PATTY corrects open chromatin bias for improved bulk and single-cell CUT&Tag profiling

Shengen Shawn Hu[1,2], Zhangli Su[3], Lin Liu[4], Qingying Chen[1], Megan C. Grieco[1], Mengxue Tian[1,5], Anindya Dutta[3], Chongzhi Zang[1,2,5,6,#]

[1] Department of Genome Sciences, University of Virginia, Charlottesville, VA 22908, USA

[2] UVA Comprehensive Cancer Center, University of Virginia, Charlottesville, VA 22908, USA

[3] Department of Genetics, University of Alabama at Birmingham, Birmingham, AL 35233, USA

[4] Institute of Natural Sciences, MOE-LSC, School of Mathematical Sciences, CMA-Shanghai, and SJTU-Yale Joint Center for Biostatistics and Data Science, Shanghai Jiao Tong University, Shanghai, 200240, China

[5] Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA 22908, USA

[6] Department of Biomedical Engineering, University of Virginia, Charlottesville, VA 22908, USA

[#] Correspondence should be addressed to: zang@virginia.edu

Abstract

Precise profiling of epigenomes is essential for better understanding chromatin biology and gene regulation. Cleavage Under Targets & Tagmentation (CUT&Tag) is an efficient epigenomic profiling technique that can be performed on a low number of cells and at the single-cell level. With its growing adoption, CUT&Tag datasets spanning diverse biological systems are rapidly accumulating in the field. CUT&Tag assays use the hyperactive transposase Tn5 for DNA tagmentation. Tn5's preference toward accessible chromatin alters CUT&Tag sequence read distributions in the genome and introduces open chromatin bias that can confound downstream analysis, an issue more substantial in sparse single-cell data. We show that open chromatin bias extensively exists in published CUT&Tag datasets, including those generated with recently optimized high-salt protocols. To address this challenge, we present PATTY (Propensity Analyzer for Tn5 Transposase Yielded bias), a comprehensive computational method that corrects open chromatin bias in CUT&Tag data by leveraging accompanying ATAC-seq. By integrating transcriptomic and epigenomic data using machine learning and integrative modeling, we demonstrate that PATTY enables accurate and robust detection of occupancy sites for both active and repressive histone modifications, including H3K27ac, H3K27me3, and H3K9me3, with experimental validation. We further develop a single-cell CUT&Tag analysis framework built on PATTY and show improved cell clustering when using bias-corrected single-cell CUT&Tag data compared to using uncorrected data. Beyond CUT&Tag, PATTY sets a foundation for further development of bias correction methods for improving data analysis for all Tn5-based high-throughput assays.

Introduction

Epigenetic mechanisms control gene expression and determine cell identity by altering chromatin states without changing DNA sequence[1]. Profiling epigenomic features, including transcription factor (TF) bindings and histone modifications (HM), across cell types, is important for understanding the molecular mechanisms of health and disease[2]. A series of high-throughput assays has been developed to generate epigenomic profiles. Advent in 2007[3,4], ChIP-seq is the first and so far the most widely used epigenomic profiling technique using next-generation sequencing. ChIP-seq uses sonication or micrococcal nuclease (MNase) to digest chromatin, requiring a large number of cells and still limited by the antibody quality and chromatin digestion efficiency, resulting in inherently high noise. ChIP-exo[5] can generate base-pair resolution TF binding patterns with a lower noise level, but requires protein-DNA crosslinking, high-quality antibodies, and precise exonuclease control, limiting its reproducibility and usability. CUT&RUN (Cleavage Under Targets and Release Using Nuclease) uses protein A fused with MNase to target antibody and improves efficiency by only digesting chromatin associated with the target protein's DNA binding sites[6]. CUT&Tag (Cleavage Under Targets and Tagmentation) was then developed to further improve the specificity and efficiency of epigenomic profiling[7]. CUT&Tag utilizes protein A (or G) fused with Tn5 transposase to specifically cleave and tag DNA at protein binding sites *in situ*, and can generate precise epigenomic profiles from native chromatin without cross-linking.

CUT&Tag is believed to outperform earlier techniques with advantages like high sensitivity, high signal-to-noise ratio, and a much easier experimental protocol, and thus has rapidly become popular. Because of its high efficiency and *in situ* nature, CUT&Tag requires low DNA input, and can even be performed on a single-cell level[7]. The versatility of CUT&Tag has been further expanded through innovative modifications and improvements, leading to the development of novel applications, such as CUTAC[8], TIP-seq[9], and HiCuT[10]. Single-cell multi-omic joint profiling

techniques were developed adapting modified CUT&Tag protocols to profile multiple histone modifications together with other modalities in the same cell, such as scCUT&Tag-pro[11] for profiling histone modifications coupled with surface protein abundance, (Droplet) Paired-tag[12–14] for profiling histone modifications and transcriptomes in parallel, CUT&Tag2for1[15], Multi-CUT&Tag[16], and nano-CUT&Tag[17,18] for profiling multiple histone modifications, RNA polymerase II, and chromatin accessibility. Spatial-CUT&Tag[19] was also developed to spatially profile genome-wide histone modifications with a spatial resolution. Assays with CUT&Tag adoption are rapidly growing, with an increasing number of studies using it to investigate epigenetic mechanisms across species, biological systems, and diseases.

To accurately characterize epigenomic and chromatin states, computational analysis should be carefully designed to account for various noises and potential biases in the data generated by these high-throughput techniques. Enzymatic DNA cleavage of Tn5 transposase has a sequence preference that causes biases in the sequencing data[20]. Another potential bias in CUT&Tag may be caused by Tn5 transposase's preference towards accessible chromatin[21–23]. These biases could lead to the misrepresentation of CUT&Tag sequence read enrichment to the true occupancy of histone modifications or transcription factor binding sites. Peak calling is commonly used for signal detection when analyzing CUT&Tag data. MACS2[24] and SICER[25] are widely used statistical model-based ChIP-seq peak calling tools, designed for narrow TF peaks and broad HM signals, respectively. SEACR[26] was developed specifically to analyze CUT&RUN data. However, these conventional peak calling methods were not designed to correct the complicated biases caused by Tn5. We previously developed SELMA to characterize and correct Tn5 enzymatic cleavage biases in ATAC-seq data[20]. However, bias correction for CUT&Tag is still challenging because the open chromatin bias in CUT&Tag is different from Tn5 cleavage bias in ATAC-seq, and for the highly sparse single-cell CUT&Tag data, the bias can

be more substantial compared to bulk data. Innovative computational methods are thus urgently needed to accurately characterize and correct Tn5 biases in CUT&Tag data.

Here, we present PATTY (Propensity Analyzer for Tn5 Transposase Yielded bias) to correct biases in CUT&Tag data at both bulk and single-cell levels, using ATAC-seq as a control. After observing widespread open chromatin bias across published CUT&Tag datasets, we design a comprehensive strategy to benchmark multiple machine learning models and optimize the bias correction approach, and implement PATTY as a pre-trained factor-specific bias correction tool. We show that PATTY can correct open chromatin bias from CUT&Tag data for both active (e.g., H3K27ac) and repressive (e.g., H3K27me3 and H3K9me3) histone modifications and generate more biologically meaningful results. We experimentally validated PATTY's performance on H3K9me3 CUT&Tag. We further develop a computational framework built on PATTY to correct open chromatin bias in single-cell CUT&Tag data, and improve single-cell clustering results.

Results

Tn5 open chromatin bias affects the H3K27me3 CUT&Tag signal

The true histone modification profile in a given cell type/state should be invariant across different epigenomic assays. Therefore, we first compared the published genome-wide profiles of H3K27me3, a repressive histone modification, in the human K562 cell line, generated by CUT&Tag and ChIP-seq. While their peaks largely overlap, a substantial subset (18% on average) of CUT&Tag peaks is unique to CUT&Tag and does not overlap with ChIP-seq peaks (Figure 1a, b). We then examined these CUT&Tag-unique H3K27me3 signals. To our surprise, we observed a clear enrichment of H3K27me3 CUT&Tag signal near the transcription start sites (TSSs) of the actively transcribed genes (Figure 1c, d). As a repressive histone modification,

H3K27me3 should not mark active gene promoters, where clear enrichments of nascent transcribed RNA and RNA Polymerase II (RNAPII) signals are observed (Figure S1a-c), confirming active transcription. In contrast, this H3K27me3 pattern is not present in ChIP-seq, consistent with the transcriptionally active state of these regions (Figure 1e). This observation indicates a potential bias in the CUT&Tag data. Interestingly, the ATAC-seq signal for chromatin accessibility showed a similar pattern to the CUT&Tag signal at these regions (Figure 1f). Indeed, the H3K27me3 peaks that uniquely appear in CUT&Tag but not in ChIP-seq are enriched at active gene promoter regions and ATAC-seq determined open chromatin regions (Figure S1d, e). Considering that active gene promoters are open chromatin and that both ATAC-seq and CUT&Tag use Tn5 to access chromatin and cleave DNA, we argue that the CUT&Tag signals for H3K27me3 at active promoter regions are false signals associated with open chromatin bias caused by Tn5 in the assay.

We then surveyed more H3K27me3 CUT&Tag datasets to assess how commonly the open chromatin bias occurs, especially with the recent protocol improvements to remove or reduce open chromatin bias by using high-salt washes. We further collected 277 H3K27me3 CUT&Tag datasets across various human cell types published since 2024, most of which reported using updated high-salt wash protocols, and examined the signals at the promoter regions of a group of ubiquitously expressed genes. We found the strong enrichment of H3K27me3 false signals across most of these CUT&Tag datasets, significantly different from ChIP-seq data from ENCODE[27] (Figure 1g-i). These observations suggest that open chromatin bias is a widespread issue across H3K27me3 CUT&Tag data, regardless of experimental protocol improvements.

Defining a ground truth benchmark for true and false H3K27me3 marked regions

To further characterize the H3K27me3 CUT&Tag signals and to create a ground truth benchmark that can be used for model training, we curated a set of true and false H3K27me3

CUT&Tag signal patterns using collected CUT&Tag data, incorporating prior biological knowledge. We assumed that true H3K27me3 should be associated with non-expressed genes, and as reciprocal conjugate modifications, H3K27me3 and H3K27ac should not coexist at the same genomic locus in any autosome in a homogeneous cell population. Based on these assumptions, we defined a true H3K27me3-marked region as a reproducible H3K27me3 CUT&Tag peak covered region around non-expressed genes, with no H3K27ac sequence reads. In the meantime, we defined a false H3K27me3-marked region as a reproducible H3K27me3 CUT&Tag peak covered region around highly expressed genes while overlapping with H3K27ac peaks (Figure 2a, b, Methods). We collected 17 high-quality H3K27me3 CUT&Tag datasets and generated such defined sets of true and false H3K27me3-marked regions in the human K562 cell line. At the 200bp nucleosomal resolution, we obtained 1428 true (285.6kb) and 1231 false (246.2kb) H3K27me3 CUT&Tag signal-marked regions. As expected, we found that the false H3K27me3 CUT&Tag regions exhibited significantly higher ATAC-seq signal levels than the true H3K27me3-marked regions, suggesting biases due to open chromatin (Figure 2c).

Evaluating machine-learning models to predict true H3K27me3-marked regions with multi-modal features

To identify true H3K27me3 marked regions from CUT&Tag data considering the open chromatin bias, we sought to develop a supervised machine-learning model to classify true and false H3K27me3 signal regions from multi-modal features on the pre-defined ground truth benchmark data from human K562 cells (Figure 3a, Methods). For each 200bp region in the genome, we considered the following potential features: the 10bp-resolution normalized signal patterns across the 1kb window centered at the 200bp bin for H3K27me3 CUT&Tag (the primary signal), ATAC-seq (to represent open chromatin), and IgG (to represent a negative control[8]), as well as the 200bp one-hot encoded DNA sequence[28,29]. We employed several commonly used

machine-learning models for this task, including panelized Logistic Regression (LR), Random Forest (RF), and Generalized Boosted Model (GBM), as well as state-of-the-arts deep neural network models, including Convolutional Neural Networks (CNN)[30–32], Multilayer Perceptron (MLP), Recurrent Neural Network (RNN), and Gated Recurrent Unit (GRU), to identify the best-performing model (Figure S2a-c, Methods). We first assessed the models' performance using a 5-fold cross-validation on the pre-defined true/false signal regions. Using all feature combinations, we found that the performance varies across different models and different feature combinations, and deep learning-based models seem to have better performance than other models (Figure S2d).

