Molecular Cell, Volume *82*

Supplemental information

Single-cell profiling of transcriptome and histone modifications with EpiDamID

Franka J. Rang, Kim L. de Luca, Sandra S. de Vries, Christian Valdes-Quezada, Ellen Boele, Phong D. Nguyen, Isabel Guerreiro, Yuko Sato, Hiroshi Kimura, Jeroen Bakkers, and Jop Kind

Supplemental Figure 1

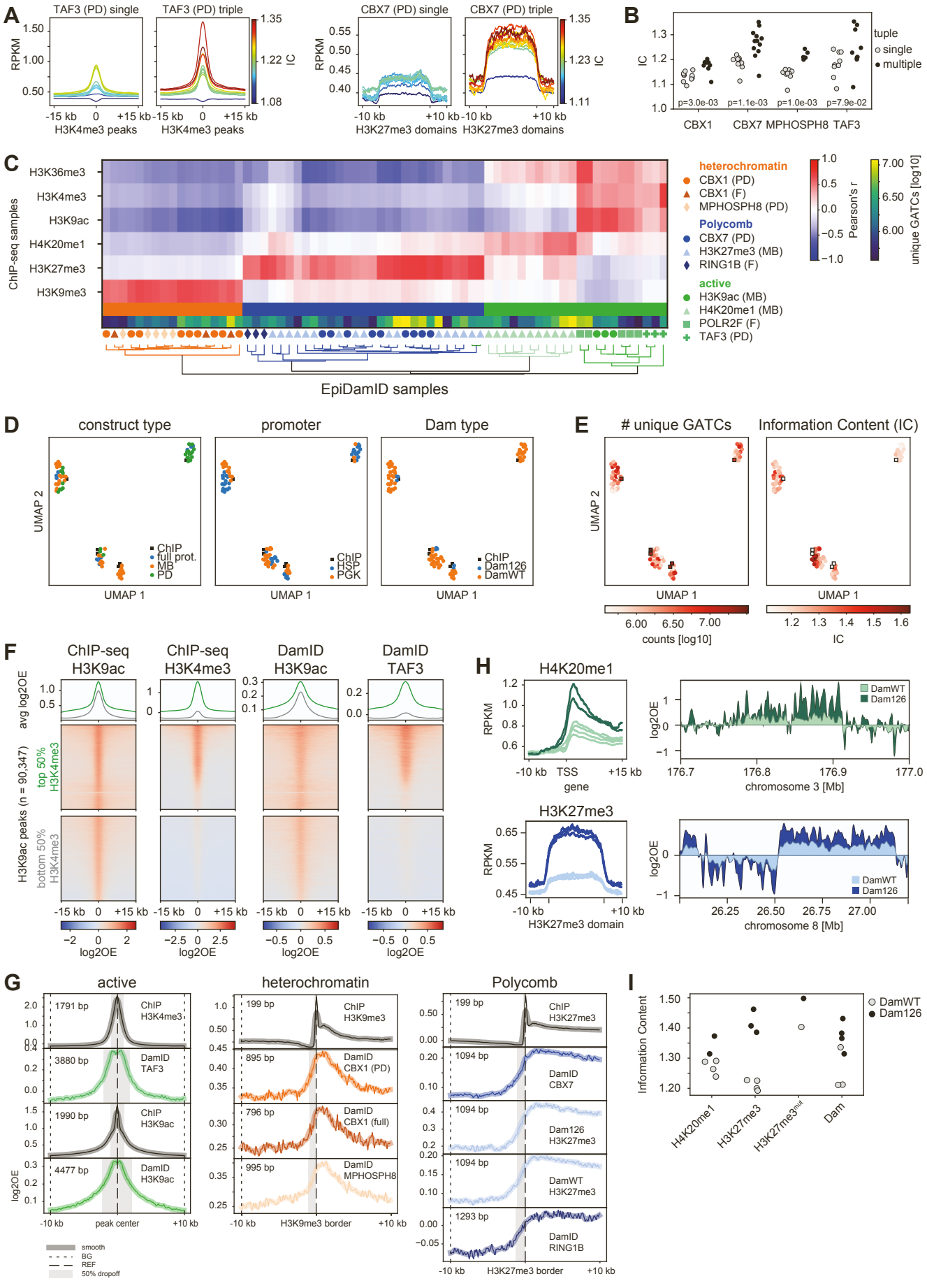

Figure S1. Technical validation of EpiDamID data, Related to Figure 1

A, Average enrichment over genomic regions of interest for TAF3 and CBX7 DamID. Left: data generated by fusing Dam to a single protein domain; Right: data generated by fusing Dam to a trimer of the same protein domain. Sample lines are colored by their Information Content (IC).

