Signatures of digital epigenetic Polycomb regulation in functional iPSC heterogeneity between individuals

Ander Movilla Miangolarra[1], Julia Kurlovich[2], Fereshteh Torabi[3], Oleksandr Dovgusha[2], Yang Cao[3], Lihan Lin[1], Ufuk Günesdogan[2,*], Stefan Schoenfelder[3,*] and Martin Howard[1,3,*]

[1] Dept. of Computational and Systems Biology, John Innes Centre, Norwich Research Park, Norwich NR4 7UH, UK

[2] University of Göttingen, Göttingen Center for Molecular Biology, Dept. of Developmental Biology, Justus-von-Liebig Weg 11, 37077 Göttingen, Germany

[3] Epigenetics Programme, Babraham Institute, Babraham Research Campus, Cambridge CB22 3AT, UK

[*] To whom correspondence should be addressed: ufuk.guenesdogan@biologie.uni-goettingen.de, stefan.schoenfelder@babraham.ac.uk and martin.howard@jic.ac.uk

Abstract

Induced pluripotent stem cells (iPSCs) show substantial heterogeneity between individuals in their capacity to differentiate into various cell types, which is a major hurdle for regenerative cell therapies and disease modelling. To explore the causes of this variability, we assembled a panel of ten human iPSC lines, for which we evaluated the differentiation potential for two different fates, and profiled the transcriptome and epigenome (chromatin accessibility and three post-translational histone modifications). Using a custom-made computational method involving a Support Vector Machine, we dissected the different regulatory modes through which chromatin states affect transcription. We found that regulation of the Polycomb-mediated histone modification H3K27me3 at specific gene loci exhibited the most consistent differences among iPSC lines, frequently displaying a digital pattern (i.e., binary ON/OFF). We developed a mathematical model based on the read-write feedback of the Polycomb complex, which explained this widespread behaviour and showed that even a small change in Polycomb feedback could lead to pronounced changes in histone modification levels. Overall, our findings demonstrate that the heterogeneity in digital Polycomb regulation mirrors the transcriptomic variability across the iPSC lines studied here, and is one of the major sources of iPSC chromatin heterogeneity. Notably, Polycomb heterogeneity was also correlated with the differentiation potentials of the iPSC lines, suggesting that chromatin variability could be one important factor influencing iPSC differentiation potential.

Introduction

All cells in a multicellular organism originate from a totipotent zygote. In mammals, the zygote develops into the blastocyst containing the pluripotent epiblast. As epiblast cells undergo differentiation, they progressively exit the pluripotent state and mature into terminally differentiated cell types. In the past two decades, strategies to induce pluripotency in terminally differentiated cells have attracted much attention [1]. These methods involve the ectopic expression of core pluripotency-associated transcription factors that reprogram somatic cells into induced pluripotent stem cells (iPSCs). iPSCs enable the *in vitro* recapitulation of developmental processes and are a cornerstone of regenerative cell therapies and disease modelling [2].

However, there is substantial variability in the ability of human induced pluripotent stem cell (hiPSC) lines to differentiate into target cell types [3-7]. This heterogeneity has been linked to genetic variability between cell lines and, specifically, between donors [8-10], as well as to differential DNA methylation [3, 5, 11]. Interestingly, genes targeted by the histone-modifying Polycomb complex are prone to considerable expression variability in hiPSCs derived from the same donor [12], suggesting a heterogeneous but significant influence of chromatin regulators. Furthermore, another study showed pronounced variability in histone post-translational modifications (PTMs) across hiPSC lines, exceeding transcriptomic variability [13].

Despite these studies, the extent to which chromatin regulation contributes to the differentiation potential of pluripotent cell lines remains unclear. To better understand the effects of chromatin heterogeneity, we analysed the differentiation potential, transcriptome and chromatin features (accessibility and histone modifications) for a panel of ten feeder- and transgene-free, fibroblast-derived, male and female hiPSC lines generated by the HipSci consortium [14].

In this study, we focused on differentially expressed genes (DEGs) across the ten hiPSC lines and sought to relate this differential expression to chromatin accessibility and histone PTM levels. To this end, we developed a computational analysis pipeline based on a support vector machine (SVM). Unlike other machine learning approaches that primarily aim to predict gene expression from epigenomic data (e.g., [15, 16]), our focus was on uncovering the logic of chromatin-mediated transcriptional regulation. Our analysis shows that the behaviour of around 70% of DEGs could be related to changes in the levels of at least one of the profiled histone marks. In contrast, chromatin accessibility around promoter regions was much less predictive of transcriptional activity. In addition, we found that the modes of chromatin regulation that involved H3K27me3, which is deposited by Polycomb Repressive Complex 2 (PRC2, [17]), displayed the most consistent differences among cell lines. The regulation of H3K27me3 showed a digital, switch-like behaviour that was consistent with a mathematical model based on a read-write feedback of the Polycomb complex [17-20]. Taken together, our results highlight the crucial role of Polycomb complexes in regulating the chromatin landscape of pluripotent stem cells, which we found correlates with their differentiation capabilities.

Results

Marked differences in the developmental potential of hiPSC lines

A previous study had highlighted the variability among hiPSC lines in their differentiation efficiency for dopaminergic neurons [7]. To test whether these differences were specific to the neuronal lineage, we selected ten hiPSC lines from the HipSci consortium that were reprogrammed from skin fibroblasts of nine healthy donors (five male and four female; Supplementary Table 1). We evaluated their competence to generate primordial germ cell-like cells (PGCLCs), for which they were first induced into mesendoderm precursors (pre-ME) and subsequently re-aggregated into embryoid bodies (EBs) in the presence of cytokines [21] (Fig. 1A). On day 4 of differentiation, EBs were analysed by immunostaining for the PGCLC markers AP2Ɣ, BLIMP1 and SOX17 (Fig. 1B, Supplementary Fig. 1A). This revealed marked differences in PGCLC differentiation efficiency among the cell lines: one cell line (*Podx1*) showed high potential (~45% PGCLCs), five exhibited intermediate potential (~15-28%), and two (*Kucg2* and *Letw5*) had low potential (~5%) (Fig. 1B, C, Supplementary Fig. 1A). Notably, two cell lines (*Sojd3* and *Yoch6*) failed to form EBs after repeated attempts (Fig. 1C, D). We also performed FACS (fluorescence-activated cell sorting) on selected cell lines using antibodies against the PGCLC surface markers INTEGRINα6 (INTα6) and EpCAM [22-24], which independently confirmed the immunofluorescence data (Supplementary Fig. 1B). Together, these results demonstrate significant inter-line variability in the differentiation potential of hiPSCs into PGCLCs. Remarkably, *Kucg2*, *Sojd3* and *Yoch6* also failed to differentiate into dopaminergic neurons in the study of Jerber and colleagues (the other seven hiPSC lines either succeeded or were not tested, [7]).

We next asked whether the low number of PGCLCs derived from *Kucg2* or *Letw5* hiPSCs is due to inefficient PGC specification or a failure to maintain the germ cell fate. To address this question, we induced PGCLCs and collected EBs on days one to four for immunofluorescence analysis. EBs derived from *Letw5* hiPSCs initially contained a high proportion of PGCLCs, comparable to cell lines with intermediate or high potential (Fig. 1E, Supplementary Fig. 2). However, *Letw5*-derived PGCLC numbers progressively declined over the time course, suggesting a failure to maintain cell identity and/or impaired proliferation. In contrast, EBs derived from *Kucg2* hiPSCs displayed a low number of PGCLCs already at day one of differentiation, indicating inefficient specification of PGCLCs. These findings suggest that *Kucg2* hiPSCs have limited developmental competence to generate PGCLCs, while *Letw5* hiPSCs are capable of PGCLC

specification but fail to sustain the germ cell fate, pointing to a defect in fate maintenance rather than in initial developmental potential.

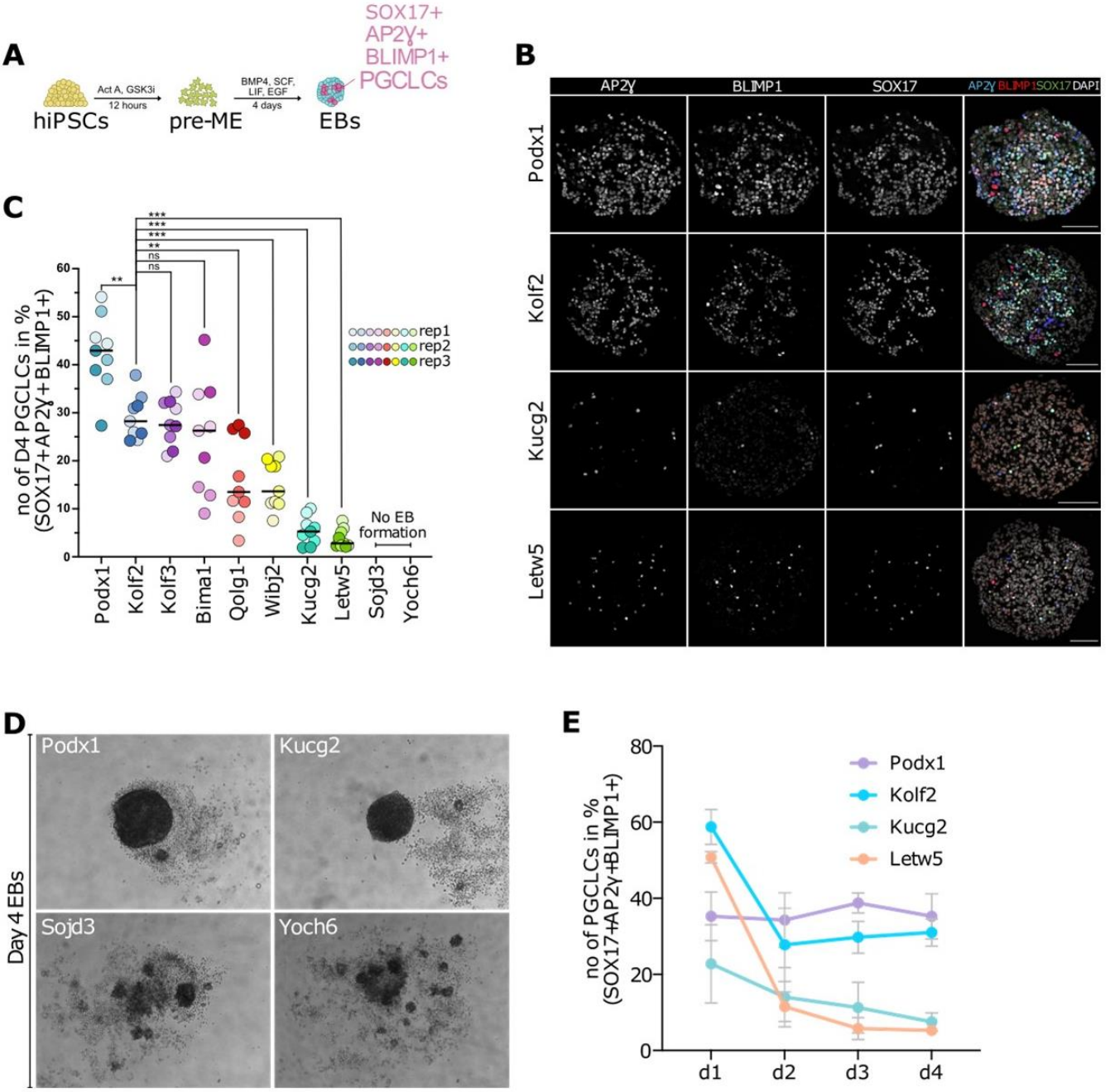

Figure 1. Variation in PGCLC differentiation potential. *A) Schematic representation of PGCLC differentiation from hiPSCs. B) Immunofluorescence images for AP2γ, BLIMP1 and SOX17 of EBs at day 4 of differentiation from indicated hiPSC lines. AP2γ+, BLIMP1+ and SOX17+ cells represent PGCLCs. C) Quantification of immunofluorescence data showing the percentage of AP2γ+, BLIMP1+ and SOX17+ cells. Median values are represented by horizontal black lines. D) Brightfield images of day 4 (d4) EBs generated from indicated hiPSC lines. E) Quantification of immunofluorescence images showing the percentage of PGCLCs (AP2γ+ BLIMP1+ SOX17+ cells) generated in day (d) 1-4 EBs. Error bars indicate standard deviation, n = 2 biological independent replicates. Scale bars, 100 µm. ActA, Activin A; GSK3i, glycogen synthase kinase 3 inhibitor; BMP4, bone morphogenetic protein 4; SCF, stem cell factor; LIF, leukaemia inhibitory factor; EGF, epidermal growth factor. Two-sided, unpaired t-test: **p <0.01; ***p <0.001; ns, non-significant.*

To determine whether the impaired differentiation potential for PGCLC fate of *Letw5* and *Kucg2* hiPSCs, and the failure of *Sojd3* and *Yoch6* hiPSCs to generate EBs, is specific or reflects a broad impairment of pluripotency, we differentiated these hiPSCs lines into definitive endoderm (DE), alongside hiPSC lines with high or intermediate PGCLC differentiation efficiency. For this, hiPSCs were first induced into mesendoderm, which was then cultured in the presence of Activin A and BMP inhibitor for an additional 48 hours to promote DE differentiation ([21]; Supplementary Fig. 3A). The resulting DE cells were

analysed using immunofluorescence for SOX17, which is a marker for both PGCLCs and DE, and the DE marker FOXA2 [25, 26]. Intriguingly, the cell lines that efficiently generated PGCLCs (*Podx1*, *Kolf2*) exhibited relatively low DE differentiation efficiency (~10-20% SOX17+ and FOXA2+ cells). In contrast, the other cell lines (*Kucg2, Letw5, Sojd3, Yoch6*) differentiated efficiently into DE (40-80% SOX17+ and FOXA2+ cells; Supplementary Fig. 3B, C), which is consistent with a previous study showing high DE differentiation efficiency of *Kucg2* and *Sojd3* hiPSCs [27]. It is noteworthy that *Kucg2* and *Letw5* hiPSCs robustly upregulated *SOX17* during DE differentiation but failed to do so during PGCLC induction. These results demonstrate that a single hiPSC line can be directed into multiple lineages with variable efficiencies.

Transcriptional and epigenetic variation among ten hiPSC lines

To further analyse the variation among hiPSC cell lines, we performed RNA-seq (RNA-sequencing). This confirmed a high degree of correlation in gene expression (measured in transcripts per million, TPM) between cell lines (Pearson correlation, excluding sex chromosomes, >0.96; see Supplementary Fig. 4A), as previously reported [13]. However, Principal Component Analysis (PCA) separated the cell lines into two distinct groups, with three cell lines (*Yoch6*, *Sojd3* and *Kucg2*) clustering on the right side of PC1 (24% of variance, Fig. 2A). It is noteworthy that these three cell lines displayed low differentiation efficiency for PGCLCs and dopaminergic neurons, but high potential for DE.