Next, we trained each model using the whole benchmark set of true/false H3K27me3 regions, and applied the trained model to the whole genome to determine whether any genomic region with H3K27me3 CUT&Tag signal has a true or false H3K27me3 mark in human K562 cells. We used an orthogonal, biology-informed approach to evaluate the model performance: We calculated the genome-wide correlation between the model prediction score (true/false H3K27me3 signal) and either RNA-seq or H3K27ac ChIP-seq signal level, and assess the model performance based on the biological ground truth that as a repressive histone modification, H3K27me3 should be negatively correlated with both gene expression and H3K27ac signal across the genome in the same cell type. As a result, the logistic regression (LR) model showed consistently the best performance, i.e., yielding the highest negative correlation between the model-predicted H3K27me3 and both gene expression (Figure 3b, S2e) and H3K27ac signals (Figure 3c, S2f). Moreover, among all feature combinations, including only the CUT&Tag and ATAC-seq signals yielded a better performance of the LR model than all other feature combinations as represented by the highest negative correlation with both gene expression (Figure 3d) and H3K27ac signal (Figure 3e). Neither IgG nor DNA sequence can further improve the model performance, while ATAC-seq is essential to the true H3K27me3

determination, indicating the importance of considering open chromatin bias in obtaining the correct H3K27me3-marked regions (Figure 3d, e). These results suggested that the LR model using CUT&Tag and ATAC-seq signal features is an optimal approach to correct open chromatin bias and generate accurate genome-wide H3K27me3-marked regions for H3K27me3 CUT&Tag data analysis. Therefore, we implement this model as a computational tool called Propensity Analyzer for Tn5 Transposase Yielded bias (PATTY), for improved analysis of histone modification CUT&Tag data by correcting the open chromatin bias (Figure 3f).

PATTY corrects open chromatin bias and improves H3K27me3 and H3K27ac CUT&Tag analysis

We applied PATTY to generate the corrected H3K27me3 profiles for the two H3K27me3 CUT&Tag datasets samples shown in Figure 1c,d, and found that the false signal pattern at the active gene promoter regions caused by the open chromatin bias was corrected throughout the genome (Figure 4a-f), exemplified at two active gene loci, NUP214[33] (Figure 4g) and GCSH[34] (Figure 4h). We further showed that the PATTY-corrected H3K27me3 CUT&Tag profile is more negatively correlated with gene expression (Figure 4i) and H3K27ac profile (Figure 4j), compared to the H3K27me3 profile generated by peak-calling with macs2, with or without considering IgG control, but without bias correction. Meanwhile, the H3K27me3 signal patterns near the repressed gene promoter regions were not much affected after PATTY correction (Figure 4g, h, S3).

It has been reported that modified CUT&Tag experimental protocols, such as using high-salt wash steps, can help reduce the open chromatin bias in H3K27me3 CUT&Tag data[23]. We then specifically tested the performance of PATTY on H3K27me3 data generated from such modified CUT&Tag protocols, including high-salt CUT&Tag[23] (Figure 5a, b) and CUTAC[8] (Figure 5c, d). Interestingly, although the open chromatin bias was supposed to be reduced by these modified

protocols, the H3K27me3 profiles after PATTY correction still attained higher negative correlations with both gene expression (Figure 5a, c) and H3K27ac (Figure 5b, d), indicating that there are still residue biases in the data even with modified experimental protocols, confirming what we observed (Figure 1g-i), and PATTY is still able to correct the bias computationally and further improve the data interpretations. These results suggest that PATTY can successfully correct the open chromatin bias in the H3K27me3 CUT&Tag data and improve signal detection.

We then tested PATTY's ability to correct open chromatin bias in CUT&Tag data for H3K27ac, an active histone modification. Unlike repressive chromatin marks, which generally do not co-occur with open chromatin, the Tn5-induced signal may confound the true active histone modification signal at the same locus, making bias correction more challenging than for repressive marks. Because the machine learning framework in PATTY is agnostic to whether the CUT&Tag is for active or repressive marks, we applied the same workflow for H3K27ac. Similar to what we did for H3K27me3, we curated a set of true and false H3K27ac CUT&Tag signal regions using collected CUT&Tag data in K562, trained the H3K27ac PATTY model using the H3K27ac CUT&Tag and ATAC-seq signal patterns as input features, and applied the trained model to generate the genome-wide corrected H3K27ac CUT&Tag profile. As a result, we observed that the PATTY-corrected H3K27ac profile has an increased positive correlation with gene expression (Figure 5e), and a higher negative correlation with its reciprocal conjugate histone modification, H3K27me3 (measured by ChIP-seq for orthogonality), compared to conventional macs2 peak calling without bias correction (Figure 5f). These results indicate that PATTY can correct open chromatin biases existing in CUT&Tag for both active and repressive histone marks.

Pretrained PATTY model can correct CUT&Tag biases across different cell types

Next, we examined whether the pre-trained PATTY model using our pre-defined ground truth datasets from one cell type can correct the open chromatin bias in CUT&Tag for the same histone modification in other cell types and improve data analysis universally. We applied the PATTY model trained with data in K562 to several CUT&Tag datasets generated in the human embryonic stem cell line H1, and observed consistently improved analysis results compared with conventional peaking calling without bias correction, for both H3K27me3 (Figure 5g, h) and H3K27ac (Figure 5i, j), as demonstrated using the same metrics such as correlation with gene expression (Figure 5g, i) or with the reciprocal conjugate histone modification (Figure 5h, j). These data suggest that using the model pre-trained from one cell type, PATTY is powerful for correcting the open chromatin bias and improving analysis of CUT&Tag data from different cell types. This result also indicates that the machine-learning model in PATTY captures the CUT&Tag intrinsic open chromatin bias specific to each histone modification rather than any cell-type-specific information.

PATTY corrects bias and improves CUT&Tag data analysis for H3K9me3

After demonstrating PATTY's performance with extensive public data for H3K27me3 and H3K27ac, we next sought to evaluate PATTY's effectiveness on our own experimental data for H3K9me3, a histone modification associated with inactive heterochromatin. We used a similar strategy to curate a ground-truth dataset of true/false H3K9me3-marked regions in K562 using CUT&Tag data for reciprocal conjugate histone modifications H3K9me3 and K3K9ac, and trained a PATTY model for H3K9me3. We performed a CUT&Tag experiment for H3K9me3 in the HCT116 cell line and tested the performance of PATTY on this in-house dataset. Similar to H3K27me3, the open chromatin bias caused a clearly false enrichment of H3K9me3 CUT&Tag signal at active gene promoter regions (Figure 6a), a pattern absent in ChIP-seq (Figure 6b). As expected, the false signals caused by open chromatin bias (Figure 6c) were corrected by PATTY (Figure 6d), exemplified by several gene loci (Figure 6e-g). Meanwhile, H3K9me3

CUT&Tag signals at non-expressed gene loci were less affected by the open chromatin bias, and therefore, they were similar to ChIP-seq signals and not altered by PATTY correction (Figure 6e-g, S4). We also examined PATTY's performance using the same genome-wide correlation metrics and found that the PATTY-corrected H3K9me3 CUT&Tag profile is negatively correlated with both gene expression (Figure 6h) and the profile of H3K9ac (Figure 6i). Such negative correlations were not detected if the H3K9me3 CUT&Tag data only underwent MACS2 peak calling without bias correction (Figure 6h, i). Therefore, our in-house experiment for H3K9me3 validated the effectiveness of PATTY in correcting the open chromatin bias in CUT&Tag data.

PATTY Improves cell type clustering from H3K27me3 single-cell CUT&Tag data

Recent innovations in single-cell CUT&Tag enable the detection of epigenomic profiles in each of hundreds or thousands of cells[11–16,18]. Single-cell CUT&Tag data are often highly sparse, and most histone modification events in individual cells can only be represented by a single DNA fragment, which could amplify the effect of the Tn5-caused open chromatin bias. We collected H3K27me3 and H3K27ac single-cell CUT&Tag data from two independent studies that employed single-cell multiome co-assays, which not only profiled multiple histone marks but also enabled cell type annotation for each individual cell to serve as ground truth. Specifically, cell type annotation was determined by cell surface markers for scCUT&Tag-pro[11] and by multi-modality information for nano-CT[18], respectively. Using two groups of cells annotated as CD8+ T cells and monocytes from the scCUT&Tag-pro dataset[11] as an example, we observed enrichments of H3K27me3 CUT&Tag signals at the promoter regions of ubiquitously expressed genes, where H3K27me3 should not occur (Figure S5a, b), and differentially expressed genes between the two cell types (Figure S5c-f), where H3K27me3 should not occur in the respective cell type highly expressing these genes (Figure S5d, e), suggesting open chromatin bias

affecting this single-cell dataset from exhibiting specific H3K27me3 signals to differentiate cell types.

We then assessed whether PATTY can correct the bias and improve the analysis of these single-cell CUT&Tag data. To reduce the data sparsity and for PATTY to work genome-wide seamlessly, we employed a meta-cell approach, to represent the signal profile of each individual cell with the average signal profile of that cell and its 10 nearest neighboring cells in the top 50 principal component (PC) space. We then applied PATTY to each cell's meta-cell-smoothed CUT&Tag signal with the ATAC-seq signal in the same cell system, and generated a PATTY-corrected single-cell CUT&Tag data matrix. We compared cell clustering results from PATTY-corrected data with uncorrected data. For both H3K27me3 and H3K27ac, PATTY-corrected data yielded more accurate cell clusters than uncorrected data in both the nano-CT dataset (Figure 7) and the scCUT&Tag-pro dataset (Figure S6), as quantified by the adjusted Rand index compared to the ground-truth cell type annotation (Figure 7d, h, S6d, h). After PATTY correction, H3K27me3 signals showed anticipated differential patterns at the differentially expressed genes between monocytes and CD8+ T cells (Figure S5g, h). These data suggest that PATTY can correct open chromatin bias in single-cell CUT&Tag and improve cell clustering.

Discussion

As CUT&Tag has been increasingly used for epigenomic profiling, characterization and correction of biases in CUT&Tag data have become inevitable. We have shown the widespread existence and severe impact of open chromatin bias in CUT&Tag, including both bulk and single-cell datasets generated with the latest protocols. We demonstrated the ability of PATTY in correcting open chromatin bias for both bulk and single-cell histone modification CUT&Tag

data, using ATAC-seq as control, and validated the results using both public and in-house experimental data. PATTY has been implemented as an open-source software tool pre-trained for H3K27me3, H3K27ac, and H3K9me3 CUT&Tag. It can be used for any cell type, biological systems, either bulk or single cell, without re-training.

We attribute the observed bias to Tn5 transposase, which plays a critical role in the CUT&Tag assay, especially in generating fragment libraries by inserting sequencing adapters into the DNA, i.e., tagmentation. Since Tn5 transposase prefers chromatin-accessible regions[11,35], utilizing Tn5 in CUT&Tag inevitably results in some read enrichment towards open chromatin regions, regardless of the feature of the target protein of CUT&Tag experiments[23]. This bias is intrinsic to Tn5 and therefore cannot be completely eliminated by experimental protocol optimization, as demonstrated in recently published datasets. We use ATAC-seq to control this bias, but cannot simply apply a conventional peaking calling method using ATAC-seq as the background control, as the chromatin input library used in ChIP-seq peak calling, because the quantitative effects caused by Tn5 transposase in CUT&Tag and ATAC-seq are not as simple as a signal-control relationship, and more sophisticated modeling is required. Therefore, we use machine learning approaches by considering the high-resolution signal patterns as input features, and indeed show that PATTY outperforms state-of-the-art peak-calling methods like macs2.

A key step in developing a well-performing machine learning model is to prepare appropriate, high-quality training data. In this work, it boils down to determining the ground truth epigenomic pattern that we can use for model training. We identified ground truth benchmark datasets for H3K27me3, H3K27ac, and H3K9me3 using publicly available data based on orthogonal biological knowledge. In the PATTY framework, the models are factor/histone modification specific due to the specific biological feature of each epigenomic mark, and the training data are

from one common cell line. However, once the model is trained, it can be used in any cell type for that same epigenomic mark. When more data becomes available for other epigenomic marks in the future, we can always use a similar strategy to identify ground truth datasets for each mark, and apply the same framework to train a PATTY model for bias correction for that mark.

The model design for PATTY highlights the importance of biological knowledge in computational biology method development. When orthogonal biological information is used for model evaluation, popular black-box deep learning models do not perform as well as simple, more interpretable logistic regression. This is consistent with a recent study showing that deep learning models did not outperform linear models in single-cell gene perturbation prediction[36]. Therefore, choosing an appropriate model is crucial. The model selection should always consider biological or scientific information for the specific goal, and should not solely rely on blindly increasing model complexities. When comparing different feature combinations, it is also interesting that considering DNA sequence features does not necessarily yield a better performance. This is possibly because the sequence bias information has already been implicitly included in the control ATAC-seq data[20]. When evaluating model performance, we use orthogonal and biologically meaningful metrics such as correlation with gene expression and with reciprocal conjugate histone modifications, rather than relying on correlation or similarity with ChIP-seq data. This is because we consider CUT&Tag as a more advanced technique, or at least an alternative, to ChIP-seq. Therefore, we view the two techniques as parallel and do not treat ChIP-seq as a ground truth or benchmark against with CUT&Tag should be judged.

Tn5 transposase is a powerful enzyme in epigenetics research. In addition to ATAC-seq and CUT&Tag, Tn5 transposase has also been used in other high-throughput assays for whole genome sequencing (LIANTI[37]), chromatin interaction detection (HiChIP[38] and HiCAR[38,39]), and

spatial epigenomic profiling (epigenomic MERFISH[40]), etc. Similar to CUT&Tag, the Tn5 preference to open chromatin could induce potential biases in data from those assays as well. Modeling and correcting the open chromatin bias and other potential enzymatic biases in these high-throughput data remain a significant task in computational biology. Our work sets a foundation for open chromatin bias correction in other data types, and more work can be built to help improve the analysis for meaningful biological interpretation.

