Analysis of H3K4me3 and H3K27me3 and their spatial organization during cell cycle progression in human embryonic stem cells

Lisa Adilijiang, Yating Wang, Zhen Zuo

University of Illinois at Urbana-Champaign, Urbana, Illinois 61801 USA

Abstract

During the eukaryotic cell cycle progression, accurate duplication of the genome as well as the epigenetic program is crucial to maintain cell identity and genome stability. While different level of regulation has been studied during cell cycle progression, few study has addressed whether epigenetic marks, especially the chromatin marks that regulate transcription, are dynamically regulated during cell cycle progression. Here, we investigated the distribution of two important chromatin marks, H3K4me3 and H3K27me3 at different stages of cell cycle and showed that these chromatin marks are dynamically regulated during different cell cycle stages. We also showed that H3K4me3 behave differently at active and repressed promoters. Our study has revealed an interesting cell cycle dynamic in the distribution of chromatin marks and can provide novel directions for subsequent studies.

Key words: chromatin, cell cycle, H3K4me3, H3K27me3

Introduction

In eukaryotic cells, the genetic information is stored in the DNA sequence. However, DNA sequence alone is not sufficient to control gene expression. The linear DNA molecules is wrapped around histone octamer and the complex of histone octamer with DNA wrapped outside is called nucleosome. Nucleosomes can be further packed into higher order chromatin structures. The chromatin structure as well as modifications on the nucleosome is inheritable and play important roles in regulating transcription.

Histone proteins are subjected to various post-translational modifications including methylation of lysine and arginine, acetylation of lysine, ubiquitination of lysine, etc. Different histone modifications function differently to control various cellular process. For example, H3K4me3 and H3 acetylation usually mark active transcription sites while H3K9me3, H4K20me3 and H3K27me3 usually correlate with repressed transcription (Allis and Jenuwein 2016). Increasing studies have suggested that multiple histone marks can coordinate to control local transcription. For example, the existence of H3K4me3 will block DNA methylation (Cedar and Bergman 2009). At heterochromatic region, multiple repressive marks such as H3K9me3 and CpG methylation often co-exit to maintain the repressive chromatin environment. It has also been shown by multiple studies that in embryonic stem (ES) cells both H3K4me3 and H3K27me3 are present at the promoters of differentiation genes (bivalent promoters) (Voigt et al. 2013). While these genes remain silenced ES cells, the bivalent marks make them "poised" to activation upon induction of differentiation. Thus, it is necessary to analyze the pattern of multiple chromatin marks together to understand how chromatin environment regulates transcription, as well as other cellular process.

During the eukaryotic cell cycle, the entire genome need to be accurately duplicated. The duplication of DNA happens during S phase while cell division happen during mitosis. During these processes, different cellular events, including DNA replication, chromosome assembly and transcription need to be highly coordinated. During the cell cycle progression process, whether and how different chromatin marks change remain unclear. It is clear that transcription is dynamic during different stages of cell cycle. During mitosis, the transcription machinery dissociates from the chromatin resulting in repressed transcription globally (Gottesfeld and Forbes 1997; Prasanth et al. 2003). Recent study suggests that there is hyperactive transcription during the mitosis-G1 transition (Hsiung et al. 2016). However, whether chromatin marks also show cell cycle regulated pattern remain unclear. In addition, during S phase, the entire genome will duplicate. While the

DNA replication machinery moves along the DNA molecules, nucleosomes will disassemble in front of the replication fork and re-assemble behind the replication fork on both daughter strands. Thus, epigenetic marks also need to be re-established on the newly assembled nucleosomes. Recent nascent chromatin capture proteomic study suggests that while some histone enzymes exist at newly replicated nucleosomes, some chromatin enzymes exist only at nucleosomes far behind the replication fork (Alabert et al. 2014). Thus, how fast can different chromatin marks be re-established behind the replication fork is also an intriguing question.

To understand whether and how different chromatin marks are dynamically regulated during different stages of cell cycle, we took advantage of a recent study where H3K4me3 and H3K27me3 were mapped at four different stages of cell cycle in human ES cells (Singh et al. 2015). Singh et al used a fluorescence based cell soring system to isolate cells at different stages of cell cycle and performed H3K4me3 and H3K27me3 ChIP-seq in these sorted cells. Using the available data, we performed clustering analysis of genes based on different chromatin marks and analyzed the distribution of H3K4me3 and H3K27me3 in different gene clusters at different stages of cell cycle. We also tried a hidden Markov modeling analysis based on the two marks of chromosome 19 to identify different epigenetic domains. Our study has unveiled an interesting dynamic pattern of chromatin marks during cell cycle progression.

Results

Data processing

In a recent study by Singh et al, the authors used a florescence based cell sorting system (FUCCI) to isolate human embryonic stem (ES) cells at different stages of cell cycle and performed ChIP-seq of different chromatin marks at different stages of cell cycle (Singh et al. 2015). The FUCCI system takes advantage of cell cycle regulated ubiquitin degradation system and integrate fusion proteins of two types of florescent proteins with either Cdt1 or Geminin degron (Sakaue-Sawano et al. 2008). Thus the florescent proteins will be degraded at different stages of cell cycle and cells at different stages can be easily separated based on their color. In this study, cells at early-G1, late-G1, S phase and G2/M phase were isolated and ChIP-seq of different chromatin marks were performed (Fig 1A). To understand whether and how different chromatin marks change across different cell cycle stages, we choose to analyze two of the most well studied histone modifications,

H3K4me3 and H3K27me3. H3K4me3 usually localizes to active promoters while H3K27me3 marks repressed promoters. Since this two marks are enriched at transcription start site (TSS), we analyzed the scores after peak calling near the TSS of each gene. Using the scores of the peaks after peak calling, we calculated the sum of peaks scores within 2kb of the TSS of each gene. Thus, each gene is assigned a score of H3K4me3 and H3K27me3 (Fig 1B). 26067 genes were assigned non-zero score for either of the two marks or either of the four cell cycle stages and these scores were used in the subsequent analysis.

Clustering analysis

Based on the assigned H3K4me3 and H3K27me3 scores of each gene at different cell cycle stages, we reached an eight-dimensional matrix for each gene. We then performed clustering analysis based on different patterns of the marks. The number of clusters was determined by the gap statistic (Fig 2A) as well as the "elbow method" (Fig 2B) and k=3 was chosen for clustering. The summary statistics of the three clusters were shown in Fig 3A. Cluster 1 represents a large group of genes with low level of both H3K4me3 and H3K27me3. Cluster 2 represents a group of genes with high level of H3K4me3 and low level of H3K27me3, suggesting that they are potentially the actively transcribed genes. Cluster 3 represents genes with low H3K4me3 and high H3K27me3 and hence are potentially repressed in human ES cells. To validate our clustering analysis, we performed gene ontology analysis of genes in Cluster 3 to address their biological function. The top DAVID clusters include genes involved in embryonic development including neurogenesis, mophogenesis and Homobox genes, etc (Fig 3B). These genes are important regulators during the differentiation process and are usually repressed in undifferentiated ES cells. Thus, our clustering analysis has successfully identified gene clusters with different chromatin signature.

