Analysis and mathematical modelling of histone trimethylation in human induced pluripotent stem cells (hiPSCs)

Student: Lihan Lin; Supervisor: Professor Martin Howard, Dr. Ander Movilla Miangolarra; John Innes Centre

Background and aim

Human induced pluripotent stem cells (hiPSCs), reprogrammed from terminal somatic cells have the capacity to differentiate into a variety of tissues. It is, thus, a valuable resource in human cell biology research and regenerative medicine. In the past decades, several hiPSC lines have been produced with varying reprogramming protocol, donor, and subsequent cloning optimization. Although all hiPSC lines are pluripotent, they display heterogeneity in their capacity to differentiate (Yokobayashi et al., 2017). The heterogenous differentiation potential is the consequence of both genetic and epigenetic factors (Nishizawa et al., 2016).

Three important types of epigenetic nucleosome modifications are histone 3 lysine 4 trimethylation (H3K4me3), histone 3 lysine 9 trimethylation (H3K9me3), and histone 3 lysine 27 trimethylation (H3K27me3). H3K4me3 recruits transcription machinery and is a positive transcriptional regulator (Wang et al., 2023); H3K9me3 and H3K27me3 promote heterochromatin formation and silence nearby genes (Yang et al., 2022,Cai et al., 2021).

This project will delve into the epigenetic side of the phenomenon: through statistical analysis and mathematical modelling, we attempt to establish the correlation between epigenetic marks and transcriptional differences between these cell lines, which will inform about their differentiation potential.

Data

Genome-wide ATAC-seq data, and CUT&TAG data for H3K4me3, H3K9me3, H3K27me3, and RNA-seq data in 8 hiPSC lines (obtained by Stefan Schoenfelder's group at Babraham Institute). All data normalized with respect to the median in each dataset (1 = median).

Results

Patterns and trends in H3K9me3 and H3K27me3

Classification of differentially enhanced regions (DERs) and differentially expressed genes (DE genes).

While the transcriptome is mostly uniform (Pearson correlation >0.95), there are regions in the genomes of these cell lines that are differentially enhanced in H3K9me3 and H3K27me3 (differentially enhanced regions, or DERs), which contribute to heterogeneity between the cell lines. These DERs are of great interest and are the subjects of subsequent analysis. DERs are characterized by two criteria: the maximum signal intensity must be greater than two to filter out noise, and the intensity in one cell line must be more than twofold greater than in another cell line in at least seven pairwise comparisons of the same region on the chromosome across all cell lines. The idea is that if the readings in a region can be classified into at least two groups, then, in at least $n-1$ pairwise comparisons, a significant difference can be observed. Consecutive DERs are merged into a single DER. The resolution of the DERs is determined by the bin size used to aggregate raw reads. Higher resolution captures finer details but also introduces more noise to the results; here, we used bin sizes ranging from 2 kbp to 100 kbp to balance these needs.

Non-DERs are regions with intensity greater than two, excluding the DERs, and serve as the null population in hypothesis tests.

Previous research in this project had identified differentially expressed (DE) genes using the DESeq2 package with a significance threshold of $p < 0.01$ and the $>(n-1)$ count criteria (Love et al., 2014).

H3K9me3 DERs are broader than H3K27me3 DERs.

Non-parametric Mann-Whitney U-test rejects the null hypothesis in favour of the alternative that H3K9me3 DERs are broader at alpha = 0.05 ($p \approx 0$).

H3K27me3 DERs relocate away from genes in certain cell lines.

The intensity of H3K27me3 in DERs in cell lines including Yoch6 and Sojd3 is significantly greater than that in other cell lines (Kruskal-Wallis H test: alpha = 0.05, $p \approx 0$). However, according to previous research, the intensity of H3K27me3 within differentially expressed (DE) genes decreases in Yoch6 and Sojd3, which appears contradictory. Notably, the cell lines with overall greater H3K27me3 intensity coincide with those that have a reduced capacity to differentiate (Stefan Schoenfelder lab, unpublished results). A reasonable hypothesis is that H3K27me3 relocates away from DE genes in these cell lines.

