Nucleic Acids Research, 2024, **52**, e11
https://doi.org/10.1093/nar/gkad1146
Advance access publication date: 6 December 2023
Methods

JMnorm: a novel joint multi-feature normalization method for integrative and comparative epigenomics

Guanjue Xiang, Yuchun Guo, David Bumcrot and Alla Sigova*

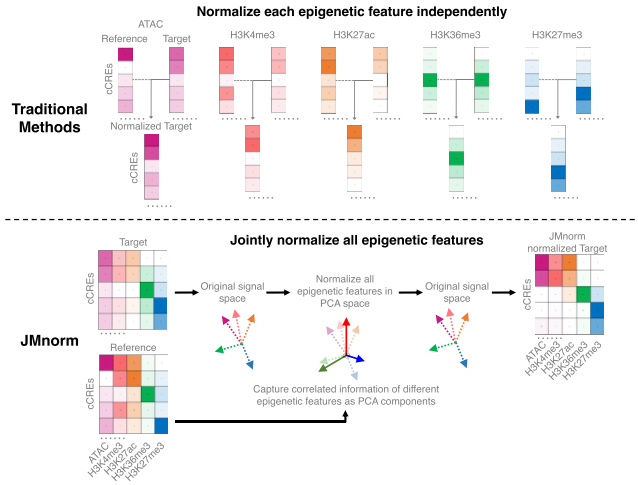
CAMP4 Therapeutics Corp., One Kendall Square, Building 1400 West, Cambridge, MA 02139, USA

*To whom correspondence should be addressed. Tel: +1 617 651 8867; Email: asigova@camp4tx.com

Abstract

Combinatorial patterns of epigenetic features reflect transcriptional states and functions of genomic regions. While many epigenetic features have correlated relationships, most existing data normalization approaches analyze each feature independently. Such strategies may distort relationships between functionally correlated epigenetic features and hinder biological interpretation. We present a novel approach named JMnorm that simultaneously normalizes multiple epigenetic features across cell types, species, and experimental conditions by leveraging information from partially correlated epigenetic features. We demonstrate that JMnorm-normalized data can better preserve cross-epigenetic-feature correlations across different cell types and enhance consistency between biological replicates than data normalized by other methods. Additionally, we show that JMnorm-normalized data can consistently improve the performance of various downstream analyses, which include candidate *cis*-regulatory element clustering, cross-cell-type gene expression prediction, detection of transcription factor binding and changes upon perturbations. These findings suggest that JMnorm effectively minimizes technical noise while preserving true biologically significant relationships between epigenetic datasets. We anticipate that JMnorm will enhance integrative and comparative epigenomics.

Graphical abstract



Introduction

Epigenetic features, including DNA accessibility and post-translational histone modifications, are thought to accurately reflect transcriptional states and help infer mechanistic insights about the regulation of gene expression in cell-type-specific contexts. Development of high-throughput sequencing techniques for genome interrogation led to the generation of hundreds of epigenetic datasets in different cell types and under various physiological conditions. Many large-scale data consortiums, such as the Encyclopedia of DNA Elements (ENCODE) and the ValIdated Systematic IntegratiON of hematopoietic epigenomes (VISION) projects, have utilized epigenetic features to identify candidate *cis*-regulatory elements (cCREs) (1–4). Follow-up studies characterized functional dynamics of epigenetic features across different cell types and conditions to elucidate their effects on transcriptional regulation (5–10). Increased sensitivity of the high-throughput sequencing methods results in amplified technical noise that can hinder the ability to extract biologically meaningful information. Therefore, to precisely quantify and compare epigenetic features across cell types, species, and experimental conditions, it is essential to develop robust epigenomic data normalization techniques to mitigate technical biases (11).

Received: July 14, 2023. **Revised:** October 25, 2023. **Editorial Decision:** November 12, 2023. **Accepted:** November 14, 2023

Many normalization methods have been developed for comparative analyses of high-throughput sequencing data (Supplementary Table 1). The two most widely used and easily implementable normalization methods are total library size normalization (TSnorm) and quantile normalization (QTnorm). TSnorm scales the signal of datasets to be compared based on the ratio of their total library sizes (12,13), whereas QTnorm transforms and equalizes signal distribution of each dataset relative to a reference distribution (14). The advantage of TSnorm is in its simplicity as it assumes that the only source of technical variation between datasets lies in differences in their sequencing read depths. QTnorm, on the other hand, can effectively remove complex technical biases by assuming different datasets share not only the same global mean but also the same global distribution. These assumptions hold true for certain types of high-throughput sequencing data such as bulk cell RNA-seq, where the majority of true biological signals are similar in different data sets. However, they are likely incorrect for epigenetic datasets with substantial variability in the number of peaks and signal intensities across different cell types (1,2) or experimental conditions (15), especially for epigenetic features with prominent signals at the cell-type-specific enhancers (1,2). Consequently, in comparative analyses of data with unequal number of biological peaks between cell types or conditions, both normalization methods tend to generate false positive or negative peak signals. Moreover, when comparing datasets produced under different treatment conditions inducing global epigenomic effects, QTnorm often distorts normalized signals relative to the true signals.

Other more specialized normalization methods, for example MAnorm and S3norm and their latest versions (16–19), are more adept for analysis of epigenetic datasets. These methods aim to eliminate biases due to differences in both sequencing depths and signal-to-noise ratios, which often arise from factors such as variations in antibody efficiency during ChIP-seq experiments (20). They utilize information from shared peak regions without/with common background regions and assume that the true signals in these regions remain consistent across different datasets and thereby, can serve as reliable anchors for data normalization. Similarly, a variant of the TSnorm method, referred to as TSnorm_cbg in this paper, computes a scaling factor using information from common background regions between two samples. It leverages the same concept as the normalization of ChIP-seq with control (NCIS) (21), which was specifically designed to address the challenge of normalizing global differences between ChIP and control datasets. However, the simple scaling factors or transformation models employed by these methods might not be adequate for addressing all technical biases, especially when they exhibit complex patterns, which are often observed in studies integrating datasets from multiple sources (3,22). Lastly, some methods can effectively model diverse signal distributions of epigenetic data by converting signals into ranks. However, these methods result in a loss of quantitative information (23).

Another significant drawback of existing approaches is that they analyze each epigenetic feature independently. These approaches may distort relationships between functionally correlated epigenetic features and hinder biological interpretation. Combinatorial patterns of multiple epigenetic features, known as epigenetic states, have been widely used for functional annotation of cCREs in different cell types, species and experimental conditions (3,24–29). Recent studies have shown that, while regulatory regions with specific epigenetic states may vary based on cellular and experimental contexts, cross-feature combinatorial patterns remain relatively conserved (4). Therefore, a normalization method that utilizes the information from functionally correlated epigenetic features could yield more accurate and biologically relevant post-normalization signals, enabling more meaningful comparison and integration of epigenetic data across conditions.

Here, we present a novel approach called Joint Multi-feature normalization (JMnorm) for simultaneous normalization of multiple epigenetic features across cell types, species, and experimental conditions by leveraging information from functionally correlated features. We demonstrate JMnorm's superior performance in preserving cross-feature correlations and improving consistency between biological replicates, as well as its better versatility and utility for various genomic applications relative to other methods. Additionally, JMnorm can increase consistency between normalized epigenetic features and orthogonal datasets, which we robustly validated across diverse types of epigenetic features, including transcription factor (TF) binding ChIP-seq and DNase-seq data.

Materials and methods

Data collection and preprocessing

We obtained epigenetic datasets primarily from two databases: the VISION (3,4,22,30–32) and the EpiMAP repository (33). For VISION datasets, we downloaded bigWig files containing average read counts for seven chromatin features (H3K4me3, H3K4me1, H3K27ac, H3K36me3, H3K27me3, H3K9me3, ATAC-seq) from 9 human and 16 mouse hematopoietic cell types. For EpiMAP datasets, we downloaded bigWig files containing average $-\log_{10}(P\text{-value})$ signal tracks for seven chromatin features (H3K4me3, H3K4me1, H3K27ac, H3K36me3, H3K27me3, CTCF ChIP-seq and DNase-seq) from 24 human cell type groups. Since EpiMAP signal tracks were originally mapped to the hg19 reference genome, we used the CrossMap package (34) with default settings to lift over these files to the hg38 reference genome. All other datasets used in this study were mapped to hg38. RNA-seq data for different cell types were also downloaded from the VISION project (log2TPM, quantile normalized) and the EpiMAP repository (log2FPKM, quantile normalized). The topologically associating domains (TAD) boundaries were downloaded from the VISION project website under the 'Hi-C' tab (https://main.genome-browser.bx.psu.edu/cgi-bin/hgTracks), and YY1 peaks (bed format) were downloaded from the Cistrome DB (http://cistrome.org/db/#/) (35,36). The links to the downloaded files and the Cistrome DB sample ID list of the downloaded YY1 peak files are provided in Supplementary Table 1. The DNase-seq data to obtain the number of DNase I Hypersensitive Sites (DHSs) in different cell types were downloaded from the Meuleman 2020 study (https://static-content.springer.com/esm/art%3A10.1038%2Fs41586-020-2559-3/MediaObjects/41586_2020_2559_MOESM3_ESM.xlsx) (37).

To prepare data for normalization and downstream analyses, we first acquired an epigenomic signal matrix for each biological replicate of each cell type, denoted as $X_{r,\,ct}$, where r is the replicate id and ct is the cell type. Each $X_{r,\,ct}$ matrix

is a *n*-by-*m* 2-dimensional matrix, where *n* is the number of cCREs and *m* is the number of epigenetic features. We note that not all cell types had datasets for all epigenetic features. The details and links to all available datasets in both species are provided in Supplementary Table 1. In the matrix $X_{r,\ ct}$, each element $x_{i,\ j}$ represents the average signal over the genomic region of *i-th* cCRE derived from the signal track of *j-th* epigenetic feature for the replicate-*r* of the cell type-*ct*. These epigenomic signal matrices ($X_{r,\ ct}$ matrics) were generated using the bigWigAverageOverBed script from the UCSC Genome Browser utilities (38) on the bigwig files of the corresponding replicates. In this study, we utilized the ENCODE cCRE regions (2) for epigenomic signal matrices, with cCRE lengths ranging from 150 to 350 bp. We used this cCRE list because it is one of the most comprehensive cCRE lists derived from uniformly preprocessed epigenomic datasets across a broad range of cell types. We also evaluated the normalization methods using an alternative strategy, which uses consecutive bins of fixed length (100 bp) along the genome.

In most of the downstream analyses, we averaged the signal matrices from all biological replicates of the same cell type (all $X_{r,ct}$ matrices of cell type-*ct*) to produce the epigenomic signal matrix for each cell type, denoted as X_{ct}, where X_{ct} is still a *n*-by-*m* matrix. Each element $x_{i,\ j}$ represents the mean of *i*-th cCRE signals across all biological replicates for the *j*th epigenetic feature of the cell type-*ct*. The only exception were the analyses performed for evaluation of signal consistency between replicates, where we kept the epigenomic signal matrices of distinct biological replicates separate and then normalized them independently (see the 'Evaluation of signal consistency between biological replicates' section).

Subsequently, we generated a reference signal matrix by averaging the epigenomic signal matrices of all cell types (each cell type has a *n*-by-*m* matrix X_{ct}) and used the resulting *n*-by-*m* matrix, denoted as X_{ref}, for all cell types to normalize against.

Generation of reference signal matrix

To normalize data across multiple cell types, we used the average signal across all cell types as reference for all cell types to normalize against. This is different from the strategy used in the normalization step of our previous genome segmentation analyses (3,17), which aimed to enhance the sensitivity of post-normalization signal by using the sample with the highest FRiP score as reference. By utilizing average signal tracks as the references in this study, the noise in individual datasets was averaged out, leading to a cleaner reference for different cell types to normalize against. Consequently, this reference selection strategy balances sensitivity and robustness, and enhances the performance of methods that are more sensitive to background noise in the reference, such as QTnorm. For each epigenetic feature, we first collected all datasets for that feature in all cell types, and then computed the average signals at each cCRE region across these cell types. The average signal vector for all cCREs was used as a reference signal vector for the specific feature. This process was performed for all features, and the resulting outputs were combined to generate the *n*-by-*m* reference signal matrix, denoted as X_{ref}.

To improve the cross-feature comparability and interpretability, we further equalized sequencing depths and signal-to-noise ratios of different features in the reference matrix using S3norm (17). Instead of using S3norm's default mode, which learns an exponential transformation model from common peak regions and common background regions of two signal vectors, we employed its cross-feature normalization mode. This mode leverages peak regions and background regions of each signal vector to learn a transformation model. The reason for this choice is that when normalizing two signal vectors of different epigenetic features, the assumption that common peak regions represent housekeeping epigenetic events with similar signal levels is not valid, particularly for features with opposing functions such as H3K27ac and H3K9me3. Specifically, we used the ratio between the top mean signal of 99% quantile of cCREs and the overall mean of all cCREs for each feature to learn the S3norm transformation model. After the normalization, the sequencing depth and the signal-to-noise ratios are equalized across all features in the reference matrix. This reference signal matrix is then used as the reference signal matrix for all cell types to normalize against in downstream analyses.

The details about JMnorm

JMnorm is designed to normalize the signal matrix of a target cell type or condition, X_{tar}, against a reference signal matrix, X_{ref}. When there are multiple cell types, each cell type is normalized against the reference data independently.

