

# Distinct DNA Sequence Preference for Histone Occupancy in Primary and Transformed Cells

Subhamoy Datta\*<sup>ORCID</sup>, Manthan Patel\*, Divyesh Patel and Umashankar Singh<sup>ORCID</sup>

HoMeCell Lab, Discipline of Biological Engineering, Indian Institute of Technology Gandhinagar, Gandhinagar, India.

Cancer Informatics  
Volume 18: 1–15  
© The Author(s) 2019  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/1176935119843835



**ABSTRACT:** Genome-wide occupancy of several histone modifications in various cell types has been studied using chromatin immunoprecipitation (ChIP) sequencing. Histone occupancy depends on DNA sequence features like inter-strand symmetry of base composition and periodic occurrence of TT/AT. However, whether DNA sequence motifs act as an additional effector of histone occupancy is not known. We have analyzed the presence of DNA sequence motifs in publicly available ChIP-sequence datasets for different histone modifications. Our results show that DNA sequence motifs are associated with histone occupancy, some of which are different between primary and transformed cells. The motifs for primary and transformed cells showed different levels of GC-richness and proximity to transcription start sites (TSSs). The TSSs associated with transformed or primary cell-specific motifs showed different levels of TSS flank transcription in primary and transformed cells. Interestingly, TSSs with a motif-linked occupancy of H2AFZ, a component of positioned nucleosomes, showed a distinct pattern of RNA Polymerase II (POLR2A) occupancy and TSS flank transcription in primary and transformed cells. These results indicate that DNA sequence features dictate differential histone occupancy in primary and transformed cells, and the DNA sequence motifs affect transcription through regulation of histone occupancy.

**KEYWORDS:** ChIP, GC-richness, TSSs, POLR2A

**RECEIVED:** March 16, 2019. **ACCEPTED:** March 24, 2019.

**TYPE:** Original Research

**FUNDING:** The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was funded by grants to US from Science and Engineering Research Board (DST; EMR/2015/001080 and ICPS T-357 Government of India), Department of Biotechnology (BT/PR15883/BRB/10/1480/2016), GSBTM (FAP SSA/4873; Government of Gujarat), and IIT Gandhinagar (RIG/0204). SD was supported by GATE fellowship (MHRD, Government of

India). MP and DP were supported by CSIR-UGC fellowship (Government of India). The authors thankfully acknowledge all databases and tools in the public domain used in this study and fellow researchers at IIT Gandhinagar.

**DECLARATION OF CONFLICTING INTERESTS:** The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**CORRESPONDING AUTHOR:** Umashankar Singh, HoMeCell Lab, Discipline of Biological Engineering, Indian Institute of Technology Gandhinagar, Gandhinagar 382355, Gujarat, India. Email: usingh@iitgn.ac.in

## Introduction

Regulation of nucleosomal localization is a complex process.<sup>1,2</sup> To facilitate transcription, replication, and DNA repair, nucleosomes slide short distances and affect local nucleosomal reorganization.<sup>3</sup> The stabilization of sliding nucleosomes has been proposed to depend on the DNA sequence and may be affected by the histone modification profile of the nucleosomes.<sup>4,5</sup> The density of nucleosomes in the vicinity may also determine the choices of DNA sequences a sliding nucleosome settles with.<sup>6,7</sup> On the other hand, there are regions which have relatively inflexible nucleosomal occupancy due to sequence features. Sequences with high dA:dT density are nucleosome-depleted.<sup>8</sup> Positioned nucleosomes flanking the transcription start sites (TSSs) and constitutive heterochromatin serve as strong examples of sequence dependence of nucleosomal occupancy.<sup>4</sup> A periodicity of TT and AT at every 10th base pair facilitates the wrapping of DNA around the histone octamer.<sup>9</sup> Stiff sequences that have GC and AT inter-strand asymmetry provide less favorable sites for nucleosomes.<sup>8</sup> For maintaining the long constitutively nucleosome-rich regions distinct from nucleosome-depleted regions, the DNA sequence plays a key role through binding of boundary element proteins, typified by CCCTC-binding factor (CTCF).<sup>10,11</sup> Nucleosome locations are neither fixed (like binding of sequence-specific transcription factor [TF] complexes) nor entirely random. Many TFs bind to specific cognate sequences which may be nucleosome-free or occupied by nucleosomes with

activating or repressive histone modifications.<sup>12,13</sup> However, the TF-chromatin interaction is also determined by the nucleosome-TFs interaction.<sup>14,15</sup> Thus, TF-DNA interaction is potentially regulated by nucleosomal occupancy in many different ways: dense nucleosomal occupancy by nucleosomes with repressive histone modifications may render the DNA inaccessible, or the presence of activating histone modifications or simply an absence of nucleosomal occupancy at specific regions may promote TF-DNA interaction. Nucleosomes with activating histone modifications interact with and recruit positive regulators of transcription, whereas the histones with repressive marks serve as negative regulators of gene expression by either associating with and recruiting repressors of transcription or by hindering the access of DNA to positive regulators of transcription.<sup>16</sup> In addition, a balanced state of positive and negative regulators of transcription, all depending on bivalent modifications of histones as nucleosomal components, may maintain genes in a poised state of transcription.<sup>16</sup> However, high nucleosomal occupancy at TSSs rich in binding sites of transcription activators occurs in differentiation, suggesting that TF-binding sites are not unanimous determinants of gene expression.<sup>17,18</sup> The cis-regulation of gene expression depends on TSS-flanking regulatory sequences. These TSS-flanking sequences determine transcription patterns of genes in a developmental stage-specific manner by being playgrounds of the epigenetic mechanisms that drive differentiation.<sup>17,18</sup> A key set of these epigenetic mechanisms includes

\*These authors have contributed equally.



regulated changes in nucleosomal occupancy, histone modifications of nucleosomes, and the accessibility of transcription machinery to DNA.<sup>19</sup> Over two-thirds of all known genes have TSSs associated with GC-rich upstream regulatory regions which are nucleosome-poor and contain binding sites for TFs.<sup>20</sup> Histones, unlike TFs, are bound to DNA genome-wide without being restricted to the GC-rich TSS-flanking cis-regulatory elements, promoters, and TF-binding sites. Identification of histone-bound sequences genome-wide, typically by chromatin immunoprecipitation sequencing (ChIP-seq) on variously modified histones, cannot detect individual nucleosome-associated DNA sequences due to systemic features of the technique and the method. Histone-binding patterns derived from single molecules get pooled for alleles and replicated chromatids in single cell analyses. In the commonly performed analyses on pools of millions of cells, deep sequencing generates sequence reads which overlap and cluster in regions characterized as histone-bound peaks. A significant part of histone ChIP-seq reads, however, do not cluster to generate peaks but nevertheless demonstrate 2 facts: that there is significant intercellular heterogeneity in nucleosomal occupancy and that nucleosomal occupancy is not random but clustered in discrete sequences identified as peaks.<sup>21,22</sup> The intercellular heterogeneity of epigenetic mechanisms thus seems to be patterned around certain signals that ensure that the specific types of histone modifications cluster at specific functional sites, such as TSSs with possible effects on transcription.

Transcription start sites and promoters with TF-binding sequences also display distinct nucleosomal patterns.<sup>23</sup> Various lines of research on nucleosomal positioning and histone occupancy have supported the notion that histone-DNA interactions are not governed by histone-specific DNA sequence motifs. While TF-binding site-rich sequences such as promoters are nucleosome-poor, the repeat-rich constitutive heterochromatin is often nucleosome-dense.<sup>19,24</sup> Although nucleosomal positioning by DNA sequence motifs is safely ruled out except for a positive correlation with W/S motifs with repeats,<sup>25</sup> it is not known if DNA sequence motifs play a role in determining nucleosomal density.

Epigenetic mechanisms are important for maintenance of cell-type identity and are altered in cellular transformation.<sup>25–28</sup> Also, disturbances in positioned nucleosomes could be caused by epigenetic changes such as cytosine methylation, activities of chromatin remodeling factors which can affect histone occupancy, and thus transcription and TSS usage.<sup>28</sup> Altered gene expression due to spatial disturbances in histone code and nucleosomal densities can contribute to tumorigenesis. Transcription start sites are sequences often associated with GC-richness, high CpG density, and association with CpG islands, as well as clusters of TF-binding motifs. Regulation of peri-TSS nucleosomal density by DNA sequence features and its consequence on gene expression remains unknown. We began with an assumption that histone-DNA interaction is not dependent on TF-binding site-like DNA sequence motifs. To test the validity of our

assumption, we have performed a systematic analysis of DNA sequence motifs hidden in ENCODE DNase-seq and histone ChIP-seq datasets. Our approach relies on identification of common ChIP-seq Peak Regions (CPR) across multiple transformed cell lines or primary cell lines, and helps us decipher any subtle primary cell-specific and transformed cell-specific signature sequence motifs that underlie altered nucleosomal positioning in cellular transformation. Our analyses reveal that CPR in primary cells (p-CPR) contain contrasting sequence features compared with the CPR in transformed cells (t-CPR). A functional consequence of peri-TSS altered nucleosomal density seems to be aberrant in RNA Polymerase II activity in transformed cells immediately downstream of TSSs.

## Methods

### *Cell lines and histone modification ChIP-seq datasets*

Eleven histone modifications for which quality-controlled ChIP-seq datasets were available for both primary and transformed cells were selected from ENCODE (Supplemental Table S1). Processed narrow ChIP-seq peak data for each cell line were downloaded either in bigBed (for hg38) or bed (for hg19) formats. No prior filtering of the data was done before using them as inputs for CPR identification.

### *Cell lines and datasets of POLR2A ChIP-seq and RNA-seq*

POLR2A ChIP-seq and RNA-seq data were not available for the same sets of cell lines as used for CPR identification based on histone modification ChIP-seq data. Disparate sets of quality-controlled RNA-seq and POLR2A ChIP-seq data from various primary and transformed cell lines were obtained from ENCODE (Supplemental Tables S6 and S7). All the RNA-seq data were from paired-end sequencing and available for both the strands as pairs of bigWig files for each cell line. POLR2A ChIP-seq peak data were available as single bigWig file per cell line.