To quantify this phenomenon, we multiplied the intensity by the distance of each DER, obtaining a measure of dispersion away from genes that can be compared between different cell lines. Figure 1 shows the p-values for pairwise comparisons of intensity-distance by Mann-Whitney U test after Bonferroni correction. Notably, Yoch6, Kucg2, and Sojd3 stand out as having significantly greater dispersion away from genes.

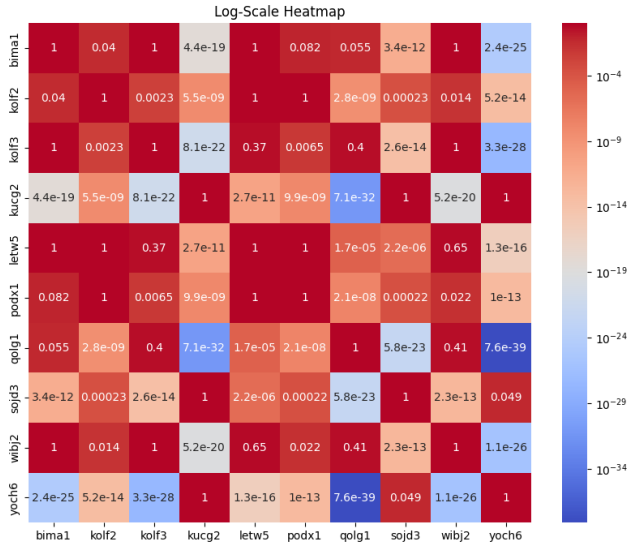


Figure 1. p-values of Mann-Whitney U-test between each cell lines after Bonferroni correction.

H3K9me3 and H3K27me3 have tendencies toward centromeric and telomeric regions.

To evaluate the relationship between differentially enhanced regions (DERs) and structural features of the chromosome, we calculated the distance from the DER to the centromere, normalizing it to 1. H3K9me3 DERs localize to both the centromeric and telomeric regions, while H3K27me3 DERs are found only in the telomeric regions (Figure 2.A). The repetitive sequences in the centromere and telomere could be the cause of significant variability in histone modification. However, no apparent correlation was found between the intensity or length of the DERs and their chromosomal distribution.

In the non-DERs of H3K9me3, the spread across the chromosome is uniform, showing no preference for specific regions. Conversely, the non-DERs of H3K27me3 share the same pattern as the DERs, localizing toward the telomeric regions.

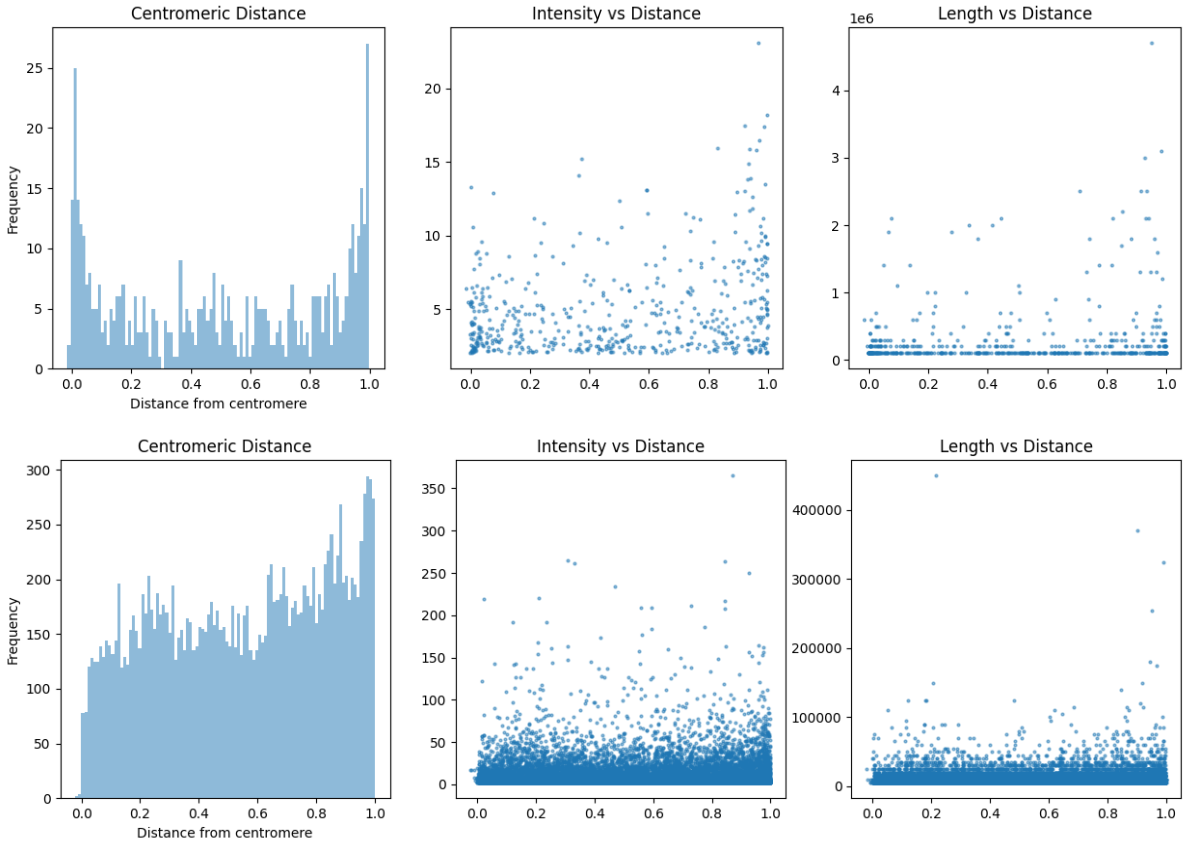
H3K27me3 DERs associate better with genes than H3K9me3 DERs