Methods

High-throughput sequencing data collection and processing

CUT&Tag, ChIP-seq, ATAC-seq, and IgG data were processed as follows: Raw sequencing reads were aligned to the GRCh38 (hg38) reference genome with bowtie2[41] (v2.2.9) (-X 2000 for paired-end data). Low-quality reads (MAPQ < 30) were discarded. For paired-end sequencing data, reads with two ends aligned to different chromosomes (chimeric reads) were discarded. For paired-end data, reads with identical 5' end positions for both ends were regarded as redundant reads and discarded. Reads mapped to mitochondrial DNA (mtDNA reads) were discarded. To generate the genome-wide signal track for downstream analysis (i.e., scan signal pattern on genome-wide bins, and generate signal heatmap across gene promoters), we extended reads from their 5'end to 146 base pairs (bp), piled them up, and normalized the piled-up signal by total read count in the sample (i.e., RPM). For CUT&Tag and ChIP-seq data, SICER (v2)[25] peaks were detected with default parameters (used in Figure 1a, b, and definition of True/False regions). For H3K27ac CUT&Tag, H3K27ac ChIP-seq, and ATAC-seq data, macs2[24] peaks were detected with additional parameters "-q 0.01 --nomodel –extsize 146". For H3K27me3 and H3K9me3 CUT&Tag and ChIP-seq data, macs2 peaks were detected with additional parameters "-q 0.01 –broad --nomodel --extsize 146".

RNA-seq was processed as follows: Raw sequencing reads were aligned to the GRCh38 (hg38) reference genome with Hisat2 (v2.2.1)[42]. Low-quality reads (MAPQ < 30) were discarded. We then estimated the expression index (RPKM) using Stringtie (v2.1.5)[43].

Exon-array data were processed with the “jetta” package[44]. We collected all exon array samples from the study (GSE19090) as the input data and used the “jetta.do.expression” function to estimate and normalize the expression index of samples. The average expression index from 3 replicates in the K562 cell line was used as the K562 gene expression.

To examine the performance of recent CUT&Tag data with the latest protocols, we collected all human H3K27me3 CUT&Tag datasets from GEO generated since 2024, in total 277 samples. For comparison, we also collected 30 H3K27me3 ChIP-seq datasets from the ENCODE project[27] (generated from two different institutes). We selected a set of ubiquitously expressed genes as the top 1000 genes with the highest minimum expression index across all cell types (normalized expression from Exon array data) and examined the pattern and enrichment of all 277 samples on the promoter (TSS ±3kb) of these genes. For each H3K27me3 dataset, we calculated the average signal pattern across the promoter regions of all ubiquitously expressed genes. Then the average patterns from the 277 samples were summarized to generate Figure 1g. To quantify the level of signal enrichment in promoter regions, we calculated the average normalized signal level at promoter center region, defined as TSS ± 300bp, and flanking promoter region, defined as 300bp-3kb from TSS on both sides, respectively, for the 1000 ubiquitously expressed genes. The center and flanking scores for all 277 datasets were visualized as a scatter plot in Figure 1h. The log2 ratio of center and flanking scores was further compared (Figure 1i).

For single-cell RNA-seq data in human PBMC[45], we downloaded the processed data and used Seurat[46] (v5.3.0) for the data processing and analysis. The cell type information was inherited from published data. The PBMC-ubiquitously expressed genes (used in Figure S5a, b) were defined with sufficient average expression index in all PBMC cell types (normalized average expression > 0.2). Two cell types with sufficient cell numbers (CD8+ T cells and monocytes) were collected, and we performed differential expression analysis (fold change >=2 and adjusted p-value < 0.05) to capture differentially expressed genes between the two cell types used in Figure S5.

For single-cell CUT&Tag data, processed data (aligned fragments and Seurat/Signac R objects) were downloaded from the original studies[11,18]. We used the ArchR package (v1.0.1)[47] for single-cell data processing. All high-quality cells provided were used. The top 25k tilling bins with the highest cross-cell signal variance were collected and split into 200bp bins for the downstream analyses (referred to as high-var bins below).

Selection of true and false H3K27me3/H3K27ac marked regions

H3K27me3 and H3K27ac CUT&Tag datasets in the K562 cell line were collected from Reference[7], the study with the most datasets, without data from other publications to avoid potential batch effects. Peak calling analysis was performed on the collected CUT&Tag datasets with SICER (v2). Peaks were then split into 200bp tiling bins as candidate H3K27me3 or H3K27ac regions.

Considered as the ground truth, the true H3K27me3 marked regions were selected based on the following criteria: 1) H3K27me3 marked regions should not be actively transcribed. So we selected candidate H3K27me3 regions (in 200bp bins obtained from SICER peaks) from each CUT&Tag H3K27me3 dataset that are overlapped with the promoter (+/- 1kb from

transcriptional start site, TSS) or gene body regions of non-expressed genes, defined as those with 0 RPKM detected from K562 RNA-seq data (8,032 genes in total). 2) H3K27me3 and H3K27ac should be mutually exclusive in a homogeneous cell population. So we filtered out the candidate H3K27me3 regions that have any H3K27ac CUT&Tag reads from any H3K27ac CUT&Tag dataset. 3) H3K27me3 regions should be reproducible from many CUT&Tag samples. So we selected those remaining candidate H3K27me3 regions (200bp bins) that reoccurred in at least 14 out of the total 17 H3K27me3 CUT&Tag samples as the final true H3K27me3 regions.

Accordingly, the false H3K27me3 regions were selected based on the following criteria: 1) H3K27me3 signals detected in actively transcribed regions are false. So we selected candidate H3K27me3 regions (in 200bp bins obtained from SICER peaks) from each CUT&Tag H3K27me3 dataset that overlapped with the promoter or gene body regions of the highly expressed genes, defined as the top 8,032 genes (on par with the non-expressed genes for selecting the true H3K27me3 regions) ranked by the expression level (in RPKM) in K562. 2) H3K27me3 and H3K27ac should be mutually exclusive in a homogeneous cell population. Thus, we further selected those candidate H3K27me3 regions that overlapped with H3K27ac CUT&Tag peaks as the false H3K27me3 regions.

The true H3K27ac regions were selected as the H3K27ac SICER peak regions that overlapped with the promoter or gene body regions of the highly expressed genes, and had no H3K27me3 CUT&Tag read detected in the candidate region or the +/-500bp flanking regions from any H3K27me3 CUT&Tag dataset. The false H3K27ac regions were selected as the H3K27ac SICER peak regions that overlapped with the promoter or gene body regions of the non-expressed genes, and also overlapped with H3K27me3 CUT&Tag SICER peaks from any H3K27me3 CUT&Tag sample.

The true H3K9me3 regions in K562 were selected as the H3K9me3 SICER peak regions that 1) have greater than or equal to 5 reads in both K562 H3K9me3 CUT&Tag samples, 2) do not overlap with the gene body or +/-10kb flanking regions of the highly expressed genes, and 3) have no H3K9ac CUT&Tag reads detected in the regions. The false H3K9me3 regions were selected as the H3K9me3 SICER peak regions from any H3K9me3 CUT&Tag sample that overlapped with the gene body or +/-10kb flanking regions of the highly expressed genes, and had at least 5 reads from any H3K9ac CUT&Tag sample.

We only included histone modification signals on the 22 autosomes and assumed homozygous chromatin states. The X chromosome was excluded to avoid potential confounding effects arising from X chromosome inactivation. The overlapping status between the two region sets was determined by BEDtools[48] with at least 1bp overlap.

Classifying true and false regions with different machine learning models and various genomic feature combinations

We built supervised models to predict the true/false classification (1/0) for the histone modification marked regions, employing a collection of machine learning techniques, including panelized logistic regression (LR), random forest (RF), gradient boosting machine (GBM), and deep neural networks (dNN)., and using various combinations of genomic features, including CUT&Tag signal pattern, ATAC-seq signal pattern, IgG signal pattern, and one-hot encoded DNA sequence (Figure 3a). For each candidate region (a 200bp bin) as an observed sample, the 10bp-resolution signal pattern (normalized DNA fragment pile-up count) across a 1kb region centered on the bin center was considered as a feature vector, for CUT&Tag, ATAC-seq, and IgG. During model training, signal patterns from different H3K27me3 CUT&Tag datasets on the same candidate region were considered independent observed samples. When multiple signal

patterns from CUT&Tag, ATAC-seq, or IgG profiles were selected as features in the dNN models, the multiple signal patterns were aligned by genomic locations as a matrix feature. The one-hot encoded DNA sequence for the entire 200bp bin was considered as the sequence feature. In the dNN models, if signal pattern features (i.e., CUT&Tag, ATAC-seq, or IgG) and DNA sequence features were both included in the model, the signal features and sequence features were considered as two branches, and their flattened outputs were concatenated later.

The panelized logistic regression (LR) model used elastic net penalty (alpha = 0.5), with maximum iterations set as 10,000. The decision tree number (n_estimators) of the random forest (RF) model was set as 100. In LR, RF, and GBM models, all selected features were flattened and horizontally concatenated as the model input.

We tested four dNN architectures in the model exploration steps (Figure 3a, S2), including Convolutional Neural Networks (CNN), Multilayer Perceptron (MLP), Recurrent Neural Networks (RNN), and Gated Recurrent Unit (GRU). For all four models, an input layer with shape (200, 4) was defined for the one-hot encoded sequence features, where 200 represents the nucleotide composition in each bp across the 200bp bin. An input layer for the signal feature had a shape (100, N), where 100 represents the 100 data points for the 10bp-resolution signal across the 1kb region, and N, chosen from 1-3, represents the selected number of signal pattern features, e.g., N=1 for CUT&Tag only, N=2 for CUT&Tag + ATAC or CUT&Tag + IgG, and N=3 for CUT&Tag + ATAC + IgG.

For CNN (Figure S2a), 1) a 1D convolutional layer with 32 filters and ReLU activation was applied. L2 regularization was used to prevent overfitting. 2) A MaxPooling1D layer with a pool of size 2 was added to reduce the dimensionality and computational complexity of the feature maps, and to extract the most salient features. All these layers were applied to both the

sequence and the signal branches. 3) The flattened outputs from both branches were concatenated. 4) A dense layer with 128 units and ReLU activation was added, and L2 regularization was applied again in this step. 5) A dropout layer for mitigating overfitting and a final dense layer with a single unit and sigmoid activation for binary classification (true/false labels) were added. The kernel size and dropout rate were tuned for optimized model performance.

For MLP (Figure S2b), 1) the input layers for both branches were passed through the flatten layer to convert the input tensors into one-dimensional vectors. 2) The vectors were then concatenated into a single vector. 3) The combined vector was fed into a dense layer with 128 units and ReLU activation. L2 regularization was applied to this layer to prevent overfitting. 4) The output from the first dense layer was then passed into a second dense layer with 64 units and ReLU activation, also with L2 regularization. 5) The final output layer was a dense layer with a single unit and a sigmoid activation function. The regularization coefficient was tuned for optimized model performance.

For RNN and GRU (Figure S2c), 1) the input layers were processed through an LSTM/GRU layer with 32 units and L2 regularization. The LSTM/GRU layer would capture temporal dependencies within both the sequence and signal branches. 2) The outputs from the two LSTM/GRU layers from the two branches were concatenated into a single vector, which was the input to a dense layer with a single unit, and a sigmoid activation function was applied to the concatenated output.

All dNN models were trained using Adam with a learning rate tuned for optimized performance evaluated by cross-validation (described below). The loss function was set to binary cross-entropy, which is suitable for binary classification tasks. We designed a 5-fold cross-validation to

evaluate the generalization performance using accuracy metrics (Figure S2a). The parameters, including kernel size and dropout rate in CNN, as well as regularization coefficient and learning rate, were also tuned in this step. For each dNN model, the hyperparameter with the optimized performance was used for the genome-wide analyses (Figure 3b-e).

All dNN models were implemented with TensorFlow and Keras (v2.11.0) packages, and other models were implemented with the sklearn (v1.0.2) packages.

Genome-wide evaluation of machine learning models

After training a model using the curated true/false datasets with cross-validation, we applied the trained model to the genome-wide 200bp bins and obtained a model prediction score for each bin (referred to as the corrected signal in Figure 3b-e). The model prediction score ranges from 0 to 1. A prediction score closer to 1 means the bin is more likely to have a true signal, and a prediction score closer to 0 means the bin is more likely not to have a true signal. We next collected all bins located within +/- 1kb from any gene's TSS (promoter bins) and examined the expression level of that gene whose promoter each bin is located in. We calculated the Pearson correlation coefficient between the corrected signal (model prediction score) and the associated gene expression across all promoter bins (e.g., Figure 3b, d). We assumed that repressive histone marks (e.g., H3K27me3, H3K9me3) should be negatively correlated with gene expression, and active histone marks (e.g., H3K27ac) should be positively correlated with gene expression. Therefore, we used the calculated correlation coefficient as a metric to evaluate the model's performance. Because the RNA-seq data in the K562 cell line had been used for selecting true/false CUT&Tag signal regions as training data, here we used the exon array data in K562 as the gene expression measurement for this genome-wide evaluation, to maintain the orthogonality and rigor.