B, Strip plot of samples comparing the IC of single (grey) and multiple (black) targeting domains. Per construct, the significance was tested with a two-sided Mann-Whitney U test.

C, Clustered heatmap showing the correlation between ChIP-seq and Dam-normalized DamID. Correlations were computed using Pearson's correlation. Samples are labeled by their targeting domain (colored shapes) and number of unique GATC counts.

D, UMAPs of samples, colored by construct properties.

E, UMAPs of samples, colored by the number of unique GATC counts and IC.

F, ChIP-seq and DamID enrichment at H3K9ac peaks (center +/- 15 kb), split into two categories according to ChIP-seq H3K4me3 occupancy (highest and lowest 50%). The heatmaps show the enrichment for each peak region, while the line plots on top show the average enrichment per H3K4me3 category.

G, Signal resolution analysis. The plots show ChIP-seq and DamID enrichment at genomic regions of interest +/- 10 kb. Left panel shows active marks; signal is centered around ChIP-seq H3K4me3 and H3K9ac peaks for DamID TAF3 and H3K9ac, respectively. Middle panel shows heterochromatin; signal is centered around ChIP-seq H3K9me3 domain borders. Right panel shows Polycomb; signal is centered around ChIP-seq H3K27me3 domain borders. Solid line indicates the mean signal at these regions, shaded line indicates smoothened signal. Large dashed line indicates the location of the highest signal in ChIP (REF); small dashed line indicates the background measuring point (BG). Grey shaded area indicates the region over which the signal at the REF point drops with 50% relative to the BG point. The size of the drop-off distance is indicated in the top left.

H, Comparison of DamWT and Dam126 signal. Left: average DamID enrichment plots over genomic regions of interest. Regions are the TSS of the top 25% H3K9ac-enriched genes for H4K20me1 (top), and ChIP-seq domains for H3K27me3 (bottom). Right: genome browser views of DamID enrichment corresponding to left panels. The data shown in H represent the combined data of all samples of a particular targeting domain.

I, Strip plot of samples comparing the IC of DamWT (grey) and Dam126 (black) targeting constructs.