Histone methylation influences euchromatin formation and transcription machinery recruitment, therefore we are interested in whether the DERs correlate with genes. Figure 2.B shows histograms of the gene coverage fraction in H3K9me3 and H3K27me3. Gene coverage is calculated by summing the lengths of DER-gene overlap normalized by DER lengths. It is an indicator of how many genes are covered by a single DER. The histogram displays multimodality with peaks at integers 0 and 1, meaning that either the DER is entirely covered by a gene, or it is absent of genes.

H3K27me3 DERs localize at genes stronger than H2K9me3 DERs do. Testing with one-sided Mann-Whitney U-test (alpha = 0.05), the above alternative (H3K27me3 DER-coverage is greater than H3K9me3 DER coverage) was accepted with $p \approx 0$.

In comparison, for non-DERs using the same test, both null hypotheses were retained (K9: p = 0.3, K27: p = 0.89).

A



B

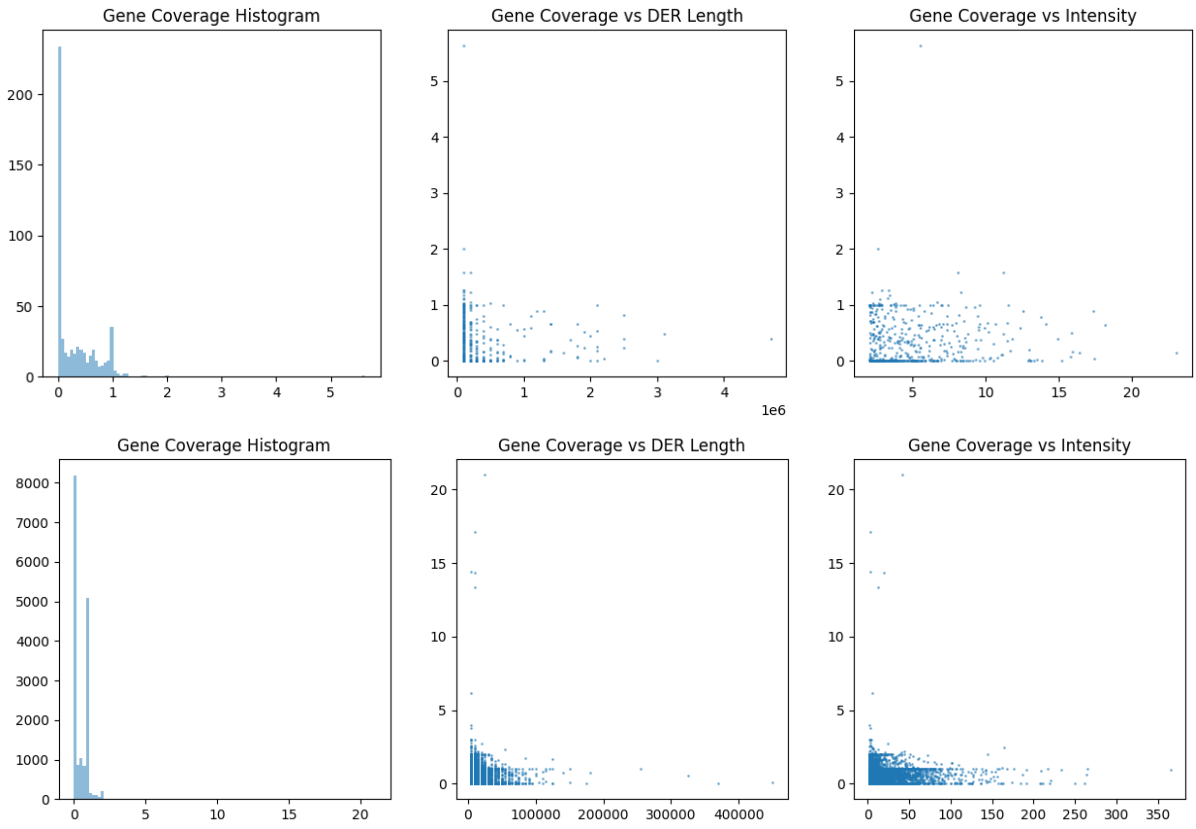


Figure 2. A: Histogram of centromeric distance of H3K9me3 DERs (up) and H3K27me3 DERs (down). B: Gene coverage histogram of H3K9me3 DERs (up) and H3K27me3 DERs (down).

Building a basic model with RNA-seq data

Otsu thresholding

DE genes obtained with DEseq2 are genes differentially expressed across the cell lines. Assuming that the genes are either expressed or silenced (binary states), the FPKM reads of gene mRNA transcripts can be casted into a digital format: 1 for expressed or 0 for silenced. This assumption is supported by the bimodality in RNA-seq data of each single gene across eight different cell lines.

The Otsu thresholding routine was applied to all DE genes. For each single gene the routine calculates the threshold that categorizes the eight RNA-seq reads into expressed or silenced.

Classifying K4-K9-K27 space with support vector machine (SVM)

