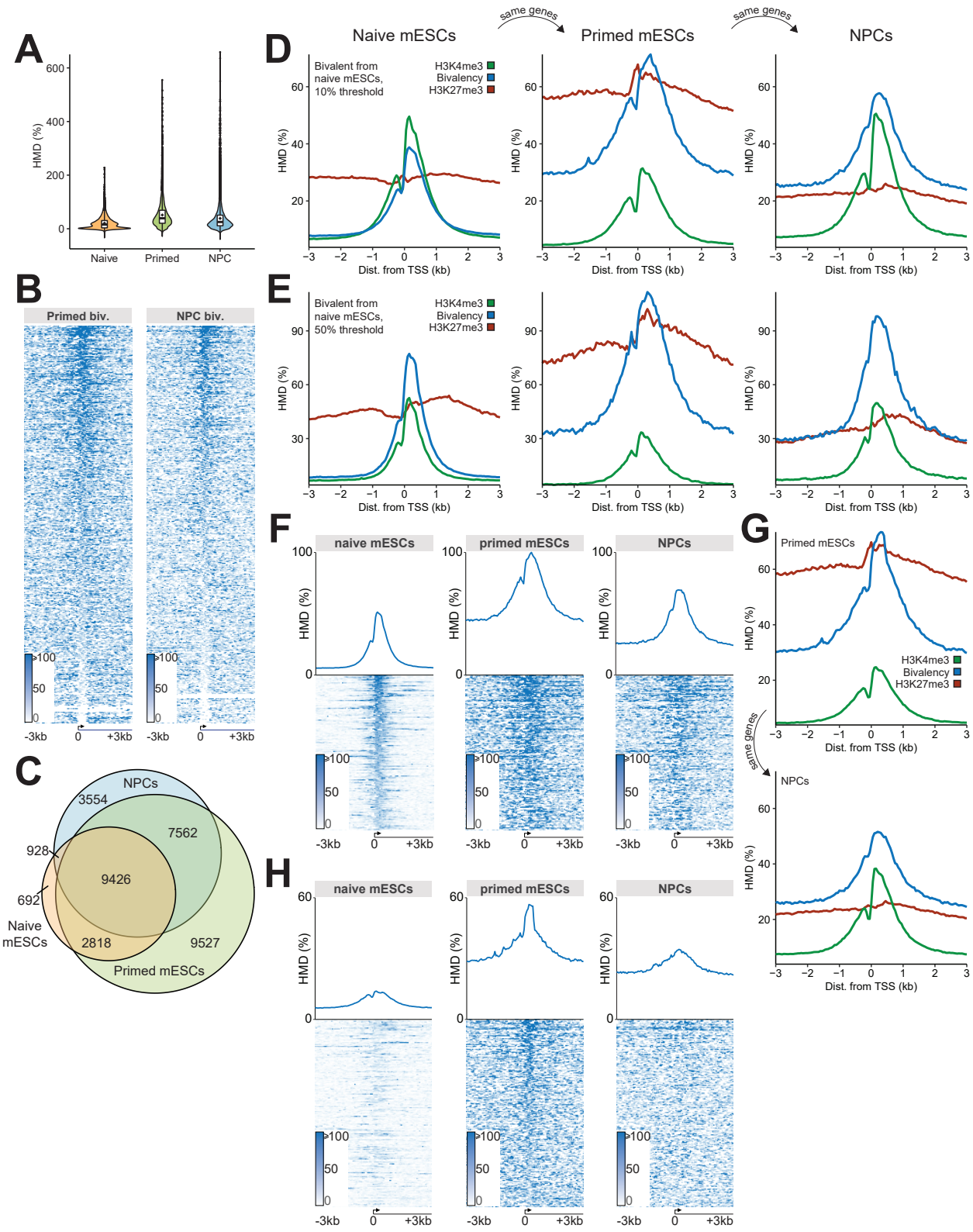


Figure 3.5: Tracking bivalent genes through differentiation.

Figure 3.5, continued:
(A) Distribution of bivalency HMDs at all Refseq promoters in three cell states. **(B)** Heatmaps of bivalency at all Refseq promoters in primed mESCs and NPCs. Genes are ordered by bivalency HMD at the promoter. **(C)** Venn diagram showing overlap of bivalent genes (25% HMD threshold) in naïve mESCs, primed mESCs, and NPCs. **(D-E)** Metaprofiles of H3K4me3, H3K27me3, and bivalency for bivalent genes in naïve mESCs with a (D) 10% or (E) 50% HMD threshold. **(F)** Heatmaps and metaprofiles of bivalent genes from naïve mESCs. **(G)** Metaprofiles of H3K4me3, H3K27me3, and bivalency at genes tracked from primed mESCs to NPCs for bivalent genes in primed mESCs (>25% HMD). **(H)** Heatmaps and metaprofiles of bivalent genes in primed mESCs that are not bivalent in naïve mESCs.

To investigate this discrepancy with the literature, we compared promoters identified as bivalent by other studies[130,188,200] to ours. The previously identified genes had 50-100% more H3K27me3 than do most bivalent genes in our set (Fig. 3.6A-B), suggesting that the previous studies undersampled H3K27me3 and thus could only identify regions with high H3K27me3 as bivalent. Accordingly, H3K27me3 dominant bivalent genes had the greatest proportional overlap with these canonical bivalent loci compared to other dominance classes (i.e. whether the bivalent genes have excess H3K27me3, excess H3K4me3, or roughly equal levels as measured by independent ICeChIP experiments for these two marks; Fig. 3.6C). The common practice of measuring bivalency as regions of overlapping H3K4me3 and H3K27me3 is also problematic, even with calibrated data[118]; many promoters with high H3K4me3 and H3K27me3 bear less than 25% bivalency (Fig. 3.6D). Notably, even for the previously identified bivalent genes, bivalency still increases relative to naïve mESCs upon differentiation. And in our datasets, this holds true across modification dominance classes – even the H3K27me3 dominant bivalent genes, which most closely resemble the canonically bivalent loci (Fig. 3.6-3.7). To the extent that any bivalency class resolves from naïve mESCs to NPCs, the largest set of genes is from the H3K4me3 dominant bivalent genes ($p = 1.78 \times 10^{-133}$; Fig. 3.4G), despite its minimal overlap with the canonical bivalent loci (Fig. 3.6C).

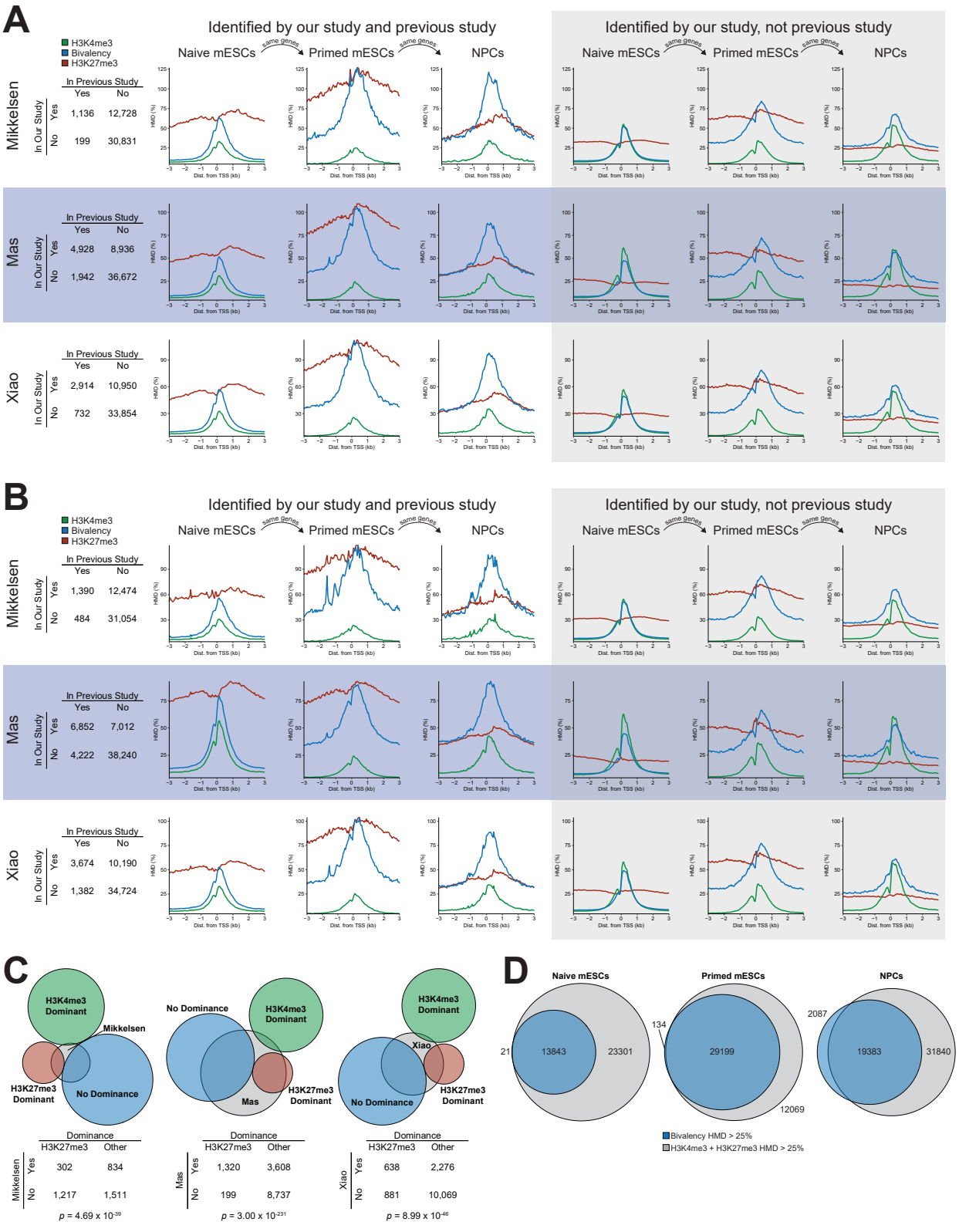


Figure 3.6: Comparing our bivalent genes to other studies.

Figure 3.6, continued:
(A-B) Contingency tables and metaprofiles for genes that are identified as >25% bivalent in our study and by Mikkelsen et al.[130], Mas et al.[200], and Xiao et al.[212], wherein: (A) gene is identified as bivalent in the external study if overlapping H3K4me3 and H3K27me3 peaks overlap the 0 to +400bp region of a gene relative to the TSS, or (B) gene is identified as bivalent in the external study if overlapping H3K4me3 and H3K27me3 peaks overlap the region from 2.5kb upstream of the TSS to the end of the gene[185]. **(C)** Overlap of bivalent genes from external datasets (as defined in part A) with each of our bivalent gene dominance classes in naïve mESCs. Significance computed by two-tailed Fisher hypergeometric test. **(D)** Overlap of genes with bivalency HMD > 25% and with H3K4me3 + H3K27me3 HMD > 25% in all three cell states.

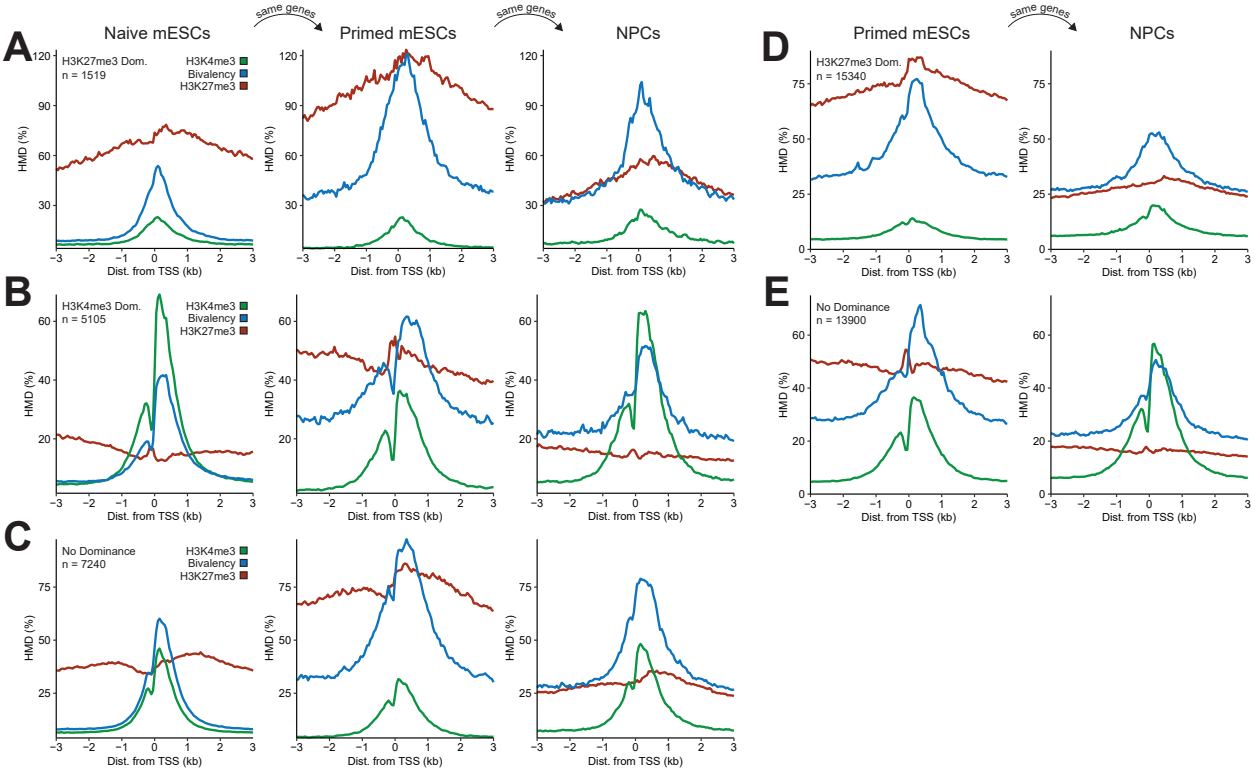


Figure 3.7: Bivalency changes across differentiation by modification dominance class.
(A-C) Metaprofiles of H3K4me3, H3K27me3, and bivalency for bivalent genes (>25% HMD) in naïve mESCs that are (A) H3K27me3 dominant (H3K27me3/H3K4me3 > e^{1}), (B) H3K4me3 dominant (H3K27me3/H3K4me3 < e^{-1}), or (C) have no dominance in naïve mESCs, tracked through three cell states. **(D-E)** Metaprofiles of H3K4me3, H3K27me3, and bivalency for bivalent genes (>25% HMD) in primed mESCs that are for indicated dominance classes tracked from primed mESCs to NPCs.

Having found that bivalency is unexpectedly common and persistent in early differentiation, we investigated the enzyme complexes that could potentially account for this ubiquity. Previous

work suggested that H3K27me3 and H3K4me3 each inhibit deposition of the other[145,210,213,214], particularly when symmetric (Supplementary Note 3.4), raising questions as to whether the pervasive bivalency we observe is plausible. To address this concern, we performed histone methyltransferase (HMTase) assays with Set1B and the full panel MLL-family core complexes (MLL1, MLL2, MLL3, MLL4), which collectively account for the bulk of H3K4 methylation in humans[215]. We find that these complexes all tolerate a wide spectrum of H3K27me3-decorated nucleosomes (Fig. 3.8), indicating that the formation of bivalent nucleosomes is not precluded by allosteric modulation of H3K4me3 installation by core factors. Although it has been suggested that Set1a[216], Mll2[217], Ezh1[218], and Ezh2[219] are all important for establishing bivalency, only Mll2 appears to be sensitive for identifying bivalent promoters in naïve mESCs, with none showing high specificity for the same (Fig. 3.9). Together, these data support the proposed specialized role for Mll2 in bivalency[217], indicate a pleiotropic role for PRC2 beyond its role in establishing bivalency, and provide plausible enzymatic avenues to the prevalent bivalency we observe by reICeChIP.

Bivalency, gene expression, and ontology

A key pillar of the bivalency hypothesis is that bivalent promoters are associated with transcriptionally repressed genes poised to be activated or terminally silenced upon differentiation[9,130,185,186]. However, bivalency is not solely found at genes with low expression in any of our measurements (Fig. 3.10A, 3.11A-B). Rather, bivalent genes had higher average expression than did non-bivalent genes or the set of all genes, and these genes display modestly higher average expression through differentiation (Fig. 3.10A), with bivalency being similar across most expression deciles (Fig. 3.11C). Bivalency associated similarly with bulk gene expression (Fig. 3.10B) and the proportion of cells expressing the associated transcripts in single cell RNA-seq (Fig. 3.10C), suggesting that

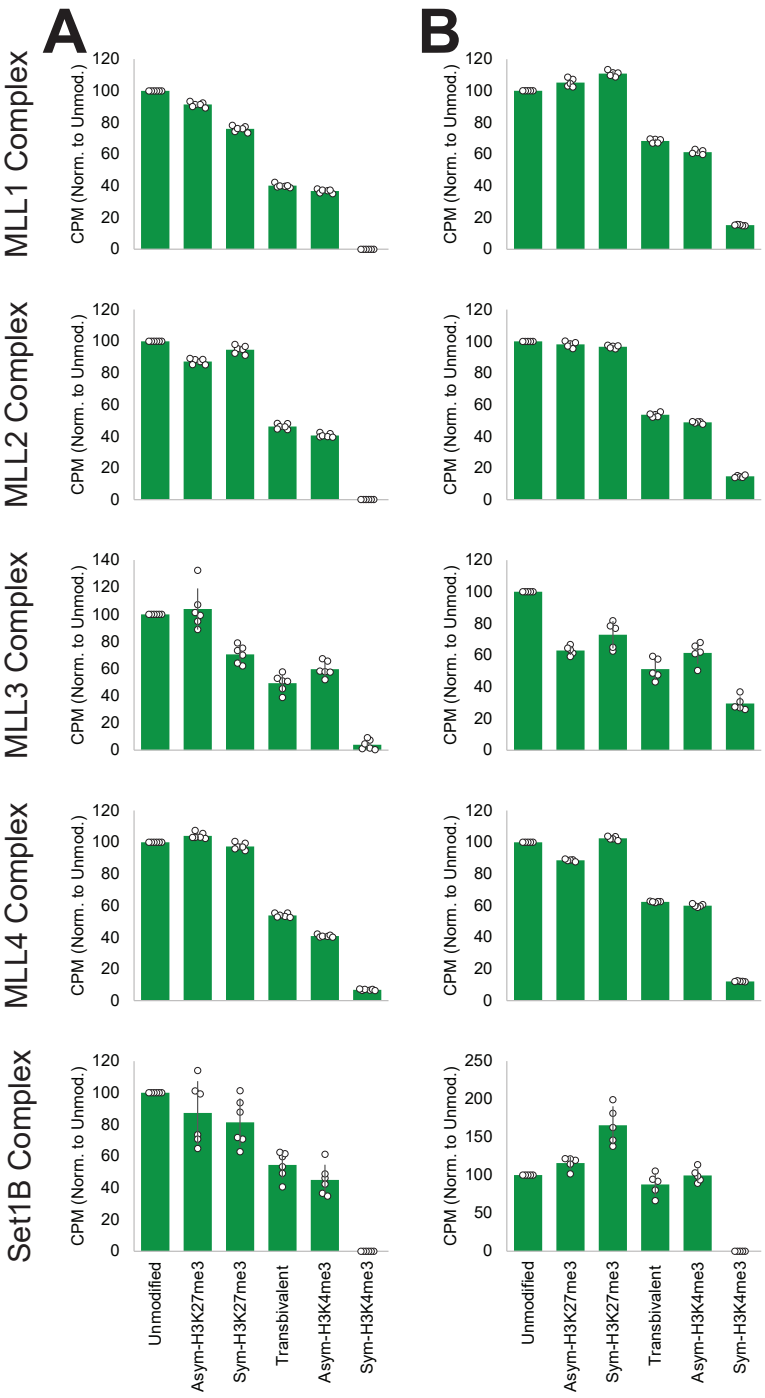


Figure 3.8: Methyltransferase assays identifying potential pathways for establishment of bivalency. **(A-B)** Methyltransferase assays for MLL1, MLL2, MLL3, MLL4, and Set1B core HMTase complexes using (A) 15 ng/uL (n=6) and (B) 20 ng/uL (n=5) semisynthetic nucleosomes as substrates for methylation. Endpoints were established at 180 min by kinetic evaluation to be sensitive to difference in activity for this panel. Signal is corrected for background and no nucleosome substrate activity. Error bars represent standard deviation.

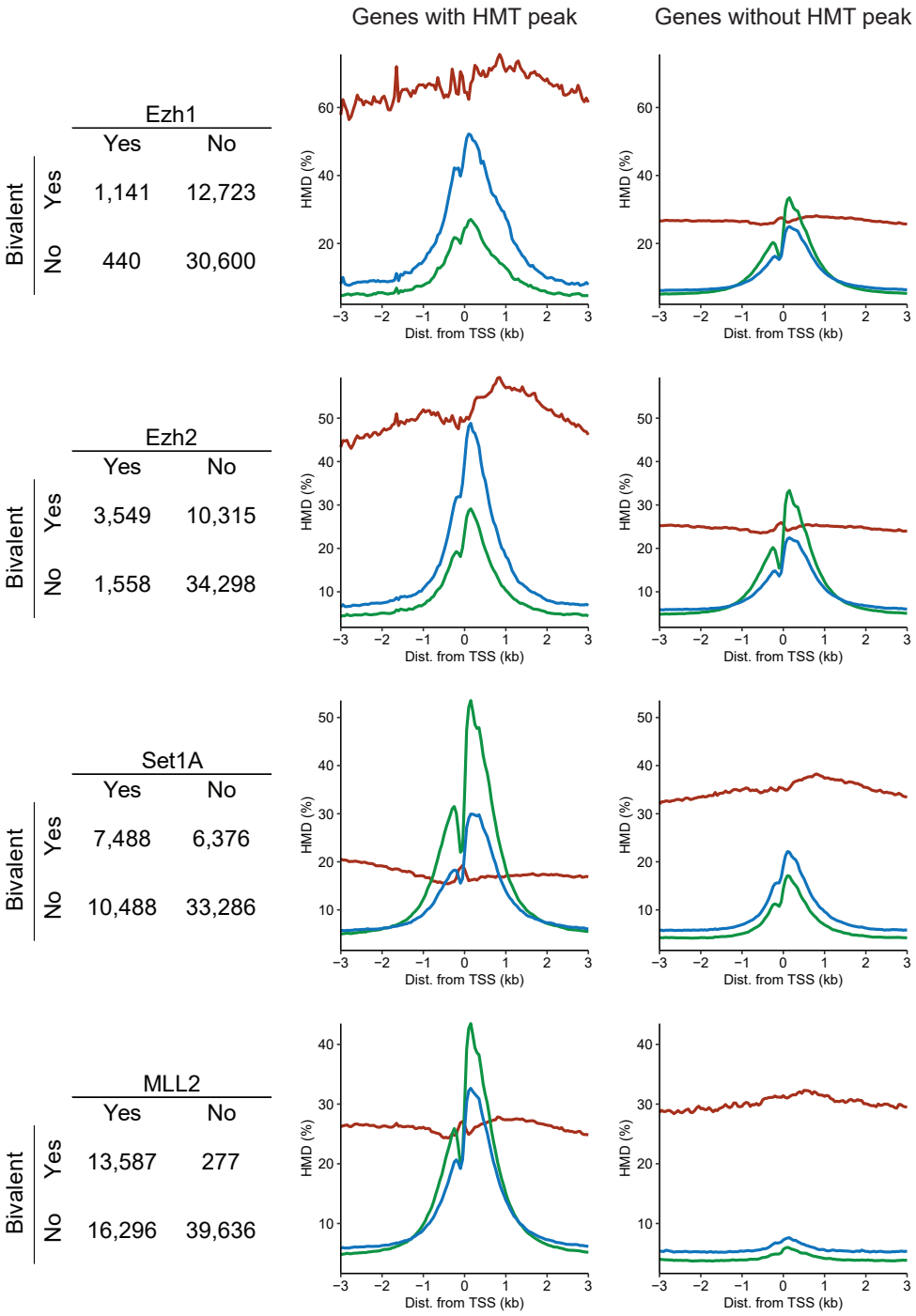


Figure 3.9: HMTase peaks and bivalency.

Contingency tables and metagene profiles in naïve mESCs for genes with and without overlapping HMT peaks. Ezh1 and Ezh2 peaks were identified as Suz12 peaks lost upon Ezh1 or Ezh2 knock-out[81]. Set1A peaks were identified by ChIP against Set1A[216]. Mll2 peaks were identified by ChIP against Mll2[206].

the association of bivalency with higher-expressed genes is not solely driven by intercellular heterogeneity. Consistent with previous observations[130], bivalency was higher at promoters with high CpG content (Fig. 3.11D) and associated with lower DNA methylation compared to non-bivalent genes (Fig. 3.11E, also holds for each dominance class). These data all suggest that bivalent genes are more highly expressed than non-bivalent genes as a whole, and this latter class is seemingly more subject to regulation by DNA methylation.

Another pillar of the bivalency model is that bivalent genes are poised to be differentially regulated through differentiation. To test this, we computed the sensitivity and specificity of different bivalency and non-bivalency classes for differentially expressed genes (DEGs; Supplementary Note 3.5). Counter to the bivalency hypothesis and previous results[9,130,186], we found that bivalency was a very poor marker of DEGs; from naïve mESCs to NPCs, bivalency was roughly as sensitive and specific for identifying DEGs as was a *lack* of bivalency (Fig. 3.10D). Though H3K27me3-dominant bivalent genes showed an increase in average gene expression (Fig. 3.11F-G), this class still only had 60% specificity for identifying DEGs, with very low sensitivity (Fig. 3.10D). Promoters of DEGs and non-DEGs from naïve mESCs to NPCs have highly similar histone modification metaprofiles in naïve mESCs (Fig. 3.10E-F) and across differentiation (Fig. 3.11H-K). Comparison of primed mESCs to NPCs displayed similar trends (Fig. 3.10G-H); though sensitivity was higher because most genes are bivalent in primed mESCs, the specificity remained similar between bivalent and non-bivalent genes. Interestingly, whether genes were upregulated, downregulated, or non-DEGs, bivalency still increased over differentiation (Fig. 3.11H-K). Collectively, these analyses show that bivalency is neither sensitive nor specifical for poised DEGs in this system.

We next examined whether bivalency is primarily associated with developmental genes, a central tenet of the original model[9,130]. The first ICeChIP study indirectly hinted that there may be at

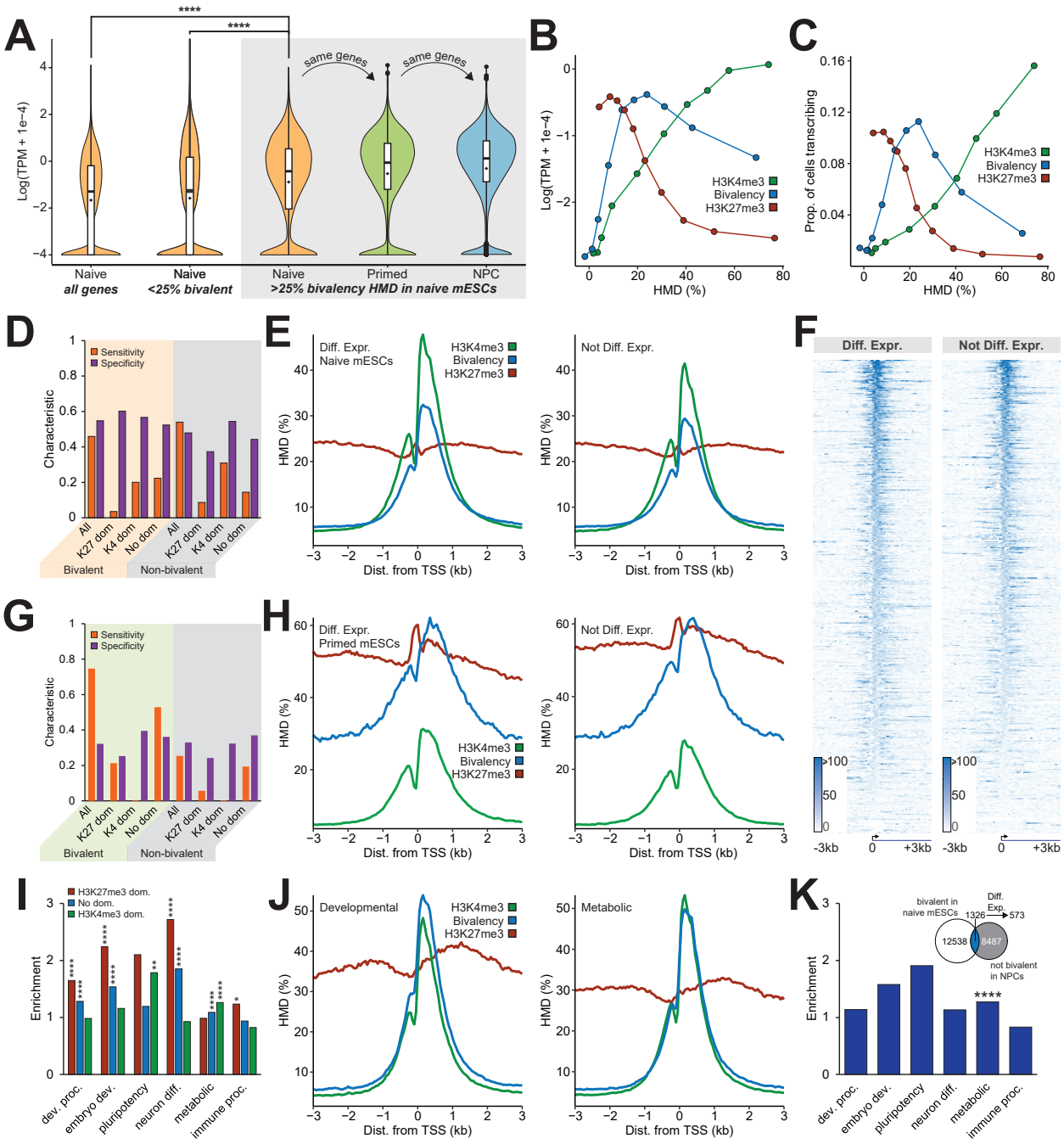


Figure 3.10: Bivalency is neither sensitive nor specific for identifying poised or developmental genes.

(A) Violin plots of gene expression[220] for all genes in naïve mESCs, non-bivalent genes (<25% HMD) in naïve mESCs, and bivalent genes (>25% HMD) tracked from naïve mESCs to the same genes in the indicated lineages. Significance computed by Welch's two-tailed *t*-test. **(B)** Gene expression vs. HMD for H3K4me3, H3K27me3, and bivalency (genes are binned into HMD deciles). **(C)** Proportion of actively transcribing cells by single-cell RNA-seq[221] vs. HMD for H3K4me3, H3K27me3, and bivalency (genes are binned into HMD deciles). **(D)** Sensitivity and

Figure 3.10, continued:
specificity (Supplementary Note 3.5) of bivalent and non-bivalent genes in naïve mESCs identifying differentially expressed genes (DEGs) from the naïve state to the NPC state. **(E)** Metaprofiles of H3K4me3, H3K27me3, and bivalency and **(f)** heatmaps of bivalency in naïve mESCs at DEGs and non-DEGs relative to NPCs. **(g)** Sensitivity and specificity of bivalent and non-bivalent genes in primed mESCs identifying DEGs from the primed state to the NPC state. **(h)** Metaprofiles of H3K4me3, H3K27me3, and bivalency in primed mESCs at DEGs and non-DEGs. **(i)** Gene ontology term enrichment of H3K27me3-dominant bivalent genes, H3K4me3-dominant bivalent genes, or bivalent genes with no clear dominance (q-value two-tailed Fisher hypergeometric test). **(j)** Metaprofiles of H3K4me3, H3K27me3, and bivalency in naïve mESCs at developmental and metabolic genes. **(k)** Gene ontology term enrichment of genes following the classic bivalency model: DEGs that lose bivalency from naïve mESCs (>25% HMD) to NPCs (<10% HMD). Significance computed by two-tailed Fisher hypergeometric test. $^{*}q < 0.05$. $^{**}q < 0.01$. $^{****}p$ or $q < 2.2x10^{-16}$.

least two classes of bivalent promoters: an H3K27me3 dominant class associated with developmental genes, and an H3K4me3 dominant class enriched for metabolic genes[118]. Direct measurements of bivalency herein unambiguously demonstrate this phenomenon more broadly (Fig. 3.4A, 3.10I). Overall, bivalent genes are enriched for a broad range of ontology terms, including developmental, metabolic, and immune system process genes (Fig. 3.10I-J), with nearly identical bivalency profiles in naïve mESCs (Fig. 3.10I, 3.12A). These classes all not only retained, but increased bivalency into NPCs – even immune system process genes, despite being seemingly unrelated to neuronal development. We only found 543 genes that *did* obey the classic bivalency model (Fig. 3.10K), representing less than 5% of the bivalent genes from naïve mESCs, with little difference in bivalency between upregulated and downregulated genes (Fig. 3.12B). Interestingly, these genes were most significantly enriched for metabolic rather than developmental genes (Fig. 3.10K). Taken together, these data suggest that bivalency is neither primarily nor specifically associated with developmental genes in this system.

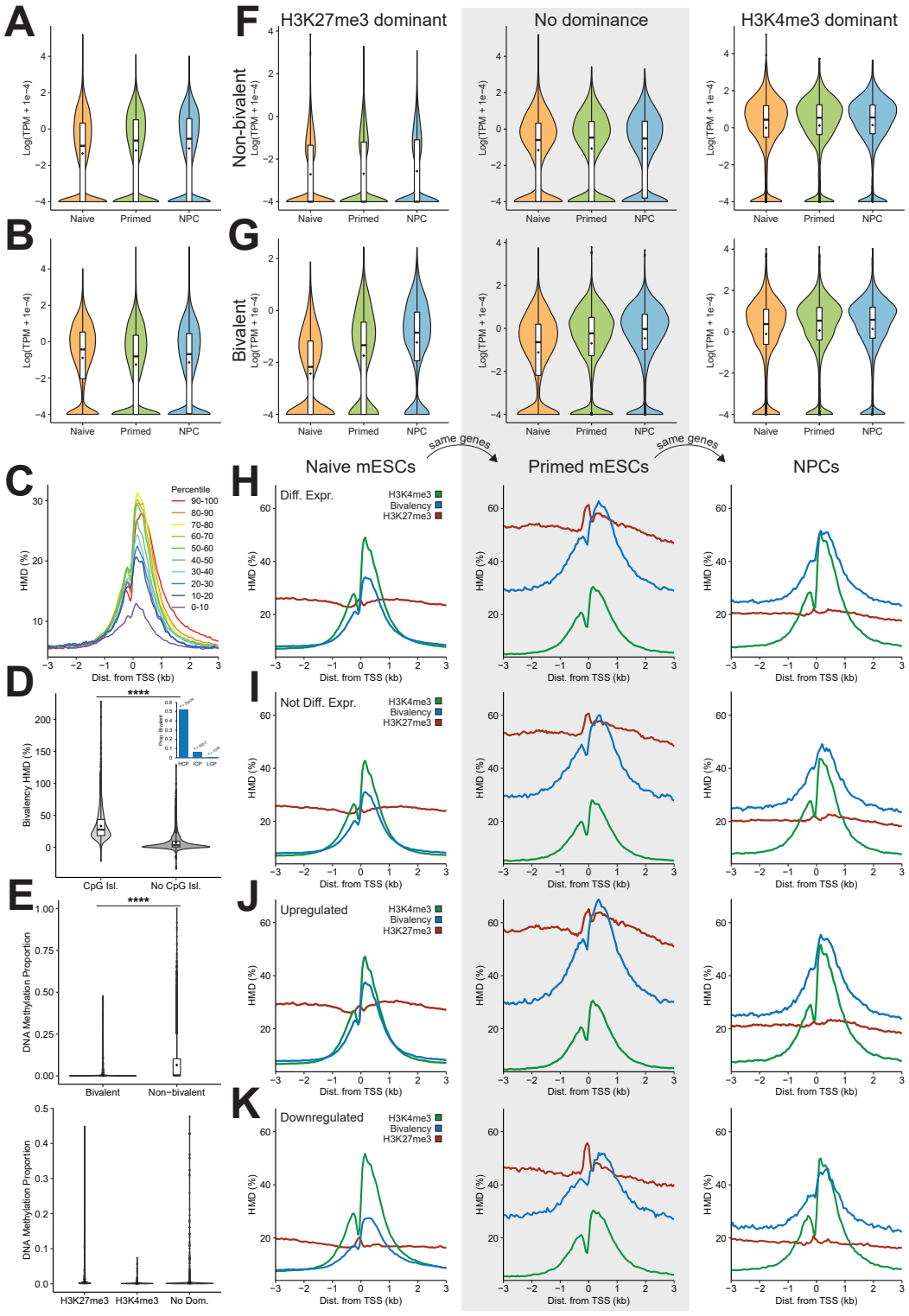


Figure 3.11: Bivalency and differential gene expression.

(A-B) Violin plots of gene expression for (A) all genes and (B) bivalent (>25% HMD) genes in each cell state. **(C)** Bivalency metaprofiles in naïve mESCs at promoters binned by gene expression decile.

Figure 3.11, continued:
(D) Violin plots of bivalency HMD in naïve mESCs at promoters with and without CpG islands. Inset shows proportion of genes that are bivalent in sets of genes classified by CpG content: high-CpG promoters (HCP), intermediate-CpG promoters (ICP), and low-CpG promoters (LCP), defined as previously described by Mikkelsen et al.[130]. Total number of genes in each class is provided as n. **(E)** Violin plots of DNA methylation at bivalent and non-bivalent genes (top), broken by dominance class for bivalent genes (bottom). **(F-G)** Violin plots of gene expression in (F) non-bivalent (<25% HMD) and (G) bivalent (>25% HMD) genes from naïve mESCs that are H3K27me3 dominant (H3K27me3/H3K4me3 > e^{1}; left), have no clear dominance (centre), or are H3K4me3 dominant (H3K27me3/H3K4me3 < e^{-1}; right). **(H-K)** Metaprofiles of H3K4me3, H3K27me3, and bivalency at genes tracked from naïve mESCs to primed mESCs to NPCs for (H) DEGs, (I) non-DEGs, (J) genes upregulated from naïve mESCs to NPCs, and (K) genes downregulated from naïve mESCs to NPCs. $^{****}p < 10^{-16}$ (Welch's two-tailed t-test).

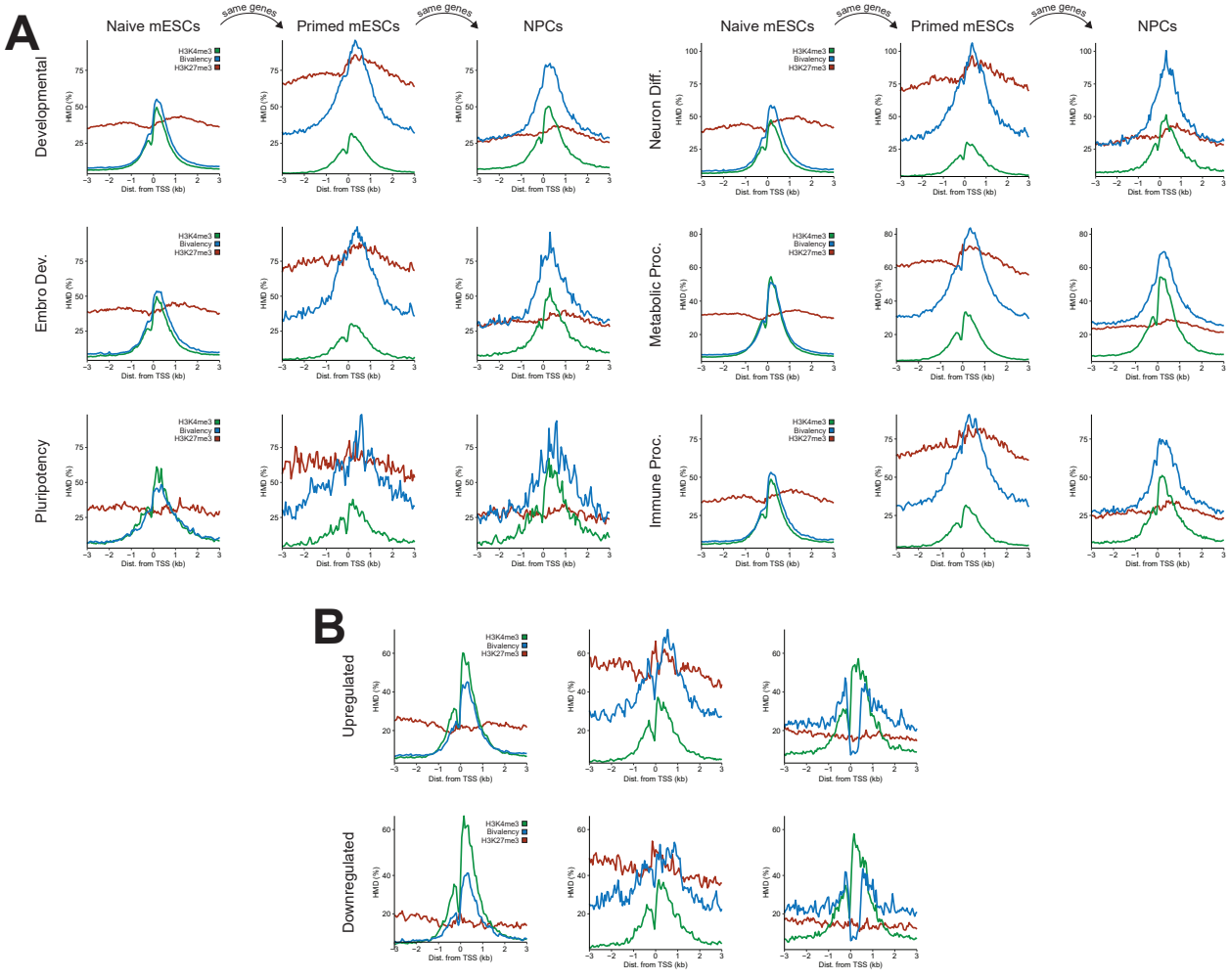


Figure 3.12: Bivalency at different classes of genes.

Figure 3.12, continued:
(A) Metaprofiles of H3K4me3, H3K27me3, and bivalency at genes tracked from naïve mESCs to primed mESCs to NPCs for bivalent genes of indicated gene ontology terms. **(B)** Metaprofiles of H3K4me3, H3K27me3, and bivalency at genes tracked across differentiation for genes that lose bivalency at the promoters (0 to +400bp relative to TSS) from naïve mESCs (>25% HMD) to NPCs (<10% HMD) and are upregulated (top) or downregulated (bottom) over differentiation.

Predicting DEGs with histone PTMs

The premise of the bivalency hypothesis is that the coexistence of H3K4me3 and H3K27me3 synergistically provides additional predictive information about the associated genes upon differentiation beyond that provided by H3K4me3 and H3K27me3 alone. With quantitative measurements of these modifications, this hypothesis can be tested by modelling. We first determined which individual parameters best identified DEGs by measuring the area under the curve (AUC) of receiver operator characteristic (ROC) curves of parameter thresholds. Of the individual histone modifications, H3K4me3 levels were best for identifying DEGs, with the highest AUC of the ROC (Fig. 3.13A, 3.14A). Bivalency was less predictive of DEGs than were either the log ratio of H3K27me3 and H3K4me3 or DNA methylation (Fig. 3.13A, 3.14A). And in primed mESCs, far from being predictive of poised genes, bivalency was *inversely* associated with DEGs upon differentiation to NPCs (Fig. 3.14A).

If bivalency provides additional information over H3K4me3 and H3K27me3, then a model without bivalency will be markedly less explanatory than a model with bivalency. To test this, we conducted logistic regressions with linear models to identify parameters most important for identifying DEGs. Bayes Information Criterion analyses preliminarily hinted that bivalency provided minimal information to this end (Fig. 3.14B; Supplementary Note 3.6). To more definitively identify whether bivalency provides meaningful predictive information, we conducted hold-out cross-validation on models with H3K4me3, H3K27me3 and either nothing else, bivalency, H3K9me3, or

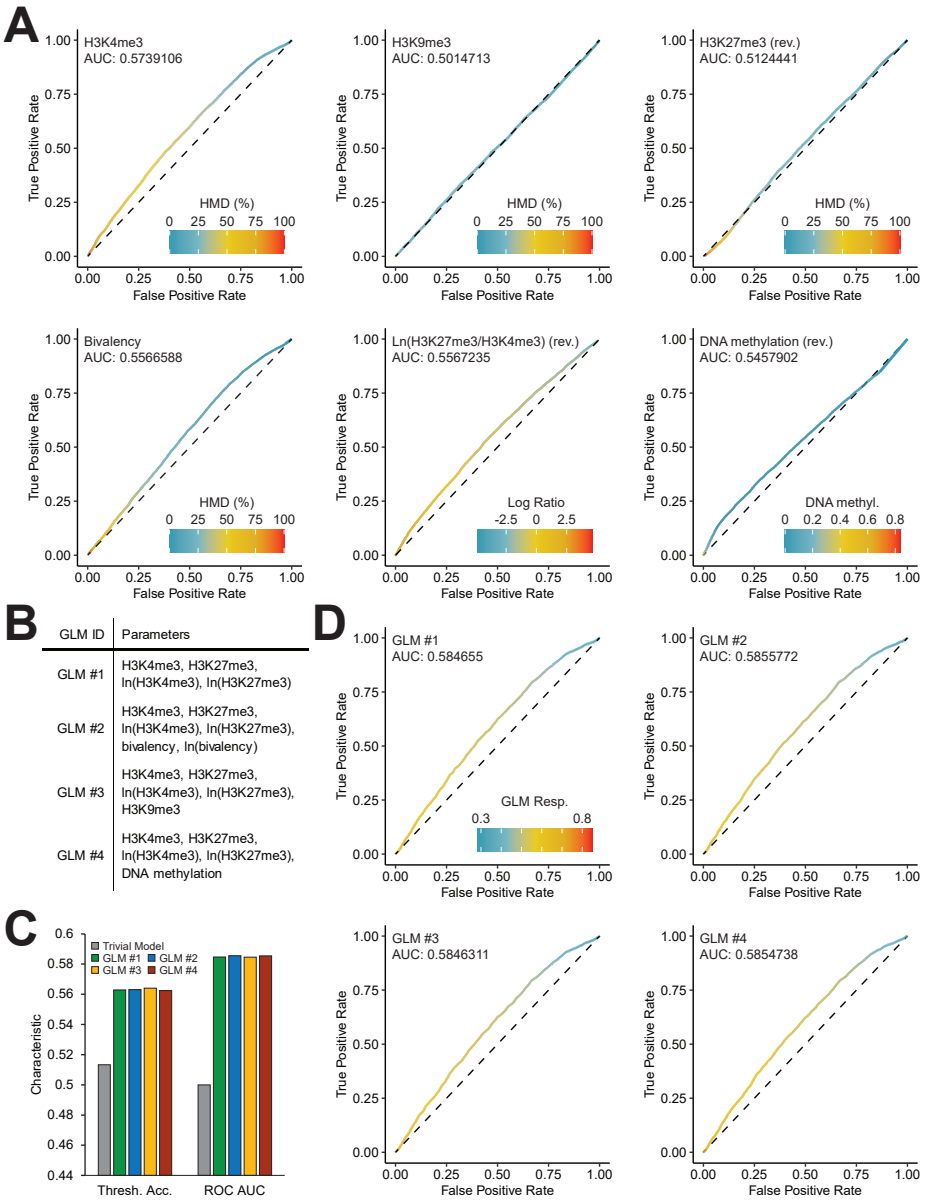
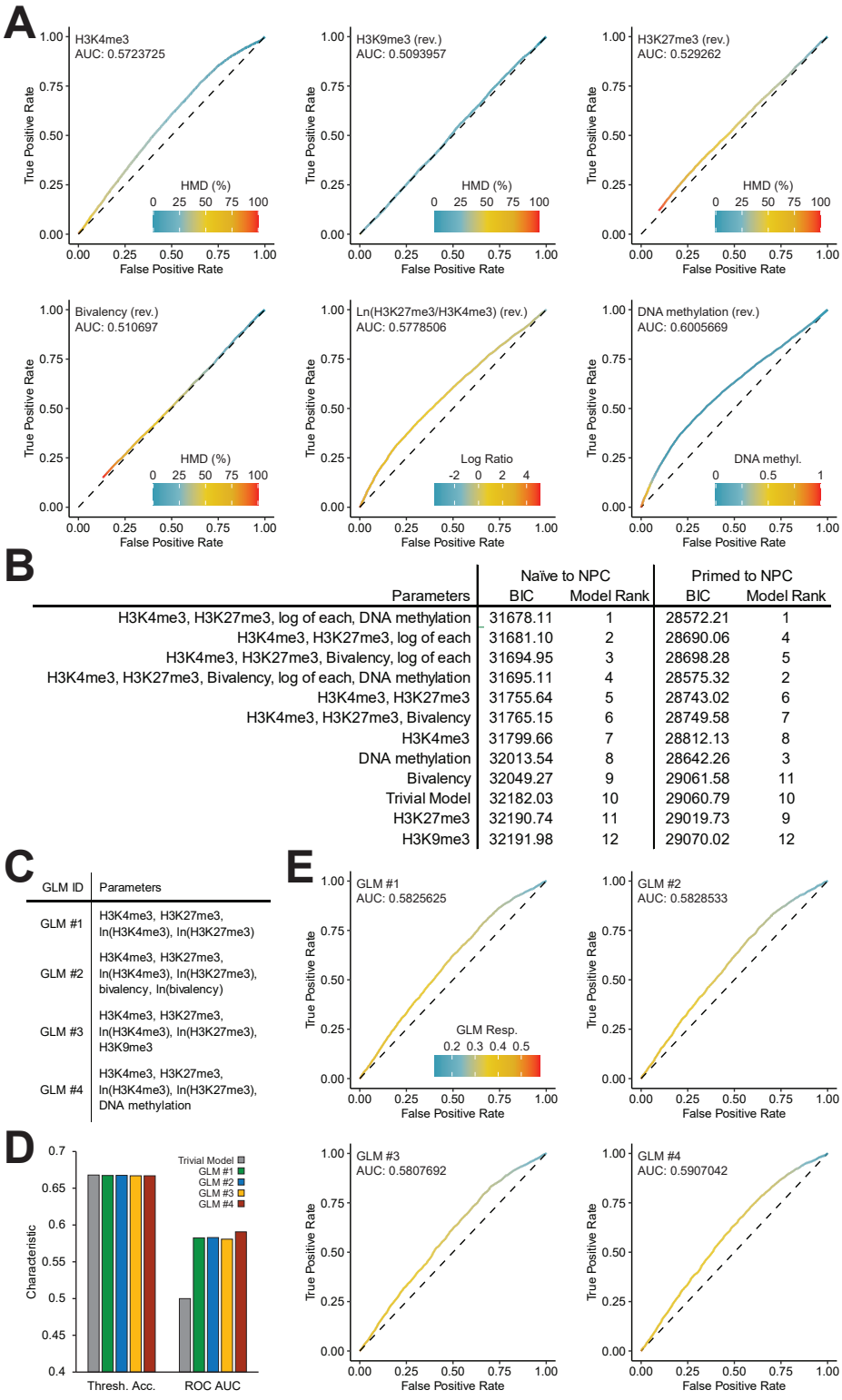


Figure 3.13: Bivalency does not provide appreciably more information than H3K4me3 and H3K27me3 alone for DEG prediction.

(A) Receiver operator characteristic (ROC) curves for identifying DEGs from naïve mESCs to NPCs by H3K4me3, H3K9me3, H3K27me3, bivalency, ln(H3K27me3/H3K4me3), or DNA methylation in naïve mESCs. For each point, parameter value threshold used to compute true positive rate (TPR) and false positive rate (FPR) is indicated by the colour. Traits with thresholds identifying non-DEGs rather than DEGs are marked with "rev." **(B)** Legend for generalized linear models (GLMs) in panels c-d. **(C)** Accuracy of trivial model and GLMs by threshold accuracy (gene identified as DEG if logistic regression > 0.5; left) and by ROC area under curve (right). **(D)** ROC curves for identifying DEGs from naïve mESCs to NPCs by different GLMs. For each point, logistic regression threshold value used to compute TPR and FPR is indicated by the colour.



Parameters	Naïve to NPC BIC	Naïve to NPC Model Rank	Primed to NPC BIC	Primed to NPC Model Rank
H3K4me3, H3K27me3, log of each, DNA methylation	31678.11	1	28572.21	1
H3K4me3, H3K27me3, log of each	31681.10	2	28690.06	4
H3K4me3, H3K27me3, Bivalency, log of each	31694.95	3	28698.28	5
H3K4me3, H3K27me3, Bivalency, log of each, DNA methylation	31695.11	4	28575.32	2
H3K4me3, H3K27me3	31755.64	5	28743.02	6
H3K4me3, H3K27me3, Bivalency	31765.15	6	28749.58	7
H3K4me3	31799.66	7	28812.13	8
DNA methylation	32013.54	8	28642.26	3
Bivalency	32049.27	9	29061.58	11
Trivial Model	32182.03	10	29060.79	10
H3K27me3	32190.74	11	29019.73	9
H3K9me3	32191.98	12	29070.02	12

Figure 3.14: Quantifying the additional information content provided by bivalency over H3K4me3 and H3K27me3 alone.

Figure 3.14, continued:
(A) ROC curves for identifying DEGs from primed mESCs to NPCs by H3K4me3, H3K9me3, H3K27me3, bivalency, ln(H3K27me3/H3K4me3), or DNA methylation in primed mESCs. For each point, parameter value threshold used to compute true positive rate (TPR) and false positive rate (FPR) is indicated by the colour. Traits with thresholds identifying non-DEGs rather than DEGs are marked with "rev." **(B)** Bayes Information Criterion (BIC) for logistic models identifying DEGs from naïve mESCs or primed mESCs to NPCs with different parameters. **(C)** Legend for generalized linear models (GLMs). **(D)** Accuracy of trivial model and GLMs by threshold accuracy (gene identified as DEG if logistic regression > 0.5; left) and by ROC area under curve (right). **(E)** ROC curves for identifying DEGs from primed mESCs to NPCs by different GLMs. For each point, logistic regression threshold value used to compute TPR and FPR is indicated by the colour.

DNA methylation (Fig. 3.13B, 3.14C; Supplementary Note 3.6). Parameters other than H3K4me3 and H3K27me3 barely improved model accuracy by two separate metrics (Fig. 3.13C-D, 3.14D-E; Supplementary Note 3.4), suggesting that those parameters provide virtually no additional information content to identify DEGs. These data suggest that, in this developmental system, there is little evidence that bivalency has emergent properties in identifying poised genes beyond the combined independent properties of H3K4me3 and H3K27me3.

Discussion

The bivalency hypothesis is one of the more influential ideas in epigenetics and molecular developmental biology. Persistent interest over the years coupled with widespread deployment and acceptance of sub-optimal bivalency measurement methods has ossified the hypothesis into dogma that extends well beyond any of the experimental data that informed it.

However, this coalescence has not been reached based on functional assays. Indeed, to the extent that functional validation of the bivalency model has been attempted, it has primarily been through deletion of enzymes with pleiotropic effects and functions throughout the genome beyond installation of bivalency[200,217,222,223]. Overwhelmingly, the prevailing views on the role of bivalency are derived from ChIP experiments. However, ChIP protocols[150] and antibodies[108,111,124,133,134] are

often highly susceptible to off-target pulldown, and uncalibrated ChIP without exogenous normalization can distort signal and the ability to compare experiments[118,124,128], leading to spurious conclusions[124]. From the quantitative and specific measurements we made with reICeChIP, we fear that this has been the case with the bivalency hypothesis, at least as far as these analyses in early mESC differentiation permit.

It has been held that bivalency is present at a small, restricted set of promoters early in development; we find that bivalency is widespread, with many thousands of promoters displaying high bivalency levels. It has been held that bivalency primarily exists early in development and resolves upon differentiation; we find that bivalency persists at least through the NPC stage and *increases* over baseline in that span. It has been held that bivalency demarcates poised, developmental genes associated with lineage commitment; we find that bivalency is neither sensitively nor specifically associated with developmental nor differentially expressed genes – and, at worst, may be *inversely* associated with the latter. Moreover, bivalent genes are predominantly not poised in an off state, but are more highly expressed than those that are not bivalent. All told, we find little evidence that bivalency provides more information in predicting poised gene status than do H3K4me3 and H3K27me3 in an independently additive manner in this system, raising questions as to whether it represents any more than a coincidental overlap of the aforementioned two marks.

Our study is not without caveats. First, we are only able to comment meaningfully on the differentiation paradigm presented here; we cannot definitively infer that these results will hold for the other developmental or clinical contexts. Although the original studies on bivalency indicated that bivalency almost entirely disappeared by the NPC stage[9,130], this stage is not terminally differentiated, so it is possible that bivalency could resolve in later stages of differentiation. Future studies will be needed to address this possibility in other developmental contexts. Second, though the extant

evidence suggests that only trans-bivalency is present at meaningful levels, our method cannot selectively distinguish between *cis*-, *trans*-, and intermediate bivalency conformations (Supplementary Note 3.1).

The reICeChIP method is not inherently restricted to the study of H3K4me3/H3K27me3 bivalency. With cleavable recombinant affinity reagents targeting other histone modifications[208,224] it could be used to quantify other combinatorial modification patterns[225–228] or modification symmetry.

Without serious changes to the standards of ChIP, the limitations of conventional ChIP-seq will continue to pose an existential challenge to the field. Indeed, the divergence between our observations of bivalency and those in the literature can be attributed to the historical lack of tools needed to make quantitative and specific measurements; in that context, the experimental designs and interpretations of the past were reasonable. Fortunately, such tools now exist. And as we have shown in this work, these methods offer a chance for the field to critically evaluate its orthodox models and pave the way for new insights on the chromatin determinants of cell identity and the regulation of development.

Acknowledgements

We would like to thank Peter Faber, Hannah Whitehurst, and Mikayla Marchuk in the University of Chicago Functional Genomics Facility for Illumina sequencing. We would also like to thank EpiCypher, Inc. for providing some of the histone octamers for this study. Adrian T. Grzybowski was supported by the Harper Dissertation Prize and the Dean's International Student Fellowship of the University of Chicago. Rohan Shah was supported by the National Institutes of Health under award number T32-HD007009-45 to the University of Chicago. Jimmy Elias was supported by the National Institutes of Health under award number T32-GM007197 and R25-GM109439 to

the University of Chicago. This study was supported by the National Institutes of Health, under award numbers R01-GM115945 to Alexander J. Ruthenburg (The University of Chicago) and R01-DA036887 to Shohei Koide (New York University); and the American Cancer Society, under award number 130230-RSG-16-248-01-DMC to Alexander J. Ruthenburg.

Supplementary Notes

Supplementary Note 3.1: Configurations of bivalent nucleosomes and impact on avidity.

As each nucleosome has two H3 protomers, there are several different configurations of bivalency that a bivalent nucleosome can theoretically adopt, each with a different avidity for ChIP pulldown with immobilized antibody. At one extreme, with the highest avidity, is the symmetric *cis*-bivalency form, where both H3K4 and both H3K27 residues are trimethylated (Fig. 3.1E). This nucleosome has the most epitopes for antibody binding and will thus have the highest avidity in pulldown reflected in apical pulldown efficiency (Fig. 3.1D). At the other extreme, with the lowest avidity, is the *trans*-bivalency form, where single H3K4me3 and H3K27me3 marks decorate different histone tails (Fig. 3.1E). This has the fewest epitopes for antibody binding and will thus have no avidity in pulldown.

This poses a theoretical challenge in normalization and calibration of a ChIP study; because we cannot separately measure *trans*-bivalency, symmetric *cis*-bivalency, nor any intermediate states, it is impossible for us to definitively state whether a given locus with a given HMD has relatively few nucleosomes that are symmetric *cis*-bivalent or whether it has relatively many nucleosomes that are trans-bivalently modified. To accommodate for this limitation, we include two different bivalent calibrants in our set of nucleosome standards: one that is symmetric *cis*-bivalent and one that is *trans*-bivalent. The bivalency sequential ChIP can then be normalized to either one of these

standards, and because these two cases represent the limits of pulldown avidity, normalization to these calibrants will define the theoretical "range" in which true bivalency HMD (i.e. the proportion of nucleosomes with some bivalent configuration) exists (Fig. 3.1E). We note that, because the signal from calibration to these standards are scalar multiples of each other, we cannot uniquely distinguish these two configurations in the genome. Absent any prior information about the dominant configuration of bivalency, the proportion of bivalently modified nucleosomes at a given locus will exist in the range defined by calibration to symmetric *cis*- or *trans*-bivalent standards (Fig. 3.3A).

In practice, there are a few reasons why this is not a major concern. First, there is no mass spectrometry evidence that H3K4me3 and H3K27me3 exist on the same histone tail, despite specific enrichment for these marks and sensitive detection limits[116,145], suggesting that configurations other than *trans*-bivalency are at most, extremely minor in abundance. Second, the scarcity of these *cis*-tail modifications is consistent with the biochemical literature prior to this work that suggests the biogenesis of these *cis*-tail modifications is enzymatically challenging due to antagonistic allosteric effects (see Supplementary Note 3.4). Third, even if symmetric *cis*-bivalency does exist at some loci, for the purposes of tracking changes in bivalency across differentiation, we can still observe an increase or decrease in bivalency by this calibration method; we simply cannot precisely discern whether the effect is driven by nucleosomes gaining/losing *trans*-bivalency, *cis*-bivalency, or some combination of the two. The overall amount of bivalency would still increase or decrease in all those scenarios, and so long as our choice of calibrant remains consistent, we can still measure that change regardless of the calibrant that we use for our normalization. Therefore, though we have generated datasets using both calibrants, we present our bivalency pulldowns as calibrated to the trans-bivalent standards.

Supplementary Note 3.2: Definition and interpretation of HMD at promoters.

Throughout this study, we have defined gene promoters to be the region from 0 to +400bp relative to the TSS, representing the +1 and +2 nucleosomes of each gene. These nucleosomes tend to be well-positioned[229] and, accordingly, are most likely to provide us with adequate read depth to robustly quantify each histone modification. This definition is conservative; we find that H3K4me3 and bivalent domains, which tend to be peak-like, have a median breadth of 550bp at bivalent genes (Fig. 3.4D).

The width of these domains raises an important point regarding the measurement of histone modification density as a continuous variable. At a given nucleosome in a single allele of a single cell, there are only three possible states for a histone modification: symmetric, asymmetric or not present. However, nucleosome readers do not typically bind only a single nucleosome at a single position; rather, the local density of the modification across multiple nucleosomes is crucial in localizing these effectors through multivalent avidity-based interactions[91,230–232]. Indeed, we find that the HMD across sequential nucleosomes relative to the TSS is well autocorrelated (Fig. 3.4E). This means that the interpretation of the HMD across a multinucleosomal span becomes more nuanced; a given histone modification may exist at one or more of those nucleosomes. Accordingly, despite the fact that a single nucleosome is essentially ternary in whether it has a given histone modification or not (i.e. HMD of 0% or 100%), a region spanning multiple nucleosomes could have an intermediate HMD; it is this latter quantity that is most relevant for the biological function imparted to the nearby genomic regions, and this is the quantity we analyse through this work. Though it is not employed in this work, similar arguments would apply to analyses of HMD over larger spans, such as gene bodies.

Supplementary Note 3.3: Limits of HMD and impacts of avidity biases.

For the datasets presented in this work, the vast majority of promoters have a histone modification density between 0-100%, representing the proportion of nucleosomes at those promoters with the modification of interest (Fig. 3.4C, 3.5A). However, at some loci, the measured HMD exceeds 100%. There are several possible reasons for this.

The most important of these possibilities is low input depth. The ICeChIP datasets are normalized to the input read depth at every genomic interval to accommodate for differences in local nucleosome density when computing the HMD. However, this means that at regions that are relatively nucleosome-depleted, there will be few reads in the input, meaning that the denominator of the HMD computation is quite small (Methods). This increased Poisson noise in these regions of low input can result in inflated apparent HMD beyond the physical limit of 100%. To accommodate for this, we can compute 95% confidence intervals for the HMD of each modification at each genomic position, and these confidence intervals virtually always overlap the physically possible range of HMD values (e.g., Fig. 3.2C). In naïve mESCs, only 0.5% of the promoters have a bivalency HMD above 100%, and for the vast majority of these promoters (86.1%), the 95% confidence interval error estimate ranges below 100%. The fact the apparent bivalency HMD calibrated by trans-bivalent standards, is broadly constrained to less than 100% further supports the idea that this choice of calibrant is appropriate and not inflationary (Supplementary Note 3.1).

There are also several other possibilities that are more challenging to accommodate for. First, some regions of the genome are known to be more artefact-prone for sequencing and mapping[233]; if the IP sample is enriched for these sequences relative to the input, then that could be disproportionately represented in the IP and have an apparent HMD greater than 100%. Second, the antibodies themselves could skew the apparent HMD. If the antibody is capturing substantial

off-target material, then that will result in systematic inflation of the IP, resulting in an inflated HMD. Though ICeChIP barcoded nucleosome standards can help monitor off-target pulldown of some nucleosome species, we can only measure the capture of the standards that we actually have spiked into the experiment. If we do not have nucleosome standards available for a potential off-target modification, then we cannot definitively state that the antibody is not capturing that material. In this context, that is likely most important for H3K27me3 pulldowns; though we cannot state this definitively due to the lack of H3K27me2 standards, it is plausible that we are pulling down some amount of H3K27me2 with these IPs, resulting in slightly inflated apparent H3K27me3 HMD. However, this may not be too problematic; H3K27me2 and H3K27me3 are thought to be recognized by many of the same proteins and to have highly similar functions[81], so the conflation of the two – if present – likely does not pose a significant problem in ascribing biologic function.

On a related note, at some loci, the bivalency HMD goes below 0%. In naïve mESCs, 8.8% of the promoters have a bivalency HMD below 0%, yet for the vast majority of these promoters (90.8%), the 95% confidence interval error estimate ranges above zero. This is because we employ *in silico* signal-correction for the bivalency dataset to remove signal that is attributable to H3K9me3. In essence, we can measure the amount of H3K9me3 pulldown in our bivalency ICeChIP dataset due to nucleosome standards employed, and we can separately measure H3K9me3 HMD by a highly specific IP. We can then a linear combination correction matrix to remove the signal that is attributable to directly measured H3K9me3 at these loci. This method can effectively reduce the impact of modest off-target binding H3K9me3, but at some loci, will result in a subzero apparent HMD due to random sampling of read depth in the two distinct pulldowns employed.

Finally, at some sets of gene promoters, the trans-bivalency HMD is shown to be greater than the H3K4me3 or H3K27me3 HMD. This apparent discrepancy has a few possible reasons.

First, there is some nuance in the interpretation of HMD in the context of single-target ICeChIP and reICeChIP. A nucleosome has two copies of each of its core histone proteins, including histone H3. This means that there are two possible sites of modification on each nucleosome for for each individual modification; if only one of those sites is modified, then that corresponds to an HMD of 50% because only half the possible modification sites are actually modified. However, this is different for the trans-bivalency HMD; by definition, only one trans-bivalency modification pattern can exist on a given nucleosome at any given time. If two "trans-bivalent" modification patterns existed on the same nucleosome simultaneously, then both H3K4 and both H3K27 residues would be trimethylated – which is symmetric *cis*-bivalency. As such, if one H3K4 and one H3K27 residue are trimethylated, then 100% of the possible trans-bivalency configurations for the nucleosome of interest are satisfied, meaning that the trans-bivalency HMD will be 100%. However, in this case, the H3K4me3 and H3K27me3 HMDs will only be 50% because only half the modifiable residues are actually modified.

The other caveat is that symmetrically modified nucleosomes will be pulled down more efficiently than asymmetrically modified nucleosomes due to avidity effects, as can be seen in the pulldown of symmetric vs. asymmetric H3K4me3 and *cis*-bivalency vs. *trans*-bivalency (Fig. 3.3), and observed previously[118]. This means that calibration to symmetric nucleosome standards will have a larger denominator in computation of HMD and thereby yield lower apparent HMDs; this can also contribute to the lower apparent HMD of H3K4me3 and H3K27me3 relative to trans-bivalency. Accommodating for this phenomenon would require detailed profiling of asymmetric H3K4me3 (which is currently difficult due to the low quality of H3K4me0 antibodies), asymmetric H3K27me3 (which is not currently possible), and distinguishing between trans-bivalency and cis-bivalency (which is also not currently possible).

However, as noted in Supplementary Note 1, so long as the method of calibration remains consistent, increases in apparent HMD will still correspond to increases in the modification of interest. Whether that increase in the target modification is due to asymmetric modification becoming symmetric or due to new gain of the modification at a previously unmodified locus in an instantaneous subpopulation remains unclear, but in both cases, modification density is still being gained at that locus. As such, even with these caveats, we can still quantitatively compare different datasets to each other as we use consistent calibration standards.