### *Other datasets and publicly available resources*

Transcription start site dataset (hg19 coordinates of Capped Analysis of Gene Expression [CAGE] peaks-verified regions) was downloaded from FANTOM database. Genome-wide GC content at 5 bp interval (gc5base.bw) was downloaded from University of California Santa Cruz (UCSC) Human Genome (hg38) Annotation database. The genomic coordinates, wherever required, were converted from hg19 to hg38 human genome assembly either through LiftOver tool in UCSC Genome database online or using a locally installed version of the same tool. Genome file for chromosome sizes and hg38 sequence (repeat-masked or unmasked) were downloaded from UCSC Genome database.

### *CPR identification and analyses of p-CPR and t-CPR*

To find CPR between different numbers of cells lines within primary and transformed cells separately, histone modification-specific multiple intersects were performed using multiintersectBed function in Bedtools. The lengths of obtained CPR across different number of combinations of cells were filtered by retaining 100bp or more.

To obtain distance distribution between CPR with p-CPR and t-CPR motifs, the closest distance between CPR from two sets was calculated using closestBed function in Bedtools. This distribution was compared against 10 iterations of distance distribution of CPR, pooled and randomly assigned to either p-CPR or t-CPR groups blindly. Significance of difference between the two distribution patterns was calculated by Kruskal-Wallis test in GraphPad Prism 7.

### *Motif discovery*

FASTA sequences for the CPR were fetched from the hg38 build of Human genome through fastaFromBed function in Bedtools. Motif discovery was performed by subjecting these sequences of CPR to non-discriminative and discriminative motif search using Discriminative Regular Expression Motif Elicitation (DREME) program from Multiple Em for Motif Elicitation (MEME) suite<sup>29,30</sup> command-line versions. The discriminative motif search was performed by using the command-line version of MEME suite application DREME. Discriminative Regular Expression Motif Elicitation performs a Fisher's exact test on the difference of a motif occurrence between given sets of positive and negative sequences. By default, DREME applies an  $E$  value cutoff of 0.05 ( $P$  value  $\times$  number of motifs analyzed) and returns maximum 100 differentially enriched motifs.

For stringency, we considered only top 10 best discriminative motifs, and all the discriminative motifs reported here correspond to  $E$  value  $< 2.6e-2$ .

For the discriminative motif search, the primary and transformed CPR FASTA sequences were used as negative files against transformed and primary CPR sequences, respectively.

### *Consensus Motif discovery*

Consensus motif identification for each histone modification from the Position Weight Matrices (PWMs) of motifs obtained by DREME was carried out using STAMP tool (a web tool for exploring DNA-binding motif similarities).<sup>29-31</sup> The PWMs obtained as STAMP output (PWMs of consensus motifs of all histone modifications) were further subjected to cell-type-specific consensus motif identification.

### *Motif prediction at CPR*

For individual histone modification, the p-CPR and t-CPR motif prediction in CPR were carried out by performing Find

Individual Motif Occurrences (FIMO) option in locally installed MEME suite by providing PWMs obtained from DREME as input ( $P \leq .0001$ ).<sup>30</sup>

### *Heat-maps and summary plots*

Conversion of bed to bigWig was done using UCSC scripts bedItemOverlapCount and bedGraphToBigWig. The bigWigs were generated for CpG density genome-wide, CPR motifs and RNA-seq reads combined for all histone modifications.

The matrices for plotting heat-maps and summary plots were generated by using computeMatrix option in Deeptools.<sup>32</sup> The plots were generated by utilizing the plotHeatmap function for heat-map and the plotProfile for summary plots in Deeptools. RNA-seq read density and POLR2A occupancy plots were generated by aggregate option in bwtool to calculate signal at base level and visualized in GraphPad Prism 7.

### *Identification of CPR-TSSs and calculation of p-CPR or t-CPR motif distances from the TSSs*

Using FIMO (MEME Suite), we first identified and filtered for further analysis a subset of CPR, which contained p-CPR motifs or t-CPR motifs ( $P \leq 0.0001$ ) in primary and transformed CPR sets, respectively. Second, by using the Bedtools Closest function, we further filtered out those CPR, the midpoints of which were a maximum of 1 kb from the nearest CpG island (CpG Island annotation bed coordinates, UCSC Genome Database). Third, using formula functions in OpenOffice Spreadsheet, we extracted the longest possible coordinates using the following 4 locations: CPR-start, CPR-end, CpG island-start, and CpG island-end. These were called chimeric CPR-CpG island coordinates. All chimeric CPR-CpG island coordinates were converted to hg19 coordinates. Finally, by Bedtools intersectBed function, we extracted all the TSSs midpoints (hg19 coordinates) that were located within the chimeric coordinates. All these steps were performed in primary and transformed CPR sets separately. The TSSs falling within the chimeric coordinates of primary and transformed groups commonly were filtered out and regarded as CPR-TSSs (hg19 coordinates). The coordinates of the midpoints of CPR-TSSs and midpoints of p-CPR or t-CPR motifs (hg19 coordinates) were used to calculate the distance of these motifs from the CPR-TSSs. For all further analyses, the hg19 coordinates of CPR-TSSs were lifted back to hg38.

All genome coordinate conversions were done using UCSC Genome browser LiftOver tool. After identifying the CPR-TSSs for each histone modification, the respective CPR-TSSs (.bed) files were concatenated and the entries were made unique to form a collective set of CPR-TSSs for all 8 histone modifications. Using Bedtools intersect function, the set of TSSs were identified (from the TSS dataset) excluding the CPR-TSSs and were termed as non-CPR-TSSs.



### *Repeat content analysis*

Repeat contents in CPR-TSSs and non-CPR-TSSs upstream flanks (Supplemental Tables S4 and S5) were measured using RepeatMasker<sup>33</sup> by running with default options. All the CPR-TSSs undergoing a p-CPR to t-CPR motif shift for the different histone modifications were collated and bed coordinates were generated for the TSSs and up to 0.5 kb upstream locations. These coordinates were made unique and merged (using Bedtools mergeBed function) before extracting sequence (hg38) and subjecting to analysis by RepeatMasker.

### *Statistical analysis*

All statistical analyses were performed in GraphPad Prism 7 on the data generated from analysis using different tools and OpenOffice Spreadsheet as mentioned.

### *Generation of shuffled genomic coordinate and mock sets of CPR*

Random selection of permuted genomic regions of similar length and chromosomal profiles of the CPR were generated against each of the histone modifications by using Bedtools shuffleBed. To obtain mock set of CPR (t-CPR and p-CPR), 10 different rounds of randomization were carried out on regions pooled from CPR-threshold of primary and transformed groups for each histone modification using the subcommand random in Bedtools. The pooled sequences were randomly segregated proportionately into mock groups.

## **Results**

### *p-CPR and t-CPR share common DNA sequence motifs*

DNA binding of histones is not known to occur at defined DNA sequence motifs. However, if DNA-histone interaction in vivo (in the form of nucleosomal occupancy) depends on genomic DNA sequence, then histone occupancy would tend to recur consistently at such sequences across different cell types. Else, if histone occupancy solely depends on tissue type and differentiation, then histone occupancy would vary between tissue types and in cellular transformation. To test these possibilities in an unbiased manner, we began by collectively analyzing histone ChIP-seq data (for a panel of transformed and primary cell lines for 11 histone modifications; chosen from ENCODE database) and finding out CPR (Supplemental Table S1).

Using the multiple intersection feature in Bedtools,<sup>34</sup> we derived CPR (minimum 100 bp long) for each histone modification in panels of primary or transformed cells separately. For each histone modification, the number of CPR decreased exponentially with an initial increase in the number of cell lines across which commonality was calculated. For

each histone modification, we subjectively chose a threshold number of cell lines (independently for the panels of transformed and primary cells) after which the increase in the number of cell lines did not have a major effect on the number of CPR (Figure 1A). The CPR for primary and transformed cells at these thresholds were called p-CPR and t-CPR, respectively. In contrast, at most of these thresholds, no common regions could be identified in shuffled genomic coordinates for any of the histone modifications (Figure 1A). This established that CPR are specifically present only in regions occupied by histones.

The ChIP-seq data used in this analysis were derived from variously modified histones, some of which co-occupy the DNA as components of nucleosomes. Expectedly, we observed overlaps between CPR for the different histone modifications (Figure 1B). A Bland-Altman plot depicting the number of CPR shared between different histone modifications clearly showed that CPR for some pairs of histone modifications occur very differently between primary and transformed cells (Figure 1B and Supplemental Table S2).

