Supplemental Notes for

MAnorm2 for quantitatively comparing groups of ChIP-seq samples

Shiqi Tu[1,2], Mushan Li[1], Haojie Chen[1,2], Fengxiang Tan[1,2], Jian Xu[3], David J. Waxman[4], Yijing Zhang[5], Zhen Shao[1]*

[1]CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China.
[2]University of Chinese Academy of Sciences, Beijing 100049, China.
[3]Children's Medical Center Research Institute, Department of Pediatrics, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA.
[4]Department of Biology and Bioinformatics Program, Boston University, Boston MA 02215, USA.
[5]National Key Laboratory of Plant Molecular Genetics, CAS Center for Excellence in Molecular Plant Sciences, Shanghai Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200032, China.

***Corresponding author:**
Dr. Zhen Shao
Email: shaozhen@picb.ac.cn

Contents

Supplemental Note S1. Differential analysis between two groups of ChIP-seq samples

1.1 Model formulation and hypothesis testing

1.2 Mean-variance curve fitting and parameter estimation

1.3 Deducing a parametric form for the mean-variance curve

1.4 Statistical simulation

Supplemental Note S2. Being adaptive to the regularity of variance structure

Supplemental Note S3. Integrating LOESS and robust linear regression into hierarchical MA normalization

Supplemental Note S4. Simultaneously comparing multiple groups of ChIP-seq samples

References

Supplemental Note S1. Differential analysis between two groups of ChIP-seq samples

Here we give a detailed description of the statistical model designed in MAnorm2 for calling differential ChIP-seq signals between two groups of samples corresponding to different biological conditions, provided that all related samples have been normalized to be comparable with each other.

1.1 Model formulation and hypothesis testing

For $j = 1,2$, suppose X_j is an $n \times m_j$ matrix recording normalized ChIP-seq signal intensities (i.e., normalized $\log_2$ read counts) at n genomic intervals for m_j samples belonging to condition j. Let $X_{i,j}$ be a column vector representing the transpose of row i of X_j. We assume

$$X_{i,j} \left| t_{i,j} \sim MVN\left(1\mu_{i,j}, S_{i,j}t_{i,j}\right)\right. \tag{1}$$

Here MVN refers to the multivariate normal distribution; $\mu_{i,j}$ and $t_{i,j}$ are two unknown scalars that parametrize the mean signal intensity of interval i in condition j and the associated signal variability, respectively; 1 is a column vector of ones; $S_{i,j}$, termed structure matrix, is an $m_j \times m_j$ matrix designed for the convenience of incorporating existing tools for modeling the precision weights of signal measurements from different samples as well as the correlations among them (Smyth 2004; Smyth et al. 2005; Law et al. 2014). All structure matrices used in the study were simply identity matrices. MAnorm2 next derives mean and variance estimates by applying the generalized least squares method:

$$\begin{aligned} \widehat{\mu_{i,j}} &= \left(1^T S_{i,j}^{-1} 1\right)^{-1} 1^T S_{i,j}^{-1} X_{i,j} \\ \widehat{t_{i,j}} &= \frac{\left(X_{i,j} - 1\widehat{\mu_{i,j}}\right)^T S_{i,j}^{-1} \left(X_{i,j} - 1\widehat{\mu_{i,j}}\right)}{m_j - 1} \end{aligned} \tag{2}$$

In practice, m_j are typically very small (most ChIP-seq data sets only have two or three biological replicates for each experiment), which results in large uncertainty associated with the $t_{i,j}$ estimators and, thus, low statistical power for the following differential tests. To improve variance estimates for individual genomic intervals, MAnorm2 borrows strength between intervals with similar signal levels and captures the underlying mean-variance dependence by fitting a smooth mean-variance curve (MVC). Specifically, it assumes the MVCs of the two groups of samples have the same shape and differ from each other only by a scaling factor. Formally, MAnorm2 defines $\sigma_{i,j}^2 = t_{i,j} / \gamma_j$, where γ_j, termed variance ratio factor, is a parameter that quantifies the global variability of ChIP-seq signals across the samples of group j. Naturally, we have $\widehat{\sigma_{i,j}^2} = \widehat{t_{i,j}} / \gamma_j$. And the complete Bayesian model that uses mean-variance trend as source of prior information is given by

$$\begin{gathered} X_{i,j} \left| \sigma_{i,j}^2 \sim MVN\left(1\mu_{i,j}, S_{i,j}\left(\gamma_j \sigma_{i,j}^2\right)\right)\right. \\ \frac{1}{\sigma_{i,j}^2} \sim \frac{1}{f\left(\mu_{i,j}\right)} \cdot \frac{\chi_{d_0}^2}{d_0} \end{gathered} \tag{3}$$

Overall, the above model is similar to limma (Smyth 2004) and an extension of it (Sartor et al. 2006), except that MAnorm2 allows for different global within-group variability between groups of samples. Here $f(\cdot)$ refers to an unscaled MVC common to the two groups of samples and $f\left(\mu_{i,j}\right)$ is called the prior variance of interval i in group j; χ_N^2 refers to the chi-squared distribution with N degrees of freedom; d_0, termed the number of prior degrees of freedom, is a hyper-parameter designed for assessing how well in general the variance of an individual interval could be predicted by its mean signal intensity. In practice, d_0 amounts to the number of extra ChIP-seq samples gained by sharing information between genomic intervals, and the use of these “samples” typically contributes significantly to the estimation of variances for individual intervals. Note also that MAnorm2 empirically determines d_0 based on the data, which renders the method adaptive to the regularity of variance structure associated with the specific data set (see the section below and also Supplemental Note S2).