Supplementary Note 3.4: Enzymology of installation of bivalency.

Intriguingly, the catalytic activity of the EZH2-PRC2 core complex on nucleosome substrates is potentiated by pre-existing H3K27me3[234,235], yet inhibited by H3K4me3, particularly when symmetric[145,210,214]. Conversely, symmetric H3K27me3 has been reported to modestly inhibit several of the human COMPASS-family complexes by qualitative assays, although only SET1 complexes were examined at the nucleosome level[213]. This presents a potential concern for our data – if the enzyme complexes that install these marks are mutually antagonized by the opposing mark, how might the widespread bivalency we observe arise? As the PRC2 effects are well established with detailed quantitative enzymology[145,210,214], which we recapitulate (data not shown), we deployed more quantitative HMTase assays with a larger panel of relevant nucleosomal substrates to evaluate the COMPASS/SET1B/MLL-family core complexes for allosteric modulation by pre-existing marks (Fig. 3.8).

Supplementary Note 3.5: Sensitivity and specificity of DEGs.

In this context, sensitivity refers to the proportion of DEGs that are represented in a specific class of genes (e.g. H3K27me3-dominant bivalent genes), whereas specificity refers to the proportion of

that class of genes that are differentially expressed. Under the prevailing bivalency model, bivalency is associated with poised genes that become upregulated or downregulated upon differentiation; as such, it should have high specificity for DEGs.

Supplementary Note 3.6: Generalized linear model evaluation and parameter selection.

The first way we evaluate different models for predicting DEGs is to compute the Bayes Information Criterion (BIC). Though not definitive, this metric estimates whether addition of a parameter to a model improves it more than expected from chance alone. When comparing two models, the model with the lower BIC will tend to have more explanatory parameters and/or fewer non-explanatory parameters than the model with the higher BIC. To this end, if BIC increases when a parameter is added, then it can be interpreted that the parameter being added contributes minimal additional explanatory power. Here, we find that adding bivalency to a model increases the BIC, meaning that it is likely (though not definitively) not contributing meaningfully more information in predicting DEG status in this differentiation paradigm.

A more definitive way to evaluate model accuracy is to use hold-out cross-validation. In this method, we split the set of all genes into two groups, one with 80% of the genes (the training set) and one with 20% of the genes (the testing set). We then train our GLMs on the training set and use the derived models to predict DEG status in the testing set. Hold-out cross-validation is a highly effective way of testing whether a model is overfit or underfit upon addition or removal of a parameter. If model accuracy increases substantially, then that would suggest the parameter has explanatory power over that provided by the other parameters. Conversely, if model accuracy decreases substantially, then that suggests that the additional parameter causes overfitting. Minimal

changes in model accuracy suggest that the additional parameter contributes little to the model over the existing parameters, positively or negatively.

There are two metrics we use to test the accuracy of the predictions in the testing set. The first is by logistic regression thresholding, in which the gene is predicted to be a DEG if the modelled probability is greater than 0.5. The second is by computing the area under the receiver operator characteristic curve to measure true and false positive rates using different modelled probabilities as the thresholds. Overall, we find that the GLM with bivalency barely changes model accuracy by either metric on hold-out cross-validation, with the magnitude of change being similar to that observed by instead adding H3K9me3 or DNA methylation. As such, we can interpret that none of these parameters – including bivalency – meaningfully contributes to the prediction of DEGs beyond that achieved with H3K4me3 and H3K27me3 in this system.

Methods and Materials

This section has been adapted from the following:

- Shah, R. N. *et al.* Examining the Roles of H3K4 Methylation States with Systematically Characterized Antibodies. *Molecular Cell* **72,** 162–177 (2018).
- Shah, R. N. *et al.* Re-evaluating the role of nucleosomal bivalency in early development. Preprint at *bioRxiv*, doi: 10.1101/2021.09.09.458948. (2021).

Cell Culture

Naïve mouse embryonic stem cells (mESCs) were grown from the mESC E14 line (129/Ola background) in high glucose DMEM (Invitrogen), supplemented with 15% (v/v) FBS (Gibco), 1% (v/v) non-essential amino acids (Gibco), 1x penicillin/streptomycin (Gibco), 0.1mM 2-mercaptoethanol

(Gibco), 2mM L-glutamine (Gibco), 1000U/mL LIF (ESG1107 Millipore), 3μM CHIR99021 (LC Laboratories), 1μM PD0325901 (LC Laboratories), sterilized using 0.1μm filter flask (Millipore), stored up to 1 week in 4°C.

Primed mESCs were grown from the mESC E14 line (129/Ola background) in high glucose DMEM (Invitrogen), supplemented with 15% (v/v) FBS (Gibco), 1% (v/v) non-essential amino acids (Gibco), 1x penicillin/streptomycin (Gibco), 0.1mM 2-mercaptoethanol (Gibco), 2mM L-glutamine (Gibco), 1000U/mL LIF (ESG1107 Millipore), sterilized using 0.1μm filter flask (Millipore), stored up to 1 week in 4°C.

Naïve and primed mESCs were grown on plates coated with 0.1% bovine gelatin (Sigma), grown to 70-90% confluence and passaged daily at a 1:3 ratio, with a media change 3 hours before passaging, supplemented with 1 vol. of fresh media 8 hours after passaging.

To start the adherent monolayer differentiation process to neuronal progenitor cells (NPCs; Day 0)[236,237], naïve mESCs cells were split onto a gelatinized 6 cm plate at 1 x 10^4 cells/cm^2 and allowed to grow for 24 hours. On Day 1, the media was switched to RHB-A (Takara, Y40001) and was subsequently changed every other day. On day 4, cells were split and plated onto Poly-L-Ornithine, laminin-treated 6-cm plates. Prior to cell seeding the plates were treated with 0.01% Poly-L-Ornithine (Millipore, A004C) for at least 20 min, followed by 5 ug/cm^2 of laminin (Fisher, CB40232) resuspended in basal RHB-A medium (Takara, Y40000). After washing off this treatment, cells were seeded in fresh RHB-A, supplemented with 10 ng/mL of bFGF (PeproTech, 100-18B) and EGF (PeproTech, 315-09). Cells were then split every 4 days at ≥ 20,000 cells/cm^2 until an appropriate amount of NPCs were cultured for ICeChIP.

Octamer Reconstitution

Symmetric H3K4me3, H3K9me3, H3K27me3, H3K36me3, H3K79me2, and H3K4me3K27me3 octamers were reconstituted from semisynthetic histones as previously described[91,118,166,167]. Recombinant core histones were expressed in BL21 (DE3) with pRARE2 and mixed to equimolarity with the relevant semisynthetic histones in freshly prepared filter sterilized Unfolding Buffer (50 mM Tris-HCl pH 8.0, 6.3 M Guanidine-HCl, 10 mM 2-mercaptoethanol, 4 mM EDTA) to a final concentration of ≥ 1 mg histone per mL. The histone reconstitution was then added to 3500 MWCO SnakeSkin dialysis tubing (Pierce) and dialyzed overnight at 4°C against 500-1000 volumes of filter sterilized Refolding Buffer (20 mM Tris-HCl pH 7.5, 2 M NaCl, 5 mM DTT, 1 mM EDTA).

After dialysis, the histone mixture was centrifuged at 18,000 g for 1 hour at 4°C, and subjected to gel filtration chromatography (Superdex 200 10/300 GL, GE Healthcare, resolved with Refolding Buffer). Each fraction that displayed a peak on the UV chromatogram was analysed by SDS-PAGE (22 mA current in 1x Laemmli Buffer for 70 minutes), stained with SYPRO Ruby (Bio-Rad) per manufacturer instructions, and imaged with a 610BP emission filter at 600V PMT setting. Octamer fractions with equimolar quantities of each core histone were pooled and concentrated (Amicon Ultra-4 Centrifugal Filters, 10,000 MWCO, Millipore) to 5-15 μM octamer, diluted with one volume of Octamer Storage Buffer, and stored at -20°C.

Asymmetrical H3K4me3 octamers were prepared as above, with modifications. Equimolar amounts of histone H2A, H2B, H3 and H4 were mixed in Unfolding Buffer to the total of 1-2mg, where 90% of histone H3 was trimethylated on Lys 4 and the remaining 10% was unmethylated and had a His_6-tag at N-terminus with TEV cleavage site. Octamers were reconstituted overnight by dialysis in 1000 volumes of Phosphate Refolding Buffer (50 mM sodium phosphate, 2 M NaCl, pH 7.5). Octamers were purified by S200 gel filtration chromatography, and his-tagged octamers

were isolated using cobalt-based immobilized metal affinity chromatography Dynabeads magnetic particles. Octamers were incubated with magnetic beads for 10 min at 4°C on a rotator, followed by two 1 ml washes with Octamer Wash Buffer (50 mM sodium phosphate, 2 M NaCl, 10 mM imidazole, pH 7.5), then eluted six times, each with 50 µL of Octamer Elution Buffer (50 mM sodium phosphate, 2 M NaCl, 250 mM imidazole, 1 mM EDTA, 1 mM DTT, pH 7.5). Fractions were characterized spectroscopically, pooled, diluted with one volume of Octamer Storage Buffer, and stored at -20°C.

Asymmetrical *trans*-bivalent H3K4me3-H3K27me3 octamers were prepared similarly to symmetrical octamers with the following differences. Histones H2A, H2B, H4, and asymmetric disulfide-linked histones H3K4me3-H3K27me3 were mixed in a 1.2:1.2:1:0.5 molar ratio. The remaining steps were done as above, but no reducing agents were used until octamer particles were formed.

Nucleosome Reconstitution

Nucleosomes were reconstituted onto 147bp DNAs composed of the core Widom 601 sequence[168] modified with a 22bp barcode on each end, with each barcode composed of two distinct 11bp sequences not found in the human or mouse genomes. The DNA and octamer were mixed to a final concentration of 1µM each in 2 M NaCl, and then dialyzed in dialysis buttons (Hampton Research) and a 10,000 MWCO SnakeSkin dialysis membrane (Pierce) against 200 mL of Refolding buffer for 10 minutes. Dialysis then continued as 2L of Buffer I0 (20 mM Tris-HCl pH 7.5, 1 mM EDTA, 1mM DTT) was added (flow rate 2-2.5 mL per minute).

Dialyzed samples were diluted with an equal volume of Nucleosome Dilution Buffer (20 mM Sodium Cacodylate pH 7.5, 10% v/v glycerol, 1 mM EDTA, 10 mM 2-mercaptoethanol, Filter

Sterilized), and 1 µl was analysed by native PAGE (100 V in 1x TBE for 30 minutes), stained with SYBR Gold in 1xTBE for one hour, and visualized with a UV transilluminator gel imager. Fractions containing nucleosomes and minimal free DNA were pooled and diluted to a working concentration of ~ 1 nM with filter sterilized Nucleosome Storage Buffer (10 mM Sodium Cacodylate pH 7.5, 100 mM NaCl, 50% v/v glycerol, 1 mM EDTA, 1x Protease Inhibitor Cocktail [1 mM PMSF, 1mM ABESF, 0.8 µM aprotinin, 20 µM leupeptin, 15 µM pepstatin A, 40 µM bestatin, 15 µM E-64 from a 200x DMSO stock]) and stored at -20°C.

Design, Expression, and Purification of 304M3B-1xHRV3C

The 304M3B-1xHRV3C Fab is based on previously described Fab 304M3B(PDB:4YHZ)[208]. The gene encoding the Fab was modified to contain HRV3C cleavage site at the C-terminus of the heavy chain. To that end, we inserted SSSLEVLFQGP (AGC AGC AGC CTT GAA GTC CTC TTT CAG GGA CCC) sequence just after the position T229 of heavy chain (numbered as in PDB:4YHZ) and before biotinylation acceptor peptide (GLNDIFEAQKIEWHE)[238]. The Fab was expressed in the 55244 strain of *E.coli* in the TBG media (Terrific Broth (FisherBrand), 0.8% (v/v) glycerol) with 100 µg/ml carbenicilin, grown for 24 hours, at 30°C, 200 rpm in the Fernbach non-baffled flasks, with constricted airflow. Fab was purified using Protein G-A1[109] affinity chromatography, followed by cation-exchange chromatography (Resource S, GE Healthcare). Purified Fab was *in vitro* biotinylated using BirA biotin ligase.

ICeChIP Input Preparation

Input was prepared for ICeChIP and reICeChIP experiments as previously described[118,124,170,171]. Briefly, cell pellets were washed twice with 5 mL of PBS, then washed twice with 5 ml of filter sterilized Buffer N, with each wash consisting of complete resuspension of the pellet, centrifugation

at 500 g for 5 minutes at 4°C, and removal of supernatant. The washed pellet was then resuspended in at least 2 packed cell volumes (PCV) of Buffer N and mixed with 1 volume of 2x Lysis Buffer and incubated on ice for 10 minutes to lyse cells.

The crude nuclei were spun down at 500 g for 5 minutes at 4°C before being resuspended in at least 6 packed nuclear volumes (PNV) of Buffer N and applied to the top of 7.5 mL of filter sterilized Sucrose Cushion N in a 15 ml centrifuge tube, then spun down at 500 g for 12 minutes at 4°C in a swinging-bucket rotor. The supernatant was discarded, and the pellet resuspended in ~ 2 PNV of Buffer N.

The nucleic acid content of the nuclei per unit volume was quantified by diluting 2 µL of nuclei suspension into 48 µL of 2 M NaCl, water-bath sonicating to solubilize DNA, and spectroscopically measuring nucleic acid concentration by Nanodrop (where one A_{280nm} = 50 ng/µL chromatin). After accounting for the 25-fold dilution of the measurement sample, the concentration of the nuclei was adjusted to 1 µg/µL of chromatin. Nuclei were dispensed to 100 µL aliquots, flash frozen, and stored at -80°C prior to use.

For use, nuclei aliquots were thawed and spiked with ~ 1 µl of each barcoded nucleosome standard per 50 µg of chromatin. This suspension was then mixed by pipette, transferred to a new tube, and warmed to 37°C for 2 minutes. 1 unit of micrococcal nuclease (MNase, Worthington) per 4.375 µg of chromatin was added, and samples incubated at 37°C while shaking at 900 rpm for 12 minutes. Digestions were stopped by adding 1/9 volume of filter sterilized 10x MNase Stop Buffer while slowly vortexing, and nuclei lysed by adding 5 M NaCl to a final concentration of 600 mM while slowly vortexing. 66 mg of HAP resin (BioRad, CHT^{TM} Ceramic Hydroxyapatite, Type I, 20 um) per 100 µg of chromatin digested was rehydrated with 200 µl of filter sterilized HAP Buffer 1 per 100 µg of chromatin digested. Lysed nuclei were centrifuged at 18,000 g for 1 minute to pellet

insoluble nuclear debris, and the soluble fraction added to the rehydrated HAP resin and incubated for 10 minutes at 4°C with rotation.

After incubation, the HAP resin slurry was added to a centrifugal filter unit (Millipore Ultrafree MC–HV Centrifugal Filter 0.45 μm) and spun at 1000 g for 30 seconds at 4°C. The HAP resin left on the filter unit was then washed 4 times with 200 μL HAP Buffer 1, and 4 times with 200 μl filter sterilized HAP Buffer 2 by spinning at 1000 g for 30 seconds at 4°C. HAP resin was eluted into a clean tube with three 100 μl solutions of filter sterilized HAP Elution Buffer. The nucleic acid content of the elution was then quantified by Nanodrop.

Antibody Preparation for ICeChIP

Antibodies and quantities used for each ICeChIP experiment are shown in Appendix A. With the exception of the 304M3B-1xHRV3C and 309M3B antibodies, the indicated amount of Protein A Dynabeads (Invitrogen) for each ICeChIP was washed with 50 μL of ChIP ChIP Buffer 1 by use of a magnetic rack, then resuspended in 50 μL of ChIP Buffer 1. In a separate set of tubes, the antibody was diluted to 100 μL with ChIP Buffer 1. The antibody and Protein A Dynabead suspensions were combined and incubated on a rotator at 4°C for at least one hour, then washed with 200 μL of ChIP Buffer 1 by use of a magnetic rack and resuspended in 50 μL of ChIP Buffer 1.

The antibodies 304M3B-1xHRV3C and 309M3B were prepared similarly with Streptavidin M-280 Dynabeads (Invitrogen) rather than Protein A Dynabeads. The beads were washed, and antibodies added and incubated as above. After incubation, the beads were washed twice with 200 μL of ChIP Buffer 1 by use of a magnetic rack. They were then washed twice with 200 μL of ChIP Buffer 1 supplemented with 5 μM of biotin by incubating for 10 minutes at 4°C on a rotator, then removing supernatant by use of a magnetic rack.

Standard ICeChIP Immunoprecipitation

After antibodies were prepared and washed, the input chromatin concentration adjusted to 20 ng/µl with filter sterilized ChIP Buffer 1, and the amount of chromatin specified in Appendix A was added to each antibody-bead conjugate and incubated for 15 minutes on a rotator at 4°C. Beads were then washed twice with filter sterilized ChIP Buffer 2 and once with filter sterilized ChIP Buffer 3, with a wash consisting of removal of the existing supernatant by use of a magnetic rack, resuspension into 150 µl of buffer, transfer to a new siliconized tube, and incubation on the rotator for 10 minutes at 4°C. After these washes, the supernatant was removed, the beads resuspended in ChIP Buffer 1, transferred to a new siliconized tube, rinsed once with 200 µl of TE before being resuspended in 50 µl of ChIP Elution Buffer and incubated at 55°C for 5 minutes.

After incubation, the supernatant was transferred to a new set of siliconized tubes, and the beads discarded. To each supernatant was then added 2 µl of 5 M NaCl, 1 µl of 500 mM EDTA, and 1 µl of 10 mg/mL Proteinase K. 15 µl of Input DNA was also diluted to 50 µl with 35 µl of ChIP Elution Buffer and was supplemented with 2 µL of 5 M NaCl, 1 µL of 500 mM EDTA, and 1 µL of 10 mg/mL Proteinase K. The IP elutions and diluted input were then incubated at 55°C for 2 hours for a Proteinase K digestion. After digestion, the DNA was purified by adding 1.5 volumes of Serapure HD, incubating at room temperature for 15 minutes, then collecting the beads on a magnetic rack, washing twice with 150 µl of 70% ethanol, and eluting into 50 µl ddH2O, which was then recovered and stored at -20°C.

reICeChIP Immunoprecipitation

After antibodies were prepared and washed, the input chromatin concentration adjusted to 20 ng/µl with filter sterilized ChIP Buffer 1, and the amount of chromatin specified in Appendix A was added

to each antibody-bead conjugate and incubated for 15 minutes on a rotator at 4°C. After incubation, the beads were washed three times with ChIP Buffer 1, with a wash consisting of removal of the existing supernatant by use of a magnetic rack, resuspension into 150 µl of buffer, transfer to a new siliconized tube, and incubation on the rotator for 10 minutes at 4°C. The chromatin was then eluted into 20 µL of ChIP Buffer 1 supplemented with 4 µg HRV3C (GE Healthcare) by incubating on ice for 60 minutes. The elution was saved and repeated once more; both elutions were then combined. With the eluted sample, ICeChIP was conducted against H3K27me3 as per the Standard ICeChIP Immunoprecipitation instructions with the antibody and resin quantities in Appendix A.

DNA Quantification and Analysis by Quantitative PCR

To assess local histone modification density and/or antibody specificity, our DNA from the ChIP experiments was quantified by quantitative PCR (qPCR). qPCR was conducted using TaqMan Gene Expression Master Mix (Applied Biosystems) using the primers and hydrolysis probes previously described[118]. These primers and probe for the barcoded sequences were previously qPCR validated for effectiveness and quality[118]. Primers were used at 900 nM; hydrolysis probe at 250 nM, in the TaqMan Gene Expression Master Mix (Applied Biosystems). The qPCR program was run at 95°C for 10 minutes, followed by 40 cycles, each consisting of 15 seconds at 95°C followed by 1 minute at 60°C and concluding with a plate read.

Cq values were analysed using the ΔΔCq method. Briefly, the Cq values for each target for each sample were averaged together to obtain the mean Cq value. Enrichment for each barcode was then computed as Enrichment $= 2^{Cq_{INPUT} - Cq_{IP}} * 10$, accounting for the 10-fold dilution of Input relative to IP and multiplying by 100% for Enrichment as a percentage of target. Off-target binding

to alternate PTMs were computed by normalizing each enrichment to that of the on-target PTM: referred to as "Specificity (% Target)".

Illumina Library Preparation and Sequencing

Illumina libraries were prepared as described[118], with minor modifications. Briefly, Serapure purified DNA was quantified using Quant-iT^{TM} PicoGreen (Thermo Fisher) as per manufacturer instructions. Libraries were then generated from up to 10 ng of each DNA sample (input or IP) with the NEBNext Ultra II DNA Library Prep kit (New England Biolabs) per manufacturer instructions. The DNA content of each library was then quantified and pooled for Illumina sequencing. Cluster generation and paired-end sequencing was conducted using standard Illumina next-generation sequencing protocols by the University of Chicago Genomics Facility on the Illumina NextSeq.

Next-Generation Sequencing Alignment and HMD Computation

To align reads, a reference genome was first created, which consisted of the either human genome (GRCh38/hg38) or the mouse genome (mm9) appended respectively by the sequences of each of the nucleosome standard barcodes. Reads were then mapped to the appropriate reference genome using Bowtie2 using the sensitive pre-set and end-to-end alignment options[172]. Using SAMTools[173], any reads which were not paired, not mapped in a proper pair, or mapped with a map quality < 20 were discarded to prevent low-quality reads from impacting downstream analyses. Reads were then flattened to create a single mapping from each matched pair of reads by retaining only one fragment per pair, and any mappings with lengths > 200bp were also discarded to ensure only mononucleosomes were being analyzed[118].

Bedgraphs of genome coverage were then generated using BEDTools[174], and IP / input genome coverage bedgraphs were merged using BEDTools[174]. The sum of reads across ladder

members for each nucleosomal standard was computed for each sample and HMD bedgraphs were then generated from the merged bedgraphs using awk to apply the following formula:

$$\text{HMD } (\%) = 100\% * \frac{\text{IP}_{\text{locus}}/\text{Input}_{\text{locus}}}{\text{IP}_{\text{barcode}}/\text{Input}_{\text{barcode}}}$$

Error and 95% confidence intervals were computed with Poisson statistics and error propagation from the merged bedgraphs using awk to apply the following formula:

$$\text{95CI Error } (\%) = 1.96 * \text{HMD } (\%) * \sqrt{\frac{1}{\text{IP}_{\text{locus}}} + \frac{1}{\text{Input}_{\text{locus}}}}$$

Bigwig files were generated for visualization using the bedGraphToBigWig tool[175]. For computation of HMD for bivalency, the *trans*-bivalency standard was used.

Correction was conducted using the H3K9me3 and *trans*-bivalency HMD datasets. using our previously described method[118] against H3K9me3 and *trans*-bivalency off-target binding. Briefly, measured HMD by each antibody can be described by a vector M, and the measured specificities by each antibody described by a square matrix S. Then, we can state, if other off-target binding is negligible, that the correct HMDs for H3K4me1, H3K4me2, H3K4me3, and H4K20me3 can be expressed by the vector C such that $M=CS$. As such, the vector C can be computed as $CSS^{-1} = C = MS^{-1}$. The elements of S^{-1} were then used to compute the HMD and Error of the corrected profiles using awk to linearly combine the two HMD profiles.

For all analyses, the HMD averaged over the N+1 and N+2 nucleosomes (taken to be 0 to +400bp into the gene body) was employed as representative of the promoter—this captures the most substantial H3K4me3 and H3K27me3 enrichment.

Genomic browser views were made using IGV. Heatmaps and gene ontology analysis was made using Homer software[178]. Further analysis and sectioning of data was conducted in R using the R code provided in Data and Software Availability.

Analysis of External Data

Bisulfite sequencing data was obtained from GEO series accession number GSE41923, dataset accession IDs GSM1027571, GSM1027572, GSM1027573, and GSM1027574. Methylation count files were obtained for each dataset and lifted to mm10. The average methylation for each promoter was then calculated for the 0 to +400bp region relative to the TSS of Refseq promoters using BEDTools.

Bulk RNA-seq data was obtained from GEO series accession numbers GSE108832 and GSE65697, dataset accession IDs GSM2913929, GSM2913930, GSM2913931, GSM1603282, GSM1603283, GSM1603284, GSM1603285, GSM1603286, and GSM1603287. Pseudoalignment was conducted against the Refseq mm10 transcriptome using kallisto[239] with fragment length mean and standard deviation of 200 and 20, respectively, and 100 iterations. Pseudoalignments were then loaded into R for differential expression analysis using sleuth[240], with correction for batch effects between primed mESCs and NPCs due to contribution to principal components of the same. Differentially expressed genes were identified as $q \leq 0.05$. Single-cell RNA-seq data was obtained from GEO series accession number GSE113417 and aligned as above with kallisto.

Suz12 ChIP data to measure PCR2 localization for WT, Ezh2 KO, and Ezh1 KO/Ezh2 KO cells was obtained from GEO series accession number GSE116603, dataset accession IDs GSM3243624, GSM3243625, and GSM3243626. Peak files were obtained for all these datasets

lifted to mm10. Ezh2 peaks were identified as peaks lost in Ezh2 KO relative to WT cells. Ezh1 peaks were identified as peaks lost in Ezh1 KO/Ezh2 KO relative to Ezh2 KO cells.

Set1A ChIP data was obtained from GEO series accession number GSE98988, dataset accession IDs GSM2629676, GSM2629677, GSM2629678, and GSM2629691. FastQ files were downloaded for the input and ChIP datasets for each replicate, then aligned to mm10 using Bowtie2 in end-to-end mode with the sensitive preset. Peak calling was then conducted on the alignments with MACS2[176], and consensus peaks for each replicate were identified.

Mll2 ChIP data was obtained from GEO series accession number GSE78708, dataset accession number GSM2073022. Peaks were obtained and lifted to mm10.

Methyltransferase assays

Enzymatic complexes were procured from Reaction Biology Corporation. Methyltransferase reactions were done using 200nM hsMLL1 (3745-3969), 200nM hsMLL2 (5319-5537), 400nM hsMLL3 (4689-4911), 200nM hsMLL4 (2490-2715), 800nM hsSet1A (1418-1707), or 800nM hsSet1B (1629-1923), in acomplex with hsWDR5 (22-334), haRbBP5 (1-538), hsAsh2L (2-534), 2x (hsDPY-30(1-99)), supplemented with 4% (v/v) RBC MLL enhancer (Reaction Biology Corp); 800nM hsEzh1 (2-747) or 120 nM hsEzh2 (2-746), in a complex with hsAEBP2 (2-517), hsEED (2-441), hsRbAp48 (2-425) and hsSUZ12 (2-739) supplemented with 3.6mM hsJarid2 (119-574) provided by Dr.Peter Lewis's laboratory. 30 ng/µl of semi-synthetic nucleosome substrate, 10µM [^{3}H]-SAM (50-80 Ci/mmol, Perkin Elmer Health Sciences), and enzymatic complexes were mixed in the Reaction Buffer (50 mM TrispH8.0, 91 mM NaCl, 5 mM $MgCl_2$, 1 mM DTT, 10% glycerol, 1 mM PMSF) and incubated at 30°C. At designated time points, 4 µl of reactions were spotted on P81 Ion Exchange Cellulose Chromatography Paper (Reaction Biology Corp). Spotted paper was

washed 4 times with 250 ml of 50 mM $NaHCO_3$ pH 9.0, for 5 minutes on a platform shaker, briefly washed with acetone, air-dried and immersed in scintillation fluid. ^3H decay rate was measured by scintillation counter (LS 6000IC, Beckman).

Data and Software Availability

ICeChIP-seq data generated for this study has been deposited at the Gene Expression Omnibus (GEO) under accession numbers GSE108747 and GSE183155. R markdown file for analysis and sectioning of datasets is provided at `https://www.github.com/shah-rohan/bivalency/`.

CHAPTER 4: QUANTIFYING INTERNAL HISTONE MODIFICATIONS WITH DENATURATIVE ICECHIP

Attributions

The ICeChIP-seq datasets used for Fig. 4.14 were generated by Bill Richter, Ph.D.'20, and previously published as: Richter, W. F. *et al.* Non-canonical H3K79me2-dependent pathways promote the survival of MLL-rearranged leukemia. *eLife* **10**, e64960 (2021). All other work was conducted by the author.

Abstract

Though valuable for pulldowns of modifications on the highly accessible histone tails, native ChIP often fails to specifically target modifications on the globular domain of the nucleosome core particle, making it difficult to understand the role of these internal modifications. Though previous reports have indicated that denaturative ChIP methods involving sonication may enable specific capture of these internal modifications, such sonication-based protocols suffer from an inability to separate the process of chromatin fragmentation from that of epitope exposure, making it challenging to reliably achieve optimal levels of both functions. Here, we present denaturative ICeChIP, a robust method to reproducibly pull down internal modifications with high specificity. We establish a novel paradigm of denaturative ChIP in which we separate the processes of fragmentation and denaturation, allowing for more complete and reproducible crosslinking and denaturation of chromatin. We further use this denaturative ChIP method to study H3K79me2, an internal modification critical for the survival and proliferation of MLL-rearranged leukemias, ultimately identifying a potential cross-talk pathway between H3K79me2 and H3K27me3. Our work thus highlights the importance

of using reproducible methods and demonstrates the power of quantitative data to identify new pathways of biological function.

Introduction

Broadly speaking, the nucleosome has two regions with distinct structural characteristics: the tails and the globular domain (Fig. 4.1)[28]. The tails, which are distal to the center of the nucleosome, are largely unstructured[241,242] and are highly accessible to solvent[243]. This has several implications, primarily driven by the ease of accessing and interacting with the histone tails. First, the tails interact with other portions of the nucleosome (including other histone protein regions or DNA elements), increasing the stability of the complex as a whole[243–247]. Second, the histone tails interact with other nucleosomes and/or histone tails on other nucleosomes, facilitating compaction and organization into nucleosome arrays and fibers[245–250]. Third, the histone tails are highly accessible to other proteins, including histone modifying complexes as well as the protein binding partners that "read" histone PTM patterns. Contained within these highly accessible tails are residues harboring some of the best-studied histone modifications, including H3K4, H3K9, H3K27, H4K16, and H4K20[28].

By contrast, the globular domain has been less well characterized. Compared to the freely accessible and poorly structured histone tails, the nucleosome globular domain has a much more clearly defined and organized structure (Fig. 4.1). The residues here are also much less accessible; many residues are buried inside the core of the globular domain, and even those that are relatively more solvent-accessible tend to be more sterically restrictive to binding than those in the disordered environment of the tails. There are still some modifications on this globular domain (i.e. internal modifications), most notably H3K79 methylation[23,251,252], but these modifications tend to be much more poorly studied than the modifications on the tail.

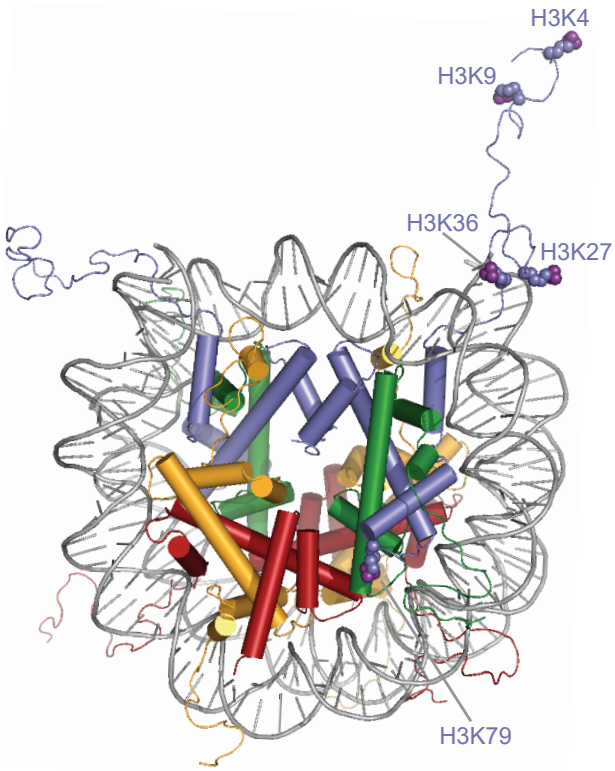


Figure 4.1: Nucleosome with select tail and internal residues highlighted.
Structure of the nucleosome with select H3 residues highlighted. H3K4, H3K9, H3K27, and H3K36 are on the histone tail. H3K79 is an internal modification and is located on the globular domain. Adapted from Werner and Ruthenburg[75].

This difference in relative understanding of tail and internal modifications is driven in part by the relative accessibility and structure of the two regions. The tail represents a highly accessible and disordered region without meaningful secondary structure[241,242]. This means that, to a reasonable first approximation, the native local structure of a tail PTM is similar to that of a peptide with the local primary structure and modification of interest. Like the histone tail, these short peptides have little to no secondary structure[253], meaning that they can present a reasonably similar binding interface to potential binding partners as that same PTM would *in vivo*. Modified peptides also have the advantage of being straightforward to synthesize in relatively large quantities with very high purity[133,254,255], offering a standardized substrate that approximates the tail modification of interest.

This structural similarity is important not just for biochemical studies (e.g. binding and competition assays), but also for genomic studies such as ChIP-seq. Antibodies against specific proteins or histone modifications are generated by immunizing an animal with a peptide bearing the modification of interest, then purifying those antibodies that bind to the modification of interest[256–259]. Because modifications on the histone tail can be structurally approximated by peptides, it is theoretically straightforward to raise an antibody that can bind to a tail modification by immunization against a readily available modified peptide. Though this is often more challenging in practice[108,111,118,124], there are nonetheless numerous antibodies that can be used to specifically recognize and purify nucleosomes bearing tail modifications in the context of a native ChIP experiment[118,124], meaning that many anti-tail-PTM antibodies can both access and recognize their targets in their native conformations.

This is not as straightforward for internal modifications. Internal modifications are located on the globular domain of the nucleosome core particle, which is much less accessible and more highly structured. This means that a short peptide, which lacks secondary structure of note, will not be as representative of the native structure of the modification in the context of the native nucleosome, meaning that antibodies that are raised against peptides may not be able to recognize the target PTM in its native, highly structured conformation. Even if the antibody is able to recognize the modification in that context, the structure of the globular domain may hinder the approach and binding of the antibody to the PTM target, making the pulldown more challenging. The result is that pulldowns of internal modifications often proceed less efficiently and less specifically than pulldowns of tail modifications (Fig. 4.2), resulting in high off-target capture and inflated apparent HMDs, often in excess of the physical limit of 100%[118].

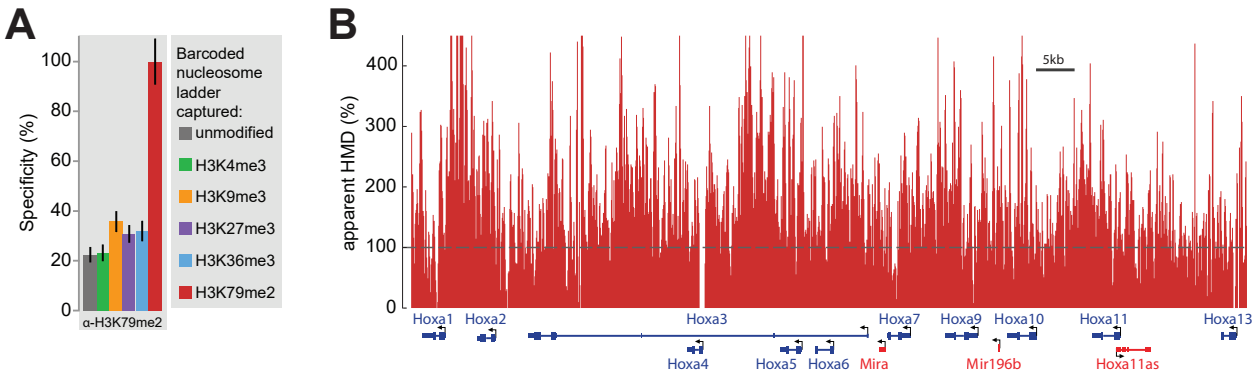


Figure 4.2: Poor measurement of H3K79me2 with native ICeChIP.
(A) Specificity of native ICeChIP pulldown of H3K79me2. **(B)** Apparent HMD from H3K79me2 pulldown by native ICeChIP, showing highly inflated apparent HMD. Dashed line represents HMD of 100%, the physical limit of histone modification density. Adapted from Grzybowski et al., 2015[118].

One modification that is particularly difficult to study for this reason is methylation of histone H3K79 (H3K79me). This mark, located on the globular domain of the nucleosome core particle (Fig. 4.1), was first described in *S. cerevisiae*, where it is installed by the highly conserved enzyme Dot1[260–263]. Dot1 or Dot1L (Dot1-like) knockouts in *S. cerevisiae*[262], *D. melanogaster*[264], or *M. musculus*[265] show global abrogation of H3K79 methylation, suggesting that it is the sole methyltransferase responsible for installing this mark. Though very abundant in yeast, this modification is relatively rare in humans, typically comprising fewer than 4% of the histones in a variety of cell lines[135]. Nonetheless, the modification has been shown to be physiologically important. Early on, H3K79 dimethylation (H3K79me2) was found to be associated with actively transcribed genes[266], and it has been shown that Dot1L disruption hampers hematological and immunological functions such as erythropoiesis[267] and antiviral immune activation[268].

Yet another clue to the function of H3K79 methylation came from the proteins associated with Dot1L in clinical contexts. One of the most prevalent classes of leukemia in infants is the MLL-rearranged leukemias, which harbor a translocation of the mixed lineage leukemia (MLL) gene[269,270].

The MLL gene itself is an H3K4 methyltransferase featuring a catalytic SET domain[271]. However, in MLL-rearranged leukemias, the SET domain is truncated from the full protein, with the remainder of the protein being translocated to a fusion partner on another protein[272,273], ultimately driving transcription at their target genes[274]. Many of the MLL fusion partners, however, were members of the Dot1L complex[275–279]. Given the association between H3K79 methylation and transcriptional activity, this suggested that H3K79 methylation was dysregulated at MLL-fusion target genes, a notion that was confirmed by subsequent studies[23]. Even more strikingly, it was later found that H3K79me2 is essential for the survival and proliferation of MLL-rearranged leukemias[251,252,269], with pharmacological inhibition of Dot1L killing leukemic cells and suppressing the tumor in preclinical[280–282] and clincial[283] studies.

And yet, despite this modification's clinical significance, H3K79me2 presents a challenge for chromatin immunoprecipitation, with native ChIP failing to specifically capture it (Fig. 4.1). However, in 2014, Orlando et al. presented new insights into ChIP for H3K79me2 in a study describing their new method ChIP-Rx[128]. ChIP-Rx is conceptually similar to ICeChIP, relying on the principle that an exogenous spike-in normalization standard is necessary to measure differences in global abundances between samples. However, rather than using the nucleosome standards used in ICeChIP, Orlando et al. spiked in chromatin from *D. melanogaster* as an exogenous reference material; this allowed them to normalize their ChIP-seq profiles across the human cells to the total number of reads mapped to the internally invariant *D. melanogaster* chromatin[128].

To test their method, Orlando et al. conducted pulldowns in pools of MV4;11 cells, a cell line derived from MLL-rearranged leukemia. They cultured cells in the presence or absence of the specific Dot1L inhibitor pinometostat at a concentration that would result in near-total ablation of H3K79me2 globally. They then mixed these cells in varying proportions and conducted ChIP-

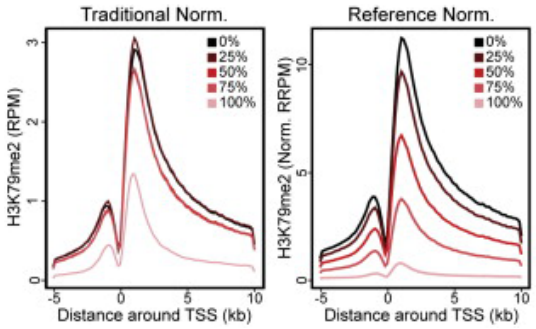


Figure 4.3: Exogenously normalized denaturative ChIP for H3K79me2.
ChIP-Rx of H3K79me2 with traditional read depth normalization (left) and normalization to exogenous read depth (right). Percentages represent the proportion of cells in each pool that were treated with Dot1L inhibitor pinometostat. Adapted from Orlando et al., 2014[128].

Rx on the samples against H3K79me2. Strikingly, they found that with their method, they could see a decrease in the exogenously normalized H3K79me2 ChIP-seq signal, concomitant with the proportion of inhibitor-treated cells in the pool used for the ChIP-seq[128] (Fig. 4.3). Notably, in the pool with 100% of cells treated with pinometostat, the This finding showed that ChIP-Rx was sensitive to global changes in histone modification abundance and could reveal such changes by use of exogenous normalization. But even more fundamentally, this result showed that it was possible to specifically immunoprecipitate H3K79me2; the fact that the 100% inhibitor-treated sample showed very little pulldown of H3K79me2 implied that there was little off-target binding and was thus indicative of apparent specificity.

The greatest apparent difference between ICeChIP and ChIP-Rx, apart from the type of calibrant used, was the form of ChIP-seq employed. ICeChIP is a native protocol, largely keeping the nucleosome in its folded state and relying on micrococcal nuclease (MNase) digestion for chromatin fragmentation[118]. Such a protocol tends to improve the pulldown specificity[150], but at the cost of inhibiting pulldowns of internal modifications. By contrast, ChIP-Rx utilized crosslinking and sonication to shear chromatin, denaturing the nucleosome in the process[128]. We hypothesized

that this denaturation was able to unfold the globular domain of the nucleosome core particle and better expose the H3K79me2 epitope, thereby making the presented antigen more closely resemble an unstructured peptide and making it more accessible.

Based on this hypothesis, we sought to develop a form of ICeChIP that utilized denaturation to expose the epitope and permit specific pulldowns of internal modifications. Here, after a close examination of sonication-based ChIP, we have developed denaturative ICeChIP, which uses thermal denaturation to reliably denature nucleosomes and specifically pull down H3K79me2. We then use denaturative ICeChIP to study the role of H3K79me2 in MLL-rearranged leukemias, finding genes with transcriptional dysregulation potentially explainable by changes in H3K79me2 and identifying new patterns of histone PTM crosstalk in that context.

Results

Sonication in denaturative ChIP

As a starting point, we first attempted to use the protocol described by Orlando et al. to assess its usefulness as a basis for our denaturative ICeChIP method. This method was previously described to have nearly 100% variability in its normalized enrichment measurements[128], and our ChIP-Rx-qPCR measurements recapitulated that finding (Fig. 4.4), indicating to us that this method did not provide sufficiently precise measurements to be useful as a basis for a quantitative ChIP protocol. Even more concerningly, we found that the trend of the change in normalized enrichment across replicates was different when normalizing to different genes, suggesting high variability in pulldown of even the invariant chromatin (Fig. 4.4). This problem was likely driven by the low enrichment at the target genes and would likely be resolvable by ChIP-Rx-seq, but the fact that the method was unreliable for ChIP-qPCR also indicated that it was suboptimal.

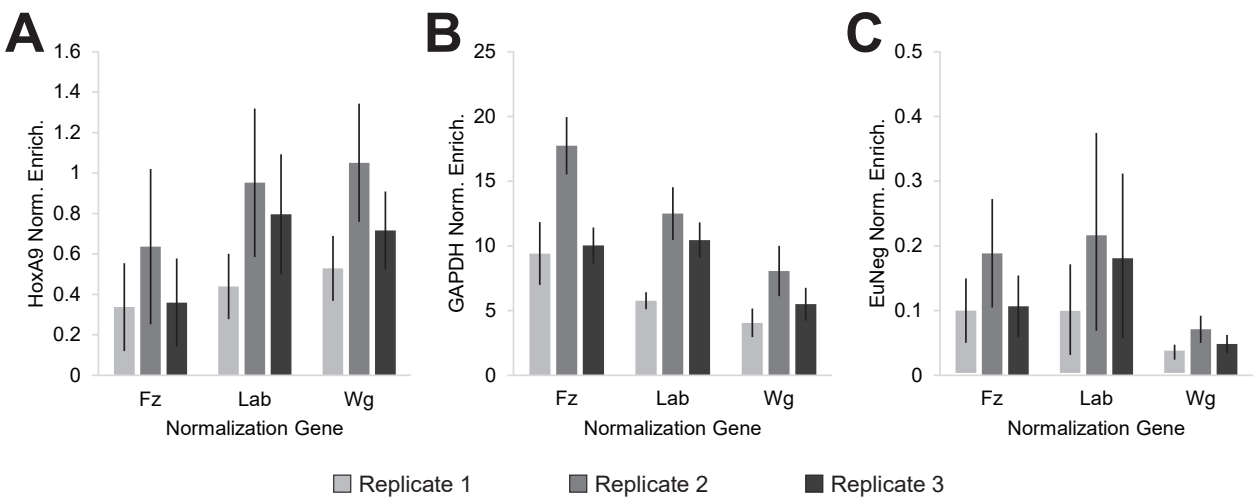


Figure 4.4: High variability in ChIP-Rx protocol.
Normalized H3K79me2 enrichment across three replicates at **(A)** HoxA9 promoter, **(B)** GAPDH promoter, and **(C)** EuNeg locus in K562 cells, normalized to *D. melanogaster* gene on X-axis.

We thus sought to develop a more robust and reproducible form of denaturative ChIP for use with our quantitative internal standards. Based on the hypothesis that the sonication of the chromatin was responsible for denaturing the nucleosome core particle and making the epitope more accessible and recognizable to the antibody, we first attempted to use sonication as our shearing method rather than MNase, much like Orlando et al.[128] (Fig. 4.5A). Rather than spiking in exogenous cells, however, we spiked in our nucleosome standards to the nuclei mixture immediately prior to cross-linking in the hopes that this would remove one element of variability from the procedure and improve reproducibility (Fig. 4.5A).

Our first goal was to better understand the effects of sonication on the efficiency and specificity of internal modification pulldowns. To do this, we sonicated our samples for 10, 20, 30, or 60 minutes to fragment and denature the chromatin. As expected, increasing the sonication time decreased the size of the fragments, with the 30-minute sonication time resulting in a roughly mononucleosomal population (Fig. 4.5B). Increasing the amount of sonication applied from 10 minutes to 30 minutes resulted in increased efficiency of H3K79me2 pulldown (Fig. 4.5C), sug-

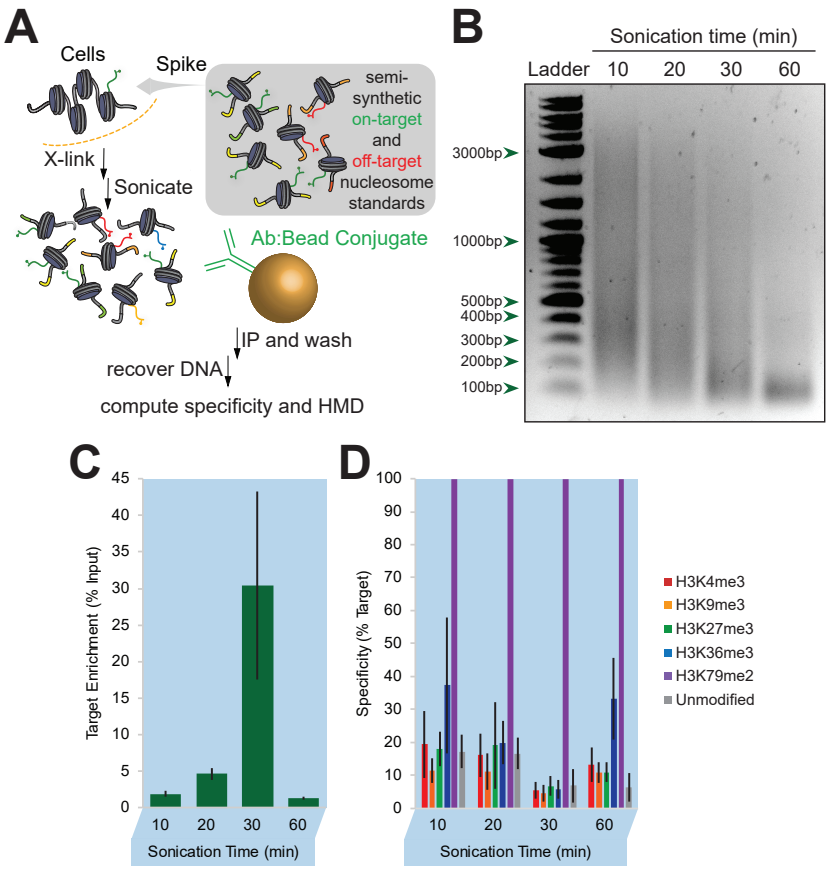


Figure 4.5: Specificity and enrichment of sonication-based denaturative ChIP.
(A) Workflow of sonication-based denaturative ICeChIP. **(B-C)** (B) Size distribution of fragments, (C) enrichment of nucleosome standards, and (D) pulldown specificity after sonication for the indicated amount of time.

gesting that the sonication did indeed denature and expose the epitope to the antibody for more efficient binding. Interestingly, this increased efficiency was accompanied by increased specificity of pulldown, with 30 minutes of sonication resulting in both the most efficient and specific IP (Fig. 4.5C-D). Though the reason for this was not entirely clear, it is possible that the increased specificity seen with the more denatured nucleosomes arises from successful competition by the H3K79me2 epitope to capture antibody and prevent off-target epitopes from binding free antibodies. All told, the H3K79me2 pulldown was markedly improved in specificity by adding sonication.

However, this method was imperfect; too much sonication could harm the quality of the pulldown. Applying 60 minutes of sonication generated a sample with subnucleosomal fragment sizes (Fig. 4.5B) with a low-efficiency pulldown (Fig. 4.5C). This indicated that oversonication could result in destruction of the nucleosome itself rather than the linker DNA, reducing the available nucleosomes for binding. Accordingly, the antibody was free to bind to off-target nucleosomes at a higher rate, reducing the apparent specificity of the ChIP (Fig. 4.5D). It thus appeared that using sonication for internal modification ChIP had a fundamental tradeoff between adequate epitope exposure and excessive destruction of the target, necessitating a balance between the two, despite the fact that such a balance may be variable between cell types.

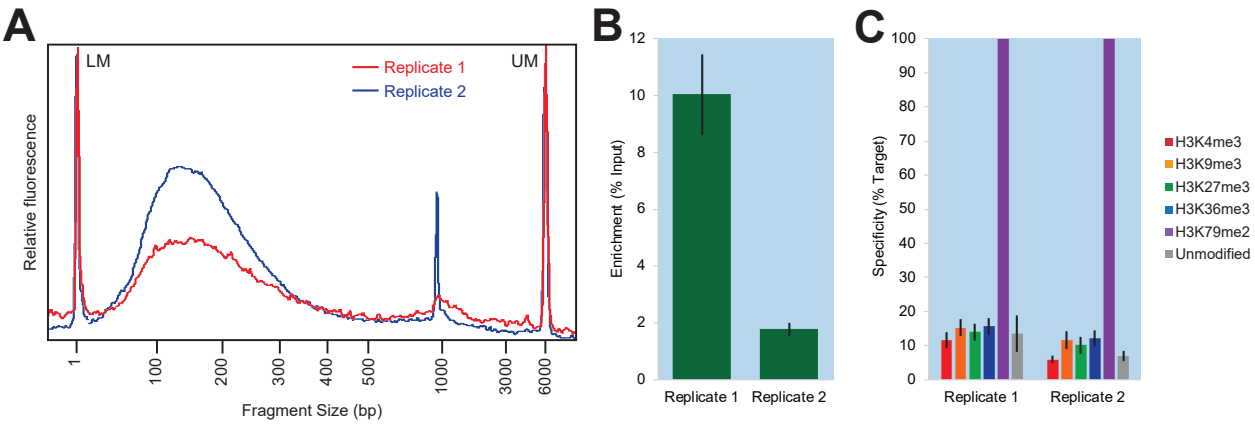


Figure 4.6: Variability of sonication-based denaturative ChIP.
(A) BioAnalyzer trace of fragment size distribution, **(B)** enrichment of nucleosome standards, and **(C)** pulldown specificity of two replicates.

With this tradeoff in mind, we next sought to test the robustness of the sonication-based denaturative ICeChIP method. To do this, we carried out our procedure in two replicates in parallel, with the cells being split immediately before crosslinking and processed simultaneously per the same protocol. And yet even between these highly standardized replicates, there were still marked differences. The relative size distributions of the fragments in these two replicates varied significantly after sonication (Fig. 4.6A), which was concerning given that the samples came from

the same cell population, were cross-linked in parallel, and sonicated using the same settings on the same machine. And after the pulldown, we observed that there was a vast difference in the pulldown efficiency of the two samples (Fig. 4.6B), with further differences between the samples in specificity as well (Fig. 4.6C). And even at its best, the off-target binding remained rather high, representing roughly 10-fold enrichment over unmodified nucleosomes (Fig. 4.6C) when the unmodified nucleosomes are roughly 20-fold more abundant[135]. Given the high apparent variability of the method and the relatively low specificity, we determined that we needed a more reproducible and specific denaturation ICeChIP protocol.

Thermal denaturation for ICeChIP

To more reliably denature the nucleosomes without excessive epitope destruction, we modified our overall denaturative ICeChIP workflow. In the previous experiments, based on the method published by Orlando et al., we crosslinked and sonicated cells or nuclei directly. However, for our new versions of denaturative ICeChIP, we instead chose to first digest chromatin with MNase and purify mononucleosomes (both genomic and spike-in) as in a native protocol. Only once the nucleosomes were purified was crosslinking and denaturation conducted (Fig. 4.7A).

This had two major advantages. First, this method ensured that the genomic and spike-in nucleosomes were subjected to the same conditions. With the previous protocols, spike-in nucleosomes were on the outside of the cell or nuclear membrane, whereas the genomic nucleosomes were inside. This meant that spike-in nucleosomes were subjected to higher effective crosslinker concentrations and greater physical stress upon sonication. By purifying nucleosomes prior to crosslinking and denaturation, the genomic and spike-in nucleosomes would be subjected to the same chemical and physical environment, making the spike-ins more representative of the genomic nucleosomes

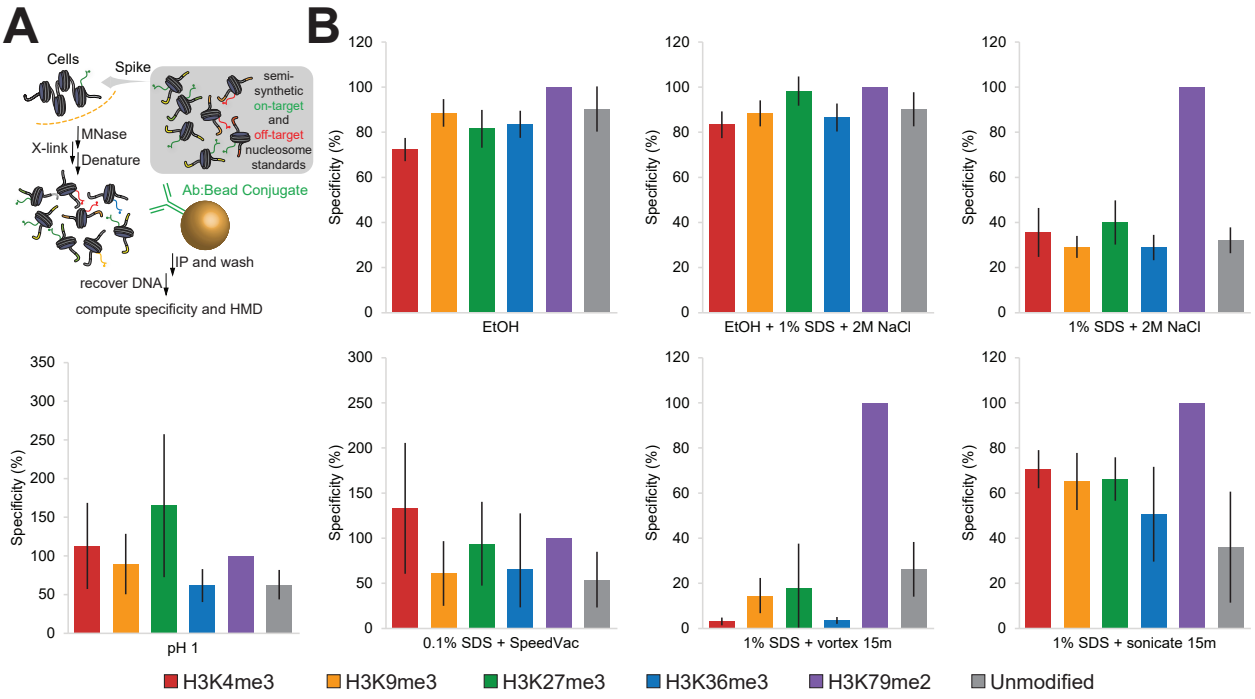


Figure 4.7: Denaturative ICeChIP workflow and panel of denaturation methods.
(A) Generalized denaturative ICeChIP workflow. **(B)** Pulldown specificity of denaturative ICeChIP with indicated denaturation methods.

and thus increasing quantitative power. Second, our new method decoupled fragmentation from denaturation. With sonication-based methods, the fragmentation and denaturation were coupled such that it was difficult to reduce or increase chromatin fragmentation without a concomitant change in denaturation. In this method, the two processes were conducted separately such that denaturation could be tuned without compromising the efficient fragmentation of chromatin into mononucleosomes.

With this framework, we tested several methods designed to denature the nucleosomes immediately prior to the pulldown. Our goal was to find a denaturation protocol that could reliably and completely denature nucleosomes in the denaturation step but could still permit pulldown by an antibody in subsequent steps. Many of the methods we tried were not successful in improving pulldown specificity (Fig. 4.7B). All of these methods had low specificity, often had low enrichment,

and would likely present a significant source of variability for the methods that required physical disruption (e.g. vortexing and water bath sonication). In particular, we noted that detergent was not inherently capable of denaturing the nucleosome and improving specificity, even in the presence of high salt (Fig. 4.7B). However, we did note that the sample with 1% SDS and vortexing had marginally higher specificity than a native pulldown (Fig. 4.7B), hinting that it may be possible to improve pulldown specificity by denaturing the nucleosome in the presence of a detergent such as SDS. Our rationale was that though the detergent itself would not be able to denature the nucleosome, it could coat a denatured protein and thereby stabilize a nucleosome that was denatured by other means. Having exhausted the other physical means of denaturing a protein, we turned to thermal denaturation.

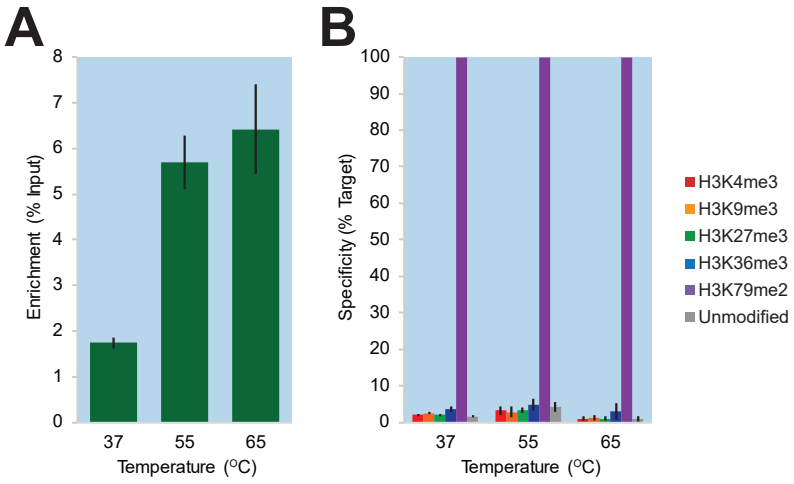


Figure 4.8: Thermal denaturation for ICeChIP.
(A) Enrichment of on-target standard and **(B)** pulldown specificity of denaturative ICeChIP against H3K79me2 using thermal denaturation with the indicated temperatures.

Thermal denaturation in the context of a crosslinked sample has theoretical drawbacks. Formaldehyde crosslink reversal is frequently done by heating the sample for an extended period of time; as such, we wanted to limit the amount of time for which we heated our crosslinked chromatin to prevent decrosslinking and, accordingly, loss of chromatin upon denaturation. As such, we heated

our samples for one minute in the presence of 1% SDS to either 37, 55, or 65 degrees Celsius, then diluted the samples ten-fold to bring the final SDS concentration down to 0.1%, which would permit antibody binding. To our surprise, this protocol was effective. Across the range of temperatures tested, the thermal denaturation permitted an H3K79me2 pulldown with reasonably high efficiency (Fig. 4.8A) and high specificity (Fig. 4.8B). Given these positive results, we moved forward with thermal denaturation for denaturative ICeChIP, with the denaturation step being conducted at 55°C to balance efficiency, specificity, and risk of decrosslinking the sample.

Evaluating reproducibility of denaturative ICeChIP

We next sought to better characterize the benefits and drawbacks of denaturative ICeChIP in different pulldown contexts. To do this, we conducted both native ICeChIP and denaturative ICeChIP on two modifications: H3K4me3 and H3K79me2. H3K4me3 is a tail modification (Fig. 4.1) that can be readily immunoprecipitated with native ChIP protocols[118,124], whereas H3K79me2 is an internal modification that requires denaturative ChIP. As expected, the H3K79me2 pulldown was markedly more specific under denaturative ICeChIP conditions than native ICeChIP conditions (Fig. 4.9A), in line with previous descriptions native ICeChIP against H3K79me2[118] (Fig. 4.2). Interestingly, however, the H3K4me3 pulldown was more specific under native conditions (Fig. 4.9A), consistent with previous reports that crosslinked material is "stickier" and that pulldowns of the same are less specific than native IPs[150]. These data suggest that both native and denaturative ICeChIP are contextually useful; for tail modifications, native ChIP will be simpler and more specific, whereas for internal modifications, denaturative ICeChIP may succeed where native will not.

Our next question was on the robustness of our denaturative ICeChIP method. The prior sonication-based method had resulted in marked variability at the level of fragment sizes, enrich-

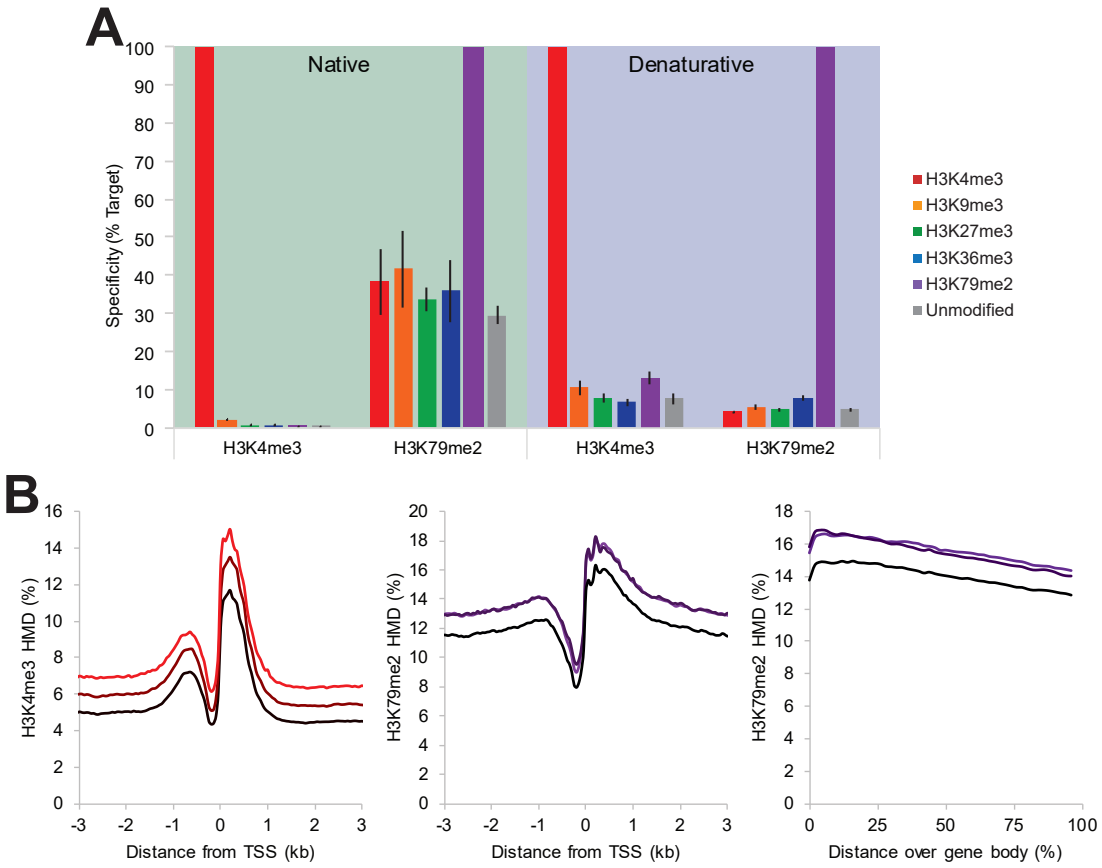


Figure 4.9: Specificity and reproducibility of denaturative ICeChIP.

(A) Specificity of native and denaturative ICeChIP against H3K4me3 and H3K79me2. Error bars represent standard deviation across three distinct biological replicates. **(B)** Metagene profiles of denaturative ICeChIP-seq signal targeting H3K4me3 (left) and H3K79me2 (center, right) at transcription start sites (TSS; left, center) and gene bodies (right), showing reproducibility of denaturative ICeChIP. Each color represents a distinct biological replicate.

ment, and specificity (Fig. 4.6). This error is markedly reduced by instead using thermal denaturation. Replicates of denaturative ICeChIP using thermal denaturation had highly similar pulldown specificities (Fig. 4.9A). Further, metagene profiles of denaturative ICeChIP-seq were highly similar across replicates (Fig. 4.9B), indicating that the pulldowns were quantitatively similar at genomic loci at well. All told, these results indicated that thermal denaturation yields high-specificity pulldowns with a high degree of reproducibility.

Given that this method worked with our nucleosome standards, we next sought to revisit the more commonly used spike-ins: exogenous chromatin from another organism, such as *D. melanogaster* or *S. cerevisiae*. To be sure, these methods are inherently suboptimal. Such a spike-in would lack the ability to measure specificity of the pulldown except in cases where the spike-in chromatin lacks the targeted histone modification entirely (e.g. H3K27 methylation in *S. cerevisiae*[284]), and even then would not indicate which modifications contribute to off-target binding. Further, exogenous genomic spike-ins are likely to have significant lot-to-lot variation, as the amount of genomic histone modification cannot be precisely controlled, whereas semisynthetic nucleosomes are precisely formulated and can thereby limit lot-to-lot variation. Nonetheless, it remains a fact that many people use exogenous chromatin as spike-ins rather than nucleosome standards[128,285], so we sought to evaluate whether our denaturative ICeChIP protocol can improve upon the high variability previously observed with these methods[128] (Fig. 4.4).

To do this, we modified the ChIP-Rx protocol to more closely resemble ICeChIP (Fig. 4.10A). First, rather than spiking crude *D. melanogaster* cells into a human cell sample, we spiked in highly purified and well-quantified *D. melanogaster* and *S. cerevisiae* nuclei into a sample of human nuclei, so the spike-in chromatin was treated more similarly to the human chromatin. Second, rather than crosslinking and shearing by sonication, we fragmented the chromatin by MNase digestion, which is more reliable. We then conducted either native ICeChIP against H3K4me3, denaturative ICeChIP against H3K4me3, or denaturative ICeChIP against H3K79me2, all in triplicate.