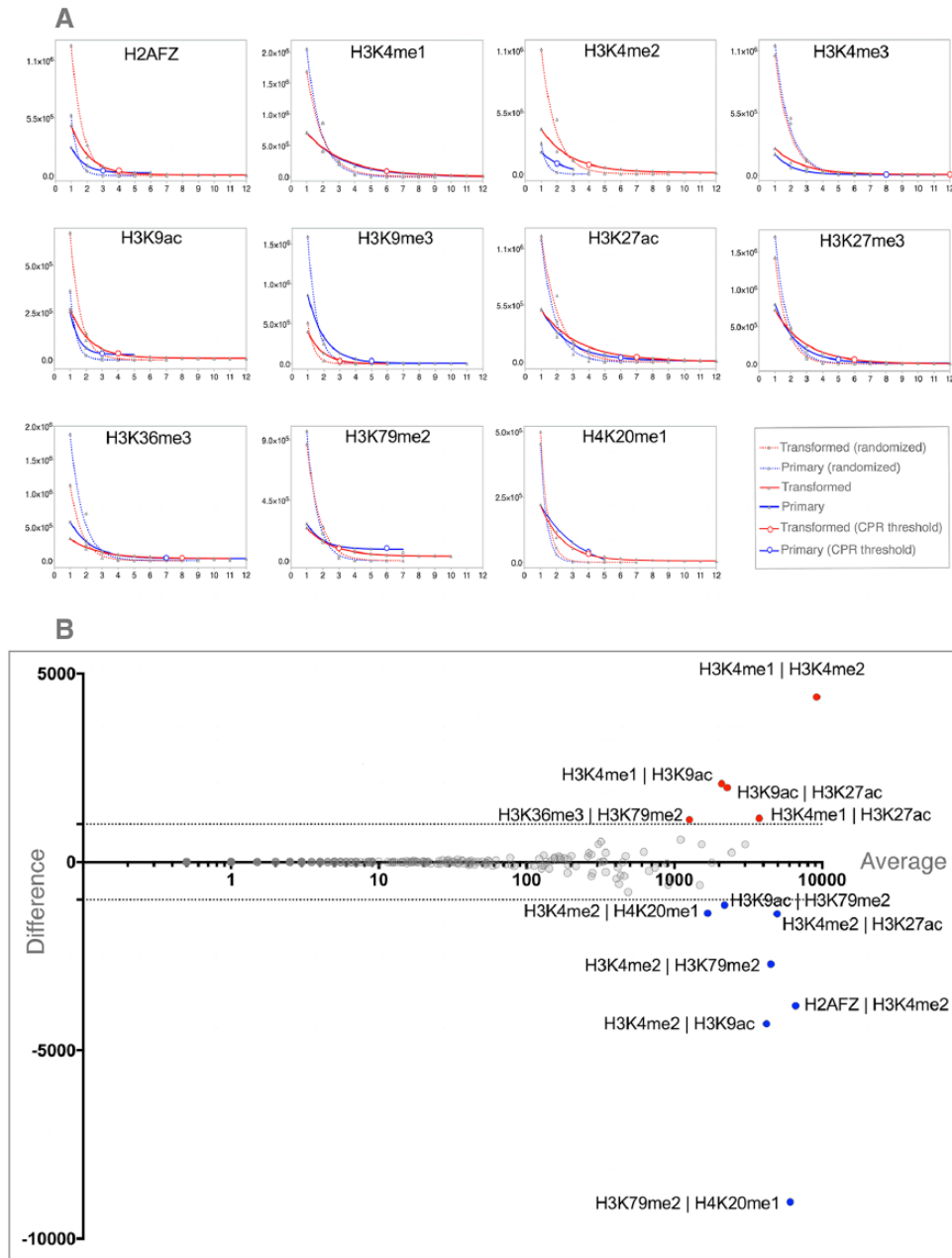
The sequences of the p-CPR and t-CPR for each histone modification were subjected to a non-discriminative motif search. Unexpectedly, a consensus DNA sequence motif was found in CPR sequences for each histone modification in both primary and transformed cell types (Figure 2A). These motifs were called “non-discriminative CPR motifs.”

These findings suggested that the motifs in CPR are sites for preferential histone occupancy in a wide spectrum of primary and transformed cells. However, histone modification and nucleosomal occupancy are disturbed in cellular transformation.<sup>35</sup> So, we next wanted to know if there are additional sequence features not revealed by the above-mentioned motif search that are subtly different between p-CPR and t-CPR.

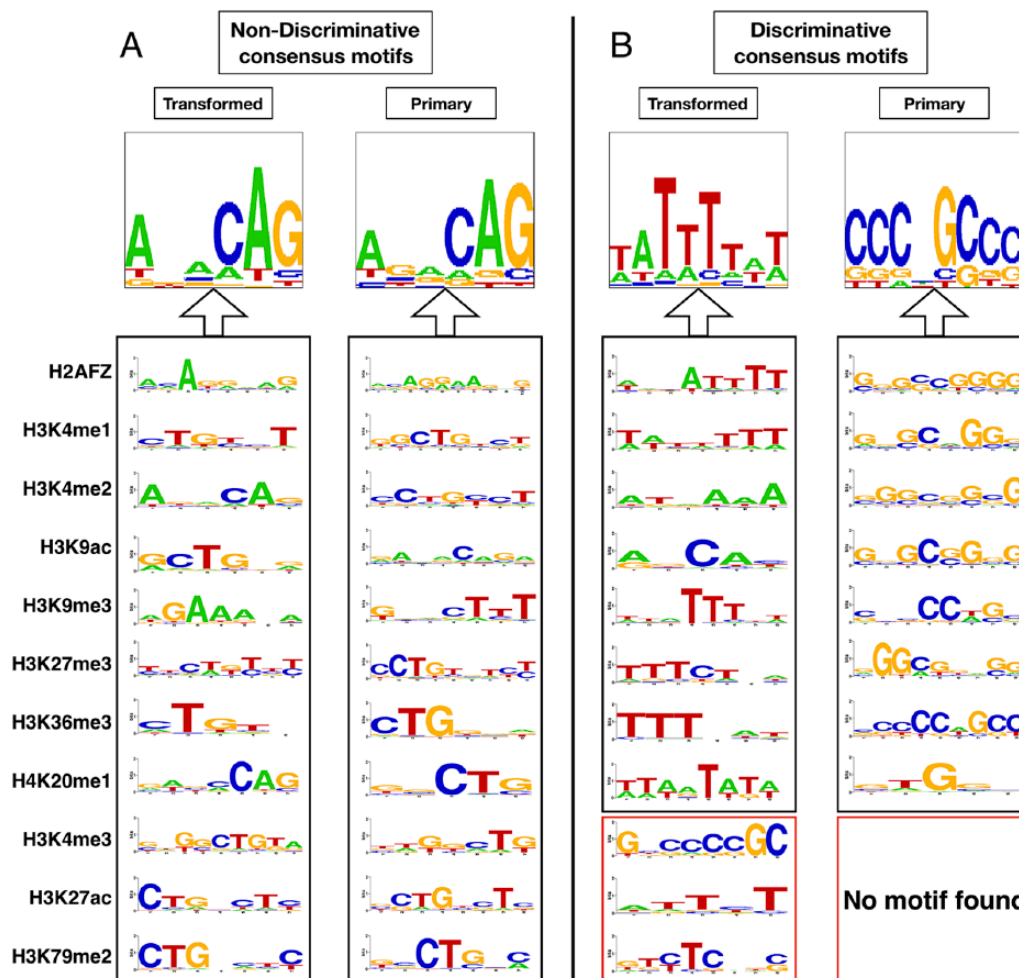
### *Discriminative motif search revealed subtle differences in t-CPR and p-CPR*

The findings described above underscored remarkable sequence similarities between the DNA sequences of p-CPR and t-CPR for any particular histone modification across different primary and transformed cell types. Such a sequence similarity between two different sets of sequences (p-CPR and t-CPR) could arise due to (1) an overlap of p-CPR and t-CPR genomic coordinates or (2) presence of same motifs in non-overlapping p-CPR and t-CPR sequences. In the former case, the distance between p-CPR and t-CPR would be lower than the latter. We tested these possibilities by measuring the closest distances between the centers of p-CPRs and t-CPR with sorted genomic coordinates.

The distribution of closest distances between p-CPR and t-CPR was obtained and compared with a mock closest distance profile of the same set of CPR after randomization between primary and transformed groups (10 different rounds of randomization of all CPR into mock sets).



**Figure 1.** Identification of histone CPR and their co-occupancy. (A) For each histone modification, the number of CPR (100bp and more) obtained after performing multiple intersections among the ChIP-seq peaks in primary and transformed cells separately was plotted (Y-axis) against the number of cell lines (X-axis). The Y-axis values corresponding to the minimum X-axis value of 1 represent the number of unique regions specific to any one cell line only. Y-axis values corresponding to X-axis values of 2, 3, and 4, for instance, represent the number of CPR common across any two, any three, or any four cell lines, respectively. On the X-axis, the deduction of CPR with successive increment of 1 cell at a time is denoted by 1-phase decay curves for primary cells (blue solid line) and transformed cells (red solid line). The candidate threshold numbers of primary and transformed cells have been highlighted by blue and red circles for primary and transformed cells, respectively. Similar decay curves were plotted with artificial CPR, generated by Bedtools shuffleBed, as controls. Then, multiple intersections between the shuffled peaks of different primary and transformed cells were performed separately using Bedtools multiintersectBed to generate artificial CPR (blue dashed lines for primary and red dashed lines for transformed cells). The artificial CPR lost commonality (Y-axis values) rapidly as the number of cell lines increased (X-axis). At the thresholds for CPR identification (on or before 12 different combinations of cells), no common regions could be retained in most of the artificial CPR showing that CPR retention in ChIP-seq datasets is a non-random phenomenon and is a specific outcome of the ChIP-seq peak location. (B) To identify if some of the 11 histone modifications co-occur, Bedtools multiintersectBed was performed between ChIP-seq peaks of all histone modifications separately for primary and transformed cells. Mean of the number of regions co-occupied by at least two different histone modifications is plotted on the X-axis, and the difference between them is plotted on the Y-axis. Larger values on X-axis depict higher levels of co-occupancy of different histone modifications. The colored points show the pairs of histone modifications that are co-abundant in either more in primary cells (blue) or in transformed cells (red) more than the threshold (1000 locations genome-wide; positive and negative values on Y-axis mean higher co-abundance of histone modification in transformed and primary cells, respectively). All of these marked histone modifications were activating pairs except the pair of H2AFZ and H3K4me2. ChIP-seq indicates chromatin immunoprecipitation sequencing; CPR, ChIP-seq Peak Regions.



**Figure 2.** CPR exhibit DNA sequence motifs both with and without cell-type specificity. (A) Highly significant motifs were identified using DREME on sequences of CPR for primary and transformed cells in a non-discriminative mode (not shown) and top 10 motifs used as inputs for generating consensus motifs (using STAMP). The consensus motifs obtained in transformed and primary cell types are indicated with the names of respective histone modifications. Core consensus motifs were deduced from the histone modification-specific consensus motifs for transformed and primary cells (A). These core consensus motifs were strikingly similar between transformed and primary cells suggesting commonality of sequence features across all CPR independent of cell-type identity and the kind of histone modification. (B) To identify the presence of any cell-type specific motif in the primary and transformed CPR, discriminative motif search was performed by providing 1 set of CPR as background (negative sequence set) for the another (positive sequence set) in DREME. The consensus motifs in the transformed cells (t-CPR motifs) were relatively AT-rich for all histone modifications (except for H3K4me3) than the consensus motifs in primary cells (p-CPR motifs) that were relatively GC-rich. However, p-CPR motifs could not be identified in H3K4me3, H3K27ac, and H3K79me2. Core consensus for the discriminative motifs deduced using STAMP tool showed that indeed the t-CPR motifs were AT-rich and the p-CPR motifs were GC-rich. Unlike the non-discriminative CPR motifs which were highly similar between primary and transformed cells, the t-CPR and p-CPR motifs showed no similarity. All the motif identification were done using default significance values of DREME and STAMP. ChIP-seq indicates chromatin immunoprecipitation sequencing; CPR, ChIP-seq Peak Regions; DREME, Discriminative Regular Expression Motif Elicitation.