While clustering different cell cycle stages of each chromatin marks, it is clear that in the case of both H3K4me3 and H3K27me3, the signal is most different during early-G1 compared with all other cell cycle stages (Fig 3C). In the case of H3K4me3, late-G1 and S phase clustered together whereas in the case of H3K27me3, S phase and G2/M phase clustered together. These results suggest different chromatin marks may have different dynamics during cell cycle progression.

Analysis of different chromatin marks across different cell cycle stages

To better understand how different chromatin marks changed during different cell cycle stages, we performed ANOVA analysis of each mark at the four cell cycle stages. Our analysis suggest that H3K4me3 showed significantly higher at early G1 compared with the other three cell cycle stages (Fig 4A). This result is consistent with previous report that H3K4me3 is transiently increased in the repressed genes in human ES cells (Singh et al. 2015). However, different from previous study, our analysis revealed increased H3K4me3 at early-G1 globally instead of at a subset of repressed genes. This is also consistent with previous report of a "transcription spike" during mitosis-G1 transition (Hsiung et al. 2016). To rule out the possibility that the increased H3K4me3 at early G1 is a dominant effect of a subset of genes. We performed ANOVA analysis of genes in Cluster 2 and Cluster 3, since these two clusters are regulated by H3K4me3 and H3K27me3. Interestingly, ANOVA analysis of genes in Cluster 2 resembles that of the total population (Fig 4B). However, the result of Cluster 3 suggest that H3K4me3 signal is higher during early-G1 as well as G2/M population (Fig 4C). This result may suggest that the increased H3K4me3 at the repressed genes may start as early as mitosis.

Our analysis of H3K27me3 also suggest increased signal at G1 phase (Fig 4D). However, unlike H3K4me3, H3K27me3 is high at early-G1 as well as late-G1. Similar analysis of genes in Cluster 2 and 3 also suggests higher H3K27me3 signal in early-G1 as well as late-G1 (Fig 4E, 4F). These results suggest that H3K4me3 and H3K27me3 both show dynamic pattern during cell cycle progression and yet they show slightly different patterns.

Hidden Markov modeling analysis of H3K4me3 and H3K27me3 on chromosome 19

To gain better understanding of the interactions between different histone marks, we decided to perform hidden Markov modeling (HMM) analysis of different chromatin marks. We performed our analysis on chromosome 19, since chr19 has the highest gene density.

To start with, we only used data of early-G1 stage. The same clustering number $K = 3$ from the previous analysis was used for our number of HMM states. HMM was performed via Java package HMMSeg, with the assumption that the distribution of different histone marks within each HMM state primarily follow a multivariate Gaussian distribution (Day et al. 2007). HMMSeg employs the expectation-maximization (EM) algorithm for estimating model parameters and the Viterbi algorithm for model selection. The Viterbi algorithm has been proven to be more computationally

efficient and essentially as accurate when compared with Forward-Backward model selection algorithms on chromosome 19 (Thurman et al. 2007; Larson and Yuan 2010).

After the HMM segmentation, we found 12, 236, and 248 domains in total within HMM state 0, state 1, and state 2, respectively in early-G1 date set (Table 1). From the emission results produced from HMMSeg log files, we gained the estimated mean and variance of Gaussian distribution of each histone marks within every HMM state. We found that state 0 correlates with low H3K4me3 and high H3K27me3 level, while state 2 correlates with high H3K4me3 and low H3K27me3 level. We thus define state 0,1 and 2 as "repressed", "null" and "active" respectively. The total size of each assigned state is 11.351Mb, 25.003Mb and 22.698MB, respectively. The maximum size was of 42 genes per domain. The average domain size among overall three states was 2.72 genes per domain.

We then perform the same analysis using data from the other three cell cycle stages. The domain size statistics of the analysis at different cell cycle stages were summarized in Table 1 and Table 2 and scheme of the three-stage domains were shown in Fig 6. Notably, while the average domain size did not change dramatically at different cell cycle stages, the proportion of different states changes during cell cycle progression. While the size of "active" domain remains similar across different cell cycle stages, late-G1 stage is marked by increased "repressed" domain and decreased "null" domain (Table 1, Fig 6A, 6B). Thus, it is likely that the spatial distribution of chromatin marks is also dynamically regulated during cell cycle progression.

Discussion

The establishment of chromatin immunoprecipitation (ChIP) method has allowed the investigation of various protein-DNA interactions. With the recent advancement in deep sequencing technology, numerous chromatin marks and chromatin-bound proteins has been mapped by ChIP-seq method. When researchers map chromatin marks, few of them consider the effect of cell cycle stages and most people perform ChIP-seq in asynchronous cell population. However, emerging evidence recently suggest that a lot of chromatin marks are indeed cell cycle regulated. For example, studies suggest that H4K20me is absent during S phase of the cell cycle (Abbas et al. 2010) and the level of H3K79me2 has also shown to be cell cycle regulated (Kim et al. 2012; Fu et al. 2013). In addition, during S phase, as the DNA being replicated, new histone proteins will also be deposited

to newly replicated DNA. All the chromatin marks need to be faithfully copied on the daughter chromosome. Thus, how fast the chromatin marks can be established on newly assembled chromosome will also affect the cell cycle distribution of chromatin marks. Mass spectrometry study suggests that H3K9me3 and H3K27me3 are reduced during S phase and gradually increase as cell progress through the cell cycle (Xu et al. 2011). Thus, it is of great importance to investigate the distribution of chromatin marks in cells synchronized at specific cell cycle stages.

In a study by Singh et al, the researchers mapped the distribution of multiple histone marks at different cell cycle stages in human ES cells (Singh et al. 2015). In this study, the authors found that the level of H3K4me3 at the promoter of key developmental genes are transiently elevated during G1 phase of the cell cycle, which may explain the fact the human ES cells at G1 phase respond to differentiation signal better than cells at other stages of the cell cycle. However, the authors did not discuss in this manuscript whether increased H3K4me3 during G1 phase is a general phenomenon or is only specific to those developmental genes. Also, it is not clear from this study whether other chromatin marks behave similarly to H3K4me3. To address whether and how different chromatin marks are cell cycle regulated, we investigate the distribution of H3K4me3 and H3K27me3 at four different cell cycle stages. Our analysis suggests that both H3K4me3 and H3K27me3 show elevated level during G1 phase. In the case of H3K4me3, H3K4me3 low regions behave slightly different form H3K4me3 high regions. In addition, we also tried to perform clustering and hidden Markov modeling analysis to investigate the change in epigenetic domains instead of focusing on individual chromatin marks.