JMnorm consists of four key steps (Figure 1). In the 1st step, it transforms the signal matrix into mutually orthogonal principal component analysis (PCA) space to consolidate correlated components across the signal vectors of different epigenetic features into the same PCA dimension. To achieve this, JMnorm learns a PCA transformation model from the reference signal matrix and then applies it to transform the target signal matrices.

Specifically, given the reference signal matrix X_{ref} of *m* epigenetic features at *n* cCRE regions, the PCA transformation can be defined as:

$$PC_{ref} = X_{ref} \times V_{ref},$$

where X_{ref} represents the *n*-by-*m* reference signal matrix in original signal space, PC_{ref} is the *n*-by-*m* reference signal matrix in PCA space, and V_{ref} is a *m*-by-*m* rotation matrix that is learned from the reference signal matrix by using singular value decomposition method.

$$PC_{tar} = X_{tar} \times V_{ref},$$

where X_{tar} is the *n*-by-*m* target signal matrix in original signal space, PC_{tar} represents the *n*-by-*m* target signal matrix in PCA space. These scores are obtained by projecting the target signal matrix into the PCA space defined using the reference signal matrix by the V_{ref} rotation matrix. This ensures that the signal matrices of all target cell types can be transferred to the same reference PCA spaces. Moreover, the PCA model captures the cross-feature correlation information from the reference signal matrix, which is used and preserved throughout the subsequent steps.

Previous studies have demonstrated that there are reproducible combinatorial patterns across various chromatin features in different cell types or species, similar to the epigenetic states identified by various genome segmentation methods such as chromHMM (24), Segway (25) and IDEAS (26), which play a role in transcriptional regulation.

In the 2nd step, JMnorm leverages this prior knowledge by first clustering the cCREs into distinct groups based on the

reference signal matrix (reference clusters). Under the assumption that these reference clusters capture the conserved cross-feature patterns of distinct cCRE groups across different cell types, each cCRE is assigned, based on the target signal matrix, to one of the reference clusters.

Specifically, the reference clusters are generated using K-means clustering based on the reference signal matrix in PCA space. The number of reference clusters (K) is automatically determined by performing hierarchical clustering on a subset of data (20 000 cCREs by default), followed by the DynamicTreeCut method, which uses a branch cutting strategy to identify clusters in a dendrogram based on its shape (cutreeDynamic function with the following parameter settings: method='hybrid' and deepSplit = 2) (39). The resulting number of DynamicTreeCut clusters is then utilized as the K for K-means clustering on the reference signal matrix. The values of K vary depending on the specific combination and number of epigenetic features involved in the analysis. In our analyses with seven epigenetic features, the K values ranged from 39 to 40.

$$C_{ref} = Kmeans\left(PC_{ref},\ K\right),$$

where C_{ref} is a vector of length n (where n is the number of cCREs) that contains the K-means cluster labels for all cCREs. Each K-means cluster is denoted as h:

$$PC_{ref,h} = PC_{ref} | C_{ref} = h$$

$$AveragePC_{ref,\ h} = ColumnMean\ \left(PC_{ref,h}\right),$$

where $PC_{ref,h}$ is a $n_{ref,h}$-by-m matrix derived from the subset of the PC_{ref} that corresponds to cCREs associated with the K-means cluster h, $n_{ref,h}$ is the number of reference cCREs that belong to the K-means cluster h, and $AveragePC_{ref,\ h}$ is the average PCA vector for K-means cluster h that contains the column mean of the $PC_{ref,h}$ matrix. The length of $AveragePC_{ref,\ h}$ vector equals to the number of epigenetic features m. The $AveragePC_{ref,\ h}$ is computed for all K clusters. After that, all $AveragePC_{ref,\ h}$ vectors are consolidated and formed into a k-by-m matrix named $AveragePC_{ref}$.

Next, the target PCA signal matrix is first normalized to the reference PCA signal matrix using quantile normalization (initial QTnorm) to mitigate complex technical biases that could result in incorrect cCRE assignments:

$$PC_{tar,QT} = QTnorm\left(PC_{tar},\ PC_{ref}\right),$$

where $PC_{tar,QT}$ represents the n-by-m PCA signal matrix after initial round of QTnorm normalization, and $QTnorm()$ represents applying the QTnorm normalization method to PC_{tar}, wherein each column of PC_{tar} is adjusted to match the distribution of its corresponding column in PC_{ref}.

A n-by-k pairwise Euclidean distance matrix $Dist_{tar}$ between the $PC_{tar,QT}$ matrix and the $AveragePC_{ref}$ matrix is then computed. The element of $Dist_{tar}$at $(i,\ j)$ is defined as:

$$d_{i,h} = \sqrt{\sum_{m}\left(X_{tar,i,m} - X_{AveRef,h,m}\right)^2},$$

where $X_{tar,i,m}$ is the ith row of the $PC_{tar,QT}$ matrix, $X_{AveRef,h,m}$ is the jth row of the $AveragePC_{ref}$ matrix, m is the number of epigenetic features, $d_{i,h}$ represents the Euclidean distance between the ith row (ith cCRE) in target signal matrix and the cluster center of hth K-means cluster in PCA space.

Each cCRE in the target signal matrix is assigned to one of the reference clusters based on the smallest Euclidean distance, and a vector C_{tar} of length n (where n is the number of cCREs) that contains the cluster labels for all cCREs in the target signal matrix is generated. It is important to note that the initial QTnorm step might introduce noise at different PCA spaces in the data. However, since different principal components (PCs) are independent, and the noise is expected to randomly appear across different PCs for each cCRE, we reasoned that the majority of PCs would still contain accurate signals, enabling correct cCRE assignment.

In the 3rd step, JMnorm normalizes the target signal matrix against the reference signal matrix within each K-means cluster in the PCA space (within-cluster QTnorm). This is achieved by using QTnorm to normalize each PC separately, removing both simple and complex technical biases that may be present in the target signal matrix.

Specifically, for each K-means cluster h:

$$PC_{ref,h} = PC_{ref} | C_{ref} = h$$

$$PC_{tar,QT,h} = PC_{tar,QT} |\ C_{tar} = h$$

$$PC_{tar,JMnorm,\ h} = \ QTnorm\left(PC_{tar,QT,h}, PC_{ref,h}\right)$$

where $PC_{ref,h}$ is a $n_{ref,h}$-by-m matrix derived from the subset of the PC_{ref} that corresponds to cluster h, $PC_{tar,QT,h}$ is a $n_{tar,h}$-by-m matrix derived from the subset of the $PC_{tar,QT}$ that corresponds to cluster h, $n_{ref,h}$ is the number of reference cCREs in the cluster h, $n_{tar,h}$ is the number of target cCREs in the cluster h, $PC_{tar,JMnorm,h}$is the $n_{tar,h}$-by-m target signal matrix after QTnorm normalization against reference signal matrix within cluster h, and $QTnorm()$ represents applying QTnorm normalization to $PC_{tar,QT,h}$, wherein each column of $PC_{tar,QT,h}$ is adjusted to match the distribution of its corresponding column in $PC_{ref,h}$. Here, we assumed that cCREs within the same cluster shared the same epigenetic state in both reference and target, and thus had the same signal distributions. However, since the number of cCREs within each cluster are often different between reference and target, they are likely to have different global distributions. The cCRE clustering followed by within-cluster QTnorm is one of the key distinctions between JMnorm and traditional QTnorm. It effectively addresses the major limitation of QTnorm, which forces all cell types to have identical global signal distributions after normalization. The $PC_{tar,JMnorm,\ h}$ for all K clusters is then computed. After that, all $PC_{tar,JMnorm,\ h}$ matrices are combined to generate the n-by-$mPC_{tar,JMnorm}$ matrix.

In the 4th step, JMnorm reconstructs the normalized target signal matrix in the original signal space. To accomplish this, a dot product is performed between the normalized target PCA signal matrix $PC_{tar,JMnorm}$ and the transposed PCA rotation matrix $t(V_{ref})$ learned from reference signal matrix in the first step:

$$X_{tar,\ JMnorm} = PC_{tar,JMnorm}\ \times\ t\left(V_{ref}\right),$$

where $X_{tar,\ JMnorm}$ is the JMnorm-normalized target signal matrix in the original signal space, and $t()$ represents the transpose operation.

Quantification of the similarity of cross-feature correlations

We assessed the ability of various methods to preserve and transfer cross-feature correlation information from reference

signal matrices to post-normalization signal matrices of target cell types. Specifically, we calculated the cross-feature correlation matrix for each post-normalization signal matrix of the target cell types and the original reference signal matrix. We then computed the mean squared error (MSE) between each target cell type's correlation matrix and the reference correlation matrix. Lower MSE values indicate the normalization method can better preserve and transfer cross-feature correlation information from the reference signal matrix to the post-normalization signal matrices of the target cell types. For the cross-species normalization comparison, S3norm's cross-feature mode was employed for S3norm normalization that only uses the peak regions and background regions from the reference and target datasets in two species to learn the transformation model. Conversely, MAnorm was excluded because it requires the common peak regions between reference and target datasets to learn its transformation model, which was not available for cross-species normalization.

Quantification of the preservation of combinatorial patterns across different cell types

To evaluate the preservation of combinatorial patterns for different epigenetic features, we first combined the signal matrices of different cell types and applied a K-means clustering method to group the data from different cell types into distinct clusters. To ensure the robustness of this analysis, we used various numbers of clusters (K = 20, 30, 40, 50) in the K-means clustering step. After K-mean clustering, we quantified the mixing of various cell types' cCREs in the clustering results using the average silhouette width (ASW), a metric ranging from 0 to 1 (40). Lower ASW values indicate a better mixing of cCREs across different cell types within each cluster relative to the between cluster distance. Here, we assumed if a normalization method effectively preserves these patterns, pooling cCREs' epigenomic signal matrices from different cell types and then clustering them should yield a good mixture of cCREs from all cell types in the resulting clusters, and thus result in lower ASW scores. Specifically, we first pooled the epigenomic signal matrices of two cell types (X_{ct1} and X_{ct2}) into one matrix $X_{ct1\&ct2}$, where $X_{ct1\&ct2}$, is a $2n$-by-m 2-dimensional matrix. Then, we ran K-means clustering on the $X_{ct1\&ct2}$ matrix. For i-th cCRE belonging to cell-type-$ct1$ within the $X_{ct1\&ct2}$ matrix, its silhouette width s_i in K-means clustering result is defined as follows:

$$s_i = \frac{b_i - a_i}{\max(a_i,\ b_i)}$$

a_i is the average dissimilarity between the i-th cCRE and all other cCREs of cell-type-$ct1$ in the cluster k to which i-th cCRE belongs, and b_i is the average dissimilarity of the i-th cCRE to all cCREs of cell-type-$ct2$ in the same cluster k. The ASW is defined as the average of s_i of all 2n cCREs in the $X_{ct1\&ct2}$ matrix.

To measure the proportion of combinatorial patterns that are robustly identified across all cell types, we first performed K-means clustering (K = 30) for each cell type independently. Specially, for each cell type c, we first employed K-means clustering (K = 30) on its epigenomic signal matrix $X_{ct:c}$. This resulted in a vector of K-means cluster labels $C_{ct:c}$. We then selected cCREs (rows) from the matrix $X_{ct:c}$ that belonged to a specific cluster h. This subset matrix was denoted as:

$$X_{ct:c,\ h} = X_{ct:c} | C_{ct:c} = h$$

For these selected cCREs, we computed the average signal vector of all epigenetic features. This is represented as $\overline{X_{ct:c,\ h}}$, where $\overline{X_{ct:c,\ h}}$ is a vector (length equals to the number of epigenetic features m) with mean values from the columns of the $X_{ct:c,\ h}$ matrix (mean signal vector). In our analysis, m equals to 7, corresponding to the 7 epigenetic features in EpiMAP data. We then pooled all mean signal vectors $\overline{X_{ct:c,\ h}}$s of all cell types into one mean signal matrix X_{ct_mean} (a K*CT-by-m matrix, where K is the number of K-means clusters (K = 30), and CT is the number of cell types). We next clustered the rows in the X_{ct_mean} matrix using the second round of K-means clustering (K = 30) (Figure 3B). The output clusters containing at least one mean signal vector from all cell types were defined as robust clusters. We used the proportion of the robust clusters to evaluate different normalization methods. A higher proportion of robust clusters would indicate that the normalization method was more effective in generating consistent combinatorial patterns across different cell types. We repeated this process multiple times with different random seeds to ensure the robustness of our conclusions.