Differential gene expression analysis based on pairwise comparisons between cell lines (Methods) uncovered 712 DEGs on autosomes among the ten hiPSC lines. Gene ontology (GO) analysis of the 712 DEGs for molecular function showed enrichment in transcriptional regulators and transmembrane receptors (Fig. 2B), which could be of importance for developmental processes and maintenance of pluripotency. GO for biological processes supports this view by showing enrichment of genes involved in pattern specification and organ development (Supplementary Fig. 4B).

Next, to gain insights into the chromatin states that may underpin differences in gene expression, we performed Cleavage Under Targets and Tagmentation (CUT&Tag, [28]) for three histone PTMs (H3K4me3, H3K9me3 and H3K27me3) in all ten hiPSC lines (2 replicates per histone PTM per cell line). In addition, we profiled chromatin accessibility data (as measured by ATAC-seq) for these cell lines from a previous study [29]. For each epigenetic variable, we performed PCA by integrating the enrichment values at each gene over the promoter region (Fig. 2C). We defined the promoter region as between 1kb up- and downstream of the gene start to better capture the peaks of H3K4me3 and ATAC-seq, similarly to other studies [30]. Of the 4 variables considered (ATAC-seq, H3K4me3, H3K9me3 and H3K27me3), only H3K27me3 mirrored the separation between cell lines observed in PCA using RNA-seq data (Fig. 2A, C). Consistently, local differential enrichment of H3K27me3 correlates with transcriptional activity of DEGs including *EPOP*, *HES2* and *CRLF1* (Fig. 2D). When considering gene bodies rather than promoters (defined as from 1kbp downstream of the gene start to the gene end, for consistency with the previous promoter definition), PCA indicated both repressive histone marks (H3K9me3 and H3K27me3) also mirrored the separation seen in the PCA for RNA-seq data (Supplementary Fig. 4C). However, for H3K9me3, this was primarily along PC2, suggesting a less pronounced segregation. We confirmed that our conclusions are robust to an alternative promoter definition (-1kb to gene start).

Overall, our analyses revealed a substantial number of DEGs between human iPSCs, and showed a genome-wide relation between transcriptional heterogeneity and the levels of repressive histone marks, particularly H3K27me3.

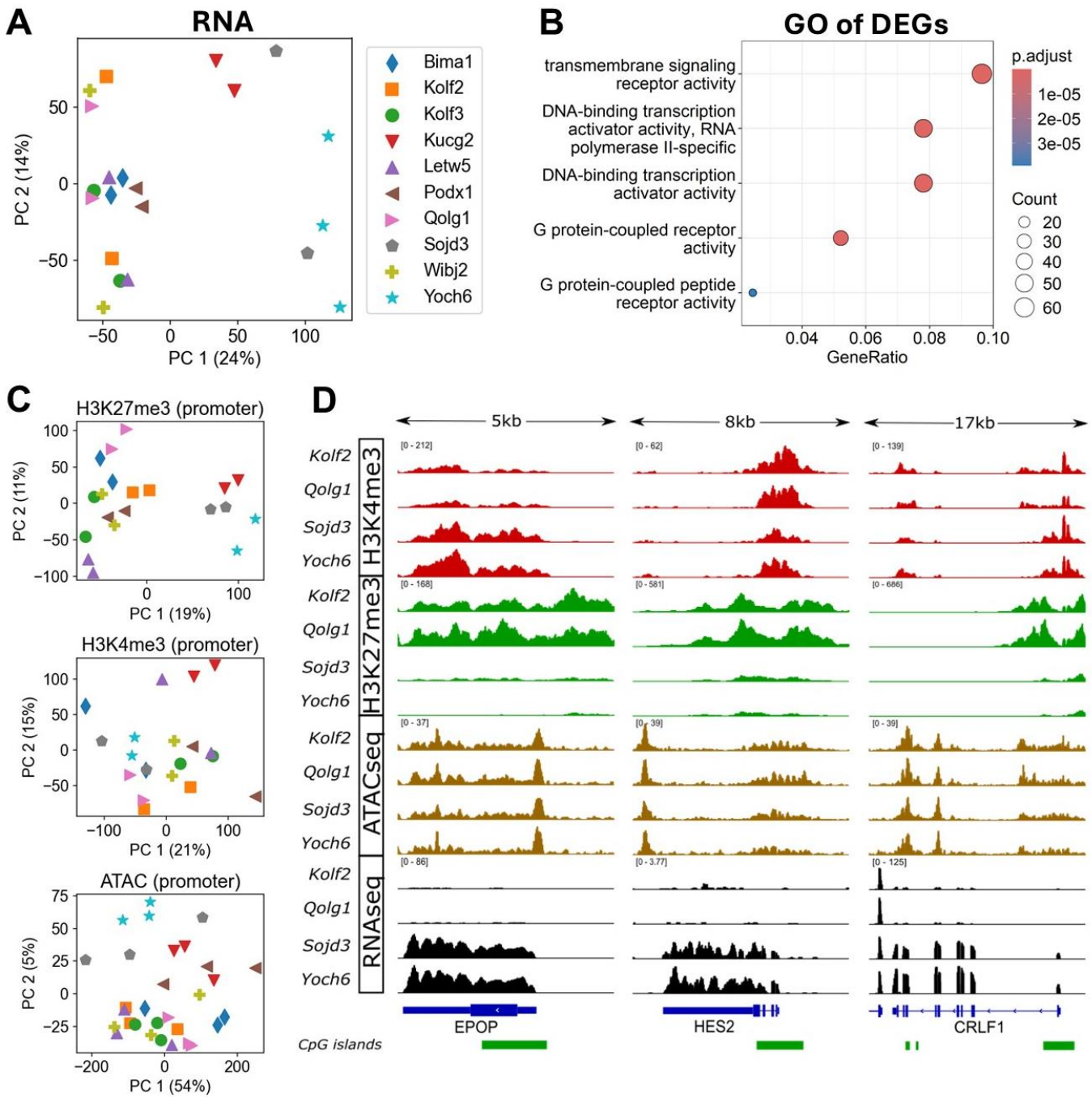

***Figure 2**. **Transcriptional and epigenetic variability in hiPSC lines**. A) PCA of the RNA-seq data for the 10 hiPSC lines. Each cell line is represented by a different marker shape and colour (see legend on the right). The variance explained by each of the principal components is given in parenthesis, next to each axis. B) Molecular function GO of the 712 DEGs. The fraction of genes (out of the total of 712) involved in a given molecular function is represented along the x-axis. C) PCA of the epigenomic profiles. For each gene, the signal was integrated over the promoter region (between 1kb up- and downstream of the gene start). See legend in panel A. D) Genome browser view of three DEGs, for four of the hiPSC lines. RNA-seq, ATAC-seq and two histone marks (H3K4me3 and H3K27me3) are shown. ATAC-seq data were obtained from [29].*

Linking variation in chromatin features with transcriptional output using SVMs

We next sought to elucidate how variability in histone PTMs or chromatin accessibility relates to transcriptional activity. To address this, we developed a computational pipeline with two main objectives: first, to predict gene expression levels based on histone PTMs and chromatin accessibility data; and second, to understand the logic linking chromatin variability to transcriptional regulation.

The pipeline takes as input the values for histone PTM enrichment and chromatin accessibility (hereafter 'epigenomic variables') and the transcription levels of each DEG for each cell line. To avoid artificially inflating the number of datapoints by combinatorially matching replicates of different epigenomic

variables, we used the merged data of each variable in each cell line. Since H3K4me3 and chromatin accessibility typically display peaks near transcription start sites, their signal were quantified over promoter regions (overwhelmingly defined as above, but see Methods for details). The signals for the repressive marks H3K9me3 and H3K27me3 were integrated over the entirety of the gene (from start to end) [31]. The pipeline's cornerstone is a Support Vector Machine (SVM), a machine learning method designed to predict a binary outcome (here, transcription) from an input (epigenomic variables). A key advantage of this technique is that, in the way used here, it is straightforward to interpret the parameters learnt from the data. Since an SVM is a binary classifier, the transcription values were binarized to distinguish between higher or lower expressed genes in each cell line, for which we used a K-means clustering approach (Fig. 3A). After binarization, the data were ready to be passed to the SVM. Given the differences in gene-specific regulation, we analysed each gene independently, by training a different SVM for each one. Although reducing transcription levels to a binary output may be an oversimplification, this approach enables robust predictions with relatively sparse datasets, as in this case with ten datapoints (cell lines) per gene.

SVMs work on the space of the input data for each gene, which is defined by four axes (three for histone PTMs and one for chromatin accessibility). Each cell line is represented as a datapoint within this four-dimensional epigenomic space and the transcriptional output, all with respect to a given gene. Once the SVM has been trained (for details, see Methods), the result is an optimal hyperplane that, if possible, separates cell lines in which the gene of interest is more highly expressed from those in which it is not. This hyperplane can then be used to predict the binary expression levels of additional datapoints (or cell lines; Fig. 3B). The vector normal to this hyperplane indicates the direction in which the epigenomic variables need to change to upregulate transcription. This vector is our gateway to identifying the regulatory mechanism, mathematically described by $\boldsymbol{n}_i = (n_{i,K4}, n_{i,K9}, n_{i,K27}, n_{i,ATAC})$. Thus, each gene i can be mapped to a point within a regulatory space, which is defined by the extent to which each epigenomic variable contributes to the prediction of transcription, in terms of the log-rescaled $\boldsymbol{n}_i$ coefficients (see Methods and Fig. 3C). Next, we clustered genes according to their regulatory mode using another K-means algorithm (see Methods and Supplementary Fig. 5 A, B), resulting in 10 clusters (named after the variables with average $|\log - \text{resc. coeff.}| > 0.2$). About 80% of the DEGs could be clustered into a biologically meaningful regulatory mode involving the chromatin state, with only a minority of about 20% of DEGs in the 'origin' cluster, for which we could not establish a strong connection between transcription and epigenomic variables (see Fig. 3D for a two-dimensional PC projection of the resulting four-dimensional regulatory space).

The regulatory space is shown in terms of the log-rescaled SVM coefficients that integrate the SVM output (hyperplane coefficients, $\boldsymbol{n}_i$) and the variation of the corresponding epigenetic variable at the gene of interest (hence, the 'rescaled' prefix). Since the data has to be normalised for the SVM training to be robust, this rescaling ensures that the original scale of variations in histone marks is taken into account (with larger variations deemed more likely to affect transcription than smaller ones). The absolute value of the log-rescaled coefficient is bounded by 1, with values closer to 1 indicating stronger regulatory association. The sign of the coefficient indicates the relationship with transcription: negative values for repression and positive values for activation. As expected, most of the log-rescaled coefficients for H3K4me3 were positive, whereas the majority of the coefficients for H3K27me3 marks were negative, consistent with their roles as activating or repressive chromatin marks, respectively [32].

The SVM-based classification of genes by chromatin regulation is only meaningful if the SVM itself performs well, that is, if it can classify correctly high or low levels of transcription from its epigenomic signals. To test this, we trained the SVM on data from nine hiPSC lines and tried to predict expression patterns for the tenth independent hiPSC line based on its epigenomic profile. On average, the SVM (and, by extension, the entire pipeline) accurately predicted gene expression for around 85% of the DEGs. This performance improved to around 90% for subsets of genes whose transcription is strongly associated with epigenetic signals (Supplementary Fig. 5C). To rigorously test the accuracy of the classification, we used Platt scaling [33] to obtain a classification score from the SVM that we could use to draw a receiver

operating characteristic (ROC) curve (see Methods). The Area Under the ROC (AUROC) curve is a standard measure for the accuracy of a classification method. When predicting gene expression in the tenth cell line using the remaining nine for training, we obtained an average AUROC of 0.89. Since a value of 1.0 represents a perfect classifier and 0.5 a random one, this result suggests that the SVM performs very well, especially given the limited training data available for each gene (nine datapoints). As an alternative approach, we also used a logistic regression instead of an SVM, for which we obtained an average AUROC of 0.85 (Supplementary Fig. 5D, E). When further interrogating this small difference, we noticed that the prediction is harder to make correctly for the three hiPSC lines that separate from the remaining cell lines in the PCA analysis for RNA-seq data (Fig. 2A; Supplementary Fig. 5D, E). In these cell lines, the AUROC of the logistic regression drops as low as 0.57, whereas the SVM still maintains a satisfactory performance of about 0.75. These results suggest that the SVM predicts high or low expression from epigenomic variables more effectively than a logistic regression, particularly in cases with limited data.

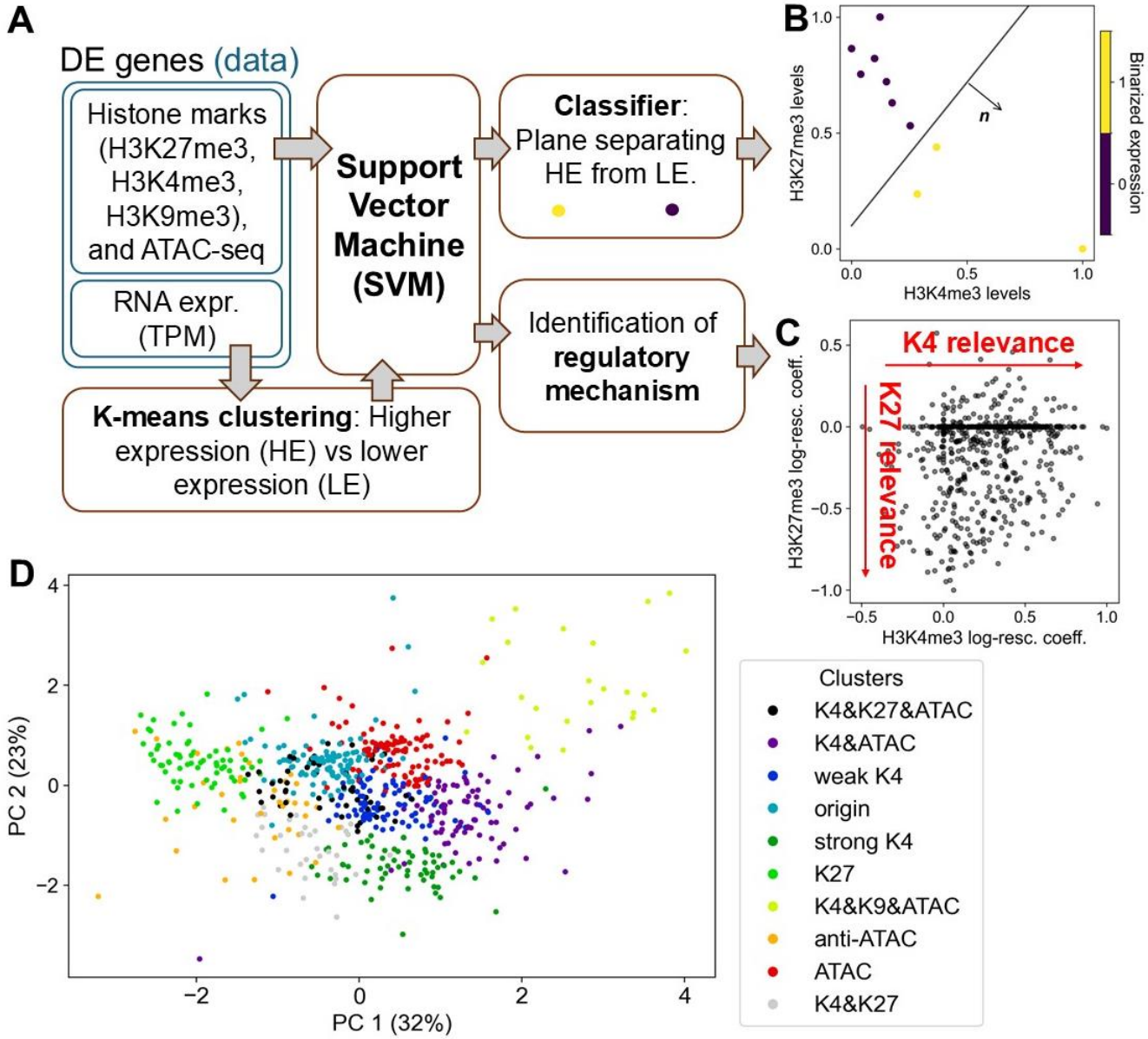

***Figure 3**. **Using an SVM to identify regulatory modes.** A) Schematic of the computational pipeline used to analyse the combined datasets. B) Example of a 2D SVM line separating cell lines with higher expression from those with lower expression for a given gene. C) Example of the regulatory space obtained from a 2D SVM. Each point is a gene, whose position in the regulatory space is determined by the relationship between transcription and the epigenomic variables. D) Two-dimensional PC projection of the 4D regulatory space, with the regulatory modes (clusters) colour coded. The variance explained by each of the principal components is given in parenthesis, next to each axis.*