Our study described here has revealed an interesting phenomenon that the level of both H3K4me3 and H3K27me3 increases during G1 phase of the cell cycle. Singh et al has reported increased H3K4me3 level at promoters of developmental genes during G1 human ES cells. Here, we show that increased H3K4me3 during G1 phase happen at both active and repressed promoters. In addition, our analysis suggests that similar to H3K4me3, the level of H3K27me3 is also elevated during G1 phase. Increased H3K4me3 level during G1 phase is consistent with recent report about "transcription spike" during mitosis-G1 transition (Hsiung et al. 2016). During mitosis, chromatin becomes condensed and transcription machinery will dissociate from the chromatin (Gottesfeld and Forbes 1997; Prasanth et al. 2003). Thus, as cells exit mitosis and enter the next G1 phase, cells need to restore the transcription program. Previous study also reported increased Pol II binding at enhancers during mitosis-G1 transition (Hsiung et al. 2016). Thus, increased H3K4me3

might be correlated with the restoration of transcription program as cell exit from mitosis. Interestingly, when we analyze H3K4me3-high and H3K3me3-low genes separately, we found that H3K4me3-low genes show increased H3K4me3 level in G2/M populations. Since it is hard to differentiate whether this increased happen at G2 phase or mitosis, this observation could be interpreted in different ways. If increased H3K4me3 happen during G2 phase, this can be explained by the re-establishment of H3K4me3 marks after DNA replication. If increased H3K4me3 happens at mitosis, it may suggest that increased H3K4me3 at repressed promoters might happen as early as late mitosis.

Similar to H3K4me3, our analysis suggest that H3K27me3 level is also elevated during G1 phase. H3K27me3 level is high at both early- and late-G1 phase, ie. H3K27me3 level is lower from S phase through mitosis. Previous study suggests that the H3K27me3 writer, PcG, and H3K27me3 level increases at polycomb response elements (PRE) prior to S phase and this histone mark will be diluted during DNA replication (Lanzuolo et al. 2011). Our results here seem to corroborate with this scenario. Recent nascent chromatin capture study suggests that H3K27me3 writer proteins are enriched at mature chromatin instead of on nascent chromatin, suggesting the re-establishment of H3K27me3 may not happen at replication fork during the DNA replication process (Alabert et al. 2014). Our results here fall in line with findings from these previous reports.

To gain better understanding of how different chromatin marks coordinate, we performed HMM analysis to investigate the change in epigenetic domains and their spatial organization, rather than individual chromatin marks, across different cell cycle stages. Our model has allowed us to partition chr 19 into three states: “repressed”, “null” and “active”. Interestingly, our result suggests that epigenetic domain is also dynamically regulated during cell cycle progression. We found that late-G1 phase is marked by increased “repressed” domain and decreased “null” domain, while during S phase, some large “active” domains disappear (Fig 6A). This result may suggest that during late-G1 stage, some of the “null” domains are converted to “repressed” domain. If this scenario is true, it supports the “dilution model” that has been proposed before, that H3K27me3 transiently increased at some heterochromatic regions before the start of DNA replication and will be diluted at the late S phase when these heterochromatic regions replicate (Lanzuolo et al. 2011). Thus, the repressive state is maintained through this “increase-dilution” mechanism in stead of adding new H3K27me3 to newly replicated chromosome. However, whether this is a general phenomenon will require a genome-wide analysis.

In the study of Singh et al, the authors isolated cells at different cell cycle stages using the FUCCI system. This system is based on the degradation of Cdt1 and Geminin at different cell cycle stages (Sakaue-Sawano et al. 2008). While the authors have isolated cells based on different fluorescence, the classification of each cell cycle stages can be discussed. The authors defined cells that has both Cdt1 and Geminin signal as "S phase". However, this may actually represent cells that are at G1-S transition, since downregulation of Cdt1 and upregulation of Geminin happen at G1-S transition. Similarly, cells with high Geminin signal and low Cdt1 signal may actually represent cells in S phase instead of G2/M and cells with low signal of both Geminin and Cdt1 may suggest that the cells are at late mitosis instead of early-G1. Thus, the result can be interpreted in a slightly different way if we classify the cell cycle stages differently.

Both H3K4me3 and H3K27me3 regulate transcription. Our study here suggests dynamic pattern of both marks as well as their spatial distribution during cell cycle progression. It would be interesting to investigate the change in transcriptome at different cell cycle stages to correlate with the H3K4me3 and H3K27me3 results discussed here. It would also be interesting to investigate the genome-wide change in epigenetic domains based on more chromatin marks as well as chromatin conformation information. In addition, the study here was performed in human ES cells. ES cells has different cell cycle program compared with somatic cells, marked by short G1 phase, lengthened S phase and absence of cell cycle checkpoints. It is also important to investigates whether chromatin marks are cell cycle regulated in somatic cells and cancerous cells.

Methods

Processing of raw data

Raw data (accession number GSE61176) were obtained from GEO (Gene Expression Omnibus) database that were public online on Jun 2015. The research paper performed chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) method for four types of histone modifications in human embryonic stem cells at four cell cycle stages: AZH (G2/M phase), AZL (S phase), DN (early G1 phase), KO2 (late G1 phase) (Singh et al. 2015). Among these histone modifications, we focused on the two histone methylations: histone H3 lysine 4 trimethylation (H3K4me3) and histone H3 lysine 27 trimethylation (H3K27me3), which were considered as having more significant roles in transcription regulation and are associated with gene

activation and repression, respectively. We downloaded the eight groups of datasets from ChIP-seq experiments that were initially analyzed by HPeak software package, among which four were for H3K27me3 across four cell cycle stages and the other four were for H3K27me3. These output datasets from HPeak algorithm were in ".txt" formats not the ".bed" formats resulted from MACS peak-calling pipeline. And these txt files also contain distinct contents with bed files. Since we need ".bed" format for following data processes, we edited the contents and adjusted the data format into ".bed" files via R programming.

Genome-wide analysis of H3K4me3 and H3K27me3

Gene-level summary score

The TSS bed file of human reference genome hg19 was download from UCSC, and the repetitive annotations for same genes were deleted. We first identified the ChIP-seq peaks within 2kb region flanking TSS site of each gene. Then we calculated the sum of peak scores in each (TSS-2kb, TSS+2kb) region and assigned it as the final score for each gene. The same process was repeated for all eight preprocessed datasets. In the end, we combined all the eight scores for the same gene into one single matrix. Thereby an eight-dimensional score vector was generated for each gene: four from H3K4me3 ChIP-Seq and four from H3K4me27 ChIP-Seq, each four were scores throughout four cell cycle stages. The matrix containing eight scores for all genes in hg19 was attained all by R programming.

Genome-wide heatmap clustering and GO annotations

After generating the score matrix, we deleted the genes that got zero scores for all eight experiments, resulting in 26067 genes left in the score matrix. Then the unsupervised learning method, *K*-means average agglomeration clustering, was performed to generate a genome-wide heatmap for two histone methylation marks in all four cell cycle stages.