Supplemental Figure 2

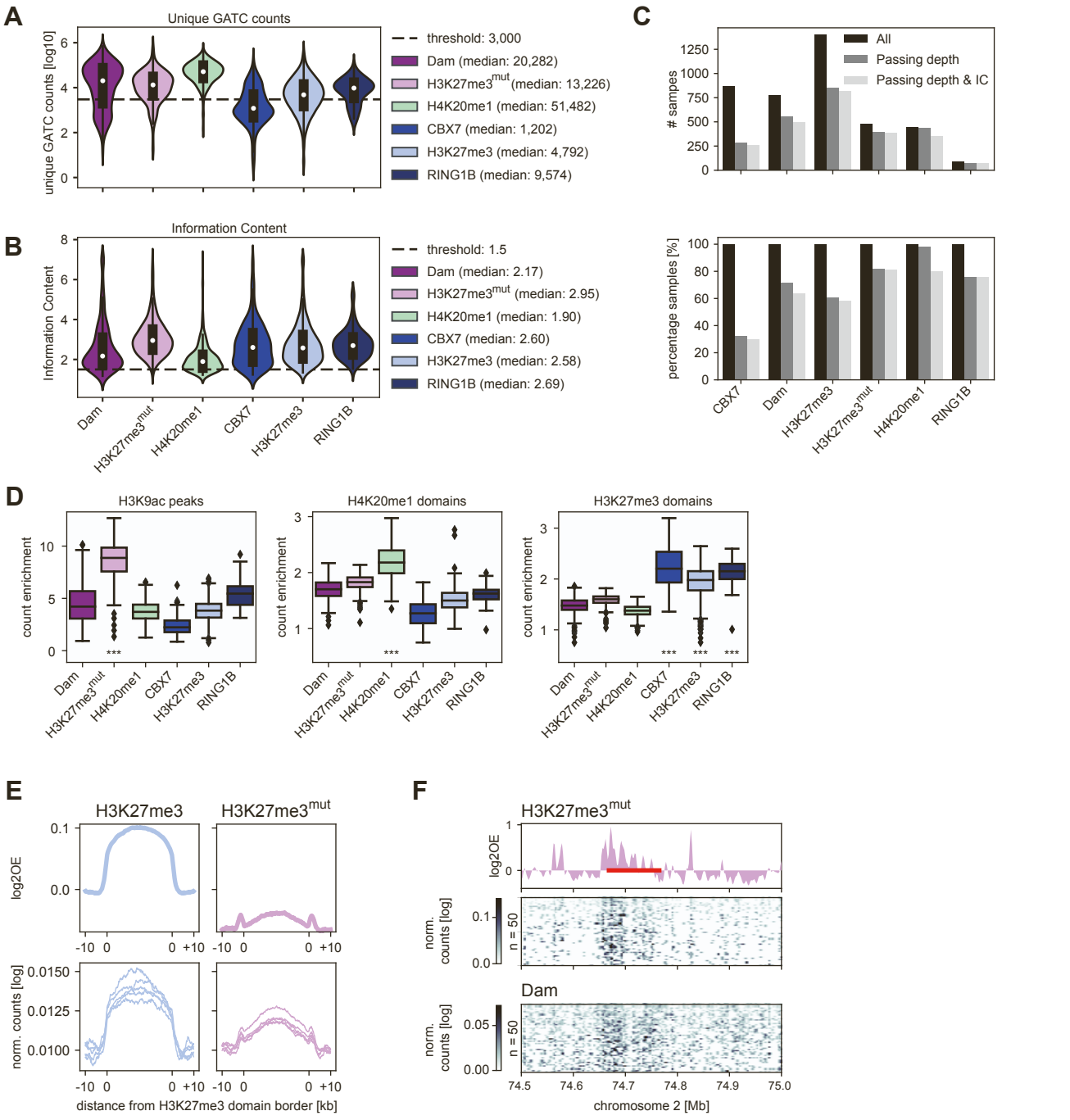

Figure S2: Detection of histone PTMs in single mouse embryonic stem cells with a single-cell implementation of EpiDamID, Related to Figure 2

A, Violin plots indicating the distribution of the number of unique GATCs detected for each cell line. The dashed line indicates the threshold used for data filtering.

B, Violin plots indicating the distribution of the Information Content (IC) after filtering on depth for each cell line. The dashed line indicates the threshold used for data filtering.

C, Overview of the number (top) and percentage (bottom) of samples retained after filtering on depth and IC.

D, Boxplots showing the count enrichment in H3K9ac ChIP-seq peaks (left), H4K20me1 ChIP-seq domains (middle), and H3K27me3 ChIP-seq domains (right) of all single cells per DamID construct. Count enrichment was computed as the fraction of GATC counts that fell within the regions, relative to the total fraction of genomic GATC positions inside these domains. In each plot, the enrichment of constructs of interest are compared to the enrichment in the Dam control. The significance of the difference was tested with a two-sided Mann-Whitney-U test. *** indicates a p-values smaller than 0.001. Constructs without an indication of significance were not tested.

E, Average signal over H3K27me3 ChIP-seq domains of H3K27me3 and H3K27me3^{mut} mintbodies. Top: *in silico* populations normalized for Dam; Bottom: five of the best single-cell samples (bottom) normalized by read depth.

F, Signal of H3K27me3^{mut} and Dam control over the *HoxD* cluster and neighboring regions. The DamID track show the Dam-normalized *in silico* populations of H3K27me3^{mut}, while the heatmaps show the depth-normalized single-cell data of the fifty richest cells for H3K27me3^{mut} and Dam. The red bar around 74.7 Mb indicates the *HoxD* cluster.