Chromatin accessibility does not always correlate with transcription

We next investigated the differences in the behaviour of genes associated with each regulatory mode in more depth. Our clustering identified 10 distinct regulatory modes. Unexpectedly, chromatin accessibility emerged as a salient feature only in a subset of these modes. Thus, we interrogated the data to clarify the relation between accessibility and transcription in each of the regulatory modes. The average log-rescaled ATAC-seq coefficient for each regulatory mode (Fig. 4A) showed that, while half of the clusters exhibit a positive log-rescaled coefficient, four clusters have coefficients close to zero, indicating minimal or no association between accessibility and gene expression. Surprisingly, in one cluster, the ATAC-seq signal at the promoter seemed to correlate with gene repression (named 'anti-ATAC'). In addition, many genes silenced by H3K27me3 either did not show significant changes in chromatin accessibility, or the changes in accessibility were not correlated with transcriptional variability. In fact, less than 40% of genes repressed by H3K27me3 displayed reduced ATAC-seq signal at their promoters. These results are consistent with reports showing that chromatin accessibility, as measured by ATAC-seq, is not always tightly linked to transcriptional activity [34-36].

H3K27me3 regulation is digital

To visualise the data within the different regulatory modes, we combined all genes within each cluster by normalising the values of transcription and the epigenomic variables between 0 and 1 (the maximum of each variable across cell lines was mapped to 1 and the minimum to 0), independently for each gene. Then, all genes within a cluster (each gene with ten datapoints, one from each cell line) were visualised within a single graph (Fig. 4B). We observed a variety of forms of transcriptional regulation: from a fairly linear relationship between RNA expression and H3K4me3 for cluster 'K4&ATAC', to a more complex link between RNA expression and two histone marks ('K4&K27' cluster). Notably, the 'K27' cluster exhibited bimodality, with large numbers of datapoints in the expressed region of the graph (low H3K27me3 and high expression) and in the silent region of the graph (high H3K27me3 and low expression), but with a largely unpopulated central region (Fig. 4B).

Based on these observations, we sought to dissect the nature of the relationships between epigenomic variables and transcription: are they continuously varying features or are there abrupt changes? A more rigorous quantitative version of the previous qualitative analysis involves computing Hartigan's Dip Statistic (HDS), which, given the distribution of a variable (e.g., histone PTM level within a regulatory mode), quantifies the distance to the nearest unimodal distribution [37]. Thus, a higher HDS indicates a greater deviation from unimodality, suggesting that the distribution is bimodal or multimodal and may therefore reflect two or more discrete states. We computed the HDS for H3K4me3 and H3K27me3 using the promoter and the gene body annotation (Fig. 4C, see also Supplementary Fig. 6A for a similar analysis with RNA-seq and ATAC-seq data, which show moderate or no bimodality at all). We found that, overall, H3K27me3 exhibits the highest degree of bimodality, particularly at gene bodies, across all three regulatory modes in which it is involved (see also Supplementary Fig. 6B for the histograms of some of these distributions, with various degrees of bimodality). However, we also found significant bimodality for H3K4me3 in the promoter regions of 'K4&ATAC' and 'K4&K9&ATAC' regulatory modes, although the latter consists of only ~20 genes. We confirmed that, for an alternative promoter definition (from -1kb to gene start), H3K27me3 is still the most bimodal mark and that its bimodality is also greater at the gene body than at the promoter.

We further visualised this bimodality in a "swarmplot" in which the normalised enrichment of the histone mark (for each datapoint of a given regulatory mode) is represented on the y axis, and the colour denotes the normalised expression (Fig. 4D). Interestingly, in the case of H3K27me3 at gene bodies, within the 'K27' regulatory mode, bimodality seems to be closely linked to transcriptional activity: the top lobe of the distribution corresponds to repressed instances, while the lower lobe represents expressed ones. A sharp transition in expression occurs around 30% of the maximum H3K27me3 signal, marked by a narrow intermediate region containing relatively few datapoints. The fact that the transition from repressed to

expressed occurs in a region in between the two main lobes of H3K27me3 enrichment (as opposed to taking place in the middle of one of the lobes) suggests a role for H3K27me3 bimodality in regulating transcription. Furthermore, the transition value of 30% allows for repression to be stably maintained even after DNA replication, when histone modification levels will be transiently halved, on average. It should also be noted that other epigenomic variables do not exhibit pronounced bimodal behaviour (see two examples in Fig. 4D, 'strong K4' and 'ATAC'). Even in cases where H3K4me3 shows some degree of bimodality (e.g. 'K4&ATAC' cluster), the bimodality is less marked and the transition from repressed to expressed states appears more gradual.

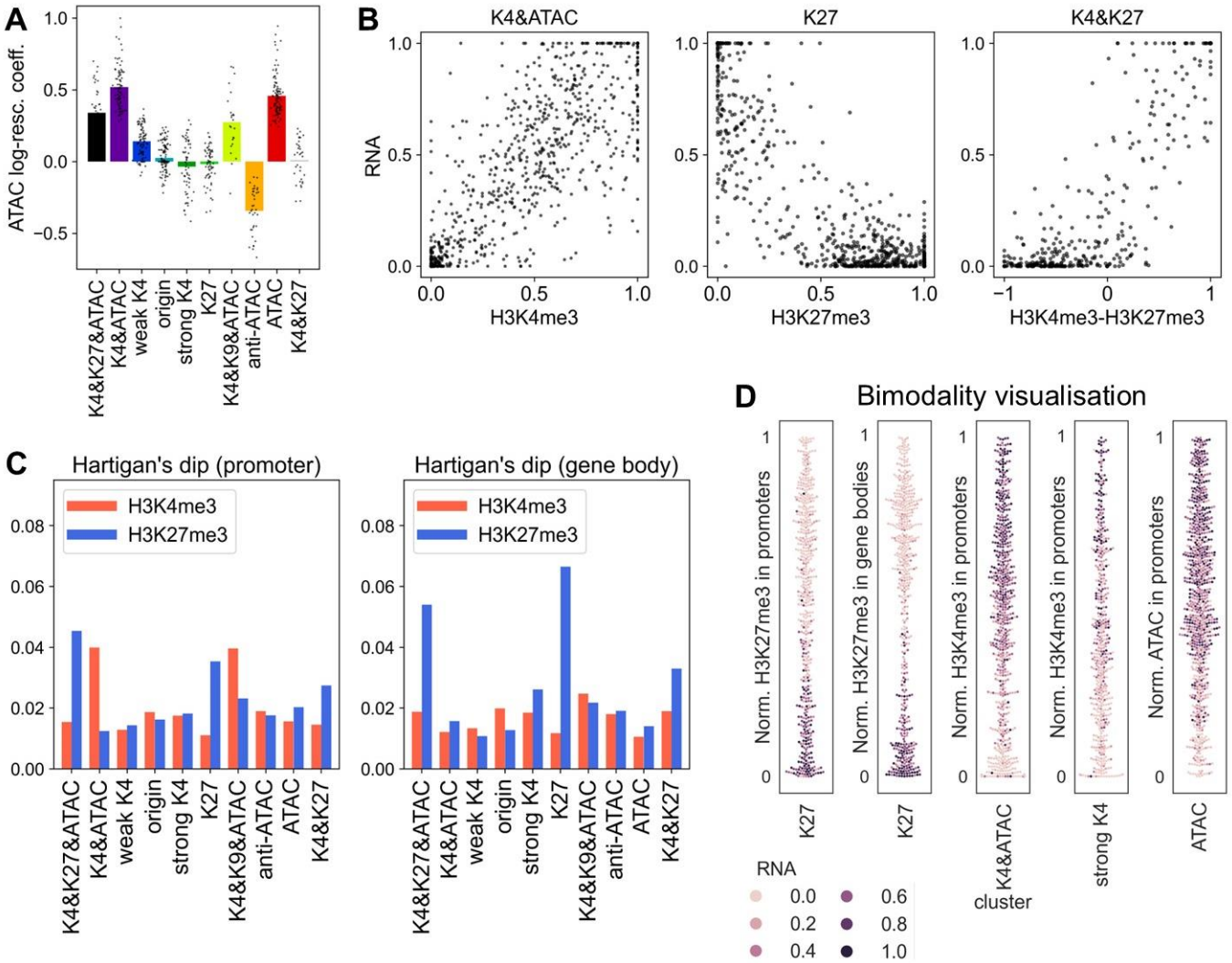

***Figure 4**. **Analysis of regulatory modes**. A) Log-rescaled coefficient for accessibility in each of the regulatory modes. Bars indicate averages and dots individual genes. B) Scatter plot of transcriptional output as a function of the relevant epigenomic variables for each regulatory mode. Both transcription and the epigenomic variables are normalised between 0 and 1 (per gene, as explained in the text). C) HDS for the histone PTMs H3K4me3 and H3K27me3, for each regulatory mode, taken over two different annotations: promoter (left panel) and gene body (right panel). For each regulatory mode the H3K4me3 HDS is depicted in red (left column) and the H3K27me3 HDS is shown in blue (right column). D) Visualisation of bimodality and its relationship with transcription, for selected regulatory modes. The normalised value of the relevant epigenomic variable of a datapoint is given by its position on the y-axis. The width of the swarm of points shows the number of datapoints in that range of enrichment values and shows clear bimodality in some cases. The colour of the datapoints represents normalised transcriptional output for a given gene and cell line.*

The behaviour of H3K27me3 can be explained using a simple mathematical model

Above, we have highlighted the bimodal behaviour of H3K27me3 in gene bodies and its tight relationship with transcriptional output. However, we also observed a crosstalk between H3K27me3 and H3K4me3 in

at least two regulatory modes, with similar but slightly different behaviour in terms of multimodality or correlations between the marks. In addition, we found that the behaviour of H3K27me3 in promoter regions may be less bimodal than in gene bodies. We therefore tried to mechanistically explain these results by means of a mathematical model.

The model considers each gene to be an independent locus where histone 3 tails at promoters can be in one of five possible states: H3K4me3, unmodified, H3K27me1, H3K27me2 and H3K27me3. At gene bodies, histone 3 tails can only be methylated at residue 27 (for a schematic of the model see Fig. 5A). Transition between states take place via basal linear rates (parametrised by k_0) or, for the addition of a methyl group to the lysine 27, via the allosteric feedback of PRC2, resulting in a nonlinear transition rate proportional to the fraction of H3K27me3 in the locus, parametrised by k_f. So far, this largely follows other well-established mathematical models in the literature [19, 20, 38, 39], except for the fact that we considered promoter regions and gene bodies separately. These two sets of histone tails only interact via the Polycomb feedback, controlled by the parameter c. We take k_f equal for promoters and gene bodies, but promoter regions are assumed to have a larger basal rate, given their higher overlap with CpG islands (known to recruit Polycomb complexes). Thus, we set $k_{0,p} = 1.5k_{0,b}$ where $k_{0,p}$ is the basal rate of promoters and $k_{0,b}$ that of gene bodies. In addition, we also considered a periodic perturbation due to DNA replication, where all histone PTMs are diluted on average by a factor of two due to the inclusion of newly synthesised unmethylated histones (shown to be important for transcriptional regulation, both experimentally [40-42] and theoretically [43, 44]). For more details on the mathematical modelling, see Methods.

To compare the output of the model to experimental data, we used an approach that creates synthetic data for qualitative and quantitative comparison. Furthermore, we assumed that most of the noise in the analysis of experimental data came from 1) technical/biological noise and 2) grouping genes (which might have different features) together, as done when clustering into regulatory modes. Therefore, we opted for deterministic modelling (ordinary differential equations) to which noise is added before (in the parameters, representing slightly different features of genes) and after (in the resulting outputs, representing technical/biological noise in experiments). Finally, the model required a parameter that can be varied to explain the differences between cell lines. In this case, we chose E, a parameter that controls both rates of methyl addition to H3K27 residues, given that the largest cell-line dependent effects occurred in H3K27me3-related regulatory modes (see next section). Thus, to obtain data for a given regulatory mode, we simulated the model as many times as there are genes that were clustered into that regulatory mode, each time with different random parameters (with a prescribed mean, see Table 1). To simulate each of the cell lines, we modulated the parameter E from 1 to 0.8 (in ten equally spaced intervals, as many as the number of cell lines). Gathering data from each simulation and adding noise to the outputs to mimic technical/biological noise, we obtained as many synthetic datapoints as experimental ones, which we then analyse in the same manner (Fig. 5B; for more details, see Methods).

With appropriate parameterisation, we found that the model could largely explain the behaviour of the three relevant regulatory modes by only varying two parameters: c (the interaction between the promoter and gene body) and R (the ratio between the feedback and basal rates $k_f/k_{0,p}$).

Furthermore, the model could quantitatively recapitulate the HDS for each of these three regulatory modes and three different variables including H3K4me3 at promoters, H3K27me3 at promoters and H3K27me3 in gene bodies (Fig. 5C, left panel). In general, gene bodies had the highest HDS (both in experiments and in the model) which, in the model, was due to having a lower basal rate than promoters, yet the same feedback rate. The difference between the HDS in gene bodies and promoters was maximal in the 'K27' cluster, which we parametrised as having a lower c. Thus, this decoupling of promoter and gene body explains the large difference in behaviour. Increasing c makes the HDS of promoters and gene bodies converge, as in the 'K4&K27&ATAC' cluster. Finally, it is worth noting that the H3K4me3 HDS was very low in both experiments and in the model. This was surprising since, in the model, the two histone PTMs cannot coexist in the same histone tail and, therefore, H3K4me3 levels should respond accordingly to

changes in H3K27me3. In other words, H3K4me3 would be expected to also be bimodal. However, since H3K4me3 levels respond to changes in all H3K27me PTMs, the resulting bimodality is substantially weaker (H3K27me1 is not as bimodal as H3K27me3) and the noise added further reduces this remnant of bimodality.