To divide and extract clusters for function enrichment analysis in each cluster, we first decided the optimal number of cluster number *K* by both measuring gap statistics and elbow method. Gap statistic and elbow method are the most widely used ways to estimate and find the appropriate number of clusters in a dataset, which both can use the output resulting from multiple kinds of algorithms, such as *K*-means algorithm and hierarchical algorithm. Here in our gap statistics analysis, the maximum number of *K* to consider was set to 4 clusters, and the number of Monte

Carlo samples (the random bootstraps permutations) was set to 20. By plotting resulting gap quantities against number of clusters *K*, we found the clear peak at the correct *K*=3, which represents the smallest *K* for which the gap statistic formula $Gap(k)-(Gap(k+1)-s_{k+1})$ becomes positive (Tibshirani et al. 2001). And in our elbow method analysis, the maximum number of *K* to consider was set to 15 clusters. By plotting the total within-cluster sum of squares against number of clusters *K*, the angle of the graph, which was defined as the elbow point, appeared at *K*=3 as well (Ketchen and Shook 1996). Hence, these two assessments both indicated that *K*=3 is the optimal number of clusters for further functional analysis. All the processes were performed via R programming. Therefore, we cut the dendrograms of all genes in the above resulting heatmap into three clusters and extracted the gene list in each cluster. The average score vector of all genes within each cluster were calculated and shown in a table of statistics summary, containing eight values for both H3K4me3 and H3K27me3 histone marks in four cell cycle stages. These processes were performed via R programming. Subsequently the genes in Cluster 3, which contains low-level of H3K4me3 and high-level of H3K27me3, were subjected to enrichment significance test via the default algorithms in online DAVID database.

One-way ANOVA test was further performed for H3K4me3 and H3K27me3 respectively, to compare their levels among four cell cycle stages. The ANOVA test was performed both for all the 26076 genes, as well as for genes in Cluster 2 and Cluster 3.

Hidden Markov model (HMM) analysis of chromosome 19

Hidden Markov Model segmentation

Since HMMSeg requires evenly-spaced input file in BED format, we re-processed the HPeak output raw datasets as described in following steps. First we extracted all the peaks on chromosome 19 from raw datasets, then generated a BED file by dividing the region chromosome 19: 69,001 – 59,120,001 into 1,000bp intervals. By summating the scores of all ChIP-Seq peaks on chr19 that were located within each interval, we assigned the score-summation as the chromosome structure-level peak-score for each interval. This process was repeated for both H3K4me3 and H3k27me3 histone marks in all four cell cycle stages. These data pre-processing were performed via R programming.

HMMSeg.jar package was downloaded from website noble.gs.washington.edu/proj/ hmmseg/, which requires Java command-line for implementation but is independent of platforms. Here we

performed HMM segmentation on Ubuntu Linux platform using following command parameters. Above all, based on the previous analysis of gap statistic and elbow method, we determined to perform three-state Hidden Markov Model segmentation. The input data files were in BED formats as described above. MODWT wavelet smoothing with a desired scale of 64000bp was performed prior to HMM segmentation (Fig 5A), as recommended in the previous research (Day et al. 2007). The algorithm for parsing the training data was set as default, which was the Viterbi algorithm. Models were initialized with randomly selected values for model parameters, which were the means and variances of each HMM state. The number of iterations for terminating expectation-maximization (EM) training was set as default, which allowed 100 iterations for each initialized model. And the number of times to re-start the EM training for different random initialized models was set to 10, therefore in this case, the model with the highest total likelihood probability was selected and subsequently reported as results in the end. Finally, the resulting information were stored in assigned log files.

At the very beginning test, we did not assign any value to the model mean and variance for initialization, letting HMMSeg to randomly select model parameters. However, the resulting HMM seems not reasonable in biological sense. The HMMSeg could not recognize the opposite regulation functions of two histone marks, resulting in the simultaneously increasing pattern of both H3K4me3 and H3K27me3 histone methylation levels from HMM state 0 to state2, which was different from the three-state HMM statistic trends in the previous paper. In this paper, the author assigned initializing model parameters that were produced by former *K*-means heatmap clustering (Larson and Yuan 2010). To tell the HMMSeg package the distinct functions of H3K4me3 and H3K27me3 histone marks, we further edited the previously generated chr19-1000bp interval score datasets of H3K27me3, by changing the peak-scores assigned to each interval to a corresponding negative one. After the editing, the same HMMSeg smoothing and segmentation processes were performed as described above.

HMM statistic summaries and domain state determination

For summarizing our HMM results, we calculated the total size of state, total domain numbers, as well as means and variances of domain size within each HMM segments, which represents the number of genes (TSS site) in each HMM domain. And then the means and variances of peak-scores for two histone methylation marks within each HMM state were obtained from resulting

HMMSeg emission probabilities. According to these primary statistical analysis, we got an overview of the resulting three-state HMM summary statistics and the Gaussian distribution parameters of both H3K4me3 and H3K27me3 histone marks within each assigned state, thereby defined the repressed, null, and active state of higher-order chromatin structure to different HMM states (Table 1).

Mapping smoothed peaks and HMM domains across chr19

MODWT smoothed peaks of each histone marks on 64kb scale were plotted across chromosome 19 from the staring positing of chr19: 69,001 to the ending position of chr19: 59,120,001. The smoothed data for H3K27me3 were plotted through the original output smoothing data before the negative value editing, thus the peaks were shown as normal positive peaks (Fig 5). And the assigned peak-score distribution plotting was only shown for early-G1 cell cycle stage as an example (Fig 5A), for comparing the score distribution before and after smoothing. On the other hand, the HMM domain mapping were performed on the datasets generated after negative correction of H3K27me3 followed by HMMSeg analysis. The results were shown for all four cell cycle stages against identical coordinates on chr19, enabling a better visualization and comparison for further analysis (Fig 5B, 5C, 5D).

References

Abbas T, Shibata E, Park J, Jha S, Karnani N, Dutta A. 2010. CRL4(Cdt2) regulates cell proliferation and histone gene expression by targeting PR-Set7/Set8 for degradation. *Mol Cell* **40**: 9-21.

Alabert C, Bukowski-Wills JC, Lee SB, Kustatscher G, Nakamura K, de Lima Alves F, Menard P, Mejlvang J, Rappsilber J, Groth A. 2014. Nascent chromatin capture proteomics determines chromatin dynamics during DNA replication and identifies unknown fork components. *Nat Cell Biol* **16**: 281-293.

Allis CD, Jenuwein T. 2016. The molecular hallmarks of epigenetic control. *Nat Rev Genet* **17**: 487-500.

Cedar H, Bergman Y. 2009. Linking DNA methylation and histone modification: patterns and paradigms. *Nat Rev Genet* **10**: 295-304.

Day N, Hemmaplardh A, Thurman RE, Stamatoyannopoulos JA, Noble WS. 2007. Unsupervised segmentation of continuous genomic data. *Bioinformatics* **23**: 1424-1426.

Fu H, Maunakea AK, Martin MM, Huang L, Zhang Y, Ryan M, Kim R, Lin CM, Zhao K, Aladjem MI. 2013. Methylation of histone H3 on lysine 79 associates with a group of replication origins and helps limit DNA replication once per cell cycle. *PLoS Genet* **9**: e1003542.

Gottesfeld JM, Forbes DJ. 1997. Mitotic repression of the transcriptional machinery. *Trends Biochem Sci* **22**: 197-202.

Hsiung CC, Bartman CR, Huang P, Ginart P, Stonestrom AJ, Keller CA, Face C, Jahn KS, Evans P, Sankaranarayanan L et al. 2016. A hyperactive transcriptional state marks genome reactivation at the mitosis-G1 transition. *Genes Dev* **30**: 1423-1439.