Cross-cell type RNA-seq predictions

Integration of various epigenetic features, such as DNA accessibility and histone modifications, has been demonstrated to be an effective approach for prediction of gene expression levels. Since RNA-seq data is independent from epigenomic data, we hypothesized that properly normalized epigenetic signal matrices could better preserve the combinatorial patterns of epigenetic events across different cell types, thus enhancing the transferability of RNA-seq prediction models based on these combinatorial patterns and improving the accuracy of gene expression predictions across different cell types. Following the model design proposed by Xiang *et al.* (3), we used the average signals of eight epigenetic features at both proximal regions (TSS ± 1 kb) and distal cCRE regions (TSS ± 500 kb, excluding proximal regions) for all genes as predictors to train regression models for RNA-seq log2 transcripts per million (log2TPM) predictions. The RNA-seq signals for different cell types were normalized by QTnorm against the average RNA-seq $\log_2$TPM across all cell types. To ensure the robustness of our evaluation, we tested three different models: a linear regression model (LM), a gradient boosting regression model (GBM) (41), and a linear regression model with gene grouping (LM-gene-grouping) (3). For gene grouping, we divided the gene set into four groups based on their average expression levels and standard deviations: (i) consistently low (mean < 0.5, sd < 0.2), (ii) differentially low (mean < 0.5, sd ≥ 0.2), (iii) differentially high (mean >= 0.5, sd ≥ 0.2), and (4) consistently high expression (mean ≥ 0.5, sd < 0.2) across cell types. For each evaluation run, we randomly selected 80% of protein-coding genes in one training cell type to train the model, then used the remaining 20% of protein-coding genes in another testing cell type for model evaluation. Model performance was evaluated using R^2, with higher R^2 values indicating more accurate cross-cell type RNA-seq predictions. We repeated this comparison multiple times with different random seeds to ensure the robustness of our conclusions.

Comparison of the post-normalization signal consistency between biological replicates

We reasoned that improved normalized signals would result in increased consistency between biological replicates. There-

fore, we first independently normalized the signals of different replicates. Specifically, we treated different biological replicates of the same cell type as if they were from distinct cell types. We then separately normalized the data of each replicate against the same reference data. This was implemented to ensure that no information from one biological replicate could affect or bias the normalization of another replicate. Then, we employed two metrics to quantify signal consistency: (i) the R^2 values, which were computed between the post-normalization signal vectors of each pair of biological replicates, and (ii) the Jaccard index values, which were calculated between the peak-calling results derived from the post-normalization signals of different replicates. Higher values for both R^2 and Jaccard indexes indicate better post-normalization signal consistency.

Peak calling from cCRE signal matrices

To compare normalization methods using peak calling results for CTCF ChIP-seq and DNase-seq, we employed an iterative *Z*-score-based approach, akin to the hotspot peak calling method (42,43), to call peaks from post-normalization cCRE signal matrices. For a specific epigenetic feature in a given cell type, we first converted normalized signals in all cCREs into *Z*-scores. We then selected cCREs with false discovery rate (FDR) adjusted *P*-values ≥ 0.1 to establish a second-round background model. Next, we recalculated the *Z*-scores and corresponding *P*-values for all cCREs using this updated background model. The cCREs with FDR adjusted *P*-values <0.1 were used as the output peak list for that chromatin feature in each cell type, which was ultimately utilized in downstream evaluations.

Given that the number of peaks can vary significantly depending on statistical thresholds and methods, we applied a non-parametric, rank-based method to identify the same number of CTCF peaks for each cell type across different normalization methods. Specifically, we first used the iterative *Z*-score-based strategy to call CTCF peaks for a specific cell type across different normalization methods. The number of peaks (N_{max}) for a specific cell type was determined based on the maximal peak count observed across the methods. Then, for different normalization methods, we used the top N_{max} cCREs based on the normalized CTCF signals as the CTCF peaks for the specific cell type.

Differential peak calling from cCRE signal matrices

As described in the 'Peak calling from cCRE signal matrices' section, for different normalization methods, we applied the rank-based method to identify the same number of differential peaks for DNase-seq datasets in different cell-types and glucocorticoid receptor (GR, gene symbol NR3C1) ChIP-seq datasets in A549 cell type with and without dexamethasone (Dex) treatment. Specifically, the top *N* peaks based on $\log_2$ fold change of DNase-seq signals in a pair of cell-types or an absolute value of $\log_2$ fold change between Dex-treated and untreated control conditions were defined as differential peaks and used for the downstream differential peak analysis.

Comparison between CTCF peaks and orthogonal data

To evaluate the ability of normalized signals to accurately capture true biological events, we compared CTCF peak sets derived from different post-normalization CTCF signal vectors with orthogonal data sets. For different normalization methods, the same number of CTCF peaks was generated using the iterative Z-score-based peak calling followed by a rank-based method, as detailed in the 'Peak calling from cCRE signal matrices' section. We then evaluated CTCF peaks using three types of orthogonal data: (1) YY1 (CTCF cofactor) peak (44,2) TAD boundaries (45–47) and (3) the CTCF binding site motifs (48,49).

We first compared the enrichment of CTCF peaks in YY1 peak regions. We reasoned that true CTCF peaks should exhibit greater enrichment in functionally related regions. Specifically, we divided the CTCF peaks into three distinct groups: (i) CTCF peaks shared by QTnorm/Harmony and JMnorm methods, (ii) CTCF peaks uniquely called from JMnorm data, (iii) CTCF peaks uniquely called from QTnorm/Harmony data. The enrichments were calculated as follows:

$$\exp_{CTCF\&YY1} = \text{obsCTCF} \times \frac{\text{YY1RegionSize}}{\text{GenomeSize}}$$

$$\text{enrichment}_{CTCF\&YY1} = \frac{\text{obs}_{CTCF\&YY1} + 1}{\exp_{CTCF\&YY1} + 1},$$

where (obsCTCF) represents the number of CTCF peaks in each group, (YY1RegionSize) represents the total number of base-pairs in the genome covered by YY1 peaks, (GenomeSize) represents the hg38 genome size, ($\text{obs}_{CTCF\&YY1}$) represents the observed number of CTCF peaks that intersect with YY1 peak regions by at least one base-pair. Due to the limited availability of YY1 peak datasets in some cell types, we pooled available YY1 peaks creating a single unified YY1 peak set for the YY1-CTCF peak intersection enrichment comparisons. Similarly, due to the availability issue for TAD boundary sets, we employed a unified TAD boundary region set for TAD boundary-CTCF peak intersection enrichment comparisons (Supplementary Table 1).

We then compared the proportion of CTCF peaks containing CTCF motif (Jaspar ID: MA0139.1) for the data normalized by different normalization methods. We used FIMO (50) to assess whether CTCF peak contained CTCF motif using a *q*-value threshold of $\Leftarrow$ 1e-03 for motif identification. The CTCF peak set was partitioned into the same three groups as described earlier for the CTCF-YY1 enrichment comparison.

Clustering cCRE based on cross-cell type DNase-seq patterns

We employed the Snapshot package with default settings to cluster cCREs based on DNase-seq signal patterns across different cell types. Snapshot was chosen for its ability to efficiently identify smaller clusters, automatically determine the ideal cluster number, and consider signal correlations across cell types when grouping cCREs into separate clusters using an indexing strategy (8). To quantify the signal-to-noise ratios for the Snapshot output clusters, we defined 'Signal' as the top quantile (100%, 90%, 80%, 70%, 60%) signal and 'Noise' as the bottom quantile (10%, 20%, 30%, 40%, 50%) signal in the meta-cluster heatmap. We calculated the signal-to-noise ratio using different combinations of 'Signal' and 'Noise' and corresponding statistical significance of the differences using the paired Wilcoxon test.

Identification of cell-type-specific DHSs

For each cell type within a cell type pair, we initially identified 10 000 cell-type-specific DHSs from the post-normalization DNase-seq signal vectors using different normalization methods, following the strategy detailed in the 'Differential peak calling from cCRE signal matrices' section. Then, using the *bedtools intersect* function with default parameters (51), these cell-type-specific DHSs were categorized into three groups: JMnorm-uniquely identified, QTnorm-uniquely identified, and those shared by both JMnorm and QTnorm methods.

Identification of differential glucocorticoid receptor (GR) ChIP-seq peaks after dexamethasone (Dex) treatment

For GR motif analyses and comparison of human phenotype term enrichments, we used the top 2000 differential peaks, identified based on the absolute value of the log2 fold changes between GR signals in A549 cells with and without Dex treatment. For different normalization methods, we analyzed the same number of differential GR ChIP-seq peaks. We computed GR motif (JASPAR ID: MA0113.1) scores in differential peaks as the $-\log_{10}P$-value using the FIMO (50) motif scanning algorithm. The method for calculating the human phenotype term enrichment score is described in the 'Human Phenotype term enrichment by GREAT analysis' section.

Human Phenotype term enrichment by GREAT analysis

To quantify enrichment of human phenotype terms for genes associated with distinct peak sets including cell-type-specific DHS and differential GR peaks post-Dex treatment, we utilized the rGREAT package (52,53). In this study, proximal regions were defined as TSS -5 kb to +1 kb, and distal regions were defined as proximal regions ± 100 kb. To focus on more specific terms, we excluded human phenotype terms linked with >1000 genes to eliminate overly general associations.

To compare the enrichment of human phenotype terms in cell-type-specific DHSs or differential GR ChIP-seq peaks after Dex treatment, we assumed that DHSs that are shared across normalization methods or differential GR ChIP-seq peaks that are identified using TSnorm_cbg corresponded to epigenetic signals representing true biological events. We identified the top 30 enriched human phenotype terms for these shared DHSs or differential GR ChIP-seq peaks. Next, for each normalization method, we quantified enrichments of the top 30 terms in cell-type-specific DHSs uniquely identified by each method and used these enrichments as a metric for method comparisons. Since relatively few differential GR peaks were unique to different normalization methods impairing the reliability of significance calculation of the human phenotype term enrichments, we compared performance of the normalization methods using enrichment of the top 30 terms in all differential GR peaks identified by each method instead.

Evaluation of computational efficiencies

To facilitate utilization of the JMnorm, we compared computation efficiencies of different normalization methods by measuring the running time and maximum memory usage for various datasets with different number of cCREs in an AWS m5.4xlarge instance. When processing 10 datasets, each comprising a signal matrix of 7 epigenetic features at one million cCREs, JMnorm completed the normalization in $\sim$9.5 min, consuming 10.9 GB of RAM (Supplementary Figure 1). In comparison, QTnorm completed normalization in $\sim$4 min, utilizing 2.4 GB of RAM. The increased maximum RAM usage observed with JMnorm arises from computing pairwise distance matrix for cCREs in the DynamicTreeCut method. It is worth noting that the RAM requirement does not escalate with the increased number of cCREs, because we have optimized this process by sampling a fixed number of rows (20000) from the data during this computation. Therefore, while JMnorm requires more computational resources and slightly longer running time relative to other methods, its requirements remain feasible for typical computer servers and laptops.

Using Harmony for cross-cell type multi-feature batch corrections

The Harmony method (54), originally designed for single-cell batch correction, transforms the input signal matrix to PCA space for batch correction. During this process, it modifies the signal matrices of both reference and target cell types in each run. As a result, in order to correct technical biases for signal matrices across cell types against a single reference matrix, both reference matrix and matrices for all cell types would need to be included in a single Harmony run, which would consume an excessive amount of computational resources. To address this issue, for each Harmony run, we assigned five times greater weight to the reference signal matrix than to the signal matrix of an individual target cell type. We reasoned that potential technical signal biases of the reference signal matrix would outweigh the biases of individual target cell type. This approach allowed for batch correction of data across all cell types against the same reference signal matrix without requiring the inclusion of all cell types' data in a single Harmony run.

Results

JMnorm overview

The goal of JMnorm is to simultaneously normalize multiple epigenetic features across cell types, species, or experimental conditions by leveraging information from functionally correlated features. The input for JMnorm is signal matrices comprising data for multiple epigenetic features for a desirable number of regulatory regions in two or more cell types or experimental conditions (reference and target cell types or conditions). To develop the method, we utilized human and mouse ENCODE cCREs (2), comprehensive collections of genome-wide regulatory elements generated in multiple cell types in the two species.