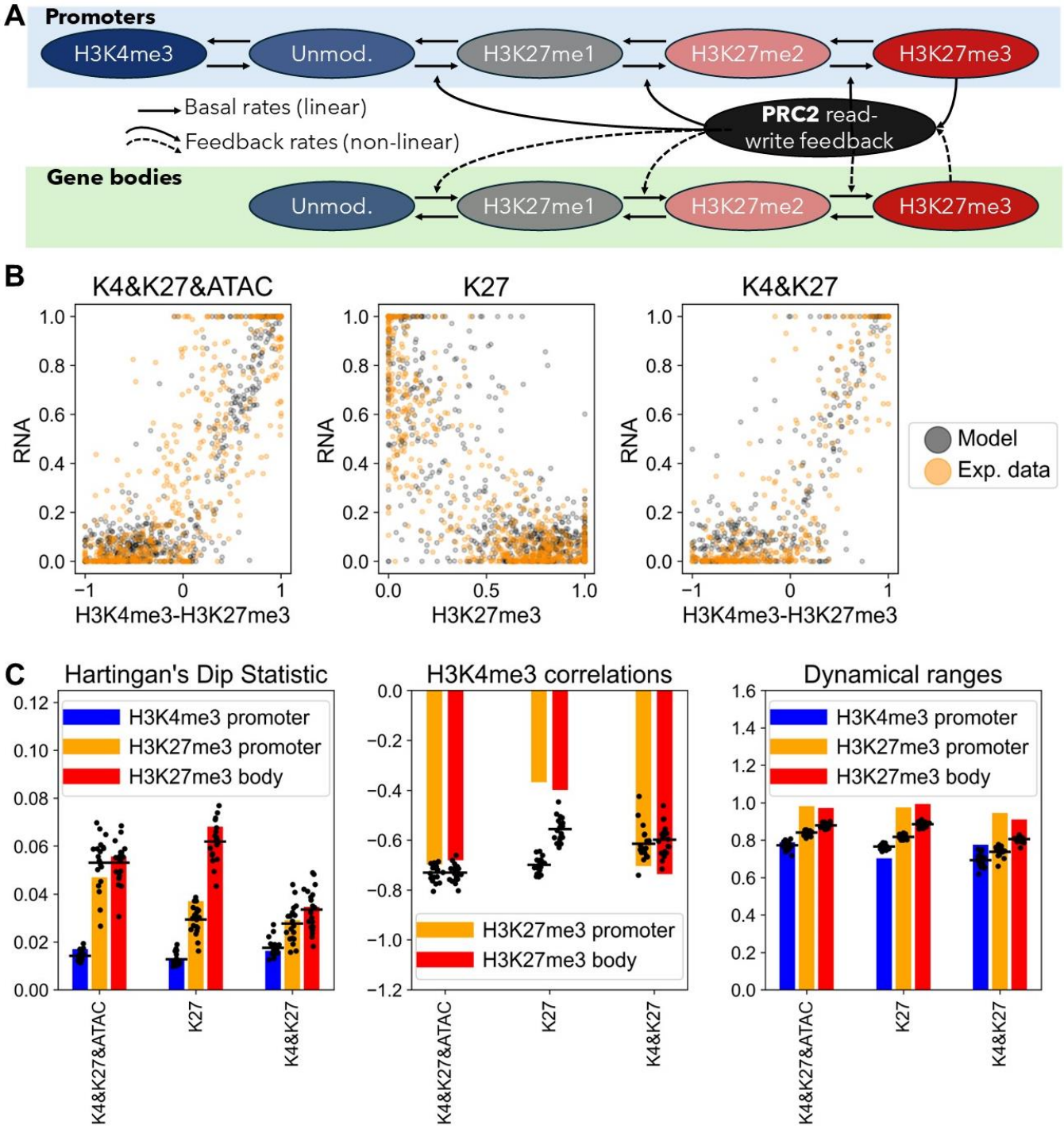

Figure 5. A mathematical model recapitulates digital H3K27me3 states. *A) Schematic of the mathematical model. B) Qualitative comparison between synthetic data (grey) and experimental data (orange), for the three regulatory modes involving H3K27me3. C) Quantitative comparison between model and experimental data for the same regulatory modes as in B). From left to right: The first panel depicts the HDS in experimental data (bars) and the model (black points, one for each of the 20 stochastic simulations, and black line for the mean of simulations). HDS is shown for three different combinations of histone mark and region of the gene. Middle: Correlation between H3K4me3 at promoters and H3K27me3 at promoters or gene bodies, for three different regulatory modes. Right: Dynamical ranges for each of the histone marks, taken as a fraction of the difference between the maximum and the minimum value of a variable (among the ten cell lines), divided by the maximum.*

In Fig. 5C we also show the dynamical ranges of these variables in the model and experiments (with relatively good agreement) and the correlation between H3K4me3 at promoters and H3K27me3 at gene

bodies or promoters (for each of the three clusters considered). This correlation was heavily constrained by the exclusive nature of the H3K4 and H3K27 marks, which correctly predicted the quantitative value of the anticorrelation in the 'K4&K27&ATAC' and the 'K4&K27' clusters. However, the model did not entirely recover the weaker correlation between H3K4me3 and H3K27me3 in the 'K27' cluster (especially at promoters, probably due to spatial inhomogeneities that we neglected, beyond separation into promoter and gene body). Nevertheless, the overall quantitative agreement between model and experimental data is very good, especially considering the relatively few parameters in the model and the varied set of quantitative outputs it explains.

Moreover, scatter plots relating epigenetic marks and transcription show that there is a qualitative agreement between the synthetic data (grey) and experimental data (orange) for these regulatory modes (Fig. 5B). Thus, the model qualitatively recapitulated the majority of the data in the H3K27me3 clusters although, due to the choice of E as a cell line-dependent parameter, it could not explain the minority of genes that were repressed in three cell lines and expressed in the other seven (see next section).

Taken together, the model highlights the critical role of the feedback rate in generating switch-like behaviour in histone PTM regulation. Promoters therefore exhibited a lower HDS than gene bodies, with the extent of this difference depending on the coupling c between promoter and gene body. Crucially, the model was capable of creating synthetic data, quantitatively and qualitatively similar to the experimental data, by assuming a modest modulation of E, the parameter that controls the overall H3K27me levels. Indeed, one of the most important consequences of the switch-like regulation of histone PTMs is that a modest variation in a parameter can cause a substantial change in histone PTM levels.

Altered H3K27me3 regulation at specific genes in hiPSC lines

Given that we saw marked transcriptional variations among cell lines (Fig. 2), we next asked if these differences were uniformly distributed across regulatory modes or whether they could be attributed to just a subset of them. To address this, we obtained heatmaps of the RNA expression (z-score, to standardise the mean and variance of each gene independently) of all ten cell lines for all genes in each cluster, using hierarchical clustering to organise genes and cell lines (Fig. 6A and Supplementary Fig. 7A). The 'K27' cluster shows distinctive behaviour with a very clear difference between *Kucg2*, *Sojd3* and *Yoch6* and the remaining hiPSC lines, which have opposite expression patterns in the vast majority of genes (Fig. 6A). Similar behaviour for these three cell lines is seen for the other two clusters involving H3K27me3 (Supplementary Fig. 7A). Although a similar trend was also apparent in the 'K4&ATAC' cluster, it was substantially less pronounced than in H3K27me3-associated clusters. We calculated the absolute value of the pairwise Pearson correlation of gene expression between two cell lines to assess which regulatory modes show more robust transcriptional differences. Higher values indicate stronger correlation or anticorrelation and, thus, stronger differences between cell lines. This analysis confirmed that the three clusters involving H3K27me3 show the strongest patterns (Fig. 6B), suggesting that this histone PTM plays a key role in the regulation of transcriptional variability in hiPSC lines. In Fig. 2D, we already saw a similar genome-wide relation for H3K27me3; here, we dissected the genes that contribute the most to this relation (see Supplementary File 1 for a list of DEGs per regulatory mode).

In addition, we compared the transcriptomes of hiPSCs with that of a conventional human embryonic stem cell (hESC) line (H9 line, from the ENCODE database [45]). The Euclidean distance calculated on z-scored expression values did not show significant differences in the distance between the hiPSC lines and the hESCs when taking all DEGs into account (Supplementary Fig. 7B). However, when dividing the DEGs into the different regulatory modes, striking differences appeared (Fig. 6C). In particular, the 'K27' cluster showed a very marked pattern that mirrored the genome-wide transcriptional heterogeneity shown in Fig. 2A (for completeness, all of the clusters are shown in Supplementary Fig. 7C). Furthermore, the only clusters where *Yoch6, Sojd3* and *Kucg2* were the three most distant lines from hESCs were those involving H3K27me3, suggesting aberrant Polycomb regulation in these cell lines. The fact that the cell lines with

diverging H3K27me3 regulation had also low (or null) efficiency to give rise to PGCLCs is unlikely to be a coincidence (p<0.04, see Methods).

Overall, these findings suggest that the chromatin mark H3K27me3 plays a major role in transcriptional heterogeneity between hiPSC lines. Moreover, genes regulated by H3K27me3 exhibited particularly clear differences between *Kucg2*, *Sojd3* and *Yoch6,* and the remaining hiPSC cell lines. These three cell lines also showed similar differentiation efficiencies for both PGCLC, neuronal and definitive endoderm fates. Crucially, we found that, by comparison with a hESC reference line, their H3K27me3 regulation was aberrant—a fact that was not apparent from a genome-wide perspective.

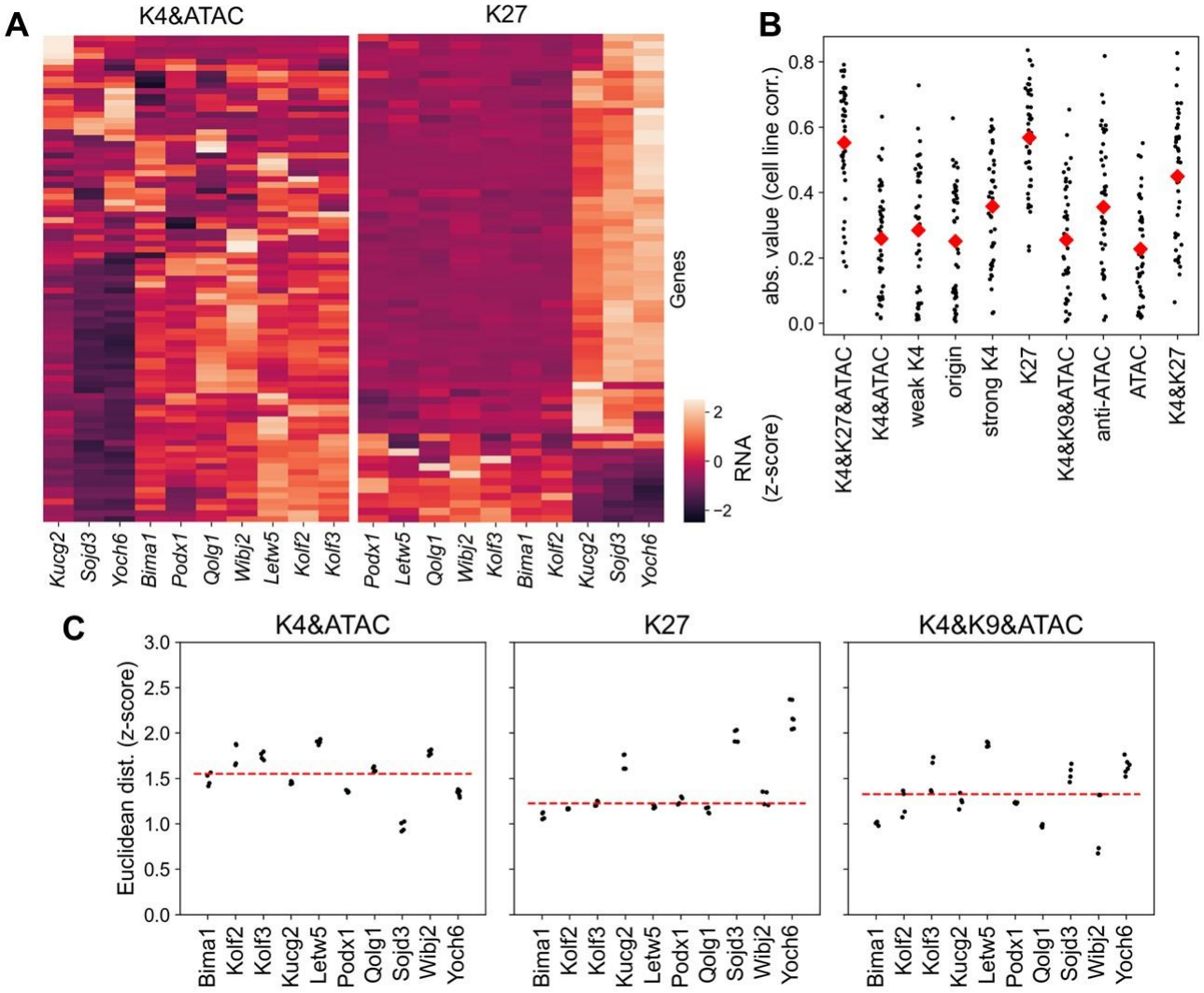

***Figure 6**. **Heterogeneity within regulatory modes between cell lines.** A) Heatmap of the transcriptional output (z-score), with purple/black representing repressed genes for a given cell line, and light orange colours representing expressed ones. B) Quantification of cell-line dependent effects, in terms of the absolute value of the pairwise correlations between cell lines, considering only genes within a given regulatory mode. Within each regulatory mode (x axis), each pairwise correlation is depicted by a black dot and the average value across all 45 possible pairwise correlations is given by the red diamond. C) Euclidean distance, in terms of the z-scores, between the hiPSC lines and H9 hESCs, using only the genes from a given regulatory mode (specified in the title of each plot). The red dashed line denotes the median distance from hiPSC lines to the hESC line.*

Discussion

In this work, we have used a combination of experimental and computational techniques to uncover the chromatin-based mechanisms that underpin transcriptome differences between hiPSC lines. We also evaluated the differentiation potential of ten hiPSC lines for two different cell fates, finding highly variable efficiencies among lines and fates. Strikingly, each cell line was capable of differentiating into at least one of the target cell types (Fig. 1C and Supplementary Fig. 3C). By profiling the transcriptome and three different histone PTMs in these lines and integrating these data with chromatin accessibility measurements,

we obtained a unique epigenomic dataset. We analysed this dataset using a SVM-based computational pipeline to dissect the different regulatory modes and concluded that *1)* the modes involving H3K27me3 displayed the most consistent differences (Fig. 6A, B), *2*) three cell lines displayed aberrant locus specific H3K27me3 regulation (taking hESCs as reference) and *3)* regulatory modes involving H3K27me3 had the capacity to create two distinct states, suggesting a digital form of histone PTM regulation (Fig. 4C, D). Crucially, this bimodality was tightly linked to the repression of the associated genes, as it divided the datapoints not only between high and low H3K27me3 levels, but also between higher and lower expression levels (Fig. 4D).

