Comparative and Functional Genomics

S3norm: simultaneous normalization of sequencing depth and signal-to-noise ratio in epigenomic data

Guanjue Xiang[1], Cheryl A. Keller[2], Belinda Giardine[2], Lin An[1], Ross C Hardison[2,*], Yu Zhang[3,*]

[1]The Bioinformatics and Genomics program, Huck Institutes of the Life Sciences, The Pennsylvania State University; [2]Department of Biochemistry and Molecular Biology, The Pennsylvania State University; [3]Department of Statistics, The Pennsylvania State University.

Abstract

Motivation: Quantitative comparison of epigenomic data across multiple cell types is a promising way to understand the biological functions of epigenetic modifications. However, differences in sequencing depth and signal-to-noise ratios ratios compromise our ability to infer real biological differences from raw epigenomic data. Proper data normalization is therefore needed before meaningful data analyses. Existing normalization methods are designed mainly for scaling signals in either whole-genome or peak regions, without considering potential difference in signal-to-noise ratios between data sets.

Results: We propose a new data normalization method called S3norm to standardize data by a monotonic nonlinear transformation to match signals in both peak regions and background regions differently, such that both sequencing depth and signal-to-noise ratio between data sets can be normalized simulatenously. We show that the epigenomic data normalized in this way can better reflect real biological variation across multiple cell types.

Contact: rch8@psu.edu, yzz2@psu.edu

Availability and implementation: https://github.com/guanjue/S3norm

1 Introduction

With advanced sequencing technologies, a large amount of epigenomic data have been generated to date (ENCODE Project Consortium, 2012; Bernstein *et al.*, 2010; Stunnenberg *et al.*, 2016; Martens and Stunnenberg, 2013). Quantitative comparisons among these data across multiple cell types is a powerful approach to study the biological functions of epigenetic modifications. Heterogeneity in the data caused by differences in aspects such as sequencing depth (SD) and signal-to-noise ratio (SNR), however, can lead to systematic biases in the observed epigenomic variation. The observed epigenetic variation may not reflect real biological variation across data sets. Proper data normalization is thus needed to correct these biases before meaningful data analyses (Shao et al., 2012; Meyer and Shirley Liu, 2014).

Many methods for normalizing sequencing data have been developed (Meyer and Liu, 2014; Tu and Shao, 2017). A commonly used approach is to calculate a scale factor. One type of scale factor is the ratio of mean signals between two data sets, e.g., a sample data set against a reference data set (Marioni *et al.*, 2008; Zhang *et al.*, 2008). Other methods such as SES and NCIS use input data to calculate the scale factor (Diaz et al., 2012; Liang and Keleş, 2012). These latter methods first identify the background regions using the input data set, and then they calculate the scale factor from the background regions only. Methods using a single scale factor assume that either the overall means or the background means across different epigenomic data sets are similar. While these methods work well for adjusting different SDs between data sets, they cannot normalize different SNRs, i.e., the different signal scales between peak regions and background regions (Fig. 1). Consequently, these methods are only suitable for normalizing data sets with same SNRs, as one scale factor is sufficient to shift the scatterplot of the two data sets onto the y = x line (Fig. 1A and B). When two data sets have different SNRs, these methods cannot match signals in both peak regions and background regions simultaneously between two data sets (Fig. 1C and D), as in such case, the different SNRs between two data sets can only be adjusted by rotating the signals in the scatter plot (Fig. 1E and F).

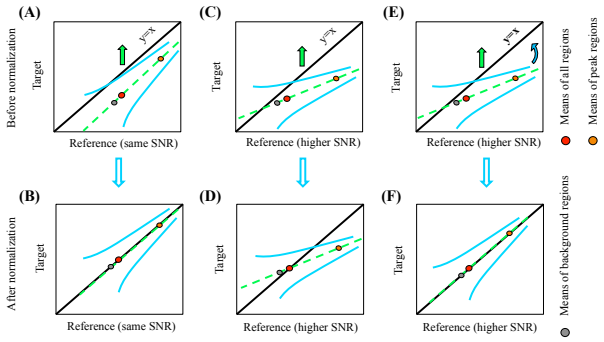

Fig 1. The common patterns of the scatterplot for the signals in two epigenomic data sets. **(A, B)** The scatterplot of two epigenomic data sets with different SDs but same SNR before **(A)** and after **(B)** total signal normalization. **(C, D)** The scatterplot of two epigenomic data sets with different SDs and different SNRs before **(C)** and after **(D)** total signal normalization. Here, the reference data set has a higher SNR relative to the target data set. **(E, F)** The intuitive way to adjust the different SNRs between two data sets is to rotate the signal (blue arrow) in the scatterplot.

A few recent methods can normalize SNRs between data sets (Tu and Shao, 2017). MAnorm (Shao et al., 2012) first identifies common peak regions between two data sets and then uses a robust linear model to fit and transform the MA plot of the common peak regions so that both SD and SNR between the two data sets are normalized. MAnorm assumes that the SNRs between data sets in the common peak regions are equal. MAnorm also assumes that in each data set the

SNR for the common peak regions is the same as the SNR for the background regions. This second assumption is unlikely to hold in practice, as peak regions and background regions have very different signal distributions. Also, this assumption is sensitive to peak calling results. If peak calling is inaccurate or if there are too few common peaks, the model fitted by MAnorm can be unreliable. Similar to MAnorm, ChIPComp also compares the signal difference between two ChIP-seq data sets (Chen *et al.*, 2015). ChIPComp treats the signal of each peak as a sum of true signal, background signal and the data-specific SNR in log scale in a statistical model . However, the method only uses the user provided peak regions to train the model, which has similar issues as in MAnorm. Other methods, such as using LOWESS to fit and adjust the MAplot (Ritchie *et al.*, 2015), may change the rank of signals after normalization, i.e., low signals may be converted to high signals, and vice versa, therefore creating false positives and false negatives. The quantile normalization method (refered to as QTnorm hereafter) adjusts both SDs and SNRs by equalizing the entire signal distributions between two data sets (Bolstad et al., 2003; Nair et al., 2012). When using QTnorm on data sets with very different numbers of peaks, however, it may create many false positive (or false negative) peaks within the data set with fewer (or more) peaks (Bolstad *et al.*, 2003). Finally, rank-based methods have been proposed to normalize data sets with different signal distributions by converting signals into ranks (Lyu and Li, 2016). However, rank-based methods will lose power by ignoring the quantitative spread among signals, and thus they are not considered in this study.