Given the previously described high reproducibility of native ICeChIP, we anticipated that the relative pulldown of both human and spike-in chromatin would be highly similar across replicates, meaning that metagene profiles of the datasets should be highly similar across replicates after normalization to the spike-ins. This was, in fact, the case; TSS metaprofiles of the native H3K4me3

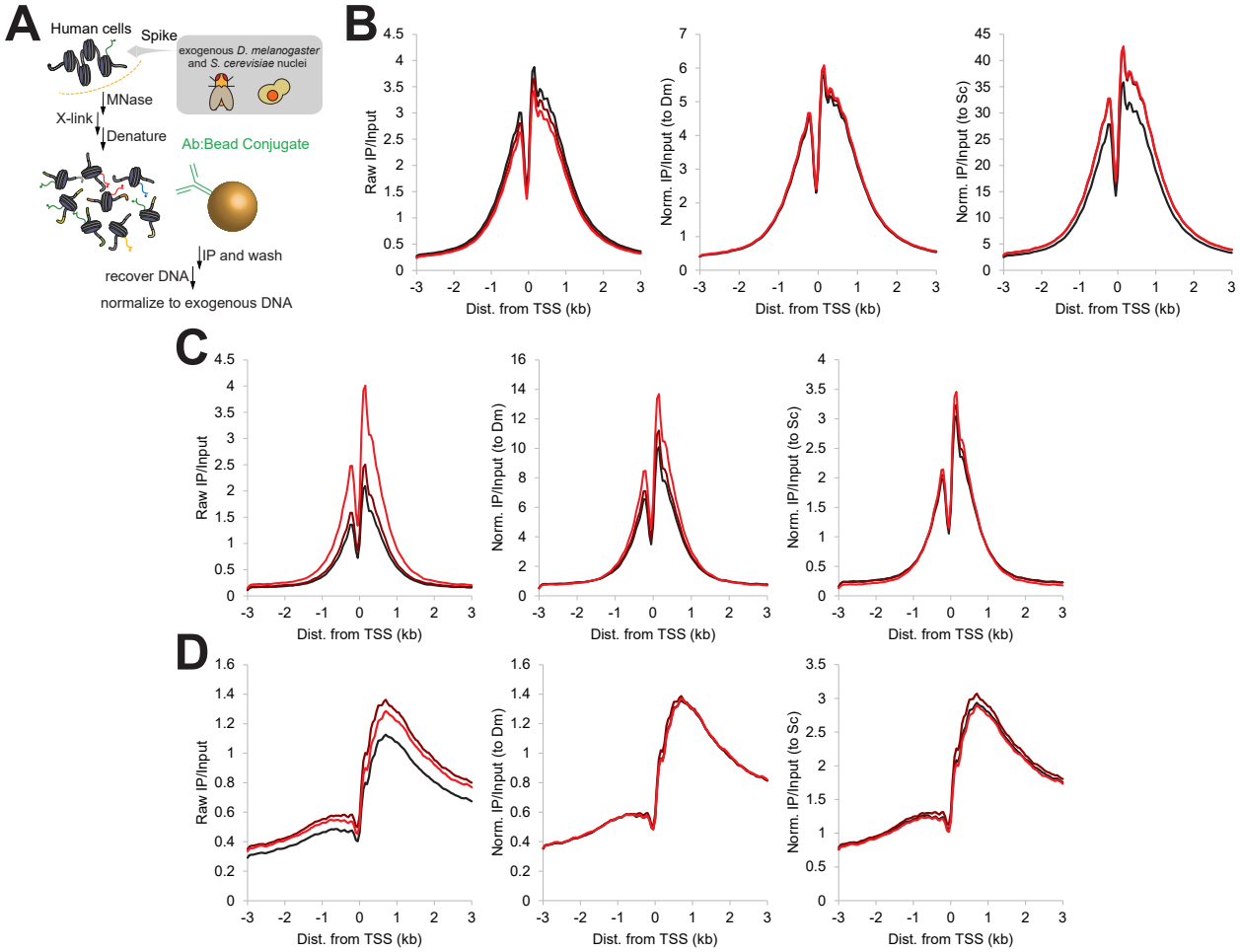


Figure 4.10: Exogenous chromatin normalization with denaturative ICeChIP protocol.
(A) Workflow for modified exogenous chromatin normalization ChIP. **(B-D)** Metagene profiles of (B) H3K4me3 native ChIP, (C) H3K4me3 denaturative ChIP, and (D) H3K79me2 denaturative ChIP, normalized to endogenous read depth (left), *D. melanogaster* read depth (center), or *S. cerevisiae* read depth (right).

pulldown were highly reproducible with exogenous chromatin normalization (Fig. 4.10B). Indeed, the native ICeChIP pulldown procedure is so robust that even the endogenously normalized data metaprofiles were quite similar (Fig. 4.10B).

Denaturative ICeChIP was less reproducible with endogenous normalization; the denaturative H3K4me3 and H3K79me2 pulldown metaprofiles showed marked variability between replicates (Fig. 4.10C-D). However, normalization to the exogenous chromatin rectified much of this

variability, resulting in highly similar metaprofiles (Fig. 4.10C-D). These data suggested that the denaturative ICeChIP pulldown protocol is also workable for use with exogenous chromatin. Again, this method is inherently suboptimal compared to nucleosome standard spike-in for the reasons stated above. Nonetheless, if a researcher is particularly inclined towards exogenous chromatin spike-ins, our denaturative ICeChIP protocol improves reproducibility even in that context.

Calibration by denaturative ICeChIP

To validate the calibration ability of our denaturative ICeChIP protocol, we compared our denaturative and native ICeChIP data targeting H3K4me3. H3K4me3, as a tail modification, is well-measured by native ICeChIP to an extent in line with mass spectrometry estimates[118]. Our expectation was that our measurements of H3K4me3, then, would be roughly similar between the denaturative and native ICeChIP protocols. To our surprise, however, this was not the case. Denaturative ICeChIP HMDs were, on average, roughly 48% that measured by native ICeChIP across TSSs (Fig. 4.11A). The discrepancy appeared to be because of the way in which the standards were pulled down; the nucleosome standards, bearing the high-affinity 601 DNA sequence[168], were pulled down roughly three times as efficiently as genomic nucleosomes in denaturative ICeChIP compared to native ICeChIP (Fig. 4.11B). This meant that the apparent enrichment that would be expected of a locus that was 100% modified with H3K4me3 was higher than appropriate for the genomic nucleosomes and, accordingly, that the denaturative H3K4me3 HMDs were systematically deflated relative to the native H3K4me3 HMDs.

We hypothesized that the nucleosome standards were being differentially captured relative to the genomic nucleosomes because of the affixed DNA sequence on the nucleosome. Our rationale was that this was the major point of difference between the standard and genomic nucleosomes; the

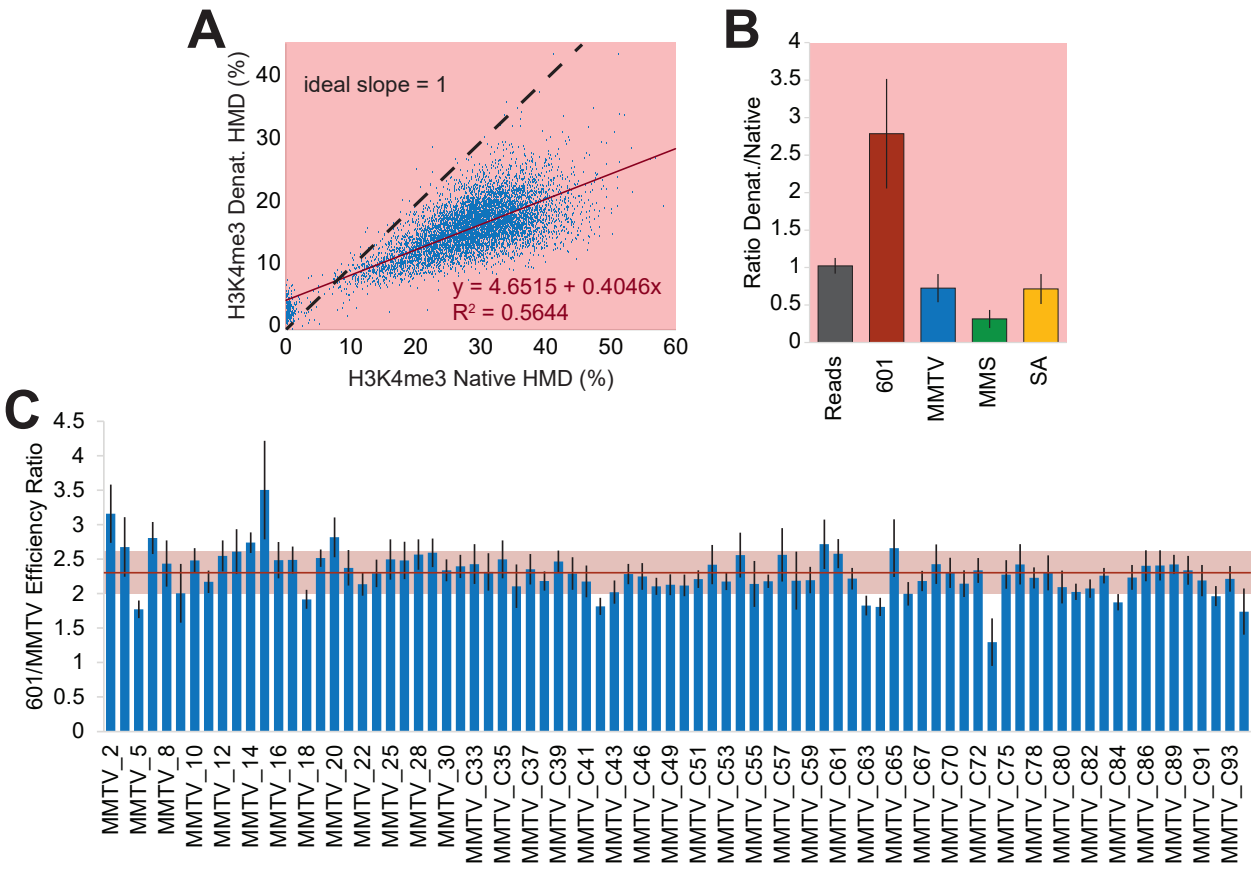


Figure 4.11: Denaturative ICeChIP HMD deflation and differential standard enrichment.
(A) Denaturative vs. native ICeChIP H3K4me3 HMD of 10,000 randomly selected genomic windows using nucleosome standards with the 601 DNA sequence. **(B)** Fold change of enrichment of nucleosome standards bearing indicated DNA sequence in denaturative/native ICeChIP against H3K4me3, normalized to read depth. **(C)** Fold change of enrichment of nucleosomes with 601 DNA sequence/indicated MMTV DNA sequence in denaturative ICeChIP against H3K4me3. Red line indicates average ratio across the different MMTV barcodes; red shaded area indicates standard deviation about average ratio.

semisynthetic histones were essentially identical to genomic histones[255], whereas the 601 sequence is designed to have supraphysiological affinity for histones[168,286].

To test this hypothesis, we developed nucleosome barcodes based on other DNA sequences that would bind to the histone octamer with lower affinity[287]. These included a sequence based on the mouse mammary tumor virus (MMTV) long terminal repeat[288], a sequence based on the mouse minor satellite (MMS)[289], and a purely synthetic sequence based on genomic unwords (Space Alien;

SA). We assembled these sequences into nucleosomes bearing H3K4me3, then measured their enrichment in denaturative and native ICeChIP-seq. The new barcode sequences had considerably lower enrichment in the denaturative ICeChIP-seq than did the 601-based standards (Fig. 4.11B), with the ratio of enrichment in the MMTV and SA standards between denaturative and native ICeChIP appearing similar to that of the genomic reads. Curiously, the MMS-based barcodes had an even lower enrichment in denaturative ICeChIP for reasons not fully understood. This was found to be applicable for MMTV sequences with a broad range of nucleosome barcodes, wherein the 601 sequence was pulled down an average of 2.27 times as efficiently as MMTV-based sequences in denaturative ICeChIP against H3K4me3 (Fig. 4.11C). These data suggested that the DNA sequence identity of the nucleosome standards affected the pulldown efficiency in denaturative ICeChIP.

What remained unclear was the reason for this difference. We first hypothesized that the nucleosomes with different DNA affinities were being differentially decrosslinked by the thermal denaturation. This would not affect the presence of the DNA in the input sample, but if the nucleosome completely fell apart, then there would be less nucleosome to pull down and, accordingly, a lower apparent enrichment. To test this, we conducted denaturative ICeChIP with our input normalization being conducted against either input DNA (as standard) or against an H3 pulldown with one of two H3 C-terminal domain (CTD) antibodies. If differential nucleosome destruction was the problem, then the H3 pulldowns (which would only measure intact nucleosomes) should resolve this difference if used as the input. Unfortunately, this was not the case; whether the input chromatin was true DNA input or an H3 CTD pulldown, the measured genomic HMD was virtually identical (Fig. 4.12A). This suggested that the difference was not driven by mere differential susceptibility to decrosslinking and destruction, but rather, by some unknown intrinsic property of the nucleosome itself.

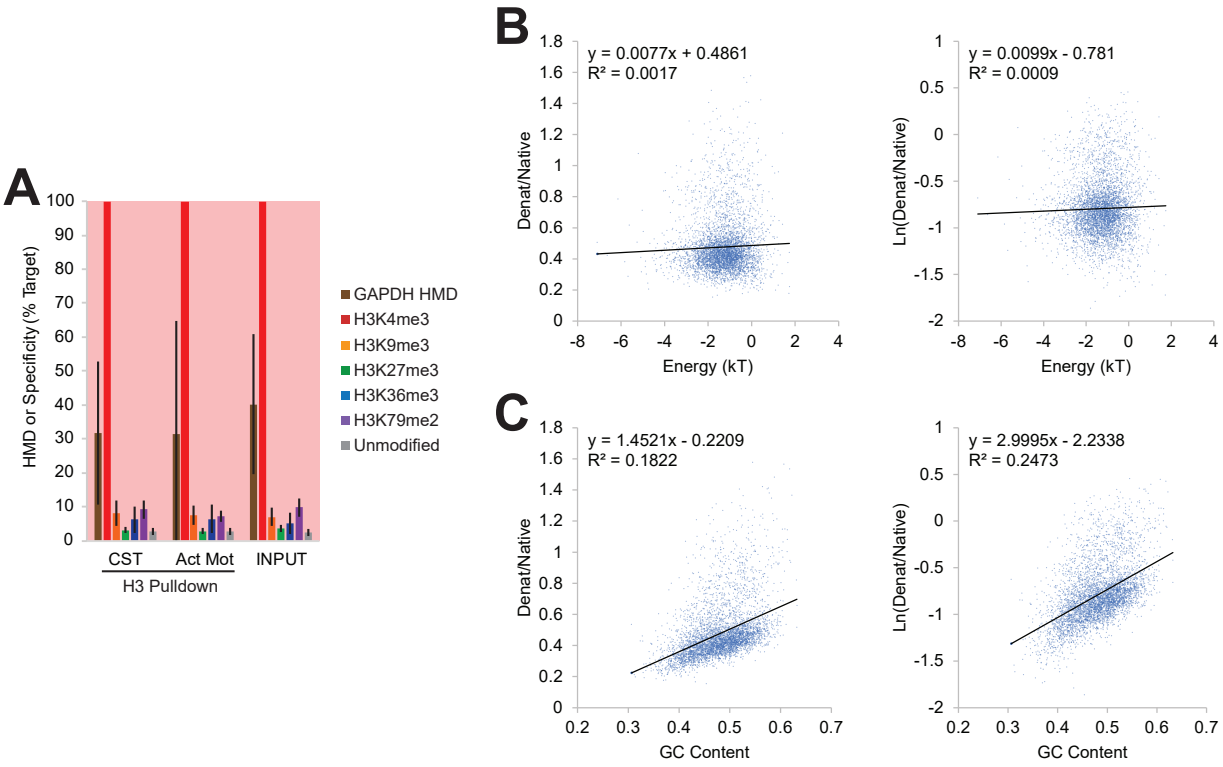


Figure 4.12: Associations with deflation of HMDs in denaturative ICeChIP.
(A) Denaturative ICeChIP against H3K4me3 using H3 pulldowns or raw DNA input as the input for computation of efficiency and specificity. **(B)** Denaturative/Native (left) or ln(Denaturative/Native) (right) ICeChIP H3K4me3 HMD vs. energy of DNA sequence binding[287] of 200bp windows across the *D. melanogaster* genome. **(C)** Denaturative/Native (left) or ln(Denaturative/Native) (right) ICeChIP H3K4me3 HMD vs. GC content of 200bp windows across the *D. melanogaster* genome.

To try to find an apparent reason for this deflation, we searched for associations between genomic deflation and different factors. Given that it appeared that DNA sequences with different nucleosome binding energies had different pulldown efficiencies in denaturative ICeChIP, we wanted to compare the extent of deflation with nucleosome binding affinity in the genome. To do this, we computed the predicted nucleosome binding energy of 200bp genomic windows[287] with the deflation ratio of HMDs at that same window (Fig. 4.12B). Interestingly, there was essentially no correlation between the predicted binding energy and the deflation ratio or a log transform therein, suggesting that binding energy is not directly responsible or associated with such HMD deflation

(Fig. 4.12B). More curiously, however, lower GC content was associated with greater deflation of HMDs than was higher GC content (Fig. 4.12C). This is consistent with our nucleosome barcode findings; the 601 sequence has relatively high GC content, so it would have higher IP efficiency in denaturative ICeChIP than would genomic loci.

This finding was startling, however, because it violated a fundamental assumption in ChIP: that the identity of the DNA bound to the nucleosome does not impact the pulldown targeting the protein component. Our findings, however, suggest that in denaturative ICeChIP, the DNA bound to the nucleosome can impact the efficiency of pulldown, meaning that denaturative ChIP is DNA-sequence biased. The extent to which this concerning discovery holds true in other denaturative ChIP paradigms remains to be seen.

After all these inquiries, we concluded that we could not easily correct for this apparent deflation of HMD. Though this makes it challenging to treat the nucleosome standards as calibrants (placing the pulldown on a biologically meaningful scale), our denaturative ICeChIP protocol and nucleosome standards still enable measurement of antibody specificity and for normalization to an invariant exogenous standard. With this caveat in mind, we proceeded to use our denaturative ICeChIP method to investigate the biology of MLL-rearranged leukemias by studying the association of H3K79me2 dysregulation with the transcriptional changes of leukemogenesis.

Examining the role of H3K79me2 in MLL-rearranged leukemia

In studying the role of H3K79me2 on the genesis of MLL-rearranged leukemias, we first sought to study its impacts on transcription. MLL-rearranged leukemias were previously described to have a distinctive transcriptional program[274] featuring activation of several genes including HOXA9 and MEIS1. Similarly, H3K79me2 is known to be dysregulated in MLL-rearranged leukemias, often

at several of the genes dysregulated transcriptionally[23,269,283]. However, determining the impact of H3K79me2 on gene expression is somewhat more complex; many of the genes that are differentially marked with H3K79me2 may be incidental and unimportant for the process of leukemogenesis; it is possible they display increased H3K79me2 as a side effect of the rearrangement rather than as a driving factor of the liquid tumor. Similarly, not all of the transcriptional changes will be directly driven by the H3K79me2 increase; some of the changes in transcription are likely to be reactive to other changes rather than primary effects.

To identify a candidate list of genes that may be dysregulated primarily as a result of H3K79me2 increase in MLL-rearranged leukemias, we searched for genes that had both dysregulation of H3K79me2 and gene expression. To do this, we sought to first identify genes that had dysregulated H3K79me2 in MLL-rearranged leukemias by conducting denaturative ICeChIP-seq against H3K79me2 in six MLL-rearranged leukemia lines and K562 cells (Fig. 4.13). We then identified genes that were differentially modified in each cell line relative to the K562 outgroup (both in absolute HMD differences and relative HMD differences). The 3834 genes that were present in all these lists were identified as the genes that were differentially modified with H3K79me2 in MLL-rearranged leukemia cells.

We next sought to identify genes that were differentially regulated in response to changes in H3K79me2 in MLL-rearranged leukemia. To do this, we reanalyzed previously published RNA-seq data in different cell lines with and without the presence of pinometostat[290]. To separate the effects of pinometostat more broadly from the effects of pinometostat in MLL-rearranged leukemia, we separated our datasets into those from MLL-rearranged or non-MLL-rearranged leukemias, then conducted separate differential expression analyses for each group and subtracted any genes that were differentially regulated in the latter, ultimately identifying 1420 genes that were differentially

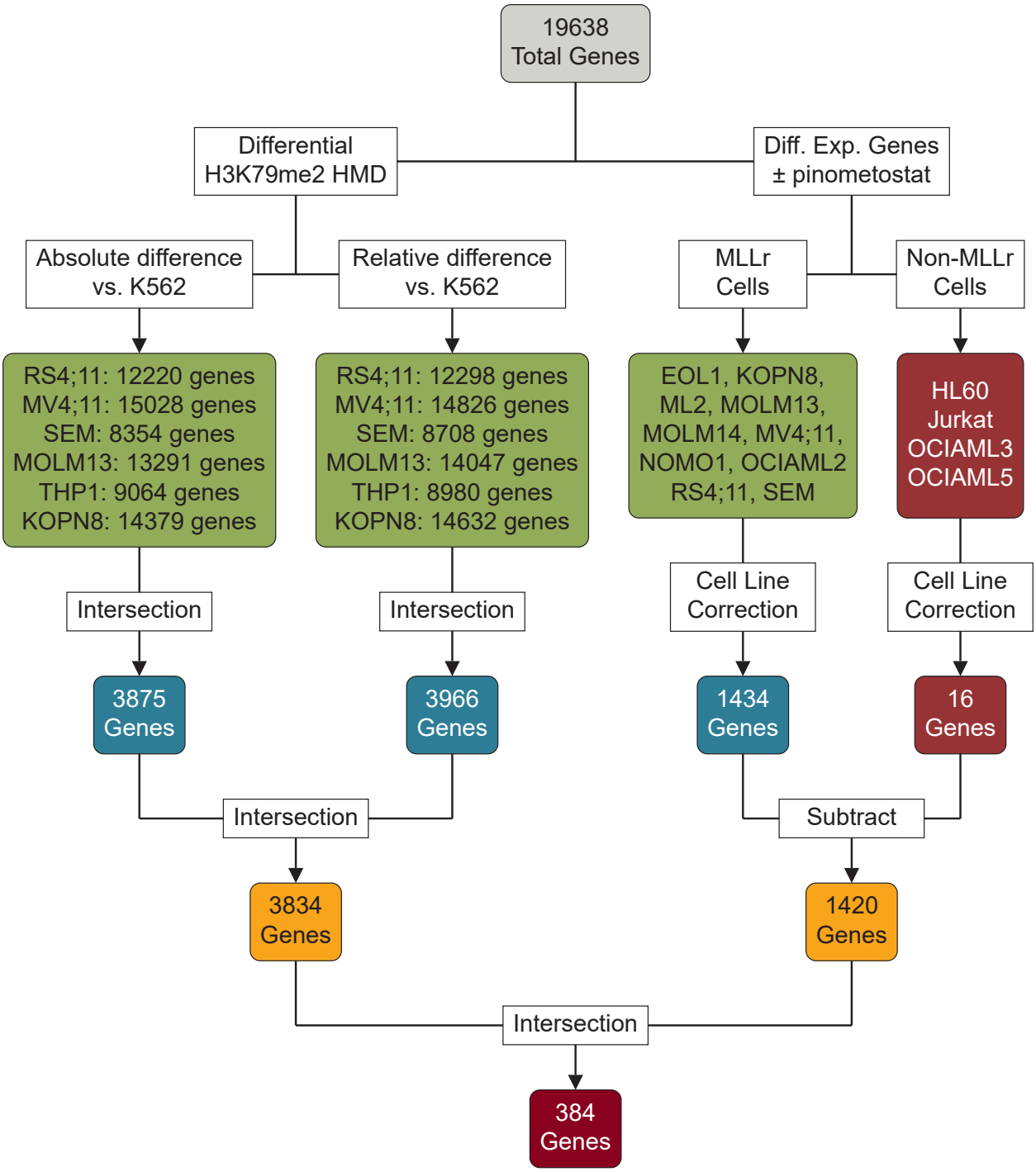


Figure 4.13: Differentially expressed and modified genes in MLL-rearranged leukemias.

Identification workflow for genes that are differentially marked by H3K79me2 HMD relative to K562s and/or differentially expressed in MLL-rearranged leukemias specifically in response to changes in H3K79me2. Overall, 384 genes were found that were differentially expressed and differentially modified with H3K79me2.

expressed in MLL-rearranged leukemia with changes in H3K79me2 levels (Fig. 4.13). Combining the H3K79me2 and RNA-seq analyses, we found 384 genes that were common between the lists of differentially marked and expressed genes (Fig. 4.13), which served as our list of candidate genes that had transcriptional dysregulation potentially as a result of H3K79me2 dysregulation.

This set of genes represented only a minority of the genes that were either differentially modified or expressed in MLL-rearranged leukemias, leading us to wish to better characterize these genes. Some of these 384 genes were among the canonical MLL-rearranged leukemia target loci, including HOXA9 and MEIS1, which was promising for the sensitivity and specificity of our analysis. To better understand the types of genes that fell into this list, we conducted gene ontology against the genes that were differentially methylated and expressed (Table 4.1), differentially methylated but not differentially expressed (Table 4.2), or differentially expressed by not differentially modified (Table 4.3). In this, several interesting trends emerged. We observed that the genes that were differentially expressed, regardless of H3K79me2 modification status, tended to be immune system process genes, with most of the top gene ontology terms having to do with immune system activation or cellular migration (Tables 4.1-2). However, the genes that were differentially modified but not differentially expressed tended to be quite different, focusing more heavily on metabolic genes (Table 4.3).

This constellation of findings had two primary interpretations. First, the genes that primarily changed their transcriptional program upon H3K79me2 changes were primarily immune system process genes, and this transcriptional program tended towards coherent changes whether the individual genes were differentially marked by H3K79me2 or not. Second, the relative paucity of immune system process genes in the “differentially modified but not differentially expressed” category suggested that many genes were only incidentally marked with H3K79me2, indicating

Table 4.1: Top twenty gene ontology terms for genes that are differentially modified by H3K79me2 and differentially expressed.

GO.ID	Term	Significant	Expected	Enrichment	Adj. p-value
GO:0050853	B cell receptor signaling pathway	11	1.22	9.02	$< 2.2 \times 10^{-16}$
GO:0050851	antigen receptor-mediated signaling pathway	20	4.03	4.96	$< 2.2 \times 10^{-16}$
GO:0002429	immune response-activating cell surface receptor signaling pathway	25	5.14	4.86	$< 2.2 \times 10^{-16}$
GO:0002757	immune response-activating signal transduction	25	5.14	4.86	$< 2.2 \times 10^{-16}$
GO:0002768	immune response-regulating cell surface receptor signaling pathway	25	5.77	4.33	$< 2.2 \times 10^{-16}$
GO:0042113	B cell activation	22	5.43	4.05	$< 2.2 \times 10^{-16}$
GO:0002253	activation of immune response	27	6.91	3.91	$< 2.2 \times 10^{-16}$
GO:0018105	peptidyl-serine phosphorylation	24	6.47	3.71	$< 2.2 \times 10^{-16}$
GO:0007015	actin filament organization	34	9.41	3.61	$< 2.2 \times 10^{-16}$
GO:0002764	immune response-regulating signaling pathway	31	9	3.44	$< 2.2 \times 10^{-16}$

that other mechanisms may be responsible for transcriptional activation rather than direct activation of gene expression by H3K79me2.

To understand some of the other mechanisms by which H3K79me2 may impact gene regulation, we examined its association with other histone modifications. To do this, we examined changes in histone modifications observed in response to pinometostat treatment in the MV4;11 cells, which was previously published[291]. As expected, H3K79me2 decreased dramatically with

Table 4.2: Top twenty gene ontology terms for genes that are differentially modified by H3K79me2 but not differentially expressed.

GO.ID	Term	Significant	Expected	Enrichment	Adj. p-value
GO:0032273	positive regulation of protein polymerization	51	25.56	2	$< 2.2 \times 10^{-16}$
GO:0048813	dendrite morphogenesis	52	26.12	1.99	$< 2.2 \times 10^{-16}$
GO:1902905	positive regulation of supramolecular fiber organization	69	38.07	1.81	$< 2.2 \times 10^{-16}$
GO:0016358	dendrite development	78	43.22	1.8	$< 2.2 \times 10^{-16}$
GO:1903311	regulation of mRNA metabolic process	102	56.64	1.8	$< 2.2 \times 10^{-16}$
GO:0051056	regulation of small GTPase mediated signal transduction	98	55.72	1.76	$< 2.2 \times 10^{-16}$
GO:0006325	chromatin organization	136	84.05	1.62	$< 2.2 \times 10^{-16}$
GO:0016570	histone modification	132	83.13	1.59	$< 2.2 \times 10^{-16}$
GO:0010975	regulation of neuron projection development	125	79.08	1.58	$< 2.2 \times 10^{-16}$
GO:0007264	small GTPase mediated signal transduction	143	91.77	1.56	$< 2.2 \times 10^{-16}$

pinometostat treatment (Fig. 4.14). Beyond that change, we observed that H3K4me3 HMD at TSSs increased markedly with pinometostat treatment (Fig. 4.14), suggesting that H3K79me2 may either oppose or compensate for H3K4me3 in this context (seeing as both modifications are associated with active transcription). However, beyond that, only modest changes were seen in H3K27me3 and H3K36me3 at TSSs (Fig. 4.14), at least with the pinometostat treatment parameters used in the treatment of the cells under these conditions.

Table 4.3: Top twenty gene ontology terms for genes that are differentially modified by H3K79me2 but not differentially expressed.

GO.ID	Term	Significant	Expected	Enrichment	Adj. p-value
GO:0032273	positive regulation of protein polymerization	51	25.56	2	$< 2.2 \times 10^{-16}$
GO:0048813	dendrite morphogenesis	52	26.12	1.99	$< 2.2 \times 10^{-16}$
GO:1902905	positive regulation of supramolecular fiber organization	69	38.07	1.81	$< 2.2 \times 10^{-16}$
GO:0016358	dendrite development	78	43.22	1.8	$< 2.2 \times 10^{-16}$
GO:1903311	regulation of mRNA metabolic process	102	56.64	1.8	$< 2.2 \times 10^{-16}$
GO:0051056	regulation of small GTPase mediated signal transduction	98	55.72	1.76	$< 2.2 \times 10^{-16}$
GO:0006325	chromatin organization	136	84.05	1.62	$< 2.2 \times 10^{-16}$
GO:0016570	histone modification	132	83.13	1.59	$< 2.2 \times 10^{-16}$
GO:0010975	regulation of neuron projection development	125	79.08	1.58	$< 2.2 \times 10^{-16}$
GO:0007264	small GTPase mediated signal transduction	143	91.77	1.56	$< 2.2 \times 10^{-16}$

Further hints at crosstalk patterns of H3K79me2, however, are present in the literature. Previous work suggested that H3K79me2 inhibition was inhibited by H3K27me3 existing on the same nucleosome; this was determined on the basis of biochemical studies showing that the Dot1L complex member AF10 is able to bind to H3K27me3 as a negative regulator to Dot1L activity[292]. However, when we conduct ICeChIP against H3K79me2 and H3K27me3, we see an even starker anticorrelation between the two modifications, with regions of high H3K27me3 bearing very little H3K79me2 and vice versa (Fig. 4.15). Indeed, the anticorrelation is so strong that there are virtually

no regions with high H3K27me3 and high H3K79me2 (Fig. 4.15). There are two possible explanations for this behavior from a systems level. First, these results could be observed if H3K79me2 is installed and removed much more rapidly than H3K27me3 and thus demonstrates a quasi-steady-state phenomenon with regards to H3K27me3. However, this is questionable – though some have claimed to have identified an H3K79me2 demethylase[293], H3K79me2 disappears at a slow rate (consistent with dilution by cellular division) in the presence of pinometostat[280]. Given this slow removal process, it becomes more likely that instead, H3K79me2 and H3K27me3 mutually inhibit the installation of the other mark, a process which would result in the strong anticorrelation observed above even with slow H3K79me2 dynamics. Thus, though more study is needed, it is plausible that H3K79me2 and H3K27me3 demonstrate mutual inhibition to a degree not previously identified – directly or otherwise.

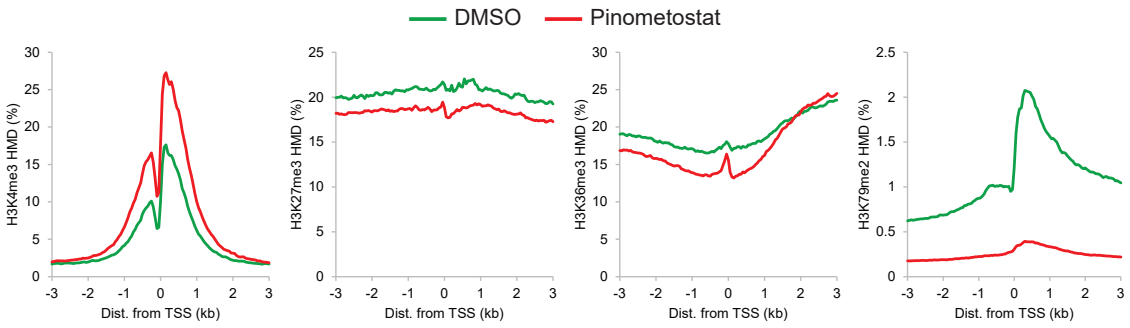


Figure 4.14: Histone modification changes with pinometostat treatment.
H3K4me3, H3K27me3, H3K36me3, and H3K79me2 HMD with and without pinometostat Dot1L inhibitor. Data taken from Richter et al., 2021[291].

Discussion

Despite their biological and clinical importance, internal histone modifications have to date represented a major blind spot of quantitative studies of histone modifications. Their structure makes it challenging to find high-confidence binding partners biochemically, and the low specificity of

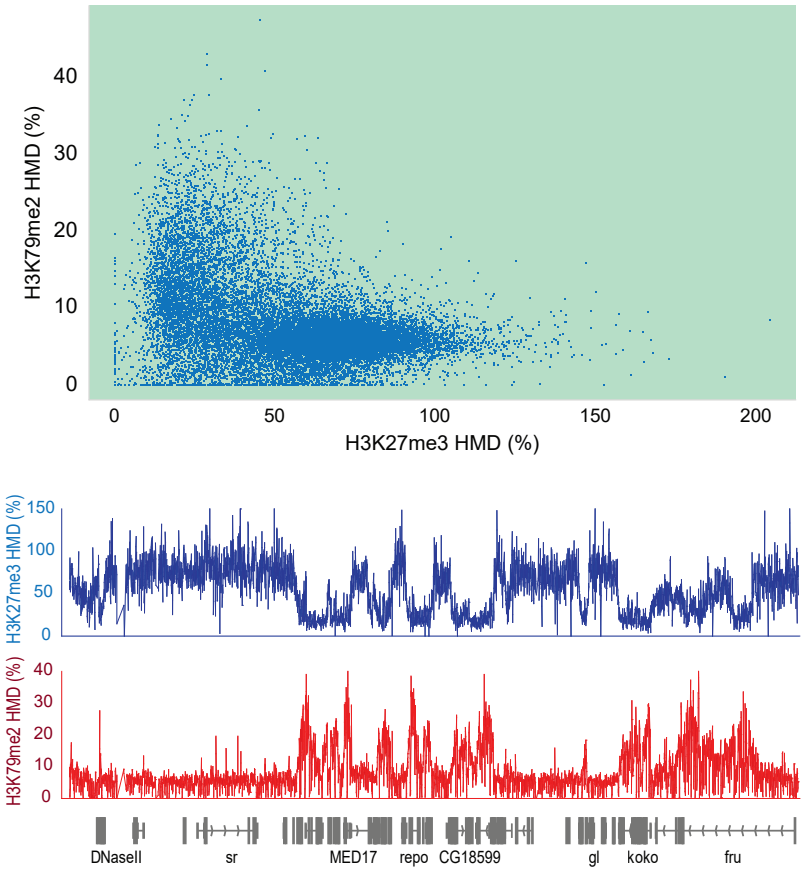


Figure 4.15: Anticorrelation of H3K79me2 and H3K27me3.
Anticorrelation of H3K79me2 and H3K27me3 at genomic windows in *D. melanogaster* S2 cells, with example genomic locus.

native pulldowns against such modifications makes genomic studies therein similarly challenging. Our work here represents a deep dive into the tunable parameters of ChIP input preparation and how their impacts on fragmentation and denaturation ultimately impacts the quality of the pulldown.

The most common methods of denaturative ChIP – including those typically used against H3K79me2[128] – involve sonication, which necessarily convolute the effects of fragmentation and denaturation (Fig. 4.5)(, making it impossible to change the extent of denaturation/epitope exposure without also changing the extent of chromatin fragmentation or epitope destruction. Further, as a physical method of denaturation that will be highly dependent on temperature, crosslinking

efficiency, and handling variation, sonication has highly variable impacts on both fragmentation and denaturation, making it suboptimal as the basis of a ChIP protocol (Fig. 4.6).

Rather than using sonication, we instead focus on deconvoluting fragmentation and denaturation so they can be tuned separately. Fragmentation, under native conditions, is straightforward to conduct reproducibly and consistently between cell types by MNase digestion[118,170]. Further, crosslinking and denaturation of nucleosomes is more reproducible starting with a highly purified sample that does not have as many "excess" proteins that can absorb the crosslinking reagent or shield the nucleosomes. Further, the fact that fragmentation occurs before denaturation means that we are able to aggressively denature without excessively digesting the chromatin to subnucleosomal fragments, an advantage that is not present for sonication-based methods. With this method, we are able to thoroughly denature nucleosomes with detergent and a short pulse of heat, allowing for high-quality, reproducible, and specific pulldowns of nucleosomes (Fig. 4.8, 4.9).

The method we present here is not without its limitations. As we noted, denaturative ICeChIP demonstrates widespread relative deflation of histone modification densities relative to native ICeChIP (Fig. 4.11) in a manner biased by sequence GC content (Fig. 4.12), suggesting that our denaturative ChIP method is at least marginally biased by DNA sequence. This is particularly concerning if it is generalizable to other denaturative ChIP paradigms; as previously noted, the fundamental premise of ChIP is that the identity of the DNA sequence is not relevant to the pulldown efficiency, allowing comparison of pulldown efficiencies at different loci. The fact that this is not necessarily true in our context is concerning ad raises questions as to whether denaturative ChIP more broadly is sequence biased. Moving forward, we would seek to better define why different regions show deflation relative to native ICeChIP. If the deflation at various genomic regions can be better defined, different nucleosome sequences can be developed to span the range of possibilities

in the genome, allowing for computational application of the appropriate calibration sequence to the relevant genomic regions.

Nonetheless, despite these limitations, we were able to use our denaturative ChIP method to explore the biology of H3K79me2 in the context of MLL-rearranged leukemias and identify new potential cross-talk pathways between H3K79me2 and H3K27me3, potentially shedding light on the role of this histone modification that, to date, has been very poorly characterized. Though imperfect, our method offers a way to avoid the irreproducibility of physical denaturation/fragmentation methods and allows for a basis for more reproducible pulldowns, in the current method and as the basis for future ones.

Acknowledgements

We wish to thank P. Faber and H. Whitehurst in the University of Chicago Functional Genomics Facility for Illumina sequencing. We also thank Dr. Rick Fehon and his laboratory for their generous gift of S2 cells. This study was supported by the National Institutes of Health under award number R01-GM115945 to Alexander J. Ruthenburg (University of Chicago).

Methods and Materials

This section has been adapted from: Shah, R. N. *et al.* Examining the Roles of H3K4 Methylation States with Systematically Characterized Antibodies. *Molecular Cell* **72,** 162–177 (2018).

Cell Culture

K562, MV4;11, RS4;11, MOLM-13, SEM, KOPN-8, and THP-1 cell lines were grown at 37°C with 5% CO_2 and 95% humidity in Dulbecco's Modified Eagle Media (DMEM, Gibco; K562 cells only) or RPMI 1640 (Gibco; other cell lines) supplemented with 10% (v/v) HyClone FBS Characterized

U.S. and 1x Penicillin/Streptomycin (Gibco). Cells were seeded into vented flasks to a density of 200,000 cells/mL of culture and were passaged at 1-2 million cells/mL of culture.

S. cerevisiae yeast (S288C strain) were cultured in YPDA on a shaker at 30°C for approximately 24 hours to an OD_{600} of approximately 1.0. S2 cells were cultured and provided as a cell pellet by the Fehon Laboratory.

Octamer Reconstitution

Symmetric H3K4me3, H3K9me3, H3K27me3, H3K36me3, H3K79me2, and H3K4me3K27me3 octamers were reconstituted from semisynthetic histones as previously described[91,118,166,167]. Recombinant core histones were expressed in BL21 (DE3) with pRARE2 and mixed to equimolarity with the relevant semisynthetic histones in freshly prepared filter sterilized Unfolding Buffer (50 mM Tris-HCl pH 8.0, 6.3 M Guanidine-HCl, 10 mM 2-mercaptoethanol, 4 mM EDTA) to a final concentration of ≥ 1 mg histone per mL. The histone reconstitution was then added to 3500 MWCO SnakeSkin dialysis tubing (Pierce) and dialyzed overnight at 4°C against 500-1000 volumes of filter sterilized Refolding Buffer (20 mM Tris-HCl pH 7.5, 2 M NaCl, 5 mM DTT, 1 mM EDTA).

After dialysis, the histone mixture was centrifuged at 18,000 g for 1 hour at 4°C, and subjected to gel filtration chromatography (Superdex 200 10/300 GL, GE Healthcare, resolved with Refolding Buffer). Each fraction that displayed a peak on the UV chromatogram was analysed by SDS-PAGE (22 mA current in 1x Laemmli Buffer for 70 minutes), stained with SYPRO Ruby (Bio-Rad) per manufacturer instructions, and imaged with a 610BP emission filter at 600V PMT setting. Octamer fractions with equimolar quantities of each core histone were pooled and concentrated (Amicon Ultra-4 Centrifugal Filters, 10,000 MWCO, Millipore) to 5-15 μM octamer, diluted with one volume of Octamer Storage Buffer, and stored at -20°C.

All other octamers were obtained from EpiCypher, Inc.

Nucleosome Reconstitution

Nucleosomes were reconstituted onto 147bp DNAs composed of the core Widom 601 sequence[168] the Mouse Mammary Tumor Virus (MMTV) long terminal repeat[288], or the Mouse minor satellite[287,289] (MMS) modified with a 22bp barcode on each end, with each barcode composed of two distinct 11bp sequences not found in the human or mouse genomes, or a fully synthetic 143bp sequence of DNA comprised of eleven 13bp distinct sequences not found in the human or mouse genomes (Space Alien sequences). The DNA and octamer were mixed to a final concentration of 1µM each in 2 M NaCl, and then dialyzed in dialysis buttons (Hampton Research) and a 10,000 MWCO SnakeSkin dialysis membrane (Pierce) against 200 mL of Refolding buffer for 10 minutes. Dialysis then continued as 2L of Buffer I0 (20 mM Tris-HCl pH 7.5, 1 mM EDTA, 1mM DTT) was added (flow rate 2-2.5 mL per minute).

Dialyzed samples were diluted with an equal volume of Nucleosome Dilution Buffer (20 mM Sodium Cacodylate pH 7.5, 10% v/v glycerol, 1 mM EDTA, 10 mM 2-mercaptoethanol, Filter Sterilized), and 1 µl was analysed by native PAGE (100 V in 1x TBE for 30 minutes), stained with SYBR Gold in 1xTBE for one hour, and visualized with a UV transilluminator gel imager. Fractions containing nucleosomes and minimal free DNA were pooled and diluted to a working concentration of ~ 1 nM with filter sterilized Nucleosome Storage Buffer (10 mM Sodium Cacodylate pH 7.5, 100 mM NaCl, 50% v/v glycerol, 1 mM EDTA, 1x Protease Inhibitor Cocktail [1 mM PMSF, 1mM ABESF, 0.8 µM aprotinin, 20 µM leupeptin, 15 µM pepstatin A, 40 µM bestatin, 15 µM E-64 from a 200x DMSO stock]) and stored at -20°C.

ICeChIP Nuclei Preparation: Mammalian and Insect Nuclei

Mammalian and insect nuclei preparation was performed as described[118,124,170,171]. Briefly, cell pellets were washed twice with 5 mL of PBS, then washed twice with 5 ml of filter sterilized Buffer N (15 mM Tris-HCl pH 7.5, 15 mM NaCl, 60 mM KCl, 8.5% w/v Sucrose, 5 mM $MgCl_2$, 1 mM $CaCl_2$, 1 mM DTT, 200 μM PMSF, 50 μg/mL BSA, 1x Roche Protease Inhibitor Cocktail), with each wash consisting of complete resuspension of the pellet, centrifugation at 500 g for 5 minutes at 4°C, and removal of supernatant. The washed pellet was then resuspended in at least 2 packed cell volumes (PCV) of Buffer N and mixed with 1 volume of 2x Lysis Buffer (Buffer N supplemented with 0.6% NP-40 Substitute) and incubated on ice for 10 minutes to lyse cells.

The crude nuclei were spun down at 500 g for 5 minutes at 4°C before being resuspended in at least 6 packed nuclear volumes (PNV) of Buffer N and applied to the top of 7.5 mL of filter sterilized Sucrose Cushion N (15 mM Tris-HCl pH 7.5, 15 mM NaCl, 60 mM KCl, 30% w/v Sucrose, 5 mM $MgCl_2$, 1 mM $CaCl_2$, 1 mM DTT, 200 μM PMSF, 50 μg/mL BSA, 1x Roche Protease Inhibitor Cocktail) in a 15 ml centrifuge tube, then spun down at 500 g for 12 minutes at 4°C in a swinging-bucket rotor. The supernatant was discarded, and the pellet resuspended in ~ 2 PNV of Buffer N.

The nucleic acid content of the nuclei per unit volume was quantified by diluting 2 μL of nuclei suspension into 48 μL of 2 M NaCl, water-bath sonicating to solubilize DNA, and spectroscopically measuring nucleic acid concentration by Nanodrop (where one A_{280nm} = 50 ng/μL chromatin). After accounting for the 25-fold dilution of the measurement sample, the concentration of the nuclei was adjusted to 1 μg/μL of chromatin. Nuclei were dispensed to 100 μL aliquots, flash frozen, and stored at -80°C prior to use.

ICeChIP Nuclei Preparation: Yeast Nuclei

Yeast were collected from culture by centrifugation at 4,500 g for 5 minutes; supernatant was discarded. The cell pellet was then resuspended into 20 mL of PBS into a single-cell suspension and pelleted by centrifugation at 500 g for 15 minutes at 4°C; supernatant was discarded. The pellet was then resuspended into Sorbitol Buffer (1.4 M sorbitol, 40 mM Tris-HCl, 0.5 mM $MgCl_2$, 1 mM PMSF, 2 mM β-mercaptoethanol, pH 7.5, filter sterilized) into a single-cell suspension and transferred into a weighed empty tube. The cells were pelleted by centrifugation at 500 g for 15 minutes at 4°C; supernatant was discarded. The cell pellet was then weighed by measuring weight of the tube and subtracting blank weight. Cells were then resuspended into 4 mL of Sorbitol Buffer per gram of cell pellet into a single-cell suspension. The suspension was placed on a shaker at 30°C for 20 minutes.

While incubating, Zymolase (Fisher) was added to the cell suspension to a final concentration of 0.5 mg/mL of Zymolase. The sample was then incubated for two hours to break down cell wall and produce spheroplasts. Spheroplasts were pelleted by centrifugation at 500 g for 15 minutes at 4°C. Spheroplasts were then washed once with Sorbitol Buffer and twice with Buffer N, with each wash consisting of complete resuspension of the pellet, centrifugation at 500 g for 5 minutes at 4°C, and removal of supernatant. The washed pellet was then resuspended in at least 2 PCV of Buffer N and mixed with 1 volume of 2x Lysis Buffer and incubated on ice for 10 minutes to lyse cells.

The crude nuclei were spun down at 3000 g for 10 minutes at 4°C. If two layers were observed in the pellet, the top layer was saved and the bottom discarded; the top layer contains the nuclei, and the bottom layer contains unlysed cells. The nuclei were resuspended into 6 PNV of Buffer N and applied to the top of 7.5 mL of filter sterilized Sucrose Cushion N in a 15 mL

tube, then spun down at 3000 g for 15 minutes at 4°C in a swinging-bucket rotor. The supernatant was discarded, and the pellet resuspended in ~ 2 PNV of Buffer N. The nucleic acid content of the nuclei per unit volume was quantified by diluting 2 µL of nuclei suspension into 48 µL of 2 M NaCl, water-bath sonicating to solubilize DNA, and spectroscopically measuring nucleic acid concentration by Nanodrop (where one A_{280nm} = 50 ng/µL chromatin). After accounting for the 25-fold dilution of the measurement sample, the concentration of the nuclei was adjusted to 1 µg/µL of chromatin. Nuclei were dispensed to 100 µL aliquots, flash frozen, and stored at -80°C.

ICeChIP Input Preparation

Input was prepared for ICeChIP and denaturative ICeChIP, and reICeChIP experiments as previously described[118,124,170,171]. For use, nuclei aliquots were thawed and spiked with ~ 1 µl of each barcoded nucleosome standard per 50 µg of chromatin. This suspension was then mixed by pipette, transferred to a new tube, and warmed to 37°C for 2 minutes. 1 unit of micrococcal nuclease (MNase, Worthington) per 4.375 µg of chromatin was added, and samples incubated at 37°C while shaking at 900 rpm for 12 minutes. Digestions were stopped by adding 1/9 volume of filter sterilized 10x MNase Stop Buffer while slowly vortexing, and nuclei lysed by adding 5 M NaCl to a final concentration of 600 mM while slowly vortexing. 66 mg of HAP resin (BioRad, CHT™ Ceramic Hydroxyapatite, Type I, 20 um) per 100 µg of chromatin digested was rehydrated with 200 µl of filter sterilized HAP Buffer 1 per 100 µg of chromatin digested. Lysed nuclei were centrifuged at 18,000 g for 1 minute to pellet insoluble nuclear debris, and the soluble fraction added to the rehydrated HAP resin and incubated for 10 minutes at 4°C with rotation.

After incubation, the HAP resin slurry was added to a centrifugal filter unit (Millipore Ultrafree MC–HV Centrifugal Filter 0.45 µm) and spun at 1000 g for 30 seconds at 4°C. The HAP

resin left on the filter unit was then washed 4 times with 200 µL HAP Buffer 1, and 4 times with 200 µl filter sterilized HAP Buffer 2 by spinning at 1000 g for 30 seconds at 4°C. HAP resin was eluted into a clean tube with three 100 µl solutions of filter sterilized HAP Elution Buffer. The nucleic acid content of the elution was then quantified by Nanodrop.

Antibody Preparation for ICeChIP

Antibodies and quantities used for each ICeChIP experiment are shown in Appendix A. The indicated amount of Protein A Dynabeads (Invitrogen) for each ICeChIP was washed with 50 µL of ChIP ChIP Buffer 1 by use of a magnetic rack, then resuspended in 50 µL of ChIP Buffer 1. In a separate set of tubes, the antibody was diluted to 100 µL with ChIP Buffer 1. The antibody and Protein A Dynabead suspensions were combined and incubated on a rotator at 4°C for at least one hour, then washed with 200 µL of ChIP Buffer 1 by use of a magnetic rack and resuspended in 50 µL of ChIP Buffer 1.

Standard ICeChIP Immunoprecipitation

After antibodies were prepared and washed, the input chromatin concentration adjusted to 20 ng/µl with filter sterilized ChIP Buffer 1, and the amount of chromatin specified in Appendix A was added to each antibody-bead conjugate and incubated for 15 minutes on a rotator at 4°C. Beads were then washed twice with filter sterilized ChIP Buffer 2 (25 mM Tris pH 7.5, 5 mM $MgCl_2$, 300 mM KCl, 10% v/v glycerol, 0.1% v/v NP-40 Substitute) and once with filter sterilized ChIP Buffer 3 (10 mM Tris pH 7.5, 250 mM LiCl, 1 mM EDTA, 0.5% Sodium Deoxycholate, 0.5% v/v NP-40 Substitute), with a wash consisting of removal of the existing supernatant by use of a magnetic rack, resuspension into 150 µl of buffer, transfer to a new siliconized tube, and incubation on the rotator for 10 minutes at 4°C. After these washes, the supernatant was removed, the beads resuspended in

ChIP Buffer 1, transferred to a new siliconized tube, rinsed once with 200 µl of TE before being resuspended in 50 µl of ChIP Elution Buffer (50 mM Tris pH 7.5, 1 mM EDTA, 1% w/v SDS, Filter Sterilized) and incubated at 55°C for 5 minutes.

After incubation, the supernatant was transferred to a new set of siliconized tubes, and the beads discarded. To each supernatant was then added 2 µl of 5 M NaCl, 1 µl of 500 mM EDTA, and 1 µl of 10 mg/mL Proteinase K. 15 µl of Input DNA was also diluted to 50 µl with 35 µl of ChIP Elution Buffer and was supplemented with 2 µL of 5 M NaCl, 1 µL of 500 mM EDTA, and 1 µL of 10 mg/mL Proteinase K. The IP elutions and diluted input were then incubated at 55°C for 2 hours for a Proteinase K digestion. After digestion, the DNA was purified by adding 1.5 volumes of Serapure HD (1:50 dilution of Sera-Mag SpeedBeads [Fisher], 20% PEG-8000, 2.5 M NaCl, 10 mM Tris pH 7.5, 1 mM EDTA, 0.05% Tween-20, Filter Sterilized prior to addition of SpeedBeads), incubating at room temperature for 15 minutes, then collecting the beads on a magnetic rack, washing twice with 150 µl of 70% ethanol, and eluting into 50 µl ddH2O, which was then recovered and stored at -20°C.

Denaturative ICeChIP Immunoprecipitation

After purification of input chromatin and preparation of antibodies, the input chromatin was cross-linked with 1/9 volume of 2.5% formaldehyde stock (final concentration 0.25% formaldehyde) on a rotator at room temperature for 8 minutes. Cross-linking was then quenched with 1/5 volume of 1 M Tris-HCl, pH 7.5 on a rotator at room temperature for 5 minutes. 50 uL cross-linked chromatin was aliquoted into a thin-walled PCR tube, and 2.5 µL of 20% SDS was added (final concentration 1% SDS). This sample was then heated to 55°C for 60 seconds, then immediately put on ice. After

cooling, the sample was diluted with 450 µL of water. The concentration of chromatin will be 7.5% of the concentration from the end of input preparation.

The amount of chromatin specified in Appendix A was added to each antibody-bead conjugate and incubated for 15 minutes on a rotator at 4°C. Beads were then washed once with filter sterilized Crosslink ChIP Buffer 1 (50 mM HEPES, 140 mM NaCl, 1 mM EDTA, 1 mM EGTA, 0.75% Triton-X-100, 0.1% SDS, 0.05% DOC, pH 7.5) and once with filter sterilized Crosslink ChIP Buffer 2 (50 mM HEPES, 500 mM NaCl, 1 mM EDTA, 1 mM EGTA, 0.75% Triton-X-100, 0.1% SDS, 0.05% DOC, pH 7.5), with a wash consisting of removal of the existing supernatant by use of a magnetic rack, resuspension into 150 µl of buffer, transfer to a new siliconized tube, and incubation on the rotator for 10 minutes at 4°C. After these washes, the supernatant was removed, the beads resuspended in ChIP Buffer 1, transferred to a new siliconized tube, rinsed once with 200 µl of TE before being resuspended in 50 µl of ChIP Elution Buffer and incubated at 55°C for 5 minutes.

After incubation, the supernatant was transferred to a new set of siliconized tubes, and the beads discarded. To each supernatant was then added 2 µl of 5 M NaCl, 1 µl of 500 mM EDTA, and 1 µl of 10 mg/mL Proteinase K. 15 µl of Input DNA was also diluted to 50 µl with 35 µl of ChIP Elution Buffer and was supplemented with 2 µL of 5 M NaCl, 1 µL of 500 mM EDTA, and 1 µL of 10 mg/mL Proteinase K. The IP elutions and diluted input were then incubated at 55°C for 2 hours for a Proteinase K digestion. After digestion, the DNA was purified by adding 1.5 volumes of Serapure HD, incubating at room temperature for 15 minutes, then collecting the beads on a magnetic rack, washing twice with 150 µl of 70% ethanol, and eluting into 50 µl ddH2O, which was then recovered and stored at -20°C.

DNA Quantification and Analysis by Quantitative PCR

To assess local histone modification density and/or antibody specificity, our DNA from the ChIP experiments was quantified by quantitative PCR (qPCR). qPCR was conducted using TaqMan Gene Expression Master Mix (Applied Biosystems) using the primers and hydrolysis probes previously described[118]. These primers and probe for the barcoded sequences were previously qPCR validated for effectiveness and quality[118]. Primers were used at 900 nM; hydrolysis probe at 250 nM, in the TaqMan Gene Expression Master Mix (Applied Biosystems). The qPCR program was run at 95°C for 10 minutes, followed by 40 cycles, each consisting of 15 seconds at 95°C followed by 1 minute at 60°C and concluding with a plate read.

Cq values were analysed using the ΔΔCq method. Briefly, the Cq values for each target for each sample were averaged together to obtain the mean Cq value. Enrichment for each barcode was then computed as Enrichment $= 2^{Cq_{INPUT} - Cq_{IP}} * 10$, accounting for the 10-fold dilution of Input relative to IP and multiplying by 100% for Enrichment as a percentage of target. Off-target binding to alternate PTMs were computed by normalizing each enrichment to that of the on-target PTM: referred to as "Specificity (% Target)".

Illumina Library Preparation and Sequencing

Illumina libraries were prepared as described[118], with minor modifications. Briefly, Serapure purified DNA was quantified using Quant-iT^{TM} PicoGreen (Thermo Fisher) as per manufacturer instructions. Libraries were then generated from up to 10 ng of each DNA sample (input or IP) with the NEBNext Ultra II DNA Library Prep kit (New England Biolabs) per manufacturer instructions. The DNA content of each library was then quantified and pooled for Illumina sequencing. Clus-

ter generation and paired-end sequencing was conducted using standard Illumina next-generation sequencing protocols by the University of Chicago Genomics Facility on the Illumina NextSeq.

Next-Generation Sequencing Alignment and HMD Computation

To align reads, a reference genome was first created, consisting of the human genome (hg38) appended respectively by the sequences of each of the nucleosome standard barcodes for the relevant barcode set. Reads were then mapped to the appropriate reference genome using Bowtie2 using the sensitive pre-set and end-to-end alignment options[172]. Using SAMTools[173], any reads which were not paired, not mapped in a proper pair, or mapped with a map quality < 20 were discarded to prevent low-quality reads from impacting downstream analyses. Reads were then flattened to create a single mapping from each matched pair of reads by retaining only one fragment per pair, and any mappings with lengths > 200bp were also discarded to ensure only mononucleosomes were being analyzed[118].

Bedgraphs of genome coverage were then generated using BEDTools[174], and IP / input genome coverage bedgraphs were merged using BEDTools[174]. The sum of reads across ladder members for each nucleosomal standard was computed for each sample and HMD bedgraphs were then generated from the merged bedgraphs using awk to apply the following formula:

$$\text{HMD } (\%) = 100\% * \frac{\text{IP}_{\text{locus}}/\text{Input}_{\text{locus}}}{\text{IP}_{\text{barcode}}/\text{Input}_{\text{barcode}}}$$

Error and 95% confidence intervals were computed with Poisson statistics and error propagation from the merged bedgraphs using awk to apply the following formula:

$$\text{95CI Error } (\%) = 1.96 * \text{HMD } (\%) * \sqrt{\frac{1}{\text{IP}_{\text{locus}}} + \frac{1}{\text{Input}_{\text{locus}}}}$$

Bigwig files were generated for visualization using the bedGraphToBigWig tool[175].

For all analyses, the HMD averaged over the N+1 and N+2 nucleosomes (taken to be 0 to +400bp into the gene body) was employed as representative of the promoter—this captures the most substantial H3K4me3 and H3K27me3 enrichment.

Genomic browser views were made using IGV. Heatmaps and gene ontology analysis was made using Homer software[178]. Further analysis and sectioning of data was conducted in R using the R code provided in Data and Software Availability.

Data and Software Availability

R markdown file for analysis and sectioning of datasets is provided at `https://www.github.com/shah-rohan/h3k79_analysis/`.

CHAPTER 5: BAYESIAN RESOLUTION OF AMBIGUOUSLY MAPPED READS

Attributions

This chapter has been adapted from: Shah, R. N. & Ruthenburg, A. J. Sequence deeper without sequencing more: Bayesian resolution of ambiguously mapped reads. *PLOS Computational Biology* **17,** e1008926 (2021). All work for this chapter was conducted by the author.

Abstract

Next-generation sequencing (NGS) has transformed molecular biology and contributed to many seminal insights into genomic regulation and function. Apart from whole-genome sequencing, an NGS workflow involves alignment of the sequencing reads to the genome of study, after which the resulting alignments can be used for downstream analyses. However, alignment is complicated by the repetitive sequences; many reads align to more than one genomic locus, with 15-30% of the genome not being uniquely mappable by short-read NGS. This problem is typically addressed by discarding reads that do not uniquely map to the genome, but this practice can lead to systematic distortion of the data. Previous studies that developed methods for handling ambiguously mapped reads were often of limited applicability or were computationally intensive, hindering their broader usage. In this work, we present SmartMap: an algorithm that augments industry-standard aligners to enable usage of ambiguously mapped reads by assigning weights to each alignment with Bayesian analysis of the read distribution and alignment quality. SmartMap is computationally efficient, utilizing far fewer weighting iterations than previously thought necessary to process alignments and, as such, analyzing more than a billion alignments of NGS reads in approximately one hour on a desktop PC. By applying SmartMap to peak-type NGS data, including MNase-seq, ChIP-seq, and ATAC-seq in three organisms, we can increase read depth by up to 53% and increase the mapped

proportion of the genome by up to 18% compared to analyses utilizing only uniquely mapped reads. We further show that SmartMap enables the analysis of more than 140,000 repetitive elements that could not be analyzed by traditional ChIP-seq workflows, and we utilize this method to gain insight into the epigenetic regulation of different classes of repetitive elements. These data emphasize both the dangers of discarding ambiguously mapped reads and their power for driving biological discovery.

Introduction

The impact of next-generation sequencing (NGS) on molecular biology can hardly be overstated. In a typical short-read NGS workflow, DNA fragments from an experiment are loaded onto a sequencer, which reports the sequence of 40-200bp of one end or both ends of each fragment (in single-end or paired-end sequencing, respectively)[294]. These reads/read pairs can then be aligned to the genome by one of several alignment tools, and the set of alignments can be used to compute the number of reads aligned to any given genomic locus. This genome-wide read depth dataset can then be used in downstream workflows.

Even beyond applications for whole genome sequencing, many critical methods have leveraged NGS to enable truly genome-wide biological studies. RNA sequencing (RNA-seq) has enabled quantification of gene expression[295] as well as the discovery and characterization of new elements of the transcriptome, such as enhancer RNAs[58,59,153,296] and chromatin-associated RNAs[70,71]. Chromatin immunoprecipitation coupled to NGS (ChIP-seq) has similarly become a mainstay of molecular biology, being used in many seminal works of the field[17–19,127,130,131,297–299]. Other common techniques, including ATAC-seq[72], Hi-C[300], CUT&RUN[121], and TAB-seq[301], similarly rely on NGS and associated workflows to provide important insights into genomic regulation.

Crucially, these workflows all rely upon alignment of each read to its corresponding genomic location. However, this can be problematic when analyzing non-unique or repetitive regions of the genome, particularly given the short window of a 40-200bp sequencing read. Indeed, some estimates suggest that a majority of the human genome is comprised by repetitive elements[22,302,303]. Accordingly, between 15-30% of the human genome is not uniquely mappable by single-end sequencing with typical read lengths[304,305], and the genomes of other model organisms, such as *M. musculus* or *D. melanogaster*, present similar challenges[304]. Paired-end sequencing can partially improve genome mappability, but of the regions that are not uniquely mappable by single-end sequencing, 70-85% will not be resolved by paired-end sequencing[304].

Many NGS pipelines address this ambiguity by masking repetitive regions to prevent alignment of reads to more than one genomic locus or by filtering only for reads that align unambiguously to the genome (hereafter referred to as unireads)[306]. This includes groups such as the ENCODE Consortium, whose ChIP-seq pipeline filters for uniquely mapped reads by default[129]. Indeed, in several of our past studies, we ourselves have utilized filters to exclude ambiguously mapping reads[118,124,171]. However, filtering out reads that map to multiple loci (hereafter referred to as multireads) sacrifices the ability to critically examine many repetitive regions of the genome, which have important roles in gene regulation[306]. Further, by definition, discarding reads reduces read depth, which makes quantitative comparisons more challenging by increasing error or the necessary sequencing depth[306]. Given the many problems with ignoring repetitive regions or ambiguous alignments, it is critical to develop and utilize methods to appropriately analyze multireads.

To date, several studies have attempted to develop methods and algorithms to resolve multiread alignments for a variety of applications. Some have targeted their analysis methods towards RNA-seq and quantifying transcripts[295,307,308]; indeed, in recent years, there has been a sharp in-

crease in the tools available for quantification of pre-defined genomic features in RNA-seq[309]. Others have developed tools designed for ChIP-seq or DNA-seq more broadly[310–314].

Despite the wide array of tools that have been previously developed for this problem, there are still several outstanding problems. First, several of the previously published tools (particularly for RNA-seq) focus on quantification of a distinct set of genomic features rather than generating truly genome-wide coverage maps[295,307–309,313,314], rending them inappropriate for ChIP-seq or other unbiased/*de novo* NGS analyses. Even amongst these remaining tools for "peak type" ChIP-seq or similar analyses, several of these tools focusing on comparison to external datasets for peak calling[313,314], leaving even fewer analysis methods for a single dataset without an exogenous reference. Second, while many existing methods use alignment weighting algorithms to allocate multiread depth, there is disagreement as to the degree to which iterative reweighting is required to properly weight the multireads without over-refining the weights; some employ no iterative reweighting at all[295,312], whereas others use up to 200 reweighting cycles[310]. In addition, most of the above methods do not consider the alignment quality when resolving read ambiguity or does so in a computationally intensive manner that would likely scale poorly with the number of reads commonly obtained from modern NGS platforms[311]. Further, these tools often focused on single-end sequencing and do not make use of the intervening length information in paired-end sequencing, limiting the scope of their applicability[310]. Finally, many of these tools do not accommodate strand-specific analyses genome-wide, limiting their application to strand-independent experiments[310,311,313,314].

In this work, we seek to resolve some of these issues. We describe SmartMap: an algorithm that uses iterative Bayesian reweighting of ambiguous mappings, with assessment of alignment quality as a factor in assigning weights to each mapping. We find that SmartMap markedly increases the number of reads that can be analyzed and thereby improves counting statistics and read depth

recovery at repetitive loci. This algorithm and software implementation is compatible with both paired-end and single-end sequencing and can be used for both strand-independent and strand-specific methods employing NGS backends to generate genome-wide read depth datasets.

Results

Development and validation of a Bayesian multiread allocation algorithm

We initially developed our SmartMap algorithm and software for application in ChIP-seq using a set of internally calibrated ChIP-seq (ICeChIP-seq) datasets. These datasets were previously generated by our lab and, with one exception, were previously published as components of past studies[118,124]. We chose to use ICeChIP-seq datasets because the included internal standards allow for computation of antibody specificity and for normalization to calculate the histone modification density (HMD), or the absolute proportion of nucleosomes at a given genomic locus bearing the targeted histone modification. These additional factors which we can compute using ICeChIP-seq datasets afford us additional points of quantitative comparison to assess differences between uniread and SmartMap analyses. However, this tool is not designed solely (or even primarily) for use with ICeChIP-seq datasets; the SmartMap algorithm does not make special use of the internal standards. Rather, this software is designed to be usable for NGS workflows more broadly.