Furthermore, our mathematical modelling revealed that this digital mode of regulation can be explained by the read-write feedback of Polycomb complexes, a positive feedback loop that propagates the H3K27me3 mark and creates two discrete states (ON/OFF). To our knowledge, this is the first instance of genome-wide quantitative modelling of digital Polycomb regulation. Previous modelling studies concerned with digital regulation have either focused narrowly on a single locus [46-48], or performed genome-wide quantitative modelling of datasets without discriminating between unimodal and bimodal behaviour [20]. Nevertheless, digital regulation of chromatin has been shown for certain genes in yeast [49, 50], plants [51] and mouse neural progenitor cells [52]. Here, we showed that a similar behaviour occurs in some genes in hiPSCs, and that it can impact transcriptional output.

Importantly, our mathematical model showed that even slight alterations to the parameters for the overall H3K27 methylation rates can cause dramatic changes in histone PTM levels. The fact that a small dysregulation of Polycomb silencing could cause such a marked effect in a subset of genes is consistent with the observation that the hiPSC lines in this study display highly similar genome-wide transcription and chromatin regulatory landscapes, with pronounced differences in both restricted to specific genomic loci. One potential candidate for this dysregulation is EPOP, which is expressed at higher levels in the cell lines with altered gene expression and chromatin profiles in H3K27me3 regulatory modes (Fig. 2D and Fig. 6A, C). EPOP is the only Polycomb accessory protein that was differentially expressed among the studied cell lines. EPOP has been shown to modestly reduce the binding of Polycomb to chromatin [53, 54], but, according to the model, this modest effect may be sufficient to completely switch the transcriptional regulation of specific target genes.

Altogether, this work has shown that hiPSC lines exhibit heterogeneous, fate-dependent differentiation potential. In fact, the four cell lines that failed to give rise to PGCLCs were found to efficiently differentiate into DE. Heterogeneity in H3K27me3 regulation correlated with this variation in differentiation potential. Specifically, three hiPSC lines (*Kucg2*, *Sojd3* and *Yoch6*) showed inefficient PGCLC induction, suggesting impaired developmental potential, concomitant with altered H3K27me3 profiles. In contrast, the fourth cell line (*Letw5*) initially succeeded in inducing PGCLCs, but then failed to maintain PGC fate. This suggests that, in this case, a distinct mechanism, independent of H3K27me3 dysregulation, may result in impaired germ cell development. Taking into account the impact of repressive histone modifications on transcription, and the number of differentially expressed transcription factors among the hiPSC lines, we propose that the dysregulation of H3K27me3 levels at specific gene loci is one important factor that influences the diverging differentiation potential between hiPSC lines. However, we also note that genetic variations between donors [9, 10], DNA methylation [5], or sex [55] may contribute to these differences. Nevertheless, emerging evidence emphasising the role of bivalent marks in regulating pluripotency supports our findings [12, 56, 57], further reinforcing the hypothesis that histone modifications modulate the capacity of pluripotent stem cells to differentiate into target cell types.

Methods

hiPSC culture. The human induced pluripotent stem cell (hiPSC) lines used in this study (HPSI1113i-bima_1, HPSI0114i-kolf_2, HPSI0114i-kolf_3, HPSI0214i-kucg_2, HPSI0514i-letw_5, HPSI1113i-podx_1, HPSI1113i-qolg_1, HPSI0314i-sojd_3, HPSI0214i-wibj_2, and HPSI0215i-yoch_6) were purchased from the Human Induced Pluripotent Stem Cells Initiative (HipSci; https://www.hipsci.org/). All cell lines were deposited by the Wellcome Trust Sanger Institute (https://www.sanger.ac.uk/), and banked at European Collection of Authenticated Cell Cultures (ECACC). ECACC's quality control testing methods are

accredited in accordance with the recognized International Standard to ISO/IEC 17025:2005. All hiPSC lines were purchased from Culture Collections of the UK Health Security Agency (https://www.culturecollections.org.uk/) and tested negative for mycoplasma contamination.

hiPSCs were cultured on vitronectin-coated (Gibco, A14700) dishes in TeSR-E8 medium (Stem Cell Technologies, #05991). The medium was changed daily, and cells were passaged as clumps every 4-5 days using Versene (Gibco, 1504033) or 0.5mM EDTA solution (Life Technologies, AM9260G) in PBS. 10µM Y-27632 2HCl (Selleckchem, S1049) was added for 1 day after passage.

Pre-ME differentiation. Induction of human PGCLCs was performed as previously described [21]. Briefly, confluent hiPSCs were washed once with 1x PBS (Gibco, 20012-019) and then trypsinised with 0.25% Trypsin-EDTA (Gibco, 25200056). The cells were then incubated at 37°C for 2.5-3 min and resuspended in MEF medium (DMEM/F-12 (Gibco, 21331-020), 10%FBS (Gibco, 10270106), 2mM L-Glutamine (Gibco, 25030081), 100 U/ml Penicillin-Streptomycin (Gibco, 15140122)) to stop the reaction. The cells were passed through a 35 µm cell strainer (Falcon, 352235) to remove the clumps and the resulting suspension was centrifuged at 200g for 4 min. The dissociated cells were then seeded in vitronectin-coated 12-well plate at a density of 200 000 cells/well. Cells were cultured in mesendoderm medium (aRB27 medium (Advanced RPMI 1640 (Gibco, 12633012), 1% B-27 supplement (Gibco, 17504001), 1x MEM NEAA (Gibco, 11140-035), 2mM L-Glutamine, 100 U/ml Penicillin-Streptomycin) supplemented with 100ng/mL Activin A, 3µM CHIR and 10 µM Y-27632 2HCl) for 12h.

PGCLC differentiation. For PGCCL induction, dissociated Pre-ME cells were washed with PBS, trypsinised for 2,5-3 min at 37°C, resuspended in MEF, passed through a strainer, centrifuged at 200g for 4 min, and resuspended in PGCLC medium (aRB27, 500 ng/ml BMP4 (R&D Systems, 314-BP), 10 ng/ml hLIF (SCI), 100 ng/ml SCF (R&D Systems, 455-MC-010), 50 ng/ml mEGF (R&D Systems, 2028-EG-200), 10 µM Y-27632 2HCl). The cells were then seeded in ultra-low-attachment 96-well plates (Costar, 7007 or Greiner, 650979) at a density of 5000 cells/well and the resulting EBs were cultured for 4 days.

DE differentiation. DE induction was performed as previously described [21]. Briefly, Pre-ME cells were cultured in ME medium for another 12 h (24 h in total), then the cells were washed once with PBS and the medium was changed to DE medium (aRB27, 100ng/mL Activin A, 0.5 µM LDN212-854). The DE cells were collected 2 days after the medium change.

EB processing, cryosectioning and immunofluorescence. For PGCLC staining, EBs were collected and washed with PBS. For fixation, EBs were incubated in 4% (w/v) paraformaldehyde (Agar Scientific, AGR1026) for 20 min at room temperature. EBs were then washed three times in PBS and incubated in 10% sucrose (Carl Roth, 57-50-1) at 4°C for either 1 h or overnight, followed by incubation in 20% sucrose at 4°C for 1h. EBs were then embedded in OCT (Cell Path, 6478.2), incubated for 30 min at 4°C and frozen at -80°C until further use. For cryosectioning, a Leica cryostat was used, and 8µm sections of EBs were collected onto charged slides (Epredia, J7800AMNT) which were stored at -80°C until further use. Slides were first washed three times for 5 min in PBS and then incubated in permeabilisation buffer (PBS, 1% BSA (Carl Roth, 8076.2), 0.1% Triton X-100 (Thermo Fisher, BP151-100)) for 30 min in a humidified chamber. The slides were then incubated with primary antibodies (Supp. Table 2) overnight at 4°C. The next day, the slides were washed three times in PBS and then incubated with secondary antibodies (Supp. Table 2) for 1 h at room temperature. Slides were then washed twice in PB, incubated with DAPI and washed again with PB before mounting with Vectashield (H-1000, Vector Laboratories). Microscopy was performed using an LSM confocal microscope (Zeiss, LSM 980). For hiPSC or DE characterisation, the cells were grown on vitronectin-coated 8-well chamber slides (Ibidi, 80821) (hiPSCs) or on glass coverslips (DE cells) and the slides were processed as described above. The Object Scan plugin (https://github.com/orgs/gurdon-institute/repositories; [58]) was used in Fiji to quantify the images.

Fluorescence-activated Cell Sorting (FACS). FACS was performed as previously described with some minor modifications [21]. Briefly, human EBs were harvested on day 4, washed with PBS and incubated in 0.25% Trypsin/EDTA in a thermocycler at 850 rpm at 37°C for 5-12 min. The reaction was then stopped with 3% FBS in PBS and the cell suspension was pipetted until dissociated into single cells and then passed through the strainer. The cells were then centrifuged, and the cell pellet was resuspended in 3% FBS in PBS. Dissociated cells were then stained for EpCAM-APC (Biolegend, 324208) and ITGalpha6-BV421 (Biolegend, 313624) for 1 h in the dark on ice. Afterwards, the cells were washed in 3% FBS in PBS and subjected to FACS using SONY Cell Sorter SH800Z. FACS data were analysed using FlowJo software.

RNA-seq. For each cell line, ~ 1 million human iPS cells were prepared per biological replicate. Cells were harvested and washed in Dulbecco's Phosphate Buffered Saline (DPBS, Cytiva HyClone, SH30028.02). Total RNA was purified using the Monarch Total RNA Miniprep Kit (NEB, T2010S). An aliquot of the total RNA was taken for integrity analysis using the Agilent RNA 6000 Pico Kit (Agilent, 5067-1513) on a 2100 Bioanalyzer (Agilent). The rest of the total RNA was stored at -80°C until library preparation. RNA-seq libraries were generated using the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (NEB, E7760L), NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB, E7490L) and NEBNext Multiplex Oligos for Illumina (Index Primers Set 1, 2, 3 and 4, NEB, E7335S, E7500S, E7710S, E7730S), using 500 ng total RNA as starting material for each preparation, and 8 final PCR cycles to obtain sequencing-ready libraries. Library concentration and quality were analysed using the Agilent High Sensitivity DNA Kit (Agilent, 5067-4626) on a 2100 Bioanalyzer (Agilent). The RNA-seq libraries were sequenced in HiSeq2500-RapidRun 50bp Paired End runs on the HiSeq sequencing platform (Illumina).

CUT&Tag. CUT&Tag was performed using the EpiCypher protocol as described previously[20] with minor modifications. hiPSCs were washed once with DPBS (Life Technologies; 14190144) and incubated with Accutase (StemCell Technologies; 07922) for 5 min at 37°C. After centrifugation for (3 min at 300 x g at RT), the cells were resuspended in 1X DPBS and counted. 2 x 10^5

cells per CUT&Tag reaction were centrifuged for 3 min at 300 x g at RT and then resuspended in 100 µl/sample of cold NE Buffer (20 mM HEPES–KOH, pH 7.9, 10 mM KCl, 0.1% Triton X-100, 20% Glycerol, 0.5 mM Spermidine (Sigma-Aldrich, 05292), 1x Roche cOmplete^{TM}, Mini, EDTA-free Protease Inhibitor) and incubated for 10 min on ice. After centrifugation (3 min at 600 x g at RT), the nuclei were resuspended in cold NE Buffer to achieve a final concentration of $1x10^6$ nuclei/ml. 1 x 10^5 nuclei per reaction were subjected to CUT&Tag. A total volume of concanavalin A (ConA) beads (EpiCypher, 21-1401) of 11 µl per sample was transferred to a 1.5 ml tube for batch processing. The ConA beads were washed twice on a magnetic stand with 100 µl/sample Bead Activation Buffer (20 mM HEPES, pH 7.9, 10 mM KCl, 1 mM $CaCl_2$, 1 mM $MnCl_2$) and resuspended in a final volume of 11 µl per sample. 10 µl beads were aliquoted into separate tubes for each sample and incubated with 100 µl of nuclei for 10 min at RT. The tubes were placed on a magnet separator (DynaMag-PCR; ThermoFisher Scientific), and the supernatant was discarded. The beads were then resuspended in 50 µl of cold Antibody150 Buffer (20 mM HEPES, pH 7.5, 150 mM NaCl, 0.5 mM Spermidine, 1x Roche cOmplete^{TM}, Mini, EDTA-free Protease Inhibitor, 0.01% digitonin (Sigma-Aldrich, D141), 2mM EDTA) containing primary antibody in a 1:50 dilution (anti-H3K9me3, Active Motif 39062; anti-H3K27me3, Active Motif; 39157; dilution 1:50; anti-H3K4me3, Active Motif; 39160). Samples were incubated overnight at 4°C on an orbital shaker. On the following day, the tubes were placed on a magnet separator, and the supernatant was discarded. The beads were incubated in Digitonin150 Buffer (20 mM HEPES, pH 7.5, 150 mM NaCl, 0.5 mM Spermidine, 1x Roche cOmplete^{TM}, Mini, EDTA-free Protease Inhibitor, 0.01% digitonin) containing 0.5 µg of secondary antibody (anti-rabbit Epicypher; 13-0047) for 1 hour at RT. The beads were then washed twice with Digitonin150 Buffer and then resuspended in 50 µl of Digitonin300 Buffer (20 mM HEPES, pH 7.5, 300 mM NaCl, 0.5 mM Spermidine, 1x Roche cOmplete^{TM}, Mini, EDTA-free Protease Inhibitor, 0.01% digitonin), and 1.25 µl of CUTANA pAG-TN5 (EpiCypher, 15-1017) was added. After incubation at RT for 1 hour, the tubes were placed on a magnet and washed twice with Digitonin300 Buffer.

The beads were then resuspended in 50 µl of chilled Tagmentation Buffer (Digitonin300 Buffer, 10 mM $MgCl_2$) and incubated for 1 hour at 37°C in thermocycler. The tube was placed on a magnetic separator and the supernatant was discarded. The beads were washed once with 50 µl RT TAPS Buffer (10 mM TAPS pH 8.5, 0.2 mM EDTA). The beads were resuspended in 5 µl of RT SDS Release Buffer (10 mM TAPS, pH 8.5, 0.1% SDS) and vortexed at maximum speed for 10 seconds followed by a brief centrifugation. The beads were then incubated for 1 hour at 58°C in a thermocycler. Following incubation, 15 µl of RT SDS Quench Buffer (0.67% Triton-X 100 in molecular grade H_2O) was added to the beads and vortexed at maximum speed for 20 seconds, followed by a brief centrifugation.