When calculating the model output's correlation with other histone mark profiles (e.g., Figure 3c, e), we used the ChIP-seq signal for the other histone mark in the same cell line. For H3K27ac and H3K9me3 CUT&Tag data, bins with at least one read were included for the correlation calculation. For H3K27me3 CUT&Tag data, bins with at least one read in at least 6 samples were included for the calculation. For the macs2 results presented for comparison, a bin overlapped with a macs2 peak was assigned a score as the Macs2 generated fold enrichment value of that peak. The other bins that are not associated with any Macs2 peak were assigned a score of 0.

PATTY method

Based on the model performance results, LR models were employed in the PATTY method. In practice, we implemented PATTY using the pre-trained LR model for each specific histone mark. The input of PATTY includes the mapped reads files (in BED format) of a CUT&Tag dataset and an ATAC-seq dataset from the same cell sample. PATTY will generate the signal tracks (for genome browser visualization) and run the LR model prediction across the genome to generate a 200bp resolution corrected signal profile. Any prediction score below 0.5 can be reduced to 0 for visualization purposes. Peak calling can be further performed on the PATTY-corrected signal profile.

Signal patterns on active/silent gene promoter regions

The top 4,000 genes with the highest expression level (measured by RPKM value from RNA-seq) and having ATAC-seq signal in their promoter regions (TSS +/-3kb) in K562 cells were selected as K562 active genes. Genes with 0 RPKM in K562 RNA-seq data and with at least 5 reads in each dataset of both H3K27me3 CUT&Tag replicates and H3K27me3 ChIP-seq were selected as silent genes. For the raw CUT&Tag, ChIP-seq, and ATAC-seq signals, sequence depth-normalized signal (described in the data collection and processing section) was used for

the aggregate plots, heatmaps, and genome browser visualizations. The PATTY corrected signal was used for the PATTY corrected plots.

Single-cell PATTY score and evaluation

To evaluate PATTY's performance on single-cell CUT&Tag data, we first developed a meta-cell approach to reduce the single-cell data sparsity before applying PATTY. In detail, we first calculated the pairwise Euclidean distances of all cell pairs in the top 50 high-variance principal component (PC) space from the ArchR preprocessed output. For each individual cell, the top 10 nearest cells (with the shortest distance) were selected as the neighboring cells. The scCUT&Tag signal patterns of these 10 neighboring cells and the target cell itself (11 cells in total) were merged together to represent the CUT&Tag signal pattern of the target cell. We collected publicly available scATAC-seq from the same cell types as the scCUT&Tag data for the ATAC-seq signal pattern, i.e., 10x Genomics PBMC scATAC-seq for scCUT&Tag-pro, and the scATAC-seq component in the nano-CT experiments for the nano-CT scCUT&Tag data. To reduce the computing complexity, we treated the scATAC-seq data as bulk ATAC-seq data. Then, for each individual cell, we used the meta-cell adjusted CUT&Tag signal pattern and the bulk ATAC-seq signal pattern as the input for PATTY to correct bias and generate the signal of the cell on all candidate bins. Finally, we can generate a PATTY score matrix for all cells across all candidate bins as the bias-corrected scCUT&Tag data matrix.

To compare the raw data matrix and PATTY-corrected data matrix, we first conducted a PCA dimensional reduction using the top 10k high-variance bins. Next, we ran a k-means clustering on the PCA output, where k was set to the same number of cell types in the ground truth, which referred to the "L2" level cell type annotation published in the original study[11,18]. Then, the clustering result was compared to the ground truth annotation, and an adjusted Rand index (ARI) was calculated to assess the clustering accuracy (Figure 7d, h, Figure S6d, h). We

applied this ARI approach to both the normalized raw data matrix and the PATTY-corrected data matrix for a fair comparison. We further visualized the cell type annotation and clustering labels (from both raw data and PATTY-corrected data) on the UMAP visualization. Only high-variance bins were included in the analysis results presented in Figure 7 and Figure S6.

Cell lines and culture

HCT116 p53-/- cell (referred to as wild-type, RRID: CVCL_S744) was a generous gift from Fred Bunz (Johns Hopkins)[49]. All cell lines were maintained in McCoy's 5A-modified medium (Corning, no. 10-050-CV) supplemented with 10% fetal bovine serum. Cell lines were authenticated by STR method and routinely tested for mycoplasma contamination by PCR.

H3K9me3 CUT&Tag library preparation

H3K9me3 CUT&Tag libraries were prepared using Active Motif CUT&Tag-IT Assay Kit (cat #2200) with H3K9me3 antibody (Diagenode #C15410193) following the manufacturer's instructions. Specifically, 500,000 cells and 1 μg antibody were used per reaction.

Data availability

The H3K9me3 CUT&Tag data in HCT116 is available at the Gene Expression Omnibus (GEO) with accession number GSE298565.

Code availability

The source code of PATTY and scripts for all analyses of this work are available at Github: https://github.com/zang-lab/PATTY

Acknowledgements

The authors thank Dr. Ye Zheng for helpful suggestions and critical reading of the manuscript, and Dr. Bingjie Zhang for assistance on scCUT&Tag-pro analysis. This work was supported by NIH grants R35GM133712 (CZ), R21HG012981 (CZ), R00CA259526 (ZS), and R01CA060499 (AD).

References

1. Allis, C. D. & Jenuwein, T. The molecular hallmarks of epigenetic control. *Nat. Rev. Genet.* **17**, 487–500 (2016).

2. Stricker, S. H., Köferle, A. & Beck, S. From profiles to function in epigenomics. Nat. Rev. Genet. **18**, 51–66 (2017).

3. Barski, A. *et al.* High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129**, 823–837 (2007).

4. Wang, Z. et al. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genetics* **40**, 897–903 (2008).

5. Rhee, H. S. & Pugh, B. F. Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. Cell **147**, 1408–1419 (2011).

6. Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife* 1–35 (2017) doi:10.7554/elife.2**185**6.001.

7. Kaya-Okur, H. S. et al. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* **10**, 1930 (2019).

8. Henikoff, S., Henikoff, J. G., Kaya-Okur, H. S. & Ahmad, K. Efficient chromatin accessibility mapping in situ by nucleosome-tethered tagmentation. *eLife* **9**, e63274 (2020).

9. Bartlett, D. A. et al. High-throughput single-cell epigenomic profiling by targeted insertion of promoters (TIP-seq). *J. Cell Biol.* **220**, e202103078 (2021).

10. Sati, S. et al. HiCuT: An efficient and low input method to identify protein-directed chromatin interactions. *PLoS Genet.* **18**, e1010121 (2022).

11. Zhang, B. et al. Characterizing cellular heterogeneity in chromatin state with scCUT&Tag-pro. *Nat. Biotechnol.* **40**, 1220–1230 (2022).

12. Xie, Y. et al. Droplet-based single-cell joint profiling of histone modifications and transcriptomes. *Nat. Struct. Mol. Biol.* **30**, 1428–14**33** (2023).

13. Wu, S. J. et al. Single-cell CUT&Tag analysis of chromatin modifications in differentiation and tumor progression. Nat. Biotechnol. **39**, 819–824 (2021).

14. Bartosovic, M., Kabbe, M. & Castelo-Branco, G. Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. Nat. Biotechnol. 39, 825–835 (2021).

15. Janssens, D. H. et al. CUT&Tag2for1: a modified method for simultaneous profiling of the accessible and silenced regulome in single cells. *Genome Biol.* **23**, 81 (2022).

16. Gopalan, S., Wang, Y., Harper, N. W., Garber, M. & Fazzio, T. G. Simultaneous profiling of multiple chromatin proteins in the same cells. *Mol. Cell* **81**, 4736-4746.e5 (2021).

17. Stuart, T. et al. Nanobody-tethered transposition enables multifactorial chromatin profiling at single-cell resolution. Nat. Biotechnol. **41**, 806–812 (2023).

18. Bartosovic, M. & Castelo-Branco, G. Multimodal chromatin profiling using nanobody-based single-cell CUT&Tag. Nat. Biotechnol. 41, 7**9**4–805 (2023).

19. Deng, Y. et al. Spatial-CUT&Tag: Spatially resolved chromatin modification profiling at the cellular level. *Science* **375**, 681–686 (2022).

20. Hu, S. S. et al. Intrinsic bias estimation for improved analysis of bulk and single-cell chromatin accessibility profiles using SELMA. Nat. Commun. 13, 5533 (2022).

21. Wang, M. & Zhang, Y. Tn5 transposase-based epigenomic profiling methods are prone to open chromatin bias. *bioRxiv* 2021.07.09.451758 (2021) doi:10.1101/2021.07.09.451758.

22. Thompson, M. D. & Byrd, A. K. Untargeted CUT&Tag reads are enriched at accessible chromatin and restrict identification of potential G4-forming sequences in G4-targeted CUT&Tag

experiments. *Nucleic Acids Res.* **53**, gkaf678 (2025).

23. Kaya-Okur, H. S., Janssens, D. H., Henikoff, J. G., Ahmad, K. & Henikoff, S. Efficient low-cost chromatin profiling with CUT&Tag. *Nat. Protoc.* **15**, 3264–3283 (2020).

24. Zhang, Y. et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biol. 9, R137 (2008).

25. Zang, C. et al. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* **25**, 1952–1958 (2009).

26. Meers, M. P., Tenenbaum, D. & Henikoff, S. Peak calling by Sparse Enrichment Analysis for CUT&RUN chromatin profiling. *Epigenetics Chromatin* **12**, 42 (2019).

27. Abascal, F. et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature **583**, 699–710 (2020).

28. Zheng, A. et al. Deep neural networks identify sequence context features predictive of transcription factor binding. *Nat. Mach. Intell.* **3**, 172–180 (2021).

29. Chen, K. M., Wong, A. K., Troyanskaya, O. G. & Zhou, J. A sequence-based global map of regulatory activity for deciphering human genetics. *Nat. Genet.* **54**, 940–949 (2022).

30. Koh, P. W., Pierson, E. & Kundaje, A. Denoising genome-wide histone ChIP-seq with convolutional neural networks. Bioinformatics **33**, i225–i233 (2017).

31. Hentges, L. D. et al. LanceOtron: a deep learning peak caller for genome sequencing experiments. Bioinformatics **38**, 4255–4263 (2022).

32. Vu, H. T. H., Zhang, Y., Tuteja, G. & Dorman, K. S. Unsupervised contrastive peak caller for ATAC-seq. *Genome Res.* **33**, gr.277677.123 (2023).

33. Port, S. A. et al. The Oncogenic Fusion Proteins SET-Nup214 and Sequestosome-1 (SQSTM1)-Nup214 Form Dynamic Nuclear Bodies and Differentially Affect Nuclear Protein and Poly(A)+ RNA Export*. *J. Biol. Chem.* **291**, 23068–23083 (2016).

34. Lai, S. et al. Quantitative Site-Specific Chemoproteomic Profiling of Protein Lipoylation. *J. Am. Chem. Soc.* **144**, 10320–10329 (2022).

35. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition

of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, **1213**–1218 (2013).

36. Ahlmann-Eltze, C., Huber, W. & Anders, S. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. Nat. Methods **22**, 1657–1661 (2025).

37. Chen, C. et al. Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). Science **356**, 18**9**–194 (2017).

38. Mumbach, M. R. et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nat. Methods 13, 919–922 (2016).

39. Wei, X. et al. HiCAR is a robust and sensitive method to analyze open-chromatin-associated genome organization. Mol. Cell **82**, 1225-1238.e6 (2022).

40. Lu, T., Ang, C. E. & Zhuang, X. Spatially resolved epigenomic profiling of single cells in complex tissues. Cell 185, 4448-4464.e17 (2022).

41. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).

42. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. Nat. Methods 12, 357–360 (2015).

43. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. 33, 290–295 (2015).

44. Seok, J., Xu, W., Gao, H., Davis, R. W. & Xiao, W. JETTA: junction and exon toolkits for transcriptome analysis. Bioinformatics 28, 1274–1275 (2012).

45. Hao, Y. et al. Integrated analysis of multimodal single-cell data. Cell **184**, 3573-3587.e29 (2021).

46. Stuart, T. et al. Comprehensive Integration of Single-Cell Data. Cell **177**, 1888-1902.e21 (2019).

47. Granja, J. M. et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nat. Genet. **53**, 403–411 (2021).

48. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).

49. Bunz, F. et al. Requirement for p53 and p21 to Sustain G2 Arrest After DNA Damage. Science **282**, 1497–1501 (1998).