For the following differential tests, we further assume that the unscaled variances of non-differential genomic intervals remain invariant across the two groups of samples.

Formally, for each interval i that satisfies $\mu_{i,1} = \mu_{i,2}$, we assume $\sigma_{i,1}^2$ equals $\sigma_{i,2}^2$ with a probability of one (i.e., they can be treated as the same random variable). This assumption is consistent with the fact that $\sigma_{i,1}^2$ and $\sigma_{i,2}^2$ follow the same prior distribution as long as $\mu_{i,1} = \mu_{i,2}$. Finally, MAnorm2 tests the null hypothesis $H_0 : \mu_{i,1} = \mu_{i,2}$ for each interval i by using the following key statistic:

$$\tilde{T}_i = \frac{\widehat{\mu_{i,2}} - \widehat{\mu_{i,1}}}{\sqrt{\left(\frac{\gamma_1}{1^T S_{i,1}^{-1} 1} + \frac{\gamma_2}{1^T S_{i,2}^{-1} 1}\right)\tilde{\sigma}_i^2}} \tag{4}$$

where

$$\tilde{\sigma}_i^2 = \frac{d_0 f\left(\frac{\widehat{\mu_{i,1}} + \widehat{\mu_{i,2}}}{2}\right) + (m_1 - 1)\widehat{\sigma_{i,1}^2} + (m_2 - 1)\widehat{\sigma_{i,2}^2}}{d_0 + m_1 + m_2 - 2} \tag{5}$$

According to the theoretical deduction presented in Smyth et al. (Smyth 2004), we assert that $\tilde{T}_i$, termed moderated *t*-statistic, follows a *t*-distribution under the null hypothesis with $(d_0 + m_1 + m_2 - 2)$ degrees of freedom, disregarding the uncertainty associated with the mean estimate for determining prior variance (i.e., $\left(\widehat{\mu_{i,1}} + \widehat{\mu_{i,2}}\right)/2$). Many existing methods derive the mean signal intensities for determining prior variances (or dispersions, if the negative binomial distribution is used) by taking the average signal intensities across individual samples (Sartor et al. 2006; Love et al. 2014), which could lead to unbalanced statistical power for identifying up-regulated signals for the two conditions, especially when the numbers of samples belonging to the two conditions differ dramatically from each other. To alleviate this effect, MAnorm2 chooses to take the average signal intensities across conditions, which is especially helpful for balancing the statistical power when d_0 is much larger than both m_1 and m_2.

The resulting moderated *t*-statistics can be used to effectively rank genomic intervals in order of statistical evidence of having differential signals between the two conditions. MAnorm2 also gives the exact (two-sided) *p*-values by

$$p_i = 2 \cdot T_{d_0+m_1+m_2-2}\left(-\left|\tilde{T}_i\right|\right) \tag{6}$$

where $T_N(\cdot)$ refers to the cumulative distribution function of the *t*-distribution with N degrees of freedom.

1.2 Mean-variance curve fitting and parameter estimation

This section discusses the estimation of f, d_0, γ_1 and γ_2.

To fit the MVC in an unbiased manner, MAnorm2 calculates mean and variance estimates separately within each group of samples and pools the resulting mean-variance pairs into a regression process, which is different from many previous methods that derive a single mean-variance pair for each individual genomic interval (Sartor et al. 2006; Law et al. 2014; Love et al. 2014). To make variance estimates comparable between the two groups of samples, we first deduce an estimate of γ_2/γ_1. Given the model formulation presented in the previous section, we have $\frac{\widehat{t_{i,2}}/\gamma_2}{\widehat{t_{i,1}}/\gamma_1} \sim F_{m_2-1,m_1-1}$ for each interval i that satisfies $\mu_{i,1} = \mu_{i,2}$, where F_{N_1,N_2} refers to the *F*-distribution with N_1 and N_2 degrees of freedom. And we give an estimator of γ_2/γ_1 as

$$\widehat{\gamma_2/\gamma_1} = \frac{median_i\left(\widehat{t_{i,2}}/\widehat{t_{i,1}}\right)}{F_{m_2-1,m_1-1}^{-1}(1/2)} \tag{7}$$

where $F_{N_1,N_2}^{-1}(\cdot)$ denotes the inverse of the cumulative distribution function of the indicated *F*-distribution. Here we use sample median instead of sample mean to perform the estimation because the former is more robust to the influence of differential intervals. To further improve the unbiasedness of the estimator, we use only the genomic intervals that are occupied by both groups of samples to calculate the median (see Methods in the main text for a detailed explanation of occupancy states of genomic intervals).