The workflows for uniread analyses (typical of ChIP-seq) and our SmartMap analysis are shown in Fig. 5.1A. For both analyses, the immunoprecipitation (IP) and MNase-seq Input sequences are aligned to the appropriate reference genome and are filtered to select for properly mapped reads in a proper pair. At that point, the two methods diverge. In the uniread analysis, which represents our published analysis pipeline for ICeChIP-seq data[118,124,171], any reads that don't

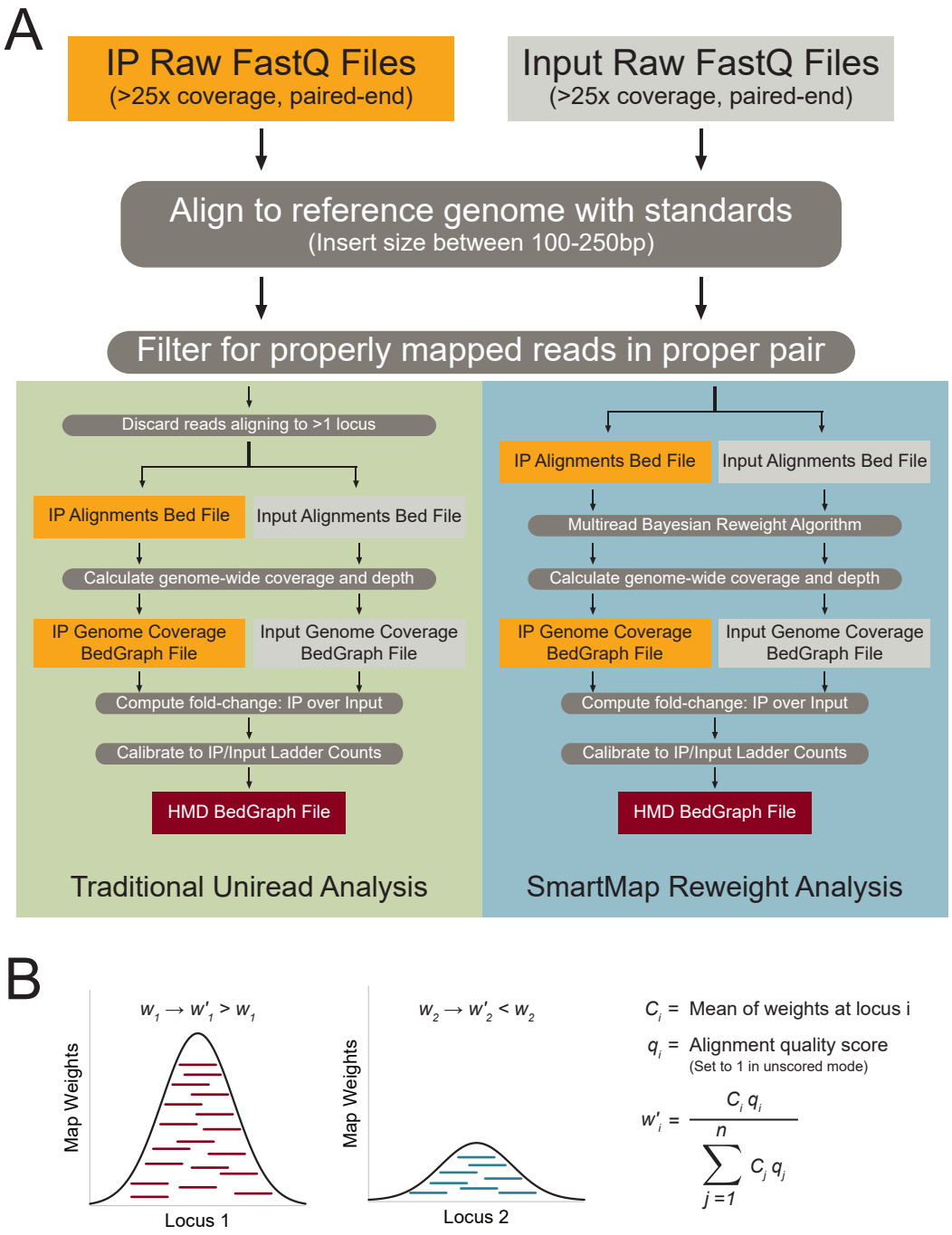


Figure 5.1: Summary of the SmartMap analysis workflow and algorithm.

(A) Flowchart outlining the workflow for traditional ChIP-seq (or ICeChIP-seq) analysis[118,124,171] utilizing only unireads (left, green) vs. the workflow for SmartMap analysis utilizing multireads with an iterative Bayesian reweighting algorithm (right, blue). **(B)** Schematic showing the Bayesian reweighting algorithm utilized in the SmartMap analysis. Each mapping associated with a read is assigned a weight such that the weight is greater for those mappings associated with loci of greater map weight density. For more detailed description of the algorithm, see Methods.

align uniquely are discarded, and the remainder are used to compute genome-wide read depth in the IP and Input, fold-change, and (if internal standards are present) HMD.

In the SmartMap analysis, however, rather than discarding ambiguously mapped reads, we instead feed our alignments into our iterative Bayesian reweighting algorithm, outlined in Fig. 5.1B. Our algorithm, like other alignment weighting algorithms[295,310–312], is motivated by the assumption that regions with more alignments are more likely to be the true source of an multiread than those with fewer alignments. In addition, like BM-Map[311], SmartMap utilizes both paired-end sequencing information and alignment quality in making these assessments. Accordingly, our tool first assigns each alignment a weight proportional to its alignment quality, computed from the alignment software output. We then iteratively reassign weights to each alignment of each read; alignments with higher alignment quality and more overlapping alignments are assigned higher weights, and those alignments with lower quality and fewer overlapping alignments are assigned lower weights (Fig. 5.1B). After the specified number of reweighting cycles, the resulting weights are used to compute the read depth for the IP and the Input genome-wide, which can then be used to compute fold-change or, if applicable, HMD in a similar manner as the uniread analysis. For computational efficiency, we use binary-indexed (Fenwick) trees to store genomic coordinates and associated alignment weights, much like the previously described CSEM[310]. Our implementation of these binary-indexed trees is modified to enable use of paired-end sequencing reads and, if needed, operate in a strand-specific manner.

To test this method, we created a set of simulated 50bp paired-end sequencing reads from a defined set of randomly selected genomic loci (the "true origin" loci) and used the simulated dataset to conduct uniread and SmartMap analyses (Fig. 5.2A). The read simulation tool produces reads with "sequencing" error and also includes coverage at off-target loci to better represent the noise

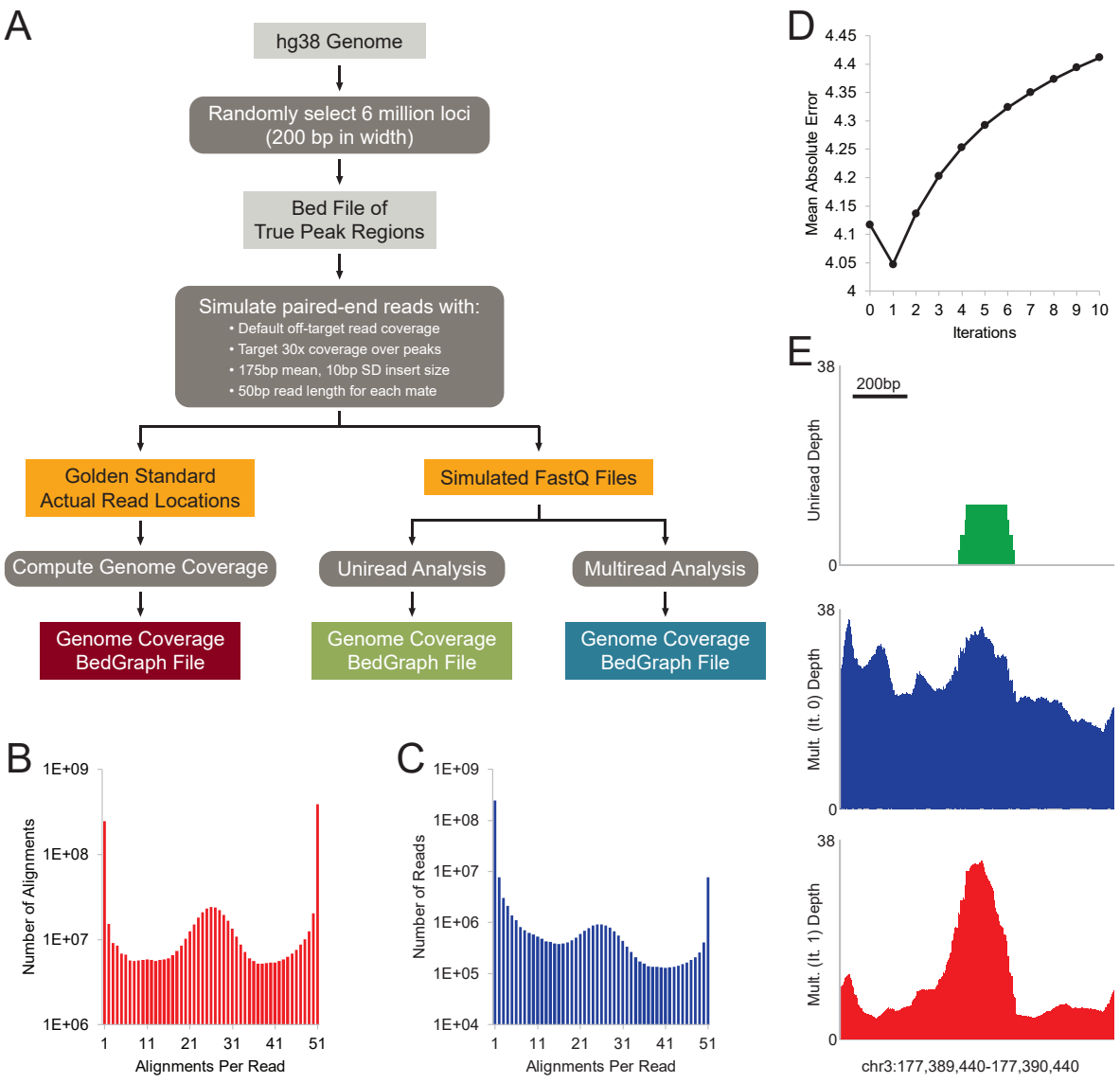


Figure 5.2: Characteristics of validation dataset.

(A) Schematic outlining the workflow to validate and optimize SmartMap. A set of six million randomly selected 200bp loci were used to simulate paired end reads. The true read depth distribution was then compared to both uniread and SmartMap analyses, with each analysis conducted in both "scored" and "unscored" modes, per Methods. **(B, C)** Number of (B) alignments or (C) reads vs. number of alignments per read for the validation datasets. **(D)** Mean absolute error of read depth at true origin loci in SmartMap scored mode vs. number of reweighting iterations **(E)** Genome browser view showing the read depth in the (top) uniread, (center) SmartMap (0 iterations), and (bottom) SmartMap (1 iteration) datasets of an example locus.

and off-target capture inherent in a biological experiment. Notably, the simulation enabled us to obtain the true distribution of reads (the Gold Standard), allowing us to compute the error associated

with each analysis method (Fig. 5.2A). This is particularly important because we wish to avoid over-refining the multiread weights with our inferential analysis; accordingly, the Gold Standard dataset allows us to evaluate the accuracy of our reweighting algorithm and reallocation.

We were particularly interested in the ability of SmartMap to recover read depth at regions of differing mappability. To investigate this relationship, we used the UMAP50 score as a measure of read mappability. The UMAP50 score for a given genomic coordinate is computed as the proportion of the 50mers covering the genomic coordinate of interest that are unique in the set of all 50mers from the genome[305]. For example, if the sequences of two of the fifty 50mers containing the genomic coordinate of interest were non-unique across the genome of study, then the UMAP50 score would be 48/50, or 0.96. As such, a genomic coordinate with a UMAP50 score closer to 1 is uniquely identified by a greater proportion of the 50mers spanning it than is a coordinate with a lower UMAP50 score, and a higher UMAP50 score can thus be interpreted as a more easily mappable region. Many of the true origin loci had low mappability scores (Fig. 5.3A), with the distribution of mappability scores being similar to that of the human genome at large (Fig. 5.3B), making this dataset useful for validating the SmartMap algorithm.

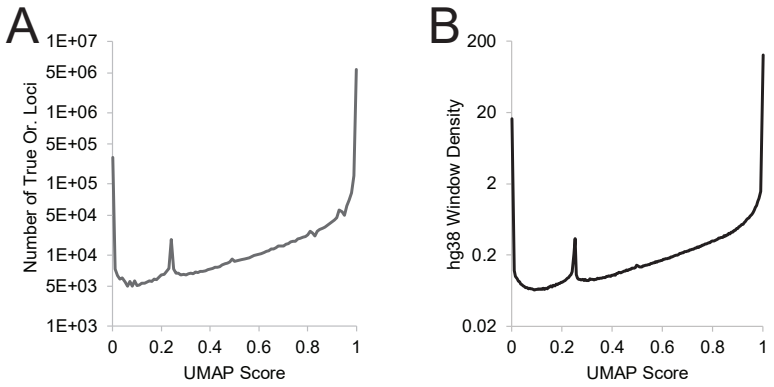


Figure 5.3: Mappability of sampled loci and human genome.
(A) Number of regions from the true origin loci vs. average mappability (UMAP50) score of the loci. **(B)** Density of UMAP50 scores of 200bp windows across the human genome (hg38).

The first step of our analysis was to align the simulated 50bp paired-end reads to the genome. We used Bowtie2[172] with a maximum of 51 alignments reported per read and charted the distributions of the number of alignments per read (Fig. 5.2B-C). Notably, we observed that there were many reads that did not uniquely align to the genome; approximately 17.1% of the simulated reads mapped to more than one locus (Fig. 5.2B-C and Table 5.1).

Our first goal was to determine the optimal number of iterations to use for our SmartMap analyses. To test this, we computed the mean absolute error of SmartMap read depth at the true origin loci with varying numbers of reweighting cycles, as compared to the Gold Standard read depth. Surprisingly, we found that the lowest error occurred after only one reweighting cycle (Fig. 5.2D), with genome browser views showing refinement of peak structure (Fig. 5.2E), which is particularly important given the importance that has been placed on peak breadth[155]. This stands in stark contrast with previous works, which have used up to 200 iterations of reweighting[310]. Our analysis here, however, shows that may be suboptimal, suggesting that applying Bayesian alignment reweighting more than once may over-refine the data.

We wanted to explore whether these increases in mean absolute error were systematic or driven by random "overshoot" of weight at each locus. In the former case, we might expect to see that the true origin loci would either show systematic increases or decreases in read weight with greater numbers of reweighting cycles. In the latter case, we would expect that the changes to each weight might increase or decrease by too much in the initial iteration, which would present as random, relatively unbiased errors.

To distinguish between these two possibilities, we conducted two analyses. First, we computed mean error of weights at the true origin loci (Fig. 5.4A) rather than the mean absolute error (Fig. 5.2D). If there was a systematic erroneous increase or decrease in the average read depth of

Table 5.1: Alignment statistics for the datasets used for multiread analysis.

	Sample	Genome	Assay	Cells	Unireads	Multireads		
						Analyzable	Un-analyzable	% Incr.
Simulated	Simulated, 50bp	hg38	Simulation	–	245,079,644	34,136,124	7,661,326	13.93%
	Simulated, -k 101	hg38	Simulation	–	244,391,815	35,520,969*	6,973,053*	14.53%
	Simulated, 100bp	hg38	Simulation	–	123,730,306	16,769,189	2,802,056	13.55%
AR7	Input Rep. 1	mm10*	MNAse-seq	mESC E14	311,090,692	85,018,787	15,184,872	27.33%
	H3K4me3 Rep. 1	mm10*	ChIP-seq	mESC E14	119,014,494	19,662,529	5,603,383	16.52%
	Input Rep. 2	mm10*	ChIP-seq	mESC E14	304,127,899	83,629,528	17,160,012	27.50%
	H3K4me3 Rep. 2	mm10*	ChIP-seq	mESC E14	91,518,104	14,549,072	4,657,032	15.90%
AR8	Input	dm3†	MNAse-seq	S2	18,678,956	7,117,520	977,776	38.10%
	H3K27me3	dm3†	ChIP-seq	S2	8,855,114	3,249,005	389,227	36.69%
AR9	Input	mm10†	MNAse-seq	mESC E14	488,503,092	131,960,514	26,577,525	27.01%
	H3K4me3	mm10†	ChIP-seq	mESC E14	169,335,369	32,089,449	7,918,756	18.95%
	H3K9me3	mm10†	ChIP-seq	mESC E14	136,008,760	73,118,061	13,012,319	53.76%
	H3K27me3	mm10†	ChIP-seq	mESC E14	155,322,021	43,508,387	9,267,806	28.01%
AR16	Input	hg38‡	MNAse-seq	K562	285,996,344	56,595,547	12,902,707	19.79%
	H3K4me1	hg38‡	ChIP-seq	K562	92,422,802	16,475,108	2,434,216	17.83%
	H3K4me2	hg38‡	ChIP-seq	K562	70,987,452	12,931,282	2,558,979	18.22%
	H3K4me3	hg38‡	ChIP-seq	K562	40,483,145	5,488,996	803,892	13.56%

Table 5.1 continues on next page.

Table 5.1, continued:

	Sample	Genome	Assay	Cells	Unireads	Multireads		% Incr.
						Analyzable	Un-analyzable	
AR17	Input	hg38‡	MNAse-seq	K562	256,373,920	48,634,887	11,216,500	18.97%
	H3K9me3	hg38‡	ChIP-seq	K562	193,011,406	40,618,196	10,337,413	21.04%
	H3K27me3	hg38‡	ChIP-seq	K562	173,915,939	32,770,085	7,107,199	18.84%
ENCODE	Snyder Rep. 1	hg38	ATAC-seq	K562	32,995,935	6,484,894	299,834	19.65%
	Snyder Rep. 2	hg38	ATAC-seq	K562	24,414,870	4,210,386	149,154	17.25%
	Gingeras Rep. 1	hg38§	RNA-seq	K562	60,184,580	20,651,064	29,231	34.31%
	Gingeras Rep. 2	hg38§	RNA-seq	K562	63,238,387	13,087,755	14,070	20.70%

For all except the ENCODE RNA-seq datasets, analysis is conducted on 200bp genomic windows. For ENCODE RNA-seq datasets, analysis is conducted on distinct Refseq genes.

% Reg. Inc.: Percent of the total regions in the SmartMap dataset with increased read depth relative to the Uniread dataset.

% Inc. Reg.: Percent increase in the number of regions with reads in the SmartMap dataset relative to the Uniread dataset.

Genome includes ICeChIP barcodes: * Series 1. † Series 2. ‡ Series 3.

§ Genome includes ENCODE ERCC standards.

each locus, then we would observe a corresponding increase or decrease in mean error with more iterations, respectively. However, what we instead observe is that the mean error is relatively stable from iterations 2-8 (Fig. 5.4A), suggesting that the marked increase in mean absolute error with increasing iterations is not primarily caused by systematic erroneous increases or decreases in locus weight depth. Put differently, it does not appear that the true loci are systematically "pulling in" or "pushing out" read depth with each reweighting cycle.

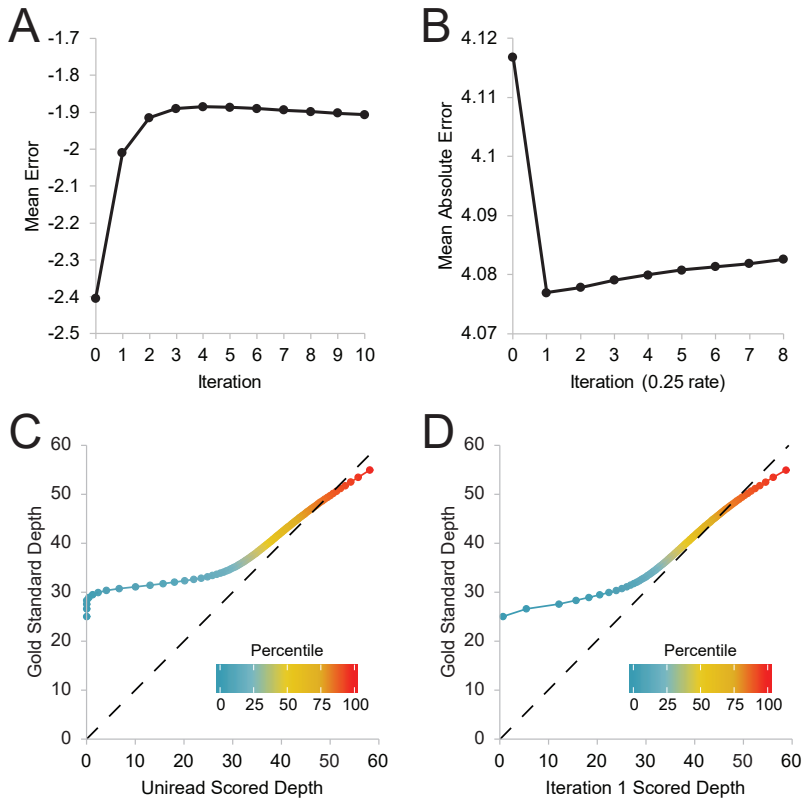


Figure 5.4: Characteristics of SmartMap with increasing iterations.
(A) Mean error of read depth at true origin loci in SmartMap scored mode vs. number of reweighting iterations. **(B)** Mean absolute error of read depth at true origin loci in SmartMap scored mode with a reweighting rate of 0.25 vs. number of reweighting iterations. **(C, D)** QQ plots of read depth in Gold Standard dataset vs. (C) uniread or (D) SmartMap (1 iteration) scored datasets. Color scale represents percentile of each point, from 1st to 99th percentiles. Dashed line represents line with slope of unity.

Second, we explored the possibility that the reweighting "overshoots" the weight adjustment for reads at random. If this was the case, then we would expect that the errors would increase relatively randomly, with both positive and negative errors. Indeed, this is what we observe in our analysis of mean error by iteration (Fig. 5.4A). In addition, we would predict slowing the rate of weight adjustment with each cycle would decrease the amount of overshoot and thereby lead to a lesser increase in error. To test this, we introduced a tunable reweighting rate parameter such that the weights could be changed less with each reweighting iteration. When applying a reweighting

rate of 0.25 (wherein the weights only change by 25% as much in normal SmartMap analysis), we found that the mean absolute error was markedly more stable after one iteration (Fig. 5.4B). Indeed, after two cycles of standard SmartMap, the mean absolute error exceeds that of the dataset with no reweighting (Fig. 5.2D); by contrast, with eight cycles of SmartMap with a reweighting rate of 0.25, the mean absolute error is considerably below that of the iteration 0 dataset and comparable to the minimum mean absolute error after one iteration (Fig. 5.4B). This suggests that the increase in error with increasing iterations observed with standard SmartMap may be due to "overshoot" of reweighting, which compounds in magnitude with further reweighting. Interestingly, we found that the mean absolute error with one iteration of standard SmartMap analysis was on par with (and even slightly lower than) that of the slow-reweighting dataset (Fig. 5.2D and 5.4B), suggesting that this potential overshoot error may not be too detrimental after only one iteration of reweighting. By command line switch, these two algorithms are both available in the SmartMap software.

After determining the optimal number of reweighting cycles, we compared SmartMap and uniread analyses of our simulated datasets (Fig. 5.2A). To determine the relative impact of using alignment quality for multiread analysis, we ran SmartMap in both scored and unscored modes. All the SmartMap analyses had greater read depth (and were closer to the Gold Standard dataset) at true origin loci than the corresponding uniread analyses (Fig. 5.5A). Interestingly, the increases in read depth were not uniform across the entire set of loci; indeed, approximately 70% of the true origin loci saw no excess read depth, defined as the difference between SmartMap and uniread read depths (Fig. 5.5B). This is similarly observed in the QQ plot comparing uniread and SmartMap analyses; a shoulder is seen at low uniread depth, with the plot converging onto a slope of unity at higher read depths (Fig. 5.4C and 5.5C), suggesting that the gains in read depth were primarily at regions of low uniread depth.

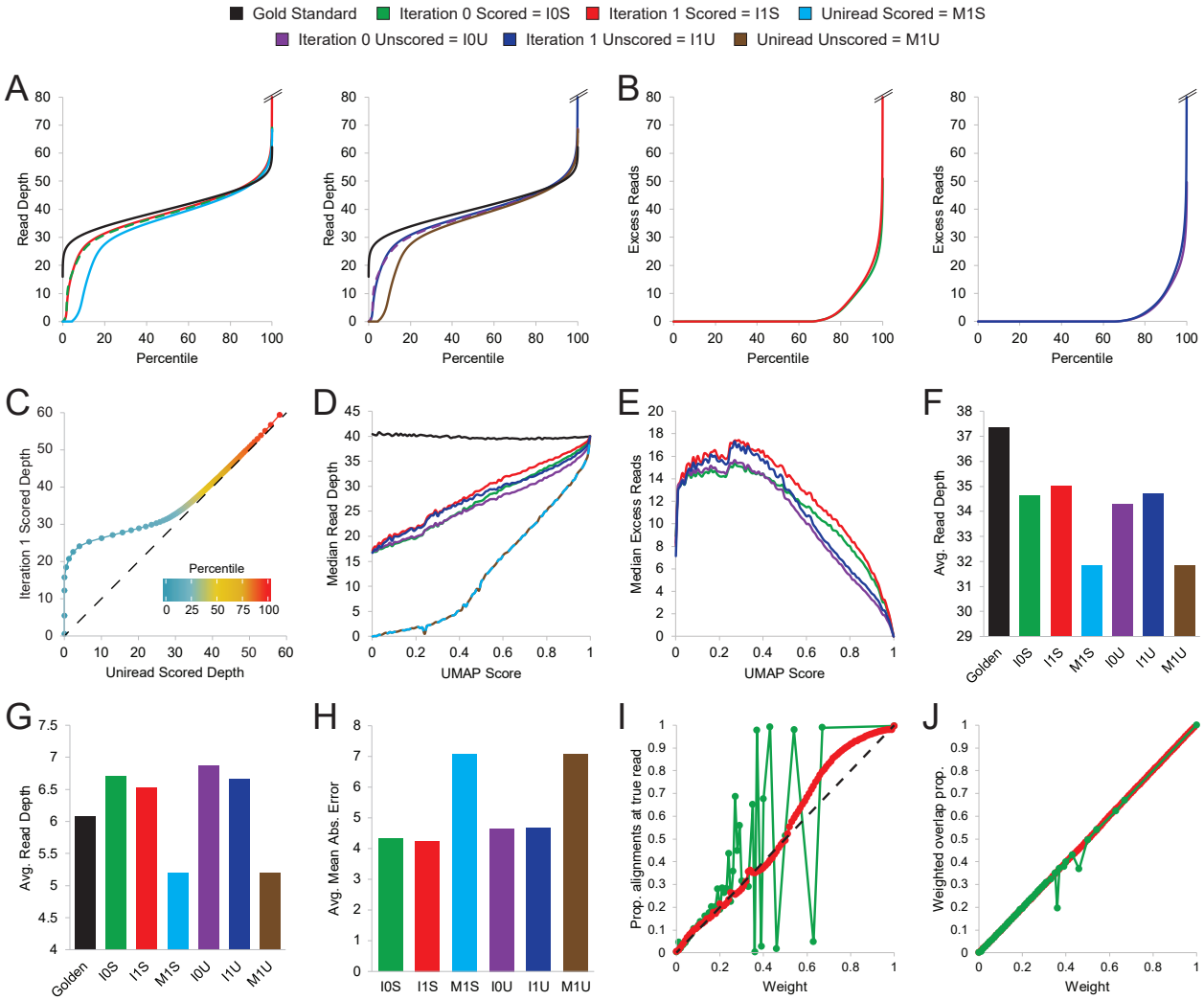


Figure 5.5: SmartMap and uniread analyses of the validation dataset.

Iteration 0 and iteration 1 refer to SmartMap analysis with 0 and 1 iterations of reweighting, respectively. Scored and unscored refer to whether alignment score was considered in analysis, per Methods. Dashed lines are presented for readability of overlapping curves rather than discontinuities in data throughout this figure. **(A)** Quantile plot of read depth at the true origin loci, with Gold Standard dataset and analysis conducted in (left) scored mode or (right) unscored mode. **(B)** Quantile plot of excess read depth in SmartMap datasets relative to corresponding uniread dataset at true origin loci in (left) scored mode and (right) unscored mode. **(C)** QQ plot of read depth at true origin loci in the SmartMap (1 iteration) scored dataset vs. uniread scored dataset. Color scale represents percentile of each point, from 1st to 99th percentiles. **(D-E)** Median (D) read depth or (E) excess read depth vs. mappability score (UMAP50)[305] of the true origin loci. **(F-G)** Average read depth (F) at true origin loci and (G) outside true origin loci. **(H)** Mean absolute error of read depth at true origin loci for each dataset, with Gold Standard as the reference point. **(I)** Mean proportion of alignments intersecting with the true read of origin for each weight after SmartMap with no reweighting (green) and one iteration of reweighting (red) in scored mode. Dashed line

Figure 5.5, continued:
represents line with slope of unity. **(J)** Mean weighted overlap proportion score between alignments intersecting the true read of origin and the true read locus for each weight after SmartMap with no reweighting (green) and one iteration of reweighting (red) in scored mode. Weighted overlap proportion score is meant to represent the proportion of a read's weight that maps to the correct location due to a particular alignment and is computed as a weighted geometric mean of the proportion of the alignment covered by the true read and the proportion of the true read covered by the alignment.

While SmartMap does not fully recover the read depth of the Gold Standard at the low end of the QQ plot, the shoulder is nonetheless much less prominent than with the uniread analysis (Fig. 5.4D), indicating considerably greater depth recovery. Consistent with that observation, the uniread analyses and the SmartMap analyses both performed well at highly mappable regions, with read depths approximately at the level of the Gold Standard (Fig. 5.5D). However, at regions of lower mappability, the SmartMap analyses recovered a markedly greater proportion of the read depth than did the uniread analyses (Fig. 5.5D-E). As expected from prior analyses (Fig. 5.4D), the SmartMap analyses with one iteration of reweighting recovered greater read depth than those with no reweighting (Fig. 5.5D-E). Importantly, though they performed similarly at regions of lower mappability, the SmartMap scored analyses recovered greater read depth than their unscored counterparts at regions with moderate mappability (Fig. 5.5D-E).

Genome-wide, SmartMap analyses had lower on-target read depth than the Gold Standard dataset but were still able to recover greater depth at the on-target loci than corresponding uniread analyses (Fig. 5.5F). Similarly, the SmartMap analyses had marginally higher off-target read depth than the Gold Standard and uniread datasets (Fig. 5.5G); however, the increased off-target depth relative to uniread datasets can be explained by the overall lower read depth in the uniread datasets (Fig. 5.6A). Consistent with the notion that improved priors enhance Bayesian predictions, the unscored SmartMap analyses had lower on-target and higher off-target read depth than the cor-

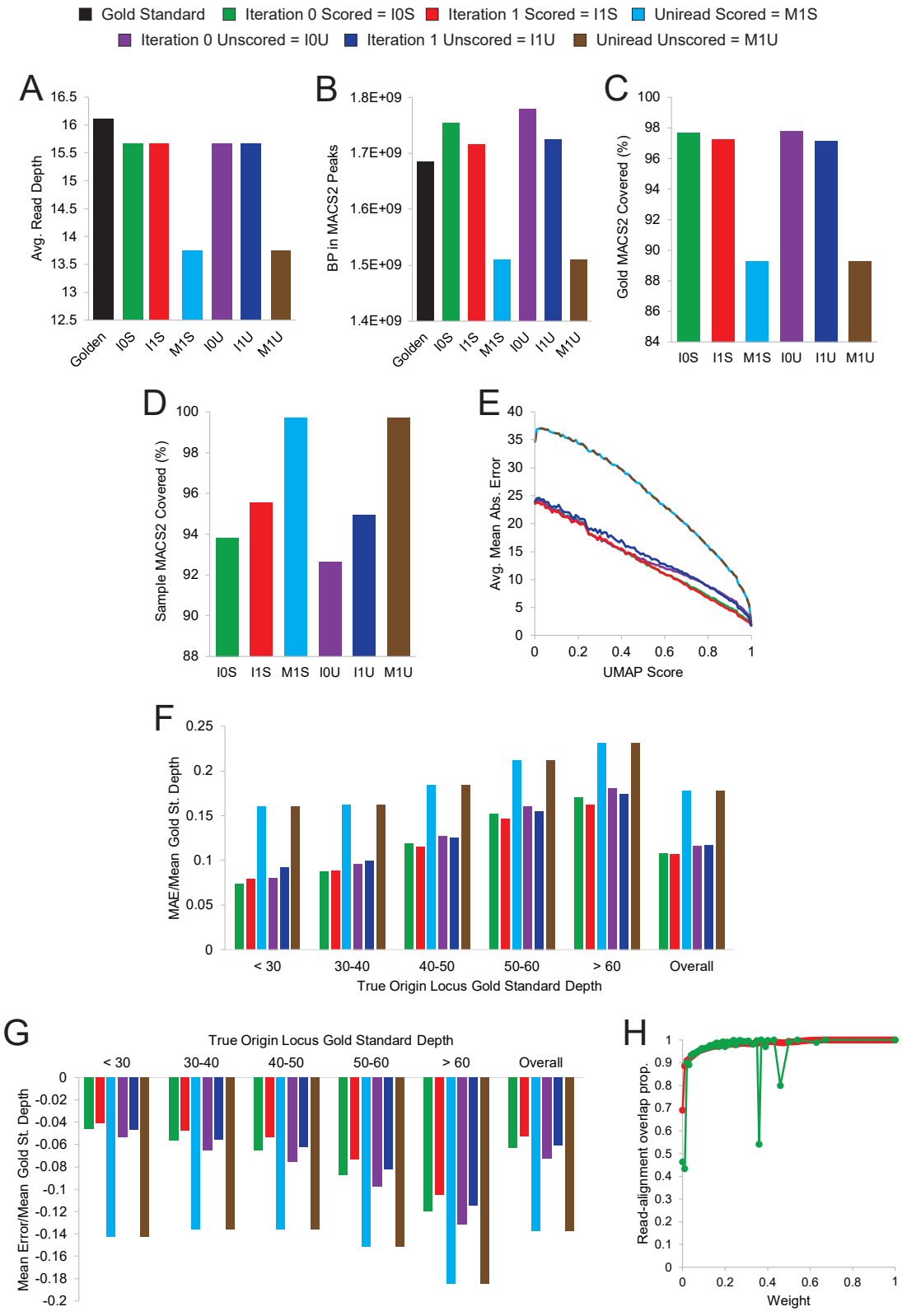


Figure 5.6: Validation and comparison of multiple mapping analysis.

Figure 5.6, continued:
(A) Average read depth of each dataset genome-wide. **(B)** Base pairs covered by MACS2 called peaks for each dataset. **(C)** Percentage of MACS2 peaks in the Gold Standard dataset intersecting with MACS2 peaks in each other analysis, as percentage of base pairs covered. **(D)** Percentage of MACS2 peaks in each analysis intersecting with MACS2 peaks in the Gold Standard dataset, as percentage of base pairs covered. **(E)** Average mean absolute error vs. mappability score (UMAP50) of each dataset. Dashed lines are presented for readability of overlapping curves rather than discontinuities in data. **(F)** Mean absolute error of read depth at true origin loci for each dataset, with Gold Standard as the reference point, stratified by average Gold Standard read depth at true origin locus. **(G)** Mean error of read depth at true origin loci for each dataset, with Gold Standard as the reference point, stratified by average Gold Standard read depth at true origin locus. **(H)** Mean unweighted overlap proportion between alignment and true read origin as a function of alignment weight for the no-iteration (green) and one-iteration (red) scored SmartMap analyses. Overlap proportion is computed as a geometric mean of the proportion of the alignment and of the true read origin that overlaps with the other.

responding scored analyses (Fig. 5.5F-G), and the no-iteration SmartMap analyses had similarly lower on-target and higher off-target read depth than their one-iteration counterparts.

As another metric to evaluate each analysis, we conducted MACS2 peak calling on each dataset and assessed the degree to which they overlap. The SmartMap analyses had similar (albeit slightly higher) base pair coverage with called peaks relative to the Gold Standard dataset and considerably higher coverage on called peaks than the uniread analyses (Fig. 5.3B), consistent with the genome-browser views that suggest a similar pattern of peak boundary sharpening (Fig. 5.3). As a measure of sensitivity, we computed the proportion of the Gold Standard peaks that were covered by SmartMap or uniread peaks (Fig. 5.6C). Conversely, to measure specificity, we computed the proportion of SmartMap or uniread peaks that were covered by Gold Standard peaks (Fig. 5.6D). As expected, there was considerably lower coverage by the uniread datasets than the SmartMap datasets, and the one-iteration SmartMap analyses had very slightly lower coverage over the Gold Standard peaks than the no-iteration analyses (Fig. 5.6C). However, the one-iteration analyses were better-covered by Gold Standard peaks than were their no-iteration counterparts (Fig.

5.6D). Together, these data suggest that SmartMap analyses with one iteration of reweighting have a marked increase in specificity relative to the no-iteration analyses at the expense of a slight decrease in sensitivity.

We then evaluated the overall mean absolute error of read depth at the true origin loci relative to Gold Standard. The uniread analyses had the highest average mean absolute error, with all SmartMap analyses outperforming all uniread analyses (Fig. 5.5H). The scored SmartMap analyses also all had lower error than did the unscored analyses, and the one-iteration SmartMap analyses slightly outperformed the no-iteration analyses (Fig. 5.5H). The error in all datasets tended to primarily be concentrated at regions of lower mappability (Fig. 5.6E). Interestingly, though SmartMap with one iteration had lower mean absolute error overall, the no-iteration modality had slightly lower mean absolute error at true origin loci of lower read depth (Fig. 5.5H and 5.6F). The reason for this difference is not clear; across all read depth classes, the one-iteration analyses had slightly less negative mean error, suggesting that there wasn't a large-scale difference in over- or underweighting after iteration as a function of read depth (Fig. 5.6G). With that said, we feel it is important to contextualize these results; these differences between the no-iteration and one-iteration analyses are small in magnitude and are comparatively dwarfed by the differences between SmartMap and uniread analyses (Fig. 5.5H and 5.6E-G). Accordingly, though there may be small differences between the no-iteration and one-iteration SmartMap analyses, the one-iteration analyses still performed better in aggregate, and both of these scored SmartMap analyses consistently outperformed their unscored or uniread counterparts.

The above analyses all focused on validating SmartMap from the perspective of the total read depth across a set of genomic intervals. However, given that we had a Gold Standard dataset listing the true positions of each read, we also wished to evaluate whether our reweighting method could

improve the estimated of the probability that an alignment was properly mapped – and, by proxy, improve the MAPQ score estimate for each alignment. Without reweighting, the probability of correct alignment ranged from 0-0.67 and 1, with no alignments with correct alignment probability between 0.67 and 1. One iteration of SmartMap reweighting expanded the spectrum of possible alignment weights to the full range of 0-1. Without reweighting, the weight of alignments did not correlate well with the proportion of alignment intersecting the true genomic position, with many large deviations seen from linearity (Fig. 5.5I). By contrast, though one iteration of reweighting still showed some deviations from linearity by this analysis, the weight of alignments more closely concorded with the proportion of the alignments intersecting the true read origin (Fig. 5.5I). This suggests that by this measure, SmartMap reweighting improved the estimates of the probability that the alignment intersects with the true genomic position of the corresponding read.

Similarly, we compared the weighted proportion of overlap between the true read positions and any intersecting alignments as a function of alignment weight. This is meant to represent the proportion of a read's weight that is mapped to the correct location due to a given alignment and incorporates both the confidence of the alignment selection (i.e. the weight) and the overlap of the alignment with the true origin of the read. In both the no-reweighting and one iteration analyses, the overlap proportion score was closely linearly related to the alignment weight, though the reweighted analysis showed a slightly smoother curve with fewer marked deviations from linearity (Fig. 5.5J). This is roughly expected, as the overlap proportion score is itself a function of weight; however, this analysis is comforting insofar as it shows that the SmartMap reweighting does not markedly inflate or deflate the expected weight contribution of a given alignment to a proper intersection with the true origin. Similarly, we find that the unweighted overlap proportion of alignments with the true origin of the read is roughly constant near one for both the no-iteration and one-iteration

datasets, though again, the one-iteration SmartMap analysis reduces the deviations from this level (Fig. 5.6H). These analyses suggested that in addition to improving measurement of read depths in aggregate, the SmartMap reweighting procedure can also improve the estimates of correct alignment for individual reads and alignments.

The biological ChIP-seq and MNase-seq datasets presented in the remainder of this work used 50bp read lengths or shorter, which is why we used 50bp read lengths for our simulated dataset. However, in recent years, 100bp read lengths have become commonplace, and indeed, the ENCODE datasets we present later in this work employed paired-end 100bp NGS. As such, we examined the degree to which SmartMap can improve recovery of sequencing depth with longer reads by conducting a similar analysis as the above with a similarly simulated dataset employing 100bp paired-end reads. For facile comparison to the other Fig. and analyses in this work, we have continued to use the UMAP50 score as our mappability score. This choice is in spite of the fact that UMAP50 measures mappability by 50mers rather than 100mers and will thus underestimate mappability by 100bp reads. Therefore, regions with lower mappability scores will often be more easily mapped than the score would indicate, blunting differences between SmartMap and uniread analyses. As such, our analyses using the UMAP50 score will offer a conservative view at the impact of SmartMap analysis on read depth recovery and error.

Despite this conservative choice of mappability score, we still see that SmartMap analysis improves sequencing depth recovery nearly as well with 100bp reads as it does with 50bp reads. The simulated dataset with 50bp reads shows a 13.9% increase in analyzable reads due to the high number of multireads (Fig. 5.2B-C and Table 5.1); the simulation with 100bp reads shows a 13.6% increase in analyzable reads and a similar proportion of multireads (Fig. 5.7A-B and Table 5.1). Along the same lines, the two simulations increase read depth over similar proportions of the genome

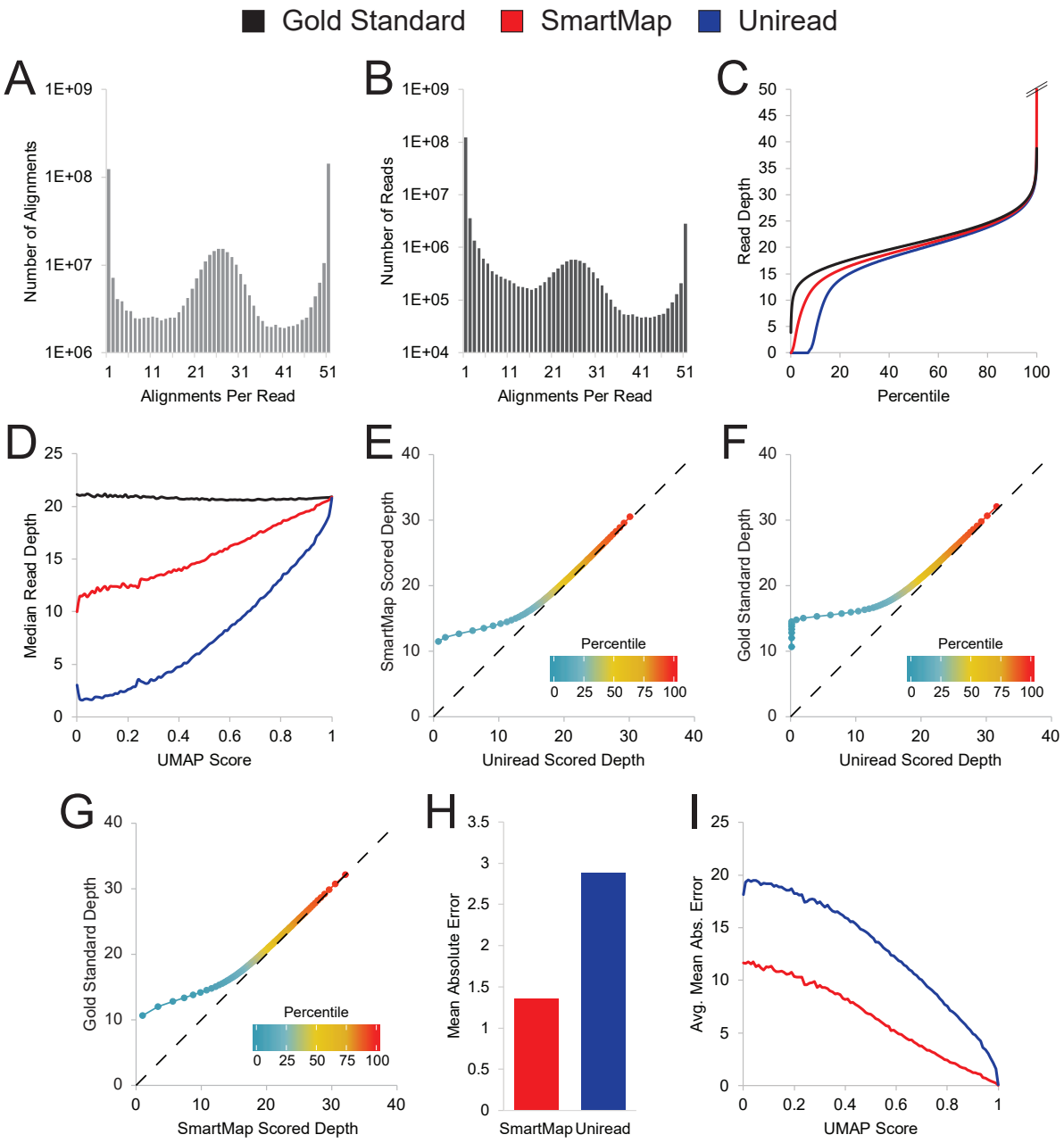


Figure 5.7: SmartMap and uniread analyses of the 100bp read length validation dataset.
(A, B) Number of (A) alignments or (B) reads vs. number of alignments per read. **(C)** Quantile plot of read depth at the true origin loci. **(D)** Median read depth vs. mappability score (UMAP50) of the true origin loci. **(E-G)** QQ plot of read depth at true origin loci in the (E) SmartMap vs. uniread, (F) Gold Standard vs. uniread, and (G) Gold Standard vs. SmartMap scored datasets. Color scale represents percentile of each point, from 1st to 99th percentiles. **(H)** Mean absolute error of read depth at true origin loci for each dataset, with Gold Standard as the reference point. **(I)** Average mean absolute error vs. mappability score (UMAP50) of each dataset.

Table 5.2: Analysis of reads across genomic windows.

	Sample	Genome	Assay	Regions	Regions with reads in: Uniread	Regions with reads in: SmartMap	% Reg. Inc.	% Inc. Reg.
Simulated	Simulated, 50bp	hg38	Simulation	15,498,848	10,486,482	11,994,872	28.74%	14.38%
	Simulated, -k 101	hg38	Simulation	15,498,848	10,463,337	12,012,046	29.07%	14.80%
	Simulated, 100bp	hg38	Simulation	15,498,848	9,956,521	11,475,027	32.77%	15.25%
AR7	Input Rep. 1	mm10*	MNAse-seq	13,654,309	12,129,867	13,243,873	33.49%	9.18%
	H3K4me3 Rep. 1	mm10*	ChIP-seq	13,654,309	11,329,858	12,999,672	27.21%	14.74%
	Input Rep. 2	mm10*	ChIP-seq	13,654,309	12,115,174	13,242,243	31.96%	9.30%
	H3K4me3 Rep. 2	mm10*	ChIP-seq	13,654,309	10,952,182	12,750,113	27.25%	16.42%
AR8	Input	dm3†	MNAse-seq	698,569	617,424	681,457	17.92%	10.37%
	H3K27me3	dm3†	ChIP-seq	698,569	612,050	680,193	17.39%	11.13%
AR9	Input	mm10†	MNAse-seq	13,654,309	12,214,070	13,245,567	35.60%	8.45%
	H3K4me3	mm10†	ChIP-seq	13,654,309	11,775,058	13,208,421	30.83%	12.17%
	H3K9me3	mm10†	ChIP-seq	13,654,309	12,027,438	13,245,567	32.11%	10.04%
	H3K27me3	mm10†	ChIP-seq	13,654,309	12,012,091	13,237,339	31.99%	10.20%
AR16	Input	hg38‡	MNAse-seq	15,498,848	13,879,635	14,629,457	34.59%	5.40%
	H3K4me1	hg38‡	ChIP-seq	15,498,848	13,310,801	14,423,602	31.07%	8.36%
	H3K4me2	hg38‡	ChIP-seq	15,498,848	13,298,178	14,443,778	30.56%	8.61%
	H3K4me3	hg38‡	ChIP-seq	15,498,848	10,338,102	12,270,858	25.24%	18.70%

Table 5.2 continues on next page.

Table 5.2, continued:

	Sample	Genome	Assay	Regions	Regions with reads in: Uniread	Regions with reads in: SmartMap	% Reg. Inc.	% Inc. Reg.
AR17	Input	hg38‡	MNAse-seq	15,498,848	13,896,029	14,634,051	34.56%	5.31%
	H3K9me3	hg38‡	ChIP-seq	15,498,848	13,856,547	14,626,552	34.14%	5.56%
	H3K27me3	hg38‡	ChIP-seq	15,498,848	13,803,814	14,618,351	33.66%	5.90%
ENCODE	Snyder Rep. 1	hg38	ATAC-seq	15,498,848	10,389,635	11,970,867	28.34%	15.22%
	Snyder Rep. 2	hg38	ATAC-seq	15,498,848	9,772,547	11,251,766	21.53%	15.14%
	Gingeras Rep. 1	hg38§	RNA-seq	41,929	21,755	25,711	22.85%	18.18%
	Gingeras Rep. 2	hg38§	RNA-seq	41,929	12,399	14,485	11.96%	16.82%

Unireads refers to the number of reads with one alignment.

For all except the "Simulated, -k 101" dataset, Analyzable Multireads refers to reads with between 2-50 alignments; Unanalyzable Multireads refers to reads with 51 reported alignments, the limit for reported alignments per read.

For the "Simulated, -k 101" dataset, Analyzable Multireads refers to reads with 2-100 alignments, and Unanalyzable Multireads refers to reads with 101 reported alignments.

% Incr.: Increase in the number of analyzable reads with SmartMap analysis, computed as the number of Analyzable Multireads as a percentage of the number of Unireads.

Genome includes ICeChIP barcodes: * Series 1. † Series 2. ‡ Series 3.

§ Genome includes ENCODE ERCC standards.

(Table 5.2). Over the true origin loci, much like the 50bp simulation, the 100bp simulated dataset shows an increase in read depth on quantile plots (Fig. 5.7C) under SmartMap analysis, with this increase in read depth primarily occurring at regions of low UMAP50 mappability score (Fig. 5.7D), conservative though this measurement of mappability is. Much like the 50bp simulated datasets, the increases in read depth under SmartMap analysis are primarily seen at regions of low mappability

and low uniread read depths; QQ plots comparing the uniread analysis with the SmartMap or Gold Standard show a shoulder at low uniread depths, with the plot converging onto a slope of unity at higher uniread depths (Fig. 5.7E-F). It should be noted that, as with the 50bp simulated dataset (Fig. 5.7D), the SmartMap dataset still fails to fully recover read depth as compared to Gold Standard with 100bp reads (Fig. 5.7G). Nonetheless, the SmartMap analysis still shows considerably lower mean absolute error than does the uniread analysis at true origin loci (Fig. 5.7H), with this decrease in error being particularly prominent at regions with lower UMAP50 mappability scores (Fig. 5.7I). In total, these analyses suggest that even for datasets employing 100bp paired-end sequencing reads, multiread analysis still has nearly undiminished importance and that SmartMap can still markedly improve read depth recovery while decreasing overall error.

The above analyses all restricted Bowtie2 to report a maximum of 51 alignments for computational efficiency. Subsequently, only those reads aligning to fewer than 51 alignments were used for SmartMap analysis. However, this practice excluded more than 7 million reads (Table 5.1), likely including reads that map to the most highly repetitive regions of the genome. Notably, this is a restriction on alignment itself, not SmartMap; there's no reason that SmartMap would inherently be unable to handle greater numbers of alignment. Nonetheless, to evaluate the impact of this restriction on the SmartMap datasets, we reanalyzed our simulated 50bp read length dataset with a maximum of 101 alignments per read (hereafter, the k101 dataset) and compared it to the previous analysis (the k51 dataset). To our surprise, the two analyses were highly similar despite the near-doubling in the maximum-alignments threshold in the former dataset. The increase in the number of analyzable reads was nearly identical between the two analyses (Fig. 5.8A-B and Table 5.1), with similar increases in depth over genomic windows (Table 5.2). At the true origin loci, the SmartMap read depths in both the k51 and k101 datasets were very similar at the level of read depth

(Fig. 5.8C-D). Mean absolute error relative to the Gold Standard was actually very slightly lower in the k51 dataset, though they were quite similar in magnitude compared to the uniread dataset (Fig. 5.8E-F). Even specifically examining repetitive elements, read depth was very similar between the k51 and k101 SmartMap analyses at all repeats (Fig. 5.8G), LINEs (Fig. 5.8H), SINEs (Fig. 5.8I), and Alu elements (Fig. 5.8J), closely approximating the Gold Standard read depth in both cases. Accordingly, though there is still a large proportion of reads that mapped to still greater numbers of loci, we find that at the range we have tested, the SmartMap analyses are robust to differences in maximum-alignments reporting thresholds and that there is little practical difference between restricting datasets to a maximum of 51 or 101 alignments per read besides the additional time and storage space needed for the latter.

To be sure, the reweighting used for SmartMap is not without concerns. In particular, one of the potential problems for SmartMap is the existence of high-signal regions, which can show falsely high read depth in NGS experiments due to sequencing or alignment error[233]. If there are regions of falsely high weight, then those regions could be skewed by the SmartMap reweighting algorithm to report even greater weights, thus exacerbating these artifactually high signals. To assess the degree to which these regions represent an issue for SmartMap, we computed the number of genomic windows with more than 60, 70, 80, or 90 reads in our simulated datasets (Table 5.3). We used these benchmarks as rough thresholds for defining high-signal regions because the Gold Standard dataset had a maximum read depth across a genomic window of approximately 83 reads. Notably, the Gold Standard did not require sequencing or mapping and should thus not be susceptible to these high-signal artifacts. Unfortunately, one iteration of SmartMap reweighting did increase the proportion of high-signal regions considerably; there were fewer than 600 genomic windows with an average depth of more than 70 in the Gold Standard, Iteration 0, and Uniread datasets, compared

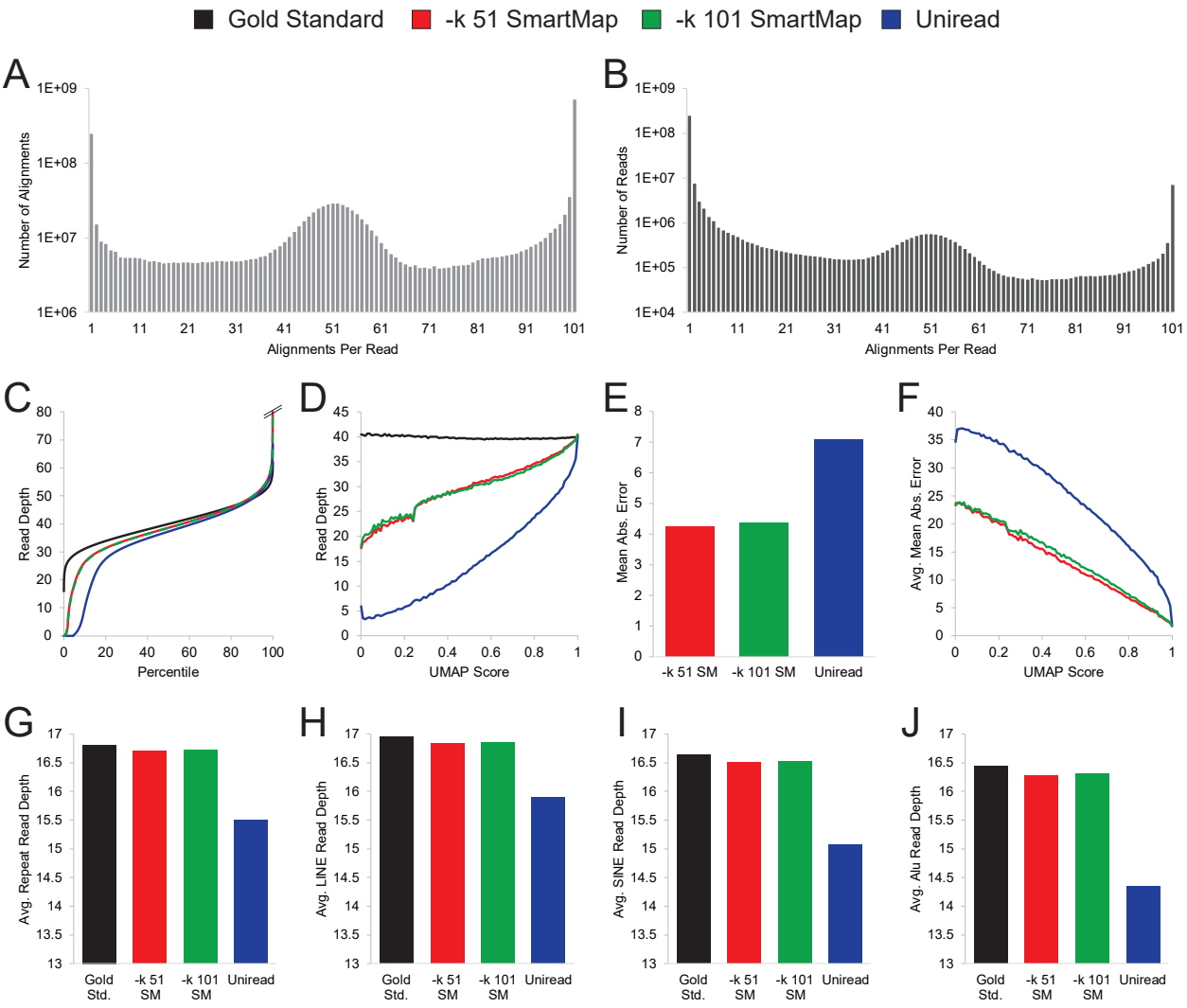


Figure 5.8: Characteristics of the -k 101 SmartMap dataset.

(A, B) Number of (A) alignments or (B) reads vs. number of alignments per read. **(C)** Quantile plot of read depth at the true origin loci. Dashed lines are presented for readability of overlapping curves rather than discontinuities in data. **(D)** Median read depth vs. mappability score (UMAP50) of the true origin loci. **(E)** Mean absolute error of read depth at true origin loci for each dataset, with Gold Standard as the reference point. **(F)** Average mean absolute error vs. mappability score (UMAP50) of each dataset. **(G-J)** Average read depth across the bodies of (G) all repetitive elements, (H) LINEs, (I) SINEs, and (J) Alu elements.

to more than 10,000 in the SmartMap dataset with one reweighting cycle. It's important to note that these regions represent a very small proportion of the genome; only 0.066% of the genomic windows had more than 70 reads on average, and even fewer had more than 80 or 90 reads (Table

5.3), leaving considerably more than 99.9% of the genome as not having abnormally high-signal attributable to SmartMap. In contrast, almost 15% of the genome is hidden from uniread analysis (Table 5.2). Nonetheless, we feel it is fair to say that the reweighting algorithm used for SmartMap will increase the weights of multiread alignments at high signal regions, which can exacerbate artifactually high read depths.

Table 5.3: Analysis of high-depth regions under SmartMap analysis.

	Number of genomic windows with:				Percent of genomic windows with:			
Dataset	>60 rds.	>70 rds.	>80 rds.	>90 rds.	>60 rds.	>70 rds.	>80 rds.	>90 rds.
Gold Std.	34,468	463	1	0	0.22	0.0030	6.5×10^{-6}	0
Iteration 0	26,969	571	85	36	0.17	0.0037	5.5×10^{-4}	2.3×10^{-4}
Iteration 1	44,185	10,193	6,337	4,296	0.29	0.066	0.041	0.028
Uniread	24,344	321	1	0	0.16	0.0021	6.5×10^{-6}	0

Number of genomic windows refers to the number of 200bp genomic windows for each dataset with an average depth or average weight greater than that indicated in each column. Percent of genomic windows refers to the number of genomic windows as a percentage of the total number of 200bp genomic windows in hg38 (15,498,848). The median read depth was 10.5 and the mean read depth was 16.1 in the Gold Standard dataset.

Even so, on the whole, these analyses suggest that SmartMap recovers read depth at a large set of loci that would otherwise be missed by the uniread analyses and that of the SmartMap analyses, one iteration of reweighting with use of alignment scores largely outperforms the other modalities. Accordingly, for the remainder of this work, we use SmartMap analysis with one iteration in scored mode as our default SmartMap method.

Using SmartMap on MNase-seq and ChIP-seq datasets

Having validated our method on the simulated dataset, we turned to the biological samples. We deployed a total of 21 datasets derived from three different organisms for our analysis (Table 5.1). Of these datasets, six were control ICeChIP Inputs, generated by MNase-seq[118,124], 11 were ICeChIP-

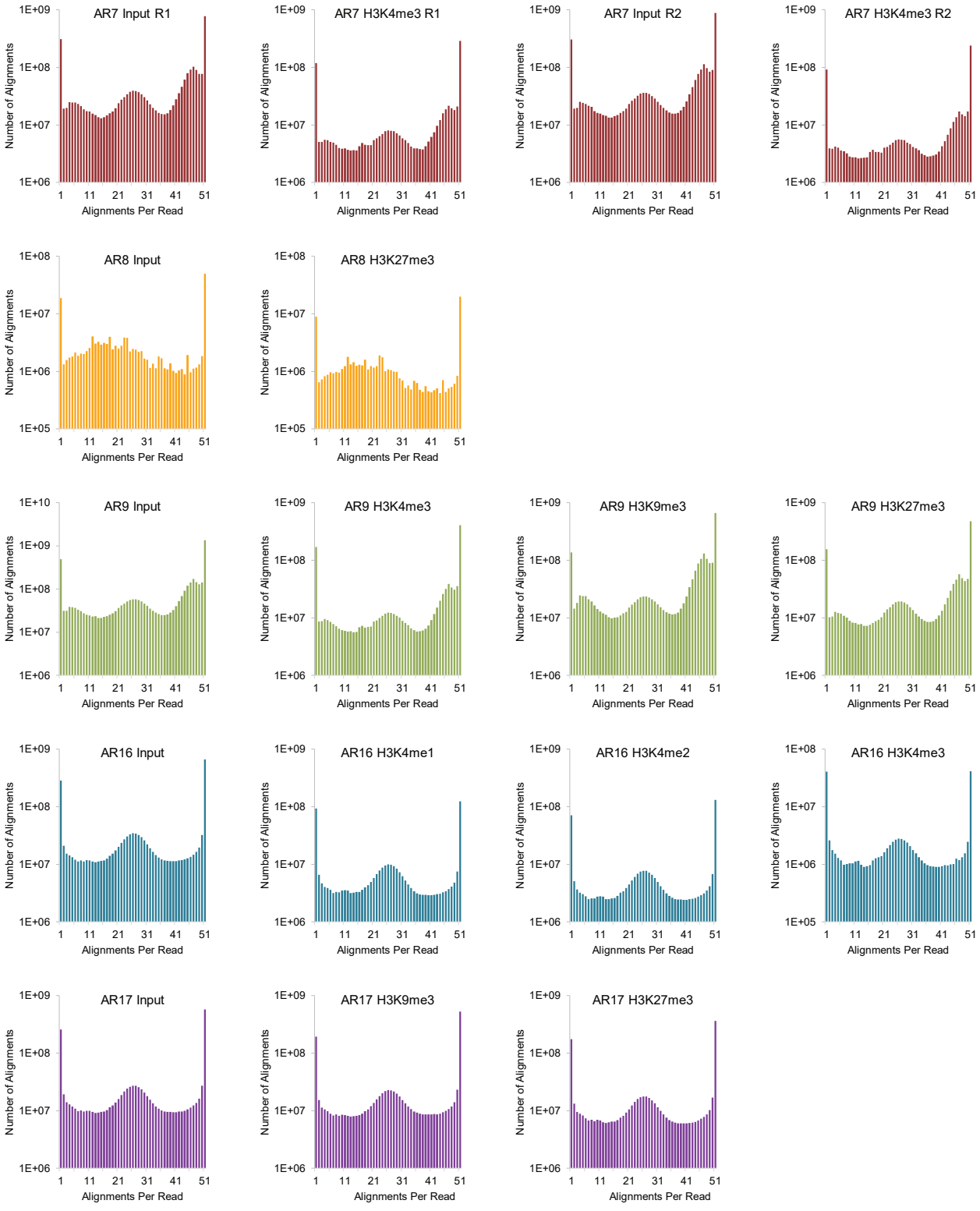


Figure 5.9: Alignments per ICeChIP-seq dataset.
Number of alignments vs. alignments per read for each ICeChIP-seq dataset analyzed.

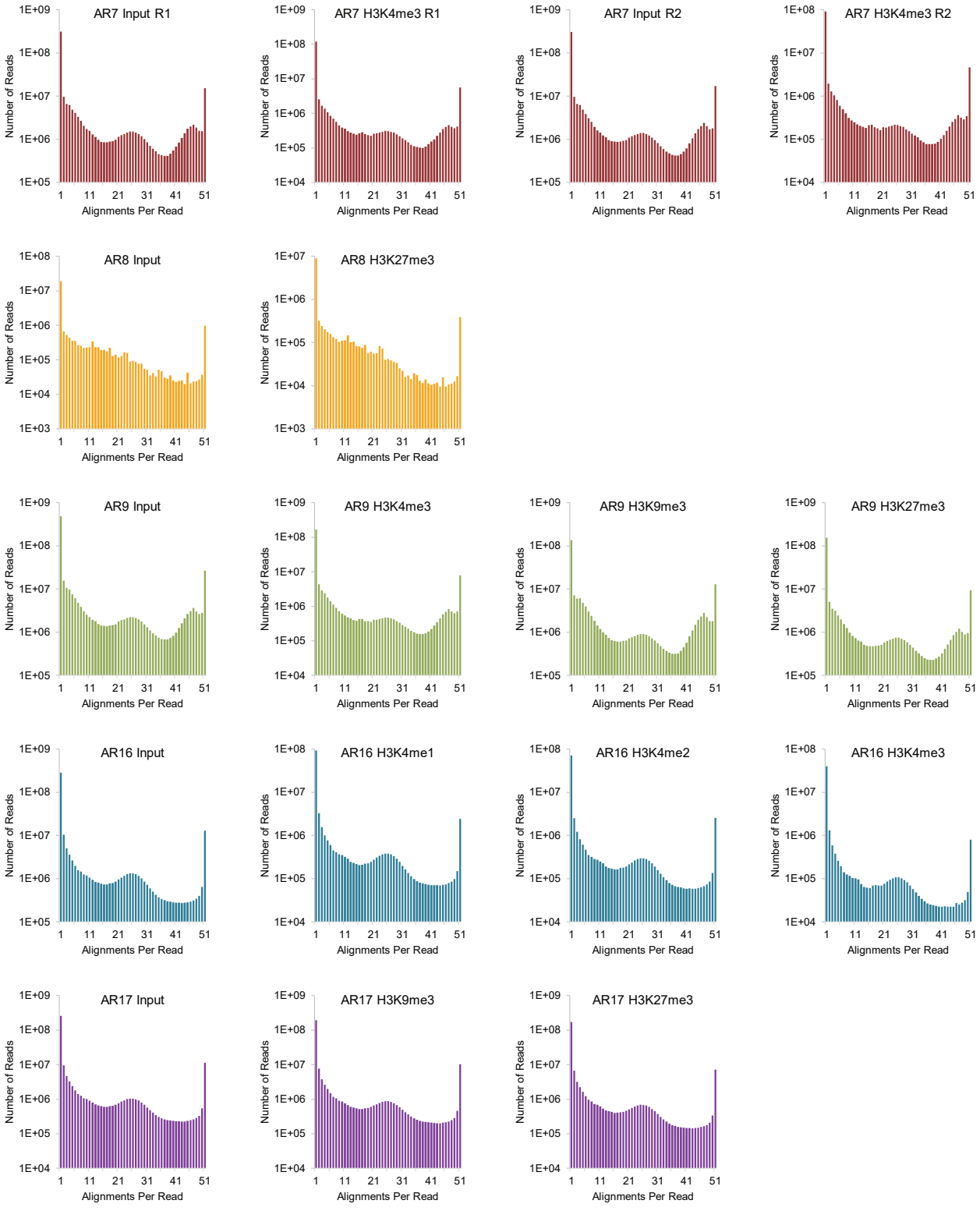


Figure 5.10: Reads per ICeChIP-seq dataset.

Number of reads vs. alignments per read for each ICeChP-seq dataset analyzed.

seq IP datasets, two were ATAC-seq datasets, and two were RNA-seq datasets. After alignment, the samples showed a 13-50% increase in the number of usable reads for SmartMap analysis relative to uniread (Fig. 5.9, 5.10 and Table 5.1).

To evaluate the impact of our algorithm on the ICeChIP-seq datasets, we first conducted SmartMap and uniread analysis on each of the Input datasets and computed the average read depth on 200bp genomic windows. As with the simulated dataset, the SmartMap analyses of the Inputs had increased read depth relative to the uniread datasets (Fig. 5.11A and 5.12A), with markedly greater depth in the SmartMap analysis at windows of lower mappability (Fig. 5.11B and 5.12B). Similarly, this excess read depth was not distributed across all reads, but rather, was concentrated onto 17-35% of windows (Fig. 5.11C and 5.12C and Table 5.2), primarily at regions of lower mappability (Fig. 5.11D and 5.12D). The QQ plots of the SmartMap vs. the uniread read depths showed a shoulder at low uniread depth (Fig. 5.11E and 5.12E), again suggesting that the increase in read depth from the SmartMap analysis is primarily at loci where the uniread analysis performs poorly. This difference in the distributions of read depths further comments on the importance of analyzing multireads.

With our Input datasets, we could also examine the reproducibility of the MNase-seq experiments under uniread and SmartMap analyses. There were three biological replicates of Input in mESC E14 cells (AR7 Replicate 1, AR7 Replicate 2, and AR9), and two biological replicates of Input in K562 cells (AR16 and AR17). For all loci with nonzero read depth, we computed the depth normalized log ratios of reads in a pairwise manner for biological replicates, shown as quantile plots in Fig. 5.11F. These plots are highly similar under SmartMap and uniread analyses across all pairwise comparisons (Fig. 5.11F). Accordingly, the average magnitudes of these ratios are similar between the two analyses – and indeed, are slightly lower in the SmartMap datasets (Fig.