Surprisingly, this closest distance distribution analysis revealed that the closest distance between p-CPR and t-CPR is higher than randomly expected (Supplemental Figure S1 and data not shown) (Kruskal-Wallis test,  $P$  value  $> 0.4543$  for significance of difference within mock CPR,  $P < 0.0001$  for difference between actual p-CPR to t-CPR distance). The mock p-CPR and t-CPR coordinates also exhibited a strong increase in overlapping coordinates as compared with the actual p-CPR and t-CPR coordinates. This was true for all histone modifications that were tested. This finding established that the motif similarity between

p-CPR and t-CPR exists even if they occur in distinct genomic coordinates.

The occurrence of p-CPR and t-CPR in distinct genomic regions raised some interesting possibilities. Are there subtle DNA sequence features that are unique to p-CPR or t-CPR? How is the CPR displacement relevant for gene expression in cellular transformation? Is cellular transformation associated with differential usage of CPR?

To identify any distinct p-CPR-specific and t-CPR-specific motifs, we performed a discriminative motif search by using 1 set as a background against the other. Results showed that for

the CPR of H2AFZ, H3K4me1, H3K4me2, H3K9ac, H3K9me3, H3K27me3, H3K36me3, and H4K20me1, distinct p-CPR-specific and t-CPR-specific motifs exist. The t-CPR-specific motifs (t-CPR motifs) for all histone modifications were AT-rich (GC-poor) unlike the p-CPR-specific motifs (p-CPR motifs), which were GC-rich in nature. However, for the CPR of H3K4me3, H3K27ac, and H3K79me2 histone modifications, only t-CPR motifs could be discovered which were AT-rich for H3K27ac and H3K79me2 and GC-rich for H3K4me3 (no p-CPR motifs could be identified). This reinforced that the GC-richness of p-CPR motifs and AT-richness of t-CPR motifs is a specific phenomenon restricted to certain histone modifications only (Figure 2B). We concluded that although non-discriminative CPR motifs are associated with the occurrence of histone modifications in primary and transformed cells both, there are also different primary-specific and transformed-specific motifs for some histone modifications.

*Unlike p-CPR motifs, the t-CPR motifs are not concentrated near CpG islands-associated TSSs*

The high GC content of p-CPR motifs as compared with the t-CPR motifs led us to argue that (1) high levels of histone occupancy in GC-rich regions, such as CpG islands, are maintained in primary cells, and (2) this relationship between CPR and GC-rich regions gets disrupted in transformed cells such that both the locations of t-CPR and their underlying DNA sequence identifier motif become different from those of p-CPR.

For studying the changes in CPR and its effect on gene expression relevant in cellular transformation, we first narrowed down to a subset of CPR for each histone modification. The midpoints of these selected CPR were located within 1 kb of the nearest CpG island. These regions were either p-CPR motif-positive or t-CPR motif-positive. All the TSSs located within the coordinates of such CPR and associated CpG islands (CPR-TSSs) were filtered out for primary and transformed cells separately. For the 8 different histone modifications for which both p-CPR and t-CPR motifs were available, different numbers of CPR-TSSs were found to be shared between primary and transformed cells (detailed in the “Methods” section). The CPR-TSSs were important because they were strong candidates for transcriptional regulation by a CpG-rich sequence in cis and yet were exhibiting a GC-rich p-CPR to GC-poor t-CPR motif preference in cellular transformation.

To understand how differential histone occupancy at CPR motifs might affect the TSSs, the distance of p-CPR and t-CPR motifs from CPR-TSSs was calculated. Genomic location plots centered at the CPR-TSSs depicting the occurrence of p-CPR motifs and t-CPR motifs showed a striking concentration of p-CPR motifs closer to the TSSs, whereas the t-CPR motifs were located at a consistently larger distance. Without any exception, for all the 8 histone modifications, the GC-rich p-CPR motifs were located closer to the TSS than the GC-poor t-CPR motifs (Figure 3). In cases where multiple

p-CPR and t-CPR motifs were present, we used the coordinates of the p-CPR and t-CPR motifs closest to the TSS for calculating the distances. These findings proved that many CpG island-associated TSSs contain p-CPR motifs and t-CPR motifs in flanking regions and that the transformed cells preferentially employ distant t-CPR motifs for histone occupancy over the proximal p-CPR motifs employed by the primary cells. These findings also indicated that the regions between p-CPR motifs and t-CPR motifs are transition zones for the GC and CpG contents.

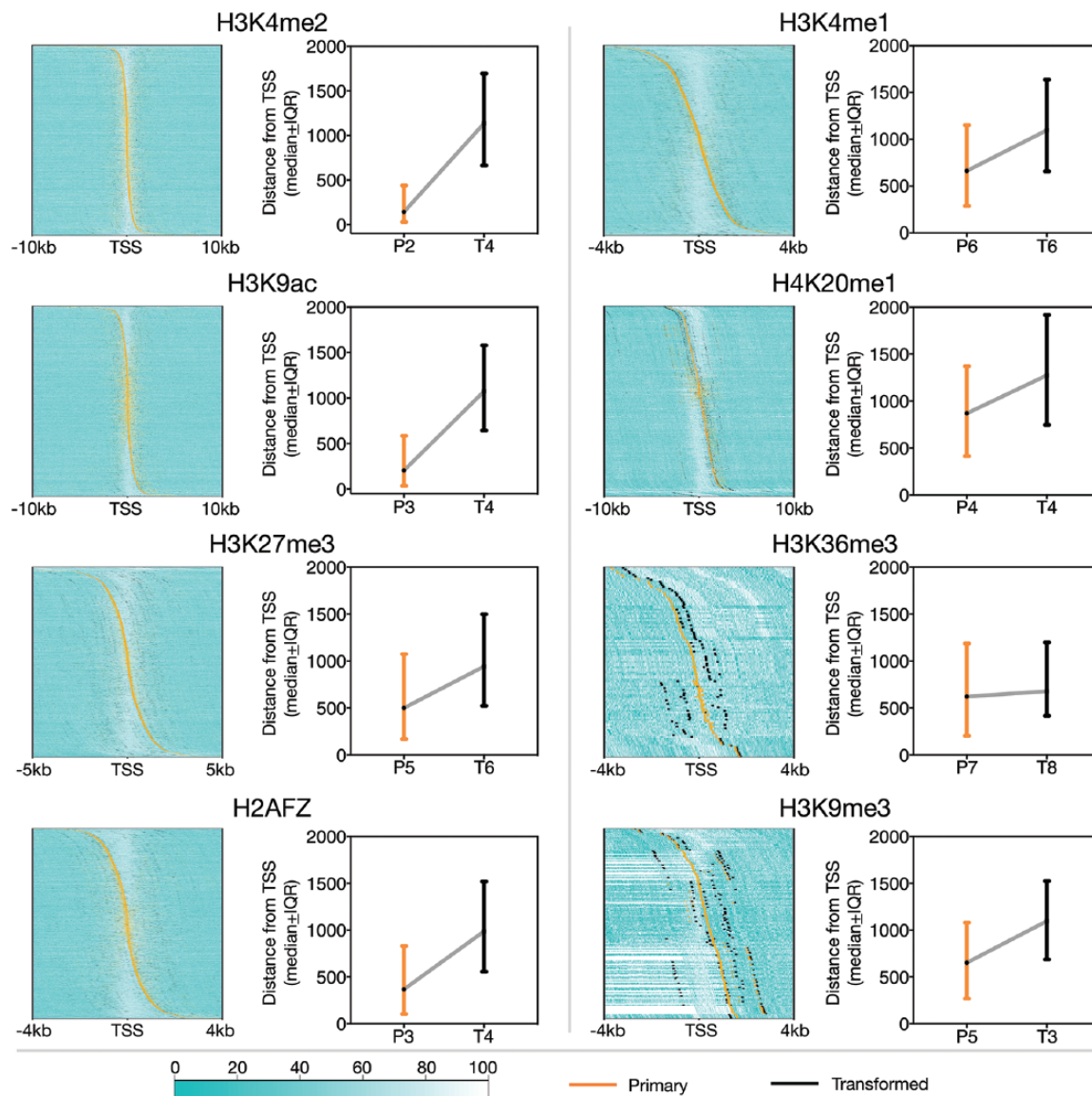
*ChIP-seq reads distribution recapitulate the patterns of p-CPR and t-CPR motif distribution around CPR-TSSs*

The occurrence of p-CPR and t-CPR motifs was deduced from the analyses of ChIP-seq peaks derived from different independent experiments. To verify the robustness of these findings, we wanted to ensure that the post-alignment read-level ChIP-seq data were in agreement with the findings obtained from peak-level analyses. We mapped the read densities for the 8 histone modification ChIP-seq datasets from comparable pairs of transformed and primary cells<sup>36–38</sup> in 5 kb flanks of CPR-TSSs. We compared histone modification datasets from A549 cell line with IMR90 cells and OCI-LY1 cell line with B-cells.<sup>38–40</sup> Reinforcing our findings that a p-CPR to t-CPR motif shift occurs at CPR-TSSs, the results showed that the ChIP-seq read densities were stronger in the immediate flanks of the CPR-TSSs in normal cells and relatively distal in transformed cells (Supplemental Figures S2 and S3). As expected, for each histone modification, a distinct read density profile was observed, but the differences in the density of reads between primary and transformed cells relative to the CPR-TSSs were consistent. In comparison, the TSSs falling in CPR that are unique either to transformed or primary cells and not containing p-CPR or t-CPR motifs (non-CPR-TSSs) showed a different pattern of ChIP-seq read densities (Supplemental Figures S2 and S3 and Supplemental Table S3). Although the read densities at non-CPR-TSSs were weaker in transformed cells compared with the primary cells, the shift away from the TSSs as observed for the CPR-TSSs was not visible. Thus, by mapping ChIP-seq read densities in representative pairs of primary and transformed cells, we could reinforce the robustness of our findings.