We propose a new method called S3norm to Simultaneously normalize SDs and SNRs of epigenomic data sets by a monotonic nonlinear transformation of signals. Our goal is to (1) match the mean signals in the common peak regions between two data sets, and at the same time (2) match the mean signals in the common background regions between two data sets. In contrast to MAnorm, which only focuses on common peak regions, S3norm uses a two-factor model to address the potential issue of different SNRs in common peak regions versus common background regions, where the imbalance between the two types of regions in the genome are reweighted such that they contribute equally to data normalization. As a result, S3norm will robustly match both SDs and SNRs between two data sets. Specifically, S3norm will increase the peak signals in the low SNR data set without inflating its background signals, which is the key difference between S3norm and existing methods. We will use epigenomic data sets with a wide range of quality to demonstrate the superior performance of S3norm than existing methods.

2 Materials and Methods

2.1 Data preprocessing and evaluation

We used the data sets generated from the **V**al**I**dated and **S**ystematic integrat**ION** of epigenomic data project (**VISION**: usevision.org), which include up to eight epigenetic marks in twenty hematopoietic cell types (Oudelaar et al., 2017; Philipsen and Hardison, 2018) (Keller, Xiang, et al. manuscript in preparation). We first used the bam file processed by the pipeline of VISION project as the input data (Langmead et al., 2009; Heuston et al., 2018) (Keller, Xiang, et al. manuscript in preparation). We then divided the mm10 genome into about 13 million 200-bp bins and counted the number of reads mapped to each bin (Quinlan, 2014). The read counts are used as the raw signal for each data set. The VISION data sets were generated in different laboratories at different time using different technologies, and thus they serve as a challenging test set for our method. For each data set, we used the number of mapped reads to quantify SD, and the Fraction of Reads in Peaks (FRiP) score to measure SNR (Landt et al., 2012). As shown in figure 2, even within the same epigenetic mark, the number of mapped reads ranged widely from <1 million reads to >10 million reads. The FRiP score ranged from <0.1 to >0.9. This wide range in both SDs and SNRs for the VISION data sets illustrates the need for normalization of both aspects.

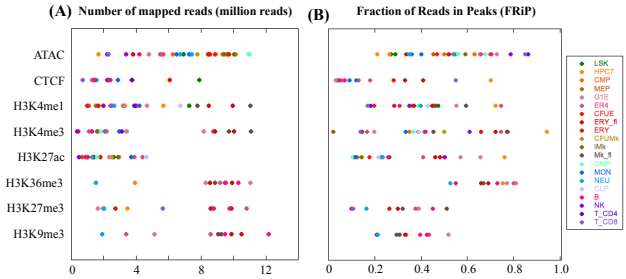

Fig 2. The evaluation of data heterogeneity in the VISION project. **(A, B)** The number of mapped reads **(A)** and the FRiP score **(B)** of each data set. Each point represents one data set. The color of the point represents the cell type of the data set.

2.2 S3norm normalization

2.2.1 Converting read counts to p-values

We first converted the raw read counts into -log10 p-values (Fig. 3A) based on a dynamic background model (Zhang et al., 2008). In ChIP-seq data, there are both background signals and peak signals. Since a globally fitted background model using all data is confounded by the peak signals, we used a two-round method adopted from the hotspot peak calling algorithm to better approximate the background model (John et al., 2011). In the first round, we learned a global background model using all data to reveal significant peak signals. This is done by computing a Negative Binomial (NB) background model (Robinson et al., 2010) for the read counts in all the 200bp bins. We often observe an excessive number of zeros in the real data, which may be due to low sequencing depth. To avoid biased estimation due to the zero bins, we only used the non-zero bins to estimate the size and the probability of success parameters of the NB background model as follows:

$$r_i \sim NB(s_1, p_1)$$

$$p_1 = \frac{\mu_{all-non-0}}{M2_{all-non-0} - \mu^2_{all-non-0} \times (1 - p_0)}$$

$$s_1 = \frac{\mu_{all-non-0} \times (1 - p_0) \times p_1}{(1 - p_1)}$$

$$M2_{all-non-0} = \frac{\sum_{i=1}^{N_{all-non-0}} {r_i}^2}{N_{all-non-0}},$$

where r_i denotes the read count in bin i, s_1 denotes the size parameter, and p_1 denotes the probability of success of the NB background model, with p_1 constrained to $0.1 \le p_1 \le 0.9$ to avoid extreme values, $\mu_{all-non-0}$ denotes the mean of read counts in all of the non-zero bins, $M2_{all-non-0}$ denotes the mean square of the read counts in all of the non-zero bins, p_0 denotes the probability of NB derived zero bins, $N_{all-non-0}$ denotes the number of the non-zero bins. Since the observed number of the zero-count bins is different from the theoretical expectation, we applied the Bayes' Rule to adjust the calculated p-value for each 200-bp bin i:

$$pval_{1,i} = p(r_i \mid r_i \sim NB(s_1, p_1))$$

$$P_{exp1,0} = p(0 \mid r_i \sim NB(s_1, p_1))$$

$$P_{obs-0} = \frac{N_{all-non-0}}{N_{all}}$$

$$pval_adj_{1,i} = pval_{1,i} \times \frac{P_{obs-0}}{P_{exp1,0}},$$

where $pval_{1,i}$ is the p-value of bin i with read count greater than or equals to r_i based on the global background model $NB(s_1, p_1)$, $P_{exp1,0}$ is the probability of zero-count bins derived from the global background model $NB(s_1, p_1)$, P_{obs-0} is the observed proportion of zero bins. $pval_adj_{1,i}$ is the adjusted p-value of each 200-bp bin based on the NB background model accounting for zero inflation. If a bin has read count 0, its p-value is set to 1.

In the second round, we removed the bins with significant p-values ($pval_adj_{1,i}$ < 1e-3) from the previous round and used the remaining bins to calculate a new NB background model. The new NB model and the corresponding p-values of read counts were calculated in the same way as in the first round, except that this time, in addition to removing significant peaks, we also replaced the global size parameter s_1 for the NB model by a local size parameter $s_{2,i}$ to adjust for local variation of input signals at each bin i. Specifically, $s_{2,i}$ is the global size parameter multiplied by the ratio between the mean input signal at bin i and the global mean of the input. Note that the input signal for each ChIP-seq experiment is utilized at this step to adjust for local variation in the background ChIP-seq signal. As shown in figure 3A, the shoulder of the peaks in raw reads count are reduced after converting to the $-\log10(pval_{2,i,j})$.