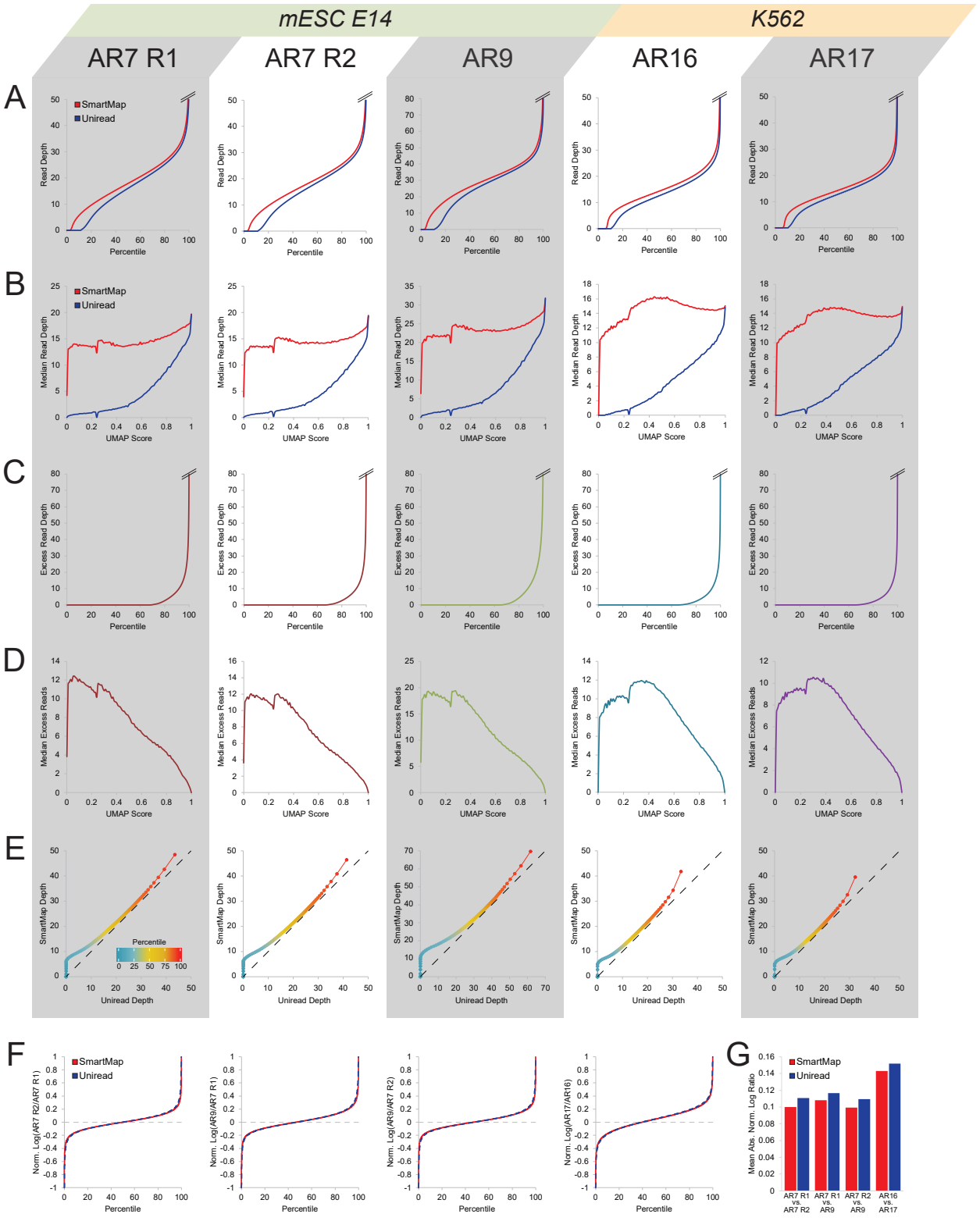


Figure 5.11: SmartMap and uniread analyses of ICeChIP-seq input depth.

Figure 5.11, continued:
All analyses conducted on 200bp genomic windows for the Inputs defined in Table 1. **(A)** Quantile plot of read depth for SmartMap and uniread analyses. **(B)** Median read depth vs. mappability score (UMAP50) for SmartMap and uniread analyses. **(C)** Quantile plot of excess read depth in SmartMap relative to uniread analysis. **(D)** Median excess read depth vs. mappability score (UMAP50). **(E)** QQ plot of read depth in SmartMap vs. uniread analysis. Color scale represents percentile of each point, from 1st to 99th percentiles. Dashed line represents line with slope of unity. **(F)** Quantile plots of depth-normalized log ratio of read depths of biological input replicates under SmartMap and uniread analysis. Graph breaks are present on both the upper and lower ends of the graphs. **(G)** Mean absolute depth-normalized log ratio for the comparisons presented in panel F.

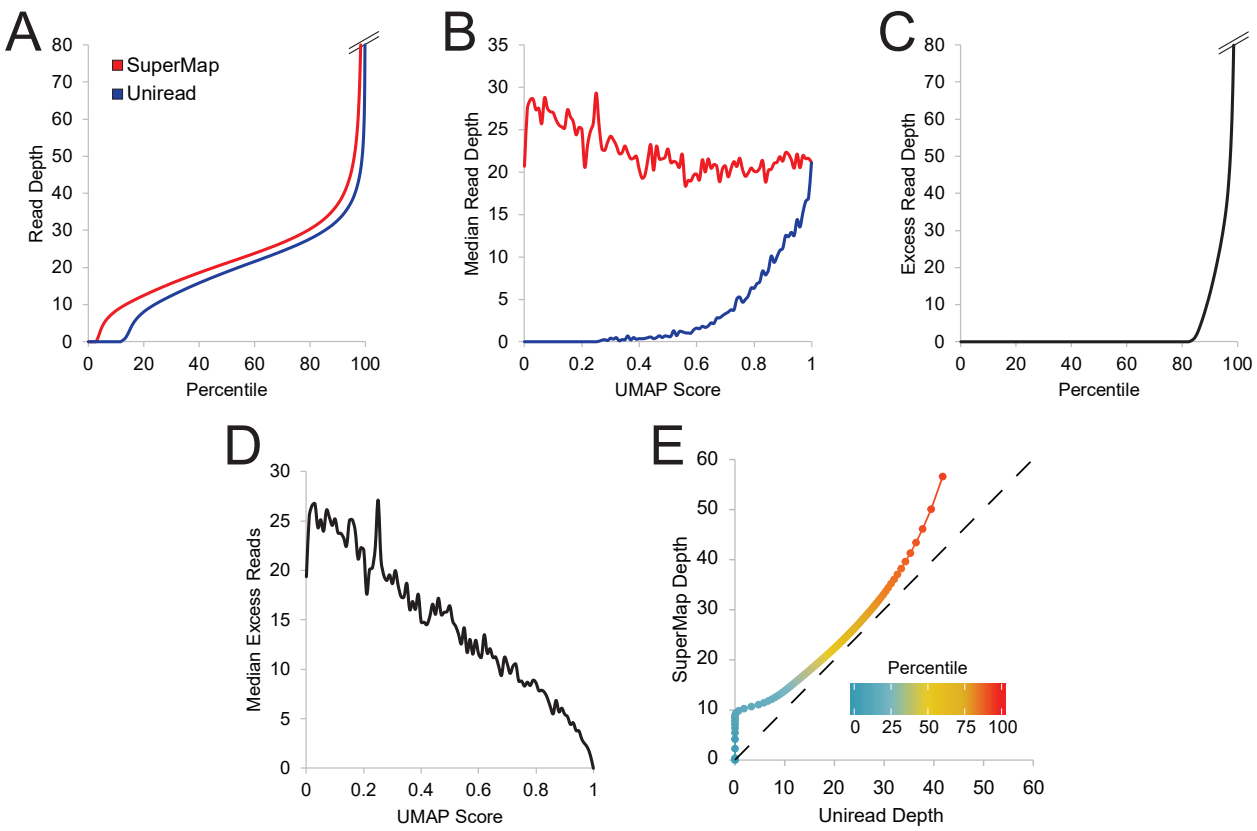


Figure 5.12: SmartMap and uniread analysis of AR8 input.
All analyses conducted on 200bp tiled genomic windows. **(A)** Quantile plot of read depth for SmartMap and uniread analyses. **(B)** Median read depth vs. mappability score (UMAP50) for SmartMap and uniread analyses. **(C)** Quantile plot of excess read depth in SmartMap relative to uniread analysis. **(D)** Median excess read depth vs. mappability score (UMAP50). **(E)** QQ plot of read depth in SmartMap vs. uniread analysis. Color scale represents percentile of each point, from 1st to 99th percentiles. Dashed line represents line with slope of unity.

5.11G). This suggests that the two modalities show highly similar estimates of reproducibility of data between biological replicates of MNase-seq.

Having examined the Input datasets, we then used the ICeChIP-seq datasets to compute histone modification densities (HMD) across 200bp genomic windows with uniread and SmartMap analyses. Interestingly, we noted that the mean HMD was quite similar between the SmartMap and uniread datasets across a broad range of mappability scores (Fig. 5.13A and 5.14A). However, the median HMD of those same datasets were divergent, with the SmartMap analyses having considerably higher median HMD across bins of low mappability than the uniread analyses (Fig. 5.13B and 5.14B). The difference between mean and median HMD may be attributable to the fact that HMD is a scaled-version of fold-change of IP over Input. We attribute the median HMD divergence to sparser distribution of read depth in the uniread dataset at lower mappability scores (Fig. 5.11B). As such, there are fewer regions with nonzero read depth in both the IP and Input. The result of this mismatch in read distribution is that more regions have an apparent HMD of zero under uniread analysis. That the mean HMDs are similar between the two analyses suggests that the ratios of the total read depths in IP over Input are similar between SmartMap and uniread analyses. Together, these data suggest that the SmartMap analyses preserve the overall HMD across a wide range of mappability scores while also enabling measurement of HMD at a broader range of loci than do uniread analyses.

One of the major benefits of using ICeChIP-seq data is the ability to measure antibody specificity[118,124,171]. In ICeChIP, internal standards bearing a variety of different histone modifications can be simultaneously spiked into an experiment, and the relative pulldown efficiency of each modification can be quantified as a proportion of the target to measure the off-target binding of the antibody. We wished to determine whether the SmartMap analyses would yield similar specificity

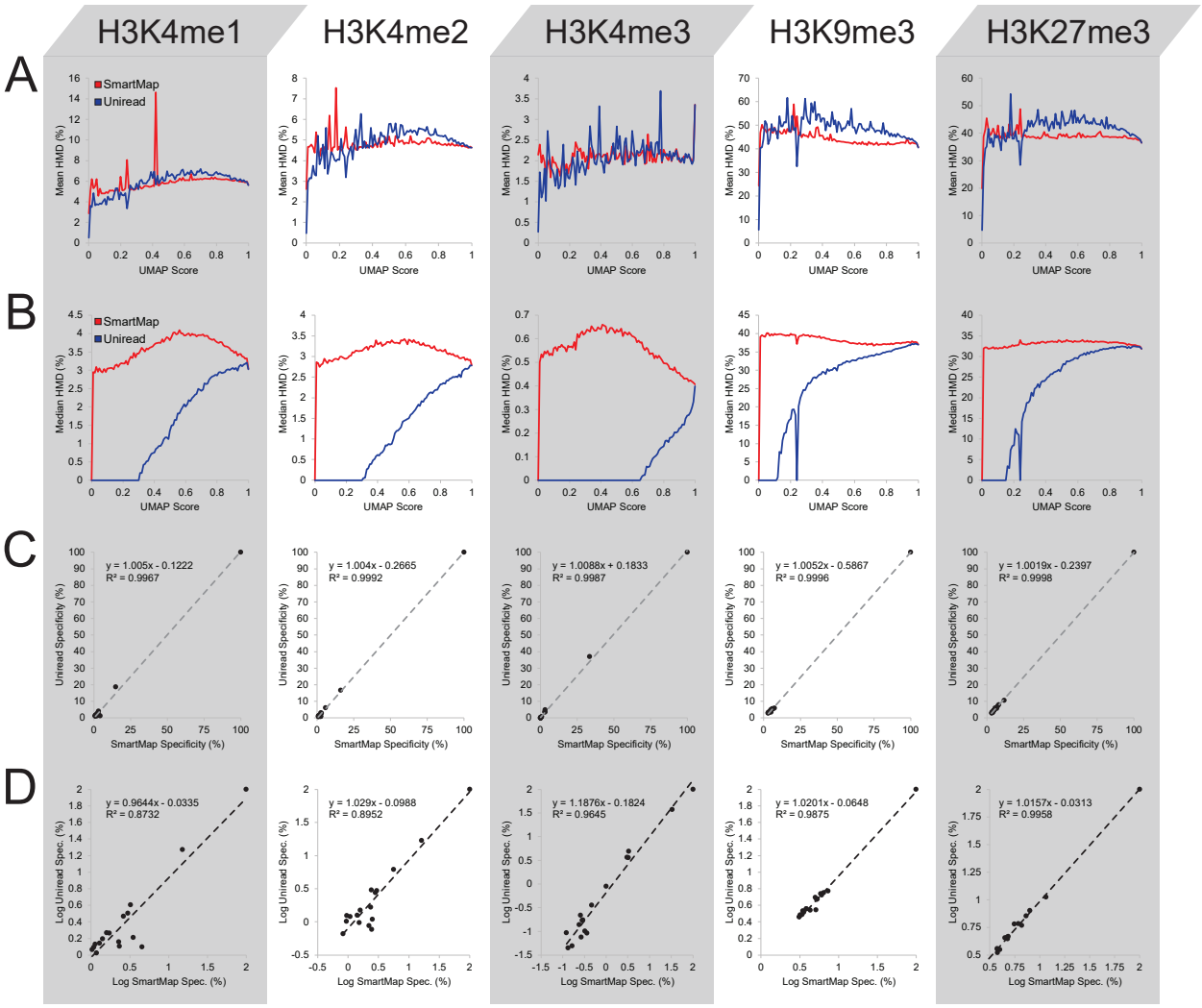


Figure 5.13: ICeChIP-seq histone modification density in SmartMap and uniread analyses. All analyses conducted on 200bp tiled genomic windows. **(A-B)** (A) Mean or (B) Median HMD vs. mappability score (UMAP50) for SmartMap and uniread analyses. **(C-D)** Scatterplots of (C) specificity or (D) log specificity for uniread vs. SmartMap analyses. Specificity is measured as the enrichment of each internal standard nucleosome as a percentage of on-target enrichment.

estimates as did the uniread analyses. First, we found that the ratio of the reads from the on-target nucleosome in the IP over the Input was highly similar between the uniread and SmartMap analyses (Table 5.4). Moreover, the scatterplots of specificity (Fig. 5.13C and 5.15A) and log specificity (Fig. 5.13D and 5.15B) under each modality show slopes close to unity and high coefficients of

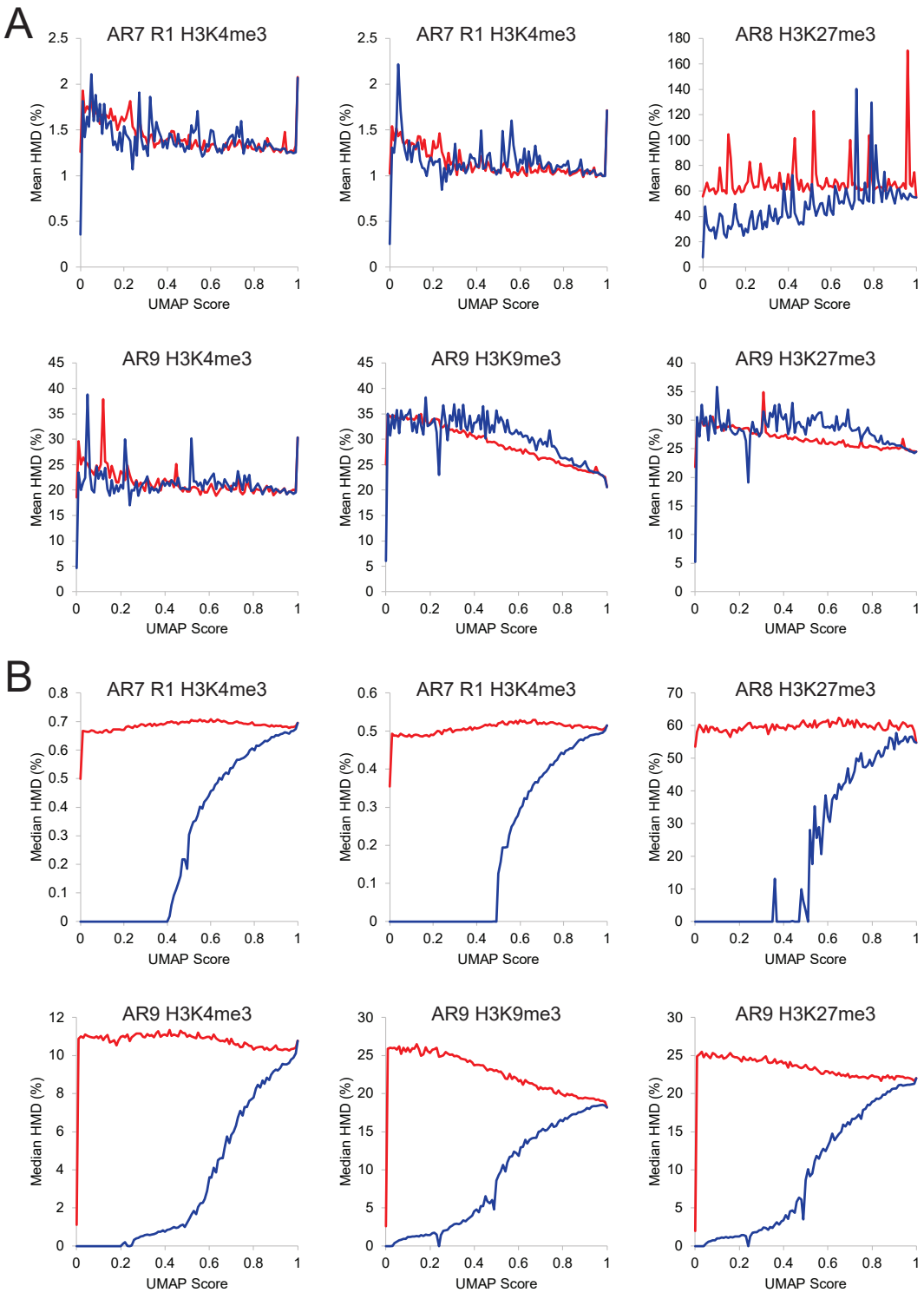


Figure 5.14: SmartMap and uniread analyses of AR7, AR8, and AR9 HMDs.

(A) Mean or **(B)** Median HMD vs. mappability score (UMAP50) for SmartMap and uniread analyses. Red line represents SmartMap analysis; blue line represents uniread analysis.

determination. This further shows that the specificity measurements in SmartMap and uniread analyses are absolutely (Fig. 5.13C and 5.15A) and relatively (Fig. 5.13D and 5.15B) similar.

Table 5.4: Analysis of ICeChIP calibrant barcodes.

				On-target IP/Input Ratio:			Specificity Plot:	
	Sample	Series	Barcodes	Uniread	SmartMap	Species	Slope	R^{-2}
AR7	H3K4me3 Rep. 1	Ser. 1	11	19.88	20.05	1	–	–
	H3K4me3 Rep. 2	Ser. 1	11	18.95	18.99	1	–	–
AR8	H3K27me3	Ser. 2	100	0.877	0.879	1	–	–
AR9	H3K4me3	Ser. 2	100	27.7	28.3	7	1.051	0.9984
	H3K9me3	Ser. 2	100	1.34	1.26	7	1.012	0.9972
	H3K27me3	Ser. 2	100	0.678	0.677	7	1.022	0.9995
AR16	H3K4me1	Ser. 3	136	4.34	4.84	17	1.005	0.9967
	H3K4me2	Ser. 3	136	3.98	3.75	17	1.004	0.9992
	H3K4me3	Ser. 3	136	32.4	31.1	17	1.009	0.9987
AR17	H3K9me3	Ser. 3	136	2.45	2.23	17	1.005	0.9996
	H3K27me3	Ser. 3	136	1.82	1.73	17	1.002	0.9998

Barcodes: the number of unique DNA barcode sequences in the ICeChIP calibrant series.

Species: the number of distinct modified nucleosomes marked by the barcodes, including the target modification and, if there is more than one species, the off-target modifications.

Specificity plot: summary of the specificity plots shown in Fig. 5.13C and 5.15A.

Extending the utility of SmartMap to ATAC-seq and RNA-seq

We also found that SmartMap could be applied to ATAC-seq data to obtain more global measurements of chromatin accessibility. To demonstrate this, we used two replicates of K562 ATAC-seq data, originally generated by the Snyder Lab as part of the ENCODE Consortium[127]. As with the ICeChIP-seq datasets, we found that SmartMap analysis could utilize 17-20% more reads than uniread analysis (Table 5.1); this increased read depth was primarily concentrated at 20-30% of

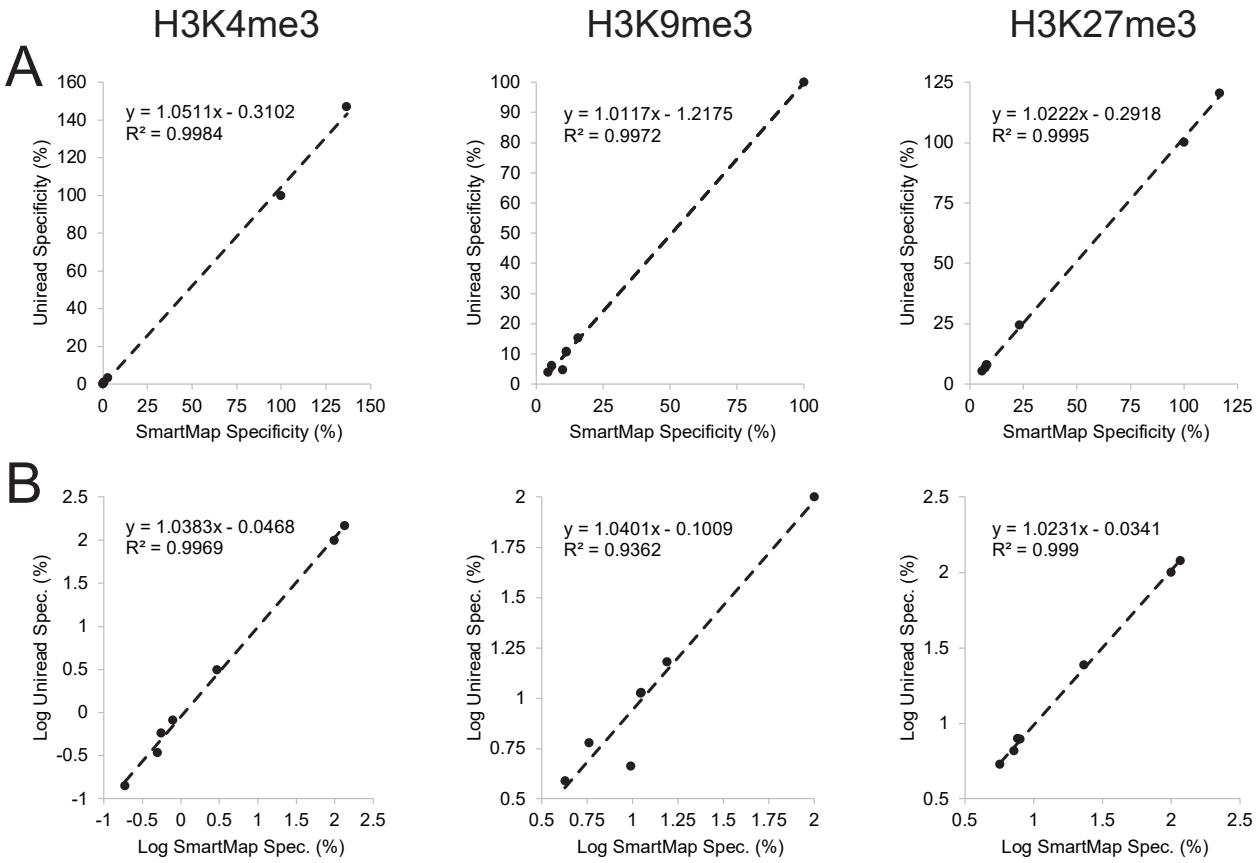


Figure 5.15: Specificity scatterplots for AR9.
Scatterplots of (A) specificity or (B) log specificity for uniread vs. SmartMap analyses. Targets of pulldowns are H3K4me3 (left), H3K9me3 (centre), and H3K27me3 (right). Specificity is measured as the enrichment of each internal standard nucleosome as a percentage of on-target enrichment.

the genome (Fig. 5.16A-C and Table 5.2), particularly those loci with low mappability scores (Fig. 5.16D-F). SmartMap and uniread analyses also showed similar levels of reproducibility between the two isogenic replicates, though SmartMap showed slightly lower reproducibility between the two datasets than did the uniread analysis (Fig. 5.16G-H). These data suggest that SmartMap is also useful for ATAC-seq data and can reveal accessible regions of the genome at poorly mappable loci that would have been missed by uniread analysis alone.

In addition to the MNase-seq, ChIP-seq, and ATAC-seq datasets, we also sought to apply our SmartMap analysis to RNA-seq experiments. Specifically, we analyzed two replicates of K562 bulk

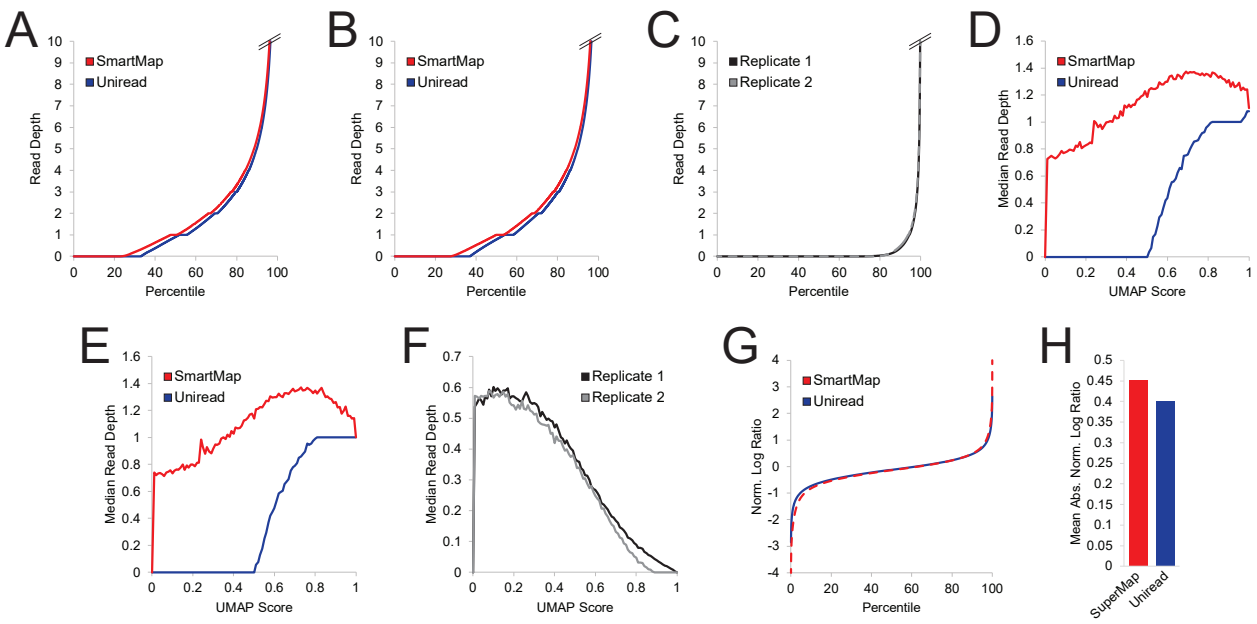


Figure 5.16: SmartMap analysis of ENCODE ATAC-seq datasets.

(A-B) Quantile plot of read depth at genomic windows in SmartMap and uniread analyses for (A) Replicate 1 or (B) Replicate 2. **(C)** Quantile plot of excess read depth in SmartMap datasets relative to corresponding uniread dataset for Replicates 1 and 2. **(D-E)** Median read depth vs. mappability score (UMAP50) in SmartMap and uniread analyses for (D) Replicate 1 or (E) Replicate 2. **(F)** Median excess read depth vs. mappability score (UMAP50). **(G)** Quantile plot of depth-normalized log ratio of read depth in Replicate 1 over Replicate 2, for SmartMap and uniread analyses. Graph breaks are present at both ends of the graph. **(H)** Mean absolute depth-normalized log ratio of the analyses shown in panel G.

RNA-seq data, originally generated by the Gingeras Lab as part of the ENCODE Consortium[127]. Our SmartMap RNA-seq analyses showed that for each replicate, relative to uniread analysis, there was a 20-35% increase in usable reads (Table 5.1) concentrated into a minority of distinct Refseq genes (Fig. 5.17A-C and Table 5.2). The reproducibility of the two datasets was also similar between the SmartMap and uniread analyses, though as with the ATAC-seq data, the SmartMap analysis showed marginally lower reproducibility between the two RNA-seq experiments than did the uniread analysis (Fig. 5.17D-E). With that said, these differences in read depth are relatively minor in magnitude, especially when normalized to differences in read depth in the SmartMap and uniread analyses. Given the other concerns with using this multiread allocation algorithm in gapped

reads or spliced transcripts (as noted in the Discussion), it is likely that SmartMap is not optimally configured for use in RNA-seq analysis.

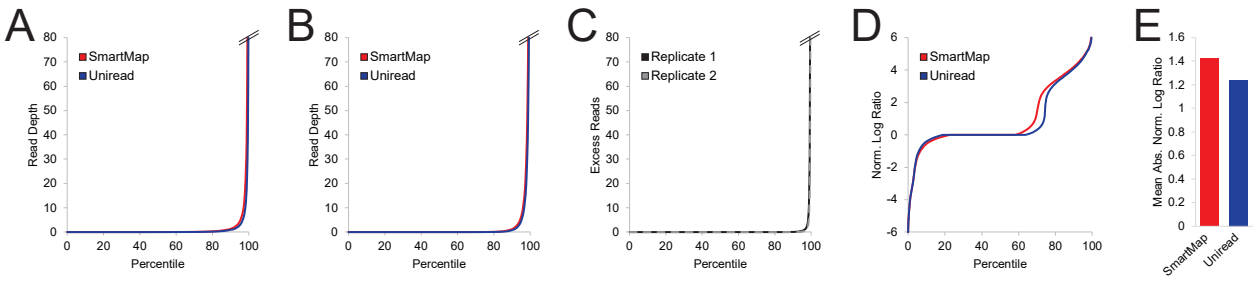


Figure 5.17: SmartMap analysis of ENCODE RNA-seq datasets.
(A-B) Quantile plot of read depth at distinct Refseq genes in SmartMap and uniread analyses for (A) Replicate 1 or (B) Replicate 2. **(C)** Quantile plot of excess read depth in SmartMap datasets relative to corresponding uniread dataset for Replicates 1 and 2. **(D)** Quantile plot of depth-normalized distinct Refseq gene log ratio of read depth in Replicate 1 over Replicate 2, for SmartMap and uniread analyses. Pseudocount of 10^{-7} was added to each gene due to the high number of genes with zero read depth. Graph breaks are present at both ends of the graph. **(E)** Mean absolute depth-normalized log ratio of the analyses shown in panel D.

SmartMap drives new biological insights about repetitive DNA elements

With this method, we sought to better explore the role of histone modifications at repetitive regions. Traditionally, the epigenetic profile of repetitive elements is viewed in light of the "genome defense" hypothesis, which suggests that regulation of repetitive elements (and particularly transposable elements) serves to silence the elements and thereby prevent transposition[39]. Consequently, much previous work on this topic has primarily pointed towards repetitive elements being enriched with heterochromatin-associated modifications such as H3K9me2[315], H3K9me3[39,101,130,316,317], and H3K27me3[39,101,318]. In recent years, some studies have described a role for canonically activating histone modifications at a subset of repetitive elements[319–324]. Indeed, this body of work has suggested that some long interspersed nuclear elements (LINEs) can bear marks such as the transcriptionally activating H3K4me3 modification, particularly early in development[319,323,324]. Similarly,

other work has suggested that a class of mammalian-wide interspersed repeats (MIRs) may be transcriptionally active and play a role in enhancer regulation[321]. Much of this work, however, has relied upon uncalibrated ChIP with antibodies of uncertain specificity, both of which can result in data distortion and biologically incorrect conclusions[124]. Further, the ChIP-seq and RNA-seq studies have used a variety of different methods of aligning and filtering for reads to reach their conclusions, none of which used a method similar to our Bayesian SmartMap analysis, which may further affect the interpretations of the experiments. As such, we sought to use our calibrated and highly specific ICeChIP-seq datasets in conjunction with SmartMap to gain new insights into the epigenetic landscape of repetitive elements and to examine the degree to which uniread analysis yields an incomplete view of the data.

To accomplish this, we examined the histone modification landscape at the promoters of all repetitive elements, LINEs, short interspersed nuclear elements (SINEs), and Simple Repeats. K-means clustering analysis on all repetitive elements revealed four classes of promoters, each with a different histone modification profile: Cluster 1, enriched for H3K27me3 and H3K9me3; Cluster 2, enriched for H3K4me1 and H3K4me2; Cluster 3, enriched for H3K4me2 and H3K4me3; and Cluster 4, which is relatively depleted of histone modifications (Fig. 5.18A). These clusters are roughly reminiscent of the functional classifications of the ENCODE hidden Markov model, where Clusters 1, 2, and 3 correspond to silenced promoters, enhancers, and active promoters, respectively[325]. Interestingly, in all but Cluster 4, a greater proportion of nucleosomes is enriched with H3K27me3 than H3K9me3, despite the previous emphasis on the latter in repetitive element silencing[39,101,130,316,317], emphasizing the importance of calibration in ChIP-seq studies for comparing different modifications[118,124]. Similar histone modification profiles are seen for the LINEs (Fig. 5.19A), SINEs (Fig. 5.19B), and Simple Repeats (Fig. 5.19C). Across all these classes,

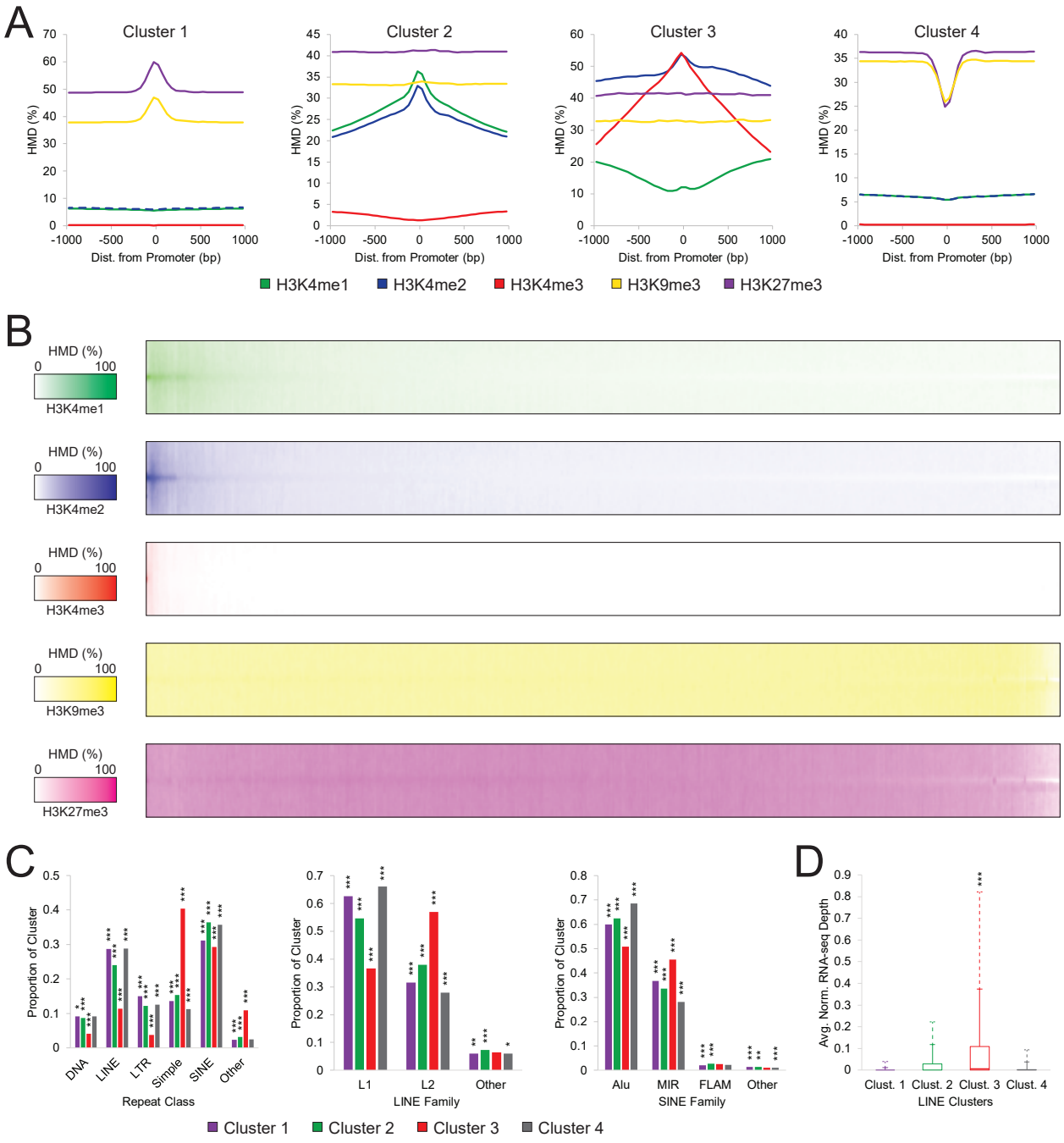


Figure 5.18: Assessment of histone modifications at promoters of repetitive DNA elements.

(A) Mean histone modification densities (HMDs) about promoters for classes of all repetitive elements, as defined by k-means clustering. Corresponding analyses of LINE, SINE, and Simple Repeat elements in S13 Fig. **(B)** Heatmap of repeat promoters with newly measurable HMD in SmartMap analysis, sorted on first principal component of repetitive elements. **(C)** Proportion of each cluster comprised by each repeat class or family for all repeats (left), LINE elements (center), and SINE elements (right). All significance tests performed as post-hoc Bonferroni-corrected

Figure 5.18, continued:
pairwise 2x2 chi-square tests. **(D)** Quantile boxplots of average normalized RNA-seq read depth across LINE elements for each LINE cluster. Solid line with marker represents 90th percentile; dashed line with marker represents 95th percentile. Significance test shows difference in median by Bonferroni-corrected pairwise Mood's median tests. Significance markers: $^{*}p < 0.01$, $^{**}p < 10^{-5}$, $^{***}p < 10^{-10}$.

Cluster 3 had the highest ATAC-seq signal (Fig. 5.19D-G), consistent with the presence of histone modifications associated with transcription and accessible chromatin[17,18,326].

Importantly, SmartMap analysis enabled us to more accurately measure HMD and assign clusters than did uniread analysis. Overall, there were 142,392 promoters with nonzero HMD in the SmartMap analysis that displayed no measurable HMD within 200bp of the TSS across all five histone modifications in the uniread dataset; similarly little HMD was detected within 1kb of the same in the uniread dataset (Fig. 5.20). This increase in HMD was substantial; most such sites had meaningful levels of histone modifications (Fig. 5.18B). A small subset was primarily H3K4me2/me3 predominant; a larger subset had high levels of H3K4me1/me2, and the remainder were primarily characterized by H3K27me3 and H3K9me3 (Fig. 5.18B). These represent promoters that would have been misclassified as histone-modification-depleted under uniread analysis; it is only through proper allocation of multireads that we can measure their HMDs and assign them to the appropriate cluster of repeat elements.

The distribution of repetitive elements across these clusters revealed interesting patterns. The distribution of the repeat classes or families across the clusters are presented in Table 5.5 for all repeats, Table 5.6 for LINEs, and Table 5.7 for SINEs, and summarized in Fig. 5.18C. Notably, amongst SINEs, MIRs were enriched in Cluster 3 (Fig. 5.18C), consistent with previous descriptions of a class of transcriptionally active MIRs[321]. In addition, Cluster 3 was enriched for Simple Repeats across all repeat promoters, consistent with descriptions of Simple Repeats in and around protein

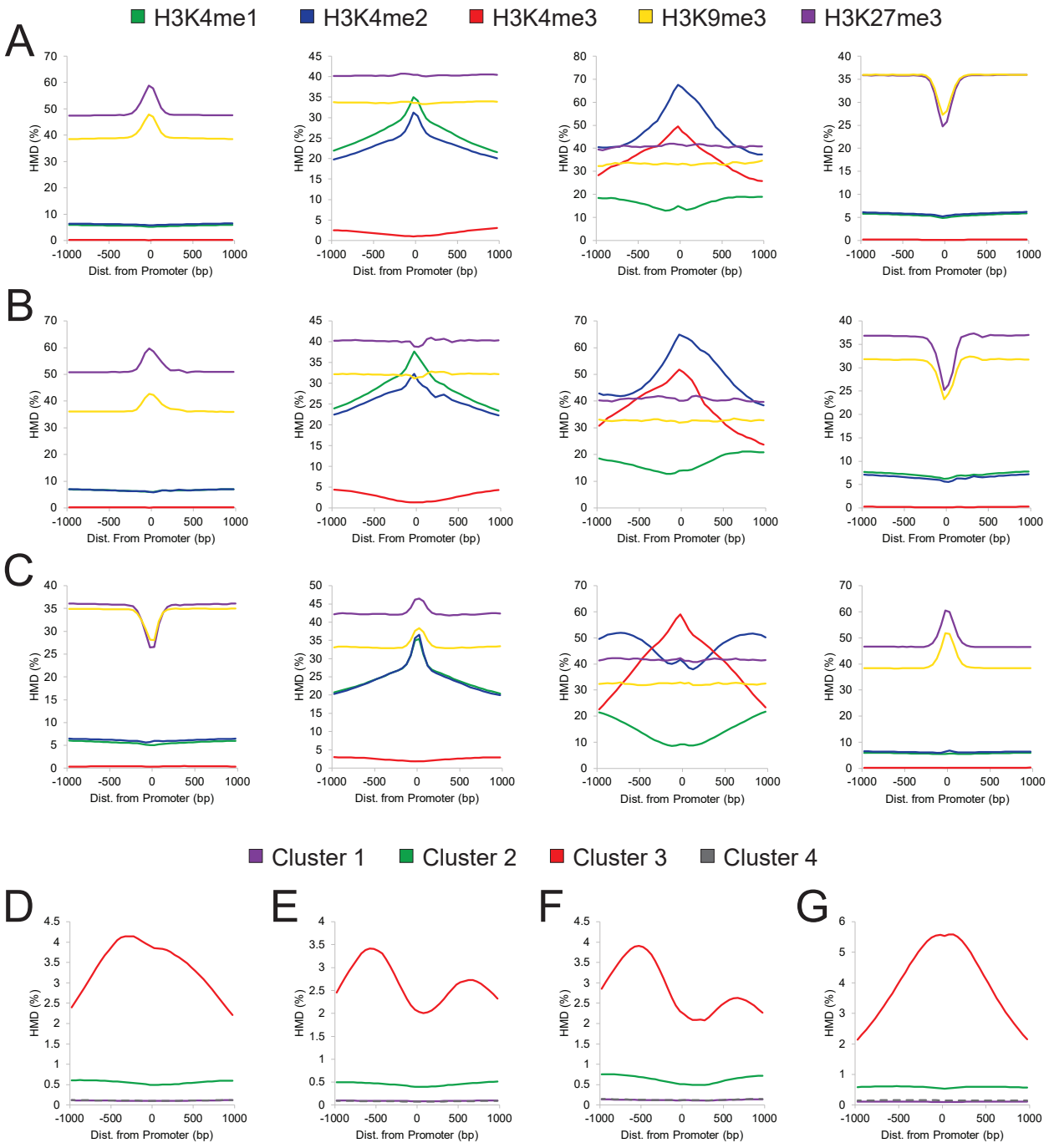


Figure 5.19: Histone modification and ATAC-seq profiles on subset clusters.

(A-C) HMDs of modifications about promoters of (A) LINEs, (B) SINEs, or (C) simple repeats separated by k-means clustering conducted on the appropriate set of repetitive elements. **(D-G)** Total ATAC-seq read depth across both replicates about promoters of (D) all repeats, (E) LINEs, (F) SINEs, or (G) simple repeats.

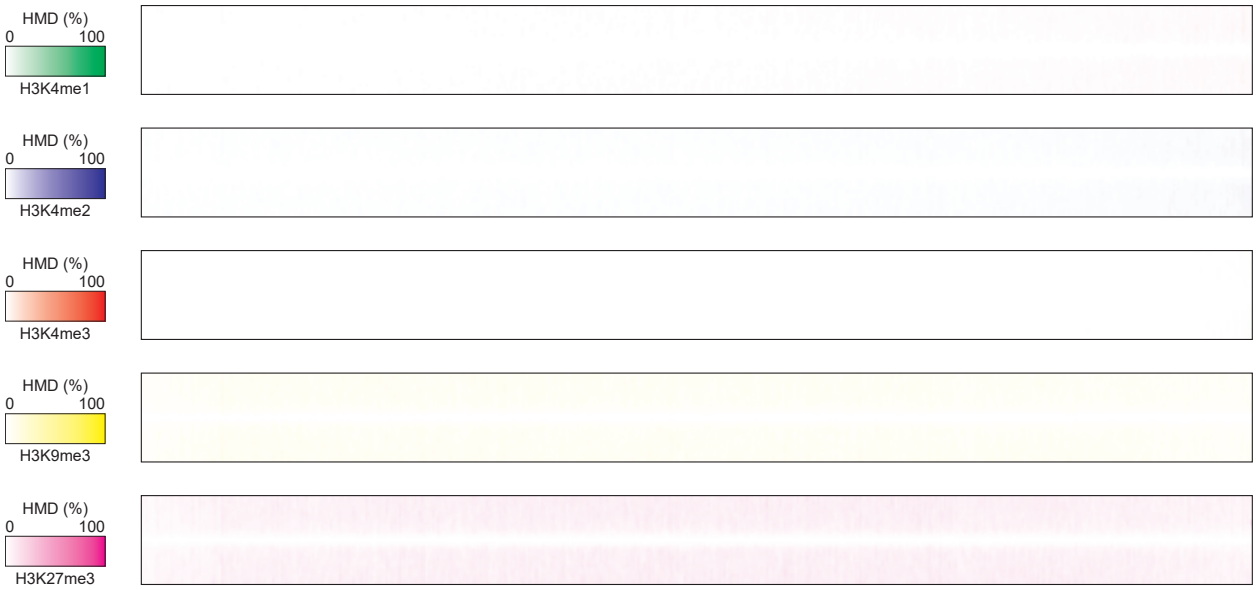


Figure 5.20: Heatmaps of repeat promoters under uniread analysis.
Heatmap of repeat promoters with measurable nonzero HMD only in SmartMap analysis, sorted on first principal component of repetitive elements.

coding genes in the literature[327]. Interestingly, Cluster 3 was enriched for the L2 subtype of LINEs, despite previous work primarily focusing on the role of H3K4me3 at L1 elements[319], representing a novel prediction of transcriptional activity of this family. To this end, using SmartMap analysis of the RNA-seq data, we found that the Cluster 3 LINEs had greater transcriptional activity than did the other clusters (Fig. 5.18D), confirming the transcriptional activity suggested by the presence of H3K4me3. Collectively, these data demonstrate the risk in only focusing on unireads – namely, the risk of missing important classes of genomic features – and highlights the role of multiread analysis of both DNA and RNA in driving new biological discovery.

Discussion

In this work, we have described a method to markedly increase sequencing depth genome-wide by analyzing ambiguously mapped reads rather than discarding them. This is of particular importance given that a significant portion of commonly studied genomes are not uniquely mappable by single-

Table 5.5: Clustering of repetitive elements.

Repeat Class	Cluster 1	Cluster 2	Cluster 3	Cluster 4	**Total**
DNA	240,787	28,465	763	232,324	**502,339**
LINE	754,991	78,670	2,133	734,556	**1,570,350**
LTR	393,797	39,911	707	319,769	**754,184**
Simple Repeat	357,734	50,395	7,555	287,900	**703,584**
SINE	818,624	119,404	5,478	908,873	**1,852,379**
Other	61,336	10,461	2,038	62,635	**136,470**
Total	**2,627,269**	**327,306**	**18,674**	**2,546,057**	**5,519,306**

Table 5.6: Clustering of LINEs.

LINE Family	Cluster 1	Cluster 2	Cluster 3	Cluster 4	**Total**
L1	463,275	46,376	828	490,788	**1,001,267**
L2	233,599	32,236	1,289	207,410	**474,534**
Other	43,993	6,221	146	44,189	**94,549**
Total	**740,867**	**84,833**	**2,263**	**742,387**	**1,570,350**

Table 5.7: Clustering of SINEs.

SINE Family	Cluster 1	Cluster 2	Cluster 3	Cluster 4	**Total**
Alu	520,927	72,223	2,880	590,748	**520,927**
MIR	319,283	38,836	2,587	241,855	**319,283**
FLAM	17,988	3,069	148	18,782	**17,988**
Other	11,943	1,634	61	9,415	**11,943**
Total	**870,141**	**115,762**	**5,676**	**860,800**	**870,141**

end or paired-end sequencing[304,305]. This difficulty arises in no small part due to the repetitiveness of the genome[22], but despite their difficulty to map, repetitive elements play critical roles in genomic regulation and function[306]. It is common discard these multireads entirely, despite these reads representing up to 30% of the sequencing depth. Works that do utilize multireads often simply select an alignment at random. We demonstrate that our SmartMap algorithm can better map reads

to the repetitive portion of the genome, facilitating better understanding their functions. Importantly, we find that the usage of alignment quality scores and paired-end sequencing can markedly increase the accuracy of alignment weights.

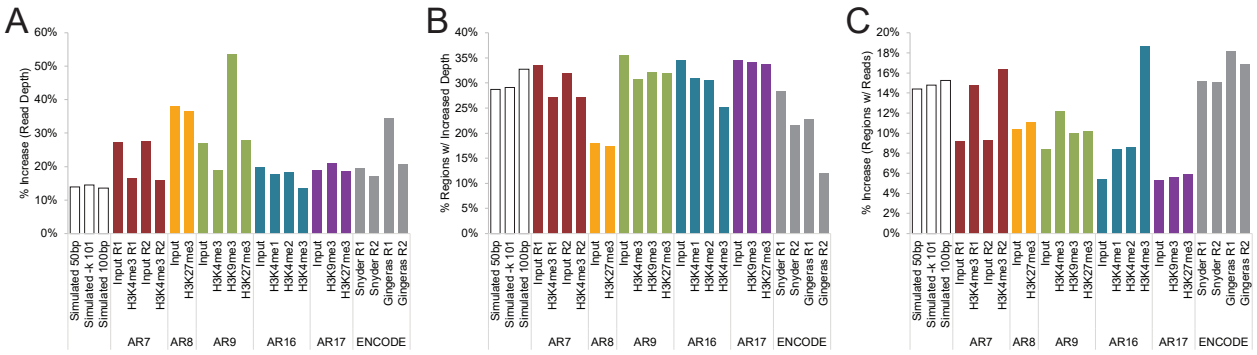


Figure 5.21: Analysis of increased usable read depth.
This figure graphically represents the data in Tables 5.1 and 5.2. **(A)** The percent increase in the number of reads usable in SmartMap analysis (reads with 1-50 alignments) relative to uniread analysis (reads with 1 alignment). **(B)** Percentage of the total number of regions with an increase in read depth in the SmartMap dataset relative to the uniread dataset. For all datasets except the ENCODE RNA-seq datasets, the list of regions analyzed is the set of 200bp genomic windows across the relevant genome (hg38, mm10, or dm3). For the ENCODE RNA-seq dataset, the list of regions analyzed is the set of distinct Refseq genes. **(C)** Percent increase in the number of regions with nonzero read depth in the SmartMap dataset relative to the uniread dataset.

Just by incorporating multireads with 2-50 alignments, we were able to increase the read depth of our samples by 13-53% (Fig. 5.21A and Table 5.1). This increase in read depth was not simply distributed across the entire genome, which is critical for the usefulness of this method. If the multireads were distributed uniformly, it would only modestly decrease error by slightly increasing read depth at all loci[118]. However, that is not the case; rather, the multireads are concentrated in a minority of the genome (Fig. 5.21B and Table 5.2), bringing regions of lower mappability to read depths comparable with highly mappable loci (Fig. 5.11A). The multiread samples have a 5-20% increase over unireads in loci with nonzero read depth (Fig. 5.21C and Table 5.2), representing a

sizable proportion of the genome that is completely ignored by uniread analysis and can be recovered only by utilizing ambiguously mapped reads.

Our method requires no particular experimental modifications or additional controls for analysis of multireads and can be applied post hoc to existing datasets. As such, SmartMap can be used to leverage the existing compendium of NGS datasets more accurately. Though we primarily used ICeChIP-seq data to demonstrate and explore the capabilities of SmartMap, this tool is not solely designed for ICeChIP and does not require the internal standards used therein. Indeed, SmartMap is designed to be a general tool for a broad range of next-generation sequencing experiments, including ChIP-seq, MNase-seq, and ATAC-seq, as we showed in this work. In addition, though we have used paired-end sequencing here, there is little reason to believe this method could not be used for a single-end sequencing experiment. In principle, an algorithm using the principles of SmartMap can be applied to any NGS experiment, past or future, that involves alignment to a genome.

Previously published methods have utilized a variety of techniques to allocate multireads; however, our analysis suggests that many of these methods may be problematic. One heuristic assumes that multireads and unireads have similar genomic distributions and, accordingly, assigned multireads weights in proportion with uniread depth[295,328]. Our data, by contrast, shows that multireads instead concentrate into a minority of loci (Table 5.2) and particularly those with low uniread depth (Fig. 5.4C-D and 5.5C). This suggests that the unireads and multireads have different genomic distributions, violating the critical assumption underlying proportional allocation of multireads. Another method of resolving multireads is to select one alignment at random for each read[39,312]. The expected value of the read distribution under this procedure converges to that of SmartMap without reweighting, which we found to have higher error than SmartMap with a Bayesian reweighting

cycle (Fig. 5.2D, 5.5H and Table 5.8). Indeed, explicit comparison to an instance of random read comparison revealed even higher error as compared to SmartMap with and without reweighting (Table 5.8).

Table 5.8: Benchmarking SmartMap software.

	Pre-algorithm Alignment and Processing				Read Allocation Algorithm			
	Read Alignment		Processing Alignments		Reading Alignments			
Algorithm	CPU Time	Wall Time	CPU Time	Proc. File Size	CPU Time	Max. Memory	Algorithm Time	Avg. MAE
SmartMap	317:30:25	6:39:46	1:34:29	59 GB	0:16:49	53 GB	0:42:38	4.04
BM-Map	317:30:25	6:39:46	N/A	820 GB	6:25:09	146 GB	ERROR	–
Iteration 0	317:30:25	6:39:46	1:34:29	59 GB	0:16:10	53 GB	0:35:16	4.12
Random	317:30:25	6:39:46	2:15:12	15 GB	0:03:58	39 GB	0:15:46	5.48
Uniread	36:08:04	0:45:34	0:17:07	13 GB	0:03:19	39 GB	0:14:43	6.50

Benchmarking conducted on computer with Ubuntu 20.04.1 LTS with 224 GB of RAM and dual Intel Xeon CPU E5-2690 v3 @ 2.60GHz processors, running on one thread except the read alignment, which used 48 threads. All times are represented in hours:minutes:seconds.

Alignment conditions are identical for all but Uniread. Parsing reads is typically conducted in parallel with alignment. File size represents the size of the required file after read parsing needed for the algorithm in question. Reading alignments is part of each algorithm and is included in the Algorithm Total Time.

Average Mean Absolute Error (Avg. MAE) is computed against the gold standard on the set of true origin loci. These benchmarking analyses were conducted separately with separate alignments from the analyses in Fig. 5.5, and the avg. MAEs vary slightly in magnitude from those presented in Fig. 5.5.

SmartMap is also computationally efficient as compared to the most similar previous algorithms and software for the assignment of multireads. This is due in part to the low number of reweighting iterations our algorithm uses, which decreases the computational burden of the software. In addition, the Fenwick tree data structure used with our method allows for more efficient processing of reads by accessing and updating of genomic weights. Previous implementations

of a scored-alignment reweighting algorithm, as done by BM-Map, have required more than five hours to process approximately seven million aligned reads after alignment in previous studies[311]. Unfortunately, we were not able to fully measure the time requirements for BM-Map for ourselves; implementing both CSEM[310] and BM-Map[311] proved challenging, as described in the Methods. However, using our simulated dataset (with 50bp reads), including more than 740 million alignments from more than 275 million reads, even just reading the alignments with BM-Map on our hardware took more than 6 hours after alignment (Table 5.8). By contrast, our algorithm can completely process that same aligned dataset in less than 2.5 hours, representing more than 100-fold less CPU time than the alignment itself (Table 5.8). As such, the low CPU-time requirements of SmartMap drastically increases our ability to use this algorithm on data from modern NGS experiments, particularly given the ever-increasing depth and decreasing cost of sequencing[329]. Though it is, admittedly, faster to solely process unireads than to conduct SmartMap (Table 5.8), the added time is not egregiously high; on our system, the full benchmarking (including alignment) required roughly eight more hours of wall time in the SmartMap analysis than in the uniread analysis.

This SmartMap method is, however, not without its limitations. Primary amongst these limitations is that rather than yielding a list of alignments, the SmartMap software either outputs the read depth at each base pair genome-wide or a list of alignments with associated weights. While this is useful for any analysis that utilizes the read depth at a given position, this makes it difficult to use downstream methods or tools that primarily utilize the full list of alignments using off-the-shelf tools. In particular, this makes it challenging to compute gene expression in RNA-seq per common methods such as FPKM, which uses the number of reads overlapping a transcript as a measure of expression rather than the read depth per base pair. This is partially alleviated by the fact that SmartMap provides the option to write lists of alignments with their corresponding weights, but

even so, incorporating these weights into existing downstream analyses, pipelines, and software may remain challenging.

In addition, the SmartMap method also may face challenges with any alignments with significant gaps relative to the alignment templates, such as those created by RNA splicing (or Hi-C experiments). Because our reweighting algorithm assigns weights based on the average read depth across an alignment, an alignment spanning a splice junction in RNA-seq may be unfairly assigned a lower weight due to decreased read depth in the intron. As such, highly spliced genes may be given a lower read depth than a similarly expressed gene with fewer introns. This could be partially accommodated by weighting with the total read depth over an alignment rather than the average read depth over the same, but this method would potentially unfairly increase weight of longer alignments, which could pose another challenge.

In addition, from a computational perspective, the SmartMap method is memory intensive. This is in large part due to the data structure used for storing genome-wide weight data. Because this tool is designed to be compatible with reweighting of paired-end reads and obtaining the total weight across a paired-end read, the data structure needs to efficiently conduct both range update and range query operations. Accordingly, for the strand-independent method, we have used a dual binary-indexed tree data structure; for strand-specific analysis, we use a dual binary-indexed tree structure for each strand, for a total of four binary-indexed trees. For this reason, for our simulated dataset, the SmartMap analysis required almost 60 GB of memory. In principle, a lower-memory method could be developed that would only use one binary-indexed tree per strand, but this would require iteration over each base position of each alignment and would thereby dramatically decrease time-efficiency. However, it's important to note that BM-Map, the only other tested software that was even able to successfully read the alignments, required almost 150 GB to read that same set

of alignments into memory (Table 5.8). In practice, with the decreasing costs of memory and the increasing availability of computational servers and clusters for a wide variety of bioinformatic tools and analyses, the memory requirements are likely workable for many users, particularly because of the low CPU time required.

Finally, even the best SmartMap analysis can only be as good as the alignment itself. In this work, we have largely restricted our Bowtie2 alignments to report a maximum of 51 alignments, with the exception of the analysis with a maximum of 101 alignments. This was conducted for feasibility; as the maximum number of reported alignments increases, so too does the computational overhead needed for alignment of the reads by Bowtie2. However, this does place an inherent limitation on our ability to look at the most repetitive regions of the genome, which can be found at hundreds of loci throughout the genome and can thus pose a significant challenge to alignment and multiread analysis. Granted, raising this threshold to a maximum of 101 alignments per reads had practically no impact on the analysis on the human genome (Fig. 5.8, Tables 5.1 and 5.2), but nonetheless, there were still nearly seven million reads that aligned to the maximum of 101 loci, representing a significant number of reads with even more potential alignments. Further, some genomes have even greater repetitiveness than does the human genome; for example, repetitive elements comprise roughly 85% of the maize genome[330], making alignment all the more challenging and raising the number of plausible alignment sites for each read. It is important to note that this is not a limitation that is inherent to SmartMap, but rather, to alignment itself. If the end user was able to generate an alignment with an arbitrarily high maximum number of reportable alignments, there is no reason that SmartMap should fail; it is not inherently capped at a maximum number of alignments per read. It should be remembered that SmartMap will not be able to "fix" an alignment with too few alignments

per read. Accordingly, it may be necessary to tune the maximum number of alignments per read to appropriately analyze data originating from some genomes despite the added computational load.

Despite these limitations, we were nonetheless able to demonstrate the usefulness of our SmartMap tool to process reads from a variety of NGS workflows (e.g. MNase-seq, ChIP-seq, ATAC-seq, and RNA-seq) and to investigate biological questions – in this case, the epigenetic regulation of repetitive elements. Just as importantly, we demonstrated the risk of using only unireads – namely, that biologically relevant regions will be hidden from analysis because the multireads have been discarded. Given the critical role that repetitive regions play in biological regulation[306], being able to analyze these regions is crucial to gaining a more complete understanding of genomic structure and function. Accordingly, we hope this method will help enable researchers to use their sequencing data more completely and thereby gain more useful information from their experiments.

Acknowledgements

We would like to thank the ENCODE Consortium for providing the RNA-seq and ATAC-seq data used in this study. In particular, we thank the Gingeras Laboratory at Cold Spring Harbor Laboratory for generating the RNA-seq data, and we thank the Snyder Laboratory at Stanford University for generating the ATAC-seq data.

Methods and Materials

Sequencing Data Sources

The ICeChIP-seq datasets analyzed in this work, with the exception of AR17 H3K27me3 IP, were sourced from previously published ICeChIP-seq datasets[118,124]. The FASTQ files for datasets

AR7, AR8, and AR9 can be obtained from Gene Expression Omnibus (GEO) Accession Number GSE60378. The FASTQ files for datasets AR16 and AR17 can be obtained from GEO Accession Number GSE103543. Inputs for each ICeChIP are generated by MNase-seq.

The AR17 H3K27me3 ICeChIP-seq was conducted at the same time as the AR17 H3K9me3 ICeChIP-seq experiment using the same AR17 Input, but was not published previously[124]. It was generated by ICeChIP-seq as previously described[124] using an anti-H3K27me3 antibody (Cell Signaling Technologies, Product Number 9733, Lot 8). This dataset has been made available at GEO Accession Number GSE103543.

RNA-seq data was obtained from the ENCODE Project[331], experiment ENCSR000AEL. The FASTQ files for Isogenic Replicate 1 was obtained from the dataset for library ENCLB053ZZZ (FASTQ accession numbers: ENCFF001RFF, ENCFF001RFE). The FASTQ files for Isogenic Replicate 2 was obtained from the dataset for library ENCLB054ZZZ (FASTQ accession numbers: ENCFF001RFD, ENCFF001RFC).

ATAC-seq data was obtained from the ENCODE Project[331], experiment ENCSR483RKN. The FASTQ files for Isogenic Replicate 1 was obtained from the dataset for library ENCLB918NXF (FASTQ accession numbers: ENCFF391BFJ, ENCFF186CQZ). The FASTQ files for Isogenic Replicate 2 was obtained from the dataset for library ENCLB758GEG (FASTQ accession numbers: ENCFF440UAD, ENCFF350ZZR).

Mappability Scores

The mappability score chosen was the UMAP50 score, which represents the proportion of 50bp kmers overlying a given point that are unique in the genome[305]. The approximate UMAP50 score of the dm3 genome was computed by computing all 50-mers in the genome and determining those

that are unique; the genome coverage of the unique 50-mers was then determined to compute approximate UMAP50 score of the genome.

Simulated Dataset

The simulated dataset was generated as followed. First, 6 million loci of length 200bp in the genome were randomly selected and designated as the target loci. Paired-end Illumina sequencing reads were then simulated using NEAT-genReads[332] using the list of target loci as the target file and the following settings: 50bp read length, 30-fold target coverage, default off-target coverage, and insert size 175bp average and 10bp standard deviation. The output list of true read locations was then used to compute a Gold Standard genome coverage BedGraph using BEDTools genomecov[174]. The average Gold Standard read depth of the target loci was then computed as described below, and the target loci with nonzero Gold Standard read depth were designated as the "true origin" loci and used for downstream analysis.

To generate the simulated dataset with 100bp read length, the same procedure was used on the same set of 6 million loci, with the NEAT-genReads tool being set to output 100bp reads rather than 50bp reads. Unless otherwise specified, this work uses "simulated dataset" or similar to refer to the simulated dataset with 50bp reads.

Computing average value of BEDGRAPH at target loci

Because the BEDTools map software does not compute base-pair-wise averages of BEDGRAPH signals, the following procedure was used to compute read depth at a list of target loci. Overlapping loci were merged using BEDTools merge, and the resultant list of loci were partitioned into 1bp windows using BEDTools makewindows. The BEDGRAPH was then mapped onto the windows

using BEDTools map, and the mapped windows were then mapped with the mean function onto the original list of target loci using BEDTools map.

Mappability estimation and binning

The mappability of a list of loci was computed by computing the average value of the UMAP50 bedgraph for the relevant genome at those loci using the method described above. To compute the number of regions per UMAP50 score, the loci were binned by average UMAP50 score in bins of width 0.01. The number of loci at each bin were then computed to determine the approximate distribution of UMAP50 score across the selected loci.

MACS2 Peak Calling

Peak calling was conducted on the simulated datasets using MACS2[176] with the bdgpeakcall function with the relevant BEDGRAPH file and default settings.

Alignment and Read Filtering and Processing

FASTQ files for ICeChIP-seq or the simulated dataset were aligned using Bowtie2[172] due to its common usage in the field and due to its ability to report alignment scores for each mate for each alignment reported as opposed to for just the best alignment. Bowtie2 alignment was run on the paired-end sequencing samples with the following settings: end-to-end alignment, very-fast preset settings, no discordant alignments, no mixed alignments, report up to 51 alignments, insert size minimum 100bp, insert size maximum 250bp. In the case of the analysis with up to 101 alignments (the k101 dataset), the above settings were used with up to 101 alignments reported per read. The genomes used for alignment were as follows: AR7, mm10 with ICeChIP barcodes series 1; AR8,

dm3 with ICeChIP barcodes series 1; AR9, mm10 with ICeChIP barcodes series 2; AR16 and AR17, hg38 with ICeChIP barcodes series 3; simulated datasets, hg38.

FASTQ files for RNA-seq were aligned using Hisat2[333] for the same reasons as the choice to use Bowtie2. Hisat2 alignment was run on the paired-end sequencing samples with the following settings: no discordant alignments, no mixed alignments, report up to 51 alignments. The genome used for alignment was hg38 with the ENCODE ERCC standards.

FASTQ files for ATAC-seq were aligned using Bowtie2[172] on the paired-end sequencing samples with the following settings: no discordant alignments, no mixed alignments, report up to 51 alignments, insert size maximum 2000bp. The genome used for alignment was hg38.

Alignments were then filtered to select for reads that are paired, mapped in a proper pair, and mate on the reverse strand, corresponding to SAM flags of 99, 163, 355, and 419. For non-strand-specific applications, the selected SAM file records were then extracted into a file containing the following fields: chromosome, start position, stop position, read ID, read alignment score (field labeled "AS:i:"), mate alignment score (field labeled "YS:i:"). For strand-specific applications, the selected SAM file records were extracted into a file containing the following fields: chromosome, start position, stop position, read ID, strand, read alignment score (field labeled "AS:i:"), mate alignment score (field labeled "YS:i:"). The reads were then split into separate BED files based on the number of alignments per read. For downstream uniread analyses, only the reads with 1 alignment were used; for downstream SmartMap analyses, reads with 1-50 alignments were used except for the k101 dataset, in which case reads with 1-100 alignments were used.

The file with 51 alignments per read (or that with 101 alignments per read for the k101 analysis) was not used for downstream analyses to prevent confounding by reads with fewer reported than possible reads.