*Distinct sequence features mark flanking sequences of TSSs with t-CPR and p-CPR motifs*

The presence of motifs of contrasting GC-richness near some CPR-TSSs suggested that in the flanks of these TSSs, the GC content is non-uniform. To understand the nature of GC-content heterogeneity, we first analyzed how the GC and CpG contents change in the region between p-CPR and t-CPR motifs. The GC-richness and CpG density were plotted for the regions between p-CPR motifs and t-CPR motifs





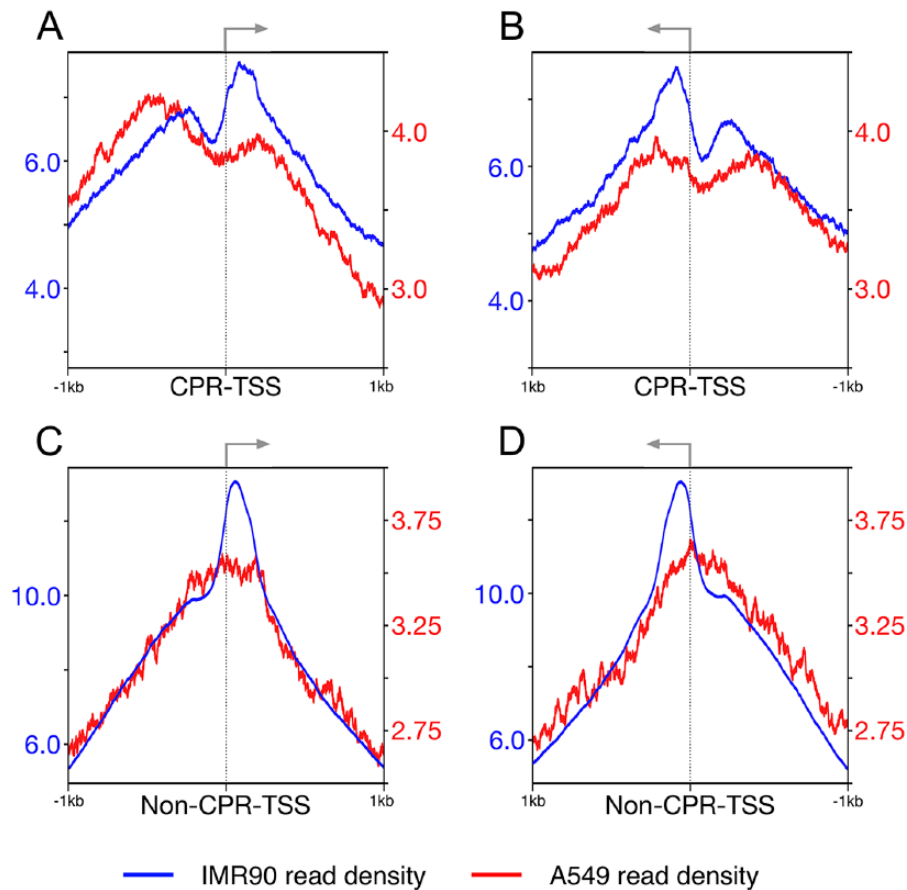
**Figure 3.** p-CPR and t-CPR motifs have distinct localization with respect to the CPR-TSSs. For each of the 8 histone modifications as indicated, the locations of individual p-CPR and t-CPR motif with respect to the associated CPR-TSSs were plotted on the heat-maps (left flank). The graphs (right flank) show median distance of the motif-midpoint from the TSS midpoint ( $\pm$ interquartile range (IQR) values) for the same data as plotted in the heat-map to the left. The X-axis of the heat-maps specifies the distance from TSS midpoint. The Y-axis of the heat-maps has all the CPR for a histone modification sorted in the increasing order of the downstream distance between a p-CPR motif and respective TSSs. The p-CPR motifs (orange) and t-CPR motifs (black) display a pattern in the heat-maps such that the p-CPR motifs are proximal to TSSs than the t-CPR motifs. This pattern is highly statistically significant as analyzed by Wilcoxon matched-pairs signed rank test on the closest pairs of p-CPR and t-CPR motif distances from the TSS midpoints ( $P < 0.0001$ ) for all histone modifications except H3K36me3 ( $P = .9296$ ). The background of the heat-maps depicts the GC content (maximum as white and minimum as aqua). For visualization, each non-overlapping motif (approximately 8bases long) is depicted as a 100-bases long signal. CPR indicates ChIP-seq Peak Regions, TSSs, transcription start sites.

(Supplemental Figure S4). It was observed that the GC content was high at the p-CPR motifs and decreased toward the t-CPR motifs. This corroborated our previous findings (Figure 3) that the p-CPR motifs are positioned closer to the TSSs in GC-rich regions unlike the t-CPR motifs. The region defined by p-CPR motifs and t-CPR motifs thus turned out to be a region within which the GC content reduces abruptly as the coordinates shift away from TSSs. This feature was also replicated when CpG density was plotted in the same regions (Supplemental Figure S4). The CpG density was higher than expected for the p-CPR motifs and dropped to expected levels

at the GC-poor t-CPR motifs (Supplemental Figure S4). These findings showed that through a motif selectivity for GC-poor t-CPR motifs, the transformed cells employ CpG-poor regions for determining histone occupancy.

The p-CPR and t-CPR motifs were all identified in repeat-masked sequences. Transcriptional regulation by them at the linked TSSs was a tantalizing possibility. However, it remained unclear how repeats and repeats-associated recurring sequence motifs could affect these TSSs. To eliminate the probability that the possible regulation of TSSs by p-CPR and t-CPR motifs could be confounded by repeat sequences, we measured repeat content in





**Figure 4.** Different patterns of H2AFZ occupancy in a representative primary and transformed cell pair at CPR-TSSs and non-CPR-TSSs suggest a role of CPR motifs in transcription. (A) and (B) CPR-TSSs on top (A) or bottom (B) strands exhibited a highly similar pattern of H2AFZ occupancy. The higher distance of H2AFZ read density peaks from the TSS midpoint for transformed cell A549 (red) than that of the primary cell IMR90 (blue) corroborates our findings that CPR motifs cluster closer to TSSs in primary cells than in transformed cells. (C) and (D) At non-CPR-TSSs, however, there was a strong difference in H2AFZ occupancy exclusively in a region immediately downstream (approximately <100bp) of the TSSs. The IMR90 cells showed a strong enrichment of H2AFZ at the site where RNA Polymerase II pausing is expected. This H2AFZ enrichment downstream of TSSs was missing in A549 cells for both the top (C) and bottom (D) strands. No qualitative difference was observed for the H2AFZ occupancy for IMR90 and A549 away from the non-CPR-TSSs. The CPR-TSSs and non-CPR-TSSs thus exhibit different patterns of H2AFZ occupancy in normal cells—compare blue curves in (A) and (C) with blue curves in (B) and (D), respectively—and this difference is further strengthened in transformed cells—compare red curves in (A) and (C) with red curves in (B) and (D), respectively. H2AFZ occupancy at CPR-TSSs for A549 exhibits a noticeable strand-specific bias due to unknown reasons. CPR indicates ChIP-seq Peak Regions, TSSs, transcription start sites.

immediate upstream regions of these TSSs. We found that the CPR-TSSs were located in regions (0.5kb upstream flank) with unexpectedly low repeat content as compared with non-CPR-TSSs (Supplemental Table S4). A strand-specific analysis of CPR-TSSs (0.3kb upstream flank) revealed that there is no strand-specific bias in the repeat contents of immediate upstream flank sequences and TSSs on both strands are equally repeat-poor (Supplemental Table S5). Together these results indicated that p-CPR and t-CPR motifs are major regulators of these TSSs and they potentially act differently between primary and transformed cells.

#### *Differential H2AFZ occupancy at CPR and non-CPR-TSSs indicates a role of CPR motifs in transcription*

The disturbances in histone occupancy at TSS flanks impact transcription. The presence of positioned nucleosomes, with H2AFZ as a component, defines the TSSs.<sup>2</sup> To understand if

the CPR-TSSs employ altered transcription at the start sites, we focused on H2AFZ occupancy at CPR-TSSs and non-CPR-TSSs in primary (IMR90) and transformed (A549) cells. H2AFZ occupancy when mapped at read-levels for these cells revealed that the TSS-positioning peaks of H2AFZ are clearly maintained in IMR90 cells but lost in A549 cells. In IMR90 cells, the H2AFZ signal was concentrated downstream of TSS (approximately <100bp) at non-CPR-TSSs. However, at CPR-TSSs, this downstream concentration of H2AFZ was at a larger distance (approximately >100bp) (Figure 4). Interestingly, this pattern of H2AFZ read density upstream of CPR-TSSs when compared with that of A549 cells showed a difference that followed the same pattern as we had found for the distribution of p-CPR and t-CPR motifs (Figure 3). Like the p-CPR motifs, the H2AFZ reads were concentrated closer to the TSSs in IMR90 (primary) cells and distal in A549 (transformed) cells. At the non-CPR-TSSs, there was no such qualitative difference in H2AFZ occupancy. Hence, the A549 cells

showed no H2AFZ enrichment distal from the non-CPR-TSSs compared with IMR90. The only difference between H2AFZ occupancy in IMR90 and A549 cells was that the latter showed a weaker occupancy of H2AFZ immediately downstream of TSSs (Figure 4). These findings collectively demonstrated that the CPR and non-CPR classification of the TSSs is justified because (1) the difference in histone occupancy at CPR-TSSs and non-CPR-TSSs is consistent across various ChIP-seq datasets, (2) CPR-TSSs and non-CPR-TSSs display distinct patterns of histone occupancy between primary and transformed cells, and (3) for a selected panel of datasets for H2AFZ, the distribution of p-CPR motifs and t-CPR motifs recapitulate the occurrence of actual H2AFZ-bound sequence reads (Figure 4).

The analyses so far demonstrated that histone occupancy, including that of H2AFZ at CPR-TSSs and non-CPR-TSSs is different between primary and transformed cells. Next, we wanted to know how the altered H2AFZ occupancy at CPR-TSSs and non-CPR-TSSs translates into transcriptional disturbances in primary and transformed cells. We applied the CPR and non-CPR classification of the TSSs to test if these were associated with different transcription patterns and RNA Polymerase II (POLR2A) occupancy across available RNA-seq and POLR2A ChIP-seq datasets, respectively.