The method consists of four key steps (Figure 1). In step 1, orthogonal transformation, JMnorm converts the correlated components of multi-dimensional epigenetic signal matrices for both reference and target cCREs into mutually orthogonal PCA dimensions. This process generates corresponding PCA matrices, effectively preserving the relationships between functionally related epigenetic features (Figure 1A) (55). In the subsequent steps 2 and 3, JMnorm performs data normalization within the PCA space, which simultaneously normalizes all features and transfers the cross-feature patterns from

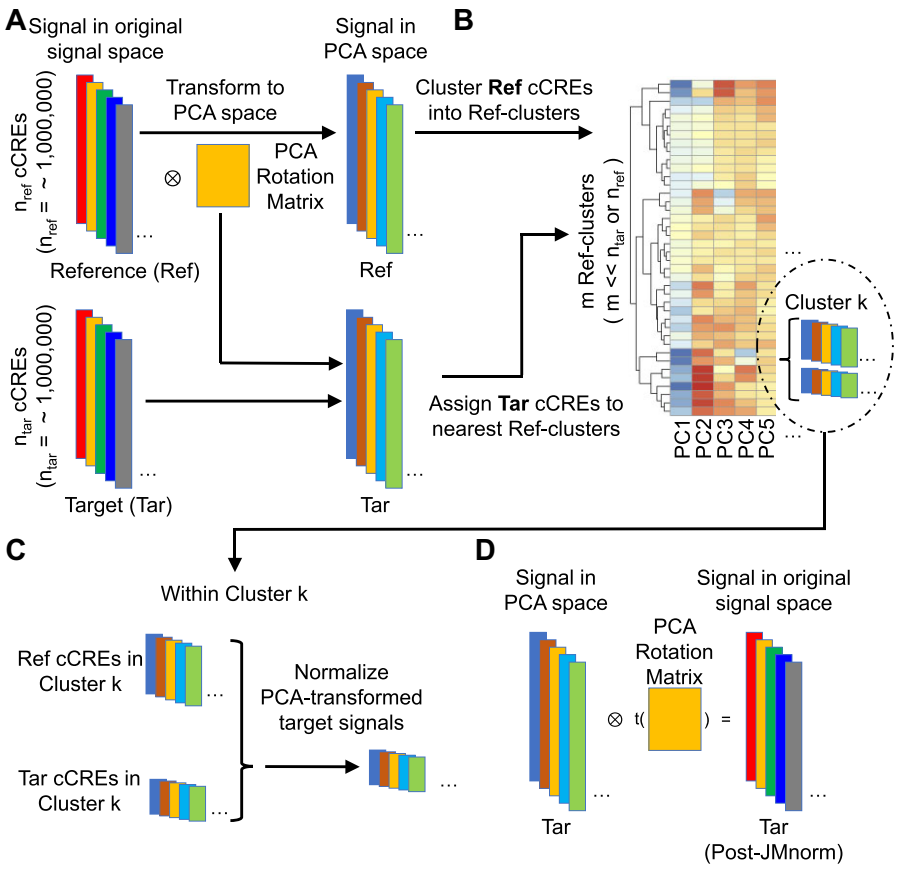


Figure 1. An overview of the four key steps in the JMnorm normalization procedure. (**A**) Step 1: orthogonal transformation. The correlated components of various epigenetic signals are transformed into mutually independent high-dimensional PCA dimensions. Each colored block on the left represents the signal vector of all epigenetic features at the n_{ref} or n_{tar} cCRE regions in reference or target samples, respectively. Colored blocks on the right denote corresponding transformed PCA epigenetic signal matrices for reference and target samples. The yellow box in the middle represents the PCA rotation matrix learned from the reference signal matrix. (**B**) Step 2: cCRE clustering. Reference cCRE clusters are generated based on the reference data in the PCA space with the average signal reference matrix shown as a heatmap. The Reference cCRE clusters were generated by *K*-means clustering method ($K = 40$). The K was determined automatically through hierarchical clustering followed by the DynamicTreeCut method (39). Detailed description of these procedures is provided in the Materials and Methods section. Target cCREs are assigned to reference clusters according to the Euclidean distances between the signal vector of the target cCRE and the average signal vectors of reference clusters in the PCA space. Within each cluster, the number of cCREs, shown as colored blocks within the insert, may vary between the reference and target samples. (**C**) Step 3: within-cluster normalization. Target signal matrix is normalized against the reference matrix using within-cluster quantile normalization as shown for Cluster *k*. (**D**) Step4: reconstruction of the JMnorm-normalized target signal matrix in the original signal space. The yellow box in the middle indicates the transposed PCA rotation matrix learned in the first step (panel A).

the reference signal matrix to the post-normalization signal matrices of the target cell types or conditions. We anticipate that this strategy can improve the performance of downstream prediction models that utilize the cross-feature patterns for tasks such as gene expression prediction (56–58), enhancer–promoter interaction prediction (59–61), or epigenomic data imputation (62–66), especially for cross-cell-type predictions. Specifically, in step 2, cCRE clustering, JMnorm first clusters (N) cCREs based on the reference PCA matrix into (M) reference clusters ($M << N$). It then assigns each of the individual target cCREs into the nearest reference cluster based on signal similarity in the PCA space (Figure 1B). In step 3, within-cluster normalization, JMnorm normalizes the target cCRE signals (PCA-transformed) to the corresponding reference cCRE signals (PCA-transformed) using quantile normalization within each cluster (Figure 1C). When performing cCRE clustering (step 2) followed by within-cluster QTnorm (step 3), we assume that (1) the combinatorial patterns of epigenetic features are conserved across cell types or conditions and (2) within each cCRE cluster, the signal distributions of PCA matrices are also conserved across cell types or conditions. It is important to note that the number of target cCREs assigned to each reference cluster can vary for different target datasets, thereby resulting in different global signal distributions across cell types or conditions after normalization. Therefore, this strategy can circumvent potential biases inherent in QTnorm, which forces identical global signal distributions across diverse datasets. Once steps 1–3 are completed, JMnorm transforms the normalized target signal matrix from the PCA space back to the original signal space in step 4 (Figure 1D). The details of these steps can be found in the Materials and Methods section.

Evaluation of preservation of cross-feature correlation

We evaluated the performance of JMnorm relative to a panel of other normalization methods (Supplementary Table 2) (TSnorm, TSnorm_cbg, MAnorm, S3norm, QTnorm), which

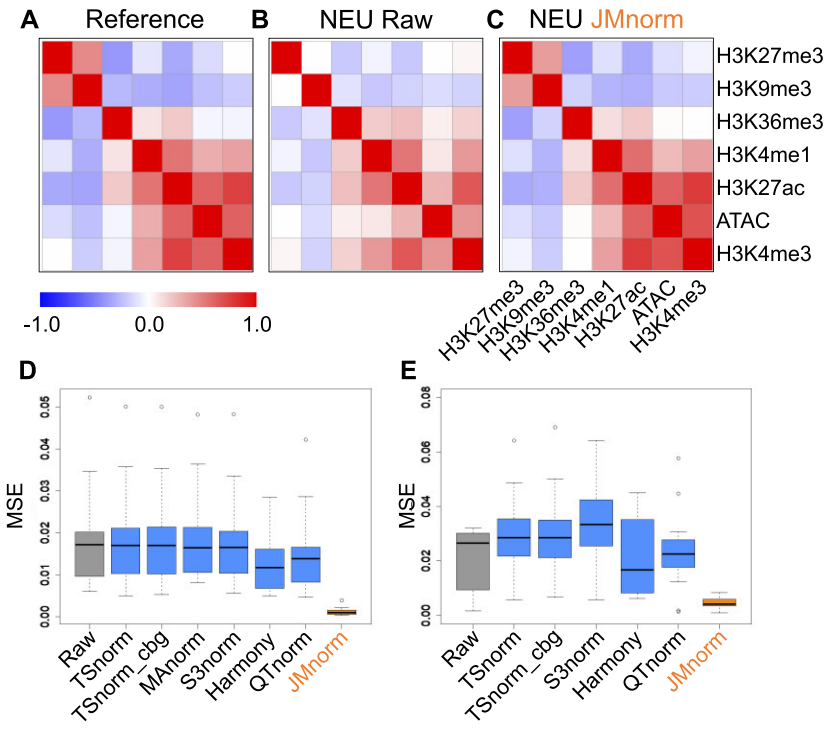


Figure 2. Evaluation of cross-feature correlation preservation. (**A**) Reference signal cross-feature correlation matrix. (**B**) Raw signal cross-feature correlation matrix in target neutrophil (NEU) cell type. (**C**) JMnorm-normalized signal cross-feature correlation matrix in target NEU cell type. (**D**) A boxplot of MSEs between human target and reference correlation matrices for the seven normalization methods. (**E**) A boxplot of MSEs between correlation matrices of mouse target cell types and the human reference for the six normalization methods.

are frequently applied as initial data normalization strategies for various downstream analyses. We also included Harmony, a widely used batch correction method for single-cell data. Similar to JMnorm, Harmony utilizes PCA transformation to harness information from correlated features within multi-dimensional data (54).

The first evaluation was to compare the performance of JMnorm and other normalization methods for their ability to preserve the cross-feature correlation among different epigenetic features across multiple cell types. While it is possible to use data from one particular cell type as a reference for normalizing data from other cell types, this approach could introduce biases from the data of the chosen cell type. To circumvent such potential biases, we generated a reference dataset by averaging each epigenetic feature across all cell types. Next, we used the selected methods to normalize raw signal matrices of target cell types against the reference signal matrix. The resulting post-normalization signal matrices were then used to compute the cross-feature correlation matrices.

We first inspected the cross-feature correlation matrices derived from the reference, the raw data, and JMnorm-normalized data from neutrophil (NEU) cells, respectively (Figure 2A, B, and C). We observed that the JMnorm-derived correlation matrix was more similar to the reference correlation matrix than to the correlation matrix of the raw data. Specifically, correlation matrices derived from both JMnorm and reference signal matrices exhibited strong positive correlations between features within the active chromatin feature group (H3K27ac, H3K4me3, H3K4me1, ATAC-seq) and the repressed chromatin feature group (H3K27me3 and H3K9me3) (Figure 2A and C). Conversely, we observed a substantial negative correlation between features in active and repressed chromatin groups (Figure 2A and C) (1,2). In contrast, the aforementioned correlation relationships are much weaker in the raw data (Figure 2B), likely due to the technical biases in the raw data. These results suggest that JMnorm is effective in reducing technical biases and preserving and transferring the cross-feature correlation information, which is better aligned with our prior knowledge, from the reference to the target cell type.

We then compared the panel of methods by measuring the MSEs between cross-feature correlation matrices of the normalized data and the reference across different cell types. JMnorm-derived correlation matrices better preserved the information in the reference correlation matrix than those derived from other methods, as indicated by its significantly lower MSEs (paired Wilcoxon test between JMnorm and the second-best performing method, P-value = 3.81e-06; Figure 2D).

Furthermore, the superior performance of JMnorm also holds true in cross-species normalizations, when mouse cCRE signal matrices of different cell types were normalized against the human reference signal matrix. Specifically, JMnorm produced significantly lower MSEs than other methods (paired Wilcoxon test between JMnorm and the second-best performing method, P-value = 2.13e-04) when comparing the correlation matrices of each mouse target cell type to the human reference (Figure 2E).

Since experimental data for large panels of epigenetic features might not always be available, we also assessed JMnorm's performance with fewer features. As demonstrated in Supplementary Figure 2, JMnorm can effectively preserve and transfer cross-feature correlation information when normalizing datasets containing as few as two epigenetic features.

Evaluation of consistency of cross-feature combinatorial patterns across cell types

Combinatorial patterns of epigenetic features, such as the epigenetic states identified in genome segmentation analysis (27), reflect functionally relevant interactions between DNA accessibility, histone modifications, and transcription factor binding under specific cellular and experimental conditions. They are often used to accurately infer transcriptional states and interpret the function of non-coding genetic variants (2,24–27,56–58). Previous studies have shown that combinatorial patterns of epigenetic features such as epigenetic states are maintained across different cell types or even species (4). With the expanding variety of epigenetic features and functional element annotations along with the growth in profiled cell types, there is a growing need for normalization methods that can effectively integrate and preserve recurring combinatorial patterns within the epigenetic states across diverse cell types.

We first compared the panel of methods for their ability to harmonize signal matrices of various epigenetic features across multiple cell types. A better normalization method should more consistently preserve common combinatorial patterns in different cell types and thus allow better data integration across cell types. To quantify the degree of preservation of common combinatorial patterns, we first pooled cCRE signal matrices of seven epigenetic features from each pair of different cell types and identified the combinatorial patterns by clustering cCREs using *K*-means ($K = 30, 20, 40, 50$). We then measured the ASW score, which ranges from 0 to 1 and a lower score indicates better sharing of combinatorial patterns between cell types (40). As shown in Figure 3A and Supplementary Figure 3, JMnorm has significantly lower ASW scores than other methods (paired Wilcoxon test between JMnorm and the second-best performing method, *P*-value = 3.05e-05, 3.05e-04, 2.14e-04, 6.23e-03 for different *K* values in *K*-means clustering), indicating that the multi-feature signal matrices normalized by JMnorm achieved better integration among cell types within each cCRE clusters.

Next, we clustered all cCREs from the 24 cell types groups available in EpiMAP database and used the resulting cCRE clusters (Figure 3B) to compare different normalization methods, using the proportion of robust cCRE clusters (4) as a metric of global preservation of epigenetic feature patterns across different cell types. Here, a robust cCRE cluster is defined as the cCRE cluster displaying high consensus across all cell types. A higher proportion of robust cCRE clusters would indicate a better preservation of combinatorial feature patterns across cell types (see Materials and method section for details). JMnorm-normalized signal matrices resulted in a significantly higher proportion of robust cCRE clusters (paired Wilcoxon test between JMnorm and the second-best performing method, *P*-value = 2.60e-04) than those obtained by other methods (Figure 3C).

Results of these comparisons demonstrate that JMnorm outperforms other methods in integrating multi-feature signal matrices and preserving cross-feature combinatorial patterns across different cell types. It is worth noting that the cross-feature correlation (Figure 2) and cross-feature consistency of combinatory patterns (Figure 3A and B) across varying cell types or conditions are expected to improve after JMnorm normalization. However, JMnorm does not directly optimize these two properties. Instead, they are naturally improved during JMnorm's normalization processes.