Uniread and SmartMap Analysis of Genome Coverage

For the uniread analysis, our SmartMap software was used with only the file containing reads with only 1 alignment per read. For the SmartMap analysis, our SmartMap software was used with the files containing reads with fewer than 51 alignments per read.

The SmartMap software uses a set of dual Binary Indexed Trees to store map counts and weights across the genome and uses an iterative Bayesian reweighting algorithm to assign weights to each of the different alignments. These steps are outlined below. Unless otherwise specified, all analyses are conducted with 1 iteration in scored mode. For the strand-specific analyses, there is a separate set of dual Binary Indexed Trees for each strand.

STORAGE OF MAP COUNTS IN THE GENOME

To facilitate efficient summation and updating of map counts and weights across the genome, each chromosome is stored as a pair of Binary Indexed Trees (BIT), also known as Fenwick Trees. The BIT is a data structure that is efficient for computing prefix sums of an ordered dataset from the beginning of the dataset to the given index. Because we used a 1-based coordinate system for the genome, the datasets to which we refer as being represented by a BIT should be assumed to be 1-based datasets unless otherwise specified.

For a dataset of length L, the BIT is represented by $L+1$ nodes, which are stored in an array. To increment a dataset represented by a BIT T at index i by the value v, the following algorithm is used. Let $T[i]$ represent the ith node of T. Let $lsb(i)$ represent the lowest significant bit in the binary representation of i. Then:

$$T[i] = T[i] + v \qquad \text{(Eqn. 5.1)}$$

$$i = i + lsb(i) \quad \text{(Eqn. 5.2)}$$

If the new value of $i \leq L + 1$, Eqn. 5.1-5.2 are iterated until $i > L + 1$. For brevity, we will refer to this operation to increment the BIT T representing the 1-based dataset by v in the value i as BITUpdate(T, i, v).

To compute the prefix sum of the dataset at index i (i.e. the sum of all values with indices $[1, i]$ of a 1-based dataset), the following algorithm is used, using the above definitions of $T[i]$ and $lsb(i)$. Let the prefix sum be represented by sum, where sum is initially set to zero. Then:

$$sum = sum + T[i] \quad \text{(Eqn. 5.3)}$$

$$i = i - lsb(i) \quad \text{(Eqn. 5.4)}$$

If the new value of $i > 0$, Eqn. 5.3-5.4 are iterated until $i \leq 0$. For brevity, we will refer to this operation to obtain the prefix sum of the T at value i as $BITSum(T, i)$.

To understand how we here use BITs to efficiently store values across the genome and efficiently sum the values across loci, consider the following.

Consider a dataset C represented by BITs T_1 and T_2. If the values of C for indices in range $[l, r)$ are incremented by v, then let the resulting dataset be represented by C'. Let the prefix sum of the resulting dataset C' at index i be represented by $PointSum(C', i)$. Then let $\Delta PointSum(C, i) = PointSum(C', i) - PointSum(C, i)$. $PointSum(C', i)$is changed in one of three ways:

Case 1: $i < l$. The increment on range $[l, r)$ will not change $PointSum(C', i)$. As such:

$$\Delta PointSum(C, i) = 0 \quad \text{(Eqn. 5.5)}$$

Case 2: $l \leq i < r$. In this case:

$$\Delta PointSum(C, i) = v * i - v * (l - 1)$$

However, per Eqn. 5.5, $\Delta PointSum(C, l - 1) = 0$. As such:

$$\Delta PointSum(C, i) = v * i - v * (l - 1) \quad \text{(Eqn. 5.6)}$$

Case 3: $i \geq r$. In this case, the increment on range $[l, r)$ will not change the values of C past index $r - 1$. Accordingly:

$$PointSum(C', i) = PointSum(C, i) + PointSum(C', r) - PointSum(C, r)$$

$$PointSum(C', i) - PointSum(C, i) = PointSum(C', r) - PointSum(C, r) \quad \text{(Eqn. 5.7)}$$

$$\Delta PointSum(C, i) = \Delta PointSum(C, r)$$

However, per Eqn. 5.6, $\Delta PointSum(C, r) = v * r - v * (l - 1)$. As such:

$$\begin{aligned} \Delta PointSum(C, i) &= (v + (-v)) * i - (v * (l - 1) - v * r) \\ &= v * r - v * (l - 1) \end{aligned} \quad \text{(Eqn. 5.8)}$$

These three cases will provide the basis for our use of BITs to store and efficiently sum values across the genome. Each chromosome C in the genome is stored as a pair of BITs, to which we shall here refer as T_1 and T_2. Let L represent the length of the chromosome. Accordingly, T_1 and T_2 have $L + 1$ nodes.

To increment the value associated with the base pairs in the range $[l, r)$ by an increment value v, the following procedure is used.

$$\begin{gathered} BITUpdate(T_1, l, v) \\ BITUpdate(T_1, r, -v) \\ BITUpdate(T_2, l, v * l) \\ BITUpdate(T_2, r, -v * r) \end{gathered} \quad \text{(Eqn. 5.9)}$$

The value associated with base pair i is then $BITSum(T_1, i)$. To obtain the prefix sum of the chromosome dataset C at base pair index i, represented by $PointSum(C, i)$, the following equation is used.

$$PointSum(C, i) = BITSum(T_1, i) * i - BITSum(T_2, i) \quad \text{(Eqn. 5.10)}$$

The sum of the values associated with each base pair in the range $[l, r) = [l, r-1]$ on chromosome C, represented by $LocusSum(C, l, r)$, can thus be described by:

$$\begin{aligned} LocusSum(C, l, r) =& PointSum(C, r-1) - PointSum(C, l-1) \\ =& BITSum(T_1, r-1) * (r-1) - BITSum(T_2, r-1) \\ & - BITSum(T_1, l-1) * (l-1) + BITSum(T_2, l-1) \end{aligned} \quad \text{(Eqn. 5.11)}$$

This dual-BIT data structure allows for efficient handling of data with respect to time complexity. The BITUpdate and BITSum steps occur with time complexity O(log L), and the updates to ranges (Eqn. 5.8) and range summations (Eqn. 5.10) use four BITUpdates and four BITSums, respectively. As such, both range updates and range queries occur with time complexity O(log L).

Iterative Bayesian Reweighting of Mapped Reads

To assess and appropriately weight reads mapped to different portions of the genome, we implemented a Bayesian approach which iteratively reweights the mappings associated with each read. For each read, we assign to each associated map a weight representative of the prior probability that the map is the origin of the associated read. We then iteratively use the distribution of the assigned maps and their weights (the prior probabilities) to determine the posterior probability for each map being the true origin of the associated read and assign that as the weight for that map, which then becomes the prior probability for the next iteration.

Let the set of all sequencing reads be represented as R, with an individual sequencing read being represented as r_i. Then $R = \{r_i \dots r_n\}$, where n represents the total number of sequencing reads obtained for the dataset in question.

Each read r_i is associated with a true genomic origin locus l_i and a set of genomic alignments $M_i = \{m_{i,1} \dots m_{i,k}\}$, where each $m_{i,j}$ represents a reported alignment of r_i and k represents the total number of alignments reported for r_i such that $k < k_{max}$, the maximum number of possible reported alignments. Each reported alignment $m_{i,j}$ is associated with an alignment score $s_{i,j}$, a weight $w_{i,j}$, and an alignment genomic locus $g_{i,j}$. We will define the set of all alignment scores associated with read r_i as $S_i = \{s_{i,1} \dots s_{i,k}\}$, with the set of all alignment weights associated with read r_i being represented as $W_i = \{w_{i,1} \dots w_{i,k}\}$, and with the set of all alignment loci associated with read r_i being defined as $G_i = \{g_{i,1} \dots g_{i,k}\}$.

For this algorithm, we assume that for each alignment m_i associated with a given read r_i, one of the associated alignment loci $g_{i,j}$ is the true origin locus l_i. Then each weight $w_{i,j}$ is defined as the probability $w_{i,j} = Pr(g_{i,j} = l_i)$, or the probability that the alignment associated with the weight $w_{i,j}$ is equal to the true origin locus.

The set of all true genomic origin loci l_i will be defined as $L = \{l_1 \dots l_n\}$. The set of all alignment scores, weights, and loci associated with every read in R will be defined as $S = \{S_1 \dots S_n\}$, $W = \{W_1 \dots W_n\}$, and $G = \{G_1 \dots G_n\}$.

These variables will define our analysis. Our observed variables are the set of alignment scores S and the set of alignment loci G. Our latent variable is the true genomic origin distribution L. We will be modeling to generate the set of alignment weights W with the goal of estimating the true read origin distribution L as the expected value of the set of reported alignments G with the

set of weights W being treated as the probability distribution of G upon which the expected value is computed.

When conducting analyses in "scored mode," we wish to consider the quality of each alignment. To do this, for each alignment $g_{i,j}$ of each read r_i, we will transform the associated alignment score $s_{i,j}$ into a pseudo-MAPQ score $z_{i,j}$ as per Bowtie2 computation of MAPQ for unireads. Let the minimum alignment score for reported alignments in Bowtie2 be represented as $s_{min} = -0.6 - 0.6 * (2 * \text{read length})$. Then:

$$z_{i,j} = \begin{cases} 42 & \text{if } \frac{s_{i,j}}{s_{min}} \in [0, 0.2] \\ 40 & \text{if } \frac{s_{i,j}}{s_{min}} \in (0.2, 0.3] \\ 24 & \text{if } \frac{s_{i,j}}{s_{min}} \in (0.3, 0.4] \\ 23 & \text{if } \frac{s_{i,j}}{s_{min}} \in (0.4, 0.5] \\ 8 & \text{if } \frac{s_{i,j}}{s_{min}} \in (0.5, 0.6] \\ 3 & \text{if } \frac{s_{i,j}}{s_{min}} \in (0.6, 0.7] \\ 0 & \text{if } \frac{s_{i,j}}{s_{min}} \in (0.7, 1] \end{cases} \quad \text{(Eqn. 5.12)}$$

If the analysis is being run in unscored mode, the quality of each alignment $q_{i,j}$ is set to 1. When conducting analyses in scored mode, from this pseudo-MAPQ score, we can compute the alignment quality $q_{i,j}$ as the probability that the alignment is aligned to the correct genomic locus from the definition of MAPQ as:

$$q_{i,j} = 1 - 10^{-z_{i,j}/10} \quad \text{(Eqn. 5.13)}$$

The set of alignment qualities associated with each read r_i is defined as $Q_i = \{q_{i,1} \dots q_{i,k}\}$. We will define our initial weights $w_{i,j}$ by setting our initial prior probabilities $Pr(g_{i,j} = l_i)$ to be proportional to the alignment quality $q_{i,j}$. Because we assume that for each read r_i, one of the

associated $g_{i,j} = l_i$, then for each read r_i, we set the associated alignment weights $w_{i,j}$ as:

$$w_{i,j} = \frac{q_{i,j}}{\sum_{q \in Q_i} q} \quad \text{(Eqn. 5.14)}$$

In scored mode, it is possible for the sum of the alignment qualities in Q_i for a given read r_i to be equal to zero; if this is the case, the read is discarded. Similarly, any alignments with alignment with a weight of zero are discarded. For all remaining reads and alignments, each weight $w_{i,j}$ is added to the appropriate chromosome dataset at the associated alignment locus $g_{i,j}$. Those reads with $k = 1$ are then removed from the list of reads over which to iterate because the weight of the associated alignment is fixed at $w = 1$.

When the initial assignment of prior probabilities as weights and addition of weights to the genome dataset is complete, then for each read r_i, the new weights can be computed as the posterior probabilities of $Pr(g_{i,j} = l_i \mid \text{total distribution of reads})$. First, we will represent the length of an alignment locus $g_{i,j}$ as $|g_{i,j}|$. Let c_b be the sum of all weights associated with all alignments containing the genomic coordinate b. Then, we define $C_{i,j}$ as the average weight across the genomic coordinates of each alignment $g_{i,j}$ by the equation:

$$C_{i,j} = \frac{1}{|g_{i,j}|} \sum_{b \in g_{i,j}} c_b \quad \text{(Eqn. 5.15)}$$

Our algorithm assumes that the probability $Pr(g_{i,j} = l_i \mid \text{total distribution of reads})$ is proportional to $C_{i,j}$ and proportional to the alignment quality $q_{i,j}$. Based on this assumption, we define our likelihood function $F_{i,j}$ as the ratio of the average quality-weighted weight across the alignment locus to the weight of the alignment itself:

$$F_{i,j} = \frac{C_{i,j}\, q_{i,j}}{w_{i,j}} \quad \text{(Eqn. 5.16)}$$

By Bayes' theorem, we then state that our posterior probability is proportional to the likelihood and to the prior probability of a given event. To accommodate for slow fitting, we will define

r as the learning rate such that if $r = 0$, the weight will not change at all, and if $r = 1$, the new weight will be defined as per Bayes' theorem. When $r = 1$, then, we thus set our new weight $w'_{i,j}$ as our posterior probability $Pr(g_{i,j} = l_i \mid \text{total distribution of reads})$ by the equation:

$$\begin{aligned} w'_{i,j} &= \frac{w_{i,j} F_{i,j}}{\sum_{j=1}^{k} w_{i,j} F_{i,j}} \\ &= \frac{C_{i,j} q_{i,j}}{\sum_{j=1}^{k} C_{i,j} q_{i,j}} \end{aligned} \qquad \text{(Eqn. 5.17)}$$

If r is not equal to 1 (i.e. if fitting is conducted more slowly or faster), then per the above definition of the learning rate:

$$\begin{aligned} w'_{i,j} &= \left(\frac{C_{i,j} q_{i,j}}{\sum_{j=1}^{k} C_{i,j} q_{i,j}} - w_{i,j} \right) r + w_{i,j} \\ &= \frac{r C_{i,j} q_{i,j}}{\sum_{j=1}^{k} C_{i,j} q_{i,j}} + (1 - r)\, w_{i,j} \end{aligned} \qquad \text{(Eqn. 5.18)}$$

Per this definition, when fitting is disabled (i.e. when $r = 0$), the new weight is not changed; when the fitting rate is set to $r = 1$, then Eqn. 5.17 is equal to Eqn. 5.16. Slower fitting can be achieved by setting $0 < r < 1$. The new weights are updated at the appropriate corresponding genomic loci, and the posterior weight $w'_{i,j}$ is treated as the prior weight $w_{i,j}$ for the next iteration. This process defined by Eqn. 5.15-5.17 is conducted iteratively for the specified number of iterations.

After the specified number of iterations are complete, the output file is written by writing the prefix sum of the BIT T_1 for each chromosome at each position. If desired, the set of reads with corresponding weights are also written.

Histone Modification Density and Specificity Computation

Because the ICeChIP-seq datasets have internal nucleosome standards bearing the targeted nucleosome modifications with uniquely identifying DNA "barcodes", we were able to calibrate our

ChIP-seq results to yield the histone modification density (HMD), or the proportion of nucleosomes bearing the modification of interest. HMD for each dataset was computed as follows.

The average value of the BEDGRAPH for each of the calibrant barcodes was computed as above, and these values were grouped by the nucleosome modification associated with the barcode and summed, as previously described[118,124] for both the IP and the Input datasets. The ratio of the summed values for the targeted modification in IP over the same in Input was designated as the target enrichment E_t.

The HMD at each genomic locus was then computed as follows, where IP and $Input$ represent the value of the IP and the Input at that genomic locus:

$$\text{HMD } (\%) = \frac{\text{IP}}{E_t * \text{Input}} * 100\% \qquad \text{(Eqn. 5.19)}$$

To generate genome-wide HMD BEDGRAPH files, the IP and corresponding Input genome coverage BEDGRAPH files outputted by the SmartMap software were merged with BEDTools unionbedg, and the HMD computation in (18) was used to compute HMD. Any region with an Input value of zero was set to an HMD of zero, as there is no nucleosome coverage to be modified at those loci.

To compute the specificity, for those ICeChIP-seq datasets with calibrants bearing more than one modification with uniquely identifying DNA barcodes (AR9, AR16, and AR17), the enrichment of every species E_i was computed analogously to the E_t, and the specificity (as percent of target enrichment) was computed as:

$$\text{Specificity}(\%\text{ target}) = \frac{E_i}{E_t} * 100\% \qquad \text{(Eqn. 5.20)}$$

Alignment Overlap Analysis

To assess for overlap of alignments, we conducted SmartMap on the simulated dataset with the read weight output setting activated. Using bedtools intersect, we then identified alignments that intersected with the true read origin in the Gold Standard dataset. From this, by weight, we were able to compute three metrics. First, we computed the number of alignments by weight that were present in the intersected dataset as a proportion of the total number of alignments by weight. Second, we computed the alignment weighted overlap proportion score, a measure of the proportion of a read's overall weight that overlaps with a given true origin of the read due to a given alignment. This is computed as the product of the weight of the alignment with the geometric mean of the proportion of overlap between the true read locus and the alignment locus. Finally, we computed the unweighted overlap proportion score, which is computed as the geometric mean of the proportion of overlap between the true read locus and the alignment locus.

Repetitive Element Analysis

Repetitive elements for hg38 were obtained from the HOMER list of repeats[178]. The promoter was defined as the most upstream portion of the annotated repeat. This dataset was used for analyzing all repeats; for analyzing LINE elements, SINE elements, or Simple Repeats, the corresponding subset of the repeats was used.

The HMD profiles in Fig. 5.15 and 5.18A were generated by computing the average HMD (from SmartMap analysis of AR16 and AR17) in 50bp windows from -1000bp to +1000bp relative to the promoter, with HMDs above 100% being set to 100% (because an HMD above 100% is definitionally impossible), and corresponding windows were averaged together to yield the average HMD profile for each set of elements.

To conduct clustering, first, the average HMD of the region -100bp to +100bp relative to each promoter in the relevant dataset was computed using the SmartMap analysis of AR16 and AR17, with HMDs above 100% being set to 100%. The data was then transformed to orthonormal basis by principal component analysis in R with scaling and centering. The resultant coordinate matrix used for k-means clustering, starting with 2 clusters and increasing the number of clusters until the decrease in total within-cluster sum of squares became markedly diminished; for each dataset (all repeats, LINE elements, SINE elements, and Simple Repeats), this occurred with 5 or more clusters and, accordingly, 4 clusters were used for each dataset.

RNA-seq analysis was conducted as follows. The average value of the RNA-seq SmartMap BEDGRAPH datasets were found across each LINE element. These values were then normalized to the SmartMap read count for each replicate (as average reads per million reads analyzed) and averaged to yield the average normalized read depth for each LINE element. These were then grouped by cluster and used to generate the quantile boxplots.

Heatmap Generation

Heatmaps of regions with nonzero HMD only in SmartMap analyses were generated as follows. The average HMD of the region -100bp to +100bp relative to each promoter was computed using both the uniread and SmartMap analyses of AR16 and AR17, with HMDs above 100% being set to 100%. Principal component analysis was conducted on the set of SmartMap HMDs in R with scaling and centering. Promoters with HMDs of zero in all of the uniread analyses and at least one nonzero SmartMap HMD were then selected and sorted by the first principal component. There were 142,392 such promoters. HMD profiles were then generated for each of the selected promoters as described above in 50bp windows from -1000bp to +1000bp relative to the promoter but were

not averaged together on corresponding windows. A field was added to the beginning of each row containing the value 100 as a calibration point for threshold adjustment.

The list of HMD profiles sorted on the first principal component was then imported into ImageJ as a Text Image. The height of the image was scaled down to 500pts with bilinear interpolation, and the thresholds were set from 0-100. The resultant image was exported as a PNG file, which was then opened in Photoshop in Indexed Color mode. The color table was then adjusted such that the lowest value was set to white and the highest value was set to the appropriate color. The leftmost point of the image (corresponding to the added field with the calibration point value of 100) was then removed from the image to generate the final heatmap.

Genome Browser Visualization

Genome browser visualization was conducted using Integrative Genomics Viewer (IGV)[334].

Comparison to Other Methods

Comparison to CSEM

Comparison was attempted against the CSEM software for multiread allocation[310] by only using the first read mate of our ICeChIP samples. However, the CSEM software returned a segmentation fault within the first minute of runtime, rendering comparison difficult.

Comparison to BM-Map

Comparison was attempted against the BM-Map software for multiread reweighting[311] by aligning the simulated read dataset with Bowtie2 per the settings used for SmartMap, followed by use of the BM-Map software with seven threads, the maximum permitted by the software. The first step of BM-Map (reading the alignments into memory) proceeded uneventfully, using one thread.

However, shortly after the second step of BM-Map began, the software returned an error and exited without returning an output. This was observed with existing binaries and with compilation of the software from source. As such, we were unable to compare to BM-Map.

Comparison to Iteration 0 and Random Alignments

The simulated dataset was aligned with Bowtie2 per the settings used for SmartMap. The reads were then parsed to yield a single extended BED file as per SmartMapPrep. For the Random Alignment selection analysis only, the reads were then split into separate files based on the number of alignments per read, and the random_read_selection.R script from the SmartMap-analysis GitHub repository was used to randomly select one alignment per read. These datasets were then used in the SmartMap software with the score set to -60.6. For the iteration 0 dataset, the number of reweighting cycles was set to zero; for the Random Alignments analysis, the number of reweighting cycles was set to one.

Comparison to Uniread

The simulated dataset was aligned with Bowtie2 per the settings used for SmartMap, with the modification that no value was specified for the option -k. Unireads were then parsed from the output SAM file by selecting for reads with MAPQ scores of: 3, 8, 23, 24, 40, 42; these are the MAPQ scores that are assigned to unireads by Bowtie2[172]. Reads were then parsed as per SmartMapPrep into a single extended BED file. This file was then used for SmartMap with one iteration, a minimum score of -60.6 and a maximum of one alignment per read.

Statistical Analyses

Statistical analysis for Fig. 5.18C was conducted first with chi-square analysis on full contingency table and with post-hoc tests on collapse contingency tables as follows. For each of the datasets in Fig. 5.18C, chi-square test for goodness-of-fit was conducted on the corresponding contingency table presented in Tables 4-6. The p-value for each of these tests was $p < 2.2x10^{-16}$ and, accordingly, post-hoc tests were conducted. The post-hoc tests consisted of collapsing each contingency table into a set of 2x2 contingency tables with the cluster of interest and family/type of interest compared to all other clusters and/or all other families within the contingency table. Chi-square goodness-of-fit tests were then conducted on each of these 2x2 contingency tables, and the p-values were Bonferroni corrected to adjust for the number of tests. These adjusted p-values for each 2x2 contingency table test were used to label the graphs in Fig. 5.18C as follows: $^{*}p < 0.01$, $^{**}p < 10^{-5}$, $^{***}p < 10^{-10}$.

Statistical analysis on Fig. 5.18D was conducted to compare median average normalized RNA-seq depth by cluster. Because the difference between cluster 3 and all of the other clusters appeared to be the most biologically meaningful, only pairwise comparisons were conducted between cluster 3 and the other clusters to limit the number of statistical comparisons and, accordingly, the degree of Bonferroni correction needed. Mood's median tests were solely conducted as pairwise comparisons between cluster 3 and each of the other clusters with Bonferroni correction to p-values with n=3 for Bonferroni correction. The adjusted p-values for each of these comparisons was $p < 2.2x10^{-16}$ and was marked appropriately on the graph.

CHAPTER 6: CONCLUSION

Histone modifications are critical epigenetic regulators, with important roles in maintaining transcriptional programs and, ultimately, helping to drive cellular identity and differentiation – or a lack thereof. To study these important marks, it is critical to be able to accurately and quantitatively measure their genomic distributions in order to identify the features with which they associate and to observe their changes across developmental or pathological states. Chromatin immunoprecipitation, as the method of choice for this task, is a powerful tool for probing the localization of a given histone modification genome-wide, but its canonical implementation has many problems that impact its accuracy and interpretability. The antibody employed may be of uncertain quality. The fragmentation and pulldown procedure may not enable high-quality IPs even with a highly specific antibody. The relative nature of traditional ChIP data may make it difficult to compare different cell types or cells with different treatment conditions. The next-generation sequencing backend may be unable to cover a large portion of the genome. As we have discussed through this work, these issues are all prevalent in the field and frequently result in major problems of interpretation that are often sufficiently severe to compromise the ultimate biological conclusions. Though ICeChIP can alleviate some of these problems, even it has limitations, and it is not yet universally employed. In this final chapter, we discuss some of the salient conclusions from this work and their implications for future practice.

Antibody specificity

In Chapter 2, we have primarily focused on the question of antibody specificity, its measurement, and the impacts of low-quality antibodies on interpretation and divination of biological meaning. We chose to focus on antibodies targeting the different methylation states of histone H3K4 – namely,

H3K4me1, H3K4me2, and H3K4me3. These different modifications are highly chemically similar; the methylation reaction consists of the replacement of a proton with a methyl group, a change that is charge-neutral and physically small in the context of the entire nucleosome (or even a small fragment therein). But despite this chemical similarity, these modifications have each canonically been associated with distinct functions, with H3K4me1 being held to mark enhancers[18,19,131], H3K4me2 being held to mark transcription factor binding sites[136], and H3K4me3 being held to mark transcriptionally active promoters[17,20,92–94,102,132], amongst many other proposed paradigms. Given that many of these conclusions were driven by chromatin immunoprecipitation, the question was whether antibodies could actually specifically discriminate between these highly similar modifications.

Fortunately, it seems that some antibodies were actually capable of this task. Though many antibodies displayed low specificity, there were many antibodies that were specific towards each methylform (Fig. 2.5). However, the most commonly used antibodies were often of low quality (Tables 2.1-2.4), including many employed (at great expense) by the ENCODE consortium in a broad variety of cellular contexts (Fig. 2.2-2.3). This is likely because of inadequate screening and antibody validation criteria. Most antibody validation procedures for ChIP involve screening by peptide arrays[108,110,112,129,143], in no small part because they provide the ability to screen a large number of modifications and combinations therein on a single plate (Fig. 2.9). However, as we showed, specificities measured by this method have little correlation with ChIP specificity (Fig. 2.7-2.8), making them less useful for validation of purported ChIP-grade antibodies.

This problem is not merely one of idle curiosity; low-quality antibodies have a demonstrable and material impact on ChIP-seq profiles. We showed that the antibodies with considerable off-target binding had excess pulldown relative to high-specificity antibodies, by developing and using a novel Fourier transform-based shape analysis method, had markedly different shapes of peaks

as compared to high-specificity antibodies (Fig. 2.10). These differences ultimately compromised the biological interpretations; as we highlighted, several studies employing low-quality antibodies came to conclusions that ultimately did not hold up to scrutiny with high-quality antibodies (Fig. 2.13-2.15). That is, ultimately, the cost of low specificity in the antibody reagents employed, and it is why the use of high-quality antibodies is of paramount importance for the field moving forward.

In recent years, particularly following the publication of our study on H3K4 methylation state antibodies[124], some antibody manufacturers and purveyors have begun employing nucleosome standard validation for their ChIP grade antibodies to validate their specificity for such an application; this is certainly an improvement in the field and one that should be employed by more manufacturers. However, even then, differences in ChIP conditions may result in differences in antibody specificity, and our work strongly suggests that researchers should take care to validate their own antibodies within the context in which they are employed.

Pulldown procedures

A high-quality antibody is necessary for a high-quality ChIP pulldown, but it is by no means sufficient. The other important factor in the specificity of the pulldown is the protocol employed. In Chapter 3 and Chapter 4, we describe the development of different pulldown protocols with distinct goals and advantages not afforded by traditional ChIP methods.

In Chapter 3, we developed a new pulldown protocol for sequential ChIP, in which the eluent of one IP is used as the substrate for a secondary IP. Here, we can already see the impact of the procedure on the IP quality; even when using the same high-quality antibodies, previously published sequential ChIP protocols had extremely low efficiency and often had low specificity (Fig.

3.1A), particularly as compared to the method that we developed involving a cleavable antibody (Fig. 3.2-3.3).

An even more dramatic representation of this protocol-dependent specificity change is found in Chapter 4, in which we examine the impact of different fragmentation and denaturation methods on the specificity and enrichment of the resulting pulldown. Indeed, the entire premise of this section is that native pulldown protocols are inadequate for assessing internal modifications, meaning that denaturative protocols are necessary. And as we observed, many of the denaturation methods, including sonication and many chemical denaturation methods, demonstrate high variability and low specificity (Fig. 4.5-4.7). It is only by employing thermal denaturation that we were able to achieve robust and specific pulldowns of the internal modification H3K79me2 (Fig. 4.8). This came at the cost of facile calibration (Fig. 4.11-4.12), but nonetheless enabled new inquiries into MLL-rearranged leukemias.

The ultimate takeaway from these studies is that the protocol employed matters considerably for the pulldown quality, and it is difficult to predict a priori whether a particular method will work or not. For example, it would not be unreasonable to hypothesize that any of the chemical denaturation methods employed in Fig. 4.7 would result in adequate denaturation of the nucleosome and a high-quality pulldown; it just so happened that this was not true. Ultimately, it is difficult to know for certain whether an IP method is actually viable or not, making it all the more critical to use internal standards to validate the method in situ.

Calibration

The other major reason to use internal standards to validate the method is for the purposes of calibration. Normalization and calibration to an exogenous reference is what enables comparisons

of ChIPs conducted in different cell types or in cells with different treatment conditions that can result in global changes in modification abundance. This exogenous reference does not, strictly speaking, have to be a set of semisynthetic nucleosome standards – as we showed in Chapter 4, the proper protocol can enable use of exogenous native chromatin as the normalization reference (Fig. 4.10). However, internal standards carry the advantage of enabling specificity measurement and making it possible to calibrate the experiment (particularly in native ICeChIP) to measure the histone modification density, enabling comparison of different IPs.

This calibration is consistently critical throughout this work. In Chapter 2, we find that previous studies about H3K4me3 at enhancers have been potentially flawed because of inadequate ability to compare H3K4me1 and H3K4me3 levels without the use of internal standard calibration (Fig. 2.13). This can also be detrimental to the ability to detect true differences; previous work on changes in H3K4me1 in cells with catalytically dead MLL3/MLL4 showed only a blunted change in H3K4me1 levels as compared to wild type cells, whereas ICeChIP revealed a much more pervasive loss of the same (Fig. 2.14). Similarly, and relevant to Chapter 4, Orlando et al. previously showed that exogenous reference normalization is necessary to detect global changes in H3K79me2 levels in response to treatment with Dot1L inhibitor[128], which we also find (Fig. 4.14).

However, the best example in this work on the importance of calibration for interpretation is presented in Chapter 3. The bivalency hypothesis was fundamentally based on finding regions that had both enrichment for H3K4me3 and H3K27me3, as measured by independent traditional ChIPs[9,130]. This is already problematic for H3K4me3 without the use of calibration, but this modification is at least distributed in a relatively peak-like manner, so regions with high absolute amounts of H3K4me3 are also likely to have high relative amounts of the same. This is not true for H3K27me3; this modification is both abundant and broadly distributed; as such, the regions with a

higher-than-baseline level of H3K27me3 (which will be the only ones detected without calibration) represent only a fraction of the H3K27me3 in the genome and ignore the remainder. To use an analogy, this can be though of as being similar to a forest, with some trees in the forest standing high above the canopy. Traditional ChIP methods, which only can find regions with relatively high amount of modification, would only detect these tallest trees and conclude that these are the only regions in which trees exist; calibration allows for the recognition that there are trees of different heights everywhere.

This was ultimately the problem for the bivalency hypothesis; because they only looked at regions with high H3K27me3 over baseline, previous studies were only able to identify a fraction of the regions of the genome that bear bivalent histone modifications (Fig. 3.6). Further, without the benefits of calibration, previous studies were unable to accurately compare bivalency levels in different cell types, meaning that they were unable to find that bivalency actually increased across differentiation rather than resolving (Fig. 3.4). There are other issues with the canonical bivalency hypothesis, as we explored in Chapter 3, but the fundamental problem in identifying bivalent regions boils down to a lack of calibration.

Calibration is not just useful for avoiding errors. As a quantitative metric, it also enables quantitative modeling to better study the impact of these modifications. In Chapter 2, for example, we used the quantitative measurements of H3K4me1, H3K4me2, and H3K4me3 to conduct modeling showing that the sum of HMDs of enhancers contacting a promoter is more associated with transcriptional control than is the average HMD (Fig. 2.12), suggesting that these enhancers may operate in concert with each other. Similarly, in Chapter 3, we use our quantitative ICeChIP data to predict DEG status and show that bivalency contributes minimal information content to such an endeavor. In Chapter 4, we use our quantitative data to identify differentially modified loci, which

we ultimately use as a comparison point against DEGs in MLL-rearranged leukemias. These are all tasks that would have been essentially impossible to conduct without quantitative data and is only possible because of the insights afforded by ICeChIP.

Collectively, this work shows both the dangers of using uncalibrated data to attempt to develop biological paradigms and shows the power of calibration to enable new quantitative analyses into the role of histone modifications in a broad range of contexts.

Next-generation sequencing read alignment

The final area that this work focuses on is the backend of the modern ChIP experiment: next-generation sequencing and its analysis. The most common method for handling reads that align to more than one locus in the genome is to systematically discard such ambiguously mapped reads. The rationale is that this ensures that the reads that remain are properly mapped, which is a fair concern, but it also results in systematic undersampling of less-mappable and highly repetitive regions of the genome (Fig. 5.2, 5.5). For obvious reasons, ignoring such a large portion of the genome is potentially problematic and, as we show, results in many repetitive regions being poorly mapped (Fig. 5.20-5.21).

To address this problem, we developed SmartMap[335], a tool that uses Bayesian reweighting of alignments to allocate reads of peak-type data that map ambiguously based on the distribution of the other reads in the dataset. This method successfully increased read depth genome-wide by up to 50% (Table 5.1-5.2), particularly at regions with lower mappability (Fig. 5.5), ultimately enabling new analyses of histone modification status and chromatin accessibility at repetitive elements (Fig. 5.18-5.21). SmartMap uses a dual-binary indexed tree structure for efficient updates and query of

the data, resulting in considerably greater efficiency and accuracy than other methods of multiread allocation (Table 5.8).

SmartMap represents a powerful tool for more completely using sequencing data to get a more truly genome-wide analysis. However, there are several directions in which SmartMap can be improved in the future. First, as we showed, it is currently not particularly useful for RNA-seq data (Fig. 5.17), in no small part because the algorithm is not compatible with large gaps such as those caused by introns; further tuning of the algorithm can address this problem. Further, SmartMap is presently not usable for trans-contact methods such as microC for a similar reason. This method can be adapted to this end, but would likely require modeling both of the contact probability as a function of distance as well as the modeling of the read distribution, adding another layer of complexity to the analysis. Nonetheless, there is nothing that, in principle, prevents the binary indexed tree data structures of SmartMap from being applied for these methods.

Significance

The problems with traditional ChIP-seq represent an existential problem for the field of molecular biology, and this work highlights the importance of using accurate and quantitative methods. As we emphasized repeatedly, many of the problems that are highlighted here occurred through no fault of the scientists conducting the previous studies; at the time, the tools that we can now use for better ChIP and ChIP-seq analysis simply did not exist, and the experimental design decisions employed in that historical context are reasonable. However, with the advent of new tools such as ICeChIP, and the broader availability of internal standards for experimental contexts more broadly, it is essential that the field changes its practices and raises its standards for data quality wholesale. Without such improvements, researchers will continue to make unreliable measurements, and they

will continue to draw incorrect conclusions from those data. But by employing the methods that are now available, researchers can make high-quality, quantitative measurements, and can ultimately use those data to drive new insights into the role of histone modifications. To not raise the standards of data quality where it is now possible is to forego those valuable opportunities for discovery.

APPENDIX A: ICECHIP PARAMETERS

This section lists antibody and chromatin used for the ICeChIP experiments in each data chapter.

Supplier Abbreviations

AB = Abcam; ABC = Abclonal; AM = Active Motif; CST = Cell Signaling Technologies; DIA = Diagenode; EMD = EMD Millipore; EPC = Epicypher; EPG = EpiGentek; KL = Koide Lab; REV = RevMAb; TF = Thermo Fisher.

Experiments from Chapter 2

Table A.1: ICeChIP Parameters from Chapter 2.

Target	Antibody ID	Antibody (µg)	Beads (µL)	Chromatin (µg)
H3K4me0	EMD 05-1341 Lot 2453179	3	12.5	3
H3K4me1	AB 8895 Lot GR305231-1	3	12.5	3
H3K4me1	ABC A2355 Lot 46694	3	12.5	3
H3K4me1	ABC A2355 Lot 46695	3	12.5	3
H3K4me1	AM 39297 Lot 01714002	3	12.5	3
H3K4me1	AM 39297 Lot 21008001	3	12.5	3
H3K4me1	AM 39635 Lot 30615011	3	12.5	3
H3K4me1	CST 5326 Lot 1	3	12.5	3
H3K4me1	CST 5326BF Lot 2	3	12.5	3
H3K4me1	DIA C15310037 Lot A399-001	3	12.5	3

Table A.1 continues on next page.

Table A.1, continued:

Target	Antibody ID	Antibody (µg)	Beads (µL)	Chromatin (µg)
H3K4me1	DIA C15410037 Lot A1657D	3	12.5	3
H3K4me1	DIA C15410194 Lot A1862D	3	12.5	3
H3K4me1	DIA C15410194 Lot A1863-001D	3	12.5	3
H3K4me1	EMD 07-436 Lot DAM1687548	3	12.5	3
H3K4me1	EPG A-4031-050 Lot 606359	3	12.5	3
H3K4me1	REV 31-1046-00 Lot P-01-00415	3	12.5	3
H3K4me1	TF 710795 Lot QL230603	3	12.5	3
H3K4me1	TF 720072 Lot RB226262	3	12.5	3
H3K4me2	AB 32356 Lot GR253788-9	3	12.5	3
H3K4me2	AB 7766 Lot GR289627-1	3	12.5	3
H3K4me2	ABC A2356 Lot 46696	3	12.5	3
H3K4me2	ABC A2356 Lot 46697	3	12.5	3
H3K4me2	AM 39141 Lot 01008001	3	12.5	3
H3K4me2	AM 39679 Lot 15515008	3	12.5	3
H3K4me2	CST 9725 Lot 9	3	12.5	3
H3K4me2	DIA C15200151 Lot 001-11	3	12.5	3
H3K4me2	DIA C15310035 Lot A391-001	3	12.5	3
H3K4me2	DIA C15410035 Lot A9360014P	3	12.5	3

Table A.1 continues on next page.

Table A.1, continued:

Target	Antibody ID	Antibody (µg)	Beads (µL)	Chromatin (µg)
H3K4me2	EMD 05-1338 Lot 2757107	3	12.5	3
H3K4me2	EMD 07-030 Lot DAM1479603	3	12.5	3
H3K4me2	EPC 13-0013 Lot 14247001	3	12.5	3
H3K4me2	EPG A-4032-050 Lot 606360	3	12.5	3
H3K4me2	TF 49-1004 Lot A391001161216	3	12.5	3
H3K4me2	TF 710796 Lot QL230606	3	12.5	3
H3K4me2	TF 720073 Lot QL226263	3	12.5	3
H3K4me3	AB 12209 Lot GR275790-1	3	12.5	3
H3K4me3	AB 8580 Lot GR190229-1	3	12.5	3
H3K4me3	AB 8580 Lot GR273043-4	3	12.5	3
H3K4me3	ABC A2357 Lot 46698	3	12.5	3
H3K4me3	ABC A2357 Lot 46699	3	12.5	3
H3K4me3	AM 39159 Lot 12613005	3	12.5	3
H3K4me3	AM 61379 Lot 24615006	3	12.5	3
H3K4me3	CST 9727 Lot 2	3	12.5	3
H3K4me3	CST 9751 Lot 9	3	12.5	3
H3K4me3	DIA C15200152 Lot 001-11	3	12.5	3
H3K4me3	DIA C15410003 Lot A1052D	3	12.5	3

Table A.1 continues on next page.

Table A.1, continued:

Target	Antibody ID	Antibody (µg)	Beads (µL)	Chromatin (µg)
H3K4me3	DIA C15410003 Lot A5051-001P	3	12.5	3
H3K4me3	EMD 05-745R Lot 2813867	3	12.5	3
H3K4me3	EMD 07-473 Lot DAM1623866	3	12.5	3
H3K4me3	EPC 13-0004 Lot 13171001	3	12.5	3
H3K4me3	EPG A-4033-050 Lot 606361	3	12.5	3
H3K4me3	REV 31-1039-00 Lot P-09-00676	3	12.5	3
H3K4me3	TF PA5-40086 Lot RL2301825	3	12.5	3
H3K4me3	KL 304M3B Lot 040416AG	3	60	3
H3K9me3	KL 309M3B Lot 072913TH	0.5	10	3

Experiments from Chapter 3

Table A.2: ICeChIP Parameters from Chapter 3.

Target	Antibody ID	Antibody (µg)	Beads (µL)	Chromatin (µg)
H3K4me3	KL 304M3B-1xHRV3C Lot 103015AG	3	60	3
H3K9me3	KL 309M3B Lot 072913TH	0.5	10	3
H3K27me3	CST 5326 Lot 8	0.6	5	0.8
H3K4me3 (reICeChIP)	KL 304M3B-1xHRV3C Lot 103015AG	3	60	3
H3K27me3 (reICeChIP)	CST 5326 Lot 8	0.6	5	Eluent of H3K4me3 IP

Experiments from Chapter 4

Table A.3: ICeChIP Parameters from Chapter 4.

Target	Antibody ID	Antibody (µg)	Beads (µL)	Chromatin (µg)
H3 CTD	EPC 13-0001 Lot 12346001	3	12.5	3
H3 CTD	EMD 05-928 Lot 2676583	3	12.5	3
H3K4me3	AM 39159 Lot 12613005	3	12.5	3
H3K27me3	CST 5326 Lot 8	0.6	5	0.8
H3K79me2	AB 3594 Lot GR173874	3	12.5	3
H3K79me2 (AR19 only)	CST 5427 Lot 4	3	12.5	3

Experiments from Chapter 5

Table A.4: ICeChIP Parameters from Chapter 5.

Target	Antibody ID	Antibody (µg)	Beads (µL)	Chromatin (µg)
H3K27me3	CST 5326 Lot 8	0.6	5	0.8

APPENDIX B: NUCLEOSOME BARCODE SEQUENCES

This section lists the sequences of the barcoded DNA applied to the nucleosome standards for use in ICeChIP.

601_CXXX Sequences

The 601_CXXX sequences are based on the Lowary and Widom 601 nucleosome binding sequence[168] with one barcode.

Table B.1: 601_CXXX nucleosome barcode sequences.

Barcode ID	Sequence
601_C002	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGCGCATAATAATCGCGC GATTTC
601_C005	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTCGTCGACGATCGTCGAA TCGTTC
601_C008	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTATACGCGTCGACGATTC GCGTTC
601_C009	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTCGCGTAATCGTTTCGAC GCGTTC
601_C010	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTTAACGTCGCGCGTTCGA ACGTTC
601_C013	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGTTTCGTACGCGCGACG TAATTC

Table B.1 continues on next page.

Table B.1, continued:

Barcode ID	Sequence
601_C014	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTCGCGTATACGTACGCGC GAATTC
601_C015	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGCGTAATACGCGCGAAA TTCGTC
601_C017	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGTACGAAACGCGTTAAC GTCGTC
601_C019	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTTACGCGTACCAACGCGT ATCGTC
601_C021	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTCGGTACGCTATCGTACG ATCGTC
601_C022	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGACGCGTATACGAATTT CGCGTC
601_C025	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTATCGCGTCGAGTGATAT CGCGTC
601_C026	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGCGTAATCGATACGTTA CGCGTC
601_C028	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTATTCGCGCGATCGCGAT TACGTC

Table B.1 continues on next page.

Table B.1, continued:

Barcode ID	Sequence
601_C029	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGATTACGCGAACGATTC GACGTC
601_C031	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTAGCGTACCGACGACGTT AACGTC
601_C032	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCATCGTCGACGAACGTTCG AACGTC
601_C033	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGAATCGACGATAGTTCG CGACTC
601_C034	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGCGACGTTAACGCGATA TCACTC
601_C037	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGCGTATCGGTCGCGTAA CGTATC
601_C038	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGACGAACGGTGTCGCGA ACTATC
601_C039	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGAACGGTCGTTTCGCGC GATATC
601_C040	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCCGACGATCGTACGACGC GATATC

Table B.1 continues on next page.

Table B.1, continued:

Barcode ID	Sequence
601_C041	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGACGTACCGTTTACGCG TCGATC
601_C042	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGTACGACGCTACGAACG TCGATC
601_C043	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCCGCGCGATATTTTCGTC GCGATC
601_C044	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTCGCGCGACATCGTAATC GCGATC
601_C046	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGCGTATTCGGTTCGTAC GCGATC
601_C047	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTCGATCGTCGGCGATCGT ACGATC
601_C049	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGACGTATCGGCGATACG ACGATC
601_C051	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGACGTAACGGACGCGAA ACGATC
601_C052	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCACGACCGTTCGCGTCGCG TTAATC

Table B.1 continues on next page.

Table B.1, continued:

Barcode ID	Sequence
601_C054	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCTCGTTCGTCGTTCGCGC GTAATC
601_C055	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCACCGTTCGTCGTCGACGC GTAATC
601_C056	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTACGTCCGTCGCGACGCG ATAATC
601_C058	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCACGGTACGTCGTTACGCG CGAATC
601_C060	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTCGCACGATCGCGCGATA CGAATC
601_C061	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTCGCCGAATCGACGCGTC GAAATC
601_C062	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTATGCGTCGCGTCGCGAC GAAATC
601_C063	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCATATCGCGCGCGTATCG CGGTTC
601_C066	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCCGAATACGCGTCGACGA CCGTTC

Table B.1 continues on next page.

Table B.1, continued:

Barcode ID	Sequence
601_C067	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTCGTACGACCGCGGTCGA ACGTTC
601_C068	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCAGCGTCGTACGTCGCGAC GAGTTC
601_C070	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGACCGATACGCGCGGTA CGATTC
601_C071	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTTCGAGCGACGCGGCGTA CGATTC
601_C073	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCGACGCGTAACGCCGCGCG TAATTC
601_C075	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGACGAACGAGTCGTATC GCGGTC
601_C076	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGTTACGCGTCTTATCGC GCGGTC
601_C077	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTAACGTCGCGCATTACGC GCGGTC
601_C078	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTACGCTCGGACTATACGC GCGGTC

Table B.1 continues on next page.

Table B.1, continued:

Barcode ID	Sequence
601_C079	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGTCGTTCGACACGACGT ACGGTC
601_C080	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCGCGCGACGTTACGATTCG ACGGTC
601_C081	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTGTCGCGCGTATACGCTC GTCGTC
601_C082	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCGTCCGAGCGTAGTATCGC GTCGTC
601_C083	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTCGCGACCGTAGTTACGC GTCGTC
601_C084	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGACGGACGTACGTATCC GTCGTC
601_C085	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGCGACGCATAGCGTTAC GTCGTC
601_C086	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCTACGCGTCGACGCGTTA GTCGTC
601_C087	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCCGACGATCGATCGGCGT ATCGTC

Table B.1 continues on next page.

Table B.1, continued:

Barcode ID	Sequence
601_C088	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGATCGTGCGACGCGACT ATCGTC
601_C089	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGATTCGGCGATGCGACG ATCGTC
601_C090	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTACGGTCGCGACCGTCGA ATCGTC
601_C091	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCATGTCGCGCGACGCGTCA ATCGTC
601_C092	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGGTCGTACGACGCGATA TGCGTC
601_C093	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTACGCGCGACACGTAATC GGCGTC
601_C094	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGTCGCTCGAATATCGGT CGCGTC
601_C096	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGCGTTACGCGCGATAGT CGCGTC
601_C097	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGCGTAACGCGGTCGTAT CGCGTC

Table B.1 continues on next page.

Table B.1, continued:

Barcode ID	Sequence
601_C098	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTCGGTACGCGCCGGATAT CGCGTC
601_C099	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCCGTCGAACGCCGCATAT CGCGTC
601_C100	CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCGCGCGTACCGATACCGAT CGCGTC

C001_CXXX Sequences

The C001_CXXX sequences are based on the Lowary and Widom 601 nucleosome binding sequence[168] with two barcodes.

Table B.2: 601_CXXX nucleosome barcode sequences.

Barcode ID	Sequence
C001_C006	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTCGTCGATTCGACGCGAA TCGTTC
C001_C008	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTATACGCGTCGACGATTC GCGTTC
C001_C009	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTCGCGTAATCGTTTCGAC GCGTTC

Table B.2 continues on next page.

Table B.2, continued:

Barcode ID	Sequence
C001_C010	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTTAACGTCGCGCGTTCGA ACGTTC
C001_C011	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGTATTACGCGAATCGCG CGATTC
C001_C014	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTCGCGTATACGTACGCGC GAATTC
C001_C015	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGCGTAATACGCGCGAAA TTCGTC
C001_C016	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCGATAGTCGACGTTATCGC GTCGTC
C001_C017	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGTACGAAACGCGTTAAC GTCGTC
C001_C018	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGTCGACTATCTCGTCGT ATCGTC
C001_C019	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTTACGCGTACCAACGCGT ATCGTC
C001_C022	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGACGCGTATACGAATTT CGCGTC

Table B.2 continues on next page.

Table B.2, continued:

Barcode ID	Sequence
C001_C023	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTCGACGCGATAATTACGT CGCGTC
C001_C024	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTCGCGCGAATATTCGTAT CGCGTC
C001_C025	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTATCGCGTCGAGTGATAT CGCGTC
C001_C028	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTATTCGCGCGATCGCGAT TACGTC
C001_C029	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGATTACGCGAACGATTC GACGTC
C001_C030	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGTATACGCGATTAACGC GACGTC
C001_C032	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCATCGTCGACGAACGTTCG AACGTC
C001_C034	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGCGACGTTAACGCGATA TCACTC
C001_C035	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCGGTACGCGTAACGCGTCG ATTATC

Table B.2 continues on next page.

Table B.2, continued:

Barcode ID	Sequence
C001_C036	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTCGCGACGTAAATTCGCG CGTATC
C001_C037	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGCGTATCGGTCGCGTAA CGTATC
C001_C038	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGACGAACGGTGTCGCGA ACTATC
C001_C039	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGAACGGTCGTTTCGCGC GATATC
C001_C040	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCCGACGATCGTACGACGC GATATC
C001_C041	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGACGTACCGTTTACGCG TCGATC
C001_C042	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGTACGACGCTACGAACG TCGATC
C001_C043	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCCGCGCGATATTTTCGTC GCGATC
C001_C044	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTCGCGCGACATCGTAATC GCGATC

Table B.2 continues on next page.

Table B.2, continued:

Barcode ID	Sequence
C001_C047	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTCGATCGTCGGCGATCGT ACGATC
C001_C048	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCACGATCGTCGGTCGTTCG ACGATC
C001_C049	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGACGTATCGGCGATACG ACGATC
C001_C050	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCATATCGCGCGGTCGTCGA ACGATC
C001_C051	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGACGTAACGGACGCGAA ACGATC
C001_C052	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCACGACCGTTCGCGTCGCG TTAATC
C001_C053	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGTATCGGTCGCGATCGC GTAATC
C001_C055	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCACCGTTCGTCGTCGACGC GTAATC
C001_C056	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTACGTCCGTCGCGACGCG ATAATC

Table B.2 continues on next page.

Table B.2, continued:

Barcode ID	Sequence
C001_C057	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCCGTTACGTCGTATCGCG CGAATC
C001_C058	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCACGGTACGTCGTTACGCG CGAATC
C001_C060	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTCGCACGATCGCGCGATA CGAATC
C001_C061	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTCGCCGAATCGACGCGTC GAAATC
C001_C063	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCATATCGCGCGCGTATCG CGGTTC
C001_C064	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGTATAGCGCGCCGTACG TCGTTC
C001_C065	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCACCGATACGCGTAGCGAC GCGTTC
C001_C066	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCCGAATACGCGTCGACGA CCGTTC
C001_C070	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGACCGATACGCGCGGTA CGATTC

Table B.2 continues on next page.

Table B.2, continued:

Barcode ID	Sequence
C001_C071	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTTCGAGCGACGCGGCGTA CGATTC
C001_C072	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCGTCGAACGACGCGGTCGA CGATTC
C001_C073	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCGACGCGTAACGCCGCGCG TAATTC
C001_C074	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTCGACGCGTAGCGCGACG CAATTC
C001_C075	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGACGAACGAGTCGTATC GCGGTC
C001_C077	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTAACGTCGCGCATTACGC GCGGTC
C001_C079	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGTCGTTCGACACGACGT ACGGTC
C001_C081	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTGTCGCGCGTATACGCTC GTCGTC
C001_C082	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCGTCCGAGCGTAGTATCGC GTCGTC

Table B.2 continues on next page.

Table B.2, continued:

Barcode ID	Sequence
C001_C085	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGCGACGCATAGCGTTAC GTCGTC
C001_C089	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGATTCGGCGATGCGACG ATCGTC
C001_C090	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTACGGTCGCGACCGTCGA ATCGTC
C001_C091	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCATGTCGCGCGACGCGTCA ATCGTC
C001_C092	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGGTCGTACGACGCGATA TGCGTC
C001_C093	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCTACGCGCGACACGTAATC GGCGTC
C001_C094	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACAGCTC TAGCACCGCTTAAACGCACGTACGCGCTGTCCCCCGCGTTTTAACCG CCAAGGGGATTACTCCCTAGTCTCCAGGCCGTCGCTCGAATATCGGT CGCGTC

MMTV_CXXX Sequences

The MMTV_CXXX sequences are based on the mouse mammary tumor virus (MMTV) long terminal repeat (LTR)[288] with one barcode.

Table B.3: MMTV_CXXX nucleosome barcode sequences.

Barcode ID	Sequence
MMTV_C001	GAAACGCGTATCGCGCGCATAATACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C002	GAAATCGCGCGATTATTATGCGCGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C003	GAACGAACGTCGAACGCGCGATATCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C004	GAACGACGCGATAATATCGCGCGTCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C005	GAACGATTCGACGATCGTCGACGACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C006	GAACGATTCGCGTCGAATCGACGACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C007	GAACGCGAAACGACGAATCGCGTACTCTTGTGTGTTTGTGTCTGTTC GCCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCC CCCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTT TTTTG
MMTV_C008	GAACGCGAATCGTCGACGCGTATACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C009	GAACGCGTCGAAACGATTACGCGACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG

Table B.3 continues on next page.

Table B.3, continued:

Barcode ID	Sequence
MMTV_C010	GAACGTTCGAACGCGCGACGTTAACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C011	GAATCGCGCGATTCGCGTAATACGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C012	GAATTACGCGCGACGCGTAATCGACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C013	GAATTACGTCGCGCGTACGAAACGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C014	GAATTCGCGCGTACGTATACGCGACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C015	GACGAATTTCGCGCGTATTACGCGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C016	GACGACGCGATAACGTCGACTATCCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C017	GACGACGTTAACGCGTTTCGTACGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C018	GACGATACGACGAGATAGTCGACGCTCTTGTGTGTTTGTGTCTGTTC GCCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCC CCCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTT TTTTG

Table B.3 continues on next page.

Table B.3, continued:

Barcode ID	Sequence
MMTV_C019	GACGATACGCGTTGGTACGCGTAACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C020	GACGATCGCGTAATACGCGATTCGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C021	GACGATCGTACGATAGCGTACCGACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C022	GACGCGAAATTCGTATACGCGTCGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C023	GACGCGACGTAATTATCGCGTCGACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C024	GACGCGATACGAATATTCGCGCGACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C025	GACGCGATATCACTCGACGCGATACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C026	GACGCGTAACGTATCGATTACGCGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C027	GACGCGTCGATTATCGCGACGTAACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG

Table B.3 continues on next page.

Table B.3, continued:

Barcode ID	Sequence
MMTV_C028	GACGTAATCGCGATCGCGCGAATACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C029	GACGTCGAATCGTTCGCGTAATCGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C030	GACGTCGCGTTAATCGCGTATACGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C031	GACGTTAACGTCGTCGGTACGCTACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C032	GACGTTCGAACGTTCGTCGACGATCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C033	GAGTCGCGAACTATCGTCGATTCGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C034	GAGTGATATCGCGTTAACGTCGCGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C035	GATAATCGACGCGTTACGCGTACCCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C036	GATACGCGCGAATTTACGTCGCGACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG

Table B.3 continues on next page.

Table B.3, continued:

Barcode ID	Sequence
MMTV_C037	GATACGTTACGCGACCGATACGCGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C038	GATAGTTCGCGACACCGTTCGTCGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C039	GATATCGCGCGAAACGACCGTTCGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C040	GATATCGCGTCGTACGATCGTCGGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C041	GATCGACGCGTAAACGGTACGTCGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C042	GATCGACGTTCGTAGCGTCGTACGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C043	GATCGCGACGAAAATATCGCGCGGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C044	GATCGCGATTACGATGTCGCGCGACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C045	GATCGCGCGTAATCATATCGCGCGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG

Table B.3 continues on next page.

Table B.3, continued:

Barcode ID	Sequence
MMTV_C046	GATCGCGTACGAACCGAATACGCGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C047	GATCGTACGATCGCCGACGATCGACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C048	GATCGTCGAACGACCGACGATCGTCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C049	GATCGTCGTATCGCCGATACGTCGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C050	GATCGTTCGACGACCGCGCGATATCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C051	GATCGTTTCGCGTCCGTTACGTCGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C052	GATTAACGCGACGCGAACGGTCGTCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C053	GATTACGCGATCGCGACCGATACGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C054	GATTACGCGCGAACGACGAACGAGCTCTTGTGTGTTTGTGTCTGTTC GCCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCC CCCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTT TTTTG

Table B.3 continues on next page.

Table B.3, continued:

Barcode ID	Sequence
MMTV_C055	GATTACGCGTCGACGACGAACGGTCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C056	GATTATCGCGTCGCGACGGACGTACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C057	GATTCGCGCGATACGACGTAACGGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C058	GATTCGCGCGTAACGACGTACCGTCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C059	GATTCGTACGCGACGACGTATCGGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C060	GATTCGTATCGCGCGATCGTGCGACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C061	GATTTCGACGCGTCGATTCGGCGACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C062	GATTTCGTCGCGACGCGACGCATACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C063	GAACCGCGATACGCGCGCGATATGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG

Table B.3 continues on next page.

Table B.3, continued:

Barcode ID	Sequence
MMTV_C064	GAACGACGTACGGCGCGCTATACGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C065	GAACGCGTCGCTACGCGTATCGGTCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C066	GAACGGTCGTCGACGCGTATTCGGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C067	GAACGTTCGACCGCGGTCGTACGACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C068	GAACTCGTCGCGACGTACGACGCTCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C069	GAATCGCGGTACGCGTATAGCGCGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C070	GAATCGTACCGCGCGTATCGGTCGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C071	GAATCGTACGCCGCGTCGCTCGAACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C072	GAATCGTCGACCGCGTCGTTCGACCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG

Table B.3 continues on next page.

Table B.3, continued:

Barcode ID	Sequence
MMTV_C073	GAATTACGCGCGGCGTTACGCGTCCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C074	GAATTGCGTCGCGCTACGCGTCGACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C075	GACCGCGATACGACTCGTTCGTCGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C076	GACCGCGCGATAAGACGCGTAACGCTCTTGTGTGTTTGTGTCTGTTC GCCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCC CCCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTT TTTTG
MMTV_C077	GACCGCGCGTAATGCGCGACGTTACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C078	GACCGCGCGTATAGTCCGAGCGTACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C079	GACCGTACGTCGTGTCGAACGACGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C080	GACCGTCGAATCGTAACGTCGCGCCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C081	GACGACGAGCGTATACGCGCGACACTCTTGTGTGTTTGTGTCTGTTC GCCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCC CCCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTT TTTTG

Table B.3 continues on next page.

Table B.3, continued:

Barcode ID	Sequence
MMTV_C082	GACGACGCGATACTACGCTCGGACCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C083	GACGACGCGTAACTACGGTCGCGACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C084	GACGACGGATACGTACGTCCGTCGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C085	GACGACGTAACGCTATGCGTCGCGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C086	GACGACTAACGCGTCGACGCGTAGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C087	GACGATACGCCGATCGATCGTCGGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C088	GACGATAGTCGCGTCGCACGATCGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C089	GACGATCGTCGCATCGCCGAATCGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C090	GACGATTCGACGGTCGCGACCGTACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG

Table B.3 continues on next page.

Table B.3, continued:

Barcode ID	Sequence
MMTV_C091	GACGATTGACGCGTCGCGCGACATCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C092	GACGCATATCGCGTCGTACGACCGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C093	GACGCCGATTACGTGTCGCGCGTACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C094	GACGCGACCGATATTCGAGCGACGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C095	GACGCGACGCAATCCGTCGAACGCCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C096	GACGCGACTATCGCGCGTAACGCGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C097	GACGCGATACGACCGCGTTACGCGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C098	GACGCGATATCCGGCGCGTACCGACTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG
MMTV_C099	GACGCGATATGCGGCGTTCGACGGCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG

Table B.3 continues on next page.

Table B.3, continued:

Barcode ID	Sequence
MMTV_C100	GACGCGATCGGTATCGGTACGCGCCTCTTGTGTGTTTGTGTCTGTTCG CCATCCCGTCTCCGCTCGTCACTTATCCTTCACTTTCCAGAGGGTCCC CCCGCAGACCCCGGCGACCCTCAGGTCGGCCGACTGCGGCACAGTTT TTTG

MMS_DXXX Sequences

The MMS_DXXX sequences are based on the mouse minor satellite (MMS) sequence[289] with one barcode.

Table B.4: MMS_DXXX nucleosome barcode sequences.

Barcode ID	Sequence
MMS_D001	ATGATATTCGTACCCGTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D002	ATGATAACGTAGACCGTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D003	ATGTAGTTCGTACGACTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D004	ATGGAAGCGAACGTATTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D005	ATGACGTCGACTATTATTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT

Table B.4 continues on next page.

Table B.4, continued:

Barcode ID	Sequence
MMS_D006	ATGCGCGATTAGACTATTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D007	ATGATGGTACGCGATTTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D008	ATGTAGATCGCGTAAGTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D009	ATGTCTAGTAACGACGTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D010	ATGTTATACCTCGCGTTTTGTAGAACAGTGTATATCAATGAGTTACAA TGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTAG ATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACGA ATGTGTTT
MMS_D011	ATGAATACGCGCGTAATTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D012	ATGGCGTTATCGTACATTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D013	ATGTGTTTAGCGAACGTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D014	ATGAGATTATCGACCGTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT

Table B.4 continues on next page.

Table B.4, continued:

Barcode ID	Sequence
MMS_D015	ATGTATAGTACGCGTCTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D016	ATGTCTATTCGGCGTATTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D017	ATGCGTCGATAACCTATTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D018	ATGCTTCGATACGTAATTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D019	ATGTCGTAACGCGAATTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D020	ATGATCGCTCTAACGTTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D021	ATGCTTATCGCGTTGATTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D022	ATGTCGTTACGTCCTATTTGTAGAACAGTGTATATCAATGAGTTACAA TGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTAG ATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACGA ATGTGTTT
MMS_D023	ATGTGAACGTCGTAGTTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT

Table B.4 continues on next page.

Table B.4, continued:

Barcode ID	Sequence
MMS_D024	ATGCGTTATACACGACTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D025	ATGTCGTACGTTAGACTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D026	ATGAACGACGGTACATTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D027	ATGTACGACGTAAGGTTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D028	ATGTACTATCGTCACGTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D029	ATGACTACGCTACGATTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D030	ATGTAATCGCGCTAACTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D031	ATGATTTAGGCGTACGTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D032	ATGTCGATAGCGTAAGTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT

Table B.4 continues on next page.

Table B.4, continued:

Barcode ID	Sequence
MMS_D033	ATGCGCGTTAGATAGTTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D034	ATGCGGTTACGCTATATTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D035	ATGTATCGCTAACTCGTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D036	ATGCGCGTAATAGTACTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D037	ATGCGTACGCTATCTATTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D038	ATGCCGCGAACTTATATTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D039	ATGTTACAATACGCGCTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D040	ATGTAGTTTACGCGAGTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D041	ATGCTCGAATTGACGTTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT

Table B.4 continues on next page.

Table B.4, continued:

Barcode ID	Sequence
MMS_D042	ATGCGTCGTACTACATTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D043	ATGCGTAATACCTACGTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D044	ATGTCATTACGATCGCTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D045	ATGGTAATGCGCGATATTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D046	ATGCGCGAATACTAAGTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D047	ATGTAACGTCCGGTAATTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D048	ATGTATTCGTATCCCGTTTGTAGAACAGTGTATATCAATGAGTTACAA TGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTAG ATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACGA ATGTGTTT
MMS_D049	ATGTAGTAACGTCGAGTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D050	ATGCCGTTATAGTACGTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT

Table B.4 continues on next page.

Table B.4, continued:

Barcode ID	Sequence
MMS_D051	ATGGATAACGCGAAACTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D052	ATGACGTAGGTATTCGTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D053	ATGCGTACTTTAGACGTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D054	ATGGAATACGCGAATCTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D055	ATGCAGTATTCGCGTATTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D056	ATGCGTACTAATCGTCTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D057	ATGGATCGCGTACTATTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D058	ATGATACGCGATGTATTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D059	ATGTTCAATACGCGACTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT

Table B.4 continues on next page.

Table B.4, continued:

Barcode ID	Sequence
MMS_D060	ATGCGAAAGACGTATCTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D061	ATGACGCCGTAATAGTTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D062	ATGCGATCGCGTATTATTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D063	ATGAGACCGATTAACGTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D064	ATGGTTCGGACGTAATTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D065	ATGAGATAGCGACGTATTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D066	ATGTATAGTATCGCGATTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D067	ATGATACTACGCCGATTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D068	ATGTATCGCGAACTTATTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT

Table B.4 continues on next page.

Table B.4, continued:

Barcode ID	Sequence
MMS_D069	ATGCTATCGAGCGATATTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D070	ATGACGTTCGAACTAGTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D071	ATGTATCGAATACGGCTTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT
MMS_D072	ATGGCGAACGTAGTTATTTGTAGAACAGTGTATATCAATGAGTTACA ATGAGAAACATGGAAAATGATAAAAACCACACTGGAGAACAGATTA GATGAGTGAGTTACACTGAAATACTACGTATCGTCCCGTTTCCAACG AATGTGTTT

Space Alien Sequences

The Space Alien (SA) sequences are synthetically designed.

Table B.5: Space Alien nucleosome barcode sequences.

Barcode ID	Sequence
S001	ATTAGCGACGTGATAATCTTTTAGACTACGTCGTGTCGCGTATAACC CGACACGTAATCGACACTACGTCGACTACGACGTGTTAGTAAAATAC GTCGCGATATTAACCGGCGTTCATTGCGCGATATAACGCGCGTCATA CTATTA
S002	ATTAGCGACGTGATAATCTTCGTACTAAGACGTATCTAGCGCGATAC CGACACGTAATCGACACTACGTCGACTATTCACGACGTATAACGTCG TACTAACTATTAACCGGCGTTCATTGCGCGATATAACGCGCGTCATA CTATTA
S003	ATTAGCGACGTGATAATCTTCGCGAGTAATAAGTACGCGAGATAGTC CGACACGTAATCGACACTACGTCGACTATACGCGATACTCATAGTAT TTCGCGATATTAACCGGCGTTCATTGCGCGATATAACGCGCGTCATA CTATTA

Table B.5 continues on next page.

Table B.5, continued:

Barcode ID	Sequence
S004	ATTAGCGACGTGATAATCTTACGAATAACGCGTCGCTATACTTCGACCGACACGTAATCGACACTACGTCGACTATACGTCGTAGAGACTACGGTCGATATTATTAACCGGCGTTCATTGCGCGATATAACGCGCGTCATACTATTA
S005	ATTAGCGACGTGATAATCTTATACGTAACCGGTAGACTTATACGCGCCGACACGTAATCGACACTACGTCGACTACGAATTACGTCGTCGTATTCGCGTTATATTAACCGGCGTTCATTGCGCGATATAACGCGCGTCATACTATTA
S006	ATTAGCGACGTGATAATCTTGATTACGACCGTTTATTCGCGAACCACCGACACGTAATCGACACTACGTCGACTAGATATATCGACCGTTACGTTCGCGATTATTAACCGGCGTTCATTGCGCGATATAACGCGCGTCATACTATTA
S007	ATTAGCGACGTGATAATCTTCGCGTTAAGTATGCGAACCGTATAGACCGACACGTAATCGACACTACGTCGACTATCTTTTCGGCGTATAGACCGCGAATATATTAACCGGCGTTCATTGCGCGATATAACGCGCGTCATACTATTA
S008	ATTAGCGACGTGATAATCTTTAGACGACCGAATTCTACTATTCGCGCCGACACGTAATCGACACTACGTCGACTAACACTATCGCGAATTATGTTCGGACGTATTAACCGGCGTTCATTGCGCGATATAACGCGCGTCATACTATTA
S009	ATTAGCGACGTGATAATCTTACGTACTACGATCTCGACGCGTAAAACCGACACGTAATCGACACTACGTCGACTAGTCGTATAACGGTAGTATACTCGCGATATTAACCGGCGTTCATTGCGCGATATAACGCGCGTCATACTATTA
S010	ATTAGCGACGTGATAATCTTCTTATTTCGCGTCACGACTAATTCCGCCGACACGTAATCGACACTACGTCGACTAGACGTAGTTACGTTTGTATACCGCGATATTAACCGGCGTTCATTGCGCGATATAACGCGCGTCATACTATTA
S011	ATTAGCGACGTGATAATCTTCGCTATACGAGAATAACGCGTCGTAACCGACACGTAATCGACACTACGTCGACTATAATCGCACGGTACATTACTCGCGAATATTAACCGGCGTTCATTGCGCGATATAACGCGCGTCATACTATTA
S012	ATTAGCGACGTGATAATCTTATAACGAGACCGAGTTCGCTTATACGCCGACACGTAATCGACACTACGTCGACTAAGGTACGACGATAATCCTACGCGTAATATTAACCGGCGTTCATTGCGCGATATAACGCGCGTCATACTATTA

Table B.5 continues on next page.