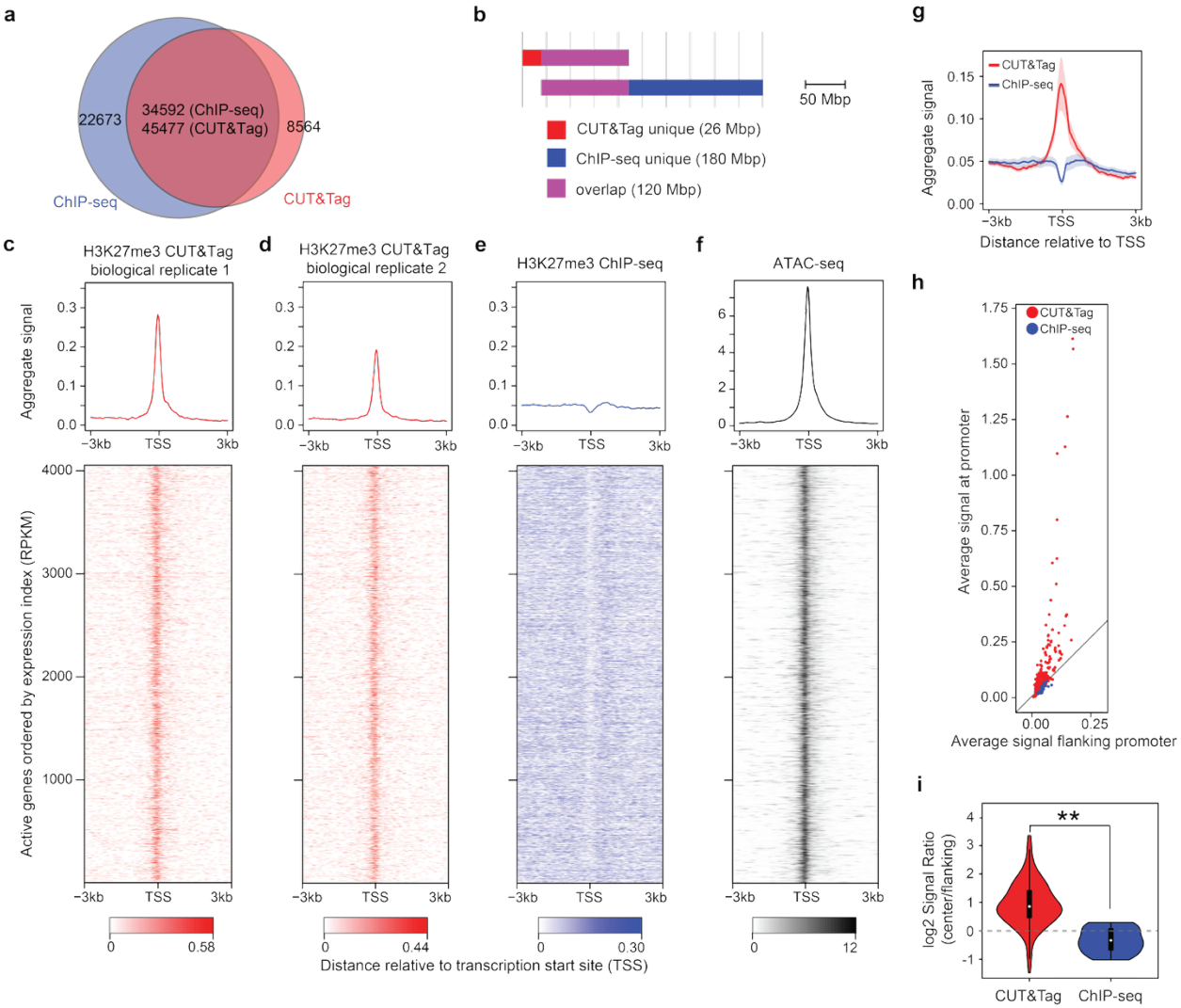

Figure 1. Aberrant CUT&Tag signals for H3K27me3. (**a**) Overlapping peak counts between CUT&Tag and ChIP-seq for H3K27me3 in human K562 cell line. (**b**) Genomic lengths of overlapping and non-overlapping peak regions between CUT&Tag and ChIP-seq for H3K27me3 in K562. (**c-f**) Signal patterns across active promoter regions in K562 cells, for two biological replicates of H3K27me3 CUT&Tag (**c**,**d**), H3K27me3 ChIP-seq (**e**), and ATAC-seq (**f**). The upper panels are the normalized aggregate signal patterns. The lower panels are heatmaps of the signal patterns at promoter regions (TSS ±3kb) of the actively transcribed genes. Rows correspond across heatmaps. (**g**) Aggregate H3K27me3 signal patterns at the promoter regions of 1000 ubiquitously expressed genes of 277 human CUT&Tag datasets (red) and 30 ENCODE human ChIP-seq datasets (blue). The shades represent the 95% confidence interval. (**h**) Average H3K27me3 signal levels at the proximal promoter region (TSS ± 300bp, y-axis) against the flanking regions (300bp – 3kb from TSS on both sides, x-axis) of the ubiquitously expressed genes for CUT&Tag (red) and ChIP-seq (blue) datasets. Each data point represents a dataset. (**i**) Distributions of log2-scaled ratio between the proximal promoter (center) region and the flanking region average signals for CUT&Tag and ChIP-seq datasets. **, $p < 0.01$, by one-sided Wilcoxon signed-rank test.

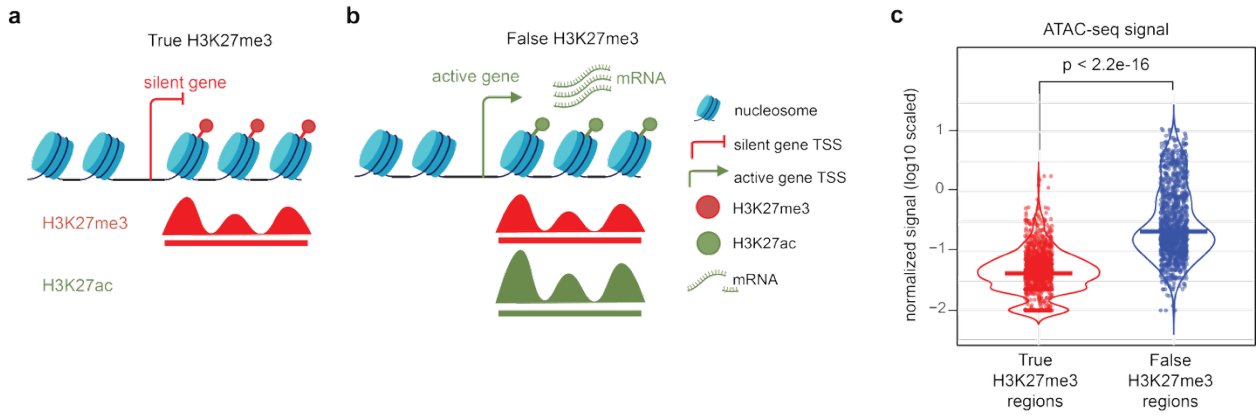

Figure 2. Selection of true and false H3K27me3 CUT&Tag signals. (**a**, **b**) Schematics of genomic regions with true (**a**) and false (**b**) H3K27me3 CUT&Tag signals. Red and green represent repressive and active chromatin features, respectively. (**c**) Distribution of chromatin accessibility level (ATAC-seq signal) on the true (red) and false (blue) H3K27me3 CUT&Tag signal regions. Each data point represents a 200-bp region. The centerline represents the median value. The P-value is calculated by one-sided Wilcoxon rank sum test.

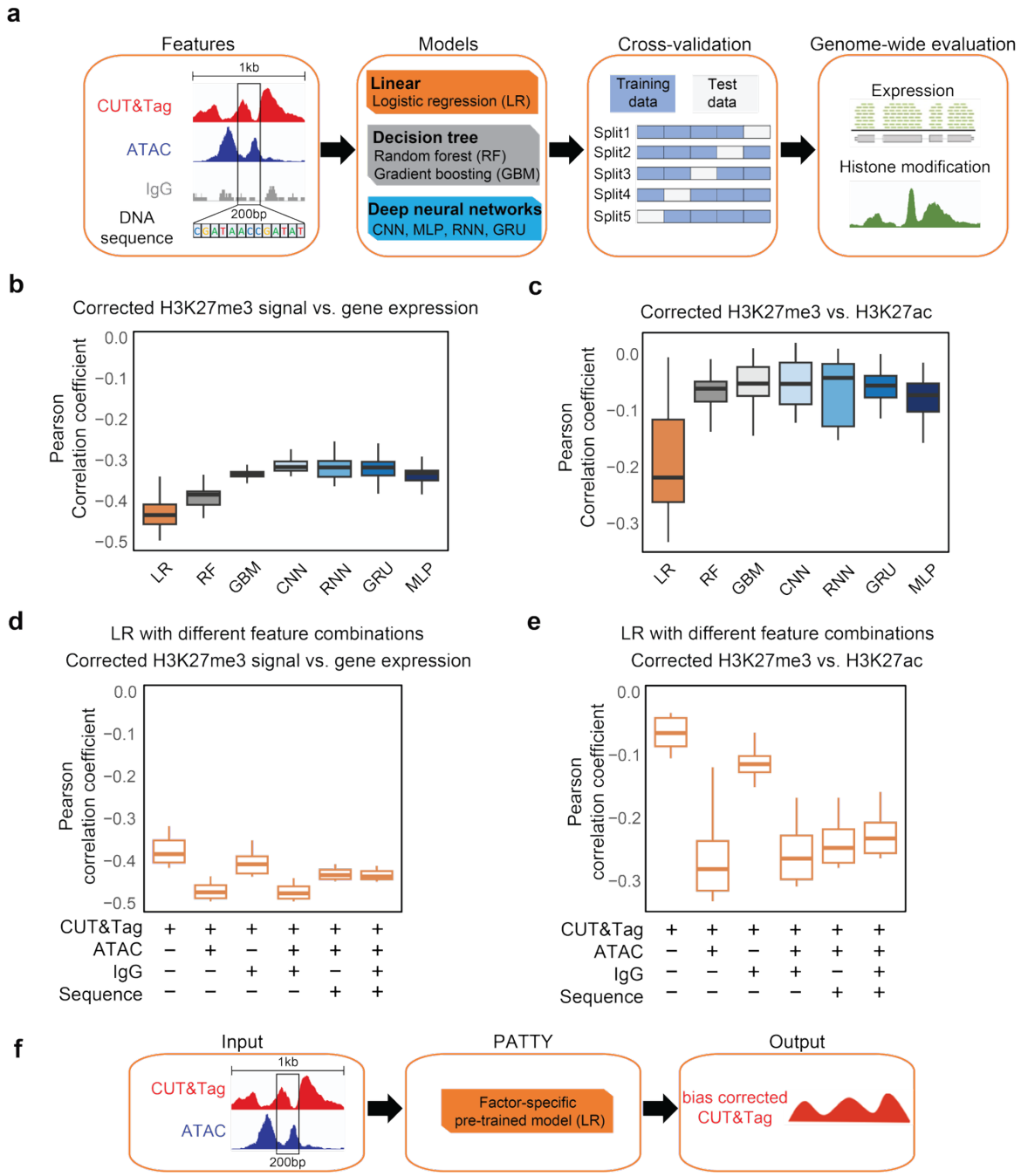

Figure 3. Machine learning models to classify true and false H3K27me3 regions with multi-dimensional features. (**a**) Workflow of model construction and evaluation. (**b**) Pearson correlation between the model derived H3K27me3 CUT&Tag score and gene expression across all gene promoter bins in the genome. (**c**) Pearson correlation between the model derived H3K27me3 CUT&Tag score and H3K27ac ChIP-seq signal level across all bins with CUT&Tag reads in the genome. Each data point in a boxplot represents a CUT&Tag sample from the model with a particular feature combination. (**d**, **e**) Similar to (**b**, **c**) respectively, for only the logistic regression model with various feature combinations labeled below the plots. The centerline, bounds of the box, vertical line bottom, and vertical line top of the boxplots represent the median, 25th to 75th percentile range, 25th percentile − 1.5 × interquartile range (IQR), and 75th percentile + 1.5 × IQR, respectively. Results from different feature combinations were combined into the same box. (**f**) Schematic of PATTY workflow.