For library amplification, 1 µl of each barcoded i5 and i7 primers (10µM stock) was used and PCR was performed using CUTANA High Fidelity PCR mix (EpiCypher, 15-1018), using the following PCR programme: 58°C for 5 min, 72°C for 5 min, 98°C for 45 s, 11-13 cycles of 98°C for 15 s, 60°C for 10 s, 72°C for 1 min. Libraries were purified twice using 1X volume of AMPure XP Beads (Beckman Coulter; A63881) and resuspended in 15 µl of 0.1X TE buffer. Following quality control using the Agilent High Sensitivity DNA Kit (Agilent, 5067-4626) on a 2100 Bioanalyzer (Agilent), the CUT&Tag libraries were multiplexed and sequenced (paired-end 150 bp) on a NovaSeq 6000 instrument (Illumina).

Sequencing data processing. Raw RNA-seq reads were assessed for quality using FastQC v0.11.8 (www.bioinformatics.babraham.ac.uk/projects/fastqc). Reads were aligned to the human reference genome GRCh38 with STAR v2.5.2b. PCR duplicates were removed using Picard v2.9.0 (https://broadinstitute.github.io/picard/). Gene-level quantification was performed using HTSeq-count v0.10.

Raw ATAC-seq reads were quality-checked using FastQC v0.11.5. Adapter sequences and low-quality bases were removed with Cutadapt v1.17. Reads were aligned to the GRCh38 genome using Bowtie2 v2.3.4.2. Alignments were filtered for mapping quality (MAPQ > 20) with SAMtools v1.9 and custom scripts. PCR duplicates, mitochondrial reads, and reads in ENCODE blacklisted regions were excluded using Picard and SAMtools. Accessible chromatin regions were identified using MACS2 v2.1.2 with the parameters -f BEDPE --keep-dup all.

CUT&Tag data were processed and aligned to the GRCh38 reference genome as previously described [28]. After alignment, reads with mapping quality below 30 were filtered out using SAMtools v1.9.

Data. The data used in this work included the transcriptome of 10 iPSC cell lines, obtained by RNA-seq. We obtained transcript counts for all ten cell lines (two independent replicates for each cell line and three in the case of Yoch6). We focused our efforts on protein-coding genes of autosomes (to avoid confounding effects that may take place in sex chromosomes). The input for the differential gene expression analysis are the raw counts per gene. For every other analysis of the transcriptome, the magnitude used was transcripts per million (TPMs), since this metric normalises by the length of each gene, making it dimensionless and more reliable [59].

In terms of the epigenome, CUT&Tag was performed for the histone marks H3K27me3, H3K4me3 and H3K9me3 in all these lines, with two replicates for each. In addition, results from ATAC-seq assays were obtained from Ref. [29]. We used the bigwig files (normalised to reads per genomic content, i.e. 1×genomic coverage) from independent replicates to establish the enrichment of a given epigenetic mark above the genome-wide average (area under the curve) at specific genomic features, namely: genes (from gene start to gene end), promoters (defined as between 1 kbp upstream and downstream of the gene start) and gene bodies (the entirety of the gene except for the first kbp). Finally, we also used transcript counts at the transcript level to identify the most used transcription start site (TSS) in each gene. Then, these TSSs were also used to quantify enrichment in a window of 1 kbp

upstream and downstream from it, similar to the procedure for promoters. Thus, it is these four types of enrichment quantifications that will be used in the data analysis below. In the case of the dataset used for the SVM, we quantified the enrichment using each of these four annotations but considering the merged bigwig file from all independent replicates.

Where appropriate, we used the ENSEMBL annotation for the human genome (hg38, release 110).

PCA. For the PCA analysis of the transcriptome, we first computed the TPMs for each gene. TPM values were normalised to unit variance (so that every gene has similar importance) and then used to plot the PCA diagram, with the coordinates of each datapoint being specified by the normalised gene TPMs for a given replicate in a particular cell line. Similar procedures were used to plot the PCA diagrams of histone marks and accessibility assays. The only difference, in these cases, is that we used the enrichment quantifications for each genomic feature (separately) obtained from the epigenomic assays instead of TPM values.

Transcriptome scatter plots. For the transcriptome scatterplots we took the TPM values of each cell line and computed the logarithm (base 2) of the TPM value and added a constant to avoid taking the logarithm of 0, for cases with zero counts. We plotted on the X axis the results for one cell line and in the Y axis the results for another one, and we computed the Pearson correlation in between the two datasets. We note that the resulting Pearson correlation between cell lines is quite sensitive to the value of the constant added: the smaller this value, the smaller the correlation. Unlike in other studies, here we chose a smaller value for this constant (0.01). Still, the Pearson correlation is quite high in most cases (over 0.96, see Supplementary Fig. 4A).

Differential expression analysis. We have data for 10 cell lines, and we are interested in any difference between any pair of cell lines. Thus, we need to perform multiple pairwise comparisons between all cell lines, to obtain a final list of differentially expressed genes (DEGs) that will be flagged for further downstream analysis. When performing multiple pairwise comparisons the resulting threshold p-value for statistical significance needs to be adjusted (decreased) to take into account the larger false discovery rate of performing multiple pairwise comparisons. For this, we used a Benjamini-Hochberg (BH) adjustment method to increase the p-value (to account for the larger false discovery rate).

Usually log2-fold changes (L2FCs) are also applied, to ensure the genes tagged as differentially expressed show substantial changes in transcription. However, the same adjustment procedure to control for multiple pairwise comparisons is not available for L2FCs. Instead, with the aim of reducing the false discovery rate, we chose a high threshold for L2FCs. The thresholds we used were 0.01 for the BH-adjusted p-values and 3 for the L2FCs. Therefore, for a gene to be considered as DE, there must be at least one pairwise comparison where both the adjusted p-value is below 0.01 and the L2FC is above 3. Additionally, we filtered out genes that had less than three samples (out of 21) with more than 10 counts. With these thresholds and filters, we identified around 700 strongly DEGs. For this analysis we used R (v 4.4.2) and DESeq2 [60] (v 1.46.0).

Gene ontology analysis. We performed a gene ontology analysis for the genes we found to be differentially expressed. We used the clusterProfiler package (v 3.0.4) and enrichGO function [61] on the 700 DEGs. The universe, or background genes against which to test the enrichment hypothesis, is the same set of genes that pass the filtering for differential expression analysis, that is, around 15,000. The p-value adjusting method was also the BH method and we tested both the molecular function and the biological process ontologies.

Comparison between hiPSC lines and hESCs. As a measure of similarity between hiPSC lines and hESC lines, we computed the Euclidean distance between the ten iPSC lines analysed here and an hESC line (H9, from the ENCODE database [45], experiment ENCSR712BRU). We took the Euclidean distances between transcriptomes in terms of the z-score. Z-score standardises means and variances for each gene independently (considering the variability among hiPSC and hESC lines), which implies that the measure is not sensitive to basal rates (as the correlation is) and it focuses only on differences. We normalised the distances by a factor of $\sqrt{N}$, where N is the number of genes considered in each instance, to control for the increase of dimensionality and maintain an stabilised metric across analyses.

Likelihood of random overlap between lines with altered H3K27me3 regulation and differentiation potential. Given that four cell lines had low PGCLC differentiation efficiency and three cell lines displayed altered transcription and H3K27me3 regulation, we wanted to know the probability that, if three cell lines are chosen at random, these three will be part of the four cell lines with impaired PGCLC differentiation. If all groups of three are equally random, we just need to count the number of cases where groups of three would fall within a group of four (4 choose 3) and divide by the total number of groups of three we can create with ten cell lines (10 choose 3). Then, the probability is:

$$p = \binom{4}{3} \Big/ \binom{10}{3} \cong 0.033.$$

Support Vector Machine analysis. SVMs are supervised classification methods, which, in their simplest form and given a set of training datapoints (each annotated with one of two possible classes), produce a hyperplane that separates the datapoints according to their class, maximising the distance from the closest datapoints to the hyperplane (also known as the margin)[62]. This technique enables two outcomes: to predict whether a gene will be highly or lowly expressed from its histone marks/accessibility, and to classify the different genes according to their regulatory mechanism. In addition, our dataset (10 very similar cell lines with transcriptional and epigenomic information) is well suited for this approach since it enables us to analyse each DEG separately (with 10 datapoints per gene), allowing us to capture the specific characteristics of each target. 10 datapoints per gene, in a four-dimensional space (accessibility and three histone marks), would usually be too small a dataset, but the robustness of SVMs nevertheless enables this approach with satisfactory precision. The procedure outlined below was performed on a per-gene basis.

Binarization of expression. The starting point is a TPM value per DEG per cell type (averaged across replicates in each cell line). For each gene, its TPM values across cell lines were binarized, that is, whether this specific gene is highly or lowly expressed in each cell line. To that end, we used a K-means algorithm [63] to cluster those 10 TPM values into zeroes and ones. To get the best description of the data, we employed both logarithmic and linear scales, i.e., we performed the K-means clustering for the TPM values in both logarithmic scale and linear scale. To decide which of the scales should be used in each case, we quantified how good the clustering was in each of the two scales and chose the better performing one. Inertia is closely related with the goodness of the clustering and is given by the sum of squares of the data to the centre of the cluster. Thus, we can compare the goodness of the clustering in between the two scales by computing, for each scale, the ratio between the inertia when clustering into two clusters over the inertia when clustering into a single cluster, as a measure of how much better the two clusters (high and low expression) are with respect to a single cluster (around the mean across all cell lines). If this ratio is smaller in one scale than in the other, it implies that the clustering is more accurate in that first scale. In general, this clustering is performed using a logarithmic scale since changes are more pronounced in terms of fold-changes than absolute differences. However, some genes have TPMs of zero for some cell lines, which makes the logarithmic scale invalid and, thus, we use the linear scale for these genes. Finally, it is worth noting that, despite this binary classification being a simplification of reality, it produces a reasonable classification that can be fed into the SVM, resulting in high accuracy predictions (see below).

SVM description. We built a support vector machine (SVM) that classifies RNA levels (high versus low expression) based on histone marks and accessibility data. For each gene and each cell line, we fed the SVM a value of H3K4me3, H3K9me3, H3K27me3 and ATAC. For the repressive marks (H3K9me3 and H3K27me3), we used the gene annotation (from gene start to gene end as annotated by ENSEMBL). For accessibility and H3K4me3 data, given that these profiles display much more pronounced and narrow peaks, we followed an alternative approach: we took the annotation that produced the maximum value for each gene among the following three: 1 kbp upstream and downstream of the gene start, 1 kbp upstream and downstream of the most used TSS, and the average over the whole gene. Usually, the whole gene average is the smallest metric, but we included it for cases where we failed to find the peak at either the gene start or TSS. This input data, together with the binarized RNA expression metric, is enough to train a SVM for each gene independently. For added robustness in the method, it is convenient to normalise all of the input data to values between zero and one. Thus, if $x_{i,k}^{j}$ is the enrichment of the *k*-th histone mark (or ATAC) around gene *i* for cell line *j*, the normalised quantity that is fed into the SVM is given by

$$\tilde{x}_{i,k}^{j} = \frac{x_{i,k}^{j} - \min_{j} x_{i,k}^{j}}{\max_{j} x_{i,k}^{j} - \min_{j} x_{i,k}^{j}}.$$

We define the scale factor as

$$s_{i,k} = \left(\max_{j} x_{i,k}^{j} - \min_{j} x_{i,k}^{j}\right)^{-1},$$

for later convenience. Once the SVM (linear kernel) is trained with the normalised data, the result is a hyperplane that separates highly expressed from lowly expressed cell lines in a four-dimensional space for the normalised values of the epigenomic variables at the gene of interest for each cell line. This is the first outcome of the SVM: a classification method to predict high/low gene expression based on its epigenetic marks/accessibility. In terms of parameters, we used a linear kernel (resulting in the hyperplane that divides high from low expression) and a high *C* parameter (*C* = 100). The *C* parameter controls the strength of the penalty for misclassification. A low *C* parameter would try to maximise the margin (distance between the hyperplane and the nearest point) instead of maximising classification accuracy. For an example of the application of the SVM, see the Supplementary Materials.

SVM as a prediction tool: binary classification. To test the performance of the SVM (and the entire analysis pipeline, from DEG identification to SVM training, passing through transcription binarization) as a predictor of transcription, we included 9 cell lines in the training set and sought to predict the transcription of the 10th (e.g., Qolg1) based on its epigenomic variables and the trained SVM. In this case, the binarization of expression is performed with the nine cell lines in the training data set and whether the 10th cell line is high or low expression (ground truth for this test) will depend on which cluster centre its expression is closest to. The accuracy of the SVM in this particular test is always above 70%, even in genes where there is little histone mark or ATAC

signal, increasing to around 90% when predicting genes with clear histone mark trends. However, it should be noted that, being a binary classification, 50% accuracy corresponds to random guesses.

Comparison with logistic regression. For benchmarking purposes, we sought to compare our SVMs with a more standard logistic regression approach. The comparison between algorithms was performed in terms of the area under the receiver operating characteristic (AUROC) curve. The ROC curve displays the true positive and false positive rates for each threshold in the binary classification and the higher the AUROC (bounded from above by 1), the more accurate the classifier.

In this context, the SVM is required to produce scores (rather than binary predictions), scores which can be interpreted as the probability that a given gene is highly expressed. To that end, we used Platt's algorithm [33] to fit the output of the SVM to a sigmoidal curve that represents a probability function, which is then evaluated at the test datapoints to obtain a probability score for each prediction. Usually, the same SVM (same parameters for the hyperplane and fit to sigmoidal) would be used to predict multiple datapoints, resulting in a vector of scores (one for each datapoint), from which we can obtain the ROC curve. In this case, due to the limited datapoints per gene, we decided to train a SVM in each gene with 9 datapoints and seek to predict the 10th. Therefore, in order to obtain a vector of scores for the ROC curve, we independently trained the SVMs for each gene (as described above) and predicted a score for the 10th datapoint in each gene. We obtained a vector of scores by pooling all genes together.

The ROC curve represents the true positive rate (number of correctly predicted positives over the total number of ground truth positives) as a function of the false positive rate (number of negatives incorrectly predicted as positives over total number of ground truth negatives). With the vector of scores from our classifier (and a vector of ground truths for the test dataset) we can obtain the ROC curve by varying the score threshold (from 0 to 1) above which a prediction would be considered positive. When the score is greater than the threshold the prediction is considered positive, allowing us to obtain a true positive rate and a false positive rate. Collectively, this set of points (one for each different threshold) in the true positive rate vs false positive rate space makes the ROC curve. It should be noted that, given that we are pooling scores from different genes and SVMs together, the ROC curve we are plotting is not the usual ROC curve, obtained from evaluating many datapoints with the same SVM. However, it still serves as a collective measure of the goodness of classification of the overall procedure.

We are now in a position to compare logistic regression to the SVM, in terms of the AUROC. We chose the logistic regression method as a benchmark, since it is a standard and efficient linear classification method. As with the SVM, we sought to predict the transcriptional outcome of the 10th cell line from the data of the remaining 9, on a per-gene basis (same preprocessing of the data as described above for the SVM procedure). Then, the ROC curve is obtained by pooling together a prediction from each gene.