Table B.5, continued:

Barcode ID	Sequence
S013	ATTAGCGACGTGATAATCTTCGATGTATCGTAGTCGGAGTACGTAAC CGACACGTAATCGACACTACGTCGACTATCGTATACTCCGATTACGC GACGTTATATTAACCGGCGTTCATTGCGCGATATAACGCGCGTCATA CTATTA
S014	ATTAGCGACGTGATAATCTTCGTAACGCGTTTAGAGTATTCGTACGC CGACACGTAATCGACACTACGTCGACTACGCGACGTATTATATAAGT CGCGTACTATTAACCGGCGTTCATTGCGCGATATAACGCGCGTCATA CTATTA
S015	ATTAGCGACGTGATAATCTTCCTTACGCGAATTCGAACTAATCACGC CGACACGTAATCGACACTACGTCGACTATTCGCGATAGTGTACCGTA AGTTCGTTATTAACCGGCGTTCATTGCGCGATATAACGCGCGTCATA CTATTA
S016	ATTAGCGACGTGATAATCTTATGATACGTCCGATACGCGTATTCGTC CGACACGTAATCGACACTACGTCGACTAAGTCAATACGCGATATAGA CGTTGCGTATTAACCGGCGTTCATTGCGCGATATAACGCGCGTCATA CTATTA
S017	ATTAGCGACGTGATAATCTTCGTACGAATTACCTTACCGTCGATTGCC GACACGTAATCGACACTACGTCGACTACGTAATATCGAGGTACTACG TCGAAATATTAACCGGCGTTCATTGCGCGATATAACGCGCGTCATAC TATTA
S018	ATTAGCGACGTGATAATCTTCATAACGGTTCGACGTACCGATGTAAC CGACACGTAATCGACACTACGTCGACTATACATCGCGTCATCGCCGA ACTATAATATTAACCGGCGTTCATTGCGCGATATAACGCGCGTCATA CTATTA
S019	ATTAGCGACGTGATAATCTTACCATTACGCGATAACTACGCACGATC CGACACGTAATCGACACTACGTCGACTAATACCGTCGTAATTTAGGT CGTCGTATATTAACCGGCGTTCATTGCGCGATATAACGCGCGTCATA CTATTA
S020	ATTAGCGACGTGATAATCTTTACGCGCACTAATGTATCGACCGTTAC CGACACGTAATCGACACTACGTCGACTAGTCGTAACGTACTACGTCT CGACATATATTAACCGGCGTTCATTGCGCGATATAACGCGCGTCATA CTATTA
S021	ATTAGCGACGTGATAATCTTCGATTAGTACTCGAATACGCTACCGTC CGACACGTAATCGACACTACGTCGACTAACTTACGTCCGTATATGTA CGGATCGTATTAACCGGCGTTCATTGCGCGATATAACGCGCGTCATA CTATTA

Table B.5 continues on next page.

Table B.5, continued:

Barcode ID	Sequence
S022	ATTAGCGACGTGATAATCTTTACGTCGGATACATATCCGCGAACTAC CGACACGTAATCGACACTACGTCGACTAATACGTCGGATTGCCGATA CTACGTATATTAACCGGCGTTCATTGCGCGATATAACGCGCGTCATA CTATTA
S023	ATTAGCGACGTGATAATCTTTATTCGATGCGGTGATTACTACGCGAC CGACACGTAATCGACACTACGTCGACTATACGGTCGTTTACAGGTCG TATCGTTTATTAACCGGCGTTCATTGCGCGATATAACGCGCGTCATAC TATTA
S024	ATTAGCGACGTGATAATCTTTCCGTAAACGACAGACGATCTCGTAAC CGACACGTAATCGACACTACGTCGACTATCGCGTCGTATTACATAAC GTTGTCGTATTAACCGGCGTTCATTGCGCGATATAACGCGCGTCATA CTATTA
S025	ATTAGCGACGTGATAATCTTTCGACGAACCTTATCGTGTAACTACGC CGACACGTAATCGACACTACGTCGACTATCGTGTCTCGATAACTATT ACTCGCGTATTAACCGGCGTTCATTGCGCGATATAACGCGCGTCATA CTATTA

APPENDIX C: PRIMERS AND PROBES

This section lists the primers used for mutagenesis of barcode sequences, as well as the primers and probes used for qPCR of genomic targets and barcode sequences.

Primers for barcode mutagenesis and amplification

Mutagenesis of 601_CXXX barcodes

Table C.1: 601_CXXX mutagenesis primers.

Barcode	Forward Primer	Reverse Primer
601 Base	CTGGAGAATCCCGGTGC	ACAGGATGTATATATCTGACACG TG
601_C002	CTGGAGAATCCCGGTGC	GAAATCGCGCGATTATTATGCGC GGCCTGGAGACTAGGGAG
601_C005	CTGGAGAATCCCGGTGC	GAACGATTCGACGATCGTCGACG AGCCTGGAGACTAGGGAG
601_C008	CTGGAGAATCCCGGTGC	GAACGCGAATCGTCGACGCGTAT AGCCTGGAGACTAGGGAG
601_C009	CTGGAGAATCCCGGTGC	GAACGCGTCGAAACGATTACGC GAGCCTGGAGACTAGGGAG
601_C010	CTGGAGAATCCCGGTGC	GAACGTTCGAACGCGCGACGTTA AGCCTGGAGACTAGGGAG
601_C013	CTGGAGAATCCCGGTGC	GAATTACGTCGCGCGTACGAAAC GGCCTGGAGACTAGGGAG
601_C014	CTGGAGAATCCCGGTGC	GAATTCGCGCGTACGTATACGCG AGCCTGGAGACTAGGGAG
601_C015	CTGGAGAATCCCGGTGC	GACGAATTTCGCGCGTATTACGC GGCCTGGAGACTAGGGAG
601_C017	CTGGAGAATCCCGGTGC	GACGACGTTAACGCGTTTCGTAC GGCCTGGAGACTAGGGAG
601_C019	CTGGAGAATCCCGGTGC	GACGATACGCGTTGGTACGCGTA AGCCTGGAGACTAGGGAG
601_C021	CTGGAGAATCCCGGTGC	GACGATCGTACGATAGCGTACCG AGCCTGGAGACTAGGGAG

Table C.1 continues on next page.

Table C.1, continued:

Barcode	Forward Primer	Reverse Primer
601_C022	CTGGAGAATCCCGGTGC	GACGCGAAATTCGTATACGCGTC GGCCTGGAGACTAGGGAG
601_C025	CTGGAGAATCCCGGTGC	GACGCGATATCACTCGACGCGAT AGCCTGGAGACTAGGGAG
601_C026	CTGGAGAATCCCGGTGC	GACGCGTAACGTATCGATTACGC GGCCTGGAGACTAGGGAG
601_C028	CTGGAGAATCCCGGTGC	GACGTAATCGCGATCGCGCGAAT AGCCTGGAGACTAGGGAG
601_C029	CTGGAGAATCCCGGTGC	GACGTCGAATCGTTCGCGTAATC GGCCTGGAGACTAGGGAG
601_C031	CTGGAGAATCCCGGTGC	GACGTTAACGTCGTCGGTACGCT AGCCTGGAGACTAGGGAG
601_C032	CTGGAGAATCCCGGTGC	GACGTTCGAACGTTCGTCGACGA TGCCTGGAGACTAGGGAG
601_C033	CTGGAGAATCCCGGTGC	GAGTCGCGAACTATCGTCGATTC GGCCTGGAGACTAGGGAG
601_C034	CTGGAGAATCCCGGTGC	GAGTGATATCGCGTTAACGTCGC GGCCTGGAGACTAGGGAG
601_C037	CTGGAGAATCCCGGTGC	GATACGTTACGCGACCGATACGC GGCCTGGAGACTAGGGAG
601_C038	CTGGAGAATCCCGGTGC	GATAGTTCGCGACACCGTTCGTC GGCCTGGAGACTAGGGAG
601_C039	CTGGAGAATCCCGGTGC	GATATCGCGCGAAACGACCGTTC GGCCTGGAGACTAGGGAG
601_C040	CTGGAGAATCCCGGTGC	GATATCGCGTCGTACGATCGTCG GGCCTGGAGACTAGGGAG
601_C041	CTGGAGAATCCCGGTGC	GATCGACGCGTAAACGGTACGTC GGCCTGGAGACTAGGGAG
601_C042	CTGGAGAATCCCGGTGC	GATCGACGTTCGTAGCGTCGTAC GGCCTGGAGACTAGGGAG
601_C043	CTGGAGAATCCCGGTGC	GATCGCGACGAAAATATCGCGC GGGCCTGGAGACTAGGGAG
601_C044	CTGGAGAATCCCGGTGC	GATCGCGATTACGATGTCGCGCG AGCCTGGAGACTAGGGAG
601_C046	CTGGAGAATCCCGGTGC	GATCGCGTACGAACCGAATACG CGGCCTGGAGACTAGGGAG

Table C.1 continues on next page.

Table C.1, continued:

Barcode	Forward Primer	Reverse Primer
601_C047	CTGGAGAATCCCGGTGC	GATCGTACGATCGCCGACGATCG AGCCTGGAGACTAGGGAG
601_C049	CTGGAGAATCCCGGTGC	GATCGTCGTATCGCCGATACGTC GGCCTGGAGACTAGGGAG
601_C051	CTGGAGAATCCCGGTGC	GATCGTTTCGCGTCCGTTACGTC GGCCTGGAGACTAGGGAG
601_C052	CTGGAGAATCCCGGTGC	GATTAACGCGACGCGAACGGTC GTGCCTGGAGACTAGGGAG
601_C054	CTGGAGAATCCCGGTGC	GATTACGCGCGAACGACGAACG AGGCCTGGAGACTAGGGAG
601_C055	CTGGAGAATCCCGGTGC	GATTACGCGTCGACGACGAACG GTGCCTGGAGACTAGGGAG
601_C056	CTGGAGAATCCCGGTGC	GATTATCGCGTCGCGACGGACGT AGCCTGGAGACTAGGGAG
601_C058	CTGGAGAATCCCGGTGC	GATTCGCGCGTAACGACGTACCG TGCCTGGAGACTAGGGAG
601_C060	CTGGAGAATCCCGGTGC	GATTCGTATCGCGCGATCGTGCG AGCCTGGAGACTAGGGAG
601_C061	CTGGAGAATCCCGGTGC	GATTTCGACGCGTCGATTCGGCG AGCCTGGAGACTAGGGAG
601_C062	CTGGAGAATCCCGGTGC	GATTTCGTCGCGACGCGACGCAT AGCCTGGAGACTAGGGAG
601_C063	CTGGAGAATCCCGGTGC	GAACCGCGATACGCGCGCGATAT GGCCTGGAGACTAGGGAG
601_C066	CTGGAGAATCCCGGTGC	GAACGGTCGTCGACGCGTATTCG GGCCTGGAGACTAGGGAG
601_C067	CTGGAGAATCCCGGTGC	GAACGTTCGACCGCGGTCGTACG AGCCTGGAGACTAGGGAG
601_C068	CTGGAGAATCCCGGTGC	GAACTCGTCGCGACGTACGACGC TGCCTGGAGACTAGGGAG
601_C070	CTGGAGAATCCCGGTGC	GAATCGTACCGCGCGTATCGGTC GGCCTGGAGACTAGGGAG
601_C071	CTGGAGAATCCCGGTGC	GAATCGTACGCCGCGTCGCTCGA AGCCTGGAGACTAGGGAG
601_C073	CTGGAGAATCCCGGTGC	GAATTACGCGCGGCGTTACGCGT CGCCTGGAGACTAGGGAG

Table C.1 continues on next page.

Table C.1, continued:

Barcode	Forward Primer	Reverse Primer
601_C075	CTGGAGAATCCCGGTGC	GACCGCGATACGACTCGTTCGTC GGCCTGGAGACTAGGGAG
601_C076	CTGGAGAATCCCGGTGC	GACCGCGCGATAAGACGCGTAA CGGCCTGGAGACTAGGGAG
601_C077	CTGGAGAATCCCGGTGC	GACCGCGCGTAATGCGCGACGTT AGCCTGGAGACTAGGGAG
601_C078	CTGGAGAATCCCGGTGC	GACCGCGCGTATAGTCCGAGCGT AGCCTGGAGACTAGGGAG
601_C079	CTGGAGAATCCCGGTGC	GACCGTACGTCGTGTCGAACGAC GGCCTGGAGACTAGGGAG
601_C080	CTGGAGAATCCCGGTGC	GACCGTCGAATCGTAACGTCGCG CGCCTGGAGACTAGGGAG
601_C081	CTGGAGAATCCCGGTGC	GACGACGAGCGTATACGCGCGA CAGCCTGGAGACTAGGGAG
601_C082	CTGGAGAATCCCGGTGC	GACGACGCGATACTACGCTCGGA CGCCTGGAGACTAGGGAG
601_C083	CTGGAGAATCCCGGTGC	GACGACGCGTAACTACGGTCGC GAGCCTGGAGACTAGGGAG
601_C084	CTGGAGAATCCCGGTGC	GACGACGGATACGTACGTCCGTC GGCCTGGAGACTAGGGAG
601_C085	CTGGAGAATCCCGGTGC	GACGACGTAACGCTATGCGTCGC GGCCTGGAGACTAGGGAG
601_C086	CTGGAGAATCCCGGTGC	GACGACTAACGCGTCGACGCGTA GGCCTGGAGACTAGGGAG
601_C087	CTGGAGAATCCCGGTGC	GACGATACGCCGATCGATCGTCG GGCCTGGAGACTAGGGAG
601_C088	CTGGAGAATCCCGGTGC	GACGATAGTCGCGTCGCACGATC GGCCTGGAGACTAGGGAG
601_C089	CTGGAGAATCCCGGTGC	GACGATCGTCGCATCGCCGAATC GGCCTGGAGACTAGGGAG
601_C090	CTGGAGAATCCCGGTGC	GACGATTCGACGGTCGCGACCGT AGCCTGGAGACTAGGGAG
601_C091	CTGGAGAATCCCGGTGC	GACGATTGACGCGTCGCGCGACA TGCCTGGAGACTAGGGAG
601_C092	CTGGAGAATCCCGGTGC	GACGCATATCGCGTCGTACGACC GGCCTGGAGACTAGGGAG

Table C.1 continues on next page.

Table C.1, continued:

Barcode	Forward Primer	Reverse Primer
601_C093	CTGGAGAATCCCGGTGC	GACGCCGATTACGTGTCGCGCGTAGCCTGGAGACTAGGGAG
601_C094	CTGGAGAATCCCGGTGC	GACGCGACCGATATTCGAGCGACGGCCTGGAGACTAGGGAG
601_C096	CTGGAGAATCCCGGTGC	GACGCGACTATCGCGCGTAACGCGGCCTGGAGACTAGGGAG
601_C097	CTGGAGAATCCCGGTGC	GACGCGATACGACCGCGTTACGCGGCCTGGAGACTAGGGAG
601_C098	CTGGAGAATCCCGGTGC	GACGCGATATCCGGCGCGTACCGAGCCTGGAGACTAGGGAG
601_C099	CTGGAGAATCCCGGTGC	GACGCGATATGCGGCGTTCGACGGGCCTGGAGACTAGGGAG
601_C100	CTGGAGAATCCCGGTGC	GACGCGATCGGTATCGGTACGCGCGCCTGGAGACTAGGGAG

Mutagenesis of C001_CXXX barcodes

Table C.2: C001_CXXX mutagenesis primers.

Barcode	Forward Primer	Reverse Primer
601 Base	CTGGAGAATCCCGGTGC	ACAGGATGTATATATCTGACACGTG
C001_C006	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GAACGATTCGCGTCGAATCGACGAGCCTGGAGACTAGGGAG
C001_C008	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GAACGCGAATCGTCGACGCGTATAGCCTGGAGACTAGGGAG
C001_C009	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GAACGCGTCGAAACGATTACGCGAGCCTGGAGACTAGGGAG
C001_C010	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GAACGTTCGAACGCGCGACGTTAAGCCTGGAGACTAGGGAG
C001_C011	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GAATCGCGCGATTCGCGTAATACGGCCTGGAGACTAGGGAG
C001_C014	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GAATTCGCGCGTACGTATACGCGAGCCTGGAGACTAGGGAG
C001_C015	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GACGAATTTCGCGCGTATTACGCGGCCTGGAGACTAGGGAG

Table C.2 continues on next page.

Table C.2, continued:

Barcode	Forward Primer	Reverse Primer
C001_C016	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GACGACGCGATAACGTCGACTATCGCCTGGAGACTAGGGAG
C001_C017	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GACGACGTTAACGCGTTTCGTACGGCCTGGAGACTAGGGAG
C001_C018	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GACGATACGACGAGATAGTCGACGGCCTGGAGACTAGGGAG
C001_C019	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GACGATACGCGTTGGTACGCGTAAGCCTGGAGACTAGGGAG
C001_C022	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GACGCGAAATTCGTATACGCGTCGGCCTGGAGACTAGGGAG
C001_C023	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GACGCGACGTAATTATCGCGTCGAGCCTGGAGACTAGGGAG
C001_C024	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GACGCGATACGAATATTCGCGCGAGCCTGGAGACTAGGGAG
C001_C025	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GACGCGATATCACTCGACGCGATAGCCTGGAGACTAGGGAG
C001_C028	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GACGTAATCGCGATCGCGCGAATAGCCTGGAGACTAGGGAG
C001_C029	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GACGTCGAATCGTTCGCGTAATCGGCCTGGAGACTAGGGAG
C001_C030	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GACGTCGCGTTAATCGCGTATACGGCCTGGAGACTAGGGAG
C001_C032	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GACGTTCGAACGTTCGTCGACGATGCCTGGAGACTAGGGAG
C001_C034	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GAGTGATATCGCGTTAACGTCGCGGCCTGGAGACTAGGGAG
C001_C035	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GATAATCGACGCGTTACGCGTACCGCCTGGAGACTAGGGAG
C001_C036	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GATACGCGCGAATTTACGTCGCGAGCCTGGAGACTAGGGAG
C001_C037	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GATACGTTACGCGACCGATACGCGGCCTGGAGACTAGGGAG
C001_C038	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GATAGTTCGCGACACCGTTCGTCGGCCTGGAGACTAGGGAG
C001_C039	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GATATCGCGCGAAACGACCGTTCGGCCTGGAGACTAGGGAG

Table C.2 continues on next page.

Table C.2, continued:

Barcode	Forward Primer	Reverse Primer
C001_C040	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GATATCGCGTCGTACGATCGTCGGGCCTGGAGACTAGGGAG
C001_C041	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GATCGACGCGTAAACGGTACGTCGGCCTGGAGACTAGGGAG
C001_C042	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GATCGACGTTCGTAGCGTCGTACGGCCTGGAGACTAGGGAG
C001_C043	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GATCGCGACGAAAATATCGCGCGGGCCTGGAGACTAGGGAG
C001_C044	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GATCGCGATTACGATGTCGCGCGAGCCTGGAGACTAGGGAG
C001_C047	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GATCGTACGATCGCCGACGATCGAGCCTGGAGACTAGGGAG
C001_C048	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GATCGTCGAACGACCGACGATCGTGCCTGGAGACTAGGGAG
C001_C049	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GATCGTCGTATCGCCGATACGTCGGCCTGGAGACTAGGGAG
C001_C050	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GATCGTTCGACGACCGCGCGATATGCCTGGAGACTAGGGAG
C001_C051	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GATCGTTTCGCGTCCGTTACGTCGGCCTGGAGACTAGGGAG
C001_C052	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GATTAACGCGACGCGAACGGTCGTGCCTGGAGACTAGGGAG
C001_C053	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GATTACGCGATCGCGACCGATACGGCCTGGAGACTAGGGAG
C001_C055	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GATTACGCGTCGACGACGAACGGTGCCTGGAGACTAGGGAG
C001_C056	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GATTATCGCGTCGCGACGGACGTAGCCTGGAGACTAGGGAG
C001_C057	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GATTCGCGCGATACGACGTAACGGGCCTGGAGACTAGGGAG
C001_C058	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GATTCGCGCGTAACGACGTACCGTGCCTGGAGACTAGGGAG
C001_C060	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GATTCGTATCGCGCGATCGTGCGAGCCTGGAGACTAGGGAG
C001_C061	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GATTTCGACGCGTCGATTCGGCGAGCCTGGAGACTAGGGAG

Table C.2 continues on next page.

Table C.2, continued:

Barcode	Forward Primer	Reverse Primer
C001_C063	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GAACCGCGATACGCGCGCGATATGGCCTGGAGACTAGGGAG
C001_C064	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GAACGACGTACGGCGCGCTATACGGCCTGGAGACTAGGGAG
C001_C065	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GAACGCGTCGCTACGCGTATCGGTGCCTGGAGACTAGGGAG
C001_C066	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GAACGGTCGTCGACGCGTATTCGGGCCTGGAGACTAGGGAG
C001_C070	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GAATCGTACCGCGCGTATCGGTCGGCCTGGAGACTAGGGAG
C001_C071	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GAATCGTACGCCGCGTCGCTCGAAGCCTGGAGACTAGGGAG
C001_C072	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GAATCGTCGACCGCGTCGTTCGACGCCTGGAGACTAGGGAG
C001_C073	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GAATTACGCGCGGCGTTACGCGTCGCCTGGAGACTAGGGAG
C001_C074	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GAATTGCGTCGCGCTACGCGTCGAGCCTGGAGACTAGGGAG
C001_C075	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GACCGCGATACGACTCGTTCGTCGGCCTGGAGACTAGGGAG
C001_C077	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GACCGCGCGTAATGCGCGACGTTAGCCTGGAGACTAGGGAG
C001_C079	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GACCGTACGTCGTGTCGAACGACGGCCTGGAGACTAGGGAG
C001_C081	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GACGACGAGCGTATACGCGCGACAGCCTGGAGACTAGGGAG
C001_C082	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GACGACGCGATACTACGCTCGGACGCCTGGAGACTAGGGAG
C001_C085	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GACGACGTAACGCTATGCGTCGCGGCCTGGAGACTAGGGAG
C001_C089	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GACGATCGTCGCATCGCCGAATCGGCCTGGAGACTAGGGAG
C001_C090	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GACGATTCGACGGTCGCGACCGTAGCCTGGAGACTAGGGAG
C001_C091	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GACGATTGACGCGTCGCGCGACATGCCTGGAGACTAGGGAG

Table C.2 continues on next page.

Table C.2, continued:

Barcode	Forward Primer	Reverse Primer
C001_C092	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GACGCATATCGCGTCGTACGACCGGCCTGGAGACTAGGGAG
C001_C093	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GACGCCGATTACGTGTCGCGCGTAGCCTGGAGACTAGGGAG
C001_C094	GAAACGCGTATCGCGCGCATAATAGCTCAATTGGTCGTAGACA	GACGCGACCGATATTCGAGCGACGGCCTGGAGACTAGGGAG

Mutagenesis of MMTV_CXXX barcodes

Table C.3: MMTV_CXXX mutagenesis primers.

Barcode	Forward Primer	Reverse Primer
MMTV_C001	GAAACGCGTATCGCGCGCATAATACTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C002	GAAATCGCGCGATTATTATGCGCGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C003	GAACGAACGTCGAACGCGCGATATCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C004	GAACGACGCGATAATATCGCGCGTCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C005	GAACGATTCGACGATCGTCGACGACTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C006	GAACGATTCGCGTCGAATCGACGACTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C007	GAACGCGAAACGACGAATCGCGTACTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C008	GAACGCGAATCGTCGACGCGTATACTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT

Table C.3 continues on next page.

Table C.3, continued:

Barcode	Forward Primer	Reverse Primer
MMTV_C009	GAACGCGTCGAAACGATTACGCGACTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C010	GAACGTTCGAACGCGCGACGTTAACTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C011	GAATCGCGCGATTCGCGTAATACGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C012	GAATTACGCGCGACGCGTAATCGACTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C013	GAATTACGTCGCGCGTACGAAACGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C014	GAATTCGCGCGTACGTATACGCGACTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C015	GACGAATTTCGCGCGTATTACGCGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C016	GACGACGCGATAACGTCGACTATCCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C017	GACGACGTTAACGCGTTTCGTACGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C018	GACGATACGACGAGATAGTCGACGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C019	GACGATACGCGTTGGTACGCGTAACTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C020	GACGATCGCGTAATACGCGATTCGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT

Table C.3 continues on next page.

Table C.3, continued:

Barcode	Forward Primer	Reverse Primer
MMTV_C021	GACGATCGTACGATAGCGTACCGACTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C022	GACGCGAAATTCGTATACGCGTCGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C023	GACGCGACGTAATTATCGCGTCGACTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C024	GACGCGATACGAATATTCGCGCGACTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C025	GACGCGATATCACTCGACGCGATACTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C026	GACGCGTAACGTATCGATTACGCGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C027	GACGCGTCGATTATCGCGACGTAACTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C028	GACGTAATCGCGATCGCGCGAATACTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C029	GACGTCGAATCGTTCGCGTAATCGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C030	GACGTCGCGTTAATCGCGTATACGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C031	GACGTTAACGTCGTCGGTACGCTACTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C032	GACGTTCGAACGTTCGTCGACGATCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT

Table C.3 continues on next page.

Table C.3, continued:

Barcode	Forward Primer	Reverse Primer
MMTV_C033	GAGTCGCGAACTATCGTCGATTC GCTCTTGTGTGTTTGTGTCTGTTC GCC	CAAAAAACTGTGCCGCAGT
MMTV_C034	GAGTGATATCGCGTTAACGTCGC GCTCTTGTGTGTTTGTGTCTGTTC GCC	CAAAAAACTGTGCCGCAGT
MMTV_C035	GATAATCGACGCGTTACGCGTAC CCTCTTGTGTGTTTGTGTCTGTTC GCC	CAAAAAACTGTGCCGCAGT
MMTV_C036	GATACGCGCGAATTTACGTCGCG ACTCTTGTGTGTTTGTGTCTGTTC GCC	CAAAAAACTGTGCCGCAGT
MMTV_C037	GATACGTTACGCGACCGATACGC GCTCTTGTGTGTTTGTGTCTGTTC GCC	CAAAAAACTGTGCCGCAGT
MMTV_C038	GATAGTTCGCGACACCGTTCGTC GCTCTTGTGTGTTTGTGTCTGTTC GCC	CAAAAAACTGTGCCGCAGT
MMTV_C039	GATATCGCGCGAAACGACCGTTC GCTCTTGTGTGTTTGTGTCTGTTC GCC	CAAAAAACTGTGCCGCAGT
MMTV_C040	GATATCGCGTCGTACGATCGTCG GCTCTTGTGTGTTTGTGTCTGTTC GCC	CAAAAAACTGTGCCGCAGT
MMTV_C041	GATCGACGCGTAAACGGTACGTC GCTCTTGTGTGTTTGTGTCTGTTC GCC	CAAAAAACTGTGCCGCAGT
MMTV_C042	GATCGACGTTCGTAGCGTCGTAC GCTCTTGTGTGTTTGTGTCTGTTC GCC	CAAAAAACTGTGCCGCAGT
MMTV_C043	GATCGCGACGAAAATATCGCGC GGCTCTTGTGTGTTTGTGTCTGTT CGCC	CAAAAAACTGTGCCGCAGT
MMTV_C044	GATCGCGATTACGATGTCGCGCG ACTCTTGTGTGTTTGTGTCTGTTC GCC	CAAAAAACTGTGCCGCAGT

Table C.3 continues on next page.

Table C.3, continued:

Barcode	Forward Primer	Reverse Primer
MMTV_C045	GATCGCGCGTAATCATATCGCGCGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C046	GATCGCGTACGAACCGAATACGCGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C047	GATCGTACGATCGCCGACGATCGACTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C048	GATCGTCGAACGACCGACGATCGTCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C049	GATCGTCGTATCGCCGATACGTCGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C050	GATCGTTCGACGACCGCGCGATATCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C051	GATCGTTTCGCGTCCGTTACGTCGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C052	GATTAACGCGACGCGAACGGTCGTCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C053	GATTACGCGATCGCGACCGATACGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C054	GATTACGCGCGAACGACGAACGAGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C055	GATTACGCGTCGACGACGAACGGTCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C056	GATTATCGCGTCGCGACGGACGTACTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT

Table C.3 continues on next page.

Table C.3, continued:

Barcode	Forward Primer	Reverse Primer
MMTV_C057	GATTCGCGCGATACGACGTAACG GCTCTTGTGTGTTTGTGTCTGTTC GCC	CAAAAAACTGTGCCGCAGT
MMTV_C058	GATTCGCGCGTAACGACGTACCG TCTCTTGTGTGTTTGTGTCTGTTC GCC	CAAAAAACTGTGCCGCAGT
MMTV_C059	GATTCGTACGCGACGACGTATCG GCTCTTGTGTGTTTGTGTCTGTTC GCC	CAAAAAACTGTGCCGCAGT
MMTV_C060	GATTCGTATCGCGCGATCGTGCG ACTCTTGTGTGTTTGTGTCTGTTC GCC	CAAAAAACTGTGCCGCAGT
MMTV_C061	GATTTCGACGCGTCGATTCGGCG ACTCTTGTGTGTTTGTGTCTGTTC GCC	CAAAAAACTGTGCCGCAGT
MMTV_C062	GATTTCGTCGCGACGCGACGCAT ACTCTTGTGTGTTTGTGTCTGTTC GCC	CAAAAAACTGTGCCGCAGT
MMTV_C063	GAACCGCGATACGCGCGCGATAT GCTCTTGTGTGTTTGTGTCTGTTC GCC	CAAAAAACTGTGCCGCAGT
MMTV_C064	GAACGACGTACGGCGCGCTATA CGCTCTTGTGTGTTTGTGTCTGTT CGCC	CAAAAAACTGTGCCGCAGT
MMTV_C065	GAACGCGTCGCTACGCGTATCGG TCTCTTGTGTGTTTGTGTCTGTTC GCC	CAAAAAACTGTGCCGCAGT
MMTV_C066	GAACGGTCGTCGACGCGTATTCG GCTCTTGTGTGTTTGTGTCTGTTC GCC	CAAAAAACTGTGCCGCAGT
MMTV_C067	GAACGTTCGACCGCGGTCGTACG ACTCTTGTGTGTTTGTGTCTGTTC GCC	CAAAAAACTGTGCCGCAGT
MMTV_C068	GAACTCGTCGCGACGTACGACGC TCTCTTGTGTGTTTGTGTCTGTTC GCC	CAAAAAACTGTGCCGCAGT

Table C.3 continues on next page.

Table C.3, continued:

Barcode	Forward Primer	Reverse Primer
MMTV_C069	GAATCGCGGTACGCGTATAGCGCGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C070	GAATCGTACCGCGCGTATCGGTCGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C071	GAATCGTACGCCGCGTCGCTCGAACTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C072	GAATCGTCGACCGCGTCGTTCGACCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C073	GAATTACGCGCGGCGTTACGCGTCCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C074	GAATTGCGTCGCGCTACGCGTCGACTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C075	GACCGCGATACGACTCGTTCGTCGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C076	GACCGCGCGATAAGACGCGTAACGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C077	GACCGCGCGTAATGCGCGACGTTACTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C078	GACCGCGCGTATAGTCCGAGCGTACTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C079	GACCGTACGTCGTGTCGAACGACGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C080	GACCGTCGAATCGTAACGTCGCGCCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT

Table C.3 continues on next page.

Table C.3, continued:

Barcode	Forward Primer	Reverse Primer
MMTV_C081	GACGACGAGCGTATACGCGCGACACTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C082	GACGACGCGATACTACGCTCGGACCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C083	GACGACGCGTAACTACGGTCGCGACTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C084	GACGACGGATACGTACGTCCGTCGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C085	GACGACGTAACGCTATGCGTCGCGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C086	GACGACTAACGCGTCGACGCGTAGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C087	GACGATACGCCGATCGATCGTCGGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C088	GACGATAGTCGCGTCGCACGATCGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C089	GACGATCGTCGCATCGCCGAATCGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C090	GACGATTCGACGGTCGCGACCGTACTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C091	GACGATTGACGCGTCGCGCGACATCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C092	GACGCATATCGCGTCGTACGACCGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT

Table C.3 continues on next page.

Table C.3, continued:

Barcode	Forward Primer	Reverse Primer
MMTV_C093	GACGCCGATTACGTGTCGCGCGTACTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C094	GACGCGACCGATATTCGAGCGACGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C095	GACGCGACGCAATCCGTCGAACGCCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C096	GACGCGACTATCGCGCGTAACGCGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C097	GACGCGATACGACCGCGTTACGCGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C098	GACGCGATATCCGGCGCGTACCGACTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C099	GACGCGATATGCGGCGTTCGACGGCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT
MMTV_C100	GACGCGATCGGTATCGGTACGCGCCTCTTGTGTGTTTGTGTCTGTTCGCC	CAAAAAACTGTGCCGCAGT

Mutagenesis of MMS_DXXX barcodes

Table C.4: MMS_DXXX mutagenesis primers.

Barcode	Forward Primer	Reverse Primer
MMS Base	TTTGTAGAACAGTGTATATCAATGAGTT	CCGTTTCCAACGAATGTGTTT
MMS_D001	ATGATATTCGTACCCGTTTGTAGAACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCCAACGAATGTGTTT
MMS_D002	ATGATAACGTAGACCGTTTGTAGAACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCCAACGAATGTGTTT

Table C.4 continues on next page.

Table C.4, continued:

Barcode	Forward Primer	Reverse Primer
MMS_D003	ATGTAGTTCGTACGACTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D004	ATGGAAGCGAACGTATTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D005	ATGACGTCGACTATTATTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D006	ATGCGCGATTAGACTATTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D007	ATGATGGTACGCGATTTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D008	ATGTAGATCGCGTAAGTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D009	ATGTCTAGTAACGACGTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D010	ATGTTATACCTCGCGTTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D011	ATGAATACGCGCGTAATTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D012	ATGGCGTTATCGTACATTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D013	ATGTGTTTAGCGAACGTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D014	ATGAGATTATCGACCGTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D015	ATGTATAGTACGCGTCTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D016	ATGTCTATTCGGCGTATTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D017	ATGCGTCGATAACCTATTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D018	ATGCTTCGATACGTAATTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D019	ATGTCGTAACGCGAATTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D020	ATGATCGCTCTAACGTTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT

Table C.4 continues on next page.

Table C.4, continued:

Barcode	Forward Primer	Reverse Primer
MMS_D021	ATGCTTATCGCGTTGATTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D022	ATGTCGTTACGTCCTATTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D023	ATGTGAACGTCGTAGTTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D024	ATGCGTTATACACGACTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D025	ATGTCGTACGTTAGACTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D026	ATGAACGACGGTACATTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D027	ATGTACGACGTAAGGTTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D028	ATGTACTATCGTCACGTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D029	ATGACTACGCTACGATTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D030	ATGTAATCGCGCTAACTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D031	ATGATTTAGGCGTACGTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D032	ATGTCGATAGCGTAAGTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D033	ATGCGCGTTAGATAGTTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D034	ATGCGGTTACGCTATATTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D035	ATGTATCGCTAACTCGTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D036	ATGCGCGTAATAGTACTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D037	ATGCGTACGCTATCTATTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D038	ATGCCGCGAACTTATATTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT

Table C.4 continues on next page.

Table C.4, continued:

Barcode	Forward Primer	Reverse Primer
MMS_D039	ATGTTACAATACGCGCTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D040	ATGTAGTTTACGCGAGTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D041	ATGCTCGAATTGACGTTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D042	ATGCGTCGTACTACATTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D043	ATGCGTAATACCTACGTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D044	ATGTCATTACGATCGCTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D045	ATGGTAATGCGCGATATTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D046	ATGCGCGAATACTAAGTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D047	ATGTAACGTCCGGTAATTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D048	ATGTATTCGTATCCCGTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D049	ATGTAGTAACGTCGAGTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D050	ATGCCGTTATAGTACGTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D051	ATGGATAACGCGAAACTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D052	ATGACGTAGGTATTCGTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D053	ATGCGTACTTTAGACGTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D054	ATGGAATACGCGAATCTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D055	ATGCAGTATTCGCGTATTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D056	ATGCGTACTAATCGTCTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT

Table C.4 continues on next page.

Table C.4, continued:

Barcode	Forward Primer	Reverse Primer
MMS_D057	ATGGATCGCGTACTATTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D058	ATGATACGCGATGTATTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D059	ATGTTCAATACGCGACTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D060	ATGCGAAAGACGTATCTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D061	ATGACGCCGTAATAGTTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D062	ATGCGATCGCGTATTATTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D063	ATGAGACCGATTAACGTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D064	ATGGTTCGGACGTAATTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D065	ATGAGATAGCGACGTATTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D066	ATGTATAGTATCGCGATTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D067	ATGATACTACGCCGATTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D068	ATGTATCGCGAACTTATTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D069	ATGCTATCGAGCGATATTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D070	ATGACGTTCGAACTAGTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D071	ATGTATCGAATACGGCTTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT
MMS_D072	ATGGCGAACGTAGTTATTTGTAG AACAGTGTATATCAATGAGTT	ATACTACGTATCGTCCCGTTTCC AACGAATGTGTTT

Amplification of Space Alien barcodes

Table C.5: Space Alien amplification primers.

Barcode	Forward Primer	Reverse Primer
Space Alien Amplification	ATTAGCGACGTGATAATCT	TAATAGTATGACGCGCG

Primers and probes for qPCR of genomic targets and barcodes

601_CXXX barcode qPCR

Table C.6: 601_CXXX barcode qPCR primers and probe.

Target	Forward Primer	Reverse Primer	Hydrolysis Probe
601_C002	GCTCAATTGGTCGTAGACAG	AATCGCGCGATTATTATGC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C005	GCTCAATTGGTCGTAGACAG	GAACGATTCGACGATCGT	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C008	GCTCAATTGGTCGTAGACAG	GAATCGTCGACGCGTATA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C009	GCTCAATTGGTCGTAGACAG	ACGCGTCGAAACGATTA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C010	GCTCAATTGGTCGTAGACAG	GAACGCGCGACGTTAA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C013	GCTCAATTGGTCGTAGACAG	TCGCGCGTACGAAAC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C014	GCTCAATTGGTCGTAGACAG	ATTCGCGCGTACGTATAC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/

Table C.6 continues on next page.

Table C.6, continued:

Target	Forward Primer	Reverse Primer	Hydrolysis Probe
601_C015	GCTCAATTGGTCGTAGACAG	CGAATTTCGCGCGTATTAC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C017	GCTCAATTGGTCGTAGACAG	CGTTAACGCGTTTCGT	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C019	GCTCAATTGGTCGTAGACAG	CGATACGCGTTGGTACG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C021	GCTCAATTGGTCGTAGACAG	CGATCGTACGATAGCGTAC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C022	GCTCAATTGGTCGTAGACAG	CGAAATTCGTATACGCGTCG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C025	GCTCAATTGGTCGTAGACAG	CGATATCACTCGACGCGATA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C026	GCTCAATTGGTCGTAGACAG	CGCGTAACGTATCGATTAC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C028	GCTCAATTGGTCGTAGACAG	GCGATCGCGCGAATA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C029	GCTCAATTGGTCGTAGACAG	TCGAATCGTTCGCGTAATC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C031	GCTCAATTGGTCGTAGACAG	GTTAACGTCGTCGGTACG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C032	GCTCAATTGGTCGTAGACAG	GAACGTTCGTCGACGAT	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C033	GCTCAATTGGTCGTAGACAG	CGCGAACTATCGTCGATTC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/

Table C.6 continues on next page.

Table C.6, continued:

Target	Forward Primer	Reverse Primer	Hydrolysis Probe
601_C034	GCTCAATTGGTCGTAGACAG	GTGATATCGCGTTAACGTCG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C037	GCTCAATTGGTCGTAGACAG	CGTTACGCGACCGATAC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C038	GCTCAATTGGTCGTAGACAG	GTTCGCGACACCGTTC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C039	GCTCAATTGGTCGTAGACAG	TATCGCGCGAAACGAC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C040	GCTCAATTGGTCGTAGACAG	TATCGCGTCGTACGATCG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C041	GCTCAATTGGTCGTAGACAG	CGCGTAAACGGTACGTC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C042	GCTCAATTGGTCGTAGACAG	ACGTTCGTAGCGTCGTA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C043	GCTCAATTGGTCGTAGACAG	ACGAAAATATCGCGCGG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C044	GCTCAATTGGTCGTAGACAG	TCGCGATTACGATGTCG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C046	GCTCAATTGGTCGTAGACAG	CGCGTACGAACCGAATAC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C047	GCTCAATTGGTCGTAGACAG	TCGTACGATCGCCGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C049	GCTCAATTGGTCGTAGACAG	TCGTCGTATCGCCGATA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/

Table C.6 continues on next page.

Table C.6, continued:

Target	Forward Primer	Reverse Primer	Hydrolysis Probe
601_C051	GCTCAATTGGTCGTAGACAG	CGTTTCGCGTCCGTTA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C052	GCTCAATTGGTCGTAGACAG	TTAACGCGACGCGAAC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C054	GCTCAATTGGTCGTAGACAG	TTACGCGCGAACGAC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C055	GCTCAATTGGTCGTAGACAG	TCGACGACGAACGGT	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C056	GCTCAATTGGTCGTAGACAG	TCGCGACGGACGTA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C058	GCTCAATTGGTCGTAGACAG	CGCGTAACGACGTACC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C060	GCTCAATTGGTCGTAGACAG	TTCGTATCGCGCGATC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C061	GCTCAATTGGTCGTAGACAG	TTTCGACGCGTCGATTC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C062	GCTCAATTGGTCGTAGACAG	CGACGCGACGCATA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C063	GCTCAATTGGTCGTAGACAG	TACGCGCGCGATATG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C066	GCTCAATTGGTCGTAGACAG	GTCGTCGACGCGTATTC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C067	GCTCAATTGGTCGTAGACAG	CCGCGGTCGTACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/

Table C.6 continues on next page.

Table C.6, continued:

Target	Forward Primer	Reverse Primer	Hydrolysis Probe
601_C068	GCTCAATTGGTCGTAGACAG	ACTCGTCGCGACGTA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C070	GCTCAATTGGTCGTAGACAG	GTACCGCGCGTATCG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C071	GCTCAATTGGTCGTAGACAG	ATCGTACGCCGCGT	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C073	GCTCAATTGGTCGTAGACAG	ATTACGCGCGGCG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C075	GCTCAATTGGTCGTAGACAG	GATACGACTCGTTCGTCG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C076	GCTCAATTGGTCGTAGACAG	GCGATAAGACGCGTAACG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C077	GCTCAATTGGTCGTAGACAG	CGTAATGCGCGACGTTA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C078	GCTCAATTGGTCGTAGACAG	GCGTATAGTCCGAGCGTA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C079	GCTCAATTGGTCGTAGACAG	CGTACGTCGTGTCGAA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C080	GCTCAATTGGTCGTAGACAG	CCGTCGAATCGTAACGTC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C081	GCTCAATTGGTCGTAGACAG	CGTATACGCGCGACA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C082	GCTCAATTGGTCGTAGACAG	GCGATACTACGCTCGGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/

Table C.6 continues on next page.

Table C.6, continued:

Target	Forward Primer	Reverse Primer	Hydrolysis Probe
601_C083	GCTCAATTGGTCGTAGACAG	CGTAACTACGGTCGCGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C084	GCTCAATTGGTCGTAGACAG	GGATACGTACGTCCGTCG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C085	GCTCAATTGGTCGTAGACAG	CGACGTAACGCTATGCG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C086	GCTCAATTGGTCGTAGACAG	CGACTAACGCGTCGAC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C087	GCTCAATTGGTCGTAGACAG	CGATACGCCGATCGATC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C088	GCTCAATTGGTCGTAGACAG	TCGCGTCGCACGAT	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C089	GCTCAATTGGTCGTAGACAG	TCGCATCGCCGAATC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C090	GCTCAATTGGTCGTAGACAG	CGATTCGACGGTCGC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C091	GCTCAATTGGTCGTAGACAG	ATTGACGCGTCGCG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C092	GCTCAATTGGTCGTAGACAG	CGCATATCGCGTCGT	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C093	GCTCAATTGGTCGTAGACAG	TACGTGTCGCGCGTA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C094	GCTCAATTGGTCGTAGACAG	CCGATATTCGAGCGACG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/

Table C.6 continues on next page.

Table C.6, continued:

Target	Forward Primer	Reverse Primer	Hydrolysis Probe
601_C096	GCTCAATTGGTCGTAGACAG	CGACTATCGCGCGTAAC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C097	GCTCAATTGGTCGTAGACAG	CGATACGACCGCGTTAC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C098	GCTCAATTGGTCGTAGACAG	TATCCGGCGCGTACC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C099	GCTCAATTGGTCGTAGACAG	ATATGCGGCGTTCGAC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
601_C100	GCTCAATTGGTCGTAGACAG	CGCGATCGGTATCGG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/

C001_CXXX barcode qPCR

Table C.7: C001_CXXX barcode qPCR primers and probe.

Target	Forward Primer	Reverse Primer	Hydrolysis Probe
C001_C006	CGTATCGCGCGCATAATA	ACGATTCGCGTCGAATC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C008	CGTATCGCGCGCATAATA	GAATCGTCGACGCGTATA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C009	CGTATCGCGCGCATAATA	ACGCGTCGAAACGATTA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C010	CGTATCGCGCGCATAATA	GAACGCGCGACGTTAA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C011	CGTATCGCGCGCATAATA	CGCGATTCGCGTAATACG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/

Table C.7 continues on next page.

Table C.7, continued:

Target	Forward Primer	Reverse Primer	Hydrolysis Probe
C001_C014	CGTATCGCGCGCATAATA	ATTCGCGCGTACGTATAC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C015	CGTATCGCGCGCATAATA	CGAATTTCGCGCGTATTAC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C016	CGTATCGCGCGCATAATA	ACGCGATAACGTCGACTA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C017	CGTATCGCGCGCATAATA	CGTTAACGCGTTTCGT	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C018	CGTATCGCGCGCATAATA	ATACGACGAGATAGTCGACG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C019	CGTATCGCGCGCATAATA	CGATACGCGTTGGTACG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C022	CGTATCGCGCGCATAATA	CGAAATTCGTATACGCGTCG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C023	CGTATCGCGCGCATAATA	GACGTAATTATCGCGTCGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C024	CGTATCGCGCGCATAATA	GATACGAATATTCGCGCGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C025	CGTATCGCGCGCATAATA	CGATATCACTCGACGCGATA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C028	CGTATCGCGCGCATAATA	GCGATCGCGCGAATA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C029	CGTATCGCGCGCATAATA	TCGAATCGTTCGCGTAATC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/

Table C.7 continues on next page.

Table C.7, continued:

Target	Forward Primer	Reverse Primer	Hydrolysis Probe
C001_C030	CGTATCGCGCGCATAATA	CGCGTTAATCGCGTATACG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C032	CGTATCGCGCGCATAATA	GAACGTTCGTCGACGAT	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C034	CGTATCGCGCGCATAATA	GTGATATCGCGTTAACGTCG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C035	CGTATCGCGCGCATAATA	TAATCGACGCGTTACGC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C036	CGTATCGCGCGCATAATA	TACGCGCGAATTTACGTC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C037	CGTATCGCGCGCATAATA	CGTTACGCGACCGATAC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C038	CGTATCGCGCGCATAATA	GTTCGCGACACCGTTC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C039	CGTATCGCGCGCATAATA	TATCGCGCGAAACGAC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C040	CGTATCGCGCGCATAATA	TATCGCGTCGTACGATCG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C041	CGTATCGCGCGCATAATA	CGCGTAAACGGTACGTC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C042	CGTATCGCGCGCATAATA	ACGTTCGTAGCGTCGTA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C043	CGTATCGCGCGCATAATA	ACGAAAATATCGCGCGG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/

Table C.7 continues on next page.

Table C.7, continued:

Target	Forward Primer	Reverse Primer	Hydrolysis Probe
C001_C044	CGTATCGCGCGCATAATA	TCGCGATTACGATGTCG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C047	CGTATCGCGCGCATAATA	TCGTACGATCGCCGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C048	CGTATCGCGCGCATAATA	TCGAACGACCGACGAT	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C049	CGTATCGCGCGCATAATA	TCGTCGTATCGCCGATA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C050	CGTATCGCGCGCATAATA	ACGACCGCGCGATAT	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C051	CGTATCGCGCGCATAATA	CGTTTCGCGTCCGTTA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C052	CGTATCGCGCGCATAATA	TTAACGCGACGCGAAC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C053	CGTATCGCGCGCATAATA	GATCGCGACCGATACG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C055	CGTATCGCGCGCATAATA	TCGACGACGAACGGT	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C056	CGTATCGCGCGCATAATA	TCGCGACGGACGTA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C057	CGTATCGCGCGCATAATA	TTCGCGCGATACGAC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C058	CGTATCGCGCGCATAATA	CGCGTAACGACGTACC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/

Table C.7 continues on next page.

Table C.7, continued:

Target	Forward Primer	Reverse Primer	Hydrolysis Probe
C001_C060	CGTATCGCGCGCATAATA	TTCGTATCGCGCGATC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C061	CGTATCGCGCGCATAATA	TTTCGACGCGTCGATTC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C063	CGTATCGCGCGCATAATA	TACGCGCGCGATATG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C064	CGTATCGCGCGCATAATA	ACGGCGCGCTATAC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C065	CGTATCGCGCGCATAATA	GCTACGCGTATCGGT	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C066	CGTATCGCGCGCATAATA	GTCGTCGACGCGTATTC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C070	CGTATCGCGCGCATAATA	GTACCGCGCGTATCG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C071	CGTATCGCGCGCATAATA	ATCGTACGCCGCGT	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C072	CGTATCGCGCGCATAATA	TCGACCGCGTCGTT	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C073	CGTATCGCGCGCATAATA	ATTACGCGCGGCG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C074	CGTATCGCGCGCATAATA	ATTGCGTCGCGCTAC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C075	CGTATCGCGCGCATAATA	GATACGACTCGTTCGTCG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/

Table C.7 continues on next page.

Table C.7, continued:

Target	Forward Primer	Reverse Primer	Hydrolysis Probe
C001_C077	CGTATCGCGCGCATAATA	CGTAATGCGCGACGTTA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C079	CGTATCGCGCGCATAATA	CGTACGTCGTGTCGAA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C081	CGTATCGCGCGCATAATA	CGTATACGCGCGACA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C082	CGTATCGCGCGCATAATA	GCGATACTACGCTCGGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C085	CGTATCGCGCGCATAATA	CGACGTAACGCTATGCG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C089	CGTATCGCGCGCATAATA	TCGCATCGCCGAATC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C090	CGTATCGCGCGCATAATA	CGATTCGACGGTCGC	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C091	CGTATCGCGCGCATAATA	ATTGACGCGTCGCG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C092	CGTATCGCGCGCATAATA	CGCATATCGCGTCGT	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C093	CGTATCGCGCGCATAATA	TACGTGTCGCGCGTA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
C001_C094	CGTATCGCGCGCATAATA	CCGATATTCGAGCGACG	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/

MMTV_CXXX barcode qPCR

Table C.8: MMTV_CXXX barcode qPCR primers and probe.

Target	Forward Primer	Reverse Primer	Hydrolysis Probe
MMTV_C001	CGTATCGCGCGCATAATA	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C002	AATCGCGCGATTATTATGC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C003	CGTCGAACGCGCGATAT	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C004	ACGACGCGATAATATCGC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C005	GAACGATTCGACGATCGT	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C006	ACGATTCGCGTCGAATC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C007	ACGCGAAACGACGAATC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C008	GAATCGTCGACGCGTATA	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C009	ACGCGTCGAAACGATTA	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C010	GAACGCGCGACGTTAA	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C011	CGCGATTCGCGTAATACG	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C012	GCGACGCGTAATCGA	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/

Table C.8 continues on next page.

Table C.8, continued:

Target	Forward Primer	Reverse Primer	Hydrolysis Probe
MMTV_C013	TCGCGCGTACGAAAC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C014	ATTCGCGCGTACGTATAC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C015	CGAATTTCGCGCGTATTAC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C016	ACGCGATAACGTCGACTA	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C017	CGTTAACGCGTTTCGT	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C018	ATACGACGAGATAGTCGACG	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C019	CGATACGCGTTGGTACG	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C020	GACGATCGCGTAATACGC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C021	CGATCGTACGATAGCGTAC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C022	CGAAATTCGTATACGCGTCG	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C023	GACGTAATTATCGCGTCGA	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C024	GATACGAATATTCGCGCGA	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/

Table C.8 continues on next page.

Table C.8, continued:

Target	Forward Primer	Reverse Primer	Hydrolysis Probe
MMTV_C025	CGATATCACTCGACGCGATA	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C026	CGCGTAACGTATCGATTAC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C027	GTCGATTATCGCGACGTAA	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C028	GCGATCGCGCGAATA	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C029	TCGAATCGTTCGCGTAATC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C030	CGCGTTAATCGCGTATACG	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C031	GTTAACGTCGTCGGTACG	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C032	GAACGTTCGTCGACGAT	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C033	CGCGAACTATCGTCGATTC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C034	GTGATATCGCGTTAACGTCG	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C035	TAATCGACGCGTTACGC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C036	TACGCGCGAATTTACGTC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/

Table C.8 continues on next page.

Table C.8, continued:

Target	Forward Primer	Reverse Primer	Hydrolysis Probe
MMTV_C037	CGTTACGCGACCGATAC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C038	GTTCGCGACACCGTTC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C039	TATCGCGCGAAACGAC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C040	TATCGCGTCGTACGATCG	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C041	CGCGTAAACGGTACGTC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C042	ACGTTCGTAGCGTCGTA	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C043	ACGAAAATATCGCGCGG	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C044	TCGCGATTACGATGTCG	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C045	TCGCGCGTAATCATATCG	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C046	CGCGTACGAACCGAATAC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C047	TCGTACGATCGCCGA	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C048	TCGAACGACCGACGAT	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/

Table C.8 continues on next page.

Table C.8, continued:

Target	Forward Primer	Reverse Primer	Hydrolysis Probe
MMTV_C049	TCGTCGTATCGCCGATA	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C050	ACGACCGCGCGATAT	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C051	CGTTTCGCGTCCGTTA	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C052	TTAACGCGACGCGAAC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C053	GATCGCGACCGATACG	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C054	TTACGCGCGAACGAC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C055	TCGACGACGAACGGT	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C056	TCGCGACGGACGTA	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C057	TTCGCGCGATACGAC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C058	CGCGTAACGACGTACC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C059	ACGCGACGACGTATC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C060	TTCGTATCGCGCGATC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/

Table C.8 continues on next page.

Table C.8, continued:

Target	Forward Primer	Reverse Primer	Hydrolysis Probe
MMTV_C061	TTTCGACGCGTCGATTC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C062	CGACGCGACGCATA	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C063	TACGCGCGCGATATG	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C064	ACGGCGCGCTATAC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C065	GCTACGCGTATCGGT	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C066	GTCGTCGACGCGTATTC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C067	CCGCGGTCGTACGA	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C068	ACTCGTCGCGACGTA	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C069	ATCGCGGTACGCGTATA	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C070	GTACCGCGCGTATCG	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C071	ATCGTACGCCGCGT	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C072	TCGACCGCGTCGTT	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/

Table C.8 continues on next page.

Table C.8, continued:

Target	Forward Primer	Reverse Primer	Hydrolysis Probe
MMTV_C073	ATTACGCGCGGCG	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C074	ATTGCGTCGCGCTAC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C075	GATACGACTCGTTCGTCG	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C076	GCGATAAGACGCGTAACG	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C077	CGTAATGCGCGACGTTA	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C078	GCGTATAGTCCGAGCGTA	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C079	CGTACGTCGTGTCGAA	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C080	CCGTCGAATCGTAACGTC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C081	CGTATACGCGCGACA	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C082	GCGATACTACGCTCGGA	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C083	CGTAACTACGGTCGCGA	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C084	GGATACGTACGTCCGTCG	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/

Table C.8 continues on next page.

Table C.8, continued:

Target	Forward Primer	Reverse Primer	Hydrolysis Probe
MMTV_C085	CGACGTAACGCTATGCG	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C086	CGACTAACGCGTCGAC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C087	CGATACGCCGATCGATC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C088	TCGCGTCGCACGAT	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C089	TCGCATCGCCGAATC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C090	CGATTCGACGGTCGC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C091	ATTGACGCGTCGCG	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C092	CGCATATCGCGTCGT	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C093	TACGTGTCGCGCGTA	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C094	CCGATATTCGAGCGACG	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C095	ACGCAATCCGTCGAAC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C096	CGACTATCGCGCGTAAC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/

Table C.8 continues on next page.

Table C.8, continued:

Target	Forward Primer	Reverse Primer	Hydrolysis Probe
MMTV_C097	CGATACGACCGCGTTAC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C098	TATCCGGCGCGTACC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C099	ATATGCGGCGTTCGAC	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/
MMTV_C100	CGCGATCGGTATCGG	TGGAAAGTGAAGGATAAGTGACGA	/56-FAM/TCTAGCACCGCTTAAACGCACGTA/3IABkFQ/

Space Alien barcode qPCR

Table C.9: Space Alien barcode qPCR primers and probe.

Target	Forward Primer	Reverse Primer	Hydrolysis Probe
S001	CGTCGTGTCGCGTATAA	CAATGAACGCCGGTTAATA	/56-FAM/ACGTAATCGACACTACGTCGACT/3IABkFQ/
S002	CGTACTAAGACGTATCTAGCG	CAATGAACGCCGGTTAATA	/56-FAM/ACGTAATCGACACTACGTCGACT/3IABkFQ/
S003	GAGTAATAAGTACGCGAGATAG	CAATGAACGCCGGTTAATA	/56-FAM/ACGTAATCGACACTACGTCGACT/3IABkFQ/
S004	ACGCGTCGCTATACTT	CAATGAACGCCGGTTAATA	/56-FAM/ACGTAATCGACACTACGTCGACT/3IABkFQ/
S005	ACGTAACCGGTAGACTTAT	CAATGAACGCCGGTTAATA	/56-FAM/ACGTAATCGACACTACGTCGACT/3IABkFQ/
S006	ATTACGACCGTTTATTCGC	CAATGAACGCCGGTTAATA	/56-FAM/ACGTAATCGACACTACGTCGACT/3IABkFQ/

Table C.9 continues on next page.

Table C.9, continued:

Target	Forward Primer	Reverse Primer	Hydrolysis Probe
S007	GTTAAGTATGCGAACCGTATAG	CAATGAACGCCGGTTAATA	/56-FAM/ACGTAATCGACACTACGTCGACT/3IABkFQ/
S008	TTAGACGACCGAATTCTACT	CAATGAACGCCGGTTAATA	/56-FAM/ACGTAATCGACACTACGTCGACT/3IABkFQ/
S009	ACGTACTACGATCTCGAC	CAATGAACGCCGGTTAATA	/56-FAM/ACGTAATCGACACTACGTCGACT/3IABkFQ/
S010	TCGCGTCACGACTAAT	CAATGAACGCCGGTTAATA	/56-FAM/ACGTAATCGACACTACGTCGACT/3IABkFQ/
S011	CGCTATACGAGAATAACGC	CAATGAACGCCGGTTAATA	/56-FAM/ACGTAATCGACACTACGTCGACT/3IABkFQ/
S012	AGACCGAGTTCGCTTATAC	CAATGAACGCCGGTTAATA	/56-FAM/ACGTAATCGACACTACGTCGACT/3IABkFQ/
S013	GATGTATCGTAGTCGGAGT	CAATGAACGCCGGTTAATA	/56-FAM/ACGTAATCGACACTACGTCGACT/3IABkFQ/
S014	CGTAACGCGTTTAGAGTATT	CAATGAACGCCGGTTAATA	/56-FAM/ACGTAATCGACACTACGTCGACT/3IABkFQ/
S015	ACGCGAATTCGAACTAATC	CAATGAACGCCGGTTAATA	/56-FAM/ACGTAATCGACACTACGTCGACT/3IABkFQ/
S016	TGATACGTCCGATACGC	CAATGAACGCCGGTTAATA	/56-FAM/ACGTAATCGACACTACGTCGACT/3IABkFQ/
S017	GAATTACCTTACCGTCGATTG	CAATGAACGCCGGTTAATA	/56-FAM/ACGTAATCGACACTACGTCGACT/3IABkFQ/
S018	TTCGACGTACCGATGTAA	CAATGAACGCCGGTTAATA	/56-FAM/ACGTAATCGACACTACGTCGACT/3IABkFQ/

Table C.9 continues on next page.

Table C.9, continued:

Target	Forward Primer	Reverse Primer	Hydrolysis Probe
S019	CCATTACGCGATAACTACG	CAATGAACGCCGGTTAATA	/56-FAM/ACGTAATCGACACTACGTCGACT/3IABkFQ/
S020	TTACGCGCACTAATGTATC	CAATGAACGCCGGTTAATA	/56-FAM/ACGTAATCGACACTACGTCGACT/3IABkFQ/
S021	AGTACTCGAATACGCTACC	CAATGAACGCCGGTTAATA	/56-FAM/ACGTAATCGACACTACGTCGACT/3IABkFQ/
S022	GGATACATATCCGCGAACT	CAATGAACGCCGGTTAATA	/56-FAM/ACGTAATCGACACTACGTCGACT/3IABkFQ/
S023	TCGATGCGGTGATTACT	CAATGAACGCCGGTTAATA	/56-FAM/ACGTAATCGACACTACGTCGACT/3IABkFQ/
S024	CGACAGACGATCTCGTAA	CAATGAACGCCGGTTAATA	/56-FAM/ACGTAATCGACACTACGTCGACT/3IABkFQ/
S025	GACGAACCTTATCGTGTAAC	CAATGAACGCCGGTTAATA	/56-FAM/ACGTAATCGACACTACGTCGACT/3IABkFQ/

Human genomic locus qPCR

Table C.10: Human genomic locus barcode qPCR primers and probe.

Target	Forward Primer	Reverse Primer	Hydrolysis Probe
HoxA9	CGCCGCTCTCATTCTCAG	GCTTGTGGTTCTCCTCCAG	/56-FAM/AAACAACCCAGCGAAGGCGC/3IABkFQ/
GAPDH	GCCTGCCGGTGACTAAC	CATCACCCGGAGGAGAAATC	/56-FAM/TAGCCTCGCTCCACCTGACTTC/3IABkFQ/
EuNeg	GCTCCTGTAACCAACCACT	CTCTGGGCTGGCTTCATT	/56-FAM/ACCATATAGAGAAAGCCTGCTT/3IABkFQ/

D. melanogaster *genomic locus qPCR*

Table C.11: *D. melanogaster* genomic locus barcode qPCR primers and probe.

Target	Forward Primer	Reverse Primer	Hydrolysis Probe
Fz3	CCTATGCCAGGCAGGTAAAT	CTCAAAGTGTGGGATCTAGAAGG	/56-FAM/ACATCAGGC/ZEN/AGAAAGCAATGAAAGT/3IABkFQ/
Wg	CAGCGGAATTAATCGCACAAATA	GCGCACTATAAATGAGGCATAATC	/56-FAM/TGAGCAGCA/ZEN/ATATCGGCATACGCA/3IABkFQ/
Lab	AAACACGACTCCCGTTGG	TCAGTCACGACTTGGTAAGC	/56-FAM/ATGACGACG/ZEN/ACGACGTGCTG/3IABkFQ/

APPENDIX D: SEQUENCING DATASETS

This section lists details of the next-generation sequencing datasets generated for this work. Datasets used beyond those here listed were generated by Adrian Grzybowski, Ph.D.'18 or Bill Richter, Ph.D.'20.

Table D.1: Next-generation sequencing dataset reference information.

Identifier	Source	ChIP	Target	Antibody
AR11-1-Input_rep1_native	*D. melanogaster* S2 cell line	None	None	None
AR11-1-Input_rep2_native	*D. melanogaster* S2 cell line	None	None	None
AR11-1-Input_rep3_native	*D. melanogaster* S2 cell line	None	None	None
AR11-1-H3K4me3_rep1_native	*D. melanogaster* S2 cell line	Native	H3K4me3	AM 39159 Lot 12613005
AR11-1-H3K4me3_rep2_native	*D. melanogaster* S2 cell line	Native	H3K4me3	AM 39159 Lot 12613005
AR11-1-H3K4me3_rep3_native	*D. melanogaster* S2 cell line	Native	H3K4me3	AM 39159 Lot 12613005
AR11-1-H3K79me2_rep1_native	*D. melanogaster* S2 cell line	Native	H3K79me2	AB 3594 Lot GR173874
AR11-1-H3K79me2_rep2_native	*D. melanogaster* S2 cell line	Native	H3K79me2	AB 3594 Lot GR173874
AR11-1-H3K79me2_rep3_native	*D. melanogaster* S2 cell line	Native	H3K79me2	AB 3594 Lot GR173874
AR11-2-Input_rep1_denat	*D. melanogaster* S2 cell line	None	None	None
AR11-2-Input_rep2_denat	*D. melanogaster* S2 cell line	None	None	None
AR11-2-Input_rep3_denat	*D. melanogaster* S2 cell line	None	None	None
AR11-2-H3K4me3_rep1_denat	*D. melanogaster* S2 cell line	Denaturative	H3K4me3	AM 39159 Lot 12613005
AR11-2-H3K4me3_rep2_denat	*D. melanogaster* S2 cell line	Denaturative	H3K4me3	AM 39159 Lot 12613005

Table D.1 continues on next page.

Table D.1, continued:

Identifier	Source	ChIP	Target	Antibody
AR11-2-H3K4me3_rep3_denat	*D. melanogaster* S2 cell line	Denaturative	H3K4me3	AM 39159 Lot 12613005
AR11-2-H3K79me2_rep1_denat	*D. melanogaster* S2 cell line	Denaturative	H3K79me2	AB 3594 Lot GR173874
AR11-2-H3K79me2_rep2_denat	*D. melanogaster* S2 cell line	Denaturative	H3K79me2	AB 3594 Lot GR173874
AR11-2-H3K79me2_rep3_denat	*D. melanogaster* S2 cell line	Denaturative	H3K79me2	AB 3594 Lot GR173874
AR15-1-H3K4me3_native_untreated_1	MV4;11 DMSO (4d)	Native	H3K4me3	AM 39159 Lot 12613005
AR15-1-H3K4me3_native_untreated_2	MV4;11 DMSO (4d)	Native	H3K4me3	AM 39159 Lot 12613005
AR15-1-H3K4me3_native_untreated_3	MV4;11 DMSO (4d)	Native	H3K4me3	AM 39159 Lot 12613005
AR15-1-H3K4me3_denat_untreated_1	MV4;11 DMSO (4d)	Denaturative	H3K4me3	AM 39159 Lot 12613005
AR15-1-H3K4me3_denat_untreated_2	MV4;11 DMSO (4d)	Denaturative	H3K4me3	AM 39159 Lot 12613005
AR15-1-H3K4me3_denat_untreated_3	MV4;11 DMSO (4d)	Denaturative	H3K4me3	AM 39159 Lot 12613005
AR15-1-H3K79me2_denat_untreated_1	MV4;11 DMSO (4d)	Denaturative	H3K79me2	AB 3594 Lot GR173874
AR15-1- H3K79me2_denat_untreated_2	MV4;11 DMSO (4d)	Denaturative	H3K79me2	AB 3594 Lot GR173874
AR15-1- H3K79me2_denat_untreated_3	MV4;11 DMSO (4d)	Denaturative	H3K79me2	AB 3594 Lot GR173874
AR15-2-H3K4me3_native_treated	MV4;11 10 μM EPZ-5676 (4d)	Native	H3K4me3	AM 39159 Lot 12613005
AR15-2- H3K79me2_denat_treated	MV4;11 10 μM EPZ-5676 (4d)	Denaturative	H3K79me2	AB 3594 Lot GR173874
AR15-3_untreated_Input	MV4;11 DMSO (4d)	None	None	None
AR15-4_treated_Input	MV4;11 10 μM EPZ-5676 (4d)	None	None	None

Table D.1 continues on next page.