Ketchen DJ, Shook CL. 1996. The application of cluster analysis in strategic management research: An analysis and critique. *Strategic Manage J* **17**: 441-458.

Kim W, Kim R, Park G, Park JW, Kim JE. 2012. Deficiency of H3K79 histone methyltransferase Dot1-like protein (DOT1L) inhibits cell proliferation. *J Biol Chem* **287**: 5588-5599.

Lanzuolo C, Lo Sardo F, Diamantini A, Orlando V. 2011. PcG complexes set the stage for epigenetic inheritance of gene silencing in early S phase before replication. *PLoS Genet* **7**: e1002370.

Larson JL, Yuan GC. 2010. Epigenetic domains found in mouse embryonic stem cells via a hidden Markov model. *BMC Bioinformatics* **11**: 557.

Prasanth KV, Sacco-Bubulya PA, Prasanth SG, Spector DL. 2003. Sequential entry of components of the gene expression machinery into daughter nuclei. *Mol Biol Cell* **14**: 1043-1057.

Sakaue-Sawano A, Kurokawa H, Morimura T, Hanyu A, Hama H, Osawa H, Kashiwagi S, Fukami K, Miyata T, Miyoshi H et al. 2008. Visualizing spatiotemporal dynamics of multicellular cell-cycle progression. *Cell* **132**: 487-498.

Singh AM, Sun Y, Li L, Zhang W, Wu T, Zhao S, Qin Z, Dalton S. 2015. Cell-Cycle Control of Bivalent Epigenetic Domains Regulates the Exit from Pluripotency. *Stem Cell Reports* **5**: 323-336.

Thurman RE, Day N, Noble WS, Stamatoyannopoulos JA. 2007. Identification of higher-order functional domains in the human ENCODE regions. *Genome Res* **17**: 917-927.

Tibshirani R, Walther G, Hastie T. 2001. Estimating the number of clusters in a data set via the gap statistic. *J Roy Stat Soc B* **63**: 411-423.

Voigt P, Tee WW, Reinberg D. 2013. A double take on bivalent promoters. *Genes Dev* **27**: 1318-1338.

Xu M, Wang W, Chen S, Zhu B. 2011. A model for mitotic inheritance of histone lysine methylation. *EMBO Rep* **13**: 60-67.

Figure Legend

Fig 1. ChIP-seq design and data processing. (A) Scheme of FUCCI system for sorting cells at different stages of cell cycle. (B) Scheme of data processing process.

Fig 2. Determination of cluster number by "gap statistic" (2A) and "elbow method" (2B). Note that only data set of early-G1 stage was used for calculation.

Fig 3. Clustering analysis of H3K4me3 and H3K27me3 at different cell cycle stages. (A) Summary of the average H3K4me3 and H3K27me3 score at each cell cycle stage. (B) DAVID analysis of genes in cluster 3. (C) Heatmap of clustering analysis.

Fig 4. Analysis of individual histone mark at different cell cycle stages. (A-C) ANOVA analysis of H3K4me3 at four cell cycle stages for total genes (4A), cluster 2 (4B) and cluster 3 (4C). (D-F) ANOVA analysis of H3K27me3 at four cell cycle stages for total genes (4D), cluster 2 (4E) and cluster 3 (4F).

Fig 5. Smoothed data of each modification at four cell cycle stages. (A) Segmentation and smoothing of early-G1 data set was shown as an example. (B-D) Smoothed data of the other three cell cycle stages.

Fig 6. Three-state HMM model at four cell cycle stages. (A) Mapping smoothed peaks and HMM domains for early-G1 data was shown as an example. (B-D) HMM domains of the other three cell cycle stages.