Evaluation of cross-cell-type gene expression predictions

Since JMnorm integrates the multi-feature signal matrices and preserves cross-feature combinatorial patterns across different cell types better than other methods, we anticipated that the gene expression prediction models utilizing data normalized by JMnorm would also exhibit improved performance in cross-cell-type predictions. Therefore, we compared normalization methods by their ability to improve cross-cell type gene expression predictions. We trained three types of regression models to learn the quantitative relationships between seven epigenetic features and RNA-seq data: a linear regression model (LM), a gradient boosting regression model (GBM) (41), and a linear regression model with gene-grouping that is based on cross-cell type average expression levels and standard deviations (LM-gene-grouping) (3). The details about training and testing of these regression models are described in the Materials and Methods section. Briefly, we randomly divided protein-coding genes into two groups: 80% of genes were used for model training (Training-Genes) and 20% of genes were used for model evaluation (Testing-Genes). The regression models were trained using the signals of Training-Genes in one Training-Cell-Type and evaluated by *R*-squared (R^2) using the signals of Testing-Genes in another Testing-Cell-Type, representing the most challenging cross-cell type hold-out gene prediction setting in gene expression prediction tasks. As shown in Figure 3D and Supplementary Figure 4, JMnorm-normalized data had significantly higher R^2 than other normalization methods across all three regression methods and different cell-type-pairs (paired Wilcoxon test between JMnorm and the second-best performing method, *P*-values are LM: 6.16e-06, GBM: 1.48e-08, LM-gene-grouping: 2.54e-07). These results demonstrate that JMnorm-normalized epigenetic signals improve performance of cross-cell type gene expression prediction models, suggesting that the JMnorm-normalized epigenetic data are more consistent with gene expression levels, an orthogonal biological data type, than those normalized by other methods.

Evaluation of signal consistency between biological replicates

We next evaluated different methods based on consistency of post-normalization signal strengths between biological replicates, measured by R^2. A better normalization method should result in higher signal consistency between the independently normalized replicates. We examined the replicate consistency of the post-normalized H3K27ac ChIP-seq signals across 9 different cell types. Six other features were used for JMnorm normalization of H3K27ac ChIP-seq data. JMnorm-normalized data had significantly higher R^2 values than data normalized by other methods (paired Wilcoxon test between JMnorm and the second-best performing method, *P*-value = 1.95e-03) (Figure 4A). Moreover, JMnorm-normalized data had the highest or the second highest R^2 values across all 7 examined epigenetic features (Figure 4A and Supplementary Figure 5), especially those that are more likely to reflect cell-type-specific epigenetic events such as ATAC-seq, H3K27ac, and H3K4me1 (1).

As an additional metric, we assessed the replicate consistency of enriched peak calling results for H3K27ac across 9 cell types. To this end, the same peak-calling method was applied to post-normalized signals generated by the panel

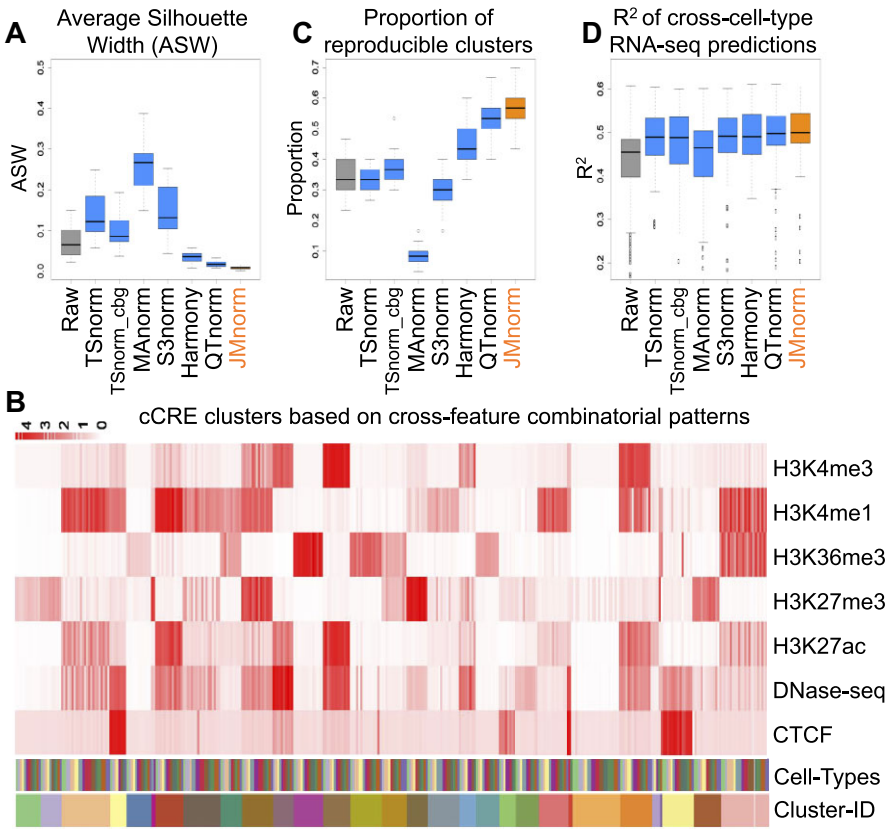


Figure 3. Evaluation of consistency of cross-feature combinatorial patterns and gene expression predictions across cell types. (**A**) A boxplot of Average Silhouette Widths (ASWs) for the seven normalization methods demonstrating quality of mixing of cCREs across different cell types in clustering outputs (*K*-means: $K = 30$). (**B**) Top. A heatmap of the average signals at cCRE clusters for seven epigenetic features in 24 EpiMAP cell type groups. Clusters are ordered by the K-means cluster label ($K = 30$). Bottom. The heatmaps with distinct colors represent labels for different cell types and *K*-means clusters. (**C**) A boxplot of proportions of cCRE clusters that are reproducible in all cell types for the seven normalization methods. (**D**) A boxplot of R^2 between observed and predicted RNA-seq values expressed as $\log_2$ transcripts per million ($\log_2$TPM) for the seven normalization methods. Linear regression model with gene grouping was used for the predictions.

of normalization methods, and an equal number of called peaks was identified for evaluation of replicate consistency using the Jaccard index. Besides the signal strengths consistency, JMnorm's peak calling results also had higher replicate consistency as indicated by significantly higher Jaccard index values compared to peaks generated from data normalized by other methods (paired Wilcoxon test between JMnorm and the second-best performing method, *P*-value = 4.5e-03; Figure 4B).

We then generated scatterplots of post-normalization signals at individual cCREs for the two biological replicates from the H3K27ac ChIP-seq experiment in $CD8^+$ T-cells (Figure 4C and Supplementary Figure 6). Compared to other methods, the scatterplot for JMnorm showed considerably less deviation from the diagonal line (indicating a higher R^2), especially for the cCREs with low or noise-like signals.

To demonstrate that JMnorm does not over-correct the data, we calculated R^2 of post-JMnorm-normalization values between biological replicates of the same cell type and biological replicates of different cell types. As shown in Supplementary Figure 7, across all seven epigenetic features, the R^2 values between replicates of the same cell type (orange boxes) were significantly higher than those between replicates from different cell types (gray boxes). The paired Wilcoxon test affirmed this finding, with p-values ranging from 2.55e-09 to 3.90e-04 for all evaluated epigenetic features. These results demonstrate our method's effectiveness in preserving both cell type-specific differences and consistency between biological replicates within the same cell type.

These results suggest that by leveraging information from functionally correlated features, JMnorm can more effectively improve replicates consistency by reducing technical noise and preserving true epigenetic signals than other normalization methods.

It is important to note that JMnorm performance is not universally superior to all other methods for normalization of all examined epigenetic features. For example, since H3K4me3 is a canonical marker of gene promoters, its ChIP-seq signals have similar global distributions across different cell types. This type of signal distribution is expected to be suitable for QTnorm, which enforces identical post-normalization signal distributions. Hence, QTnorm exhibited a slightly higher R^2 between biological replicates for H3K4me3 ChIP-seq signal than JMnorm (Supplementary Figure 5). Based on the overall performance, we conclude that JMnorm surpasses other normalization methods in enhancing signal and minimizing technical noise and improves consistency between biological replicates.

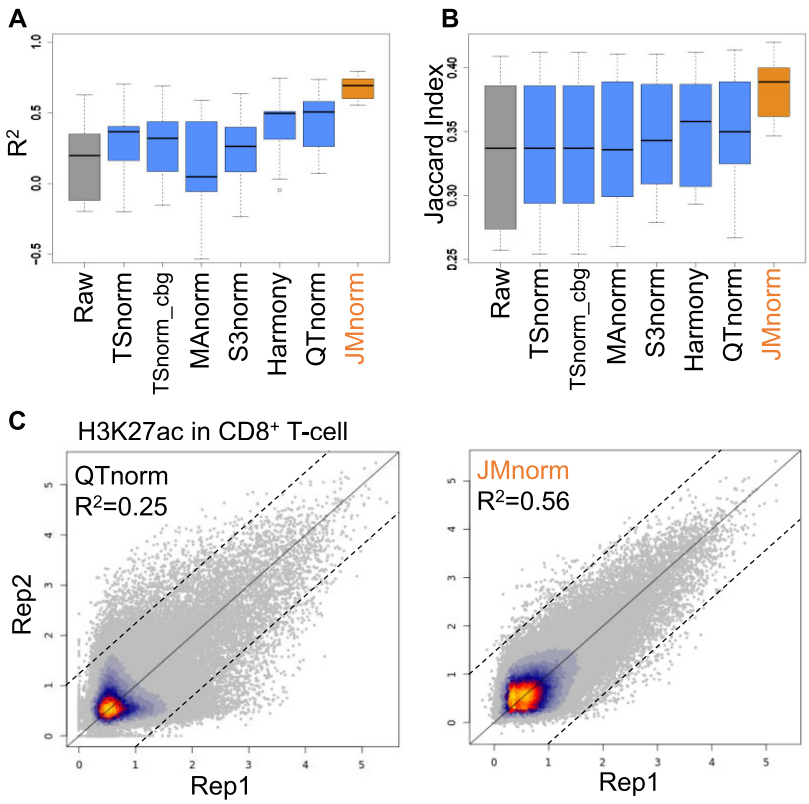


Figure 4. Evaluation of the post-normalization signal consistency between biological replicates. (**A**) A boxplot of R^2 values between biological replicates' signals in multiple cell types for the seven normalization methods. (**B**) A boxplot of Jaccard indexes comparing peak calling results between biological replicates of multiple cell types for the seven normalization methods. (**C**) Scatterplots of H3K27ac ChIP-seq signal in CD8+ T-cells in two biological replicates normalized by QTnorm (left) and JMnorm (right). Data are shown on log2 scale. Bright orange and gray colors indicate higher and lower data point density, respectively.

Evaluation of quality of peak calling results

Based on the performance evaluations results above, JMnorm, QTnorm, and Harmony outperformed the other four methods. This outperformance is expected since JMnorm incorporates the strengths of both QTnorm and Harmony: as QTnorm, it minimizes the complex technical biases by equalizing the signals of the same rank, and as Harmony, it combines the highly correlated variables in the PCA space. To more closely evaluate the seven normalization methods, we compared their performance by the quality of peak calling results (Figure 5 and Supplementary Figures 8 and 9). We hypothesized that better normalization could improve the accuracy of peak calling results by reducing both false positives and negatives, thereby yielding peaks more closely associated with true biologically relevant events.

For this test, we first selected TF ChIP-seq data for CTCF (67) and YY1 (44) because the role of these TFs in transcription regulation at gene promoters, enhancers, and topologically associating domain (TAD) boundaries is well understood and their DNA-binding motifs had been extensively characterized. Specifically, CTCF and YY1 often co-occupy the same genomic regions at the promoters and enhancers of active genes (44), whereas TAD boundaries are uniquely bound by CTCF (45–47). Moreover, approximately 80% of previously validated CTCF ChIP-seq peaks contain CTCF binding motifs (48,49). To evaluate the quality of CTCF ChIP-seq peak calling results, we applied the same peak-calling method to the post-normalized signals generated by the seven normalization methods, yielding an equal number of CTCF peaks for each method (see Materials and Methods section for more details). The resulting CTCF peaks were evaluated using the following three metrics: (i) enrichment at YY1 peak regions, (ii) enrichment at TAD boundaries and (iii) fraction of the CTCF peaks with CTCF DNA-binding motif. We found that CTCF peaks, uniquely called using JMnorm-normalized data, exhibited significantly higher enrichment at YY1 peak regions than the ones generated by other methods (paired Wilcoxon test P-values < 0.05) (Figure 5). Similarly, JMnorm-normalized CTCF ChIP-seq peaks had significantly higher (paired Wilcoxon test P-values < 0.05) or comparable enrichment at TAD boundaries than other methods' peaks (Supplementary Figure 8). Lastly, there is no significant difference in proportion of CTCF peaks containing CTCF motifs (Jaspar (68) ID: MA0139.1) with the JMnorm-normalized data relative to other methods' data (Supplementary Figure 9). In sum, these findings suggest that JMnorm exhibits improved performance in terms of quality and biological significance of the resulting TF ChIP-seq peaks.