SVM as an explanatory tool: coefficients and clustering. For the second outcome (identification of regulatory mechanisms), we leverage the coefficients of the hyperplane defined by the SVM procedure. The coefficients of the hyperplane define the vector normal to the surface, which in this case and given our conventions, is the direction of upregulation of transcription. Thus, in our four-dimensional case, for each gene *i*, we obtain $\boldsymbol{n}_i = (n_{i,K4}, n_{i,K9}, n_{i,K27}, n_{i,ATAC})$. The combination of the hyperplane coefficients (how epigenetic marks/accessibility relate to transcription) and the dynamic ranges (how much variability each epigenetic mark/accessibility has in each gene) can be very informative for the regulatory logic, as they contain information on the nature of the relationship with transcription and on the magnitude of the effect. Nevertheless, in its raw form, as an output of the SVM, the coefficients of the SVM analysis cannot be directly used and need to undergo some preprocessing.

We need a quantity that incorporates both the directionality information of the SVM hyperplane, combined with the dynamical range of the original data (in logarithmic scale). The inclusion of the dynamical range is important to gauge the strength of the effect, e.g., a dynamical range just above the genome-wide mean for that mark is unlikely to have a strong effect in transcription, even if, due to some spurious correlation/noise, the SVM normal vector points in that direction. To that end, first, each of the four coefficients that define the hyperplane was normalised using the summed square of the four coefficients. This allows us to ascertain the relative contribution of each variable and permits an interpretation as a vector within a sphere of radius one. The normalised hyperplane coefficients are:

$$\tilde{n}_{i,k} = \frac{n_{i,k}}{\sqrt{\sum_k n_{i,k}^2}}.$$

Second, the information lost when normalising to feed epigenomic data to the SVM needs to be recovered. To that end, we take the logarithm of the inverse of the scale factor $s_{i,k}$ (if $s_{i,k} > 1$ we take the log to be zero to avoid negative results). If $s_{i,k} \geq 1$, the dynamical range is the same as the genome-wide mean or smaller (due to the normalisation of the raw data), implying that any changes in that mark, at that gene, are relatively small and we therefore neglect such changes. This leads to the quantity we are looking for, the log-rescaled coefficient:

$$R_{i,k} = \tilde{n}_{i,k} \log\left(s_{i,k}^{-1}\right),$$

which combines the information of the SVM with the dynamical ranges in logarithmic scale, while preserving the signs of the SVM vector, which are informative of transcriptional output. Note that, while the epigenomic data was normalised by $s_{i,k}$ before

feeding it to the SVM, the hyperplanes coefficients had not (i.e., this is not directly undoing an earlier operation). Finally, to avoid biasing our analysis towards quantities that (due to their own nature) show greater dynamical ranges, we normalise the maximum of log-rescaled coefficients across all genes, for each histone mark/accessibility assay: $\widetilde{R_{\iota,k}} = \frac{\tilde{n}_{i,k} \log(s_{i,k}^{-1})}{\max_i[\tilde{n}_{i,k} \log(s_{i,k}^{-1})]}$.

The rationale for using the logarithm of the normalisation constant is that we believe that fold changes are more instructive than absolute differences. The reason for using the normalised hyperplane coefficients rather than their logarithm is that the hyperplane coefficients might be negative, so cannot themselves be recast on a logarithmic scale. In addition, the signs of the hyperplane coefficients should be kept, since they specify whether a given epigenetic mark is activating (positive) or repressing (negative) with respect to transcription.

After the processing of the SVM outputs, we can now proceed to clustering DEGs into clusters according to their regulatory logic. To that end, we use again a K-means algorithm for clustering the log-rescaled coefficients $\widetilde{R_{\iota,k}}$. We clustered the genes into 2-20 clusters (1000 independent random initialisations) and examined different measures for the goodness of clustering (Silhouette score, Calinski-Harabasz score and Davies-Bouldin score). The Calinski-Harabasz score identified 4 as the optimal number of clusters, while the Davies-Bouldin and Silhouette scores identified 5 as the optimal cluster number (see Supplementary Fig. 5A). However, the Silhouette and Davies-Bouldin scores also had another local optimum for 10 or 11 clusters, with similar scores as for the cluster of 5. Given that clustering into only 4 or 5 clusters is too coarse for the downstream analyses we are aiming for, we chose 10 as the number of clusters to keep.

Z-score RNA cluster plot. We plotted the RNA TPM values of the DEGs in each cluster for all 10 cell lines. Using Seaborn's clustermap function, the genes and cell lines were hierarchically clustered and plotted (with the mean expression of each gene set to 0 and the variance normalised to 1, i.e., using the z-score).

Bimodality. To ascertain whether the datasets we are working with are bimodal in the different regulatory contexts, we computed the Hartingan's dip statistic (HDS) for each histone mark in each of the different clusters. The HDS measures the distance from the empirical distribution to the closest unimodal distribution. Thus, the higher the statistic, the further away it is from a unimodal distribution. To compute an empirical HDS for each cluster from a given set of genes, it is necessary to normalise the dynamical range of each variable at each gene to the [0,1] interval. When normalising each gene into the [0,1] interval, there will be one value that is mapped to 0 and one that is mapped to 1 (out of a total of 10 values). When pooling together dozens of genes, this results in an artificial over-representation of 0s and 1s that could bias the results of bimodality metrics. To avoid this issue, for each gene, one 0 value and one 1 value is dropped from the dataset (in total, 2 values per gene are dropped). Pooling the resulting values within each cluster, we obtain an empirical distribution for which an HDS value can be obtained. We note that this procedure (since it discards the extremes for each gene) results in a conservative metric for bimodality.

The scripts in this section were written and executed in Python 3.10.11, using scikit-learn [63] version 1.4.1, pandas version is 2.2.1, numpy version is 1.26.2, scipy version 1.12.0 and seaborn version is 0.13.2.

Mathematical model. The model describes the fraction of lysine residues 4 and 27 of histone 3 that are methylated in two different regions: promoter (±1kbp around the gene start) and gene body (from +1kbp of the gene start to the gene end). The definitions of these annotations follow those of the main text and the computational analysis of experimental data.

For lysine residue 27, since it has been shown to be important for theoretical modelling, we consider four states: H3K27me0, H3K27me1, H3K27me2 and H3K27me3. Methylation of the H3K27 residue is possible both at the promoter and gene body (variables for promoter and gene body are distinguished by a subscript, p or b, respectively). For lysine residue 4, we describe two states: either unmodified (H3K4me0) or trimethylated (H3K4me3). We only consider two states for H3K4, since there is little evidence for any feedback in these states and methylation at this residue occurs on a faster timescale [64]. We assume methylation of residue H3K4 is only possible at promoter regions, as H3K4me3 in the gene body is usually negligible. In addition, we assume that methylation at H3K4 and at H3K27 are mutually exclusive, since they cannot take place on the same histone tail [65].

The fractions of histones with each modification are represented by $[\cdot]$ and their quantitative dynamics are given by:

$$\frac{\mathrm{d}[\mathrm{K4me3}]_x}{\mathrm{d}t} = k_a\,[\mathrm{K27me0}]_x - k_{da}[\mathrm{K4me3}]_x \tag{1}$$

$$\frac{\mathrm{d}[\mathrm{K27me0}]_x}{\mathrm{d}t} = k_{da}[\mathrm{K4me3}]_x - k_a[\mathrm{K27me0}]_x + k_{dm}\,[\mathrm{K27me1}]_x - [\mathrm{K27me0}]_x\left(k_{0,x} + F_x\,k_f\right)E \tag{2}$$

$$\frac{\mathrm{d}[\mathrm{K27me1}]_x}{\mathrm{d}t} = k_{dm}[\mathrm{K27me2}]_x - k_{dm}[\mathrm{K27me1}]_x + ([\mathrm{K27me0}]_x - [\mathrm{K27me1}]_x)\left(k_{0,x} + F_x\,k_f\right)E \tag{3}$$

$$\frac{\mathrm{d}[\mathrm{K27me2}]_x}{\mathrm{d}t} = k_{dm}[\mathrm{K27me3}]_x - k_{dm}[\mathrm{K27me2}]_x + ([\mathrm{K27me1}]_x - [\mathrm{K27me2}]_x)\left(k_{0,x} + F_x\,k_f\right)E \tag{4}$$

$$\frac{\mathrm{d}[\mathrm{K27me3}]_x}{\mathrm{d}t} = k_{dm}[\mathrm{K27me3}]_x + [\mathrm{K27me2}]_x\left(k_{0,x} + F_x\, k_f\right) E, \tag{5}$$

where $[\mathrm{K4me3}]_x + \sum_i [\mathrm{K27me}i]_x = 1$ (due to normalisation of fractions), and

$$F_x = \frac{[\mathrm{K27me3}]_x + c\,[\mathrm{K27me3}]_y}{1+c}, \tag{6}$$

$x,y \in (p,b)$ and $x \neq y$. Thus, the only coupling between the two regions is due to the feedback term F, parametrised by c (which is always smaller than 1, ensuring that these regions interact more within themselves than with each other). The construction of the model reflects the fact that if interactions are stronger with the other region they have to be weaker within the same region, as a locus can make a limited number of contacts. In addition, the cell cycle is also simulated, by halving all histone marks by a factor of 2 (and increasing consequently the fraction of H3K27me0 histones) every period T.

The H3K27 demethylation rates are parametrised by k_{dm}. $k_{0,x}$ represents the strength of the basal rate at region x, and k_f, that of the feedback rate. k_a is the rate of H3K4 methylation, and $k_a \neq 0$ only at the promoter region. k_{da} is the rate of H3K4me3 demethylation. The parameter E is cell-line dependent. Therefore, in order to describe the different configuration in which a gene from any of the Polycomb clusters can be found, the parameter E will be varied.

Finally, similar to previous models [19], we assumed that the transcriptional response takes the form:

$$\left[f^p_{K4me3} - \frac{f^p_{K27me3} + f^b_{K27me3}}{2}\right] \cdot\ \theta\left(1 - f^b_{K27me3} - f^p_{K27me3}\right), \tag{7}$$

provided that $f^p_{K4me3} > \frac{f^p_{K27me3} + f^b_{K27me3}}{2}$ (otherwise the transcriptional output is 0). In Eq. (7), θ is the Heaviside (step) function and f denotes the normalised and noise-added outputs of the simulations (for details see below).

Generation of synthetic data. First, we obtained parameters for a deterministic simulation, based on the underlying parameters for the cluster (see Table 1). For each cluster of genes, we obtained as many sets of random parameters as genes in the cluster. The random parameters were normally distributed, with mean at the cluster values and with a standard deviation of 10% of the mean value.

For each of these parameter sets, we obtained 10 datapoints (since we have 10 cell lines), by changing the parameter E. Thus, the simulations are deterministic once the parameters are fixed. The simulations are run (according to the equations of Methods—*Mathematical model*, starting from a silent—high H3K27me3—configuration) for 20 cell cycles to equilibrate, and sampled 100 times during the 21st cell cycle. The average (for each variable) of the samples in the last cell cycle is the only value stored from each simulation. The rationale behind this choice is that, in experiments, cell cycles within the population were not synchronised and, hence, their outcome is an average of cells at different points of the cell cycle.

To simulate the spread of the data more realistically (simulating experimental noise), some noise was added to the output obtained from each simulation independently. The random variables for this noise were drawn from a normal distribution (with mean 0.1 and variance 0.05) and added to the histone mark fractions (note that these fractions are naturally bounded from above by 1). We took absolute values of this sum, for the unlikely event that it results on a negative value.

The 10 noisy datapoints that are the output of the simulation are then normalised between 0 and 1 (exactly the same procedure as described in the methods for the experimental data, to make them comparable). This procedure yields normalised outputs f, which are used to compute the transcriptional output, as per Eq. (7). Finally, we also add noise to this transcriptional output, by adding a normally distributed variable (again, with mean 0.1 and variance 0.05) and taking absolute values. Transcriptional output is also resized to span the [0,1] segment (as was done for the experimental data).

From here, any further downstream quantification (e.g., HDS) is made to compare with experimental data and, thus, follows exactly the same procedure as described in the methodology for the analysis of the experimental data.

Parameters. We set $k_{da} = 2\ \mathrm{hr}^{-1}$, to reflect the more rapid turnover of active marks (compared with the K27 demethylation rate $k_{dm} = 1\ \mathrm{hr}^{-1}$). To simulate the cell-line specific differences, the parameter E is varied from 1 to 0.8 (0.8 representing weaker Polycomb activity, e.g., *Yoch6* cell line). As stated in the main text, we assume that $k_{0,p} = 1.5k_{0,b}$ in all three clusters. All of the fixed parameters of the model are given in Table 1.

Table 1. *Parameters of the mathematical model, for each of the regulatory modes modelled. In boldface, the parameters that change depending on the regulatory mode.*

	$\boldsymbol{k_{0,p}}$ (hr^{-1})	$\boldsymbol{k_f}$ (hr^{-1})	$\boldsymbol{c}$	k_{da} (hr^{-1})	k_{dm} (hr^{-1})	k_a (hr^{-1})	T (hr)
K4&K27&ATAC	0.42	3.6	0.3	2	1	0.8	22
K27	0.42	3.6	0.03	2	1	0.8	22
K4&K27	0.52	3.2	0.3	2	1	0.8	22

Data and code availability. Data will be made publicly available upon publication and the code used to perform the SVM analysis and mathematical model is available at https://github.com/AMovillaMiangolarra/hiPSCchromatin-SVM

Acknowledgements. AMM and MH thank the BBSRC Institute Strategic Programme (BB/P013511/1) and the Wellcome Human Developmental Biology Initiative for funding. LL thanks the Biochemical Society for funding. UG was supported by a Sofja Kovalevskaja Award of the Humboldt foundation. OD was supported by the International Max Planck Research School for Genome Science. JK was supported by the International Max Planck Research School for Molecular Biology. SS was supported by a UKRI MRC Rutherford Fund Fellowship (MR/T016787/1). FT, YC, and SS were supported by the BBSRC Babraham Institute's Epigenetics Strategic Programme Grant (BBS/E/B/000C0421) and a BBSRC Flexible Talent and Mobility Account (FTMA) 3 grant.

Competing interests. SS is a co-founder, shareholder, and employee of Enhanc3D Genomics. The rest of the authors declare no competing interest.