Fig 1

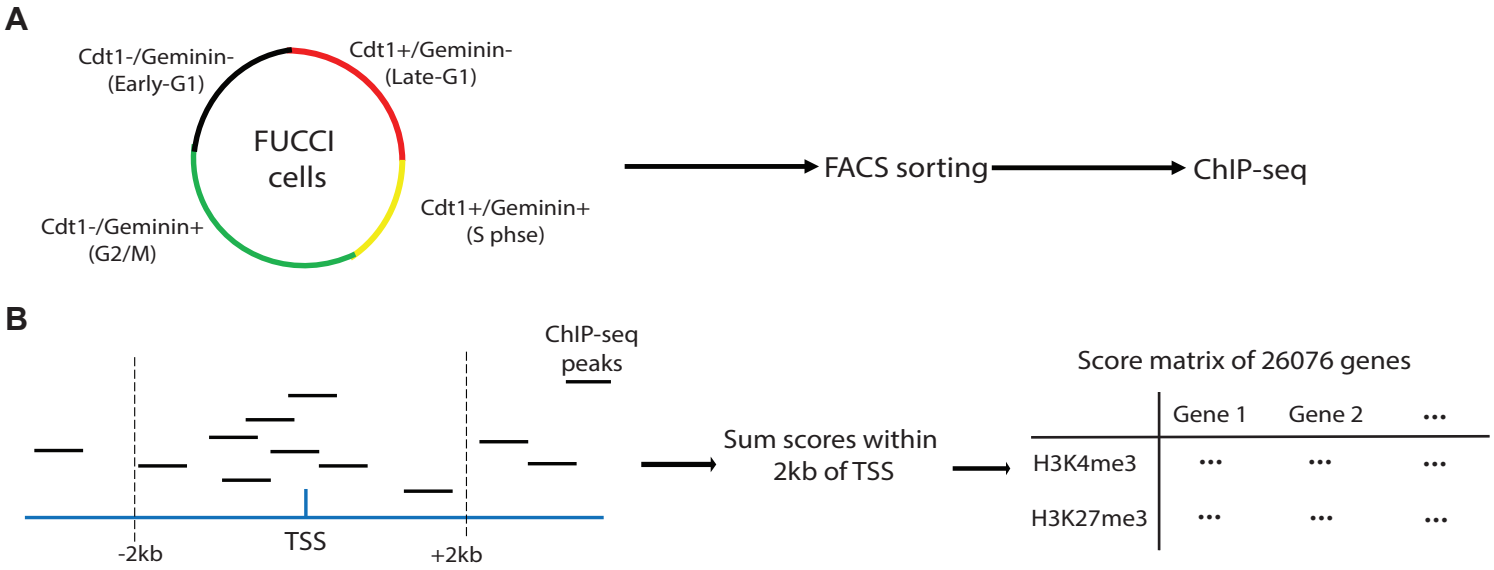

Fig 2

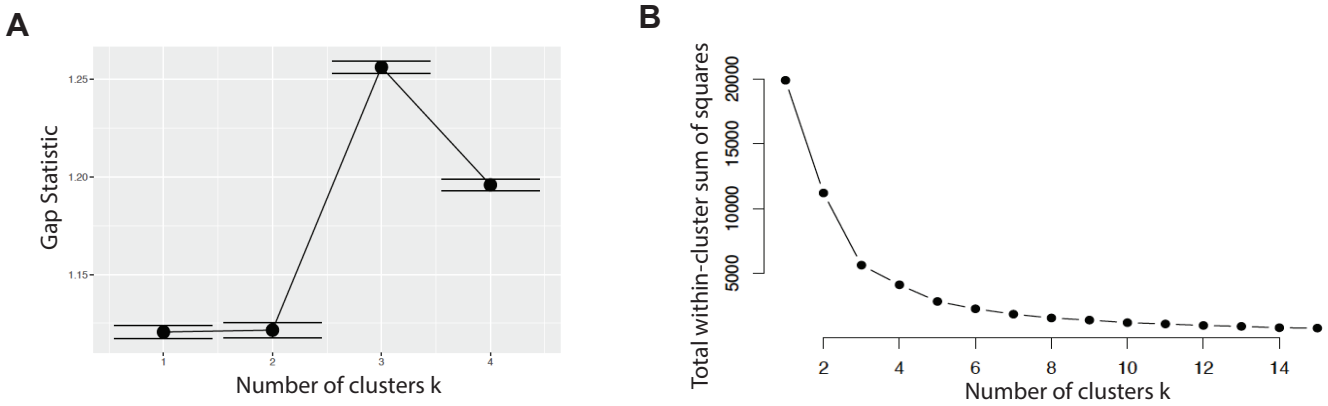

Fig 3

A

	Early-G1 H3K4me3	Late-G1 H3K4me3	S phase H3K4me3	G2/M H3K4me3	Early-G1 H3K27me3	Late-G1 H3K27me3	S phase H3K27me3	G2/M H3K27me3
Cluster 1 (n=9017)	34.63863	16.71324	18.39855	19.84596	6.85496	6.73232	3.562235	4.19643
Cluster 2 (n=13403)	324.4011	159.95852	190.12426	195.94051	4.44991	4.63125	2.13707	2.68194
Cluster 3 (n=3647)	57.84548	48.53753	43.89344	56.58964	209.1457	193.6241	126.9322	126.5912