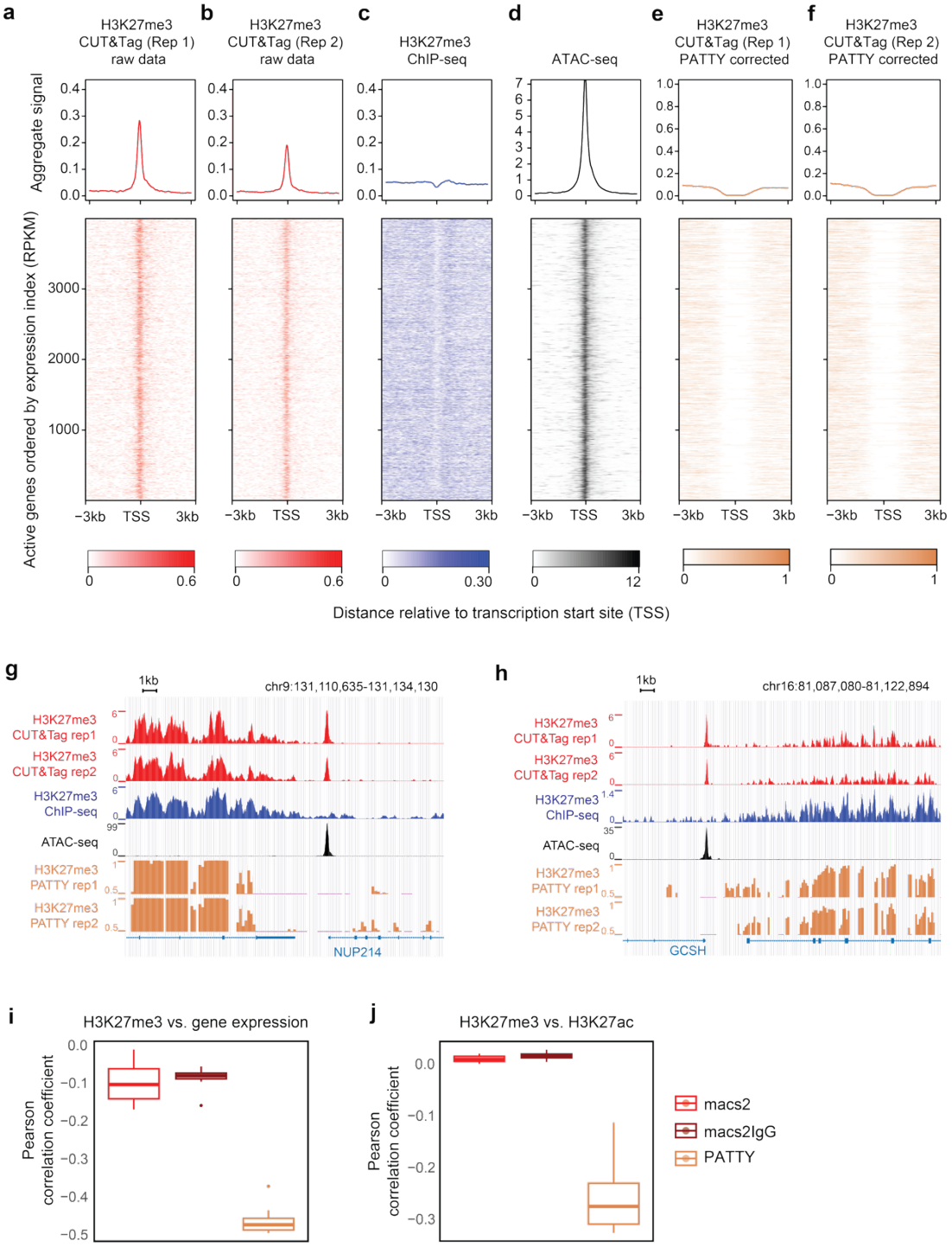

Figure 4. PATTY corrects intrinsic biases in H3K27me3 CUT&Tag data. (**a-f**) H3K27me3 CUT&Tag signal patterns across active gene promoter regions before (**a**, **b**) and after (**e**, **f**) bias correction by PATTY. ChIP-seq (**c**) and ATAC-seq (**d**) signals across the same regions are shown for reference (similar to **Figure 1e,f**). (**g**, **h**) Genome browser snapshots at the loci of two active genes, NUP214 (**g**) and GCSH (**h**), in K562. (**i**, **j**) Pearson correlation between H3K27me3 CUT&Tag signal, processed by various methods, and gene expression across all gene promoter bins in the genome (**i**) or H3K27ac ChIP-seq signal across all bins with CUT&Tag reads in the genome (**j**), in the K562 cell line. Each data point in a boxplot represents a different H3K27me3 CUT&Tag sample.

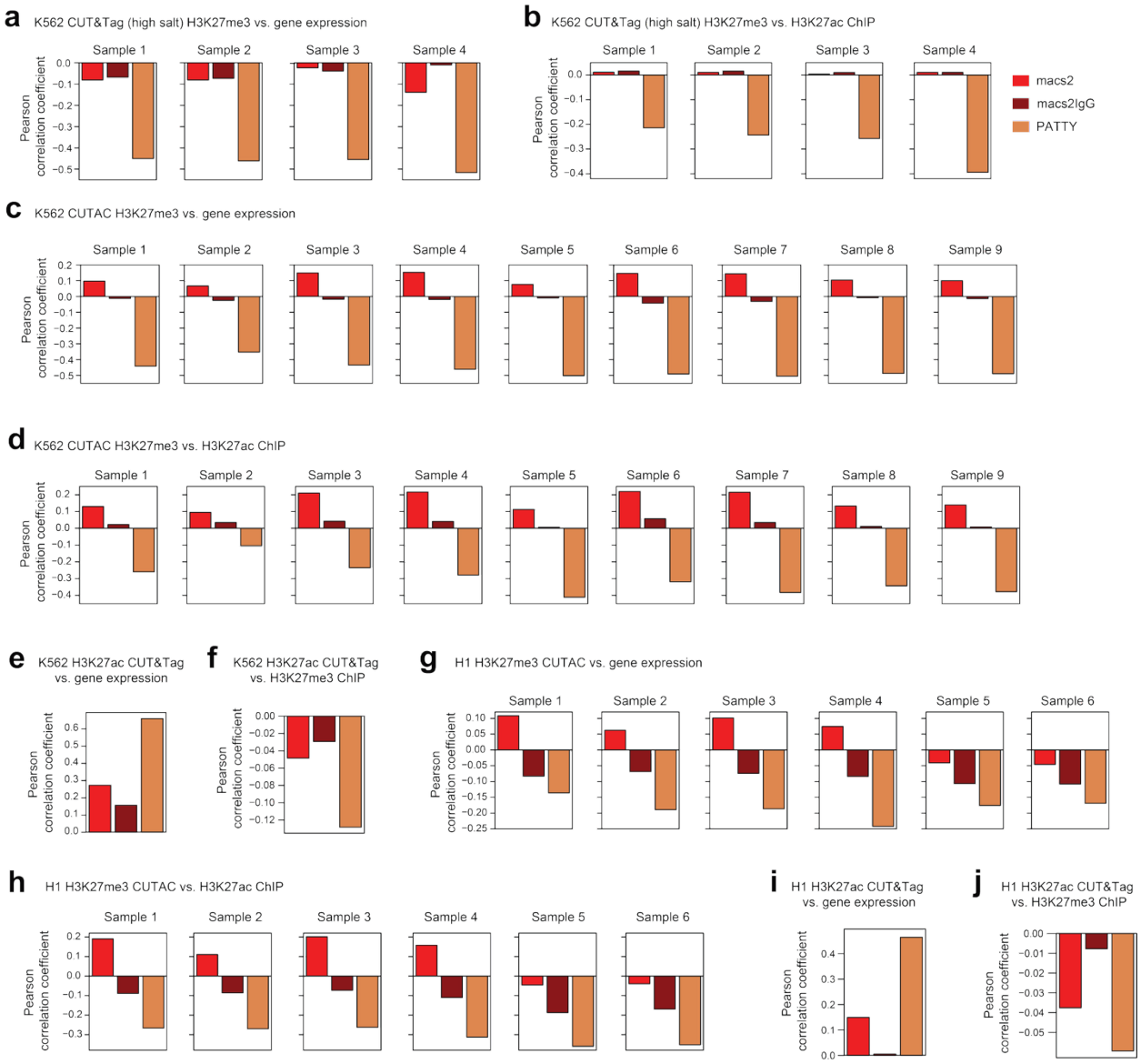

Figure 5. PATTY corrects CUT&Tag biases for H3K27me3 and H3K27ac across various samples and cell types. (**a-d**) Pearson correlation between processed H3K27me3 CUT&Tag signal and gene expression across all genes (**a**, **c**) or H3K27ac ChIP-seq signal across all 200-bp bins in the genome (**b**, **d**) in the K562 cell line. CUT&Tag samples in (**a**, **b**) were prepared in high-salt concentration. Samples in (**c**, **d**) were generated using the CUTAC assay. (**e**, **f**) Pearson correlation between processed H3K27ac CUT&Tag signal and gene expression across all genes (**e**) or H3K27me3 ChIP-seq signal across all 200-bp bins genome-wide (**f**), in K562. (**g-j**) Similar to (**a**, **b**, **e**, **f**) but for H1 human embryonic stem cell (hESC) line.

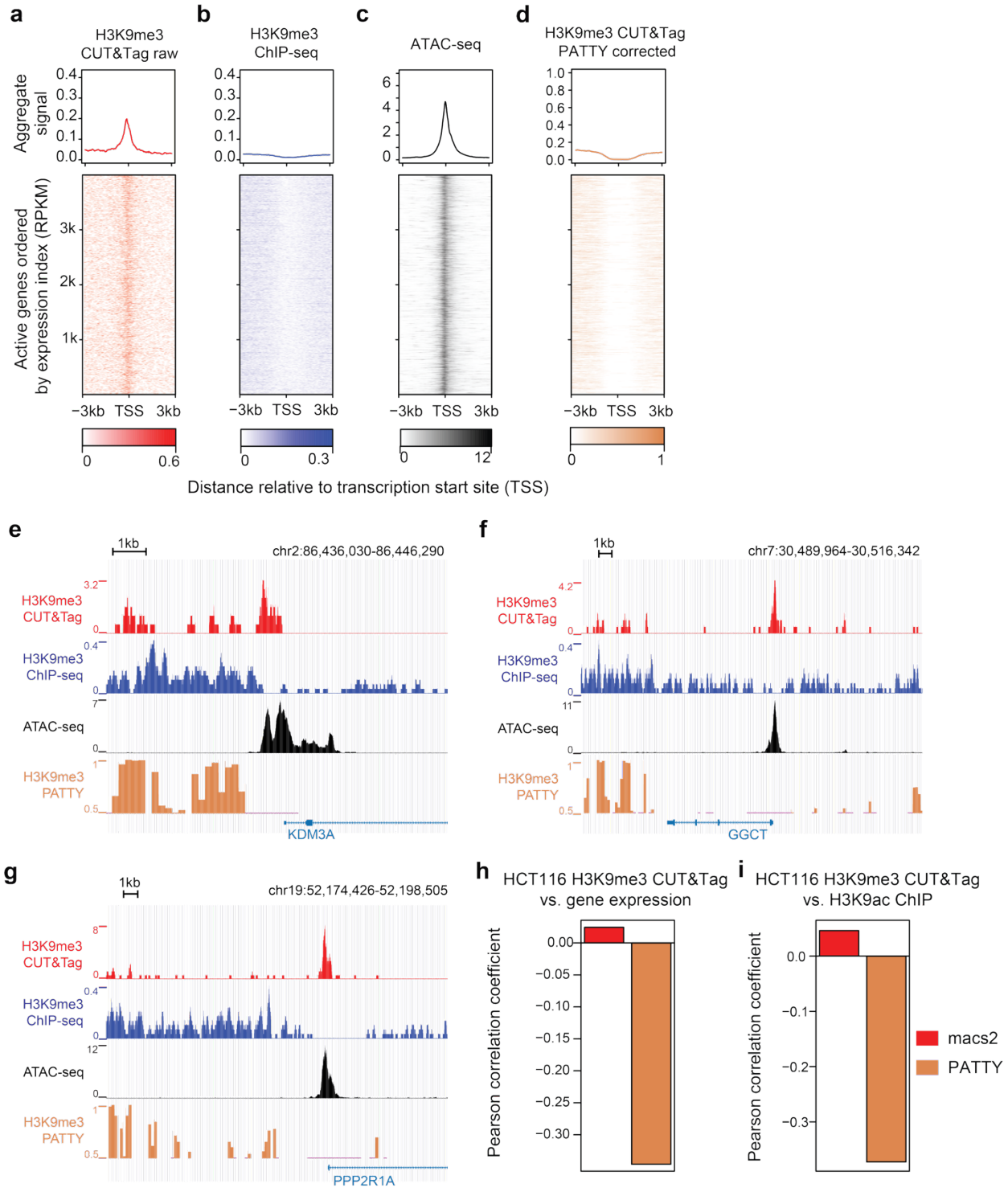

Figure 6. PATTY corrects intrinsic biases in H3K9me3 CUT&Tag data. (**a-d**) H3K9me3 CUT&Tag signal patterns across active gene promoter regions before (**a**) and after (**d**) bias correction by PATTY. ChIP-seq (**b**) and ATAC-seq (**c**) signals across the same regions are shown for reference (similar to **Figure 4a-f**), in human HCT116 cell line. (**e-g**) Genome browser snapshots at the loci of several active genes in HCT116. (**h**, **i**) Pearson correlation between H3K9me3 CUT&Tag signals and gene expression across all gene promoter bins in the genome (**i**) or H3K27ac ChIP-seq signal across all bins with CUT&Tag reads in the genome (**j**), in the HCT116 cell line.