Supplemental Figure 3

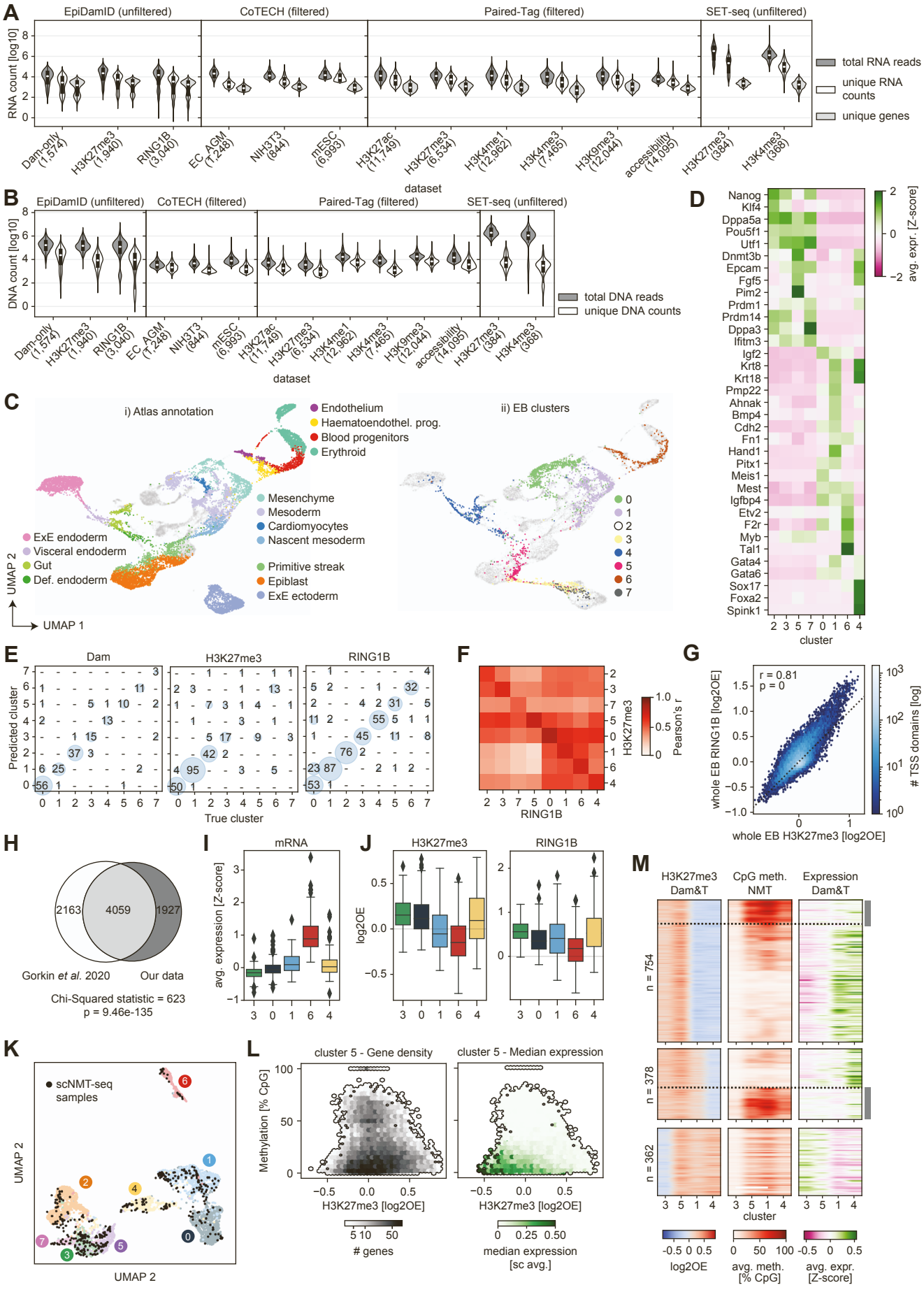

Figure S3. Validation and characterization of scDam&T-seq data in mouse embryoid bodies, Related to Figure 3

A-B, Overview of the RNA and DNA outputs for a number of datasets generated by recent single-cell multimodal omics techniques. A: The number of raw RNA-derived reads, unique transcripts and unique genes. B: The number of raw DNA-derived reads and unique counts. The included techniques are CoTECH (Xiong et al., 2021), Paired-Tag (Zhu et al., 2021) and SET-seq (Sun et al., 2021). The EpiDamID (scDam&T-seq) data shows the statistics of the embryoid body (EB) dataset. Some techniques show only the statistics of cells that passed quality thresholds ("filtered"), while others show the statistics of all obtained cells ("unfiltered"). The labels on the x-axis indicate the name of the various datasets, with the number of samples shown in paratheses. **C,** UMAPs of samples based on the integration of our EB transcription data with single-cell RNA-seq mouse embryonic data (Pijuan-Sala et al., 2019), colored by reference-annotated cell type (i) and EB-annotated cluster (ii). For atlas integration, the day 0 (i.e., mESC) time point was excluded.

D, Average expression of known marker genes. Expression was standardized over single-cells and the per-cluster average was computed.

E, Confusion plots showing the performance of the LDA classifier during training, for each construct.

F, Pearson correlation between the combined H3K27me3 and RING1B DamID signal at the TSS of all genes per transcriptional cluster.

G, Correlation of combined H3K27me3 and RING1B DamID signal at the TSS of all genes. Data of all single-cell samples passing DamID thresholds was combined for each construct. The correlation was computed using Pearson's correlation.

H, Overlap between a published set of PRC targets during mouse development (Gorkin et al., 2020) and our PRC targets. Only genes represented in both datasets could be compared. Significance of the overlap was computed with a Chi-squared test.

I, Boxplots showing the expression (averaged *Z*-score) of genes identified as significantly upregulated in cluster 6.

J, Boxplots showing the H3K27me3 (left) and RING1B (right) DamID signal at the TSS of the subset of genes shown in **I** that are PRC targets.

K, UMAPs of samples based on the integration of our EB transcription data with the transcriptional readout of the EB scNMT-seq data generated by Argelaguet *et al*. (Argelaguet et al., 2019). EpiDamID samples are colored by the transcriptional clusters determine previously; scNMT-seq samples are indicated in black.