MAnorm2 next pools the mean-variance pairs of the form $\left(\widehat{\mu_{i,1}}, \widehat{t_{i,1}}\right)$ or $\left(\widehat{\mu_{i,2}}, \frac{\widehat{t_{i,2}}}{\widehat{\gamma_2/\gamma_1}}\right)$ into a weighted gamma-family regression process, with $(m_1 - 1)$

and $(m_2 - 1)$ as the weights of observations from group 1 and 2, respectively. Note that, to enhance the regularity of the data on which the regression is performed, MAnorm2 selects for each group of samples only the genomic intervals that are occupied by it to calculate mean-variance pairs. Currently, we have devised two candidate schemes for performing the regression. One of them uses a theoretically derived parametric form to fit a generalized linear model (see the following section for the specific form as well as the deduction of it). This method is most suited to data sets with a highly regular variance structure, in which the mean-variance relationship could be expected to be well profiled by the presumed formula (e.g., when the samples of each group are biological replicates for the same experiment). The other method adopts a local regression procedure implemented in the locfit package (Loader 1999). This method allows for more general mean-variance relationships (Anders and Huber 2010). Whichever method is chosen, MAnorm2 iteratively fits an MVC and detects outliers by using 1e-4 and 15 as lower and upper bounds of residuals (i.e., the ratios of observed variances to fitted ones), respectively (Love et al. 2014). Outliers detected in a round of iteration are removed from the fitting of MVC in the next round, and the whole regression process finishes as soon as the set of outliers fixes.

In the rest of this section, we ignore the uncertainty associated with the estimate of f, since it is typically fitted on a great number of observations. The method for estimating d_0 is similar to the one used in limma (Smyth 2004), except that MAnorm2 integrates two estimation results that are respectively derived from the two groups of samples. Formally, we define $z_{i,j} = \log \frac{\widehat{t_{i,j}}}{f\left(\widehat{\mu_{i,j}}\right)}$. Given $\frac{\widehat{t_{i,j}}}{\gamma_j f\left(\mu_{i,j}\right)} \sim F_{m_j - 1, d_0}$, which can be deduced from equation (2) and (3), we assert that the marginal distribution of $z_{i,j}$ is a scaled Fisher's *z*-distribution plus a constant (Aroian 1941), disregarding the uncertainty associated with $\widehat{\mu_{i,j}}$. And we have

$$\begin{aligned} E\left[z_{i,j}\right] &\approx \log \gamma_j + \psi\left(\frac{m_j - 1}{2}\right) - \psi\left(\frac{d_0}{2}\right) + \log \frac{d_0}{m_j - 1} \\ \operatorname{var}\left[z_{i,j}\right] &\approx \psi'\left(\frac{m_j - 1}{2}\right) + \psi'\left(\frac{d_0}{2}\right) \end{aligned} \tag{8}$$

where $\psi(\cdot)$ and $\psi'(\cdot)$ are the digamma and trigamma functions, respectively.

MAnorm2 next uses these two moments to estimate γ_j and d_0.

Noticing that $z_{i,j}$ from the same group of samples (approximately) have the same expectation and variance, we give, using the data associated with each group j, an estimate of $\psi'\left(d_0/2\right)$ by

$$D_j = \frac{\sum_i \left(z_{i,j} - \sum_{i'} z_{i',j} / n_j\right)^2}{n_j - 1} - \psi'\left(\frac{m_j - 1}{2}\right) \tag{9}$$

Note that MAnorm2 only uses the genomic intervals that have been used for fitting f to estimate γ_j and d_0. Therefore, each of the sum operators in equation (9) is applied only to the intervals whose mean-variance pairs in group j have been involved in the regression process, and n_j denotes the number of such intervals. n_j varies with j, as MAnorm2 selects only occupied intervals from each group of samples to fit f. The final estimate of d_0 is obtained by solving

$$\psi'\left(\frac{\widehat{d_0}}{2}\right) = \frac{\sum_j \left(n_j - 1\right) D_j}{\sum_j \left(n_j - 1\right)} \tag{10}$$

whose right-hand side has a form similar to the pooled sample variance in a two-sample *t*-test. Note that $\widehat{d_0}$ is set to positive infinity if the right-hand side of equation (10) is less than or equal to 0, since in such cases there is no evidence supporting the variation of the underlying variance (i.e., $\sigma_{i,j}^2$) across genomic intervals with the same mean signal intensity. Note also that the marginal distribution of each $\widehat{t_{i,j}}$ is a scaled chi-squared distribution when d_0 is positive infinity (as the number of denominator degrees of freedom of an *F*-distribution approaches positive infinity, the *F*-distribution converges to a scaled chi-squared distribution), which is consistent with the use of a gamma-family regression procedure to fit f.

Finally, we derive for each group j an estimate of γ_j by

$$\widehat{\gamma_j} = \exp\left\{ \frac{\sum_i z_{i,j}}{n_j} - \psi\left(\frac{m_j - 1}{2}\right) + \psi\left(\frac{\widehat{d_0}}{2}\right) - \log\frac{\widehat{d_0}}{m_j - 1} \right\} \tag{11}$$

Again, the above sum operator is applied only to the genomic intervals whose mean-variance pairs in group j have been used for fitting f.