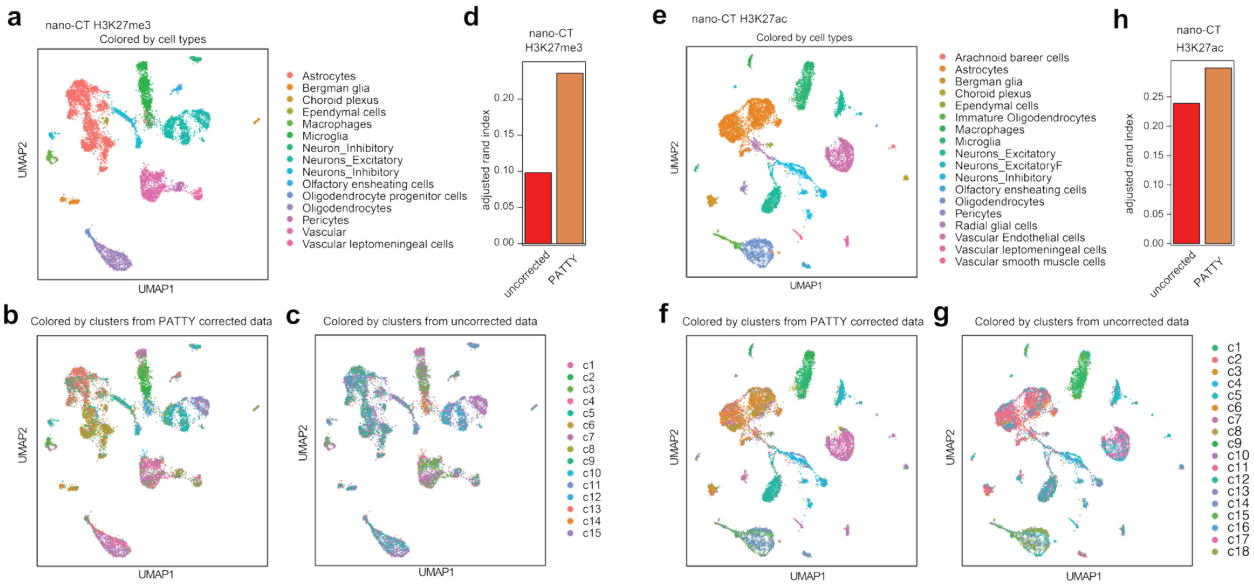

Figure 7. PATTY improves single cell clustering using nano-CT data. (**a**-**c**, **e**-**g**) UMAP visualization of H3K27me3 (**a**-**c**) and H3K27ac (**e**-**g**) nano-CT single cell data with cells colored by published cell type annotation as ground truth (**a**, **e**), cluster label from PATTY corrected data (**b**, **f**), and cluster label from uncorrected data (**c**, **g**). (**d**, **h**) Adjusted rand index between the clustering results and the ground truth cell type annotation for H3K27me3 (**d**) and H3K27ac (**h**) nano-CT data.

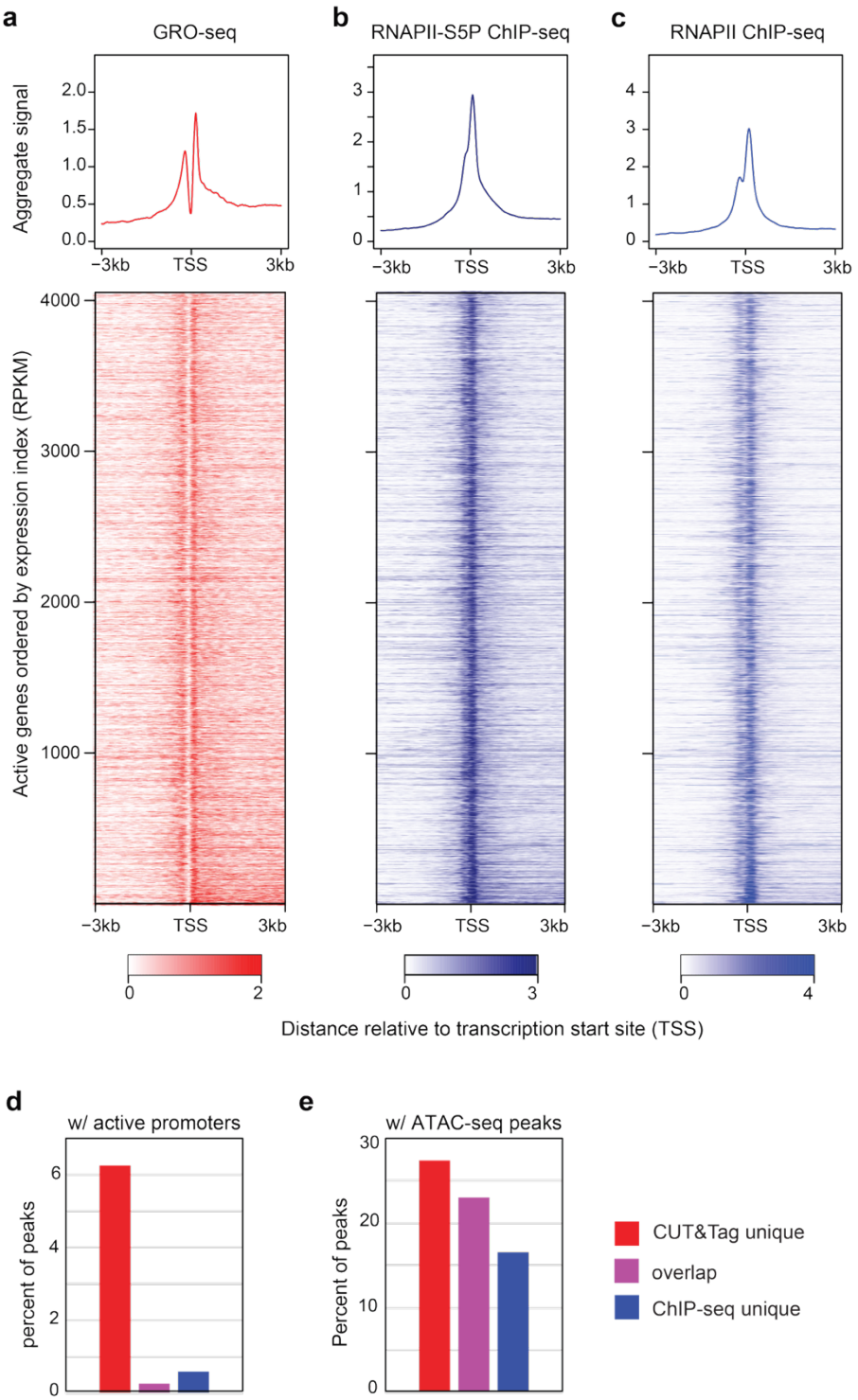

Figure S1. Enrichment of H3K27me3 CUT&Tag signal on active gene promoters. (**a**-**c**) Signal patterns across active promoter regions in K562 cells, for two GRO-seq (**a**), RNAPII-s5p ChIP-seq (**b**), and RNAPII ChIP-seq (**c**). The upper panels are the normalized aggregate signal patterns. The lower panels are heatmaps of the signal patterns at promoter regions (TSS ±3kb) of the actively transcribed genes. Rows correspond across heatmaps. (**d**, **e**) Percentage of CUT&Tag-unique (red), CUT&Tag-ChIP-overlap (purple), and ChIP-seq-unique (blue) peaks that overlap with active gene promoter regions (TSS ±1kb) (**d**) and with ATAC-seq peaks (**e**).

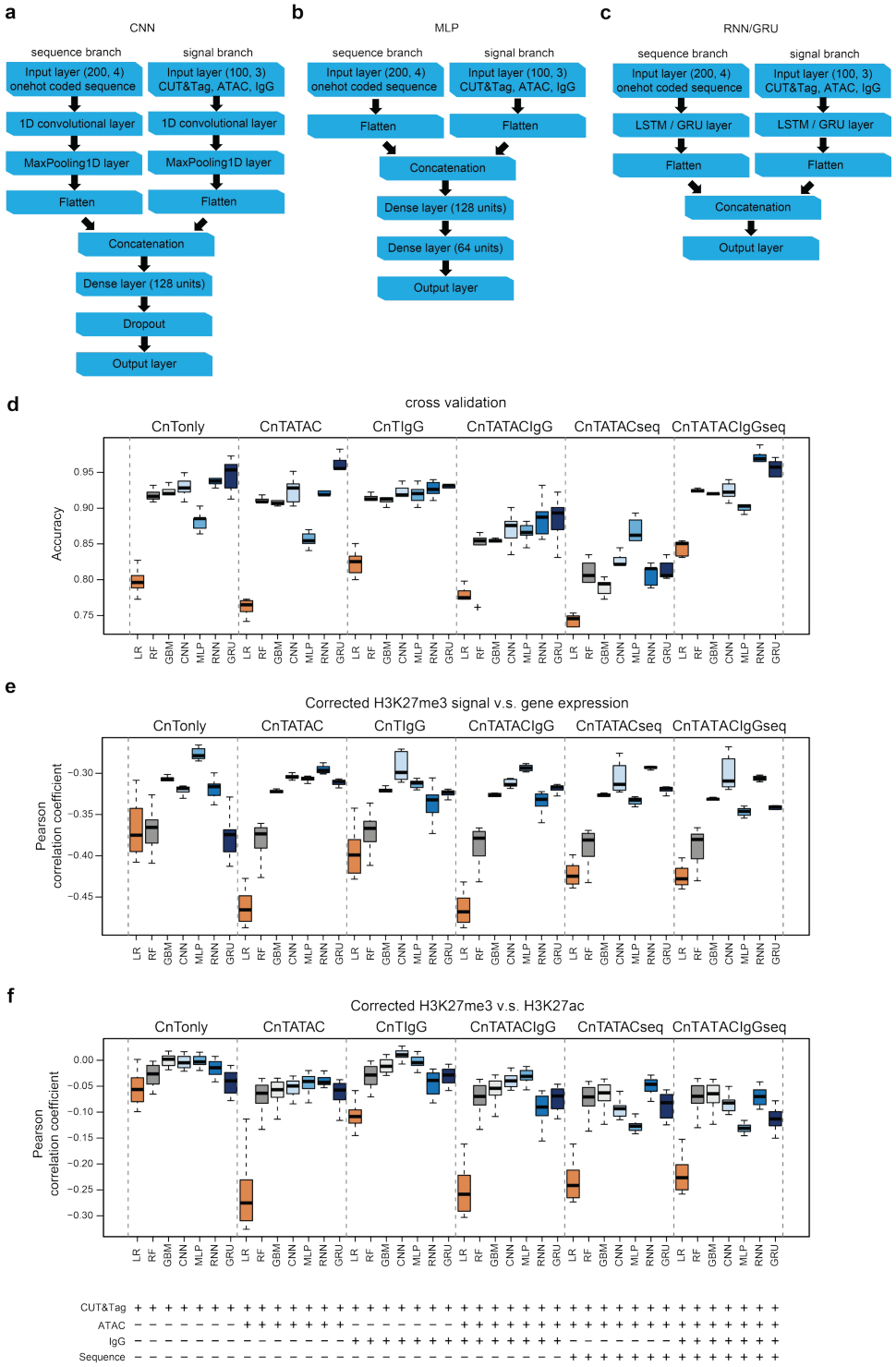

Figure S2. Performance of different models for H3K27me3 CUT&Tag. (**a-c**) Schematic of the (**a**) CNN, (**b**) MLP, (**c**) RNN, and GRU models. (**d**) The accuracy of different prediction models was evaluated using 5-fold cross-validation results. (**e**) Pearson correlation coefficient (y-axis) between the model-derived H3K27me3 score and gene expression across all gene promoter bins in the genome. (**f**) Pearson correlation between the model derived H3K27me3 CUT&Tag score and H3K27ac ChIP-seq signal level across all bins with CUT&Tag reads in the genome. Models with the same feature combination were displayed in the same block. Each data point in a boxplot represents a CUT&Tag sample.

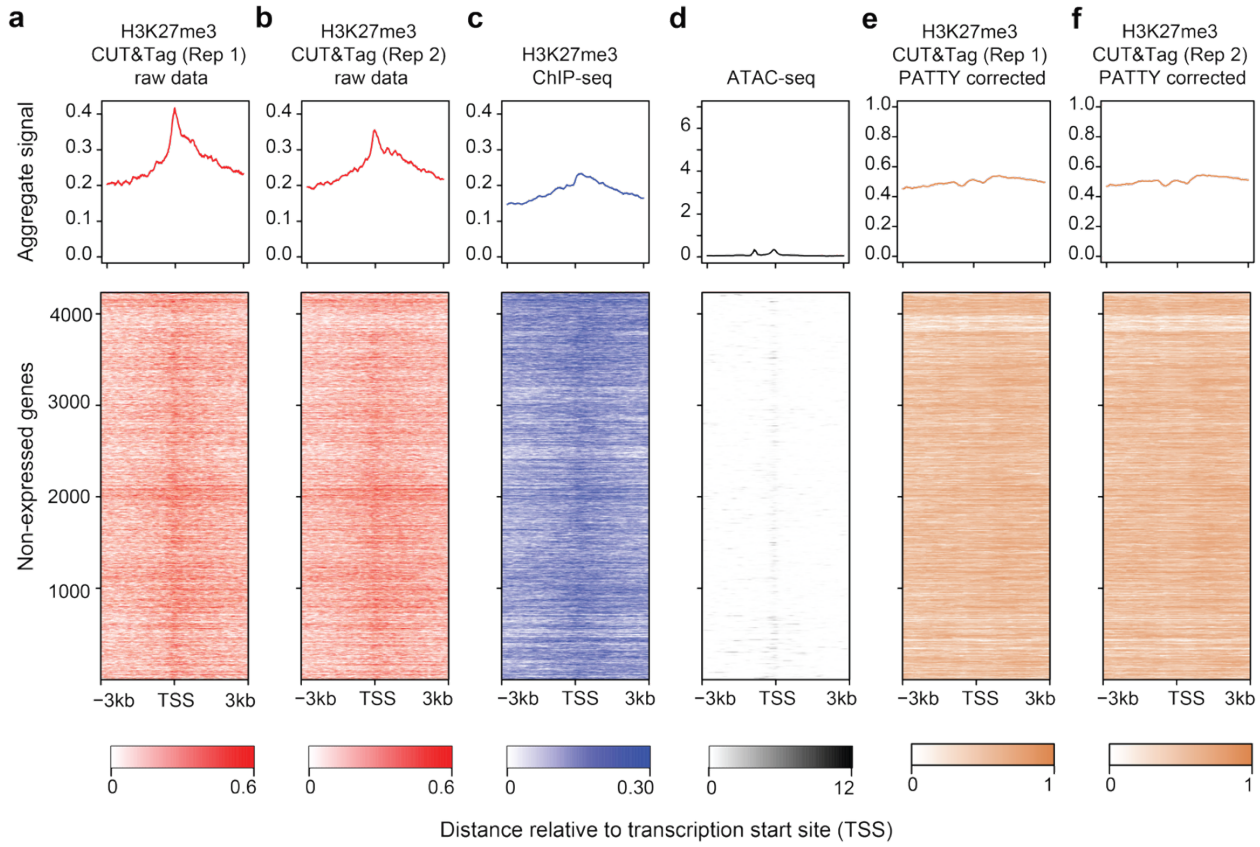

Figure S3. H3K27me3 CUT&Tag signal patterns on repressive genes before and after PATTY correction.