If *k* replicate data sets are available, we merge the p-values of these replicates by the Fisher's method (Fisher, 1925) as follows:

$$fs_i = -2 \times \sum_{j=1}^{k} \ln(pval_{2,i,j})$$

$$fs_i \sim \chi^2_{2k},$$

where $pval_{2,i,j}$ denotes the p-value of replicate j in bin i obtained by the two-step method explained above. The p-value $pval_{merge,i}$ of fs_i is then calculated from a χ^2_{2k} distribution. The -log10 $pval_{merge,i}$ was used as the processed signal at each position along the genome.

2.2.2 Selecting reference data set

We assume that a data set with higher SNR is more likely to detect real epigenetic signal. Therefore, we used the data set with the highest SNR as our reference data set in S3norm normalization (Fig. 3B). To calculate the SNR of a data set, we first divided the genome into peak regions and background regions, using a threshold of FDR=0.05. For this step, users can alternatively provide their own peak calling results. If the number of peak regions is less than 10,000, the top 10,000 bins with the highest signal will be used as peak regions. We then calculated the SNR of each data set as the ratio of the mean signal (μ_{peak}) in peak regions divided by the mean signal ($\mu_{background}$) in the background regions:

$$SNR = \frac{\mu_{peak}+1}{\mu_{background}+1},$$

2.2.3 Simultaneous normalization of SDs and SNRs

The third step of S3norm is to match the SDs and the SNRs of each data set with those of the reference data set (Fig. 3C). Since different data sets may have unique peaks, we only used the signals in common peak regions and common background regions to fit a normalization model in S3norm. Our assumption is that the SDs and the SNRs of common peak regions and common background regions should be the same between data sets after proper normalization. We thus aim to match the means of both common peak regions and common background regions between the two data sets. In particular, we used the following monotonic nonlinear function to transform the signals in the target data set:

$$Y_{norm,i} = \alpha(Y_i + \varepsilon)^{\beta} - \varepsilon,$$

where Y_i denotes the signal in the target data set before normalization, and $Y_{norm,i}$ denotes the signal in the target data set after normalization. Constants α, β and ε are three positive parameters to be learned by comparing common peak and common background regions between the two data sets. Specifically, α is a scalar that is similar to the scale factor used in the total mean normalization method, which can shift the signals of the target data set in log scale; β is a power transformation parameter that can increase the SNR of the target data set by lifting the high signals in peak regions more than the low signals in background regions; when $\beta \neq 1$, the signals of the target data set is rotated in log scale; when $\beta = 1$, the method becomes a linear transformation. Since we are only matching SDs and SNRs between data sets, using α and β alone would be sufficient to achieve the goal. However, this occasionally will lead to a very large β due to an excessive number of zeros in the data. We therefore added a small positive value ε to all bins in the target data set to attenuate the impact of zeros in S3norm. We calculated ε as:

$$\varepsilon = \frac{\mu_{cpk1} \times \mu_{cbg2} - \mu_{cbg1} \times \mu_{cpk2}}{(\mu_{cbg1} - \mu_{cpk1}) - (\mu_{cbg2} - \mu_{cpk2})},$$

where μ_{cpk1} and μ_{cbg1} are the means of common peak and common background regions, respectively, in the reference data set, and μ_{cpk2} and μ_{cbg2} are the means of common peak and common background regions in the target data set. The rationale is that, by adding ε to the target data set, the SDs and SNRs between the target and the reference data sets may be matched at β =1, i.e., a total mean normalization would suffice. However we bound ε to be $0.1 \leq \varepsilon \leq 1$. That is, S3norm is equivalent to the total mean normalization (β =1) by adding a small value ε to the target data set, when ε is small enough. On the other hand, when the needed ε is too large (>1), we only add 1 to the target data set and alternatively use $\beta \neq 1$ (i.e., non-linear transformation) to achieve the same goals of matching SDs and SNRs. We used the Newton-Raphson method to learn α and β (Ypma, 1995). The initial values for α and β are 1.0 and 2.0, respectively.

2.3 Evaluating normalized H3K4me3 signal by RNA-seq

H3K4me3 is a histone modification associated with active promoters in gene expression. To test which normalization method better reflects the real biological signal strength, we compared the difference of normalized H3K4me3 signals with the difference of the RNA-seq signal between cell type pairs. We assumed that the direction of the better normalized H3K4me3 changes should be consistent with the direction of RNA-seq changes; a similar idea was used in the evaluation of the MAnorm method (Shao *et al.*, 2012). Here, we used the ENCODE pipeline to process the RNA-seq data (ENCODE Project Consortium, 2012). Then, we used the TMM method in the edgeR package to normalize the library size (Robinson and Oshlack, 2010; Robinson et al., 2010). We calculated the mean signal of H3K4me3 ChIP-seq at TSS (+-5 KB) for each protein-

coding gene, and we calculated the Reads Per Kilobase Per Million mapped reads (RPKM) of RNA-seq as the expression level for each protein coding gene (Mortazavi et al., 2008). To reduce the noise generated from the non-expressed genes, we restricted the analysis to the genes expressed in at least one cell type. The distribution of the maximum log2-RPKM of each gene across cell types was bi-modal with a valley at log2-RPKM= -1.5. Thus, we selected as expressed the genes whose maximum log2-RPKM values were greater than or equal to -1.5. After the filtration, 12130 protein coding genes remained. To scale the RNA-seq signal and the ChIP-seq signal to the same level, we subtracted the mean signal and divided by the standard deviation of RNA-seq of all genes across all cell types for RNA-seq. For ChIP-seq, we used the same method for scaling.