References

1. González F, Boué S, Belmonte JCI. Methods for making induced pluripotent stem cells: reprogramming à la carte. Nature Reviews Genetics. 2011;12(4):231-42. doi: 10.1038/nrg2937.
2. Takahashi K, Yamanaka S. Induced pluripotent stem cells in medicine and biology. Development. 2013;140(12):2457-61. doi: 10.1242/dev.092551.
3. Koyanagi-Aoi M, Ohnuki M, Takahashi K, Okita K, Noma H, Sawamura Y, et al. Differentiation-defective phenotypes revealed by large-scale analyses of human pluripotent stem cells. Proceedings of the National Academy of Sciences. 2013;110(51):20569-74. doi: 10.1073/pnas.1319061110.
4. Pantazis CB, Yang A, Lara E, McDonough JA, Blauwendraat C, Peng L, et al. A reference human induced pluripotent stem cell line for large-scale collaborative studies. Cell Stem Cell. 2022;29(12):1685-702.e22. doi: https://doi.org/10.1016/j.stem.2022.11.004.
5. Nishizawa M, Chonabayashi K, Nomura M, Tanaka A, Nakamura M, Inagaki A, et al. Epigenetic Variation between Human Induced Pluripotent Stem Cell Lines Is an Indicator of Differentiation Capacity. Cell Stem Cell. 2016;19(3):341-54. doi: https://doi.org/10.1016/j.stem.2016.06.019.
6. Yokobayashi S, Okita K, Nakagawa M, Nakamura T, Yabuta Y, Yamamoto T, et al. Clonal variation of human induced pluripotent stem cells for induction into the germ cell fate†. Biology of Reproduction. 2017;96(6):1154-66. doi: 10.1093/biolre/iox038.
7. Jerber J, Seaton DD, Cuomo ASE, Kumasaka N, Haldane J, Steer J, et al. Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. Nature Genetics. 2021;53(3):304-12. doi: 10.1038/s41588-021-00801-6.
8. Choi J, Lee S, Mallard W, Clement K, Tagliazucchi GM, Lim H, et al. A comparison of genetically matched cell lines reveals the equivalence of human iPSCs and ESCs. Nature Biotechnology. 2015;33(11):1173-81. doi: 10.1038/nbt.3388.
9. Kajiwara M, Aoi T, Okita K, Takahashi R, Inoue H, Takayama N, et al. Donor-dependent variations in hepatic differentiation from human-induced pluripotent stem cells. Proceedings of the National Academy of Sciences. 2012;109(31):12538-43. doi: 10.1073/pnas.1209979109.
10. Kilpinen H, Goncalves A, Leha A, Afzal V, Alasoo K, Ashford S, et al. Common genetic variation drives molecular heterogeneity in human iPSCs. Nature. 2017;546(7658):370-5. doi: 10.1038/nature22403.
11. Kim K, Doi A, Wen B, Ng K, Zhao R, Cahan P, et al. Epigenetic memory in induced pluripotent stem cells. Nature. 2010;467(7313):285-90. doi: 10.1038/nature09342.
12. Carcamo-Orive I, Hoffman GE, Cundiff P, Beckmann ND, D'Souza SL, Knowles JW, et al. Analysis of Transcriptional Variability in a Large Human iPSC Library Reveals Genetic and Non-genetic Determinants of Heterogeneity. Cell Stem Cell. 2017;20(4):518-32.e9. doi: https://doi.org/10.1016/j.stem.2016.11.005.
13. Yokobayashi S, Yabuta Y, Nakagawa M, Okita K, Hu B, Murase Y, et al. Inherent genomic properties underlie the epigenomic heterogeneity of human induced pluripotent stem cells. Cell Reports. 2021;37(5). doi: 10.1016/j.celrep.2021.109909.
14. Streeter I, Harrison PW, Faulconbridge A, The HipSci Consortium, Flicek P, Parkinson H, et al. The human-induced pluripotent stem cell initiative—data resources for cellular genetics. Nucleic Acids Research. 2016;45(D1):D691-D7. doi: 10.1093/nar/gkw928.
15. Cheng C, Yan K-K, Yip KY, Rozowsky J, Alexander R, Shou C, et al. A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. Genome Biology. 2011;12(2):R15. doi: 10.1186/gb-2011-12-2-r15.

16. Lu W, Tang Y, Liu Y, Lin S, Shuai Q, Liang B, et al. CatLearning: highly accurate gene expression prediction from histone mark. Briefings in Bioinformatics. 2024;25(5):bbae373. doi: 10.1093/bib/bbae373.
17. Tamburri S, Rustichelli S, Amato S, Pasini D. Navigating the complexity of Polycomb repression: Enzymatic cores and regulatory modules. Molecular Cell. 2024;84(18):3381-405. doi: 10.1016/j.molcel.2024.07.030.
18. Margueron R, Justin N, Ohno K, Sharpe ML, Son J, Drury Iii WJ, et al. Role of the polycomb protein EED in the propagation of repressive histone marks. Nature. 2009;461(7265):762-7. doi: 10.1038/nature08398.
19. Berry S, Dean C, Howard M. Slow chromatin dynamics allow polycomb target genes to filter fluctuations in transcription factor activity. Cell systems. 2017;4(4):445-57. e8.
20. Newar K, Abdulla AZ, Salari H, Fanchon E, Jost D. Dynamical modeling of the H3K27 epigenetic landscape in mouse embryonic stem cells. PLOS Computational Biology. 2022;18(9):e1010450.
21. Kobayashi T, Zhang H, Tang WWC, Irie N, Withey S, Klisch D, et al. Principles of early human development and germ cell program from conserved model systems. Nature. 2017;546(7658):416-20. doi: 10.1038/nature22812.
22. Chen D, Liu W, Lukianchikov A, Hancock GV, Zimmerman J, Lowe MG, et al. Germline competency of human embryonic stem cells depends on eomesodermin†. Biology of Reproduction. 2017;97(6):850-61. doi: 10.1093/biolre/iox138.
23. Gell JJ, Liu W, Sosa E, Chialastri A, Hancock G, Tao Y, et al. An Extended Culture System that Supports Human Primordial Germ Cell-like Cell Survival and Initiation of DNA Methylation Erasure. Stem Cell Reports. 2020;14(3):433-46. doi: https://doi.org/10.1016/j.stemcr.2020.01.009.
24. Sasaki K, Yokobayashi S, Nakamura T, Okamoto I, Yabuta Y, Kurimoto K, et al. Robust In Vitro Induction of Human Germ Cell Fate from Pluripotent Stem Cells. Cell Stem Cell. 2015;17(2):178-94. doi: https://doi.org/10.1016/j.stem.2015.06.014.
25. Kanai-Azuma M, Kanai Y, Gad JM, Tajima Y, Taya C, Kurohmaru M, et al. Depletion of definitive gut endoderm in Sox17-null mutant mice. Development. 2002;129(10):2367-79. doi: 10.1242/dev.129.10.2367.
26. Séguin CA, Draper JS, Nagy A, Rossant J. Establishment of Endoderm Progenitors by SOX Transcription Factor Expression in Human Embryonic Stem Cells. Cell Stem Cell. 2008;3(2):182-95. doi: 10.1016/j.stem.2008.06.018.
27. Cuomo ASE, Seaton DD, McCarthy DJ, Martinez I, Bonder MJ, Garcia-Bernardo J, et al. Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. Nature Communications. 2020;11(1):810. doi: 10.1038/s41467-020-14457-z.
28. Kaya-Okur HS, Wu SJ, Codomo CA, Pledger ES, Bryson TD, Henikoff JG, et al. CUT&Tag for efficient epigenomic profiling of small samples and single cells. Nature Communications. 2019;10(1):1930. doi: 10.1038/s41467-019-09982-5.
29. Todd CD, Ijaz J, Torabi F, Dovgusha O, Bevan S, Cracknell O, et al. Epigenetic priming of mammalian embryonic enhancer elements coordinates developmental gene networks. Genome Biology. 2025;26(1):214. doi: 10.1186/s13059-025-03658-8.
30. Tang WWC, Castillo-Venzor A, Gruhn WH, Kobayashi T, Penfold CA, Morgan MD, et al. Sequential enhancer state remodelling defines human germline competence and specification. Nature Cell Biology. 2022;24(4):448-60. doi: 10.1038/s41556-022-00878-z.
31. Pauler FM, Sloane MA, Huang R, Regha K, Koerner MV, Tamir I, et al. H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome. Genome Research. 2009;19(2):221-33.
32. Policarpi C, Munafò M, Tsagkris S, Carlini V, Hackett JA. Systematic epigenome editing captures the context-dependent instructive function of chromatin modifications. Nature Genetics. 2024;56(6):1168-80. doi: 10.1038/s41588-024-01706-w.
33. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers. 1999;10(3):61-74.
34. Kiani K, Sanford EM, Goyal Y, Raj A. Changes in chromatin accessibility are not concordant with transcriptional changes for single-factor perturbations. Molecular Systems Biology. 2022;18(9):e10979. doi: 10.15252/msb.202210979.
35. Tu Z, Bi Y, Mao T, Wang H, Gao S, Wang Y. Discordance between chromatin accessibility and transcriptional activity during the human primed-to-naïve pluripotency transition process. Cell Regeneration. 2023;12(1):35. doi: 10.1186/s13619-023-00179-2.
36. Cardamone F, Piva A, Löser E, Eichenberger B, Romero-Mulero MC, Zenk F, et al. Chromatin landscape at cis-regulatory elements orchestrates cell fate decisions in early embryogenesis. Nature Communications. 2025;16(1):3007. doi: 10.1038/s41467-025-57719-4.
37. Hartigan JA, Hartigan PM. The Dip Test of Unimodality. The Annals of Statistics. 1985;13(1):70-84.

38. Briffa A, Menon G, Movilla Miangolarra A, Howard M. Dissecting mechanisms of epigenetic memory through computational modeling. Annual Review of Plant Biology. 2024;75.
39. Sneppen K, Ringrose L. Theoretical analysis of Polycomb-Trithorax systems predicts that poised chromatin is bistable and not bivalent. Nature Communications. 2019;10(1):2133. doi: 10.1038/s41467-019-10130-2.
40. Goodnight D, Rine J. S-phase-independent silencing establishment in Saccharomyces cerevisiae. Elife. 2020;9:e58910.
41. Schlissel G, Rine J. The nucleosome core particle remembers its position through DNA replication and RNA transcription. Proceedings of the National Academy of Sciences. 2019;116(41):20605-11. doi: 10.1073/pnas.1911943116.
42. Alabert C, Barth TK, Reverón-Gómez N, Sidoli S, Schmidt A, Jensen ON, et al. Two distinct modes for propagation of histone PTMs across the cell cycle. Genes Dev. 2015;29(6):585-90.
43. Movilla Miangolarra A, Howard M. Theory of epigenetic switching due to stochastic histone mark loss during DNA replication. Physical Biology. 2025;22(1):016005. doi: 10.1088/1478-3975/ad942c.
44. Zerihun MB, Vaillant C, Jost D. Effect of replication on epigenetic memory and consequences on gene transcription. Physical Biology. 2015;12(2):026007. doi: 10.1088/1478-3975/12/2/026007.
45. Abascal F, Acosta R, Addleman NJ, Adrian J, Afzal V, Ai R, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;583(7818):699-710. doi: 10.1038/s41586-020-2493-4.
46. Movilla Miangolarra A, Saxton DS, Yan Z, Rine J, Howard M. Two-way feedback between chromatin compaction and histone modification state explains Saccharomyces cerevisiae heterochromatin bistability. Proceedings of the National Academy of Sciences. 2024;121(16):e2403316121. doi: 10.1073/pnas.2403316121.
47. Nickels JF, Edwards AK, Charlton SJ, Mortensen AM, Hougaard SCL, Trusina A, et al. Establishment of heterochromatin in domain-size-dependent bursts. Proceedings of the National Academy of Sciences. 2021;118(15):e2022887118.
48. Lövkvist C, Mikulski P, Reeck S, Hartley M, Dean C, Howard M. Hybrid protein assembly-histone modification mechanism for PRC2-based epigenetic switching and memory. eLife. 2021;10:e66454. doi: 10.7554/eLife.66454.
49. Saxton DS, Rine J. Distinct silencer states generate epigenetic states of heterochromatin. Molecular Cell. 2022;82(19):3566-79.
50. Xu EY, Zawadzki KA, Broach JR. Single-Cell Observations Reveal Intermediate Transcriptional Silencing States. Molecular Cell. 2006;23(2):219-29. doi: https://doi.org/10.1016/j.molcel.2006.05.035.
51. Antoniou-Kourounioti RL, Meschichi A, Reeck S, Berry S, Menon G, Zhao Y, et al. Integrating analog and digital modes of gene expression at Arabidopsis FLC. Elife. 2023;12:e79743.
52. Holoch D, Wassef M, Lövkvist C, Zielinski D, Aflaki S, Lombard B, et al. A cis-acting mechanism mediates transcriptional memory at Polycomb target genes in mammals. Nature Genetics. 2021;53(12):1686-97. doi: 10.1038/s41588-021-00964-2.
53. Beringer M, Pisano P, Di Carlo V, Blanco E, Chammas P, Vizán P, et al. EPOP Functionally Links Elongin and Polycomb in Pluripotent Stem Cells. Molecular Cell. 2016;64(4):645-58. doi: https://doi.org/10.1016/j.molcel.2016.10.018.
54. Gong L, Liu X, Yang X, Yu Z, Chen S, Xing C, et al. EPOP Restricts PRC2.1 Targeting to Chromatin by Directly Modulating Enzyme Complex Dimerization. bioRxiv. 2024:2024.09.10.612337. doi: 10.1101/2024.09.10.612337.
55. D'Antonio-Chronowska A, Donovan MKR, Young Greenwald WW, Nguyen JP, Fujita K, Hashem S, et al. Association of Human iPSC Gene Signatures and X Chromosome Dosage with Two Distinct Cardiac Differentiation Trajectories. Stem Cell Reports. 2019;13(5):924-38. doi: https://doi.org/10.1016/j.stemcr.2019.09.011.
56. Sutcliffe MA, Wingett SW, Morris CAJ, Wong E, Schoenfelder S, Lancaster MA. Epigenetic restoration of differentiation competency via reversal of epiblast regionalisation. bioRxiv. 2024:2024.12.27.630149. doi: 10.1101/2024.12.27.630149.
57. Lim B, Domsch K, Mall M, Lohmann I. Canalizing cell fate by transcriptional repression. Molecular Systems Biology. 2024;20(3):144-61-61. doi: https://doi.org/10.1038/s44320-024-00014-z.
58. Murakami K, Günesdogan U, Zylicz JJ, Tang WWC, Sengupta R, Kobayashi T, et al. NANOG alone induces germ cells in primed epiblast in vitro by activation of enhancers. Nature. 2016;529(7586):403-7. doi: 10.1038/nature16480.
59. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics. 2009;26(4):493-500. doi: 10.1093/bioinformatics/btp692.
60. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;15(12):550. doi: 10.1186/s13059-014-0550-8.
61. Xu S, Hu E, Cai Y, Xie Z, Luo X, Zhan L, et al. Using clusterProfiler to characterize multiomics data. Nature Protocols. 2024;19(11):3292-320. doi: 10.1038/s41596-024-01020-z.

62. Bishop CM, Nasrabadi NM. Pattern recognition and machine learning: Springer; 2006.
63. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825--30.
64. Zheng Y, Tipton JD, Thomas PM, Kelleher NL, Sweet SMM. Site-specific human histone H3 methylation stability: fast K4me3 turnover. PROTEOMICS. 2014;14(19):2190-9. doi: https://doi.org/10.1002/pmic.201400060.
65. Voigt P, LeRoy G, Drury William J, Zee Barry M, Son J, Beck David B, et al. Asymmetrically Modified Nucleosomes. Cell. 2012;151(1):181-93. doi: https://doi.org/10.1016/j.cell.2012.09.002.