1.3 Deducing a parametric form for the mean-variance curve

MAnorm2 deduces an explicit formula for profiling mean-variance trend by assuming a quadratic relationship between expectations and variances of read count variables, which has been proposed by several previous studies (Robinson et al. 2010; Law et al. 2014; Love et al. 2014). Formally, suppose Y is a random variable standing for a read count and that it satisfies

$$\mathrm{var}[Y] = \beta_0 E^2[Y] + \beta_1 E[Y] \tag{12}$$

Of note, the whole framework designed in MAnorm2 is for analyzing continuous measurements, which are typically obtained by applying a $\log_2$ transformation to read counts. We next deduce an approximate formula that connects the variance of $\log_2 Y$ with its expectation by using the delta method (Oehlert 1992). Formally, by defining $X = \log_2 Y$ and investigating its one-order Taylor expansion at $Y = E[Y]$, we have $X \approx \log_2 E[Y] + (Y - E[Y]) \cdot \frac{dX}{dY}|_{Y=E[Y]}$. It follows that

$$E[X] \approx \log_2 E[Y] \tag{13}$$

$$\mathrm{var}[X] \approx \frac{\mathrm{var}[Y]}{E^2[Y]} \cdot \frac{1}{(\log 2)^2} \tag{14}$$

Finally, by substituting the right-hand side of equation (12) for $\mathrm{var}[Y]$ in (14) and using (13) to replace $E[Y]$, we have

$$\mathrm{var}[X] \approx \beta_0' + \beta_1' 2^{-E[X]} \tag{15}$$

where $\beta_i' = \frac{\beta_i}{(\log 2)^2}$ for $i = 0,1$. This form shall be used by MAnorm2 for performing a gamma-family generalized linear regression with identity link.

1.4 Statistical simulation

We performed statistical simulation across various settings to verify the effectiveness of the whole parameter estimation framework. In each simulation, data (i.e., normalized ChIP-seq signal intensities) were generated based on the model assumption of MAnorm2, and the occupancy states of genomic intervals were ignored (or equivalently, each genomic interval was considered to be occupied by each group of samples).

Each simulation was about a comparison between two groups of samples, and the data for each genomic interval were independently generated. Here, we give a formal description of the process for generating data for a specific interval i (we keep using the notations defined in 1.1). The point is to first determine $\mu_{i,j}$ and $t_{i,j}$ for $j = 1, 2$. Then, the data are generated by equation (1) (all structure matrices used in statistical simulation were identity matrices). If interval i is non-differential, $\mu_{i,j}$ and $t_{i,j}$ are determined by

$$\begin{aligned}
&\mu_{i,1} \sim U\left(A_{lower}, A_{upper}\right) \\
&\mu_{i,2} = \mu_{i,1} \\
&\frac{1}{\sigma_{i,1}^2} \sim \frac{1}{f\left(\mu_{i,1}\right)} \cdot \frac{\chi_{d_0}^2}{d_0} \\
&\sigma_{i,2}^2 = \sigma_{i,1}^2 \\
&t_{i,1} = \gamma_1 \sigma_{i,1}^2 \\
&t_{i,2} = \gamma_2 \sigma_{i,2}^2
\end{aligned} \tag{16}$$

where $U(a,b)$ refers to the uniform distribution with a and b as lower and upper bounds, respectively. Otherwise, $\mu_{i,j}$ and $t_{i,j}$ are determined by

$$\begin{aligned}
&A_i \sim U\left(A_{lower}, A_{upper}\right) \\
&M_i \sim N\left(0, \sigma_M^2\right) \\
&\mu_{i,1} = A_i - M_i/2 \\
&\mu_{i,2} = A_i + M_i/2 \\
&\frac{1}{t_{i,1}} \sim \frac{1}{\gamma_1 f\left(\mu_{i,1}\right)} \cdot \frac{\chi_{d_0}^2}{d_0} \\
&\frac{1}{t_{i,2}} \sim \frac{1}{\gamma_2 f\left(\mu_{i,2}\right)} \cdot \frac{\chi_{d_0}^2}{d_0}
\end{aligned} \tag{17}$$