We next focused on comparing the JMnorm and QTnorm by the quality of the DNase-seq peak calling results across cell types (Figure 6 and Supplementary Figure 10). Both methods are designed to minimize technical biases by equalizing the signals of the same rank. However, a critical technical bias of QTnorm is that it imposes an identical signal distribution across cell types, which might lead peak-calling algorithms to identify the same number of peaks for all cell types. We asked

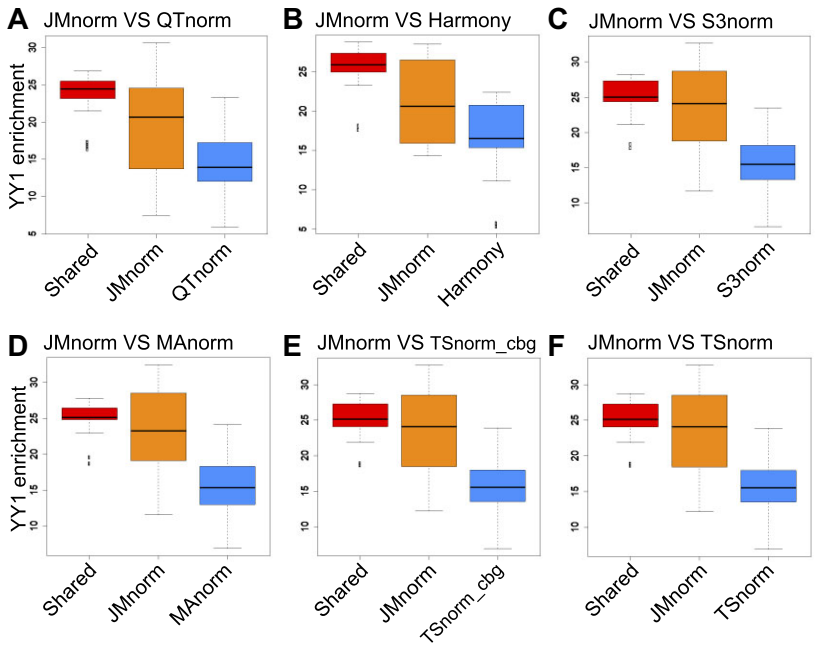


Figure 5. Evaluation of quality of TF peak calling results. Boxplots comparing JMnorm's and six other methods' performance as determined by CTCF peak enrichment at YY1 peak regions. (**A**) Comparison between JMnorm and QTnorm, (**B**) JMnorm and Harmony, (**C**) JMnorm and S3norm, (**D**) JMnorm and MAnorm, (**E**) JMnorm and TSnorm_cbg, and (**F**) JMnorm and TSnorm. Red box plots represent enrichments for CTCF peaks that are shared between JMnorm and a corresponding alternative method. Orange box plots represent enrichments for CTCF peaks uniquely identified by JMnorm. Blue box plots represent enrichments for CTCF peaks uniquely identified by the respective alternative method.

if JMnorm could circumvent this technical bias since it applies QTnorm only within each cCRE cluster, which may contain a different number of cCREs per cell type, thereby leading to different global signal distributions across cell types and varying numbers of peaks for the normalized data.

To this end, we first compared the performance of QTnorm and JMnorm on the same set of DNase-seq data generated in 24 cell types and found that QTnorm calls the same number of peaks across cell types, whereas JMnorm effectively identifies a different number of peaks per cell type (Supplementary Figure 10A). As expected, most cell types close to stem cells exhibited a larger number of peaks (denoted by orange coloring) than other cell types, consistent with prior observations by Meuleman *et al.* (37) (Supplementary Figure 10B).

We then evaluated the noise level for the two methods. Since QTnorm imposes an identical signal distribution across various cell types, we expected more false positive peaks in the cell types with fewer true biological peaks. Clustering of cCREs based on these signals would reveal noisier cross-cell type patterns with increased presence of weaker signals in many cell types. To test the validity of these expectations, we utilized the Snapshot clustering algorithm to group cCREs using normalized DNase-seq signals (8) and then evaluated the quality of the resulting clusters. As expected, QTnorm clusters exhibited relatively noisier cross-cell type patterns (Supplementary Figure 10C and D) with significantly lower signal-to-noise ratios than JMnorm's clusters (paired Wilcoxon test *P*-value = 2.38e-07) (Supplementary Figure 10E) (see Materials and methods section for more details), indicating that JMnorm-normalized DNase-seq signals had less noise.

Lastly, to determine whether the JMnorm-normalized DNase-seq peaks contained true biologically meaningful information, we conducted pairwise comparisons of post-normalized DNase-seq peaks in different cell types and assessed the enrichment of human phenotype terms (69,70) relevant to the respective cell types in the differential peaks. We reasoned that cell-type-specific differential peaks would be more likely to reflect true biological differences between cell types and corresponding top-enriched human phenotype terms could be used to highlight the cell-type-specific functional differences. An initial analysis of the differential peaks identified in heart and lung cell types revealed greater enrichment of heart-relevant terms in heart-specific peaks uniquely identified using JMnorm-normalized DNase-seq data than QTnorm-normalized data (Figure 6A). Next, we conducted the same pairwise DNase-seq peak analysis for 100 randomly selected pairs of cell types. Differential peaks unique to JMnorm had a significantly higher enrichment (paired Wilcoxon test *P*-value < 0.05) in cell-type-specific functional terms than peaks uniquely identified using other methods (Figure 6B-G), indicating that JMnorm-derived differential peaks may more accurately capture true biological information.

These results demonstrate that JMnorm improves the accuracy and biological relevance of the peak calling results.

Evaluation of quality of differential peak calling in response to perturbations

Small molecule perturbations often make a substantial and global impact on the epigenome (71,72) presenting a substantial challenge for data normalization. That is because many normalization methods assume that differences in the overall signal (12–14) or common peak regions (16,17) between different datasets are caused primarily by technical inconsistencies rather than true biological changes in global epigenetic signals. We compared the performance of seven normalization methods by the quality of differential glucocorticoid

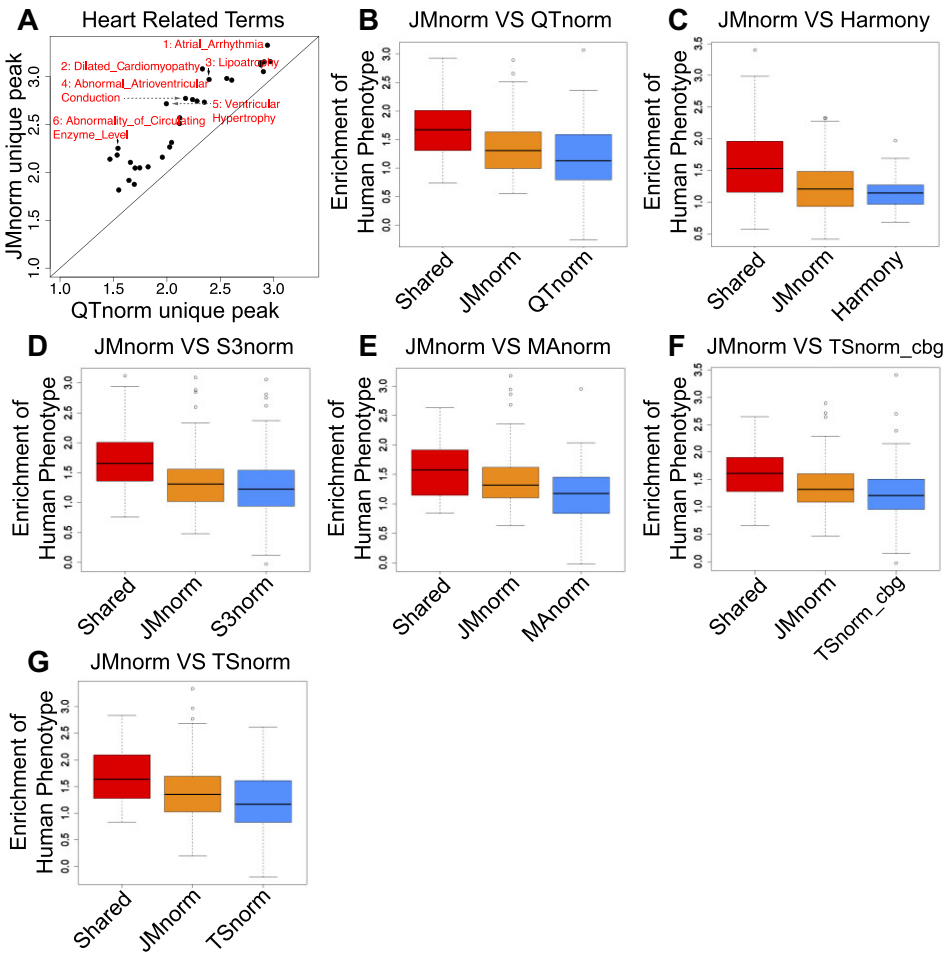


Figure 6. Comparative analysis of JMnorm and other methods using DNase-seq peak calling results. (**A**) A scatterplot of enrichments of the Human Phenotype terms in heart-specific peaks (relative to lung) uniquely identified by QTnorm (x-axis) or JMnorm (y-axis). The displayed terms are the top 30 most significantly enriched terms identified by both methods. (**B**) A box plot of Human Phenotype terms enriched in shared and uniquely identified QTnorm's or JMnorm's peaks across 100 randomly selected cell type pairs. (C–G) same as (B), for comparison between (**C**) JMnorm and Harmony, (**D**) JMnorm and S3norm, (**E**) JMnorm and MAnorm, (**F**) JMnorm and TSnorm_cbg, and (**G**) JMnorm and TSnorm.

receptor (GR, gene symbol NR3C1) ChIP-seq peak calling results in the context of dexamethasone (Dex) treatment of A549 cells (15). We selected GR ChIP-seq data for this evaluation because GR is a well-characterized receptor for Dex, GR's binding to DNA increases globally with transient Dex treatment, and its DNA-binding motif is also well known (73).

For this analysis, we first normalized the GR ChIP-seq data for both the Dex treatment condition and the no-treatment control using each of the seven methods and then identified differential GR peaks using the post-normalization signals. For JMnorm and Harmony, the GR ChIP-seq data were normalized in conjunction with four additional epigenetic features (ATAC, H3K27ac, H3K4me1 and H3K4me3). We benchmarked the performance of different normalization methods relative to TSnorm_cbg. The TSnorm_cbg leverages the same concept as the normalization of ChIP-seq with control (NCIS) (21) method that was specifically developed to address the challenge of normalizing global differences between ChIP and control datasets, by calculating a scale factor based on the information from common background regions between the two. We hypothesized that better normalization could improve the accuracy of differential peak calling results, yielding peaks that are more closely associated with true biologically relevant changes induced by the perturbagen.

As expected, most of the differential GR peaks (top 2000 differential peaks based on absolute value of $\log_2$ fold-change between Dex treatment sample and no treatment sample) called using TSnorm_cbg-normalized data were upregulated after Dex treatment (Figure 7A), indicating an increase of GR binding to the genome. For JMnorm, most differential peaks also exhibited increased GR signals (Figure 7B), similarly supporting our biological understanding. Conversely, only approximately 50% of differential peaks called using QTnorm-normalized data were upregulated with Dex treatment (Figure 7C).

To further characterize the differential GR peaks identified using data normalized by the six methods relative to TSnorm_cbg-normalized data, we evaluated the quality of the GR motif (Jaspar ID: MA0113.1) (Figure 7D) and the extent of Human Phenotype term enrichment (Figure 7E and Supplementary Figure 11) in differential peaks. As expected, GR motif scores and the degree of GR-related process term enrichment were significantly lower in differential peaks called using QTnorm-, S3norm-, and MAnorm-normalized data as compared to those derived from the JM-

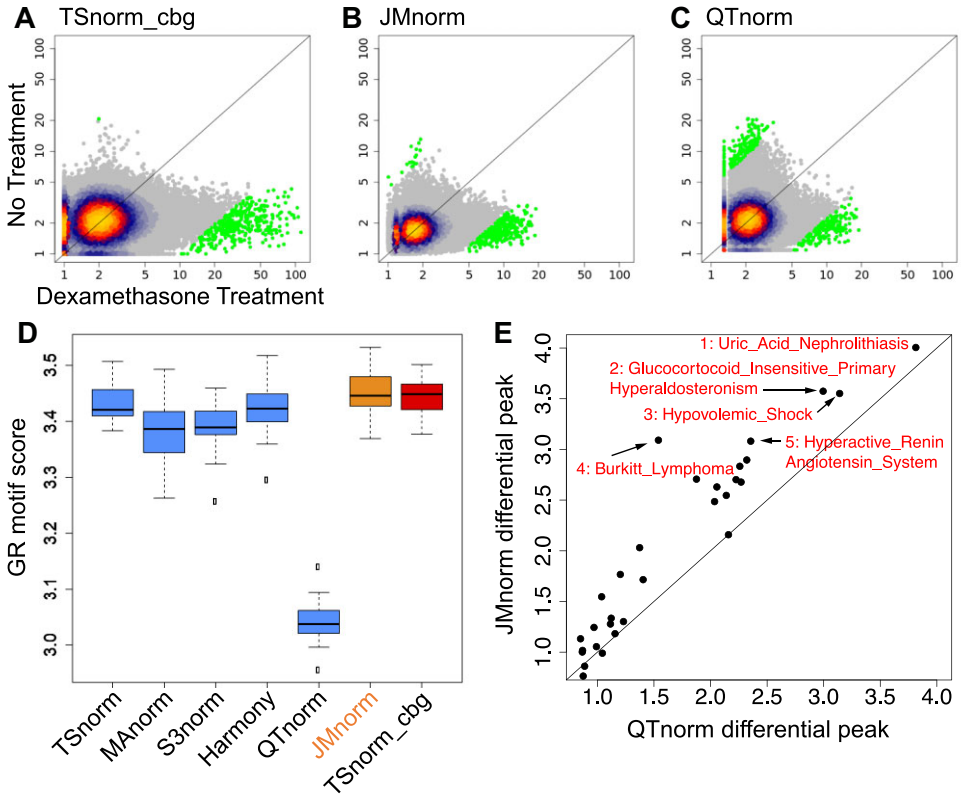


Figure 7. Evaluation of quality of differential peak calling in response to perturbations. (A–C) Scatter plots of GR ChIP-seq signals at ENCODE cCREs with (x-axis) and without (y-axis) Dex treatment for (**A**) TSnorm_cbg, (**B**) JMnorm, and (**C**) QTnorm. Bright orange and gray colors indicate higher and lower data point density, respectively. Green color represents GR ChIP-seq signals for the top 2000 differential peaks based on absolute value of $\log_2$ fold change of signals between Dex treatment and no treatment control. (**D**) A boxplot of FIMO scores for GR motif (Jaspar ID: MA0113.1) found in differential peaks for the seven normalization methods. (**E**) A scatterplot of enrichments of the Human Phenotype terms in differential GR ChIP-seq peaks identified by QTnorm (x-axis) or JMnorm (y-axis).

norm (paired Wilcoxon test P-value < 0.05). These results suggest that an improper matching of overall distribution or signal-to-noise ratio can undermine the association between differential GR peaks and Dex treatment. In conclusion, JMnorm improves the accuracy and biological significance of the differential peak calling results relative to QTnorm, S3norm, and MAnorm, and demonstrates a comparable performance relative to the other approaches including the state-of-the-art method (TSnorm_cbg) for normalization of epigenetic datasets with global differences induced by perturbations.