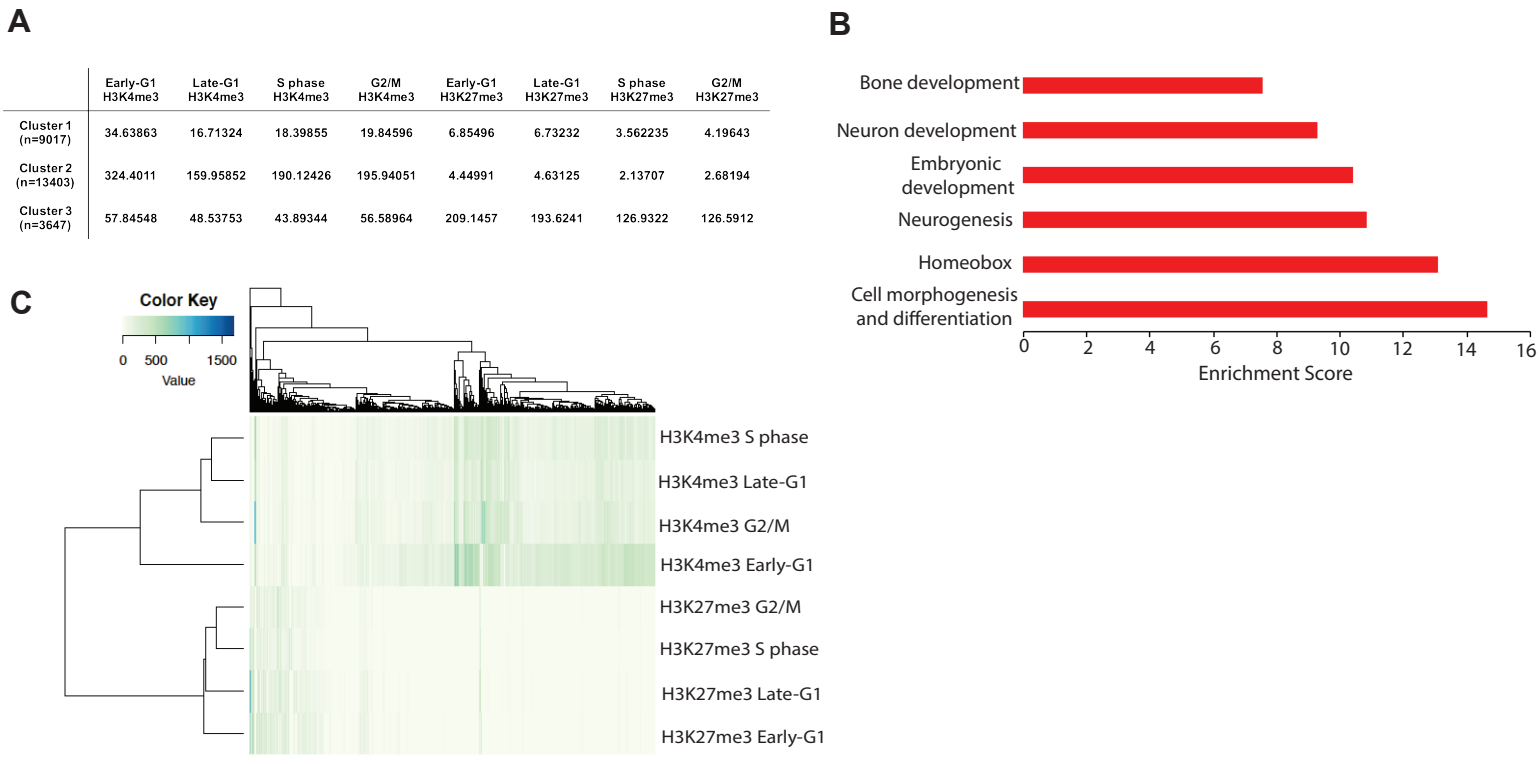

Fig 4

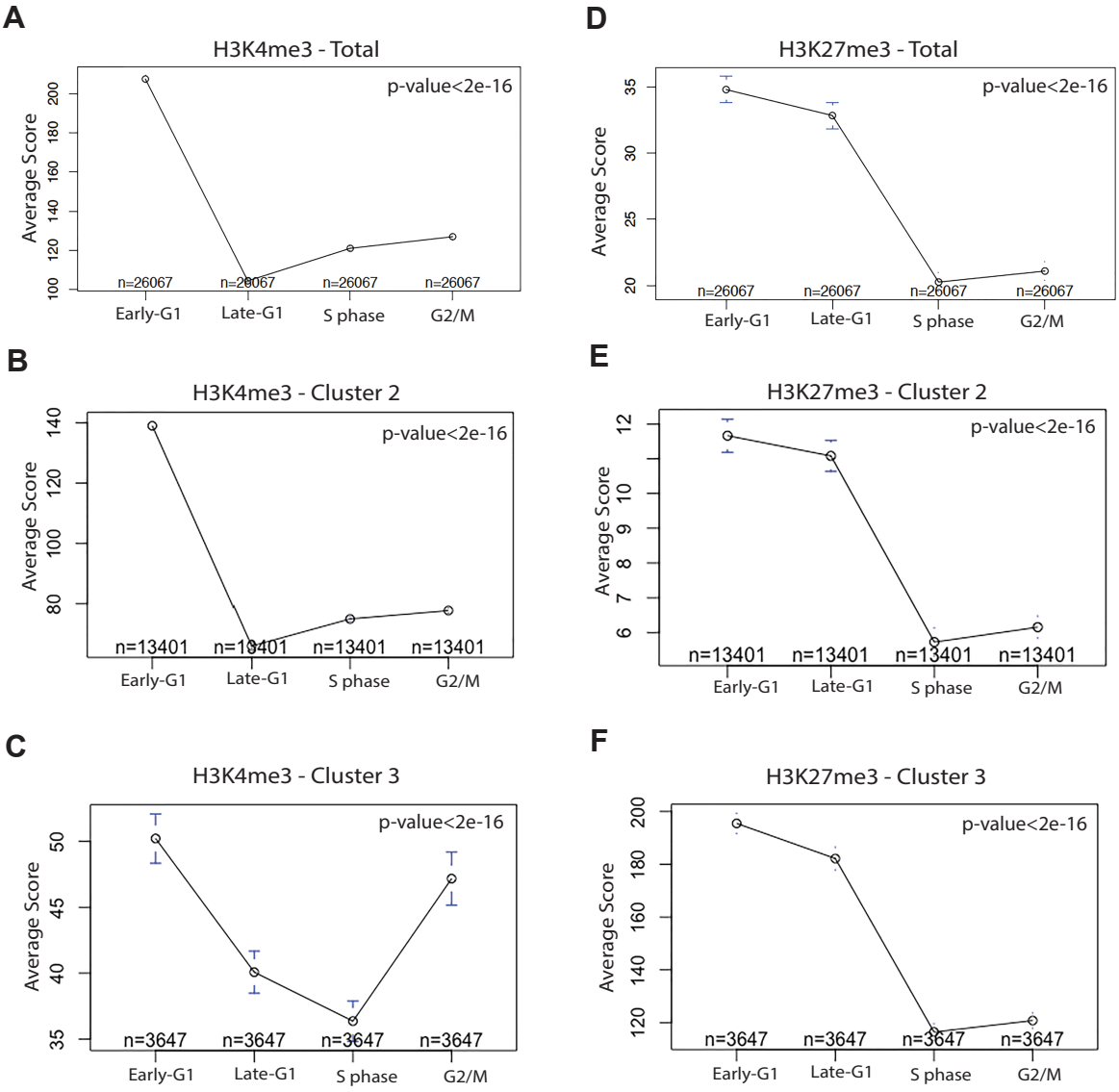
A
H3K4me3 - Total
p-value<2e-16
Average Score
n=26067
Early-G1
Late-G1
S phase
G2/M
D
H3K27me3 - Total
p-value<2e-16
Average Score
n=26067
Early-G1
Late-G1
S phase
G2/M
B
H3K4me3 - Cluster 2
p-value<2e-16
Average Score
n=13401
Early-G1
Late-G1
S phase
G2/M
E
H3K27me3 - Cluster 2
p-value<2e-16
Average Score
n=13401
Early-G1
Late-G1
S phase
G2/M
C
H3K4me3 - Cluster 3
p-value<2e-16
Average Score
n=3647
Early-G1
Late-G1
S phase
G2/M
F
H3K27me3 - Cluster 3
p-value<2e-16
Average Score
n=3647
Early-G1
Late-G1
S phase
G2/M

Fig 5

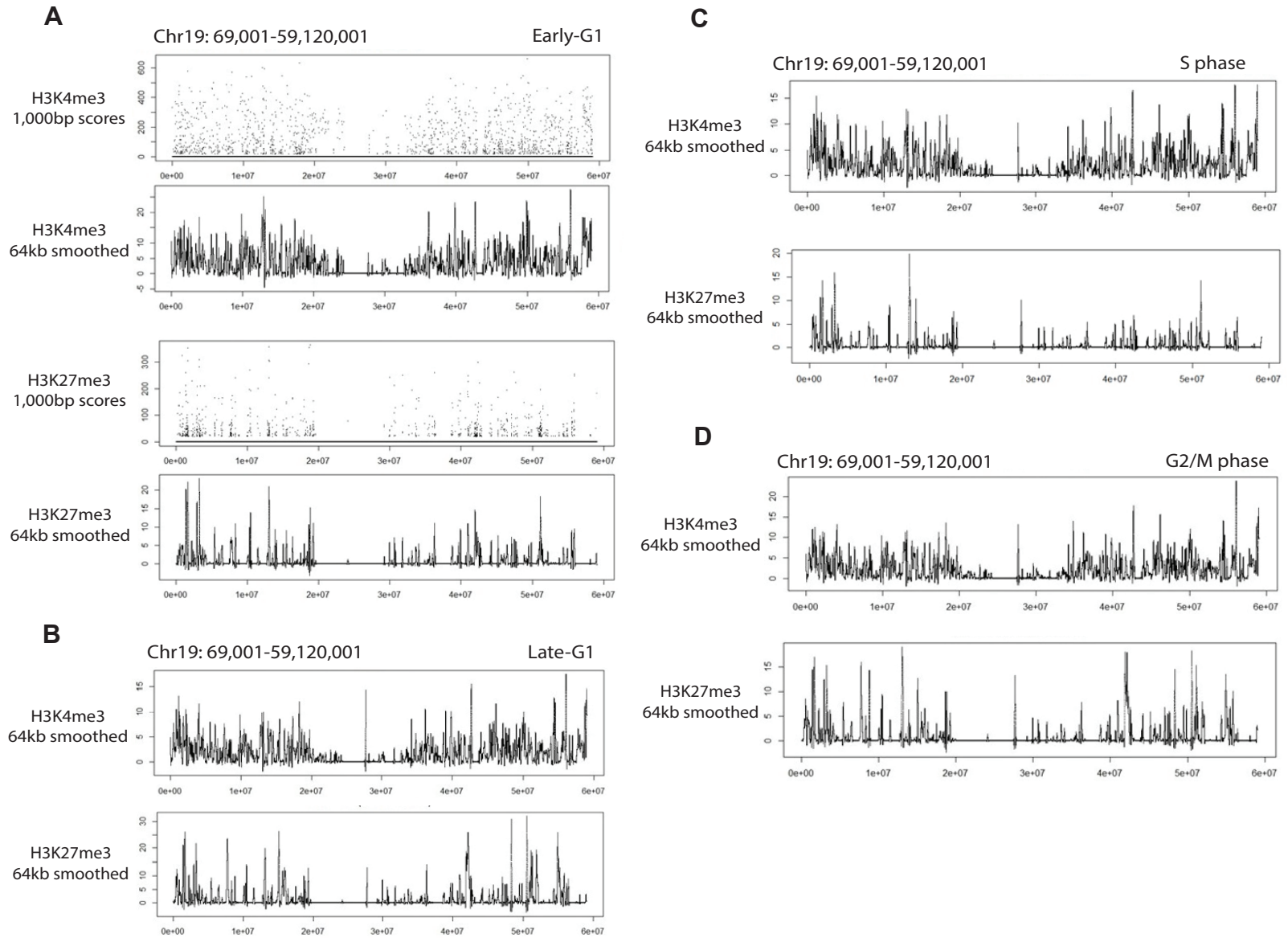
A
Chr19: 69,001-59,120,001
Early-G1
H3K4me3
1,000bp scores
H3K4me3
64kb smoothed
H3K27me3
1,000bp scores
H3K27me3
64kb smoothed
B
Chr19: 69,001-59,120,001
Late-G1
H3K4me3
64kb smoothed
H3K27me3
64kb smoothed
C
Chr19: 69,001-59,120,001
S phase
H3K4me3
64kb smoothed
H3K27me3
64kb smoothed
D
Chr19: 69,001-59,120,001
G2/M phase
H3K4me3
64kb smoothed
H3K27me3
64kb smoothed