L, Relationship between promoter CpG methylation, promoter H3K27me3 enrichment and gene expression of all genes in cells belonging to cluster 5 (epiblast-like). The left plot shows the relationship between promoter CpG methylation (+/- 2 kb around TSS) and H3K27me3 enrichment (-5 kb/+3 kb around TSS) for all genes. The right plot shows the same relationship, but the color scale indicates the median expression of genes in each region of the plot.

M, Heatmaps indicating the promoter H3K27me3 enrichment, promoter CpG methylation and gene expression for three groups of PRC targets with variable H3K27me3 enrichment. Rows are genes; columns are transcription clusters. Enrichment is shown for the 4 clusters that contained sufficient scNMT-seq samples (cluster 3: 31 cells; cluster 5: 21 cells; cluster 1: 37 cells; cluster 4: 43 cells). Genes are sorted by hierarchical cluster based on their CpG methylation levels. Examples of genes where H3K27me3 and CpG methylation complementary repress genes are indicated with a dotted line and a grey box.

Supplemental Figure 4

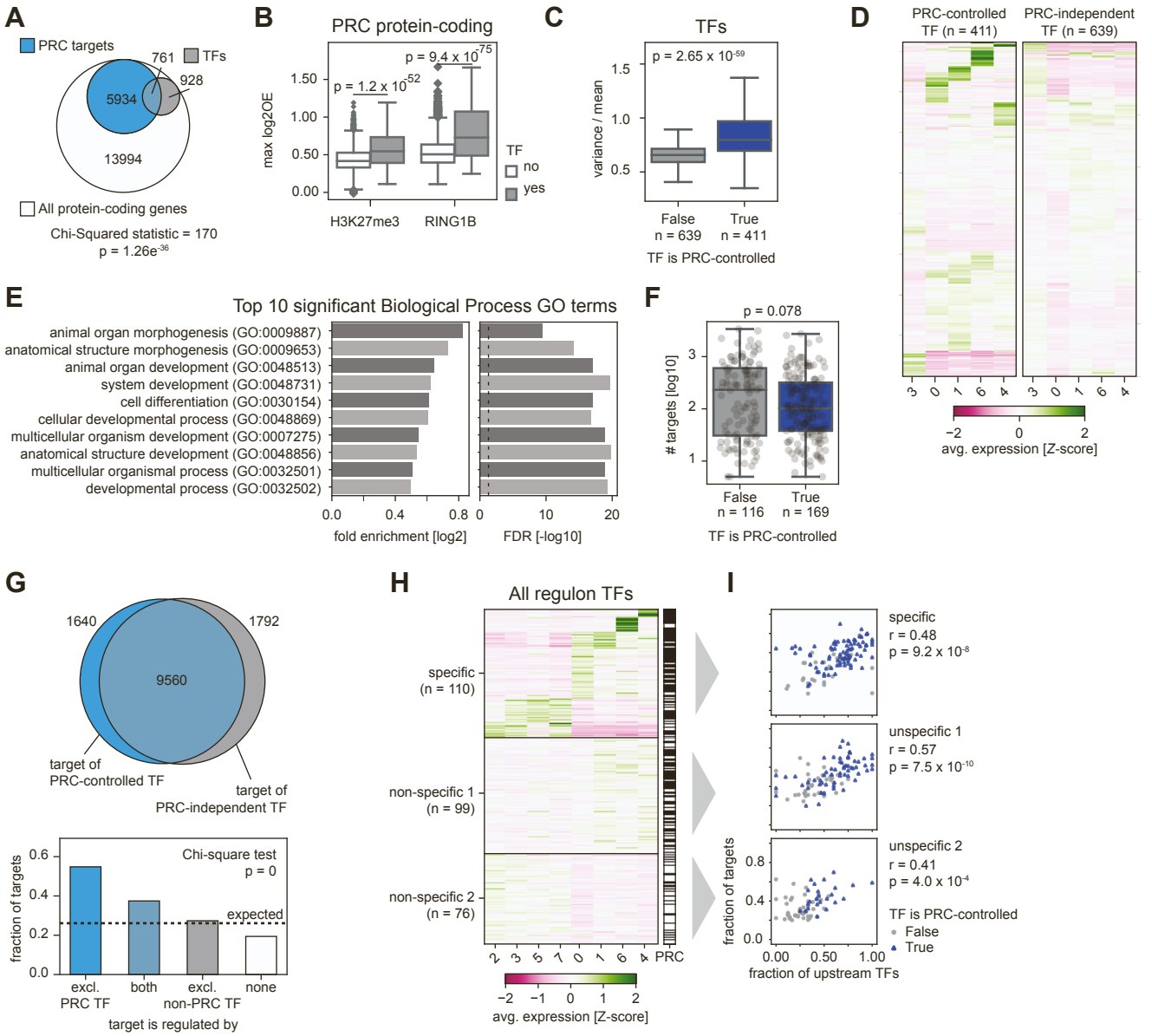

Figure S4. Characterization of the Polycomb-regulated regulatory network, Related to Figure 4

A, Venn diagram showing the overlap between PRC-controlled protein-coding genes (blue) and transcription factors (TF) (grey) in the context of all protein-coding genes (white). The significance of the overlap between PRC targets and TFs was computed using a Chi-squared test.