Plotting the trimethylation intensities of eight cell lines at a single gene yields Figure 3. By labelling the points with as categorized by the Otsu threshold, the dataset can be used to train support vector machines (SVM, linear kernel) that predicts gene expression. The SVM generates three coefficients for H3K4me3, H3K9me3, and H3K27me3, producing a hyperplane that separates the space into expressed and silenced. The intensities of trimethylations are normalized to the (0,1) segment prior to SVM training.

After SVM training, the scaling factors are multiplied back into the coefficients to restore the relationship between the absolute intensities of histone modifications. For each gene we now have three coefficients that indicate the influence on transcription by H3K4me3, H3K9me3, and H3K27me3, see Figure 4. From the plot it is evident that H3K9me3 has very little impact on the expression of DE genes, H3K27me3 negatively impacts DE gene expression, and H3K4me3 promotes gene expression. Applying a manual threshold, the DE genes were classified into 4 groups based on their major source of influence (Figure 5).

Our major interest is in group 3 where the expression of DE genes is influenced by both H3K4me3 and H3K27me3. Column 2 in Figure 6 plots the relationship between H3K4me3, H3K9me3, mRNA signal in group 3. Each line in the plot consists of eight points for eight cell lines at a single gene. The bimodality in K4-K27 plot (bimodality coefficient = 0.70) suggests bistability under certain parameter conditions.

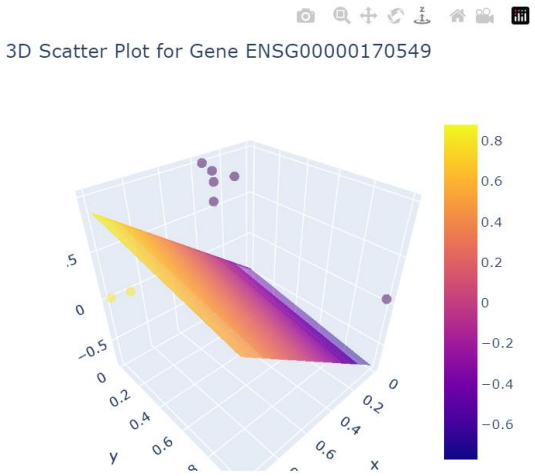


Figure 3. x: H3K4me3 intensity; y: H3K9me3 intensity; z: H3K27me3 intensity (all normalised to the (0,1) segement). Yellow for expressed genes; purple for silenced genes. The hyperplane denotes the boundary of expression/silencing of the gene.