For the comparison, we used R^2 to evaluate the relationship between the difference of normalized H3K4me3 signal and the difference of RNA-seq signal. We used the difference of the normalized H3K4me3 signal as our predictor and the RNA-seq difference as our response variable. The performance was measured by the R^2:

$$R^2 = 1 - \frac{\sum_i^n (\mathrm{RNAdif_i} - \mathrm{H3K4me3dif_i})^2}{\sum_i^n \left(\mathrm{RNAdif_i} - \mu_{\mathrm{RNAdif_i}}\right)^2},$$

where $\mathrm{RNAdif_i}$ is the difference of RNA-seq signal between two cell types, and $\mathrm{H3K4me3dif_i}$ is the difference of normalized H3K4me3 signal between cell types. The theoretical best score of the R^2 equals to 1 which means the difference of the normalized H3K4me3 signal perfectly reflected the difference of the RNA-seq between each cell type pair. For better visualization, we plotted the R^2 in the boxplot for the comparison of different normalization methods.

3 Results

3.1 S3norm overview

The epigenomic data generated from different laboratories or techniques can have very different SDs and SNRs which may result from batch effects or technical issues rather than real biological differences (Fig. 2). Existing methods are mainly suited for normalizing different SDs of data sets while assuming the same SNRs. Some methods can alternatively adjust SNRs by using common peak regions in different data sets (MAnorm) or by forcing a common signal distribution across data sets (QTnorm) (Shao *et al.*, 2012; Bolstad *et al.*, 2003). These methods all have their drawbacks, which motivated us to develop a new method called S3norm that uses a nonlinear transformation to normalize both SDs and SNRs simultaneously.

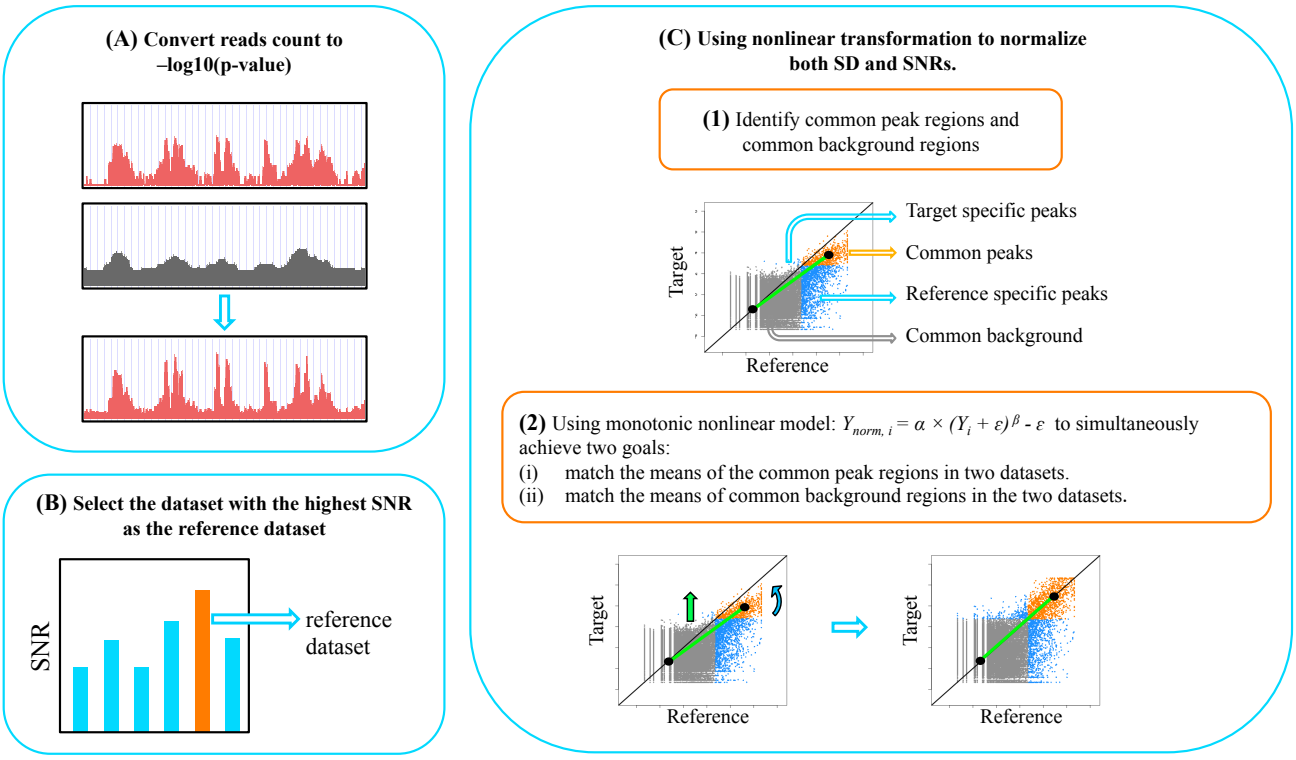

Fig 3. The overview of the S3norm normalization method. There are three major steps in S3norm. **(A)** The 1st step is to convert read counts in the 200-bp bins to a -log10(NB-p-value) for each epigenomic data set. The first track is the raw reads count of the target data set. The second track is the reads count of input. The third tract is the -log10(p-value) determined by two rounds of fitting to a dynamic background model. **(B)** The 2nd step is selecting the data set with the highest SNR (orange bar) as the reference data set for the S3norm normalization. **(C)** The 3rd step is using a monotonic nonlinear data transformation model to normalize both the SNR and SD between the two data sets. The **(1)** part is identifing common peak regions (orange points) and the common background regions (gray points) between the two data sets. The **(2)** part is using the monotonic nonlinear data transformation model to shift and rotate the signal of the target data set. The blue data points represent the data set-specific peaks.

The workflow of the S3norm method is shown in Figure 3. The three major steps in S3norm are described in detail in the Methods. The first step convrted read counts to -log10 p-values using two rounds of dynamic modeling of the background to a negative binomial distribution (Fig. 3A). The second round of this modeling included an adjustment for local background signal variation. In the second step, the data set with the highest SNR was chosen as the reference against which all data sets will be normalized (Fig. 3B). The third step used a nonlinear transformation to match SDs and SNRs between each data set and the reference data set (Fig. 3C). As shown in Fig. 3C, the two black dots represent the means of common peak regions and common background regions respectively. The goal of S3norm was to match these means between the two data sets, i.e., move the two points to the

x=y line in the scatter plot. By doing this, both SDs and SNRs of two data sets are guaranteed to match after the normalization.