*A p-CPR to t-CPR motif preference for H2AFZ in transformed cells coincides with deregulated transcription in TSS flanks*

ChIP-seq Peak Regions-transcription start sites exhibit a shift of p-CPR motif to t-CPR motif for histone occupancy. Such a shift at the TSSs for positioned nucleosomes which consist of H2AFZ can potentially affect the POLR2A occupancy at CPR-TSSs and non-CPR-TSSs differently. We analyzed H2AFZ occupancy, POLR2A occupancy, and RNA-seq read density in datasets from a panel of primary and transformed cell lines (Supplemental Tables S6 and S7) to test these possibilities.

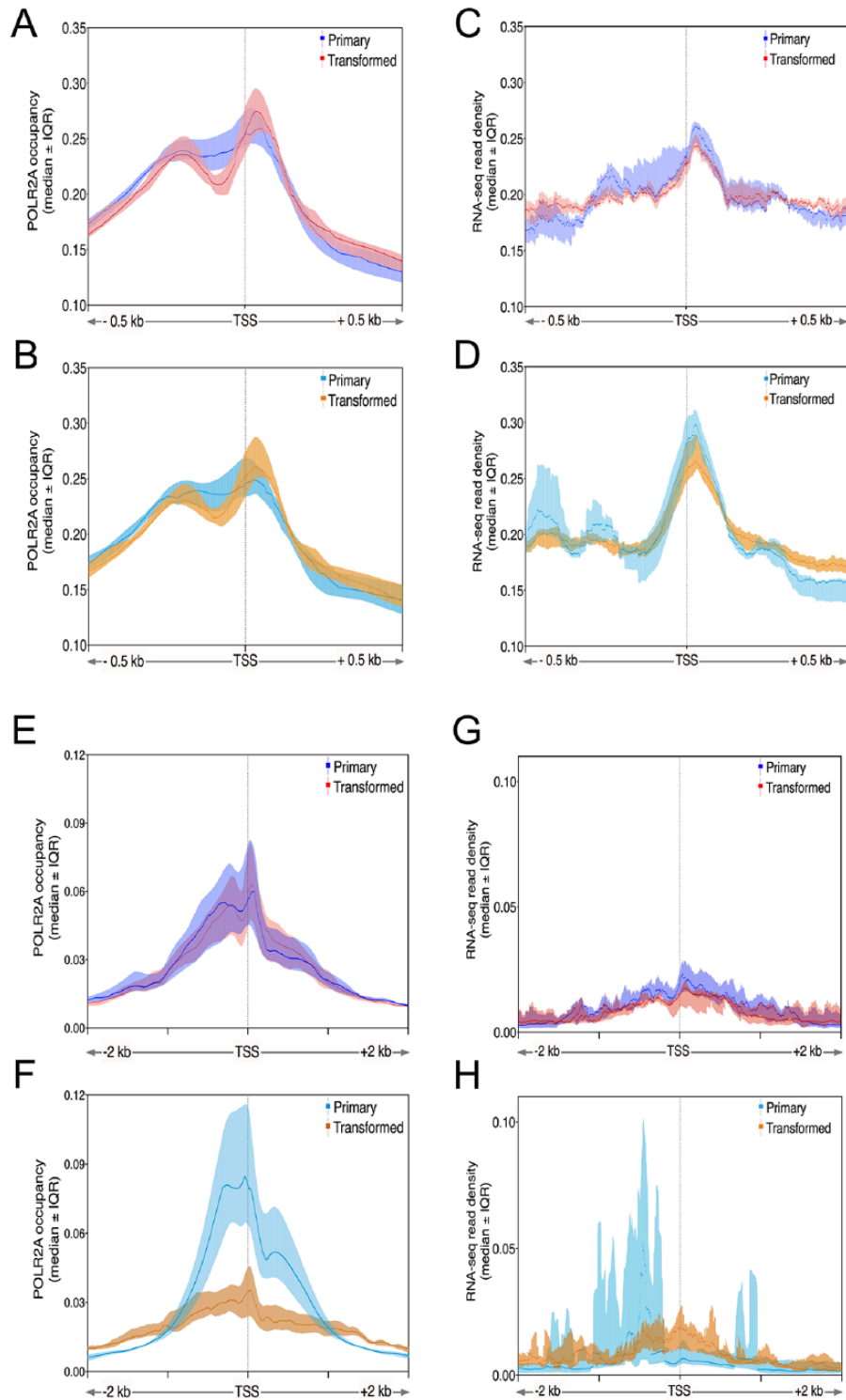
POLR2A occupancy was plotted in 0.5 kb flanks of CPR-TSSs as well as for a control set of TSSs that did not show any association with the CPR motifs. POLR2A occupancy exhibited similar pattern at CPR-TSSs in primary and transformed cells. Thus, by employing a p-CPR to t-CPR motif preference, the CPR-TSSs escaped the transformation-associated anomalous POLR2A recruitment (Figure 5A and B). A highly similar pattern of RNA-seq read density distribution in primary and transformed cells was observed in both these sets of coordinates (Figure 5C and D).

Next, we performed a targeted analysis of POLR2A occupancy and RNA-seq read density in CPR-TSSs and non-CPR-TSSs, specifically for H2AFZ (Figure 5E-H). POLR2A occupancy was highly similar between primary and transformed cells at H2AFZ CPR-TSSs (Figure 5E). On the other hand, H2AFZ non-CPR-TSSs showed highly different POLR2A

occupancy (Figure 5F). The transformed cells showed strikingly low POLR2A occupancy both upstream and downstream of non-CPR-TSSs in the direction of transcription (red and orange samples in Figure 5E and F, respectively). The primary cells, on the other hand, showed a robust POLR2A recruitment at non-CPR-TSSs as expected with a subpeak of POLR2A read density depicting upstream binding, and a downstream subpeak depicting the pausing of POLR2A. Interestingly, this difference in POLR2A recruitment between primary and transformed cells was not observed in CPR-TSSs (blue sample in Figure 5E) and was restricted only to the non-CPR-TSSs (aqua sample in Figure 5F). The classification of TSSs as CPR-TSSs and non-CPR-TSSs thus proved to be instrumental in identifying groups of TSSs that maintain distinct histone occupancy in transformed and primary cells, thereby affecting POLR2A occupancy.

We argued that if different levels of POLR2A occupancy at H2AFZ-bound CPR-TSSs and non-CPR-TSSs are functionally relevant, then it would affect transcriptional activity in the flanks of these TSSs. The functional impact of different POLR2A occupancy at H2AFZ-bound CPR-TSSs and non-CPR-TSSs could thus be tested by analyzing RNA-seq datasets of different cell lines and correlating transcript levels with the POLR2A occupancy. We measured the RNA-seq read density at H2AFZ-bound CPR-TSSs and non-CPR-TSSs (same as the ones on which POLR2A occupancy was plotted). The CPR-TSSs in both primary and transformed cells showed RNA-seq reads distributed on both sides of the TSSs, as was expected from the POLR2A distribution in these regions (Figure 5G and H). In transformed cells, the RNA-seq and POLR2A occupancy showed a strong correlation in 1 kb flanks of CPR-TSSs (Supplemental Figure S5A and B) and non-CPR-TSSs (Supplemental Figure S5C and D). Clearly, the H2AFZ-bound CPR-TSSs maintained similar pattern of transcript levels in primary and transformed cells than the non-CPR-TSSs. This correlated RNA-seq and POLR2A occupancy in transformed cells were same as seen in primary cells for the CPR-TSSs only (Supplemental Figure S5E and F). The correlated RNA-seq and POLR2A occupancy in non-CPR-TSSs seen in transformed cells (Supplemental Figure S5C and D) were different from the generally uncorrelated pattern expected in primary cells (Supplemental Figure S5G and H).

These findings were consistent with the knowledge of RNA Polymerase II pausing immediately downstream of the TSSs. Because such a pausing of RNA Polymerase II is expected to be its non-productive retention, in primary cells the RNA-seq and POLR2A read density are expected to be uncorrelated. In transformed cells, however, this pattern was lost and POLR2A occupancy seemed to be coupled with transcription, and hence positively correlating with RNA-seq read density. These data indicated that unlike CPR-TSSs, the non-CPR-TSSs in transformed cells are prone to runaway transcriptional activity by RNA Polymerase II. At the CPR-TSSs, however, a p-CPR to



**Figure 5.** H2AFZ non-CPR-TSSs display distinct patterns of POLR2A and RNA-seq reads density in primary and transformed cells. (A) and (B) POLR2A ChIP-seq read density at all CPR-TSSs (A) and all non-CPR-TSSs (B) are highly similar. (C) and (D) RNA-seq read density in the same regions as plotted in (A) and (B), respectively, also shows a similar pattern between CPR (C) and non-CPR-TSSs (D) as well as primary and transformed cells, especially in immediate flanks where  $-1$  and  $+1$  positioned nucleosomes are expected to be located. (E) and (F) POLR2A occupancy at CPR-TSSs (E) and non-CPR-TSSs (F) specifically for H2AFZ, a positioned nucleosome component, showed striking dissimilarity that stands out of the pooled analysis of all 8 histone modifications (A) and (B). The POLR2A showed enrichment at approximately 100 to 200bp flanks of H2AFZ CPR-TSSs (E) which increased to approximately 300 to 500bp flanks of H2AFZ non-CPR-TSSs (F). At the H2AFZ non-CPR-TSSs, the primary and transformed cells showed the strongest difference in POLR2A occupancy. (G) and (H) In the same regions as in (E) and (F), RNA-seq reads were plotted. At H2AFZ CPR-TSSs (G), the RNA-seq reads recapitulated the pattern of POLR2A occupancy (E) in both primary and transformed cells. However, as shown in (H), RNA-seq reads in the H2AFZ non-CPR-TSSs showed good correlation only for transformed cells, not primary cells. Values plotted on Y-axis are median and interquartile range (IQR) for 13 primary and 9 transformed cell lines for (A), (B), (E), and (F), and 9 primary and 18 transformed cell lines for (C), (D), (G), and (H). Correlation plots and Pearson coefficients for pairs of POLR2A and RNA-seq data plotted here are shown in Supplemental Figure S5. CPR indicates ChIP-seq Peak Regions, TSSs, transcription start sites.



t-CPR motif shift mitigates such aberrant transcriptional activity.