Table D.1, continued:

Identifier	Source	ChIP	Target	Antibody
AR16-1-Input	K562	None	None	None
AR16-2-AB-8895	K562	Native	H3K4me1	AB 8895 Lot GR305231-1
AR16-2-EMD-05-1338	K562	Native	H3K4me2	EMD 05-1338 Lot 2757107
AR16-2-TF-710795	K562	Native	H3K4me1	TF 710795 Lot QL230603
AR16-3-AB-7766	K562	Native	H3K4me2	AB 7766 Lot GR289627-1
AR16-3-AM-39635	K562	Native	H3K4me1	AM 39635 Lot 30615011
AR16-3-CST-9725	K562	Native	H3K4me2	CST 9725 Lot 9
AR16-4-AB-12209	K562	Native	H3K4me3	AB 12209 Lot GR275790-1
AR16-4-AB-8580	K562	Native	H3K4me3	AB 8580 Lot GR190229-1
AR16-4-ABC-46698	K562	Native	H3K4me3	ABC A2357 Lot 46698
AR16-4-EMD-07-473	K562	Native	H3K4me3	EMD 07-473 Lot DAM1623866
AR16-4-TF-PA5-40086	K562	Native	H3K4me3	TF PA5-40086 Lot RL2301825
AR16-5-CST-5326BF	K562	Native	H3K4me1	CST 5326BF Lot 2
AR16-5-EMD-05-745R	K562	Native	H3K4me3	EMD 05-745R Lot 2813867
AR16-5-EPG-A-4031-050	K562	Native	H3K4me1	EPG A-4031-050 Lot 606359
AR16-5-TF-710796	K562	Native	H3K4me2	TF 710796 Lot QL230606
AR17-1_K562_Input	K562	None	None	None
AR17-2-CST-C36B11-K562	K562	Native	H3K27me3	CST 5326 Lot 8
AR17-2-DIA-15410003-K562	K562	Native	H3K4me3	DIA C15410003 Lot A1052D

Table D.1 continues on next page.

Table D.1, continued:

Identifier	Source	ChIP	Target	Antibody
AR17-3-EPG-A-4033-053-K562	K562	Native	H3K4me3	EPG A-4033-050 Lot 606361
AR17-3-Taka-309M3B-K562	K562	Native	H3K9me3	KL 309M3B Lot 072913TH
AR17-4and5_WT-R1-mESC-Input	mESC R1	None	None	None
AR17-6and7_dCD-R1-mESC_Input	mESC R1 MLL3/4 dCD	None	None	None
AR17-8-CST-5326-WT-mESC	mESC R1	Native	H3K4me1	CST 5326 Lot 1
AR17-8-CST-5326-dCD-mESC	mESC R1 MLL3/4 dCD	Native	H3K4me1	CST 5326 Lot 1
AR18-1-Primed_Input	mESC E14 Serum/LIF	None	None	None
AR18-2-Primed_Bivalent	mESC E14 Serum/LIF	reICeChIP	H3K4me3/ H3K27me3	KL 304M3B-1xHRV3C/ CST 5326
AR18-2-Primed_H3K27me3	mESC E14 Serum/LIF	Native	H3K27me3	CST 5326 Lost 8
AR18-2-Primed_H3K4me3	mESC E14 Serum/LIF	Native	H3K4me3	KL 304M3B-1xHRV3C Lot 103015AG
AR18-2-Primed_H3K9me3	mESC E14 Serum/LIF	Native	H3K9me3	KL 309M3B Lot 072913TH
AR18-3-NPC_Input	NPC	None	None	None
AR18-4-NPC_Bivalent	NPC	reICeChIP	H3K4me3/ H3K27me3	KL 304M3B-1xHRV3C/ CST 5326
AR18-4-NPC_H3K27me3	NPC	Native	H3K27me3	CST 5326 Lost 8
AR18-4-NPC_H3K4me3	NPC	Native	H3K4me3	KL 304M3B-1xHRV3C Lot 103015AG
AR18-4-NPC_H3K9me3	NPC	Native	H3K9me3	KL 309M3B Lot 072913TH
AR19-1-RS411-Input	RS4;11	None	None	None

Table D.1 continues on next page.

Table D.1, continued:

Identifier	Source	ChIP	Target	Antibody
AR19-1-RS411-H3K79me2	RS4;11	Denaturative	H3K79me2	CST 5427 Lot 4
AR19-2-Kopn8-Input	Kopn8	None	None	None
AR19-2-Kopn8-H3K79me2	Kopn8	Denaturative	H3K79me2	CST 5427 Lot 4
AR19-3-K562-Input	K562	None	None	None
AR19-3-K562-H3K79me2	K562	Denaturative	H3K79me2	CST 5427 Lot 4
AR19-4-Molm13-Input	Molm13	None	None	None
AR19-4-Molm13-H3K79me2	Molm13	Denaturative	H3K79me2	CST 5427 Lot 4
AR19-5-THP1-Input	THP1	None	None	None
AR19-5-THP1-H3K79me2	THP1	Denaturative	H3K79me2	CST 5427 Lot 4
AR19-6-SEM-Input	SEM	None	None	None
AR19-6-SEM-H3K79me2	SEM	Denaturative	H3K79me2	CST 5427 Lot 4

REFERENCES

1. Felsenfeld, G. A Brief History of Epigenetics. *Cold Spring Harb Perspect Biol* **6,** a018200 (2014).

2. Morgan, T. H. An Attempt to Analyze the Constitution of the Chromosomes on the Basis of Sex-Limited Inheritance in Drosophila. *J Exp Zool* **11,** 365–413 (1911).

3. Avery, O. T., MacLeod, C. M. & McCarty, M. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types. *J Exp Med* **79,** 137–158 (1944).

4. Hershey, A. D. & Chase, M. Independent Functions of Viral Protein and Nucleic Acid in Growth of Bacteriophage. *J Gen Physiol* **36,** 39–56 (1952).

5. Waddington, C. H. *The Strategy of the Genes* 262 pp. (Allen & Unwin, London, 1957).

6. Laskey, R. A. & Gurdon, J. B. Genetic Content of Adult Somatic Cells Tested by Nuclear Transplantation from Cultured Cells. *Nature* **228,** 1332–1334 (1970).

7. Bird, A. Perceptions of Epigenetics. *Nature* **447,** 396–398 (2007).

8. Riggs, A. D., Martienssen, R. A. & Russo, V. E. in *Epigenetic Mechanisms of Gene Regulation* 1–4 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1996).

9. Bernstein, B. E. *et al.* A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell* **125,** 315–326 (2006).

10. Bonn, S. *et al.* Tissue-Specific Analysis of Chromatin State Identifies Temporal Signatures of Enhancer Activity during Embryonic Development. *Nat Genet* **44,** 148–156 (2012).

11. Boyer, L. A. *et al.* Polycomb Complexes Repress Developmental Regulators in Murine Embryonic Stem Cells. *Nature* **441,** 349–353 (2006).

12. Creyghton, M. P. *et al.* Histone H3K27ac Separates Active from Poised Enhancers and Predicts Developmental State. *Proc Natl Acad Sci USA* **107,** 21931–21936 (2010).

13. Liyanage, V. R. B. & Rastegar, M. Rett Syndrome and MeCP2. *Neuromolecular Med* **16,** 231–264 (2014).

14. Vastenhouw, N. L. & Schier, A. F. Bivalent Histone Modifications in Early Embryogenesis. *Curr Opin Cell Biol* **24,** 374–386 (2012).

15. Vastenhouw, N. L. *et al.* Chromatin Signature of Embryonic Pluripotency Is Established during Genome Activation. *Nature* **464,** 922–926 (2010).

16. Francis, N. J., Kingston, R. E. & Woodcock, C. L. Chromatin Compaction by a Polycomb Group Protein Complex. *Science* **306,** 1574–1577 (2004).

17. Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R. & Young, R. A. A Chromatin Landmark and Transcription Initiation at Most Promoters in Human Cells. *Cell* **130,** 77–88 (2007).

18. Heintzman, N. D. *et al.* Distinct and Predictive Chromatin Signatures of Transcriptional Promoters and Enhancers in the Human Genome. *Nat Genet* **39,** 311–318 (2007).

19. Heintzman, N. D. *et al.* Histone Modifications at Human Enhancers Reflect Global Cell-Type-Specific Gene Expression. *Nature* **459,** 108–112 (2009).

20. Santos-Rosa, H. *et al.* Active Genes Are Tri-Methylated at K4 of Histone H3. *Nature* **419,** 407–411 (2002).

21. Bennett, R. L. *et al.* A Mutation in Histone H2B Represents a New Class of Oncogenic Driver. *Cancer Discov* **9,** 1438–1451 (2019).

22. Criscione, S. W., Zhang, Y., Thompson, W., Sedivy, J. M. & Neretti, N. Transcriptional Landscape of Repetitive Elements in Normal and Cancer Human Cells. *BMC Genomics* **15,** 583 (2014).

23. Krivtsov, A. V. *et al.* H3K79 Methylation Profiles Define Murine and Human MLL-AF4 Leukemias. *Cancer Cell* **14,** 355–368 (2008).

24. Mitelman, F., Johansson, B. & Mertens, F. The Impact of Translocations and Gene Fusions on Cancer Causation. *Nat Rev Cancer* **7,** 233–245 (2007).

25. Patani, H. *et al.* Transition to Naïve Human Pluripotency Mirrors Pan-Cancer DNA Hypermethylation. *Nat Commun* **11,** 3671 (2020).

26. Kornberg, R. D. & Thonmas, J. O. Chromatin Structure: Oligomers of the Histones. *Science* **184,** 865–868 (1974).

27. Kornberg, R. D. & Lorch, Y. Twenty-Five Years of the Nucleosome, Fundamental Particle of the Eukaryote Chromosome. *Cell* **98,** 285–294 (1999).

28. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal Structure of the Nucleosome Core Particle at 2.8 A Resolution. *Nature* **389,** 18 (1997).

29. Luger, K., Dechassa, M. L. & Tremethick, D. J. New Insights into Nucleosome and Chromatin Structure: An Ordered State or a Disordered Affair? *Nat Rev Mol Cell Biol* **13,** 436–447 (2012).

30. Ou, H. D. *et al.* ChromEMT: Visualizing 3D Chromatin Structure and Compaction in Interphase and Mitotic Cells. *Science* **357,** eaag0025 (2017).

31. Jones, P. A. *et al.* Moving AHEAD with an International Human Epigenome Project. *Nature* **454,** 711–715 (2008).

32. Kumar, S., Chinnusamy, V. & Mohapatra, T. Epigenetics of Modified DNA Bases: 5-Methylcytosine and Beyond. *Front Genet* **9,** 640 (2018).

33. Lyko, F. The DNA Methyltransferase Family: A Versatile Toolkit for Epigenetic Regulation. *Nat Rev Genet* **19,** 81–92 (2018).

34. Bird, A. P. Use of Restriction Enzymes to Study Eukaryotic DNA Methylation: II. The Symmetry of Methylated Sites Supports Semi-Conservative Copying of the Methylation Pattern. *J Mol Biol* **118,** 49–60 (1978).

35. Bird, A. P. & Southern, E. M. Use of Restriction Enzymes to Study Eukaryotic DNA Methylation: I. The Methylation Pattern in Ribosomal DNA from Xenopus Laevis. *J Mol Biol* **118,** 27–47 (1978).

36. Holliday, R. & Pugh, J. E. DNA Modification Mechanisms and Gene Activity during Development. *Science* **187,** 226–232 (1975).

37. Riggs, A. D. X Inactivation, Differentiation, and DNA Methylation. *Cytogenet Genome Res* **14,** 9–25 (1975).

38. Suelves, M., Carrió, E., Núñez-Álvarez, Y. & Peinado, M. A. DNA Methylation Dynamics in Cellular Commitment and Differentiation. *Brief Funct Genomics* **15,** 443–453 (2016).

39. Huda, A., Mariño-Ramírez, L. & Jordan, I. K. Epigenetic Histone Modifications of Human Transposable Elements: Genome Defense versus Exaptation. *Mobile DNA* **1,** 2 (2010).

40. Bell, A. C. & Felsenfeld, G. Methylation of a CTCF-dependent Boundary Controls Imprinted Expression of the Igf2 Gene. *Nature* **405,** 482–485 (2000).

41. Hark, A. T. *et al.* CTCF Mediates Methylation-Sensitive Enhancer-Blocking Activity at the H19/Igf2 Locus. *Nature* **405,** 486–489 (2000).

42. Kanduri, C. *et al.* Functional Association of CTCF with the Insulator Upstream of the H19 Gene Is Parent of Origin-Specific and Methylation-Sensitive. *Curr Biol* **10,** 853–856 (2000).

43. Greenberg, M. V. C. & Bourc'his, D. The Diverse Roles of DNA Methylation in Mammalian Development and Disease. *Nat Rev Mol Cell Biol* **20,** 590–607 (2019).

44. Branco, M. R., Ficz, G. & Reik, W. Uncovering the Role of 5-Hydroxymethylcytosine in the Epigenome. *Nat Rev Genet* **13,** 7–13 (2012).

45. Ficz, G. *et al.* Dynamic Regulation of 5-Hydroxymethylcytosine in Mouse ES Cells and during Differentiation. *Nature* **473,** 398–402 (2011).

46. Hahn, M. A., Szabó, P. E. & Pfeifer, G. P. 5-Hydroxymethylcytosine: A Stable or Transient DNA Modification? *Genomics* **104,** 314–323 (2014).

47. Kriaucionis, S. & Heintz, N. The Nuclear DNA Base 5-Hydroxymethylcytosine Is Present in Purkinje Neurons and the Brain. *Science* **324,** 929–930 (2009).

48. Pastor, W. A. *et al.* Genome-Wide Mapping of 5-Hydroxymethylcytosine in Embryonic Stem Cells. *Nature* **473,** 394–397 (2011).

49. Tahiliani, M. *et al.* Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1. *Science* **324,** 930–935 (2009).

50. Li, Z., Zhao, P. & Xia, Q. Epigenetic Methylations on N6-Adenine and N6-Adenosine with the Same Input but Different Output. *Int J Mol Sci* **20,** 2931 (2019).

51. Hangauer, M. J., Vaughn, I. W. & McManus, M. T. Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding RNAs. *PLOS Genet* **9,** e1003569 (2013).

52. Jarroux, J., Morillon, A. & Pinskaya, M. in *Long Non Coding RNA Biology* (ed Rao, M.) 1–46 (Springer, Singapore, 2017). ISBN: 978-981-10-5203-3.

53. Dinger, M. E., Amaral, P. P., Mercer, T. R. & Mattick, J. S. Pervasive Transcription of the Eukaryotic Genome: Functional Indices and Conceptual Implications. *Brief Funct Genomics* **8,** 407–423 (2009).

54. Djebali, S. *et al.* Landscape of Transcription in Human Cells. *Nature* **489,** 101–108 (2012).

55. Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S. & Brockdorff, N. Requirement for Xist in X Chromosome Inactivation. *Nature* **379,** 131–137 (1996).

56. Kalantry, S., Purushothaman, S., Bowen, R. B., Starmer, J. & Magnuson, T. Evidence of Xist RNA-independent Initiation of Mouse Imprinted X-chromosome Inactivation. *Nature* **460,** 647–651 (2009).

57. Loda, A. & Heard, E. Xist RNA in Action: Past, Present, and Future. *PLOS Genet* **15,** e1008333 (2019).

58. Core, L. J. *et al.* Analysis of Nascent RNA Identifies a Unified Architecture of Initiation Regions at Mammalian Promoters and Enhancers. *Nat Genet* **46,** 1311–1320 (2014).

59. Kim, T.-K. *et al.* Widespread Transcription at Neuronal Activity-Regulated Enhancers. *Nature* **465,** 182–187 (2010).

60. Ørom, U. A. *et al.* Long Noncoding RNAs with Enhancer-like Function in Human Cells. *Cell* **143,** 46–58 (2010).

61. Arnold, P. R., Wells, A. D. & Li, X. C. Diversity and Emerging Roles of Enhancer RNA in Regulation of Gene Expression and Cell Fate. *Front Cell Dev Biol* **7,** 377 (2020).

62. Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. Identification of Novel Genes Coding for Small Expressed RNAs. *Science* **294,** 853–858 (2001).

63. Lee, R. C. & Ambros, V. An Extensive Class of Small RNAs in Caenorhabditis Elegans. *Science* **294,** 862–864 (2001).

64. Lee, R. C., Feinbaum, R. L. & Ambros, V. The C. Elegans Heterochronic Gene Lin-4 Encodes Small RNAs with Antisense Complementarity to Lin-14. *Cell* **75,** 843–854 (1993).

65. Lee, H.-C. *et al.* C. Elegans piRNAs Mediate the Genome-wide Surveillance of Germline Transcripts. *Cell* **150,** 78–87 (2012).

66. Zhang, P., Wu, W., Chen, Q. & Chen, M. Non-Coding RNAs and Their Integrated Networks. *J Integr Bioinform* **16,** 20190027 (2019).

67. Zhang, D. *et al.* The piRNA Targeting Rules and the Resistance to piRNA Silencing in Endogenous Genes. *Science* **359,** 587–592 (2018).

68. Sun, X., Liu, J., Xu, C., Tang, S.-C. & Ren, H. The Insights of Let-7 miRNAs in Oncogenesis and Stem Cell Potency. *J Cell Mol Med* **20,** 1779–1788 (2016).

69. Sun, X. *et al.* Chromatin-Enriched RNAs Mark Active and Repressive Cis-Regulation: An Analysis of Nuclear RNA-seq. *PLOS Comput Biol* **16,** e1007119 (2020).

70. Werner, M. S. *et al.* Chromatin-Enriched lncRNAs Can Act as Cell-Type Specific Activators of Proximal Gene Transcription. *Nat Struct Mol Biol* **24,** 596–603 (2017).

71. Werner, M. S. & Ruthenburg, A. J. Nuclear Fractionation Reveals Thousands of Chromatin-Tethered Noncoding RNAs Adjacent to Active Genes. *Cell Rep* **12,** 1089–1098 (2015).

72. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-binding Proteins and Nucleosome Position. *Nat Meth* **10,** 1213–1218 (2013).

73. Kouzarides, T. Chromatin Modifications and Their Function. *Cell* **128,** 693–705 (2007).

74. Strahl, B. D. & Allis, C. D. The Language of Covalent Histone Modifications. *Nature* **403,** 41–45 (2000).

75. Werner, M. & Ruthenburg, A. J. The United States of Histone Ubiquitylation and Methylation. *Mol Cell* **43,** 5–7 (2011).

76. Bannister, A. J. & Kouzarides, T. Regulation of Chromatin by Histone Modifications. *Cell Res* **21,** 381–395 (2011).

77. Sawicka, A. & Seiser, C. Histone H3 Phosphorylation – A Versatile Chromatin Modification for Different Occasions. *Biochimie* **94,** 2193–2201 (2012).

78. Chen, H., Lin, R. J., Xie, W., Wilpitz, D. & Evans, R. M. Regulation of Hormone-Induced Histone Hyperacetylation and Gene Activation via Acetylation of an Acetylase. *Cell* **98,** 675–686 (1999).

79. Grunstein, M. Histone Acetylation in Chromatin Structure and Transcription. *Nature* **389,** 349–352 (1997).

80. Hebbes, T. R., Thorne, A. W. & Crane-Robinson, C. A Direct Link between Core Histone Acetylation and Transcriptionally Active Chromatin. *EMBO J* **7,** 1395–1402 (1988).

81. Lavarone, E., Barbieri, C. M. & Pasini, D. Dissecting the Role of H3K27 Acetylation and Methylation in PRC2 Mediated Control of Cellular Identity. *Nat Commun* **10,** 1679 (2019).

82. Weake, V. M. & Workman, J. L. Histone Ubiquitination: Triggering Gene Activity. *Mol Cell* **29,** 653–663 (2008).

83. Wojcik, F. *et al.* Functional Crosstalk between Histone H2B Ubiquitylation and H2A Modifications and Variants. *Nat Commun* **9,** 1394 (2018).

84. Azad, G. K., Swagatika, S., Kumawat, M., Kumawat, R. & Tomar, R. S. Modifying Chromatin by Histone Tail Clipping. *J Mol Biol* **430,** 3051–3067 (2018).

85. Duncan, E. M. *et al.* Cathepsin L Proteolytically Processes Histone H3 During Mouse Embryonic Stem Cell Differentiation. *Cell* **135,** 284–294 (2008).

86. Santos-Rosa, H. *et al.* Histone H3 Tail Clipping Regulates Gene Expression. *Nat Struct Mol Biol* **16,** 17–22 (2009).

87. Hassan, A. H. *et al.* Function and Selectivity of Bromodomains in Anchoring Chromatin-Modifying Complexes to Promoter Nucleosomes. *Cell* **111,** 369–379 (2002).

88. Sanchez, R. & Zhou, M.-M. The Role of Human Bromodomains in Chromatin Biology and Gene Transcription. *Curr Opin Drug Discov Devel* **12,** 659–665 (2009).

89. Lorton, B. M. *et al.* A Binary Arginine Methylation Switch on Histone H3 Arginine 2 Regulates Its Interaction with WDR5. *Biochemistry* **59,** 3696–3708 (2020).

90. Lorton, B. M. & Shechter, D. Cellular Consequences of Arginine Methylation. *Cell Mol Life Sci* **76,** 2933–2956 (2019).

91. Ruthenburg, A. J. *et al.* Recognition of a Mononucleosomal Histone Modification Pattern by BPTF via Multivalent Interactions. *Cell* **145,** 692–706 (2011).

92. Wysocka, J. *et al.* A PHD Finger of NURF Couples Histone H3 Lysine 4 Trimethylation with Chromatin Remodelling. *Nature* **442,** 86–90 (2006).

93. Vermeulen, M. *et al.* Selective Anchoring of TFIID to Nucleosomes by Trimethylation of Histone H3 Lysine 4. *Cell* **131,** 58–69 (2007).

94. Lauberth, S. M. *et al.* H3K4me3 Interactions with TAF3 Regulate Preinitiation Complex Assembly and Selective Gene Activation. *Cell* **152,** 1021–1036 (2013).

95. Kang, X. *et al.* SUMO-Specific Protease 2 Is Essential for Suppression of Polycomb Group Protein-Mediated Gene Silencing during Embryonic Development. *Mol Cell* **38,** 191–201 (2010).

96. Bernstein, E. *et al.* Mouse Polycomb Proteins Bind Differentially to Methylated Histone H3 and RNA and Are Enriched in Facultative Heterochromatin. *Mol Cell Biol* **26,** 2560–2569 (2006).

97. Morey, L. *et al.* Nonoverlapping Functions of the Polycomb Group Cbx Family of Proteins in Embryonic Stem Cells. *Cell Stem Cell* **10,** 47–62 (2012).

98. Al-Sady, B., Madhani, H. D. & Narlikar, G. J. Division of Labor between the Chromodomains of HP1 and Suv39 Methylase Enables Coordination of Heterochromatin Spread. *Mol Cell* **51,** 80–91 (2013).

99. Bannister, A. J. *et al.* Selective Recognition of Methylated Lysine 9 on Histone H3 by the HP1 Chromo Domain. *Nature* **410,** 120–124 (2001).

100. Lomberk, G., Wallrath, L. & Urrutia, R. The Heterochromatin Protein 1 Family. *Genome Biol* **7,** 228 (2006).

101. Martens, J. H. *et al.* The Profile of Repeat-Associated Histone Lysine Methylation States in the Mouse Epigenome. *EMBO J* **24,** 800–812 (2005).

102. Ruthenburg, A. J., Allis, C. D. & Wysocka, J. Methylation of Lysine 4 on Histone H3: Intricacy of Writing and Reading a Single Epigenetic Mark. *Mol Cell* **25,** 15–30 (2007).

103. Chervona, Y. & Costa, M. Histone Modifications and Cancer: Biomarkers of Prognosis? *Am J Cancer Res* **2,** 589–597 (2012).

104. Dawson, M. A. & Kouzarides, T. Cancer Epigenetics: From Mechanism to Therapy. *Cell* **150,** 12–27 (2012).

105. Mahmood, T. & Yang, P.-C. Western Blot: Technique, Theory, and Trouble Shooting. *N Am J Med* **4,** 429 (2012).

106. Pillai-Kastoori, L., Schutz-Geschwender, A. R. & Harford, J. A. A Systematic Approach to Quantitative Western Blot Analysis. *Anal Biochem* **593,** 113608 (2020).

107. Taylor, S. C. & Posch, A. The Design of a Quantitative Western Blot Experiment. *Biomed Res Int* **2014,** 361590 (2014).

108. Egelhofer, T. A. *et al.* An Assessment of Histone-Modification Antibody Quality. *Nat Struct Mol Biol* **18,** 91–93 (2011).

109. Bailey, L. J. *et al.* Applications for an Engineered Protein-G Variant with a pH Controllable Affinity to Antibody Fragments. *J Immunol Meth* **415,** 24–30 (2014).

110. Cornett, E. M., Dickson, B. M. & Rothbart, S. B. Analysis of Histone Antibody Specificity with Peptide Microarrays. *J Vis Exp,* e55912–e55912 (2017).

111. Fuchs, S. M., Krajewski, K., Baker, R. W., Miller, V. L. & Strahl, B. D. Influence of Combinatorial Histone Modifications on Antibody and Effector Protein Recognition. *Curr Biol* **21,** 53–58 (2011).

112. Rothbart, S. B. *et al.* An Interactive Database for the Assessment of Histone Antibody Specificity. *Mol Cell* **59,** 502–511 (2015).

113. Noberini, R., Robusti, G. & Bonaldi, T. Mass Spectrometry-Based Characterization of Histones in Clinical Samples: Applications, Progresses, and Challenges. *FEBS J* (2021).

114. Su, X., Ren, C. & Freitas, M. A. Mass Spectrometry-Based Strategies for Characterization of Histones and Their Post-Translational Modifications. *Expert Rev Proteomic* **4,** 211–225 (2007).

115. Garcia, B. A. *et al.* Chemical Derivatization of Histones for Facilitated Analysis by Mass Spectrometry. *Nat Protoc* **2,** 933–938 (2007).

116. Young, N. L. *et al.* High Throughput Characterization of Combinatorial Histone Codes. *Mol Cell Proteom* **8,** 2266–2284 (2009).

117. Schachner, L. F. *et al.* Decoding the Protein Composition of Whole Nucleosomes with Nuc-MS. *Nat Meth* **18,** 303–308 (2021).

118. Grzybowski, A. T., Chen, Z. & Ruthenburg, A. J. Calibrating ChIP-Seq with Nucleosomal Internal Standards to Measure Histone Modification Density Genome Wide. *Mol Cell* **58,** 886–899 (2015).

119. Shema, E. *et al.* Single-Molecule Decoding of Combinatorially Modified Nucleosomes. *Science* **352,** 717–721 (2016).

120. Solomon, M. J., Larsen, P. L. & Varshavsky, A. Mapping proteinDNA Interactions in Vivo with Formaldehyde: Evidence That Histone H4 Is Retained on a Highly Transcribed Gene. *Cell* **53,** 937–947 (1988).

121. Skene, P. J. & Henikoff, S. An Efficient Targeted Nuclease Strategy for High-Resolution Mapping of DNA Binding Sites. *eLife* **6** (ed Reinberg, D.) e21856 (2017).

122. Kaya-Okur, H. S. *et al.* CUT&Tag for Efficient Epigenomic Profiling of Small Samples and Single Cells. *Nat Commun* **10,** 1930 (2019).

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

124. Shah, R. N. *et al.* Examining the Roles of H3K4 Methylation States with Systematically Characterized Antibodies. *Mol Cell* **72,** 162–177 (2018).

125. Gilmour, D. S. & Lis, J. T. Detecting Protein-DNA Interactions in Vivo: Distribution of RNA Polymerase on Specific Bacterial Genes. *Proc Natl Acad Sci USA* **81,** 4275–4279 (1984).

126. Barski, A. *et al.* High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129,** 823–837 (2007).

127. Consortium, T. E. P. An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature* **489,** 57–74 (2012).

128. Orlando, D. A. *et al.* Quantitative ChIP-Seq Normalization Reveals Global Modulation of the Epigenome. *Cell Rep* **9,** 1163–1170 (2014).

129. Landt, S. G. *et al.* ChIP-seq Guidelines and Practices of the ENCODE and modENCODE Consortia. *Genome Res* **22,** 1813–1831 (2012).

130. Mikkelsen, T. S. *et al.* Genome-Wide Maps of Chromatin State in Pluripotent and Lineage-Committed Cells. *Nature* **448,** 553–560 (2007).

131. Rada-Iglesias, A. *et al.* A Unique Chromatin Signature Uncovers Early Developmental Enhancers in Humans. *Nature* **470,** 279–283 (2011).

132. Schübeler, D. *et al.* The Histone Modification Pattern of Active Genes Revealed through Genome-Wide Chromatin Analysis of a Higher Eukaryote. *Genes Dev* **18,** 1263–1271 (2004).

133. Bock, I. *et al.* Detailed Specificity Analysis of Antibodies Binding to Modified Histone Tails with Peptide Arrays. *Epigenetics* **6,** 256–263 (2011).

134. Nishikori, S. *et al.* Broad Ranges of Affinity and Specificity of Anti-Histone Antibodies Revealed by a Quantitative Peptide Immunoprecipitation Assay. *J Mol Biol* **424,** 391–399 (2012).

135. LeRoy, G. *et al.* A Quantitative Atlas of Histone Modification Signatures from Human Cancer Cells. *Epigenetics Chromatin* **6,** 20 (2013).

136. Wang, Y., Li, X. & Hu, H. H3K4me2 Reliably Defines Transcription Factor Binding Regions in Different Cells. *Genomics* **103,** 222–228 (2014).

137. Fang, R. *et al.* Human LSD2/KDM1b/AOF1 Regulates Gene Transcription by Modulating Intragenic H3K4me2 Methylation. *Mol Cell* **39,** 222–233 (2010).

138. Pekowska, A., Benoukraf, T., Ferrier, P. & Spicuglia, S. A Unique H3K4me2 Profile Marks Tissue-Specific Gene Regulation. *Genome Res* **20,** 1493–1502 (2010).

139. Matthews, A. G. W. *et al.* RAG2 PHD Finger Couples Histone H3 Lysine 4 Trimethylation with V(D)J Recombination. *Nature* **450,** 1106 (2007).

140. Baudat, F., Imai, Y. & de Massy, B. Meiotic Recombination in Mammals: Localization and Regulation. *Nat Rev Genet* **14,** 794–806 (2013).

141. Bieberstein, N. I., Carrillo Oesterreich, F., Straube, K. & Neugebauer, K. M. First Exon Length Controls Active Chromatin Signatures and Transcription. *Cell Rep* **2,** 62–68 (2012).

142. Sims III, R. J. *et al.* Recognition of Trimethylated Histone H3 Lysine 4 Facilitates the Recruitment of Transcription Postinitiation Factors and Pre-mRNA Splicing. *Mol Cell* **28,** 665–676 (2007).

143. Uhlen, M. *et al.* A Proposal for Validation of Antibodies. *Nat Meth* **13,** 823–827 (2016).

144. Taverna, S. D. *et al.* Yng1 PHD Finger Binding to H3 Trimethylated at K4 Promotes NuA3 HAT Activity at K14 of H3 and Transcription at a Subset of Targeted ORFs. *Mol Cell* **24,** 785–796 (2006).

145. Voigt, P. *et al.* Asymmetrically Modified Nucleosomes. *Cell* **151,** 181–193 (2012).

146. Sanders, Y. Y. *et al.* Histone Modifications in Senescence-Associated Resistance to Apoptosis by Oxidative Stress. *Redox Biol* **1,** 8–16 (2013).

147. Li, G. *et al.* Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation. *Cell* **148,** 84–98 (2012).

148. Pekowska, A. *et al.* H3K4 Tri-methylation Provides an Epigenetic Signature of Active Enhancers. *EMBO J* **30,** 4198–4210 (2011).

149. Fan, X. & Struhl, K. Where Does Mediator Bind In Vivo? *PLOS ONE* **4,** e5029 (2009).

150. Kasinathan, S., Orsi, G. A., Zentner, G. E., Ahmad, K. & Henikoff, S. High-Resolution Mapping of Transcription Factor Binding Sites on Native Chromatin. *Nat Meth* **11,** 203–209 (2014).

151. Teytelman, L., Thurtle, D. M., Rine, J. & van Oudenaarden, A. Highly Expressed Loci Are Vulnerable to Misleading ChIP Localization of Multiple Unrelated Proteins. *Proc Natl Acad Sci USA* **110,** 18602–18607 (2013).

152. O'Neill, L. P. & Turner, B. M. Immunoprecipitation of Native Chromatin: NChIP. *Methods. Histone Modifications* **31,** 76–82 (2003).

153. Dorighi, K. M. *et al.* Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. *Mol Cell* **66,** 568–576.e4 (2017).

154. Koch, F. *et al.* Transcription Initiation Platforms and GTF Recruitment at Tissue-Specific Enhancers and Promoters. *Nat Struct Mol Biol* **18,** 956–963 (2011).

155. Benayoun, B. A. *et al.* H3K4me3 Breadth Is Linked to Cell Identity and Transcriptional Consistency. *Cell* **158,** 673–688 (2014).

156. Chen, K. *et al.* Broad H3K4me3 Is Associated with Increased Transcription Elongation and Enhancer Activity at Tumor-Suppressor Genes. *Nat Genet* **47,** 1149–1157 (2015).

157. Dahl, J. A. *et al.* Broad Histone H3K4me3 Domains in Mouse Oocytes Modulate Maternal-to-Zygotic Transition. *Nature* **537,** 548–552 (2016).

158. Rothbart, S. B. *et al.* Poly-Acetylated Chromatin Signatures Are Preferred Epitopes for Site-Specific Histone H4 Acetyl Antibodies. *Sci Rep* **2,** 489 (2012).

159. Kinkley, S. *et al.* reChIP-seq Reveals Widespread Bivalency of H3K4me3 and H3K27me3 in $CD4^+$ Memory T Cells. *Nat Commun* **7,** 12514 (2016).

160. Seenundun, S. *et al.* UTX Mediates Demethylation of H3K27me3 at Muscle-Specific Genes during Myogenesis. *EMBO J* **29,** 1401–1411 (2010).

161. Sen, S., Block, K. F., Pasini, A., Baylin, S. B. & Easwaran, H. Genome-Wide Positioning of Bivalent Mononucleosomes. *BMC Med Genomics* **9,** 60 (2016).

162. Weiner, A. *et al.* Co-ChIP Enables Genome-Wide Mapping of Histone Mark Co-Occurrence at Single-Molecule Resolution. *Nat Biotechnol* **34,** 953–961 (2016).

163. Jin, C. & Felsenfeld, G. Nucleosome Stability Mediated by Histone Variants H3.3 and H2A.Z. *Genes Dev* **21,** 1519–1529 (2007).

164. Neumann, H. *et al.* A Method for Genetically Installing Site-Specific Acetylation in Recombinant Histones Defines the Effects of H3 K56 Acetylation. *Mol Cell* **36,** 153–163 (2009).

165. Baker, M. Reproducibility Crisis: Blame It on the Antibodies. *Nat News* **521,** 274 (2015).

166. Dyer, P. N. *et al.* Reconstitution of Nucleosome Core Particles from Recombinant Histones and DNA. *Meth Enzymol* **375,** 23–44 (2004).

167. Luger, K., Rechsteiner, T. J. & Richmond, T. J. Expression and Purification of Recombinant Histones and Nucleosome Reconstitution. *Meth Mol Biol* **119,** 1–16 (1999).

168. Lowary, P. T. & Widom, J. New DNA Sequence Rules for High Affinity Binding to Histone Octamer and Sequence-Directed Nucleosome Positioning. *J Mol Biol* **276,** 19–42 (1998).

169. Dickson, B. M., Cornett, E. M., Ramjan, Z. & Rothbart, S. B. ArrayNinja: An Open Source Platform for Unified Planning and Analysis of Microarray Experiments. *Meth Enzymol* **574,** 53–77 (2016).

170. Brand, M., Rampalli, S., Chaturvedi, C.-P. & Dilworth, F. J. Analysis of Epigenetic Modifications of Chromatin at Specific Gene Loci by Native Chromatin Immunoprecipitation of Nucleosomes Isolated Using Hydroxyapatite Chromatography. *Nat Protoc* **3,** 398–409 (2008).

171. Grzybowski, A. T., Shah, R. N., Richter, W. F. & Ruthenburg, A. J. Native Internally Calibrated Chromatin Immunoprecipitation for Quantitative Studies of Histone Post-Translational Modifications. *Nat Protoc* **14,** 3275–3302 (2019).

172. Langmead, B. & Salzberg, S. L. Fast Gapped-Read Alignment with Bowtie 2. *Nat Meth* **9,** 357–359 (2012).

173. Li, H. *et al.* The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

174. Quinlan, A. R. & Hall, I. M. BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features. *Bioinformatics* **26,** 841–842 (2010).

175. Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: Enabling Browsing of Large Distributed Datasets. *Bioinformatics* **26,** 2204–2207 (2010).

176. Zhang, Y. *et al.* Model-Based Analysis of ChIP-Seq (MACS). *Genome Biol* **9,** R137 (2008).

177. Buenrostro, J. D. *et al.* Single-Cell Chromatin Accessibility Reveals Principles of Regulatory Variation. *Nature* **523,** 486–490 (2015).

178. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol Cell* **38,** 576–589 (2010).

179. Mi, H., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. Large-Scale Gene Function Analysis with the PANTHER Classification System. *Nat Protoc* **8,** 1551–1566 (2013).

180. O'Carroll, D. *et al.* The Polycomb-Group GeneEzh2 Is Required for Early Mouse Development. *Mol Cell Biol* **21,** 4330–4336 (2001).

181. Pasini, D., Bracken, A. P., Jensen, M. R., Lazzerini Denchi, E. & Helin, K. Suz12 Is Essential for Mouse Development and for EZH2 Histone Methyltransferase Activity. *EMBO J* **23,** 4061–4071 (2004).

182. Pasini, D., Bracken, A. P., Hansen, J. B., Capillo, M. & Helin, K. The Polycomb Group Protein Suz12 Is Required for Embryonic Stem Cell Differentiation. *Mol Cell Biol* **27,** 3769–3779 (2007).

183. Ringrose, L., Ehret, H. & Paro, R. Distinct Contributions of Histone H3 Lysine 9 and 27 Methylation to Locus-Specific Stability of Polycomb Complexes. *Mol Cell* **16,** 641–653 (2004).

184. Azuara, V. *et al.* Chromatin Signatures of Pluripotent Cell Lines. *Nat Cell Biol* **8,** 532–538 (2006).

185. Blanco, E., González-Ramírez, M., Alcaine-Colet, A., Aranda, S. & Di Croce, L. The Bivalent Genome: Characterization, Structure, and Regulation. *Trends Genet* **36,** 118–131 (2020).

186. Voigt, P., Tee, W.-W. & Reinberg, D. A Double Take on Bivalent Promoters. *Genes Dev* **27,** 1318–1338 (2013).

187. Gifford, C. A. *et al.* Transcriptional and Epigenetic Dynamics during Specification of Human Embryonic Stem Cells. *Cell* **153,** 1149–1163 (2013).

188. Xie, W. *et al.* Epigenomic Analysis of Multilineage Differentiation of Human Embryonic Stem Cells. *Cell* **153,** 1134–1148 (2013).

189. Abraham, B. J., Cui, K., Tang, Q. & Zhao, K. Dynamic Regulation of Epigenomic Landscapes during Hematopoiesis. *BMC Genomics* **14,** 193 (2013).

190. Burney, M. J. *et al.* An Epigenetic Signature of Developmental Potential in Neural Stem Cells and Early Neurons. *Stem Cells* **31,** 1868–1880 (2013).

191. Cui, K. *et al.* Chromatin Signatures in Multipotent Human Hematopoietic Stem Cells Indicate the Fate of Bivalent Genes during Differentiation. *Cell Stem Cell* **4,** 80–93 (2009).

192. Dhar, S. S. *et al.* An Essential Role for UTX in Resolution and Activation of Bivalent Promoters. *Nucleic Acids Res* **44,** 3659–3674 (2016).

193. Kanayama, K. *et al.* Genome-Wide Mapping of Bivalent Histone Modifications in Hepatic Stem/Progenitor Cells. *Stem Cells Int.* **2019,** e9789240 (2019).

194. Tanimura, K., Suzuki, T., Vargas, D., Shibata, H. & Inagaki, T. Epigenetic Regulation of Beige Adipocyte Fate by Histone Methylation. *Endocr J* **66,** 115–125 (2019).

195. Zhou, Y. *et al.* Bivalent Histone Codes on WNT5A during Odontogenic Differentiation. *J Dent Res* **97,** 99–107 (2018).

196. Alder, O. *et al.* Ring1B and Suv39h1 Delineate Distinct Chromatin States at Bivalent Genes during Early Mouse Lineage Commitment. *Development* **137,** 2483–2492 (2010).

197. Dahl, J. A., Reiner, A. H., Klungland, A., Wakayama, T. & Collas, P. Histone H3 Lysine 27 Methylation Asymmetry on Developmentally-Regulated Promoters Distinguish the First Two Lineages in Mouse Preimplantation Embryos. *PLOS ONE* **5,** e9150 (2010).

198. Denholtz, M. *et al.* Long-Range Chromatin Contacts in Embryonic Stem Cells Reveal a Role for Pluripotency Factors and Polycomb Proteins in Genome Organization. *Cell Stem Cell* **13,** 602–616 (2013).

199. Kundu, S. *et al.* Polycomb Repressive Complex 1 Generates Discrete Compacted Domains That Change during Differentiation. *Mol Cell* **65,** 432–446.e5 (2017).

200. Mas, G. *et al.* Promoter Bivalency Favors an Open Chromatin Architecture in Embryonic Stem Cells. *Nat Genet* **50,** 1452–1462 (2018).

201. Vieux-Rochas, M., Fabre, P. J., Leleu, M., Duboule, D. & Noordermeer, D. Clustering of Mammalian Hox Genes with Other H3K27me3 Targets within an Active Nuclear Domain. *Proc Natl Acad Sci USA* **112,** 4672–4677 (2015).

202. Burr, M. L. *et al.* An Evolutionarily Conserved Function of Polycomb Silences the MHC Class I Antigen Presentation Pathway and Enables Immune Evasion in Cancer. *Cancer Cell* **36,** 385–401.e8 (2019).

203. Dunican, D. S. *et al.* Bivalent Promoter Hypermethylation in Cancer Is Linked to the H327me3/H3K4me3 Ratio in Embryonic Stem Cells. *BMC Biol* **18,** 25 (2020).

204. Kaukonen, D. *et al.* Analysis of H3K4me3 and H3K27me3 Bivalent Promotors in HER2+ Breast Cancer Cell Lines Reveals Variations Depending on Estrogen Receptor Status and Significantly Correlates with Gene Expression. *BMC Med Genomics* **13,** 92 (2020).

205. Messier, T. L. *et al.* Oncofetal Epigenetic Bivalency in Breast Cancer Cells: H3K4 and H3K27 Tri-Methylation as a Biomarker for Phenotypic Plasticity. *J Cell Physiol* **231,** 2474–2481 (2016).

206. Hu, D. *et al.* Not All H3K4 Methylations Are Created Equal: Mll2/COMPASS Dependency in Primordial Germ Cell Specification. *Mol Cell* **65,** 460–475.e6 (2017).

207. Hu, D. *et al.* The Mll2 Branch of the COMPASS Family Regulates Bivalent Promoters in Mouse Embryonic Stem Cells. *Nat Struct Mol Biol* **20,** 1093–1097 (2013).

208. Hattori, T. *et al.* Antigen Clasping by Two Antigen-Binding Sites of an Exceptionally Specific Antibody for Histone Methylation. *Proc Natl Acad Sci USA* **113,** 2092–2097 (2016).

209. Raran-Kurussi, S., Tözsér, J., Cherry, S., Tropea, J. E. & Waugh, D. S. Differential Temperature Dependence of Tobacco Etch Virus and Rhinovirus 3C Proteases. *Anal Biochem* **436,** 142–144 (2013).

210. Lechner, C. C., Agashe, N. D. & Fierz, B. Traceless Synthesis of Asymmetrically Modified Bivalent Nucleosomes. *Angew. Chem. Int. Ed.* **55,** 2903–2906 (2016).

211. Marks, H. *et al.* The Transcriptional and Epigenomic Foundations of Ground State Pluripotency. *Cell* **149,** 590–604 (2012).

212. Xiao, S. *et al.* Comparative Epigenomic Annotation of Regulatory DNA. *Cell* **149,** 1381–1392 (2012).

213. Kim, D.-H. *et al.* Histone H3K27 Trimethylation Inhibits H3 Binding and Function of SET1-Like H3K4 Methyltransferase Complexes. *Mol Cell Biol* **33,** 4936–4946 (2013).

214. Schmitges, F. W. *et al.* Histone Methylation by PRC2 Is Inhibited by Active Chromatin Marks. *Mol Cell* **42,** 330–341 (2011).

215. Shilatifard, A. The COMPASS Family of Histone H3K4 Methylases: Mechanisms of Regulation in Development and Disease Pathogenesis. *Annu Rev Biochem* **81,** 65–95 (2012).

216. Sze, C. C. *et al.* Histone H3K4 Methylation-Dependent and -Independent Functions of Set1A/COMPASS in Embryonic Stem Cell Self-Renewal and Differentiation. *Genes Dev* **31,** 1732–1737 (2017).

217. Denissov, S. *et al.* Mll2 Is Required for H3K4 Trimethylation on Bivalent Promoters in Embryonic Stem Cells, Whereas Mll1 Is Redundant. *Development* **141,** 526–537 (2014).

218. Aoyama, K. *et al.* Ezh1 Targets Bivalent Genes to Maintain Self-Renewing Stem Cells in Ezh2-Insufficient Myelodysplastic Syndrome. *iScience* **9,** 161–174 (2018).

219. Béguelin, W. *et al.* EZH2 and BCL6 Cooperate to Assemble CBX8-BCOR Complex to Repress Bivalent Promoters, Mediate Germinal Center Formation and Lymphomagenesis. *Cancer Cell* **30,** 197–213 (2016).

220. Terranova, C. *et al.* Global Developmental Gene Programing Involves a Nuclear Form of Fibroblast Growth Factor Receptor-1 (FGFR1). *PLOS ONE* **10,** e0123380 (2015).

221. Xing, Q. R. *et al.* Parallel Bimodal Single-Cell Sequencing of Transcriptome and Chromatin Accessibility. *Genome Res* **30,** 1027–1039 (2020).

222. Eckersley-Maslin, M. A. *et al.* Epigenetic Priming by Dppa2 and 4 in Pluripotency Facilitates Multi-Lineage Commitment. *Nat Struct Mol Biol* **27,** 696–705 (2020).

223. Tan, H. K. *et al.* DNMT3B Shapes the mCA Landscape and Regulates mCG for Promoter Bivalency in Human Embryonic Stem Cells. *Nucleic Acids Res* **47,** 7460–7475 (2019).

224. Hattori, T. *et al.* Recombinant Antibodies to Histone Post-Translational Modifications. *Nat Meth* **10,** 992–995 (2013).

225. Allis, C. D. & Jenuwein, T. The Molecular Hallmarks of Epigenetic Control. *Nat Rev Genet* **17,** 487–500 (2016).

226. Henikoff, S. Histone Modifications: Combinatorial Complexity or Cumulative Simplicity? *Proc Natl Acad Sci USA* **102,** 5308–5309 (2005).

227. Lu, C., Coradin, M., Janssen, K. A., Sidoli, S. & Garcia, B. A. Combinatorial Histone H3 Modifications Are Dynamically Altered in Distinct Cell Cycle Phases. *J Am Soc Mass Spectrom* **32,** 1300–1311 (2021).

228. Suganuma, T. & Workman, J. L. Signals and Combinatorial Functions of Histone Modifications. *Annu Rev Biochem* **80,** 473–499 (2011).

229. Radman-Livaja, M. & Rando, O. J. Nucleosome Positioning: How Is It Established, and Why Does It Matter? *Dev Biol* **339,** 258–266 (2010).

230. Rothbart, S. B. *et al.* Multivalent Histone Engagement by the Linked Tandem Tudor and PHD Domains of UHRF1 Is Required for the Epigenetic Inheritance of DNA Methylation. *Genes Dev* **27,** 1288–1298 (2013).

231. Savitsky, P. *et al.* Multivalent Histone and DNA Engagement by a PHD/BRD/PWWP Triple Reader Cassette Recruits ZMYND8 to K14ac-Rich Chromatin. *Cell Rep* **17,** 2724–2737 (2016).

232. Yun, M., Wu, J., Workman, J. L. & Li, B. Readers of Histone Modifications. *Cell Res* **21,** 564–578 (2011).

233. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9,** 9354 (2019).

234. Margueron, R. *et al.* Role of the Polycomb Protein EED in the Propagation of Repressive Histone Marks. *Nature* **461,** 762–767 (2009).

235. Yuan, W. *et al.* Dense Chromatin Activates Polycomb Repressive Complex 2 to Regulate H3 Lysine 27 Methylation. *Science* **337,** 971–975 (2012).

236. Abranches, E. *et al.* Neural Differentiation of Embryonic Stem Cells In Vitro: A Road Map to Neurogenesis in the Embryo. *PLOS ONE* **4,** e6286 (2009).

237. Conti, L. *et al.* Niche-Independent Symmetrical Self-Renewal of a Mammalian Tissue Stem Cell. *PLOS Biol* **3,** e283 (2005).

238. Beckett, D., Kovaleva, E. & Schatz, P. J. A Minimal Peptide Substrate in Biotin Holoenzyme Synthetase-Catalyzed Biotinylation. *Protein Sci* **8,** 921–929 (1999).

239. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-Optimal Probabilistic RNA-seq Quantification. *Nat Biotechnol* **34,** 525–527 (2016).

240. Pimentel, H., Bray, N. L., Puente, S., Melsted, P. & Pachter, L. Differential Analysis of RNA-seq Incorporating Quantification Uncertainty. *Nat Meth* **14,** 687–690 (2017).

241. Smith, R. M. & Rill, R. L. Mobile Historie Tails in Nucleosomes: Assignments of Mobile Segments and Investigations of Their Role in Chromatin Folding. *J Biol Chem* **264,** 10574–10581 (1989).

242. Walker, I. O. Differential Dissociation of Histone Tails from Core Chromatin. *Biochemistry* **23,** 5622–5628 (1984).

243. Davey, C. A., Sargent, D. F., Luger, K., Maeder, A. W. & Richmond, T. J. Solvent Mediated Interactions in the Structure of the Nucleosome Core Particle at 1.9Å Resolution. *J Mol Biol* **319,** 1097–1113 (2002).

244. Ghoneim, M., Fuchs, H. A. & Musselman, C. A. Histone Tail Conformations: A Fuzzy Affair with DNA. *Trends Biochem Sci* **46,** 564–578 (2021).

245. Kan, P.-Y., Lu, X., Hansen, J. C. & Hayes, J. J. The H3 Tail Domain Participates in Multiple Interactions during Folding and Self-Association of Nucleosome Arrays. *Mol Cell Biol* **27,** 2084–2091 (2007).

246. Kan, P.-Y., Caterino, T. L. & Hayes, J. J. The H4 Tail Domain Participates in Intra- and Internucleosome Interactions with Protein and DNA during Folding and Oligomerization of Nucleosome Arrays. *Mol Cell Biol* (2009).

247. Pepenella, S., Murphy, K. J. & Hayes, J. J. Intra- and Inter-nucleosome Interactions of the Core Histone Tail Domains in Higher-Order Chromatin Structure. *Chromosoma* **123,** 3–13 (2014).

248. Dorigo, B., Schalch, T., Bystricky, K. & Richmond, T. J. Chromatin Fiber Folding: Requirement for the Histone H4 N-terminal Tail. *J Mol Biol* **327,** 85–96 (2003).

249. Dorigo, B. *et al.* Nucleosome Arrays Reveal the Two-Start Organization of the Chromatin Fiber. *Science* **306,** 1571–1573 (2004).

250. Sinha, D. & Shogren-Knaak, M. A. Role of Direct Interactions between the Histone H4 Tail and the H2A Core in Long Range Nucleosome Contacts. *J Biol Chem* **285,** 16572–16581 (2010).

251. Farooq, Z., Banday, S., Pandita, T. K. & Altaf, M. The Many Faces of Histone H3K79 Methylation. *Mutat Res Rev Mutat Res* **768,** 46–52 (2016).

252. Nguyen, A. T. & Zhang, Y. The Diverse Functions of Dot1 and H3K79 Methylation. *Genes Dev* **25,** 1345–1358 (2011).

253. Satoshi, O., Yoshiko, K., Robert, P., Mark, M. & Tsutomu, A. Structural Characteristics of Short Peptides in Solution. *Protein Pept Lett* **20,** 1308–1323 (2013).

254. Behrendt, R., White, P. & Offer, J. Advances in Fmoc Solid-phase Peptide Synthesis. *J Pept Sci* **22,** 4–27 (2016).

255. Chen, Z., Grzybowski, A. T. & Ruthenburg, A. J. Traceless Semisynthesis of a Set of Histone 3 Species Bearing Specific Lysine Methylation Marks. *Chembiochem* **15,** 2071–2075 (2014).

256. Hattori, T. & Koide, S. Next-Generation Antibodies for Post-Translational Modifications. *Curr Opin Struct Biol* **51,** 141–148 (2018).

257. Hebbes, T. R., Turner, C. H., Thorne, A. W. & Crane-Robinson, C. A "Minimal Epitope" Anti-Protein Antibody That Recognises a Single Modified Amino Acid. *Mol Immunol* **26,** 865–873 (1989).

258. Posnett, D. N., McGrath, H. & Tam, J. P. A Novel Method for Producing Anti-Peptide Antibodies. Production of Site-Specific Antibodies to the T Cell Antigen Receptor Beta-Chain. *J Biol Chem* **263,** 1719–1725 (1988).

259. Sutcliffe, J. G. *et al.* Chemical Synthesis of a Polypeptide Predicted from Nucleotide Sequence Allows Detection of a New Retroviral Gene Product. *Nature* **287,** 801–805 (1980).

260. Feng, Q. *et al.* Methylation of H3-Lysine 79 Is Mediated by a New Family of HMTases without a SET Domain. *Curr Biol* **12,** 1052–1058 (2002).

261. Lacoste, N., Utley, R. T., Hunter, J. M., Poirier, G. G. & Côte, J. Disruptor of Telomeric Silencing-1 Is a Chromatin-Specific Histone H3 Methyltransferase. *J Biol Chem* **277,** 30421–30424 (2002).

262. Van Leeuwen, F., Gafken, P. R. & Gottschling, D. E. Dot1p Modulates Silencing in Yeast by Methylation of the Nucleosome Core. *Cell* **109,** 745–756 (2002).

263. Ng, H. H. *et al.* Lysine Methylation within the Globular Domain of Histone H3 by Dot1 Is Important for Telomeric Silencing and Sir Protein Association. *Genes Dev* **16,** 1518–1527 (2002).

264. Shanower, G. A. *et al.* Characterization of the Grappa Gene, the Drosophila Histone H3 Lysine 79 Methyltransferase. *Genetics* **169,** 173–184 (2005).

265. Jones, B. *et al.* The Histone H3K79 Methyltransferase Dot1L Is Essential for Mammalian Development and Heterochromatin Structure. *PLOS Genet* **4,** e1000190 (2008).

266. Kouskouti, A. & Talianidis, I. Histone Modifications Defining Active Genes Persist after Transcriptional and Mitotic Inactivation. *EMBO J* **24,** 347–357 (2005).

267. Feng, Y. *et al.* Early Mammalian Erythropoiesis Requires the Dot1L Methyltransferase. *Blood* **116,** 4483–4491 (2010).

268. Marcos-Villar, L. & Nieto, A. The DOT1L Inhibitor Pinometostat Decreases the Host-Response against Infections: Considerations about Its Use in Human Therapy. *Sci Rep* **9,** 16862 (2019).

269. Bernt, K. M. *et al.* MLL-Rearranged Leukemia Is Dependent on Aberrant H3K79 Methylation by DOT1L. *Cancer Cell* **20,** 66–78 (2011).

270. Hess, J. L. MLL: A Histone Methyltransferase Disrupted in Leukemia. *Trends Mol Med* **10,** 500–507 (2004).

271. Milne, T. A. *et al.* MLL Targets SET Domain Methyltransferase Activity to Hox Gene Promoters. *Mol Cell* **10,** 1107–1117 (2002).

272. Ayton, P. M. & Cleary, M. L. Molecular Mechanisms of Leukemogenesis Mediated by MLL Fusion Proteins. *Oncogene* **20** (2001).

273. Krivtsov, A. V. & Armstrong, S. A. MLL Translocations, Histone Modifications and Leukaemia Stem-Cell Development. *Nat Rev Cancer* **7,** 823–833 (2007).

274. Armstrong, S. A. *et al.* MLL Translocations Specify a Distinct Gene Expression Profile That Distinguishes a Unique Leukemia. *Nat Genet* **30,** 41–47 (2002).

275. Bitoun, E., Oliver, P. L. & Davies, K. E. The Mixed-Lineage Leukemia Fusion Partner AF4 Stimulates RNA Polymerase II Transcriptional Elongation and Mediates Coordinated Chromatin Remodeling. *Hum Mol Genet* **16,** 92–106 (2007).

276. Mueller, D. *et al.* A Role for the MLL Fusion Partner ENL in Transcriptional Elongation and Chromatin Modification. *Blood* **110,** 4445–4454 (2007).

277. Mueller, D. *et al.* Misguided Transcriptional Elongation Causes Mixed Lineage Leukemia. *PLOS Biol* **7,** e1000249 (2009).

278. Okada, Y. *et al.* hDOT1L Links Histone Methylation to Leukemogenesis. *Cell* **121,** 167–178 (2005).

279. Zhang, W., Xia, X., Reisenauer, M. R., Hemenway, C. S. & Kone, B. C. Dot1a-AF9 Complex Mediates Histone H3 Lys-79 Hypermethylation and Repression of ENaCalpha in an Aldosterone-Sensitive Manner. *J Biol Chem* **281,** 18059–18068 (2006).

280. Daigle, S. R. *et al.* Potent Inhibition of DOT1L as Treatment of MLL-fusion Leukemia. *Blood* **122,** 1017–1025 (2013).

281. Klaus, C. R. *et al.* DOT1L Inhibitor EPZ-5676 Displays Synergistic Antiproliferative Activity in Combination with Standard of Care Drugs and Hypomethylating Agents in MLL-Rearranged Leukemia Cells. *J Pharmacol Exp Ther* **350,** 646–656 (2014).

282. Waters, N. J. *et al.* Exploring Drug Delivery for the DOT1L Inhibitor Pinometostat (EPZ-5676): Subcutaneous Administration as an Alternative to Continuous IV Infusion, in the Pursuit of an Epigenetic Target. *J Control Release* **220,** 758–765 (2015).

283. Stein, E. M. *et al.* The DOT1L Inhibitor Pinometostat Reduces H3K79 Methylation and Has Modest Clinical Activity in Adult Acute Leukemia. *Blood* **131,** 2661–2669 (2018).

284. Freitag, M. Histone Methylation by SET Domain Proteins in Fungi. *Annu Rev Microbiol* **71,** 413–439 (2017).

285. Niu, K., Liu, R. & Liu, N. Quantitative ChIP-seq by Adding Spike-in from Another Species. *Bio Protoc* **8,** e2981 (2018).

286. Thåström, A., Lowary, P. T. & Widom, J. Measurement of Histone–DNA Interaction Free Energy in Nucleosomes. *Methods* **33,** 33–44 (2004).

287. Van der Heijden, T., van Vugt, J. J., Logie, C. & van Noort, J. Sequence-Based Prediction of Single Nucleosome Positioning and Genome-Wide Nucleosome Occupancy. *Proc Natl Acad Sci USA* **109,** E2514–E2522 (2012).

288. Fragoso, G., John, S., Roberts, M. S. & Hager, G. L. Nucleosome Positioning on the MMTV LTR Results from the Frequency-Biased Occupancy of Multiple Frames. *Genes Dev* **9,** 1933–1947 (1995).

289. Widlund, H. R. *et al.* Identification and Characterization of Genomic Nucleosome-Positioning Sequences. *J Mol Biol* **267,** 807–817 (1997).

290. Brzezinka, K. *et al.* Functional Diversity of Inhibitors Tackling the Differentiation Blockage of MLL-rearranged Leukemia. *J Hematol Oncol* **12,** 66 (2019).

291. Richter, W. F., Shah, R. N. & Ruthenburg, A. J. Non-Canonical H3K79me2-dependent Pathways Promote the Survival of MLL-rearranged Leukemia. *eLife* **10,** e64960 (2021).

292. Chen, S. *et al.* The PZP Domain of AF10 Senses Unmodified H3K27 to Regulate DOT1L-Mediated Methylation of H3K79. *Mol Cell* **60,** 319–327 (2015).

293. Kang, J.-Y. *et al.* KDM2B Is a Histone H3K79 Demethylase and Induces Transcriptional Repression via Sirtuin-1-Mediated Chromatin Silencing. *EMBO J* **32,** 5737–5750 (2018).

294. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of Age: Ten Years of next-Generation Sequencing Technologies. *Nat Rev Genet* **17,** 333–351 (2016).

295. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq. *Nat Meth* **5,** 621–628 (2008).

296. Sartorelli, V. & Lauberth, S. M. Enhancer RNAs Are an Important Regulatory Layer of the Epigenome. *Nat Struct Mol Biol* **27,** 521–528 (2020).

297. Guenther, M. G. *et al.* Aberrant Chromatin at Genes Encoding Stem Cell Regulators in Human Mixed-Lineage Leukemia. *Genes Dev* **22,** 3403–3408 (2008).

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

299. Valouev, A. *et al.* Genome-Wide Analysis of Transcription Factor Binding Sites Based on ChIP-Seq Data. *Nat Meth* **5,** 829–834 (2008).

300. Lieberman-Aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326,** 289–293 (2009).

301. Yu, M. *et al.* Base-Resolution Analysis of 5-Hydroxymethylcytosine in the Mammalian Genome. *Cell* **149,** 1368–1380 (2012).

302. Lander, E. S. *et al.* Initial Sequencing and Analysis of the Human Genome. *Nature* **409,** 860–921 (2001).

303. Wheeler, T. J. *et al.* Dfam: A Database of Repetitive DNA Based on Profile Hidden Markov Models. *Nucleic Acids Res* **41,** D70–D82 (2013).

304. Derrien, T. *et al.* Fast Computation and Applications of Genome Mappability. *PLOS ONE* **7,** e30377 (2012).

305. Karimzadeh, M., Ernst, C., Kundaje, A. & Hoffman, M. M. Umap and Bismap: Quantifying Genome and Methylome Mappability. *Nucleic Acids Res* **46,** e120–e120 (2018).

306. Slotkin, R. K. The Case for Not Masking Away Repetitive DNA. *Mobile DNA* **9,** 15 (2018).

307. Consiglio, A. *et al.* A Fuzzy Method for RNA-Seq Differential Expression Analysis in Presence of Multireads. *BMC Bioinform* **17,** 345 (2016).

308. Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A. & Dewey, C. N. RNA-Seq Gene Expression Estimation with Read Mapping Uncertainty. *Bioinformatics* **26,** 493–500 (2010).

309. Lanciano, S. & Cristofari, G. Measuring and Interpreting Transposable Element Expression. *Nat Rev Genet* **21,** 721–736 (2020).

310. Chung, D. *et al.* Discovering Transcription Factor Binding Sites in Highly Repetitive Regions of Genomes with Multi-Read Analysis of ChIP-Seq Data. *PLOS Comput Biol* **7,** e1002111 (2011).

311. Ji, Y. *et al.* BM-Map: Bayesian Mapping of Multireads for Next-Generation Sequencing Data. *Biometrics* **67,** 1215–1224 (2011).

312. Liu, Y. *et al.* An Enrichment Method for Mapping Ambiguous Reads to the Reference Genome for NGS Analysis. *J Bioinform Comput Biol* (2020).

313. Newkirk, D., Biesinger, J., Chon, A., Yokomori, K. & Xie, X. AREM: Aligning Short Reads from ChIP-Sequencing by Expectation Maximization. *J Comput Biol* **18,** 1495–1505 (2011).

314. Zeng, X. *et al.* Perm-Seq: Mapping Protein-DNA Interactions in Segmental Duplication and Highly Repetitive Regions of Genomes with Prior-Enhanced Read Mapping. *PLOS Comput Biol* **11,** e1004491 (2015).

315. Kondo, Y. & Issa, J.-P. J. Enrichment for Histone H3 Lysine 9 Methylation at Alu Repeats in Human Cells. *J Biol Chem* **278,** 27658–27662 (2003).

316. Bulut-Karslioglu, A. *et al.* Suv39h-Dependent H3K9me3 Marks Intact Retrotransposons and Silences LINE Elements in Mouse Embryonic Stem Cells. *Mol Cell* **55,** 277–290 (2014).

317. Pezic, D., Manakov, S. A., Sachidanandam, R. & Aravin, A. A. piRNA Pathway Targets Active LINE1 Elements to Establish the Repressive H3K9me3 Mark in Germ Cells. *Genes Dev* **28,** 1410–1428 (2014).

318. Pauler, F. M. *et al.* H3K27me3 Forms BLOCs over Silent Genes and Intergenic Regions and Specifies a Histone Banding Pattern on a Mouse Autosomal Chromosome. *Genome Res* **19,** 221–233 (2009).

319. Fadloun, A. *et al.* Chromatin Signatures and Retrotransposon Profiling in Mouse Embryos Reveal Regulation of LINE-1 by RNA. *Nat Struct Mol Biol* **20,** 332–338 (2013).

320. He, J. *et al.* Transposable Elements Are Regulated by Context-Specific Patterns of Chromatin Marks in Mouse Embryonic Stem Cells. *Nat Commun* **10,** 34 (2019).

321. Jjingo, D. *et al.* Mammalian-Wide Interspersed Repeat (MIR)-Derived Enhancers and the Regulation of Human Gene Expression. *Mobile DNA* **5,** 14 (2014).

322. Mravinac, B. *et al.* Histone Modifications within the Human X Centromere Region. *PLOS ONE* **4,** e6602 (2009).

323. Ward, M. C. *et al.* Latent Regulatory Potential of Human-Specific Repetitive Elements. *Mol Cell* **49,** 262–272 (2013).

324. Zhang, B. *et al.* Allelic Reprogramming of the Histone Modification H3K4me3 in Early Mammalian Development. *Nature* **537,** 553–557 (2016).

325. Ernst, J. *et al.* Mapping and Analysis of Chromatin State Dynamics in Nine Human Cell Types. *Nature* **473,** 43–49 (2011).

326. Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From Reads to Insight: A Hitchhiker's Guide to ATAC-seq Data Analysis. *Genome Biol* **21,** 22 (2020).

327. Li, Y.-C., Korol, A. B., Fahima, T. & Nevo, E. Microsatellites Within Genes: Structure, Function, and Evolution. *Mol Biol Evol* **21,** 991–1007 (2004).

328. Johnson, N. R., Yeoh, J. M., Coruh, C. & Axtell, M. J. Improved Placement of Multi-mapping Small RNAs. *G3: Genes Genomes Genet* **6,** 2103–2111 (2016).

329. Muir, P. *et al.* The Real Cost of Sequencing: Scaling Computation to Keep Pace with Data Generation. *Genome Biol* **17,** 53 (2016).

330. Kato, A., Lamb, J. C. & Birchler, J. A. Chromosome Painting Using Repetitive DNA Sequences as Probes for Somatic Chromosome Identification in Maize. *Proc Natl Acad Sci USA* **101,** 13554–13559 (2004).

331. Moore, J. E. *et al.* Expanded Encyclopaedias of DNA Elements in the Human and Mouse Genomes. *Nature* **583,** 699–710 (2020).

332. Stephens, Z. D. *et al.* Simulating Next-Generation Sequencing Datasets from Empirical Mutation and Sequencing Models. *PLOS ONE* **11,** e0167047 (2016).

333. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-Based Genome Alignment and Genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37,** 907–915 (2019).

334. Robinson, J. T. *et al.* Integrative Genomics Viewer. *Nat Biotechnol* **29,** 24–26 (2011).

335. Shah, R. N. & Ruthenburg, A. J. Sequence Deeper without Sequencing More: Bayesian Resolution of Ambiguously Mapped Reads. *PLOS Comput Biol* **17,** e1008926 (2021).

INDEX