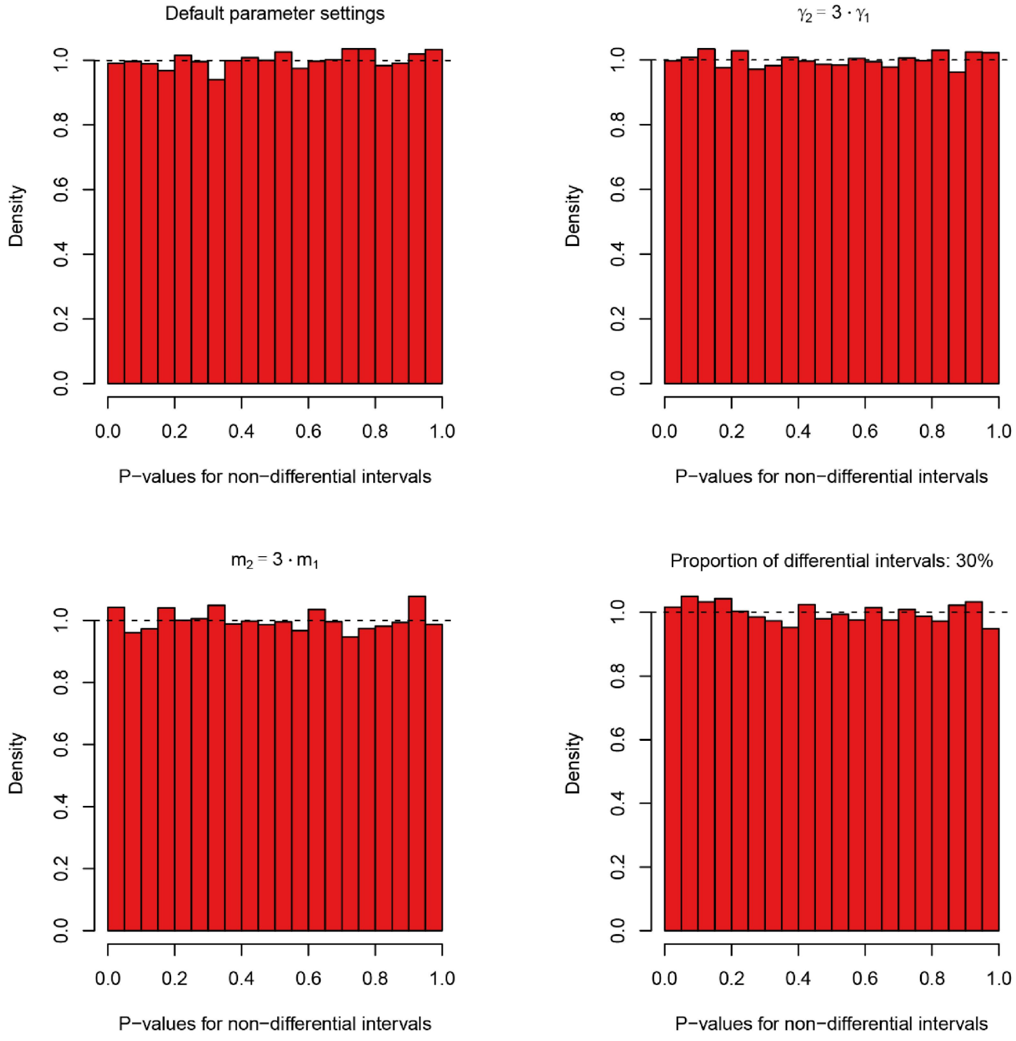
where N refers to the normal distribution.

For each simulation, we generated data for 30,000 genomic intervals, and the default proportion of differential intervals was 10%. Default settings of the other parameters were as follows:

$$\begin{aligned}
&A_{lower} = 3, A_{upper} = 11 \\
&\sigma_M^2 = 4 \\
&f(x) = 0.005 + 20 \cdot 2^{-x} \\
&d_0 = 10 \\
&\gamma_2 = \gamma_1 = 1 \\
&m_2 = m_1 = 3
\end{aligned} \tag{18}$$

These settings generally matched our observations from practical ChIP-seq data sets.

We used default parameter settings with modification of at most one parameter in each simulation, and we inspected the distribution of *p*-values assigned to non-differential intervals for an overall assessment of the parameter estimation framework:



These simulations involved cases with distinct global within-group variability between groups of samples, unbalanced group sizes, or abundant differential intervals. It can be seen that the *p*-value distribution associated with non-differential intervals is very uniform on $(0,1)$ across different parameter settings, which indicates the effectiveness of the parameter estimation framework.

Complete R source code for performing the simulations and generating the histograms can be found in Supplemental Code.

Supplemental Note S2. Being adaptive to the regularity of variance structure

MAnorm2 introduces the notion of prior degrees of freedom in the modeling of mean-variance trend, which is for achieving a wide applicability to data sets of various characteristics. Here by characteristics of a data set we specifically refer to the regularity of the associated variance structure, which we assess by quantifying the variability of variance residual (from the regression of variance on mean signal intensity) across different genomic intervals (i.e., the variability of $z_{i,j}$ across different i for some fixed j; see Supplemental Note S1). For clarification, if this variability is low for a data set, it is considered to be of high regularity. Note that MAnorm2 typically derives large numbers of prior degrees of freedom for data sets of high regularity (see equation (9) and (10) in Supplemental Note S1).

In practice, the regularity of variance structure as well as the number of prior degrees of freedom derived by MAnorm2 (denoted by $\widehat{d_0}$ in the following) varies significantly across data sets. This is because there are quite a few factors that could influence the variability of ChIP-seq signals across samples (Steinhauser et al. 2016), including technical noise, biological variation, batch effects, and so on. As more such factors are involved in a data set, it becomes harder to predict the variances for individual genomic intervals depending solely on their mean signal intensities, and the associated $\widehat{d_0}$ decreases accordingly. For example, we have performed all pairwise comparisons of H3K4me3 ChIP-seq signals among four human lymphoblastoid cell lines (LCLs; GM12890, GM12891, GM12892 and SNYDER) derived from different individuals of the Caucasian population (Supplemental Fig. S2). In this analysis, each group of ChIP-seq samples belonging to the same biological condition was comprised of biological replicates for an individual LCL, and the associated variance structure could be expected to be highly regular. As a result, $\widehat{d_0}$ derived by MAnorm2 for these comparisons ranged from 9.5 to 137.5 and had a median of 18.8. For contrast, we have also made a between-sex comparison (of H3K4me3 levels) by classifying the four LCLs into two males (GM12891 and SNYDER) and two females (GM12890 and GM12892; Supplemental Fig. S1B, D). Note that, for this analysis, we created a reference profile for each individual LCL by taking the average ChIP-seq signals across its replicates (see Methods in the main text for details about reference profile). Compared to the previous scenario, ChIP-seq profiles grouped together in this case were additionally associated with epigenetic variation across human individuals (Kasowski et al. 2013) and, thus, the associated variance structure should become less regular. Consistently, MAnorm2 derived a $\widehat{d_0}$ of 4.2.