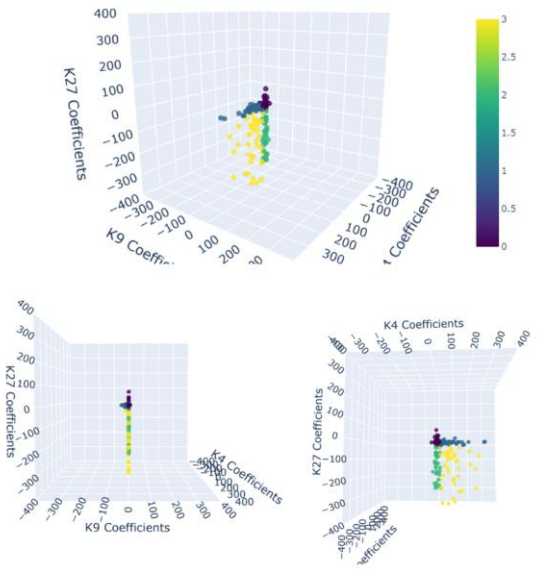


Figure 4: Scatter plot of DE genes plotted using coefficients calculated with SVM. Most effects are from H3K4me3 (positive coefficients) and H3K27me3 (negative coefficients).

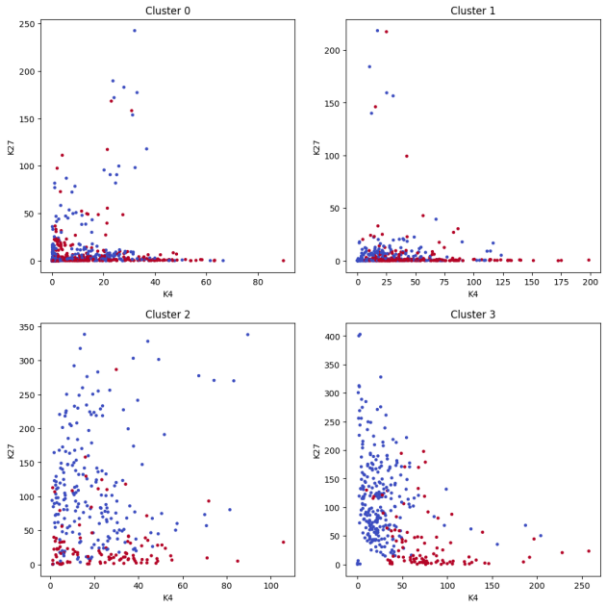


Figure 5. Intensity plot of each cell line in each gene. Blue points are silenced, red points are expressed. Categorized into 4 clusters based on SVM.