## Discussion

It has been established that DNA sequence can affect positioning of individual nucleosomes. Within nucleosomes, histones co-occur in different combinations of post-translational modifications. These post-translational modifications of histones affect the regulation of nucleosomal as well as higher order chromatin structure and function. A large amount of data exist (typically histone modification ChIP-seq data) in the public domain that describes the association of variously modified histones with human genomic DNA in different types of cells. As the histones typically interact with DNA as components of nucleosomes, histone occupancy can indicate nucleosomal occupancy. Histone occupancy data from pools of cells, with each cell having its own unique histone occupancy profile, do not contain information about nucleosomal boundaries. However, if site variability of histone occupancy has a tendency such that certain genomic regions tend to have more histone occupancy than others in a population of cells, then this information is embedded in such histone modification ChIP-seq data. In this work, we have systematically mined the histone modification ChIP-seq data regardless of age and gender in an unbiased manner for primary and transformed cells and identified genomic regions that tend to have high histone occupancy and described their unexpected sequence features. Our comparative approach led us to discover that genomic regions with a high tendency of histone occupancy have functionally relevant subtle differences in sequence features between primary and transformed cells.

Nucleosomal occupancy has also been studied using DNase-seq. However, the footprint of DNase digestion depends on the binding of several different protein complexes to the DNA other than histones as well. This difference between DNase-seq and histone ChIP-seq datasets is strongly manifested in the narrow peaks. In the narrow peaks, the nucleosomes are present in low density, making the DNA accessible to DNA-binding proteins. Indeed, we have found (data not shown) that similar to CPR, non-discriminative motif-containing common peak regions can also be identified in DNase-seq datasets, albeit with a much lower frequency. These findings were obtained from 35 primary and 25 transformed cell lines (ENCODE).

The publicly available histone modification ChIP-seq data are at the level of peaks as well as reads, wherein each peak represents a central tendency of a cluster of reads. Traditionally, the ChIP-seq peaks for sequence-specific DNA-binding proteins such as TFs, not histones, are probed for central enrichment of sequence motifs. However, establishment and maintenance of chromatin conformation by sequence-specific binding factors also regulate the histone occupancy and modification profile of the target region. In such a condition, genomic regions with a strong tendency to be histone-bound will also harbor DNA sequence motifs for such chromatin organizer proteins. This is a

tangible possibility because DNA sequence-specific binding of protein such as CTCF and YY1 can regulate occupancy of specifically modified histone modifications.<sup>41–45</sup>

Our non-candidate approach has involved ChIP-seq datasets from several primary and transformed cell lines available on ENCODE. Derivation of common peak regions from such a diverse group of multiple datasets from independent experiments makes our approach more robust and increases both statistical strength and biological relevance of our findings.

To eliminate the arbitrariness in selecting the threshold number of cell lines for calling of CPR, we calculated the commonly retained regions for different numbers of independent samples, with no prior filtering of the published datasets. To further strengthen the confidence in the CPR identification, we used ChIP-seq peaks as input for our analyses, where the reads are reported to cluster over and above the background in different samples. We also applied the same CPR identification protocol on a mock set of sequences obtained by randomly selecting genomic regions of similar length and chromosomal profiles as of the peaks used for actual CPR calling. Thus, we ensured that the CPR could be called only with actual peak sequences and not with a random set of sequences, thereby establishing the specificity of CPR. Also, by selecting a threshold number of cell lines such that comparable number of CPR are called between primary and transformed cells, we ensured that any downstream processing of the data is not affected by different CPR counts between the two sets.

Given that we were analyzing data from completely disparate sets of cell types, in which histone occupancy was expected to be widely different due to different lineages of differentiation and cellular transformation, identification of any CPR was unexpected. To the contrary, we identified thousands of recurrent sequences in multiple datasets. That these are over 0.1 kb long sequences again reinforces that the CPR are not accidentally short occurring sequences due to unidentified random probabilities. Thus, variously modified histones do gravitate toward these CPR and thus, in disparate datasets, exhibit a tendency to be associated with them.

Histone modifications and their occupancy undergo global deregulation in transformed cells.<sup>28</sup> Thus, we derived the CPR and performed downstream analyses in primary and transformed cells separately. The CPR, the different motifs, and sequence properties that we have reported are hence representatives of primary and transformed cell types and not of cell types, tissue types, or modes of cellular transformation. We have thus separated out the CPR findings that do not differentiate transformed and primary cells from those which are distinct between primary and transformed cells.

To identify common sequence motifs in CPR, the sequences of CPR for each histone modification were subjected to motif identification. The majority of CPR, independently discovered for 11 out of 11 histone modifications, had similar motifs between primary and transformed cells. The ChIP-seq peaks used as input

in this analysis were all from the repeat-masked genome, thereby eliminating the possibility that the CPR motifs identified are repetitive sequences. These unexpected findings show that global histone occupancy for several histone modifications is primarily regulated by robust mechanisms that are universal for different tissue/cell types and remain unaltered in cellular transformation. However, these results did not rule out the role of additional subtler tissue/cell-type-specific or transformation-specific mechanisms. A prior classification of the samples into primary or transformed groups helped us to discover such hidden DNA sequence elements associated with histone occupancy in these 2 cell-type groups distinctly. The usage of similar sizes of sequences as motif-negative sets in a discriminative motif search led us to highly similar motifs in CPR of 8 out of 11 histone modifications. That the CPR for 2 modifications did not follow the same pattern and an additional third histone modification showed opposite patterns for only cell types suggesting indirectly that CPR of some histone modifications do not follow the discriminative motif pattern we have identified. Thus, although the p-CPR and t-CPR motifs are common to CPR of 8 histone modifications, they are not universal features of sequences that differentially regulate histone occupancy in transformation.

The CPR were first called at the level of individual histone modifications and then analyzed at the cell-type-specific level. At a single cell level, it is expected that CPR of co-occurring histone modifications (as histones bind to DNA as components of nucleosomes) will be overlapping regions. Although identifying such an overlap in data derived from a mass of cells is not possible, it is also expected that the regions with high nucleosomal occupancy will nonetheless recur. Expecting a physical overlap between CPR coordinates of numerous independent and diverse samples is prone to type II error. However, any location-independent sequence features common to CPR could be identified free from such systematic errors, as proven by our results. The most unexpected feature of the CPR is that they harbor two different types of motifs: (1) the universal CPR motifs present in most histone modifications and most cell types analyzed, and (2) a restrictive set of motifs that are present in the CPR but preferentially employed by either primary or transformed cells only. A combined synthesis of this finding is that while the universal CPR indiscriminately favor histone occupancy, the motifs specific for p-CPR and t-CPR fine-tune the CPR around the universal motifs differently in primary or transformed cells. The p-CPR motifs are GC-rich, and CpG hypermethylation in gene promoters is an often observed epigenetic feature of cancer cells.<sup>46</sup> Interestingly, the sequence-specific DNA-binding proteins that regulate histone occupancy (for example, CTCF) may lose DNA binding upon cytosine methylation.<sup>41</sup> Thus, the GC-rich p-CPR motifs are not preferred in transformed cells. The AT-richness of the t-CPR motifs could either be an alternate binding site for the histone occupancy regulating motif-specific DNA-binding protein or a simple consequence of unavailability of the GC-rich sequences in transformed cells.

While the cause of the motif-type preference switch in transformed cells is an area for further investigation, our results show that it has a functional consequence. Transcription start sites are some of the best characterized functional landmarks in human genome and so we chose to study the functional effects of a motif preference shift with respect to the TSSs. By analyzing motifs with respect to those TSSs that are associated in cis with both p-CPR and t-CPR motifs, we found a surprising consistency in the location of t-CPR motifs distal from the TSSs compared with the p-CPR upstream in the direction of transcription. The effects of shift in histone localization could directly affect transcriptional activity and even alternative start site usage. The features we have observed to be differentially associated with p-CPR and t-CPR motifs (namely, different GC and AT-contents) are also important determinants of histone occupancy.<sup>47</sup> These positioned nucleosomes limit divergent transcription<sup>47–50</sup> and affect RNA Polymerase II occupancy and activity.<sup>51</sup> The differences in RNA-seq read pattern differences we observe between primary and transformed cells could thus be a result of differential histone occupancy.

The combined effect of p-/t-CPR motif shift at any TSS will be a combination of the profiles of the co-occurring histone modifications in a particular CPR and other factors governing the transcription at that locus. For example, motif shift for CPR of H2AFZ, a component of positioned nucleosomes at TSSs, could be more informative about any wobble in the TSS usage than that for other histone modifications, which are not components of positioned nucleosomes or are prevalent in gene bodies. Thus, our analyses of CPR-TSSs and non-CPR-TSSs specifically for H2AFZ CPR-TSSs have been instrumental in uncovering transcriptional disturbances in TSS flanks. Our results obtained from analyses of disparate sets of RNA-seq data show that H2AFZ occupancy at the flanks of CPR-TSSs is susceptible to disruption in cellular transformation leading to aberrant transcription. Transcription start sites that undergo a p-CPR motif to t-CPR motif shift of H2AFZ occupancy in transformed cells maintain TSS flank transcription normally. However, the CPR-TSSs that do not undergo p-CPR to t-CPR motif shift exhibit aberrant upstream as well as downstream transcription. The RNA-seq read prevalence pattern in the flanks of p- or t-CPR-TSSs is correlated with the POLR2A occupancy. However, whereas the non-CPR-TSSs, which did not exhibit the motif shift, the POLR2A occupancy and RNA-seq read abundance were uncorrelated. Collectively, the findings suggest that these TSSs are prone to divergent transcription between  $-1$  and  $+1$  nucleosomes,<sup>48</sup> and the antisense transcription toward  $-1$  nucleosome is what we observe in our results. Interestingly, it seems that p-CPR to t-CPR motif transition is a mechanism that helps transformed cells to maintain same transcription pattern at the CPR-TSSs as the primary cells. The functional annotation of H2AFZ CPR-TSSs (as the target set) revealed specific enrichment of genes involved in RNA Polymerase II regulatory region sequence-specific DNA binding (GO:0000977) as compared with H2AFZ non-CPR-TSSs (as

the background set) using Gene Ontology enRIchment anaLy-sis and visualizAtion (GORilla) tool<sup>29</sup> (not shown).