Evaluation of JMnorm performance at the 100 bp genomic bin resolution

To evaluate the robustness of JMnorm's performance at a higher genomic resolution, normalization methods were compared across multiple benchmarks at the 100 bp genomic bin resolution. As shown in Supplementary Figure 12A, the cross-feature correlation matrix computed using JMnorm results had significantly lower MSEs (paired Wilcoxon test between JMnorm and the second-best performing method, P-value $<$ 2.2e-16), indicating higher consistency between the cross-feature correlation matrices of the target and reference samples. Similarly, the cross-feature combinatory patterns were more similar across different cell types resulting in lower ASW score (paired Wilcoxon test between JMnorm and the second-best performing method, P-value = 4.2e-03) and higher proportion of robust cCRE clusters (paired Wilcoxon test between JMnorm and the second-best performing method, P-value = 5.7e-02) (Supplementary Figure 12B and C). Since better preservation of cross-feature combinatorial patterns across different cell types should lead to a better performance in the cross-cell type RNA-seq prediction models, we further compared different normalization methods based on the accuracy of cross-cell type RNA-seq predictions using post-normalization epigenetic signals as an input. As shown in Supplementary Figure 12D, JMnorm indeed had significantly higher R^2 between the predicted and observed RNA-seq signals than other methods (paired Wilcoxon test between JMnorm and the second-best performing method, P-value = 4.8e-03). Next, we evaluated the signal and peak calling consistency between the biological replicates. Results shown in Supplementary Figure 12E and F demonstrate significantly higher consistency for JMnorm-normalized data based on both signal R^2 (paired Wilcoxon test between JMnorm and the second-best performing method, P-value = 7.28e-09) and Jaccard Index of peak calling results (paired Wilcoxon test between JMnorm and the second-best performing method, P-value = 2.67e-09).

In addition, we evaluated the JMnorm-normalized signals by using the quality of its peak calling and differential peak calling results as a metric, measured by comparing the consistency between the peak calling and differential peak calling results and orthogonal information. As demonstrated in Supplementary Figure 13A-E, when comparing CTCF peaks

uniquely identified using JMnorm-normalized data with peaks called using data normalized by other methods, CTCF peaks derived from JMnorm consistently exhibited significantly higher enrichments at the YY1 peak regions (paired Wilcoxon test P-value < 0.05) than those derived from other methods. Similarly, differential DNase peaks uniquely identified from the JMnorm-normalized data showed higher enrichments in cell type-specific functional terms compared to peaks uniquely identified by other methods (Supplementary Figure 13F–J).

Finally, we also evaluated JMnorm by the quality of differential peak calling results in response to perturbations. The scores for GR motif were higher for differential peaks obtained using JMnorm-normalized data in comparison with QTnorm-, S3norm-, or MAnorm-normalized data (Supplementary Figure 13K). Similarly, enrichment levels of the GR-related process terms for JMnorm-specific GR differential peaks were significantly higher than those for QTnorm-specific GR differential peaks (paired Wilcoxon test P-value = 9.71e-06) (Supplementary Figure 13L).

In conclusion, based on results of evaluations across multiple benchmarks, we found that JMnorm's superior performance was robust at higher resolution (100 bp genomic bins) used for generating the epigenomic signal matrices.

Discussion

We present a novel approach named JMnorm that simultaneously normalizes multiple epigenetic features across cell types, species, and experimental conditions by leveraging information from functionally correlated features. JMnorm presents several methodological advances over existing normalization approaches that analyze each feature independently and thus may distort relationships between epigenetic features. Specifically, JMnorm normalizes multiple epigenetic features jointly in the PCA space that preserves correlations between different features. Secondly, using a two-step process of initial cCRE clustering followed by within-cluster quantile normalization, JMnorm effectively reduces technical biases without imposing identical global signal distributions across different cell types. Following the principle of Occam's razor, we have implemented JMnorm using simple yet sufficiently effective statistical techniques. By employing these techniques, we aimed to enhance the robustness of JMnorm and enable its wide application for various downstream analyses. Future investigations could potentially benefit from exploring advanced techniques, such as orthogonal transformation and cCRE clustering. We have demonstrated JMnorm's improved capabilities by comparing cross-feature correlation matrices before and post-normalization, analyzing the extent of preservation of combinatorial patterns of epigenetic features across different cell types, and evaluating consistency between biological replicates. Furthermore, in several use cases of epigenomic analyses, such as prediction of gene expression, peak calling, and differential TF binding, we showed that JMnorm achieves better results than other methods when being validated against various types of orthogonal biological data. Altogether, these improvements underscore the strength of JMnorm in reducing noise and preserving true biologically meaningful information in epigenomic data.

Given the superior performance of JMnorm across diverse types of epigenetic features and genomic tasks, we anticipate that it can be used across a broad range of applications. In addition to the applications described in this work, a potential application for JMnorm is in normalizing single-cell gene module signal matrices. Single-cell gene expression analyses (74,75) often involve data transformation that converts relatively noisy individual gene expression information into gene module scores (76–80). By leveraging information from correlated gene modules, JMnorm may provide a more accurate representation of cells or meta cells in different gene module spaces across various cell groups, conditions, or time points.

Because JMnorm normalizes data by leveraging information from functionally correlated datasets, higher correlations among the features can result in better improvements with normalization. To help users more effectively select features with high correlations for JMnorm analyses, we computed pairwise cross-feature correlations for 538 epigenetic features in K562 cells using data from the ENCODE Consortium (1,2) (Supplementary Figure 14A). Considering the substantial size of the output correlation matrix and the difficulties with visualization, we also set up a Shiny app (81) visualization tool for convenient and interactive exploration of the correlation matrix (Supplementary Figure 14B and Supplementary Table 3).

Lastly, we would like to point out a potential limitation: JMnorm might not be able to adequately handle global changes across all functionally correlated features under certain conditions. This could result in losing some global changes at the within-cluster quantile normalization step. For these scenarios, other normalization strategies that take into account global changes, such as normalization using only common background regions (21) or a spike-in reference (71,72), may be more appropriate.

In summary, JMnorm introduces a novel approach for multi-feature normalization of epigenetic data. This method has a straightforward design and is effective in reducing technical biases and preserving cross-feature correlations across different cell types. With the continuing development of high-throughput sequencing technologies for genome interrogation and the growing number of epigenetic datasets generated in different cell types, species, and under various physiological conditions, we anticipate JMnorm becoming a crucial tool for data normalization in integrative and comparative epigenomics studies.

Data availability

The JMnorm package is available at GitHub (https://github.com/camp4tx/JMnorm) and Zenodo (https://doi.org/10.5281/zenodo.10119105) and released under GNU General Public License, version 2.0 or later. The main part of JMnorm was implemented in R. We also provided a conda environment that can be deployed in both MacOS and Linux operating systems. The signal tracks used in this project were mainly downloaded from VISION project data portal (*https://usevision.org/data/*) and the EpiMAP repository (*https://epigenome.wustl.edu/epimap/data/*). The detailed cell types included in each cell type group in the EpiMAP data can be found in the EpiMAP metadata file (*https://personal.broadinstitute.org/cboix/epimap/metadata/Short_Metadata.html*). Human and mouse cCRE lists were downloaded from ENCODE-SCREEN data portal (*https://screen.encodeproject.org/*) (82). The list of links for the files used in this paper can be found in Supplementary Table 1.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

We would like to express our gratitude to the ENCODE, VISION and EpiMAP data Consortium projects for their valuable resources. We thank Gokul Ramaswami, Wei Jiang, and Qiuhai Zeng for their valuable suggestions and help.

Funding

No external funding.

Conflict of interest statement

All authors are employees and equity holders of CAMP4 Therapeutics Corporation.

References

1. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
2. Moore,J.E., Purcaro,M.J., Pratt,H.E., Epstein,C.B., Shoresh,N., Adrian,J., Kawli,T., Davis,C.A., Dobin,A., Kaul,R., *et al.* (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, **583**, 699–710.
3. Xiang,G., Keller,C.A., Heuston,E., Giardine,B.M., An,L., Wixom,A.Q., Miller,A., Cockburn,A., Sauria,M.E.G., Weaver,K., *et al.* (2020) An integrative view of the regulatory and transcriptional landscapes in mouse hematopoiesis. *Genome Res.*, **30**, 472–484.
4. Xiang,G., He,X., Giardine,B.M., Weaver,K.J., Taylor,D.J., Mccoy,R.C., Jansen,C., Keller,C.A., Wixom,A.Q., Cockburn,A., *et al.* (2023) Interspecies regulatory landscapes and elements revealed by novel joint systematic integration of human and mouse blood cell epigenomes. bioRxiv doi: https://doi.org/10.1101/2023.04.02.535219, 21 November 2023, preprint: not peer reviewed.
5. Vu,H. and Ernst,J. (2022) Universal annotation of the human genome through integration of over a thousand epigenomic datasets. *Genome Biol.*, **23**, 9.
6. Luan,J., Xiang,G., Gómez-García,P.A., Tome,J.M., Zhang,Z., Vermunt,M.W., Zhang,H., Huang,A., Keller,C.A., Giardine,B.M., *et al.* (2021) Distinct properties and functions of CTCF revealed by a rapidly inducible degron system. *Cell Rep.*, **34**, 108783.
7. Koch,H., Keller,C.A., Xiang,G., Giardine,B., Zhang,F., Wang,Y., Hardison,R.C. and Li,Q. (2022) CLIMB: high-dimensional association detection in large scale genomic data. *Nat. Commun.*, **13**, 6874.
8. Xiang,G., Giardine,B., An,L., Sun,C., Keller,C.A., Heuston,E.F., Anderson,S.M., Kirby,M., Bodine,D., Zhang,Y., *et al.* (2023) Snapshot: a package for clustering and visualizing epigenetic history during cell differentiation. *BMC Bioinf.*, **24**, 102.
9. Bernstein,B.E., Stamatoyannopoulos,J.A., Costello,J.F., Ren,B., Milosavljevic,A., Meissner,A., Kellis,M., Marra,M.A., Beaudet,A.L., Ecker,J.R., *et al.* (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.*, **28**, 1045–1048.
10. McDowell,I.C., Manandhar,D., Vockley,C.M., Schmid,A.K., Reddy,T.E. and Engelhardt,B.E. (2018) Clustering gene expression time series data using an infinite Gaussian process mixture model. *PLoS Comput. Biol.*, **14**, e1005896.
11. Meyer,C.A. and Liu,X.S. (2014) Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat. Rev. Genet.*, **15**, 709–721.
12. Marioni,J.C., Mason,C.E., Mane,S.M., Stephens,M. and Gilad,Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.
13. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nussbaum,C., Myers,R.M., Brown,M., Li,W., *et al.* (2008) Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
14. Bolstad,B.M., Irizarry,R.A., Astrand,M. and Speed,T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.
15. Reddy,T.E., Pauli,F., Sprouse,R.O., Neff,N.F., Newberry,K.M., Garabedian,M.J. and Myers,R.M. (2009) Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation. *Genome Res.*, **19**, 2163–2171.
16. Shao,Z., Zhang,Y., Yuan,G.-C., Orkin,S.H. and Waxman,D.J. (2012) MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. *Genome Biol.*, **13**, R16.
17. Xiang,G., Keller,C.A., Giardine,B., An,L., Li,Q., Zhang,Y. and Hardison,R.C. (2020) S3norm: simultaneous normalization of sequencing depth and signal-to-noise ratio in epigenomic data. *Nucleic Acids Res.*, **48**, e43.
18. Xiang,G., Giardine,B.M., Mahony,S., Zhang,Y. and Hardison,R.C. (2021) S3V2-IDEAS: a package for normalizing, denoising and integrating epigenomic datasets across different cell types. *Bioinformatics*, **37**, 3011–3013.
19. Diaz,A., Nellore,A. and Song,J.S. (2012) CHANCE: comprehensive software for quality control and validation of ChIP-seq data. *Genome Biol.*, **13**, R98.
20. Park,P.J. (2009) ChIP–seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **10**, 669–680.
21. Liang,K. and Keleş,S. (2012) Normalization of ChIP-seq data with control. *BMC Bioinf.*, **13**, 199.
22. Hardison,R.C., Zhang,Y., Keller,C.A., Xiang,G., Heuston,E.F., An,L., Lichtenberg,J., Giardine,B.M., Bodine,D., Mahony,S., *et al.* (2020) Systematic integration of GATA transcription factors and epigenomes via IDEAS paints the regulatory landscape of hematopoietic cells. *IUBMB Life*, **72**, 27–38.
23. Lyu,Y. and Li,Q. (2016) A semi-parametric statistical model for integrating gene expression profiles across different platforms. *BMC Bioinf.*, **17**, S5.
24. Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.
25. Hoffman,M.M., Buske,O.J., Wang,J., Weng,Z., Bilmes,J.A. and Noble,W.S. (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods*, **9**, 473–476.
26. Zhang,Y., An,L., Yue,F. and Hardison,R.C. (2016) Jointly characterizing epigenetic dynamics across multiple human cell types. *Nucleic Acids Res.*, **44**, 6721–6731.
27. Libbrecht,M.W., Chan,R.C.W. and Hoffman,M.M. (2021) Segmentation and genome annotation algorithms for identifying chromatin state and other genomic patterns. *PLoS Comput. Biol.*, **17**, e1009423.
28. Zhang,Y. and Hardison,R.C. (2017) Accurate and reproducible functional maps in 127 human cell types via 2D genome segmentation. *Nucleic Acids Res.*, **45**, 9823–9836.
29. Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J., Ziller,M.J., *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
30. Oudelaar,A.M., Hanssen,L.L.P., Hardison,R.C., Kassouf,M.T., Hughes,J.R. and Higgs,D.R. (2017) Between form and function: the complexity of genome folding. *Hum. Mol. Genet.*, **26**, R208–R215.
31. Philipsen,S. and Hardison,R.C. (2018) Evolution of hemoglobin loci and their regulatory elements. *Blood Cells Mol. Dis.*, **70**, 2–12.