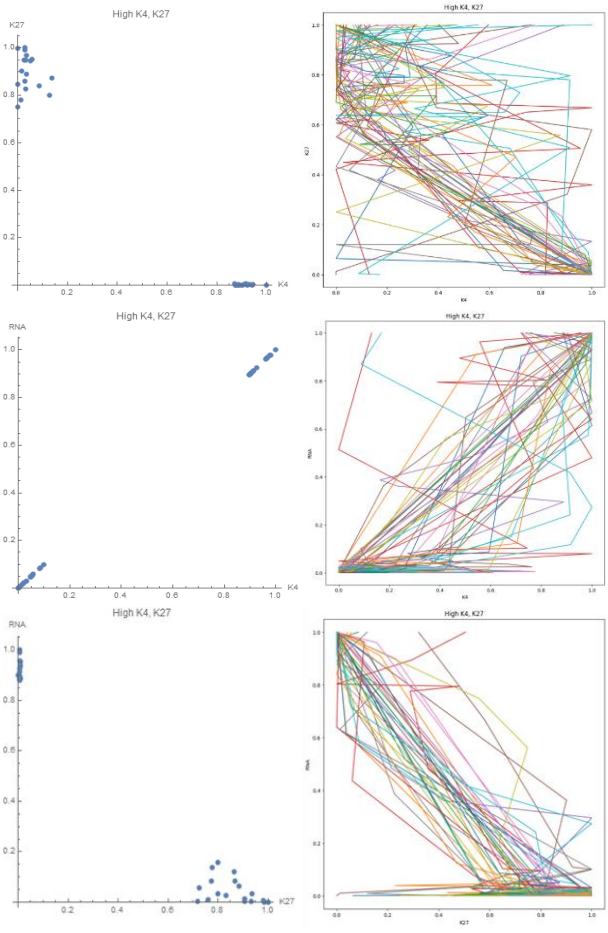


Figure 6. Relationship between H3K4me3, H3K27me3 and gene expression level in Group 3. Column 1 is model output; column 2 is actual data, each line represents 8 cell lines of a single gene.

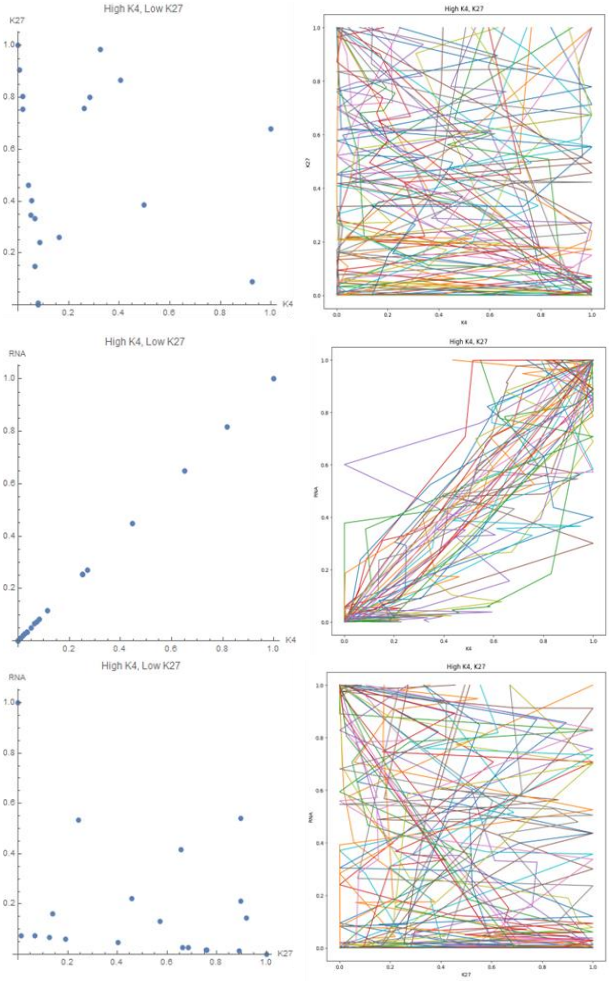
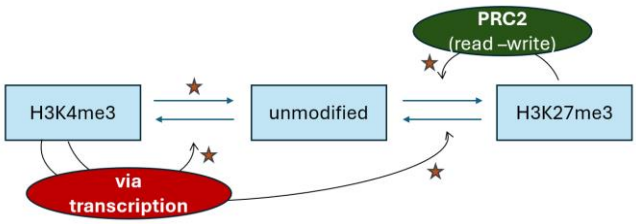


Figure 7. Relationship between H3K4me3, H3K27me3 and gene expression level in Group 1. Column 1 is model output; column 2 is actual data, each line represents 8 cell lines of a single gene.