As an example, we used the nuclease accessibility data (ATAC-seq) at some histone genes to demonstrate the effect of S3norm normalization in four cell types (Fig. 4) (Buenrostro et al., 2013). Histone genes are expected to have similar activities across all cell types since active production of histones is required for cell replication, and thus we expect their nuclease landscapes in the neighborhood are similar across cell types. As shown in Figure 4A, the raw ATAC-seq signals at these histone genes were clearly weaker in the G1E cells (a model for GATA1-dependent erythroid maturation) and T-CD8 cells than in the common myeloid progenitor cells (CMP) and the erythroblasts cells (ERY). After S3norm normalization (Fig. 4B), the signals of the peak regions in G1E and T-CD8 were increased to a comparable level as in CMP and ERY, while the signals in the background regions were remain low.

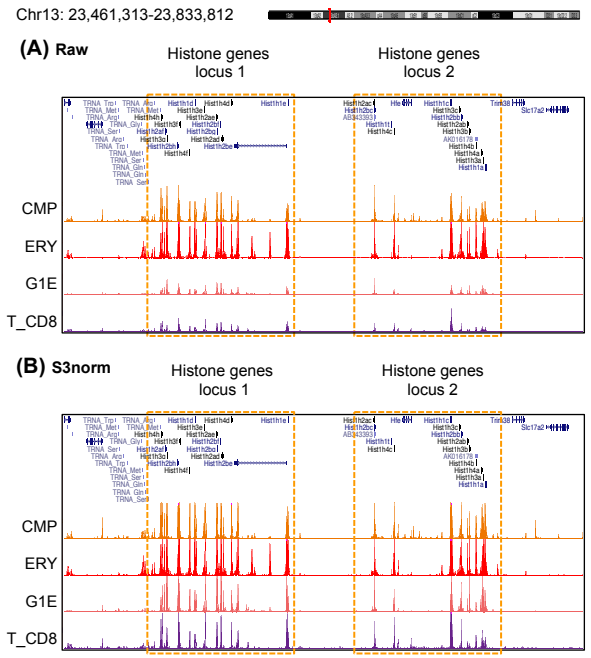

Fig 4. The ATAC-seq/DNase-seq signal at two histone gene loci (orange dash boxes) before **(A)** and after **(B)** S3norm data normalization. The tracks are shown in the same scale, from 0 to 100 –log10 p-value.

3.2 Evaluation of SD and SNR normalization

We compared S3norm with other methods in terms of normalizing SD and SNR. We used the CTCF ChIP-seq data sets in erythroblasts from mouse fetal liver (ERY_fl) and G1E cells from the VISION project for illustration. We chose these two data sets because they have similar total signals (i.e., similar SDs) but different SNRs (Fig. 5A). Since the ERY_fl data set has a higher SNR than the G1E data set, we used the ERY_fl as the reference. As shown in Figure 5A, the SNRs for the two data sets were clearly different, as reflected by the angle between the green line and the y = x line. The purpose of S3norm is to shift and rotate the green line to the y = x line, as is shown in Figure 5E after S3norm normalization. In comparison, the total signal normalization (refered to as TSnorm hereafter) method was able to adjust for SDs, as indicated by the red dot on the y=x line, but it did not match SNRs between the two data sets, as shown by the orange dots away from the y=x line in Figure 5B. The latter observation showed that the peak signals in ERY-fl were still consistently higher than the peak signals in G1E after TSnorm normalization.

We also compared S3norm with MAnorm. MAnorm is one of the first methods to normalize SNRs in sequencing data. Here, we used MAnorm to normalize the G1E data set. The SDs and SNRs in common peak regions between the two data sets were similar after MAnorm normalization (Fig. 5D). Since MAnorm did not consider background regions, however, the MAnorm normalized data may be inconsistent in the background regions. To test this hypothesis, we compared SNRs of all 11 CTCF ChIP-seq data in the VISION project. All data sets were normalized by the four methods. We used the data set with the highest SNR as the reference. We calculated 1 minus the absolute difference between the SNR of the reference data and the SNR of the target data, so that the method with larger statistics are the one that can better normalize SNRs. As shown in Figure 5F, due to the difference of signals in the background region, the SNRs were quite different across all data sets normalized by TSnorm and MAnorm, whereas the SNRs were much more homogeneous across data sets normalized by QTnorm and S3norm. For the TSnorm results, the SNRs did not appear to improve at all compared to the raw data, confirming that single factor normalization cannot normalize SNRs. In comparison, the S3norm and QTnorm effectively adjusted SNRs so that those measurements became much more similar to each other across multiple data sets after normalization.

3.3 Evaluation of false positives and false negatives

Both S3norm and QTnorm methods can match SDs and SNRs simultaneously. QTnorm however aggressively forces different data sets to have identical signal distributions after normalization. Qtnorm may be overly aggressive by forcing a similar number of peaks between two data sets, which will be inappropriate when the true number of peaks differs substantially between two data sets. In our study, the SNRs of the two data sets indeed became very similar after QTnorm normalization (Fig 5C). Simultaneously, the number of CTCF peaks called in G1E (36,901 peaks) and ERY-fl (31,425 peaks) also became similar after QTnorm normalization, despite the fact that the numbers of CTCF peaks called from the raw data sets were substantially lower in G1E cells (11,349 peaks) than in ERY_fl (31,425 peaks). In contrast, the number of peaks called on S3norm normalizaed data were still different between G1E (21,015 peaks) and ERY-fl (31,425 peaks). Since both S3norm and QTnorm are monotonic normalizations, the ranks of signals did not change after normalization. Thus, the 21,015 peaks called on S3norm results are the top 21,015 peaks of the 36,901 peaks called on QTnorm results. To check whether the QTnorm specific peaks are likely false positive peaks, we used FIMO (Grant et al., 2011) in its default setting to scan CTCF binding site motifs from CIS-BP database (Weirauch et al., 2014) in the QTnorm specific peaks. While >80% of the S3norm peaks had a CTCF binding site motif in both ERY-fl (89.9%) and G1E (82.1%), only 57.8% of the QTnorm peaks and 22.9% of the QTnorm unique peaks had CTCF motif in G1E. These results support the hypothesis that QTnorm may have created many false positive peaks due to its aggressive normalization based on false assumptions on the signal distributions between two data sets, while S3norm is substantially more robust to this effect.