(**a**-**f**) H3K27me3 CUT&Tag signal patterns across repressive gene promoter regions before (**a**, **b**) and after (**e**, **f**) bias correction by PATTY. ChIP-seq (**c**) and ATAC-seq (**d**) signals across the same regions are shown for reference. The upper panels are the normalized aggregate signal patterns. The lower panels are heatmaps of the signal patterns at promoter regions (TSS ±3kb) of the repressive genes. Rows correspond across heatmaps.

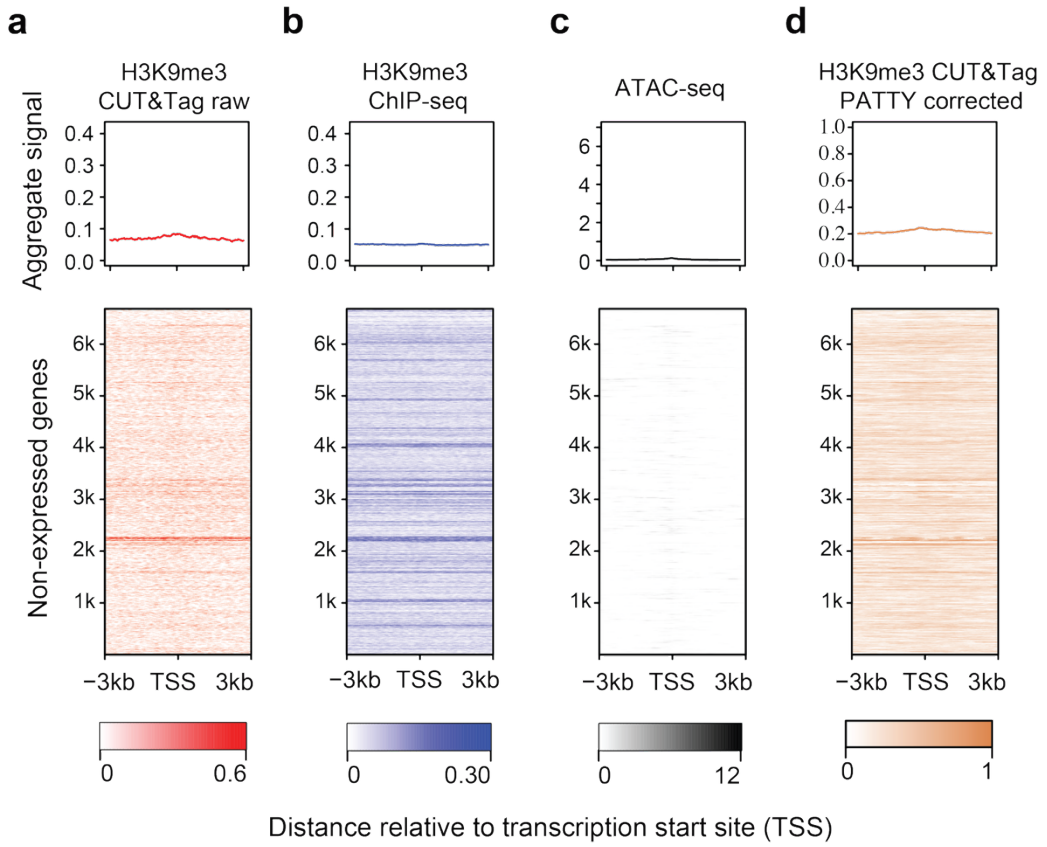

Figure S4. H3K9me3 CUT&Tag signal patterns on repressive genes before and after PATTY correction. Similar to **Figure S3** but for H3K9me3 CUT&Tag data across repressive genes in the HCT116 cell line**.**

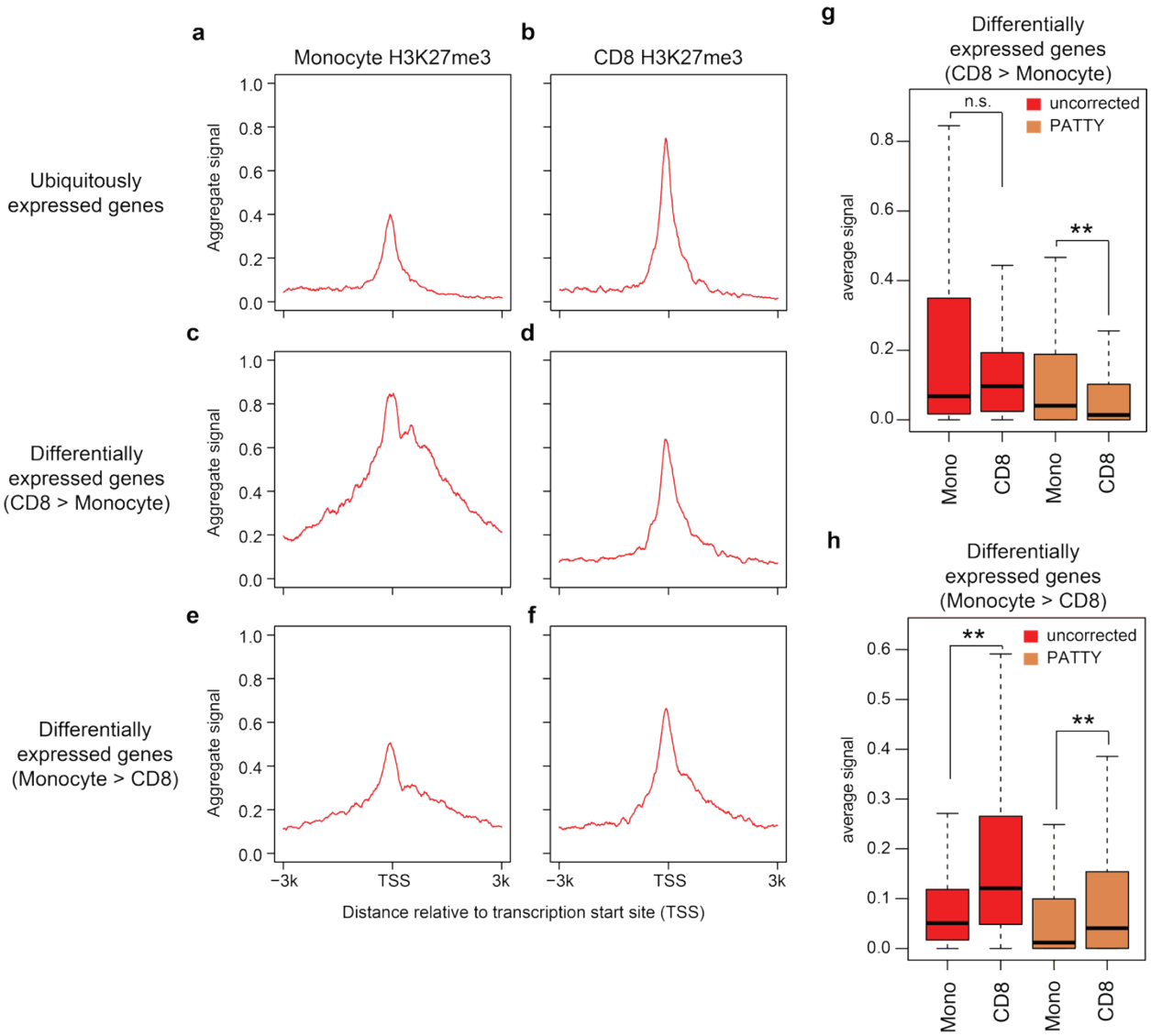

Figure S5. H3K27me3 CUT&Tag signal in scCUT&Tag-pro data. (**a**-**f**) Aggregate H3K27me3 scCUT&Tag-pro signal patterns across the promoter regions (TSS ±3kb) of PBMC-ubiquitously expressed genes (**a**, **b**), differentially expressed genes (DEGs) between Monocytes and CD8+ T cells that are higher in CD8 than Monocytes (**c**, **d**), and higher in Monocytes than CD8 (**e**, **f**), for scCUT&Tag-pro signal as pseudo-bulk in Monocytes (**a**, **c**, **e**) and in CD8+ T cells (**b**, **d**, **f**). (**g**, **h**) Average H3K27me3 scCUT&Tag-pro signal levels in a pseudo-bulk of cell group (labeled under x-axis) on the promoter regions (TSS ±3 kb) of DEGs high in CD8 (**g**) and DEGs high in Monocytes (**h**), comparing uncorrected data (red) with PATTY-corrected data (yellow). Each data point in a boxplot represents the H3K27me3 scCUT&Tag-pro signal for a gene. **, $p < 0.01$, by one-sided Wilcoxon signed-rank test.

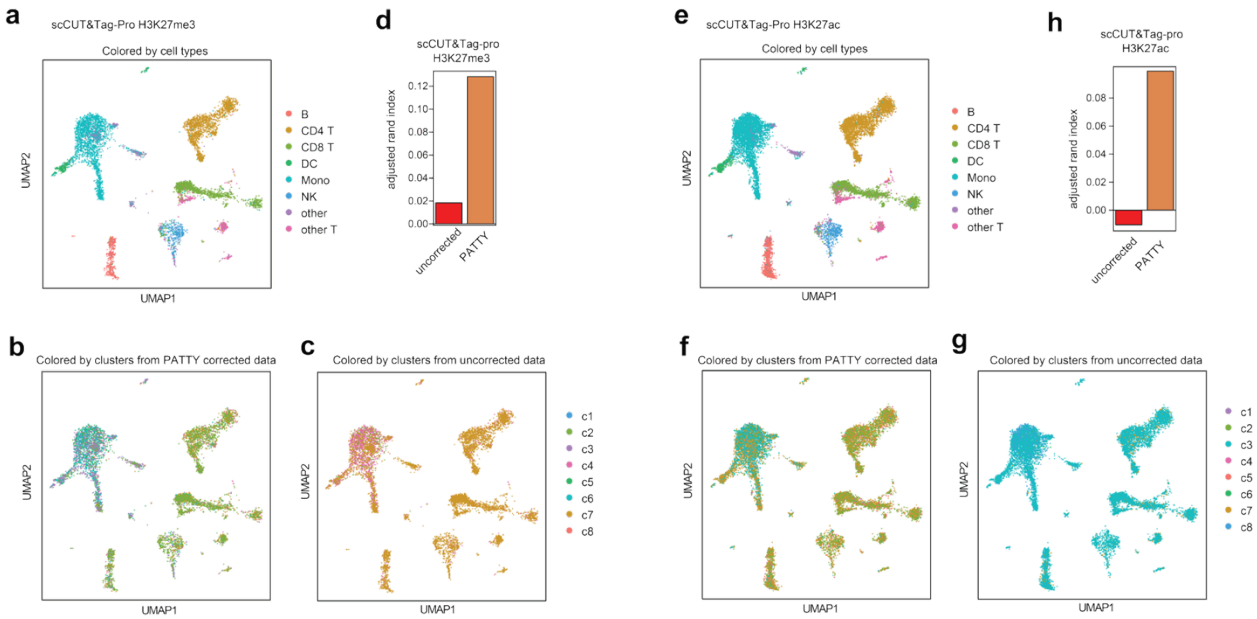

Figure S6. PATTY improves cell clustering of scCUT&Tag-pro. (**a**-**c**, **e**-**g**) UMAP visualization of H3K27me3 (**a**-**c**) and H3K27ac (**e**-**g**) scCUT&Tag-pro single cell data with cells colored by published cell type annotation as ground truth (**a**, **e**), cluster label from the PATTY corrected data (**b**, **f**), and cluster label from the raw count data (**c**, **g**). (**d**, **h**) Adjusted rand index between the clustering results and the ground truth cell type annotation for H3K27me3 (**d**) and H3K27ac (**h**) scCUT&Tag-pro data.