B, Boxplots showing the maximum observed H3K27me3 and RING1B DamID signal across transcriptional clusters for PRC-controlled TFs (grey) and the remaining PRC-controlled protein-coding genes (white). The significance of the difference between TFs and other genes was tested with a two-sided Mann-Whitney-U test.

C, Quantification of variability in gene expression of PRC-regulated and PRC-independent TFs (only expressed genes are included). Boxplots show variance over mean across all single cells. Significance was computed using a two-sided Mann-Whitney U test.

D, Clustered heatmaps showing mRNA expression (averaged Z-score) per cluster, of Polycomb-regulated TFs (left) and Polycomb-independent TFs (right). Only expressed genes are included in this plot.

E, The ten most significant Biological Process GO terms between PRC-controlled and PRC-independent TFs.

F, Number of targets of each regulon TF, split by whether or not the TF is PRC-regulated. The significance of the difference between the two groups was tested with a two-sided Mann-Whitney U test.

G, Top: Venn diagram displaying the overlap between genes that are targets of a PRC-controlled TF (blue) and genes that are targets of a PRC-independent TF (grey).

G, Bottom: Bar plot showing the fraction of targets in each category that is PRC-regulated. The dotted line indicates the expected fraction, i.e., the fraction of all genes that is a PRC target. A Chi-square test was performed to evaluate whether the deviation from the expected frequencies is significant.

H, Clustered heatmap showing mRNA expression (averaged *Z*-score) per cluster, of all regulon TFs, grouped by lineage-specific or non-specific genes. TFs are annotated as PRC-controlled (black) or PRC-independent (white).

I, Scatter plot showing the relationship between the fraction of Polycomb-controlled targets and regulators of a regulon TF. Regulon TFs that are PRC controlled are indicated in blue; regulon TFs that are PRC independent are indicated in grey. Regulon TFs are split based on the groups indicated in **H**. Correlation was computed using Pearson's correlation.