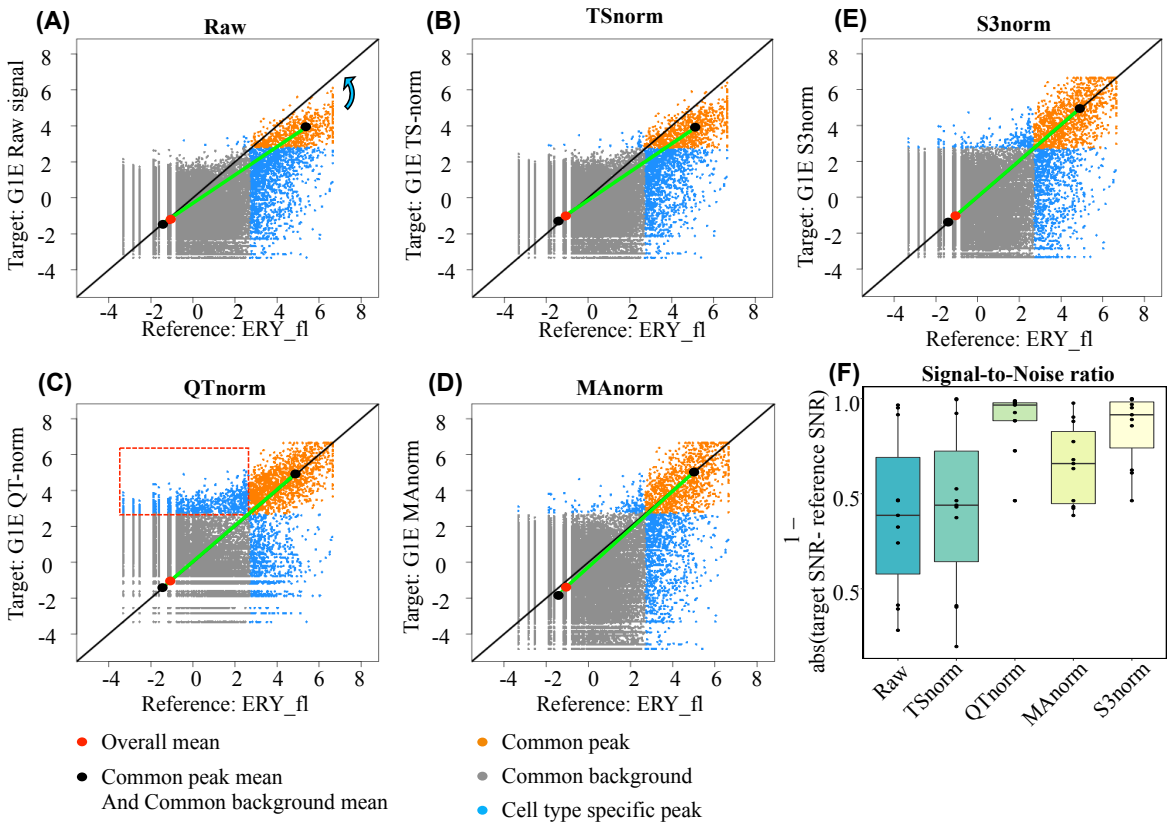

Fig 5. The scatterplot of ERY-fl and G1E CTCF ChIP-seq signal (log2 scale). The reference data set is the ERY-fl and the target data set is G1E. The data were split into three groups: the common peak regions (orange), the common background regions (gray), and the data set specific peaks (blue). The red point represents the overall mean of the two data sets. The two black points represent the mean signal of the common peak and the mean of the common background respectively. **(A)** The scatterplot of raw signal between reference data set and target data set. Intuitively, a proper normalization should rotate the data counterclockwise (blue arrow). **(B)** The scatterplot of the signal after TSnorm normalization. **(C)** The scatterplot of the signal after QTnorm normalization. There were more data points were called as the G1E-specific peaks (blue points in red dash box). **(D)** The scatterplot of the signal after MAnorm normalization. **(E)** The scatterplot of the signal after S3norm normalization. **(F)** The boxplot of one minus the absolute difference of the the SNR between target data sets and reference data set of all CTCF ChIP-seq. Compared with other methods, the SNRs of all data sets become closer to the reference data set after the S3norm and the QTnorm normalizations.

We also compared S3norm and QTnorm in ATAC-seq data sets. Higher ATAC-seq signals and greater numbers of ATAC-seq peaks usually indicate more local regions being accessible in a cell type. We compared the ATAC-seq signals of a population of CMP cells and G1E cells from the VISION project (Fig. 6). The CMP cell population has the ability to differentiate into several different lineages of blood cells, whereas G1E cells are committed to a single lineage (erythroid). This greater potentiality of CMP cells may be reflected in more extensive accessibility of the genome in CMP cells than in G1E cells. We used CMP as the reference and G1E as the target data. After QTnorm, we again obtained similar numbers of ATAC-seq peaks between CMP (77,237 peaks, Fig. 6A gray bar) and G1E (82,282 peaks, Fig. 6A black bar). In contrast, using S3norm we identified substantially fewer ATAC-seq peaks in G1E (35,202 peaks, Fig. 6A blue bar) compared to those in CMP cells. We next calculated the correlations of S3norm signals, and the correlations of QTnorm signals, between two biological replicates using the top N peaks ranked by S3norm signals (blue curve) and QTnorm signals (black curve), respectively (Fig. 6B). Since both S3norm and QTnorm do not change the ranks of signals, the top N peaks in S3norm signals must also be the top N peaks in QTnorm signals. As shown in Figure 6B, for the very top ranked peaks, the signal correlations between the replicates of S3norm and QTnorm results are similar. As more peaks were included, however, the correlation curve of QTnorm data decreased much faster than the curve of S3norm data. This result indicates that QTnorm results were less consistent than S3norm results as the signals become weaker. For example, when we consider all G1E ATAC-seq peaks (82,282 peaks) called by either S3norm or QTnorm, the correlation of the peaks between two resplicates using S3norm was 0.63, whereas the correlation using Qtnorm was only 0.27. Further shown in Figure 6 C and D are the scatterplots of the G1E ATAC-seq peaks identified by either S3norm or QTnorm. Many peaks (mostly QTnorm unique peaks) with lower signals were likely inflated noise by QTnorm, as indicated by the low correlation between two biological replicates. In contrast, using S3norm to normalize data in cell types with very different numbers of peaks is much more robust.