MAnorm2 improves its adaptivity to the specific data set by empirically estimating d_0 and using the estimation result to determine the relative contributions of observed variances and prior ones to the final variance estimates for individual intervals (see equation (5) in Supplemental Note S1). Based on the corresponding gene expression data, we compared the performance of MAnorm2 with two variants of it in the above two scenarios of differential ChIP-seq analysis. Note that we used the comparison between GM12890 and SNYDER LCLs (2 vs. 2 biological replicates) as representative of the first scenario, as this comparison was associated with the same group sizes as the between-sex comparison (2 vs. 2 LCLs). The two variants of MAnorm2, referred to as no-MVC and MVC-only respectively, adopt different strategies to derive final variance estimates. Specifically, no-MVC and MVC-only directly use observed and prior variances as final variance estimates, respectively, while MAnorm2 integrates the two types of variances by taking a weighted average, with the weights depending on the d_0 estimated from the data (technically, no-MVC and MVC-only are equivalent to always treating d_0 as 0 and positive infinity, respectively). As described above, the variance structure in the first scenario was of high regularity, and there was a low variability of variance across genomic intervals with the same mean signal intensity (Supplemental Fig. S1A). In this case, prior variances alone can serve as reliable variance estimates for individual intervals. By contrast, the observed variances were associated with large uncertainty due to the small numbers of biological replicates, which resulted in a much worse performance of no-MVC compared to MVC-only (Supplemental Fig. S1C). As for MAnorm2, it derived a $\widehat{d_0}$ of 14.6, which was more than seven times the number of observed degrees of freedom (i.e., the number of free signal measurements for calculating the observed variance of each interval, which equals the total number of ChIP-seq samples minus the number of groups). Consequently, the final variance estimates used by MAnorm2 were dominated by prior variances, and MAnorm2 therefore exhibited virtually the same performance as MVC-only. In comparison with the first scenario, the variance structure in the second one was considerably less regular, and the $\widehat{d_0}$ derived by MAnorm2 was only 4.2 (Supplemental Fig. S1B), which was comparable to the number of observed degrees of freedom. In this scenario, no-MVC continued to suffer from small group sizes, and MVC-only completely ignored the high fluctuation of variance across intervals with the same mean signal intensity. As a result, both no-MVC and MVC-only were clearly outperformed by MAnorm2 (Supplemental Fig. S1D). Together, these results demonstrated the adaptivity of MAnorm2 and suggested a wide applicability of it.

Supplemental Note S3. Integrating LOESS and robust linear regression into hierarchical MA normalization

In principle, any regression process allowing extrapolation can be integrated into the hierarchical MA normalization framework (see Methods in the main text). For this integration, the original framework is largely retained, only that the underlying technique for normalizing an individual sample or a reference profile against another is changed, which is now achieved by designing abundance-dependent offsets to remove M-A trend.

Suppose that X and Y are two vectors of $\log_2$ read counts (we used an offset of 0.5 in the study) representing raw signal intensities of two ChIP-seq samples in a list of genomic intervals. Let M and A values be defined by $M = Y - X$ and $A = \frac{1}{2}\left(X + Y\right)$, respectively. We now normalize Y against X by using an arbitrary regression process to fit M-A trend. Specifically, we perform a regression by using $M_{\_}$ and $A_{\_}$ as responses and predictor values, respectively, where $\_$ indicates the vectors are subsetted to common peak regions (i.e., the intervals occupied by both samples). Let $f\left(\cdot\right)$ be the resulting (vectorized) regression function. The normalization is accomplished by applying the following transformation:

$$Y^{*} = Y - f\left(A\right) \tag{19}$$

For normalizing a reference profile against another, the offset vector (i.e., $f\left(A\right)$) derived for the former is equally applied to each individual sample of the corresponding group.

In the study, we have separately used LOESS (local polynomial regression) and robust linear regression to implement this normalization algorithm, and the resulting two normalization methods were referred to as loess and rlm-offset respectively (Supplemental Fig. S15). We applied LOESS by using the loess function provided by the R package stats, with `control=loess.control(surface="direct")` to allow extrapolation (R Core Team 2018). For robust linear regression, we used the rlm function of the MASS package with default parameters (Venables and Ripley 2002).

We have also used another way to integrate robust linear regression into the hierarchical MA normalization framework, which was referred to as rlm-linear. For this integration, f is still fitted as described above. Given the nature of linear regression, f is determined by an intercept and a slope, and a connection between M and A values can therefore be established as

$$M \sim \beta_0 + \beta_1 A \tag{20}$$

This connection can be expanded to

$$Y - X \sim \beta_0 + \frac{\beta_1}{2}\left(X + Y\right) \tag{21}$$

which is then rearranged as

$$X \sim \frac{1-\beta_1/2}{1+\beta_1/2}Y - \frac{\beta_0}{1+\beta_1/2} \tag{22}$$

Finally, the linear transformation given by the right hand of equation (22) is applied to Y to finish the normalization. Again, for normalizing a reference profile against another, the linear transformation derived for the former is equally applied to each individual sample of the corresponding group.