Supplemental Figure 5

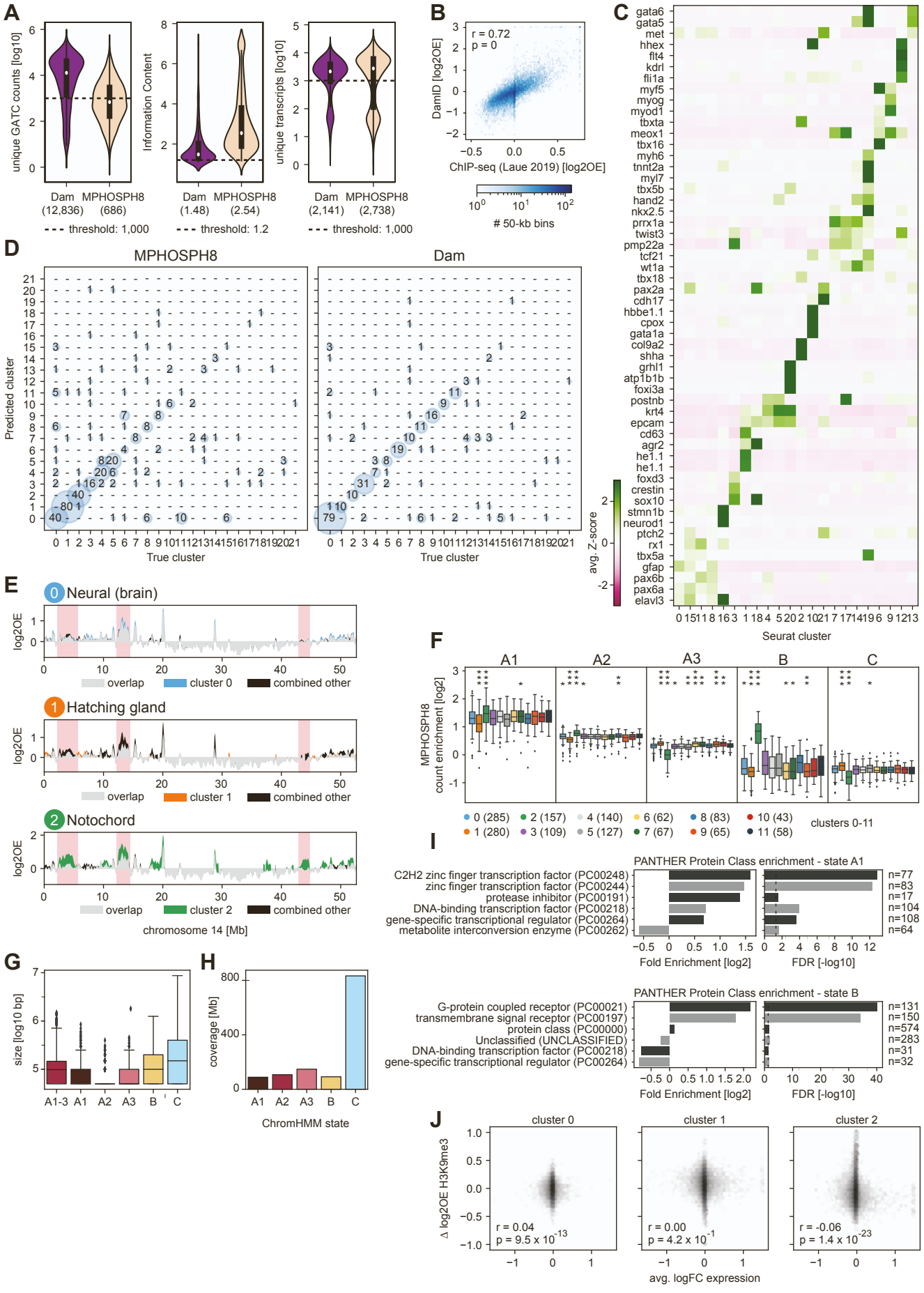

Figure S5: Characterization of transcriptomic clusters and associated genomic H3K9me3 enrichments, Related to Figure 5

A, Violin plots showing the total number of unique GATC counts, the information content (IC) and total number of unique transcripts obtained for all cells in the zebrafish dataset.

B, Comparison of our data with a published H3K9me3 ChIP-seq dataset of the 6-hpf zebrafish embryo (Laue et al., 2019). All single-cell MPHOSPH8 and Dam samples were combined to generate an *in silico* whole-embryo data set; DamID data is the log2OE of MPHOSPH8 signal over Dam is shown; ChIP-seq is the log2OE of H3K9me3 over input control. The correlation was computed using Pearson's correlation.

C, Expression of marker genes over all clusters, ordered by cell type. The average single-cell Z-scores are shown.

D, Confusion plots showing the performance of the LDA classifier during training, for each construct.

E, Genomic H3K9me3 signal over chromosome 14. For clusters 0-2, the cluster-specific signal (color) is compared to the combined signal from all other clusters (black). Each set indicates the overlay, where overlapping regions are colored grey.

F, Boxplots showing the enrichment of counts within genomic regions belonging to each of the five ChromHMM states for all cells belonging to transcriptional clusters 0-11. Count enrichment was computed as the fraction of GATC counts that fell within the regions, relative to the total fraction of genomic GATC positions inside these domains. Per state, the count enrichment of a cluster was compared to the enrichment of cells in all other clusters using a two-sided Mann-Whitney-U test. *** = $p < 0.001$; ** = $p < 0.01$; * = $p < 0.1$; no indication means the result was insignificant.

G, Distribution of domain sizes per ChromHMM state and for states A1-3 combined.

H, Total genomic coverage per ChromHMM state.

I, PANTHER protein-class enrichments (Mi et al., 2013) for genes found in state A1 (top) and B (bottom).

J, Plot displaying the relationship between differential gene expression and differential H3K9me3 enrichment. The x-axis shows the average log-foldchange in gene expression of cells in one cluster relative to all other cells; the y-axis shows the differential log2OE H3K9me3 at these genes of one cluster relative to all other cells. H3K9me3 at a gene was measured as the log2OE value of the 50-kb genomic window containing the TSS of the gene. The relationship between the variables was tested with a Pearson's correlation test.