Three-State Model

The SVM analysis of histone modifications suggests a three-state model with positive read-write feedback to allow the existence of two equilibrium points.



$$\frac{d[K4]}{dt} = r_{me4}[unm] + k_{me4}\alpha[K4][unm] - r_{dm4}\,[K4] \qquad \text{Eq1}$$

$$\frac{d[K27]}{dt} = r_{me27}[unm] + k_{me27}\,[unm]\,[K27] - r_{dm27}\,[K27] - k_{dm27}\alpha[K4][K27] \qquad \text{Eq2}$$

$$\frac{d[mRNA]}{dt} = \alpha[K4] - \beta[mRNA] \qquad \text{Eq3}$$

$$1 = [K4] + [unm] + [K27] \qquad \text{Eq4}$$

The model is based on the knowledge that H3K4me3, co-transcriptionally, promotes its own deposition (Woo et al., 2017) and H3K27me3 have a positive feedback loop through PRC2 (Uckelmann & Davidovich, 2021). Eq1 describes the rate of change in H3K4me3, with two linear terms for background conversion between H3K4me3 and non-methylated state and a non-linear term to account for the feedback through transcription. Eq2 describes the rate of change in H3K27me3, with two linear terms for background rates and two non-linear terms for PRC2 feedback and H3K4me3 regulation. Assuming exclusivity between the states, the three states are summed to one to address the fixed number of nucleosomes.

The table below explains the parameters in the model. The background demethylation rate of H3K4me3 and the ratio between H3K27me3 feedback methylation and background demethylation sets the equilibria of the system. When H3K27me3 read-write methylation rate decreases, the system falls from being bistable to monostable in high K4.

Parameters		Bistable	Monostable in high K4
$k_{me}4$	K4 feedback methylation rate		
$r_{me}4$	K4 background methylation rate (noise)	0~1	0~1
$r_{dm}4$	K4 background demethylation rate		
$k_{me}27$	K27 feedback methylation rate		
$r_{me}27$	K27 background methylation rate (noise)	0~0.1	0~0.1
$r_{dm}27$	K27 background demethylation rate		
$k_{dm}27$	K27 feedback demethylation rate		
alpha	Transcription rate scaler		
beta	mRNA degradation rate		
$k_{me}4$*alpha		20	20
alpha*$k_{dm}27/r_{dm}27$		5	5
$r_{dm}4$		8	x < 20
$k_{me}27/r_{dm}27$		2.5 < x < 10	0.5

Column 1 of Figure 6 shows Group 3 data points generated from the model with varying $k_{me}27/r_{dm}27$ and noise in the background. The predicted data resembles the actual data and managed to recapitulate scattering along the H3K4me3 axis from noise. However the model failed to capture noise in H3K27me3 because the transcription equation is oversimplified and does not have a H3K27me3 term. This is simplification is valid as H3K4me3 and H3K27me3 are exclusive of each other, but has the above downsides.

Figure 7 shows Group 1 of DE genes where expression is influenced by H3K4me3 only. According to the model, this is when feedback strength of H3K27me3 methylation is weak and the system becomes monostable: only one equilibrium with high H3K4me3 and low H3K27me3. The boundary conditions are visualized in Figure 8. We can see that the model recapitulated the linear relationship between H3K4me3 and gene expression which was revealed through SMV analysis.

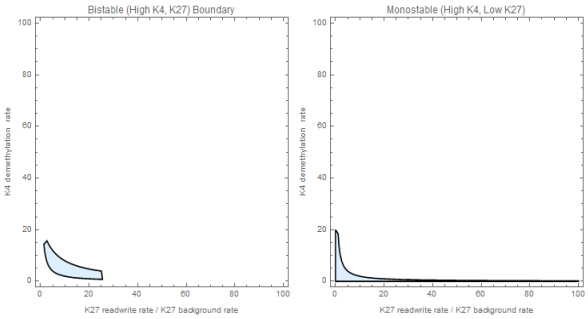


Figure 8. Boundary parameter conditions when the system falls into monostable from bistable. The key parameters are feedback K27 methylation rate / background K27 demethylation rate (x-axis) and K4 background demethylation rate (y-axis).

Conclusion and future directions

In this project, we analysed patterns and trends in H3K9me3 and H3K27me3, the observation of H3K27me3 relocating outside genes is insightful for further exploration of H3K27me3 forming mechanism. This difference in spatial distribution could be an indicator of hiPSC pluripotency. The SVM analysis of histone modification and gene expression successfully classified the genes by their regulatory mechanism. By inspecting this classification we can better understand the relationships between histone marks and transcription, and develop more detailed hypotheses regarding its misregulation. The model is simple and lack many of the finer details in the dataset. However, it suggests the possibility of a bistable system at the DE genes. This hypothesis could be validated in the future with single-cell chromatin modification and RNA-seq data.