32. Heuston,E.F., Keller,C.A., Lichtenberg,J., Giardine,B., Anderson,S.M., Hardison,R.C. and Bodine,D.M. (2018) Establishment of regulatory elements during erythro-megakaryopoiesis identifies hematopoietic lineage-commitment points. *Epigenetics Chromatin*, **11**, 22.
33. Hoon,D.S.B., Rahimzadeh,N. and Bustos,M.A. (2021) EpiMap: fine-tuning integrative epigenomics maps to understand complex human regulatory genomic circuitry. *Signal Transduct Target Ther*, **6**, 179.
34. Zhao,H., Sun,Z., Wang,J., Huang,H., Kocher,J.P. and Wang,L. (2014) CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*, **30**, 1006–1007.
35. Zheng,R., Wan,C., Mei,S., Qin,Q., Wu,Q., Sun,H., Chen,C.H., Brown,M., Zhang,X., Meyer,C.A., *et al.* (2019) Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Res.*, **47**, D729–D735.
36. Mei,S., Qin,Q., Wu,Q., Sun,H., Zheng,R., Zang,C., Zhu,M., Wu,J., Shi,X., Taing,L., *et al.* (2017) Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res.*, **45**, D658–D662.
37. Meuleman,W., Muratov,A., Rynes,E., Halow,J., Lee,K., Bates,D., Diegel,M., Dunn,D., Neri,F., Teodosiadis,A., *et al.* (2020) Index and biological spectrum of human DNase I hypersensitive sites. *Nature*, **584**, 244–251.
38. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,a.D. (2002) The Human Genome Browser at UCSC. *Genome Res.*, **12**, 996–1006.
39. Langfelder,P., Zhang,B. and Horvath,S. (2008) Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*, **24**, 719–720.
40. Tran,H.T.N., Ang,K.S., Chevrier,M., Zhang,X., Lee,N.Y.S., Goh,M. and Chen,J. (2020) A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.*, **21**, 12.
41. Friedman,J.H. (2002) Stochastic gradient boosting. *Comput. Stat. Data Anal.*, **38**, 367–378.
42. Koohy,H., Down,T.A., Spivakov,M. and Hubbard,T. (2014) A comparison of peak callers used for DNase-Seq data. *PLoS One*, **9**, e96303.
43. John,S., Sabo,P.J., Thurman,R.E., Sung,M.-H., Biddie,S.C., Johnson,T.A., Hager,G.L. and Stamatoyannopoulos,J.A. (2011) Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.*, **43**, 264–268.
44. Weintraub,A.S., Li,C.H., Zamudio,A.V., Sigova,A.A., Hannett,N.M., Day,D.S., Abraham,B.J., Cohen,M.A., Nabet,B., Buckley,D.L., *et al.* (2017) YY1 is a structural regulator of enhancer-promoter loops. *Cell*, **171**, 1573–1588.
45. An,L., Yang,T., Yang,J., Nuebler,J., Xiang,G., Hardison,R.C., Li,Q. and Zhang,Y. (2019) OnTAD: hierarchical domain structure reveals the divergence of activity among TADs and boundaries. *Genome Biol.*, **20**, 282.
46. Dixon,J.R., Gorkin,D.U. and Ren,B. (2016) Chromatin domains: the unit of chromosome organization. *Mol. Cell*, **62**, 668–680.
47. Dixon,J.R., Selvaraj,S., Yue,F., Kim,A., Li,Y., Shen,Y., Hu,M., Liu,J.S. and Ren,B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.
48. Ghirlando,R. and Felsenfeld,G. (2016) CTCF: making the right connections. *Genes Dev.*, **30**, 881–891.
49. Nakahashi,H., Kwon,K.-R.K., Resch,W., Vian,L., Dose,M., Stavreva,D., Hakim,O., Pruett,N., Nelson,S., Yamane,A., *et al.* (2013) A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Rep.*, **3**, 1678–1689.
50. Grant,C.E., Bailey,T.L. and Noble,W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.
51. Quinlan,A.R. (2014) BEDTools: the Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinformatics*, **47**, 11.12.1–11.12.34.
52. Gu,Z. and Hübschmann,D. (2023) rGREAT: an R/bioconductor package for functional enrichment on genomic regions. *Bioinformatics*, **39**, btac745.
53. McLean,C.Y., Bristor,D., Hiller,M., Clarke,S.L., Schaar,B.T., Lowe,C.B., Wenger,A.M. and Bejerano,G. (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, **28**, 495–501.
54. Korsunsky,I., Millard,N., Fan,J., Slowikowski,K., Zhang,F., Wei,K., Baglaenko,Y., Brenner,M., Loh,P. and Raychaudhuri,S. (2019) Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods*, **16**, 1289–1296.
55. Jolliffe,I.T. and Basilevsky,A. (1997) Statistical factor analysis and related methods: Theory and applications. *Biometrics*, **53**, 97–182.
56. Dong,X., Greven,M.C., Kundaje,A., Djebali,S., Brown,J.B., Cheng,C., Gingeras,T.R., Gerstein,M., Guigó,R., Birney,E., *et al.* (2012) Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol.*, **13**, R53.
57. Karlic,R., Chung,H.-R., Lasserre,J., Vlahovicek,K. and Vingron,M. (2010) Histone modification levels are predictive for gene expression. *Proc. Natl. Acad. Sci.*, **107**, 2926–2931.
58. Singh,R., Lanchantin,J., Robins,G. and Qi,Y. (2016) DeepChrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics*, **32**, i639–i648.
59. Moore,J.E., Pratt,H.E., Purcaro,M.J. and Weng,Z. (2020) A curated benchmark of enhancer-gene interactions for evaluating enhancer-target gene prediction methods. *Genome Biol.*, **21**, 17.
60. Whalen,S., Truty,R.M. and Pollard,K.S. (2016) Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.*, **48**, 488–496.
61. Fulco,C.P., Nasser,J., Jones,T.R., Munson,G., Bergman,D.T., Subramanian,V., Grossman,S.R., Anyoha,R., Doughty,B.R., Patwardhan,T.A., *et al.* (2019) Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.*, **51**, 1664–1669.
62. Schreiber,J., Durham,T., Bilmes,J. and Noble,W.S. (2020) Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome. *Genome Biol.*, **21**, 81.
63. Ernst,J. and Kellis,M. (2015) Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat. Biotechnol.*, **33**, 364–376.
64. Durham,T.J., Libbrecht,M.W., Howbert,J.J., Bilmes,J. and Noble,W.S. (2018) PREDICTD PaRallel epigenomics data imputation with cloud-based tensor decomposition. *Nat. Commun.*, **9**, 1402.
65. Li,H. and Guan,Y. (2022) Asymmetric predictive relationships across histone modifications. *Nat Mach Intell*, **4**, 288–299.
66. Schreiber,J., Boix,C., wook Lee,J., Li,H., Guan,Y., Chang,C.-C., Chang,J.-C., Hawkins-Hooker,A., Schölkopf,B., Schweikert,G., *et al.* (2023) The ENCODE Imputation Challenge: a critical assessment of methods for cross-cell type imputation of epigenomic profiles. *Genome Biol.*, **24**, 79.
67. Ong,C.T. and Corces,V.G. (2014) CTCF: an architectural protein bridging genome topology and function. *Nat. Rev. Genet.*, **15**, 234–246.
68. Mathelier,A., Zhao,X., Zhang,A.W., Parcy,F., Worsley-Hunt,R., Arenillas,D.J., Buchman,S., Chen,C., Chou,A., Ienasescu,H., *et al.* (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **42**, D142–D147.
69. Robinson,P.N. and Mundlos,S. (2010) The human phenotype ontology. *Clin. Genet.*, **77**, 525–534.
70. Köhler,S., Gargano,M., Matentzoglu,N., Carmody,L.C., Lewis-Smith,D., Vasilevsky,N.A., Danis,D., Balagura,G., Baynam,G., Brower,A.M., *et al.* (2021) The human phenotype ontology in 2021. *Nucleic Acids Res.*, **49**, D1207–D1217.
71. Orlando,D.A., Chen,M.W., Brown,V.E., Solanki,S., Choi,Y.J., Olson,E.R., Fritz,C.C., Bradner,J.E. and Guenther,M.G. (2014) Quantitative ChIP-Seq normalization reveals global modulation of the epigenome. *Cell Rep.*, **9**, 1163–1170.

72. Risso,D., Ngai,J., Speed,T.P. and Dudoit,S. (2014) Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.*, **32**, 896–902.
73. McDowell,I.C., Barrera,A., D'Ippolito,A.M., Vockley,C.M., Hong,L.K., Leichter,S.M., Bartelt,L.C., Majoros,W.H., Song,L., Safi,A., *et al.* (2018) Glucocorticoid receptor recruits to enhancers and drives activation by motif-directed binding. *Genome Res.*, **28**, 1272–1284.
74. Satija,R., Farrell,J.A., Gennert,D., Schier,A.F. and Regev,A. (2015) Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.*, **33**, 495–502.
75. Heumos,L., Schaar,A.C., Lance,C., Litinetskaya,A., Drost,F., Zappia,L., Lücken,M.D., Strobl,D.C., Henao,J., Curion,F., *et al.* (2023) Best practices for single-cell analysis across modalities. *Nat. Rev. Genet.*, **24**, 550–572.
76. Lange,M., Bergen,V., Klein,M., Setty,M., Reuter,B., Bakhti,M., Lickert,H., Ansari,M., Schniering,J., Schiller,H.B., *et al.* (2022) CellRank for directed single-cell fate mapping. *Nat. Methods*, **19**, 159–170.
77. La Manno,G., Soldatov,R., Zeisel,A., Braun,E., Hochgerner,H., Petukhov,V., Lidschreiber,K., Kastriti,M.E., Lönnerberg,P., Furlan,A., *et al.* (2018) RNA velocity of single cells. *Nature*, **560**, 494–498.
78. Zhang,Y., Xiang,G., Jiang,A.Y., Lynch,A., Zeng,Z., Wang,C., Zhang,W., Fan,J., Kang,J., Gu,S.S., *et al.*MetaTiME integrates single-cell gene expression to characterize the meta-components of the tumor immune microenvironment. *Nat. Commun.*, **14**, 2634.
79. Wang,W., Tan,H., Sun,M., Han,Y., Chen,W., Qiu,S., Zheng,K., Wei,G. and Ni,T. (2021) Independent component analysis based gene co-expression network inference (ICAnet) to decipher functional modules for better single-cell clustering and batch integration. *Nucleic Acids Res.*, **49**, e54.
80. Pelka,K., Hofree,M., Chen,J.H., Sarkizova,S., Pirl,J.D., Jorgji,V., Bejnood,A., Dionne,D., Ge,W.H., Xu,K.H., *et al.* (2021) Spatially organized multicellular immune hubs in human colorectal cancer. *Cell*, **184**, 4734–4752.
81. Chang,W., Cheng,J., Allaire,J.J., Xie,Y. and McPherson,J. (2015) shiny: web application framework for R. R package version 0.11. 1. *Google Scholar*, https://CRAN.R-project.org/package=shiny.
82. Luo,Y., Hitz,B.C., Gabdank,I., Hilton,J.A., Kagda,M.S., Lam,B., Myers,Z., Sud,P., Jou,J., Lin,K., *et al.* (2020) New developments on the encyclopedia of DNA elements (ENCODE) data portal. *Nucleic Acids Res.*, **48**, D882–D889.

Received: July 14, 2023. **Revised:** October 25, 2023. **Editorial Decision:** November 12, 2023. **Accepted:** November 14, 2023