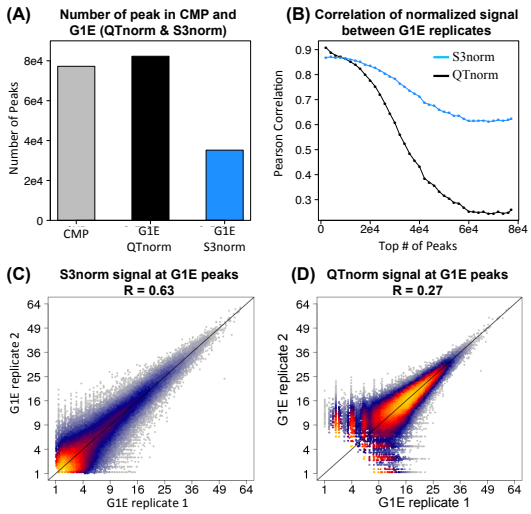

Fig 6. The comparison of ATAC-seq after S3norm and QTnorm. **(A)** The number of peak in CMP cell type (reference data set; gray bar) and G1E cell type (target data set) after QTnorm (balck bar) and S3norm (blue bar). **(B)** The Pearson correlations of S3norm signal and the correlations of QTnorm signal between two biological replicates at the top N peaks with the highest S3norm signals (blue curve) or the highest QTnorm signals (black curve). **(C)** The heatscatterplot of S3norm signal of the two biolgical replicates of the G1E ATAC-seq peaks (82,282 peaks) that are identified in either S3norm or QTnorm. The color represent the density of the scatterplot. The warmer color represents higher density and the colder color represents lower density. **(D)** The heatscatterplot of QTnorm signal of the two biolgical replicates of the G1E ATAC-seq peaks (82,282 peaks) that are identified in either S3norm or QTnorm.

3.4 Evaluation by gene expression

H3K4me3 is positively associated with active promoters, and strongly correlated with expression levels of genes (Jones, 2012). We hypothesize that the changes in better normalized H3K4me3 signals at transcription start site (TSS) should correlate better with the changes of gene expression levels across multiple cell types. If this hypothesis is true, we may use gene expression to evaluate and compare different normalization methods. Here, we used 11 cell types (55 cell-pairs in total) which have both H3K4me3 ChIP-seq data sets and RNA-seq data sets from the VISION project to compare normalization methods. To reduce the noise generated from the non-expressed genes, we selected the genes expressed in at least one cell type based on the log2-RPKM bi-modal distribution (Fig. 7A). As an example, we show the signal difference between the colony-forming unit-erythroid cell population (CFUE) and G1E cells (Fig. 7B-F). RNA-seq signal difference is calculated as RNA-seq signal (log2 RPKM) in CFUE minus the RNA-seq signal in G1E. The ChIP-seq signal difference is calculated as ChIP-seq signal (S3norm normalized -log10 of p-value) in CFUE minus the ChIP-seq signal in G1E. The correlation between H3K4me3 differences and RNA-seq differences for each expressed genes was determined before and after ChIP-seq data normalization by four methods (Fig. 7C-F). For the difference of RNA-seq (x-axis), the points are centered as 0, indicating that most of the genes between the two cell types have similar expression levels. For the difference of H3K4me3 ChIP-seq signals, the S3norm and QTnorm results produced a much higher correlation with the difference of the RNA-seq signal than did TSnorm or MAnorm. In addition, we compared with three additional measures of ChIP-seq signals: raw Read Counts (RC), -log10 p-value based on a dynamic Poisson model (POIS) (Zhang et al., 2008), and Z score of the read counts (Z). As shown in figure 7G, among all methods, the S3norm and QTnorm methods performed the best, i.e., the difference of S3norm normalized H3K4me3 at TSS between cell-pairs were best correlated with the difference of gene expression between cell-pairs.

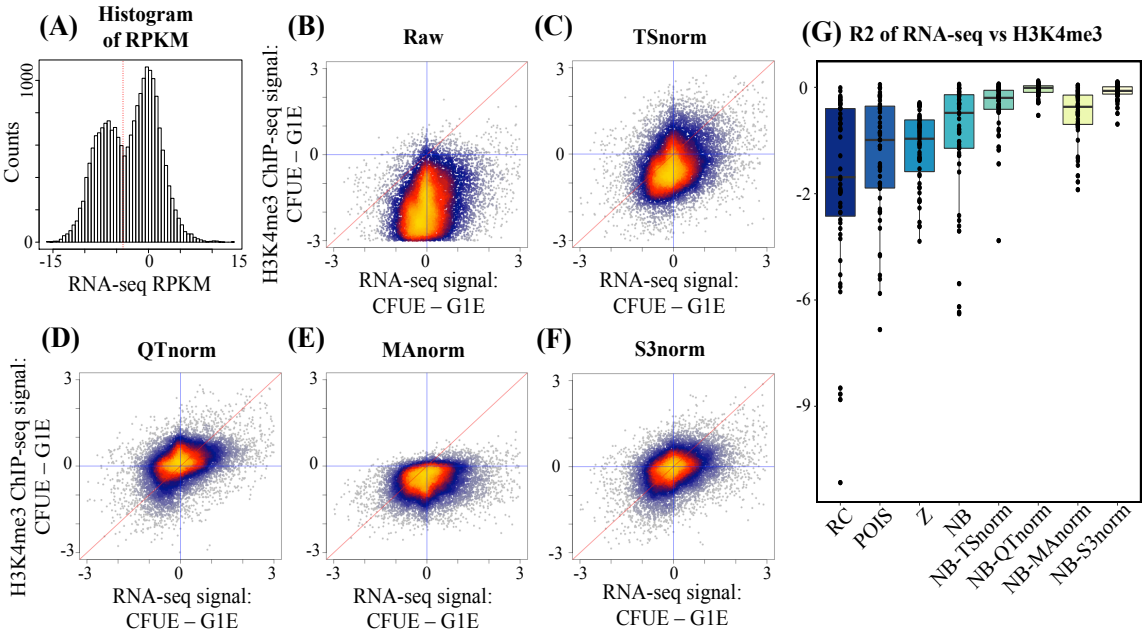

Fig 7. (A) The distribution of maximun RNA-seq RPKM across all cell type in the VISION project. The red line (log2-RPKM = -1.5) is used as the threshold to select genes for the downstream analysis. **(B-F)** The signal difference between CFUE cell type and the G1E cell type. The y-axis is the RNA-seq signal difference. The x-axis is the H3K4me3 ChIP-seq signal difference. The scatterplot is in log2 scale. The color of the point represents the density of the scatterplot. The warmer color represents higher density and the colder color represents lower density. The red line represents the y = x line. **(G)** The boxplot of performance scores of all 55 cell-pairs after different kinds of normalization. The performance is defined as the 1 – SEE. The higher performance score means the difference of RNA-seq between the cell-pair is better explained by the difference of the H3K4me3 ChIP-seq.