Value of studentship

Research group: The student undertook a genomic data analysis project related to a collaboration between the Howard lab and Stefan Schoenfelder (Babraham Institute). He drove the project forward in two main aspects: the in-depth analysis of genomic regions differentially enriched in H3K9me3, and the classification and modelling of how bivalent histone marks (H3K27me3 and H3K4me3) regulate transcription in the different targets. Altogether, the overall research project was significantly developed during Lihan's studentship. The Howard group thanks the Biochemical Society for their support.

Student: Through the studentship, I improved my coding skills and learned common mathematical biology strategies

in dealing with large genomic data. Apart from becoming more confident in answering biological questions with statistic and mathematical tools, I realized the importance of finding and asking questions in research. I also developed my communication skills by presenting results and attending talks and symposiums with the group. The chance of being immersed in an academic setting helped me acquire research mindset and gain a clearer image of what it's like to conduct research, which will be extremely useful in my future career.

Acknowledgements

Thank you to the Biochemical society for funding the studentship. Thank you to the members of Professor Martin Howard's lab at John Innes Centre for hosting my studentship and being so friendly and supportive. I'm especially thankful to Dr. Ander Movilla Miangolarra for overseeing and guiding me through the project.

References

Berry, S., Dean, C., & Howard, M. (2017). Slow chromatin dynamics allow polycomb target genes to filter fluctuations in transcription factor activity. Cell Systems, 4(4). https://doi.org/10.1016/j.cels.2017.02.013

Cai, Y., Zhang, Y., Loh, Y. P., Tng, J. Q., Lim, M. C., Cao, Z., Raju, A., Lieberman Aiden, E., Li, S., Manikandan, L., Tergaonkar, V., Tucker-Kellogg, G., & Fullwood, M. J. (2021). H3K27me3-rich genomic regions can function as silencers to repress gene expression via chromatin interactions. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-20940-y

Love, M. I., Huber, W., & Anders, S. (2014). *Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with Deseq2*. https://doi.org/10.1101/002832

Nishizawa, M., Chonabayashi, K., Nomura, M., Tanaka, A., Nakamura, M., Inagaki, A., Nishikawa, M., Takei, I., Oishi, A., Tanabe, K., Ohnuki, M., Yokota, H., Koyanagi-Aoi, M., Okita, K., Watanabe, A., Takaori-Kondo, A., Yamanaka, S., & Yoshida, Y. (2016). Epigenetic variation between human induced pluripotent stem cell lines is an indicator of differentiation capacity. *Cell Stem Cell*, *19*(3), 341–354. https://doi.org/10.1016/j.stem.2016.06.019

Uckelmann, M., & Davidovich, C. (2021). Not just a writer: PRC2 as a chromatin reader. *Biochemical Society Transactions*, *49*(3), 1159–1170. https://doi.org/10.1042/bst20200728

Wang, H., Fan, Z., Shliaha, P. V., Miele, M., Hendrickson, R. C., Jiang, X., & Helin, K. (2023). H3K4me3 regulates RNA polymerase II promoter-proximal pause-release. *Nature*, *615*(7951), 339–348. https://doi.org/10.1038/s41586-023-05780-8

Woo, H., Dam Ha, S., Lee, S. B., Buratowski, S., & Kim, T. (2017). Modulation of gene expression dynamics by co-transcriptional histone methylations. *Experimental & Molecular Medicine*, *49*(4). https://doi.org/10.1038/emm.2017.19

Yang, H., Bai, D., Li, Y., Yu, Z., Wang, C., Sheng, Y., Liu, W., Gao, S., & Zhang, Y. (2022). Allele-specific h3k9me3 and DNA methylation co-marked CPG-rich regions serve as potential imprinting control regions in pre-implantation embryo. *Nature Cell Biology*, *24*(5), 783–792. https://doi.org/10.1038/s41556-022-00900-4

Yokobayashi, S., Okita, K., Nakagawa, M., Nakamura, T., Yabuta, Y., Yamamoto, T., & Saitou, M. (2017). Clonal variation of human induced pluripotent stem cells for induction into the germ cell fate†. *Biology of Reproduction*, *96*(6), 1154–1166. https://doi.org/10.1093/biolre/iox038