Supplemental Figure 6

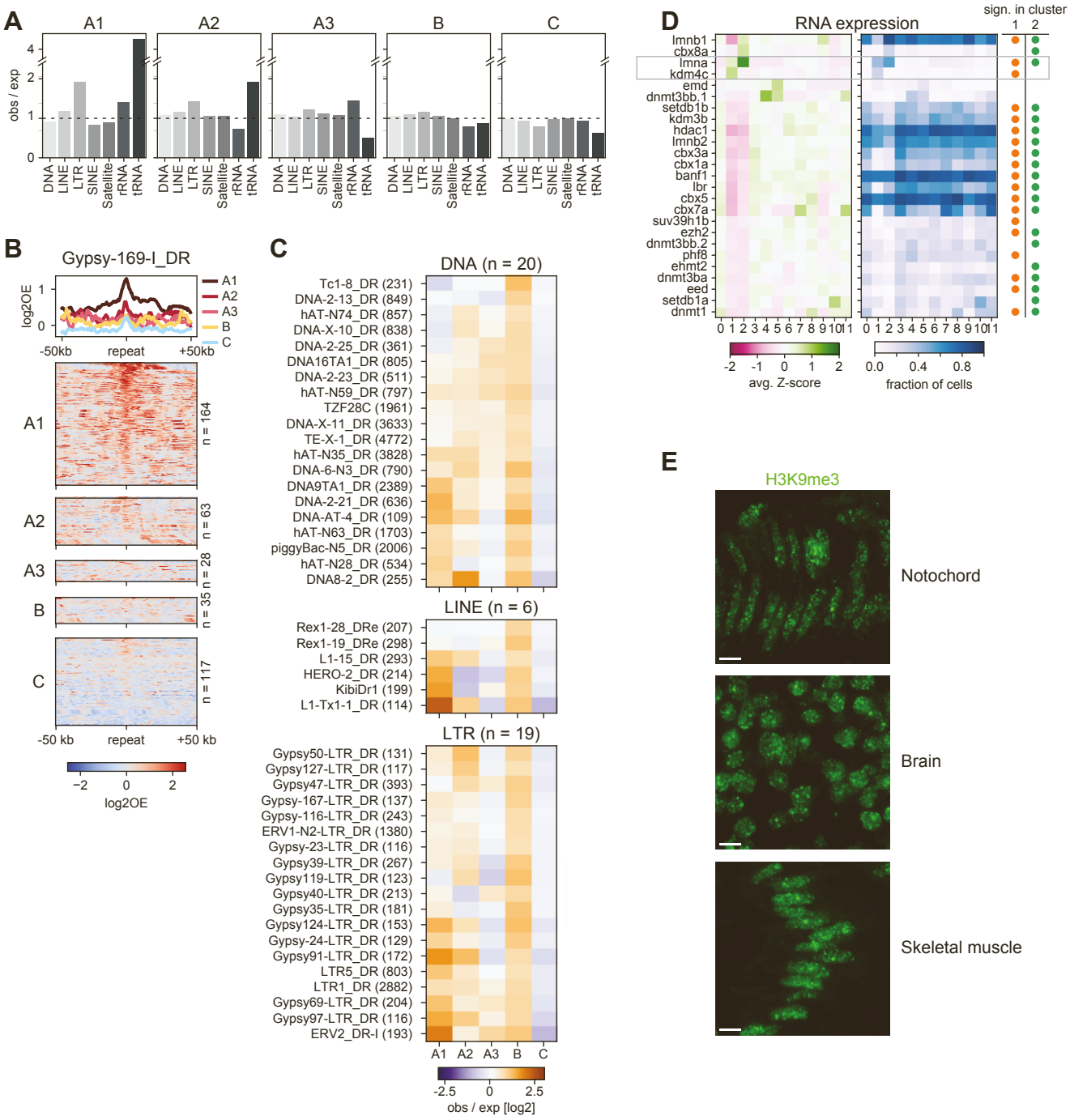

Figure S6: Characterization of repeat content, expression of chromatin factors and nuclear localization of H3K9me3 chromatin, Related to Figure 5

A, Enrichment of repeats per class for all ChromHMM states. Enrichment is computed as the observed number of repeats within a state relative to the expected number based on the genome coverage of each state.

B, H3K9me3 enrichment at Gypsy-169-I_DR repeats across ChromHMM states. The heatmaps show the enrichment per individual repeat instance, while the line plot shows the average enrichment per state.

C, Enrichment of repeats in ChromHMM states as in Figure 5I. Only repeats having at least 100 copies throughout the genome and an enrichment $\geq$1.5 in state B are included. Enrichment is computed as the observed number of repeats in a stated compared to the expected number based on the genome coverage of that state.

D, RNA expression of various chromatin factors across clusters 0-11. The left heatmap shows the average single-cell expression (Z-score); the right heatmaps shows the fraction of cells in each cluster with at least one transcript of each gene. Only factors that are expressed in at least 10% of cells of at least one cluster are shown.

E, Representative images of H3K9me3 staining in cryosections of notochord (left), brain (middle), and skeletal muscle (right) in 15-somite embryos. Scale bars represent 4 μm.