Fig 6

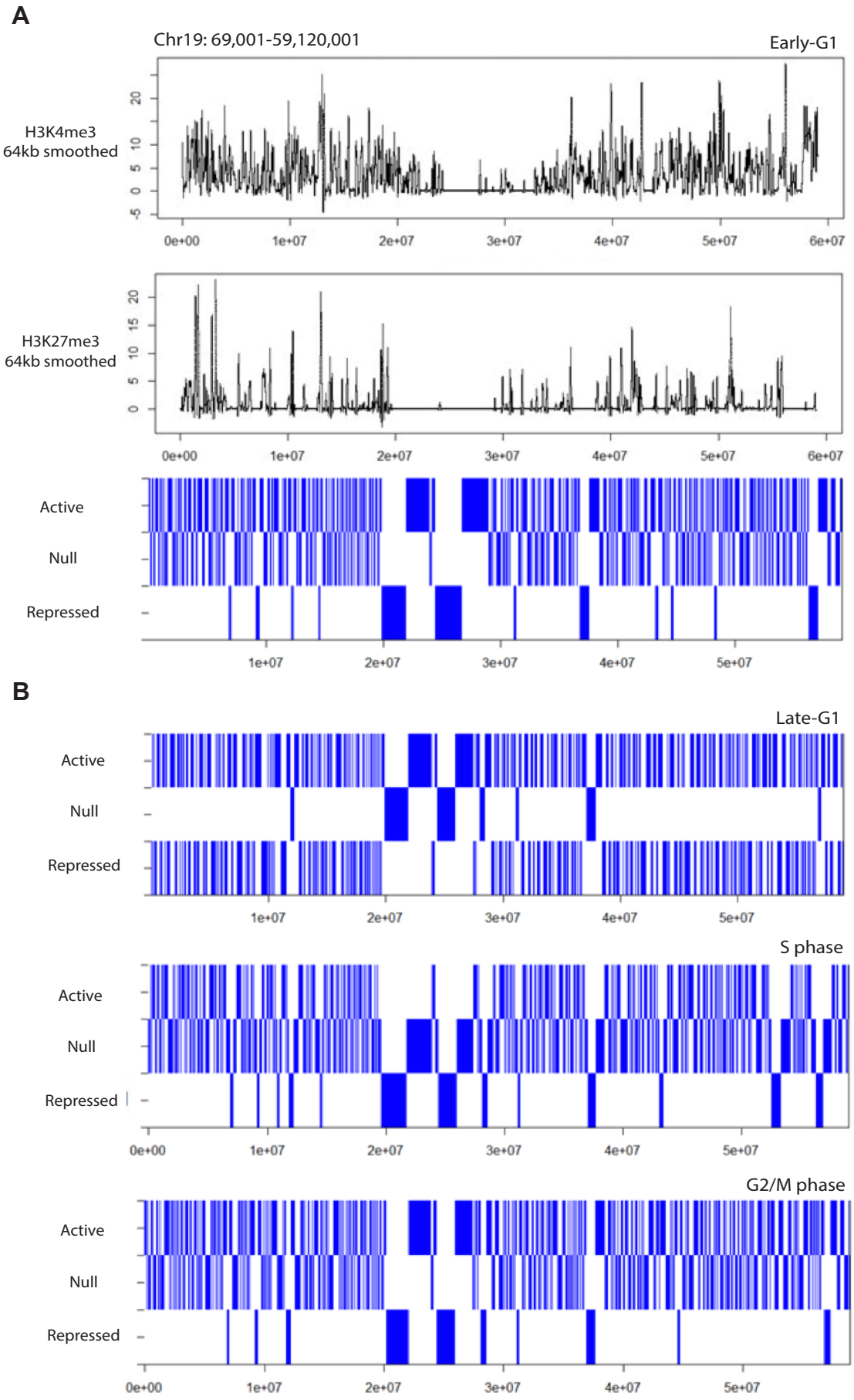
A
Chr19: 69,001-59,120,001
Early-G1
H3K4me3
64kb smoothed
H3K27me3
64kb smoothed
Active
Null
Repressed
B
Late-G1
S phase
G2/M phase

Table 1: Summary statistics of domain size in three-state HMM

Cell Cycle Stage	HMM State	Total size of state (in Mb)	Total domain numbers	Maximum domain size	Mean of domain size	Variance of domain size
Early-G1	**Repressed**	11.351	12	36	7.58	120.08
	Null	25.003	236	30	3.51	25.01
	Active	22.698	248	42	1.75	17.00
	Total	59.052	496	42	2.72	24.36
Late-G1	**Repressed**	25.053	191	30	4.92	38.03
	Null	8.396	7	30	4.49	38.03
	Active	25.603	197	24	2.25	16.96
	Total	59.052	395	30	3.42	29.23
S phase	**Repressed**	11.542	13	31	9.00	131.00
	Null	22.902	212	20	1.89	11.60
	Active	24.608	208	35	3.93	37.24
	Total	59.052	442	35	3.06	28.91
G2/M	**Repressed**	9.097	10	28	6.90	80.54
	Null	26.629	207	31	4.38	38.42
	Active	23.326	217	19	1.73	11.67
	Total	59.052	434	31	3.12	27.82

Table 2: Mean and variance of histone modification distribution within each HMM state

Cell Cycle Stage	Modification	Repressed State	Null State	Active State
Early-G1	H3K4me3	1.291 (4.485)	4.561 (21.522)	3.522 (17.875)
	H3K27me3	0.0 (1.96E-90)	-2.210 (11.567)	-4.76E-4 (0.003)
Late-G1	H3K4me3	2.775 (8.936)	0.429 (0.607)	1.445 (2.729)
	H3K27me3	-3.758 (27.687)	0.0 (1.96E-90)	-0.009 (0.017)
S phase	H3K4me3	0.769 (1.447)	1.680 (4.122)	3.212 (11.228)
	H3K27me3	0.0 (1.96E-90)	-0.002 (0.001)	-1.363 (5.534)
G2/M	H3K4me3	0.521 (0.914)	3.266 (11.451)	1.646 (3.810)
	H3K27me3	0.0 (1.96E-90)	-2.232 (11.642)	4.59E-4 (0.003)