Supplemental Note S4. Simultaneously comparing multiple groups of ChIP-seq samples

This note gives a formal description of the statistical model designed in MAnorm2 for simultaneously comparing more than two groups of ChIP-seq samples corresponding to different biological conditions. We keep using the notations defined in Supplemental Note S1, except that the group index j now takes integers from 1 through C, where C is the total number of groups to be compared. To be rigorous, we give a succinct but still self-contained description of the related model formulation and hypothesis testing.

For each genomic interval i in each group j, we assume

$$\begin{aligned} X_{i,j} \left| \sigma_{i,j}^2 \right. &\sim MVN\left(1\mu_{i,j}, S_{i,j}\left(\gamma_j \sigma_{i,j}^2\right)\right) \\ \frac{1}{\sigma_{i,j}^2} &\sim \frac{1}{f\left(\mu_{i,j}\right)} \cdot \frac{\chi_{d_0}^2}{d_0} \end{aligned} \tag{23}$$

We further assume that the unscaled variance of each non-differential genomic interval remains invariant across groups. Formally, for each interval i that satisfies $\mu_{i,1} = \mu_{i,2} = \ldots = \mu_{i,C}$, we assume $\sigma_{i,1}^2 = \sigma_{i,2}^2 = \ldots = \sigma_{i,C}^2$ happens with a probability of one (i.e., they can be treated as the same random variable). This assumption is consistent with the fact that $\sigma_{i,1}^2, \sigma_{i,2}^2, \ldots, \sigma_{i,C}^2$ follow the same prior distribution as long as $\mu_{i,1} = \mu_{i,2} = \ldots = \mu_{i,C}$. For later use, we derive expressions of mean and variance estimators by applying the generalized least squares method:

$$\begin{aligned} \widehat{\mu_{i,j}} &= \left(1^T S_{i,j}^{-1} 1\right)^{-1} 1^T S_{i,j}^{-1} X_{i,j} \\ \widehat{t_{i,j}} &= \frac{\left(X_{i,j} - 1\widehat{\mu_{i,j}}\right)^T S_{i,j}^{-1} \left(X_{i,j} - 1\widehat{\mu_{i,j}}\right)}{m_j - 1} \end{aligned} \tag{24}$$

Methods detailed in Supplemental Note S1 for estimating f, d_0, and γ_j can be naturally extended with few modifications to cases involving more than two groups of samples. Specifically, for fitting f, we select a group as baseline and derive an estimate of γ_j / γ_b for each $j \neq b$ by using equation (7), where b refers to the selected baseline group. Then, the mean-variance pairs having a form of $\left(\widehat{\mu_{i,b}},\widehat{t_{i,b}}\right)$ or $\left(\widehat{\mu_{i,j}},\frac{\widehat{t_{i,j}}}{\gamma_j/\gamma_b}\right)$ with $j \neq b$ are pooled into a weighted gamma-family regression process, with $\left(m_j - 1\right)$ as the weight of observations from group j. For the selection of baseline group, MAnorm2 utilizes an algorithm similar to the one for selecting a baseline ChIP-seq sample to normalize a group of samples (see Methods in the main text). Specifically, it first picks out the genomic intervals that are occupied by all the C groups and uses their $\widehat{t_{i,j}}$ to construct a matrix, whose rows and columns correspond to the intervals and the groups, respectively. MAnorm2 then applies the median-ratio strategy (Anders and Huber 2010) to the matrix and derives the "size factor" of each group. Finally, the group whose $\log_2$ size factor is closest to 0 is selected as baseline. After fitting f, the estimation of d_0 and γ_j is accomplished by using equation (9), (10) and (11). The resulting estimates of f, d_0, and γ_j are treated as non-stochastic in subsequent statistical tests, as they are typically derived based on a great number of observations.

We next detail the procedure for testing the null hypothesis $H_0 : \mu_{i,1} = \mu_{i,2} = \ldots = \mu_{i,C}$ for each interval i. As in the one-way analysis of variance (ANOVA), we first fit the full model and calculate the corresponding residual sum of squares (RSS):

$$RSS_i = \sum_{j=1}^{C} \frac{\left(m_j - 1\right)\widehat{t_{i,j}}}{\gamma_j} \tag{25}$$

We then fit a reduced model by assuming all the C biological conditions are associated with the same mean signal intensity in interval i:

$$\mu_i^{(0)} = \frac{\sum_{j=1}^{C}\left(\frac{1^T S_{i,j}^{-1} 1}{\gamma_j}\right)\widehat{\mu_{i,j}}}{\sum_{j=1}^{C}\frac{1^T S_{i,j}^{-1} 1}{\gamma_j}} \tag{26}$$

where $\mu_i^{(0)}$ is intrinsically a weighted average of mean estimates from different groups, with the weights being inversely proportional to their variances. And the associated RSS can be derived by

$$RSS_i^{(0)} = \sum_{j=1}^{C} \frac{\left(X_{i,j} - 1\mu_i^{(0)}\right)^T S_{i,j}^{-1} \left(X_{i,j} - 1\mu_i^{(0)}\right)}{\gamma_j} \tag{27}$$