The motif shift, although identified for 8 histone modifications here, was analyzed for functional consequence only for H2AFZ and restricted only to TSSs and transcription. ChIP-seq Peak Regions of other histone modifications, occurring at non-TSS regions, could have functional consequences beyond transcription. In addition, it will be informative to analyze what kinds of sequence-specific DNA-binding factors could interact with CPR, and thus, what roles do CPR play in chromatin organization and cellular fate in differentiation and transformation.

## Conclusions

Our study reveals that the t-CPR motifs and p-CPR motifs function differently in normal and transformed cells with an impact on histone occupancy and transcription. This is an interesting epigenetic difference between normal and cancer cells not reported till date. Out of 8 histone modifications, we restricted the functional analysis of CPR only to H2AFZ. Our work sets the platform for future investigations as to why histone modifications have different CPR preferences in primary and transformed cells. The robustness of our findings lie in (1) the number and diversity of the panel of cell types chosen for this analysis and (2) that the patterns identified by us hold true across different sequencing and data-analysis platforms. Also, the CPR identified from ChIP-seq data at the level of DNA correlate well with unrelated datasets at the level of RNA-seq. Our work sets a precedence for analyzing and comparing seemingly unrelated datasets to identify patterns that cannot be otherwise propounded by a hypothesis-driven approach only. This work is thus of pertinent relevance to analytics of next-generation datasets, to fellow researchers in the field of epigenetics as well as cancer biology. An obvious line of work to emanate from these data is to identify the epigenetic functional nature of the CPR, something that relies upon experiments and is hence beyond the scope of the current piece of work.

## Author Contributions


SD, MP, DP, and US performed analyses and wrote the manuscript. US supervised the work. SD and MP contributed equally to this work. All authors read and approved the final manuscript.

## Supplemental Material

Supplemental material for this article is available online.

## ORCID iDs

Subhamoy Datta  <https://orcid.org/0000-0002-9573-1070>

Umashankar Singh  <https://orcid.org/0000-0001-8578-8201>

## REFERENCES

- Jiang C, Pugh BF. Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet.* 2009;10:161–172.
- Lai WKM, Pugh BF. Understanding nucleosome dynamics and their links to gene expression and DNA replication. *Nat Rev Mol Cell Biol.* 2017;18:548–562.
- Mueller-Planitz F, Klinker H, Becker PB. Nucleosome sliding mechanisms: new twists in a looped history. *Nat Struct Mol Biol.* 2013;20:1026–1032.
- Radman-Livaja M, Rando OJ. Nucleosome positioning: how is it established, and why does it matter? *Dev Biol.* 2010;339:258–266.
- Struhl K, Segal E. Determinants of nucleosome positioning. *Nat Struct Mol Biol.* 2013;20:267–273.
- Chereji RV, Ocampo J, Burke T, et al. Major determinants of nucleosome organization. *Biophys J.* 2016;110:68a.
- Clark DJ. Nucleosome positioning, nucleosome spacing and the nucleosome code. *J Biomol Struct Dyn.* 2010;27:781–793.
- Segal E, Widom J. Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr Opin Struct Biol.* 2009;19:65–71.
- Anselmi C, Bocchinfuso G, De Santis P, Savino M, Scipioni A. Dual role of DNA intrinsic curvature and flexibility in determining nucleosome stability. *J Mol Biol.* 1999;286:1293–1301.
- Cuddapah S, Jothi R, Schones DE, Roh TY, Cui K, Zhao K. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res.* 2009;19:24–32.
- Fu Y, Sinha M, Peterson CL, Weng Z. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet.* 2008;4:e1000138.
- Nie Y, Liu H, Sun X. The patterns of histone modifications in the vicinity of transcription factor binding sites in human lymphoblastoid cell lines. *PLoS ONE.* 2013;8:e60002.
- Benveniste D, Sonntag H-J, Sanguinetti G, Sproul D. Transcription factor binding predicts histone modifications in human cell lines. *Proc Natl Acad Sci U S A.* 2014;111:13367–13372.
- Hayes JJ, Wolffe AP. The interaction of transcription factors with nucleosomal DNA. *Bioessays.* 1992;14:597–603.
- Mutskov V, Gerber D, Angelov D, Ausio J, Workman J, Dimitrov S. Persistent interactions of core histone tails with nucleosomal DNA following acetylation and transcription factor binding. *Mol Cell Biol.* 1998;18:6293–6304.
- Nature ENCODE: Nature Publishing Group. RNA and chromatin modification patterns around promoters. Website. <http://www.nature.com/encode/threads/rna-and-chromatin-modification-patterns-around-promoters>. Accessed January 18, 2019.
- West JA, Cook A, Alver BH, et al. Nucleosomal occupancy changes locally over key regulatory regions during cell differentiation and reprogramming. *Nat Commun.* 2014;5:4719.
- Chen T, Dent SYR. Chromatin modifiers and remodellers: regulators of cellular differentiation. *Nat Rev Genet.* 2014;15:93–106.
- Segal E, Fondufe-Mittendorf Y, Chen L, et al. A genomic code for nucleosome positioning. *Nature.* 2006;442:772–778.
- Deaton AM, Bird A. CpG islands and the regulation of transcription. *Genes Dev.* 2011;25:1010–1022.
- Nakato R, Shirahige K. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. *Brief Bioinform.* 2017;18:279–290.
- Computational methodology for ChIP-seq analysis. Website. <https://www.ncbi.nlm.nih.gov/pubmed/25741452>. Accessed January 18, 2019.
- Nature ENCODE: Nature Publishing Group. Chromatin patterns at transcription factor binding sites. Website. <http://www.nature.com/encode/threads/chromatin-patterns-at-transcription-factor-binding-sites>. Accessed January 18, 2019.
- Huda A, Mariño-Ramírez L, Landsman D, Jordan IK. Repetitive DNA elements, nucleosome binding and human gene expression. *Gene.* 2009;436:12–22.
- Li W, Sosa D, Jose MV. Human repetitive sequence densities are mostly negatively correlated with R/Y-based nucleosome-positioning motifs and positively correlated with W/S-based motifs. *Genomics.* 2013;101:125–133.
- Farman FU, Iqbal M, Azam M, Saeed M. Nucleosomes positioning around transcriptional start site of tumor suppressor (Rb1/p130) gene in breast cancer. *Mol Biol Rep.* 2018;45:185–194.
- Cohen I, Poreba E, Kamieniarz K, et al. Histone modifiers in cancer: friends or foes? *Genes Cancer.* 2011;2:631–647.
- Sharma S, Kelly TK, Jones PA. Epigenetics in cancer. *Carcinogenesis.* 2009;31:27–36.
- Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GORilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics.* 2009;10:48.
- Bailey TL, Boden M, Buske FA, et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 2009;37:W202–208.
- Mahony S, Benos PV. STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.* 2007;35:W253–W258.
- Ramírez F, Dündar F, Diehl S, Gruning BA, Manke T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.* 2014;42:W187–W191.
- Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics.* 2004. Website. <https://www.ncbi.nlm.nih.gov/pubmed/18428725>.
- Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010;26:841–842.



35. Lin JC, Jeong S, Liang G, et al. Role of nucleosomal occupancy in the epigenetic silencing of the MLH1 CpG island. *Cancer Cell*. 2007;12:432–444.
36. Love MI, Huska MR, Jurk M, et al. Role of the chromatin landscape and sequence in determining cell type-specific genomic glucocorticoid receptor binding and gene regulation. *Nucleic Acids Res*. 2017;45:1805–1819.
37. Kanaji N, Yokohira M, Nakano-Narusawa Y, et al. Hepatocyte growth factor produced in lung fibroblasts enhances non-small cell lung cancer cell survival and tumor progression. *Respir Res*. 2017;18:118.
38. Hodson DJ, Shaffer AL, Xiao W, et al. Regulation of normal B-cell differentiation and malignant B-cell survival by OCT2. *Proc Natl Acad Sci U S A*. 2016;113:E2039–E2046.
39. Caramuta S, Lee L, Ozata DM, et al. Role of microRNAs and microRNA machinery in the pathogenesis of diffuse large B-cell lymphoma. *Blood Cancer J*. 2013;3:e152.
40. Jiang Y, Soong TD, Wang L, Melnick AM, Elemento O. Genome-wide detection of genes targeted by non-Ig somatic hypermutation in lymphoma. *PLoS ONE*. 2012;7:e40332.
41. Weth O, Paprotka C, Günther K, et al. CTCF induces histone variant incorporation, erases the H3K27me3 histone mark and opens chromatin. *Nucleic Acids Res*. 2014;42:11941–11951.
42. Splinter E, Heath H, Kooren J, et al. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes Dev*. 2006;20:2349–2354.
43. Zlatanova J, Caiafa P. CTCF and its protein partners: divide and rule? *J Cell Sci*. 2009;122:1275–1284.
44. Aoyama T, Okamoto T, Fukiage K, et al. Histone modifiers, YY1 and p300, regulate the expression of cartilage-specific gene, chondromodulin-I, in mesenchymal stem cells. *J Biol Chem*. 2010;285:29842–29850.
45. Rezai-Zadeh N, Zhang X, Namour F, et al. Targeted recruitment of a histone H4-specific methyltransferase by the transcription factor YY1. *Genes Dev*. 2003;17:1019–1029.
46. DNA methylation and cancer. Website. <https://www.ncbi.nlm.nih.gov/pubmed/20920744>. Accessed January 18, 2019.
47. Peckham HE, Thurman RE, Fu Y, et al. Nucleosome positioning signals in genomic DNA. *Genome Res*. 2007;17:1170–1177.
48. Seila AC, Core LJ, Lis JT, Sharp PA. Divergent transcription: a new feature of active promoters. *Cell Cycle*. 2009;8:2557–2564.
49. Rhee HS, Bataille AR, Zhang L, Pugh BF. Subnucleosomal structures and nucleosome asymmetry across a genome. *Cell*. 2014;159:1377–1388.
50. Weber CM, Ramachandran S, Henikoff S. Nucleosomes are context-specific, H2A.Z-modulated barriers to RNA polymerase. *Mol Cell*. 2014;53:819–830.
51. Chen Y, Jørgensen M, Kolde R, et al. Prediction of RNA Polymerase II recruitment, elongation and stalling from histone modification data. *BMC Genomics*. 2011;12:544.