4 Discussion

We introduce a simple and robust method to simultaneously normalize SDs and SNRs between epigenomic data sets. The essence of this method is to use a nonlinear transformation to rotate the signal of the target data set to that of the reference data set, so that the mean signals of both common peak regions and common background regions are matched simultaneously between two data sets. Unlike existing methods, S3norm is robust to bias due to either substantial proportion of background regions in the genome, which affects the performance of the total mean normalization and the QTnorm method, or using data in peak regions only, which affects the performance of the MAnorm method. S3norm can be trained even on data sets with a very small number of peaks such as in spike-in data sets (Jiang et al., 2011). In addition, S3norm has very few parameters to be trained from the data, which makes the method resilient to false positives and false negatives.

The key assumption of the S3norm method is that data sets appropriately adjusted to account for differences in SD and SNR will provide more accurate biological insights. For some data sets in which we expect to have different SNRs, the S3norm method may not be directly applicable. For example, for ChIP-seq data sets of transcription factors (TF) whose abundance is changing over the course of an experiment (e.g. in a targeted degradation protocol), we

expect the SNRs of the TF ChIP-seq data sets to change over time. In such cases, one should not use any of the data sets in the time course as a reference for S3norm normalization. Instead, a spike-in control should be paired with the ChIP-seq data at each time point in order for S3norm to work properly.

In summary, S3norm is a robust way to normalize both SDs and SNRs across multiple data sets. The results of its normalization can better reflect real biological differences than existing methods. As epigenomic data sets continue to be generated in many different laboratories and consortia, S3norm can become increasingly useful to normalize signals across these diverse and heterogeneous epigenomic data sets and better highlight true epigenetic changes against technical bias. Improved normalization will better facilitate the studies that analyze data sets across multiple cell types, such as differential gene regulation, genome segmentation (Zhang et al., 2016; Ernst and Kellis, 2012) , joint peak calling (Stark and Brown, 2011), predicting gene expression (Singh et al., 2016), and detecting transcription factor binding events (Alipanahi et al., 2015).

Data Availability

Files for raw signals, p-value converted signals, and signals from S3norm are available both for download and for viewing from the VISION website (http:// usevision.org).

Acknowledgements

The work is supported by NIH grants GM121613, DK106766, and CA178393.

Reference

Alipanahi,B. *et al.* (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*

Bernstein,B.E. *et al.* (2010) The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.*

Bolstad,B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*.

Buenrostro,J.D. *et al.* (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*.

Chen,L. *et al.* (2015) A novel statistical method for quantitative comparison of multiple ChIP-seq datasets. *Bioinformatics*.

Diaz,A. *et al.* (2012) CHANCE: comprehensive software for quality control and validation of ChIP-seq data. *Genome Biol.*

ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*.

Ernst,J. and Kellis,M. (2012) ChromHMM: Automating chromatin-state discovery and characterization. *Nat. Methods*.

Fisher,R. (1925) Statistical methods for research workers. *Biol. Monogr. manuals*.

Grant,C.E. *et al.* (2011) FIMO: Scanning for occurrences of a given motif. *Bioinformatics*.

Heuston,E.F. *et al.* (2018) Establishment of regulatory elements during erythro-megakaryopoiesis identifies hematopoietic lineage-commitment points. *Epigenetics and Chromatin*.

Jiang,L. *et al.* (2011) Synthetic spike-in standards for RNA-seq experiments. *Genome Res.*

John,S. *et al.* (2011) Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.*

Jones,P.A. (2012) Functions of DNA methylation: Islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.*

Landt,S.G. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*

Liang,K. and Keleş,S. (2012) Normalization of ChIP-seq data with control. *BMC Bioinformatics*.

Lyu,Y. and Li,Q. (2016) A semi-parametric statistical model for integrating gene expression profiles across different platforms. *BMC Bioinformatics*.

Marioni,J.C. *et al.* (2008) RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*

Martens,J.H.A. and Stunnenberg,H.G. (2013) BLUEPRINT: Mapping human blood cell epigenomes. *Haematologica*.

Meyer,C.A. and Liu,X.S. (2014) Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat. Rev. Genet.*

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*.

Nair,N.U. *et al.* (2012) Chipnorm: A statistical method for normalizing and identifying differential regions in histone modification chip-seq libraries. *PLoS One*.

Oudelaar,A.M. *et al.* (2017) Between form and function: The complexity of genome folding. *Hum. Mol. Genet.*

Philipsen,S. and Hardison,R.C. (2018) Evolution of hemoglobin loci and their regulatory elements. *Blood Cells, Mol. Dis.*

Quinlan,A.R. (2014) BEDTools: The Swiss-Army tool for genome feature analysis. *Curr. Protoc. Bioinforma.*

Ritchie,M.E. *et al.* (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*

Robinson,M.D. *et al.* (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*.

Robinson,M.D. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*

Shao,Z. *et al.* (2012) MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. *Genome Biol.*

Singh,R. *et al.* (2016) DeepChrome: Deep-learning for predicting gene expression from histone modifications. In, *Bioinformatics*.

Stark,R. and Brown,G. (2011) DiffBind : differential binding analysis of ChIP-Seq peak data. *Bioconductor*.

Stunnenberg,H.G. *et al.* (2016) The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell*.

Tu,S. and Shao,Z. (2017) An introduction to computational tools for differential binding analysis with ChIP-seq data. *Quant. Biol.*

Weirauch,M.T. *et al.* (2014) Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell*.

Ypma,T.J. (1995) Historical Development of the Newton–Raphson Method. *SIAM Rev.*

Zhang,Y. *et al.* (2016) Jointly characterizing epigenetic dynamics across multiple human cell types. *Nucleic Acids Res.*

Zhang,Y. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*