Before defining the final key statistic for testing the null hypothesis, we summarize some facts regarding the distributions of associated random variables as follows. Under the H_0, we have $\mu_{i,1} = \mu_{i,2} = \ldots = \mu_{i,C}$ (denoted by μ_i in the following) and that $\sigma_{i,1}^2, \sigma_{i,2}^2, \ldots, \sigma_{i,C}^2$ refer to the same random variable (denoted by σ_i^2 in the following). Based on equation (23) and previous studies of one-way ANOVA, we have (under the H_0)

$$\begin{aligned} &\frac{1}{\sigma_i^2} \sim \frac{1}{f(\mu_i)} \cdot \frac{\chi_{d_0}^2}{d_0} \\ &RSS_i \left| \sigma_i^2 \perp \left(RSS_i^{(0)} - RSS_i \right) \right| \sigma_i^2 \\ &RSS_i \left| \sigma_i^2 \sim \sigma_i^2 \cdot \chi_{\sum_j m_j - C}^2 \right. \\ &\left(RSS_i^{(0)} - RSS_i \right) \left| \sigma_i^2 \sim \sigma_i^2 \cdot \chi_{C-1}^2 \right. \end{aligned} \tag{28}$$

in which the second formula indicates that the two random variables are conditionally (on σ_i^2) independent of each other. Equation (28) gives all the results that are necessary for us to derive

$$\frac{\left(RSS_i^{(0)} - RSS_i \right) \Big/ (C-1)}{\left(RSS_i + d_0 f(\mu_i) \right) \Big/ \left(\sum_j m_j - C + d_0 \right)} \sim F_{C-1, \sum_j m_j - C + d_0} \tag{29}$$

Finally, we define a moderated *F*-statistic for interval i as

$$\tilde{F}_i = \frac{\left(RSS_i^{(0)} - RSS_i \right) \Big/ (C-1)}{\left(RSS_i + d_0 f\left(\sum_j \widehat{\mu_{i,j}} \Big/ C \right) \right) \Big/ \left(\sum_j m_j - C + d_0 \right)} \tag{30}$$

which *approximately* follows $F_{C-1, \sum_j m_j - C + d_0}$ under the null hypothesis, considering the uncertainty of the mean estimate for deducing the prior variance of interval i (i.e., $\sum_j \widehat{\mu_{i,j}} \Big/ C$). $\tilde{F}_i$ has a form similar to the classical *F*-statistic in one-way ANOVA, except that its variance estimate (i.e., the denominator) has incorporated additional information regarding σ_i^2 , which is exactly obtained by modeling mean-variance dependence. In practice, this incorporation of prior variances helps stabilizing variance estimates for individual intervals as well as increasing the statistical power for identifying differential signals, which can be seen from the increased number of denominator degrees of freedom associated with $\tilde{F}_i$. Note also that $\tilde{F}_i$ is similar to the moderated *F*-statistic designed in limma (Smyth 2004), except that the latter uses a constant prior variance for all genomic intervals and does not take mean-variance dependence into account. As explained in Supplemental Note S1, here we derive the mean estimates for determining prior variances by taking the average signal intensities across groups of samples rather than individual samples, which is for avoiding biasing the mean estimates towards the groups that have more samples than the others. In practice, such biases typically lead to stronger statistical power for identifying up-regulated signals in the conditions with more samples. Taking the average signal intensities across groups is especially effective for alleviating the unbalanced statistical power when d_0 is much larger than $\left(\sum_j m_j - C\right)$.

Accordingly, MAnorm2 gives the *p*-value of the statistical test for interval i by

$$p_i = 1 - F_{C-1,\sum_j m_j - C + d_0}\left(\tilde{F}_i\right) \tag{31}$$

where $F_{N_1,N_2}(\cdot)$ refers to the cumulative distribution function of the *F*-distribution with N_1 and N_2 degrees of freedom.

References

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome biology* **11**(10): R106.

Aroian LA. 1941. A Study of R. A. Fisher's *z* Distribution and the Related F Distribution. *Ann Math Statist* **12**(4): 429-448.

Kasowski M, Kyriazopoulou-Panagiotopoulou S, Grubert F, Zaugg JB, Kundaje A, Liu Y, Boyle AP, Zhang QC, Zakharia F, Spacek DV et al. 2013. Extensive variation in chromatin states across humans. *Science* **342**(6159): 750-752.

Law CW, Chen Y, Shi W, Smyth GK. 2014. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology* **15**(2): R29.

Loader C. 1999. *Local Regression and Likelihood*. Springer, New York.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**(12): 550.

Oehlert GW. 1992. A Note on the Delta Method. *Am Stat* **46**(1): 27-29.

R Core Team. 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**(1): 139-140.

Sartor MA, Tomlinson CR, Wesselkamper SC, Sivaganesan S, Leikauf GD, Medvedovic M. 2006. Intensity-based hierarchical Bayes method improves testing for differentially expressed genes in microarray experiments. *BMC bioinformatics* **7**: 538.

Smyth GK. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology* **3**: Article3.

Smyth GK, Michaud J, Scott HS. 2005. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics* **21**(9): 2067-2075.

Steinhauser S, Kurzawa N, Eils R, Herrmann C. 2016. A comprehensive comparison of tools for differential ChIP-seq analysis. *Briefings in bioinformatics* **17**(6): 953-966.

Venables WN, Ripley BD. 2002. *Modern Applied Statistics with S*. Springer, New York.