Supplementary Material of the manuscript

"Multi-scale Poisson process approaches for differential expression analysis of high-throughput sequencing data"

Heejung Shim, Zhengrong Xing, Ester Pantaleo, Francesca Luca, Roger Pique-Regi, Matthew Stephens

1 Supplementary Material Table 1

		Exploit high-resolution information in data?	
		No - overall expression method	Yes - multi-scale approach
Model the count nature of data?	No	Window-based approaches from Pickrell *et al.* (2010); Degner *et al.* (2012)	WaveQTL (Shim and Stephens, 2015), WFMM (Lee and Morris, 2016)
	Yes	DESeq2 (Love *et al.*, 2014), edgeR (Robinson *et al.*, 2010)	multiseq

Table 1: Four categories of methods that identify differentially expressed molecular phenotypes between multiple groups of samples using $*$-seq data. This table presents four categories of differential expression methods according to two features: 1) overall expression method vs. multi-scale methods that use high-resolution information in data (in addition to overall expression) by combining information across different scales; 2) whether it directly model the count nature of the data. Note that some differential expression methods belong to none of the four categories (e.g., derfinder (Frazee *et al.*, 2014)).

2 Supplementary Material Figure 1

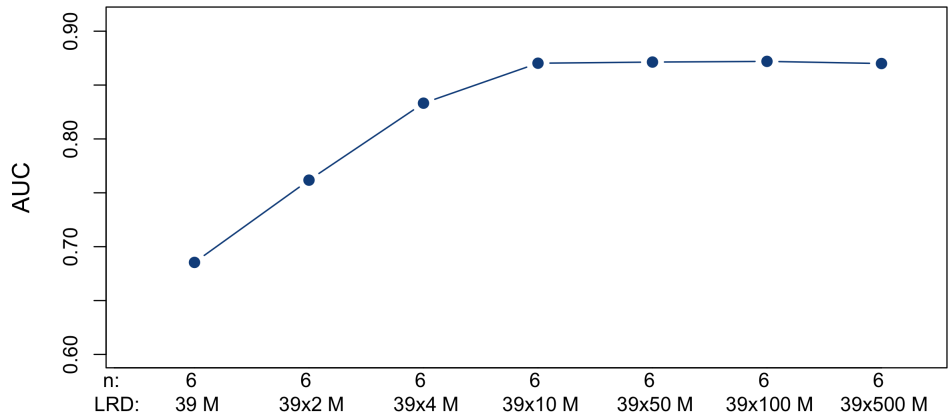

Figure 1: **There is an upper bound for performance that multiseq can attain by increasing read depth.** Performance of multiseq (AUC) is measured at sample size 6 ($n = 6$) as an expected library read depth (LRD) increases up to 39×500 M.

3 Supplementary Material Figure 2

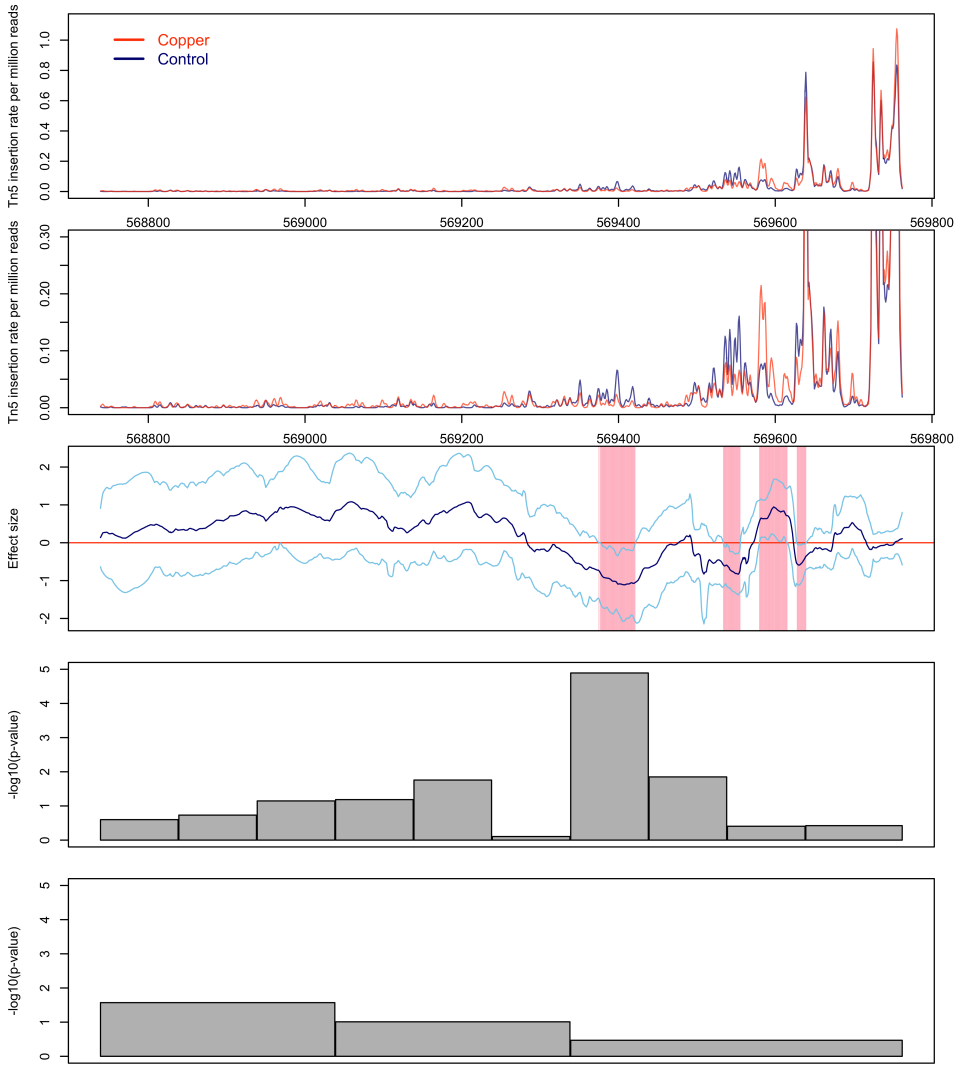

Figure 2: **Example of DERs identified by multiseq, but not by DESeq2 (chr1:568739-569762)** The top three panels are the same as Figure 3 of the main text. The last two panels show $-\log_{10}$(p-value) from DESeq2 for each bin when we use different bin sizes (100bp for the forth panel; 300bp for the bottom panel; See Supplementary Material section 11.4 for DESeq2 analysis with different bin sizes). Note that the p-value from DESeq2 for each bin is different from p-value from the empirical distribution of the test statistic. For multiseq, $\log \Lambda \approx 98.91$; p-value ≈ 0.0000034. For DESeq2, p-values from the empirical distribution are ≈ 0.19 for 1024 bin, ≈ 0.0025 for 300 bin, and ≈ 0.0000029 for 100 bin. In the forth panel, the seventh 100bp bin fully overlaps with the first signal (consistent in direction over about 50bp), so DESeq2 with 100 bin successfully detects the signal.

4 Supplementary Material Figure 3

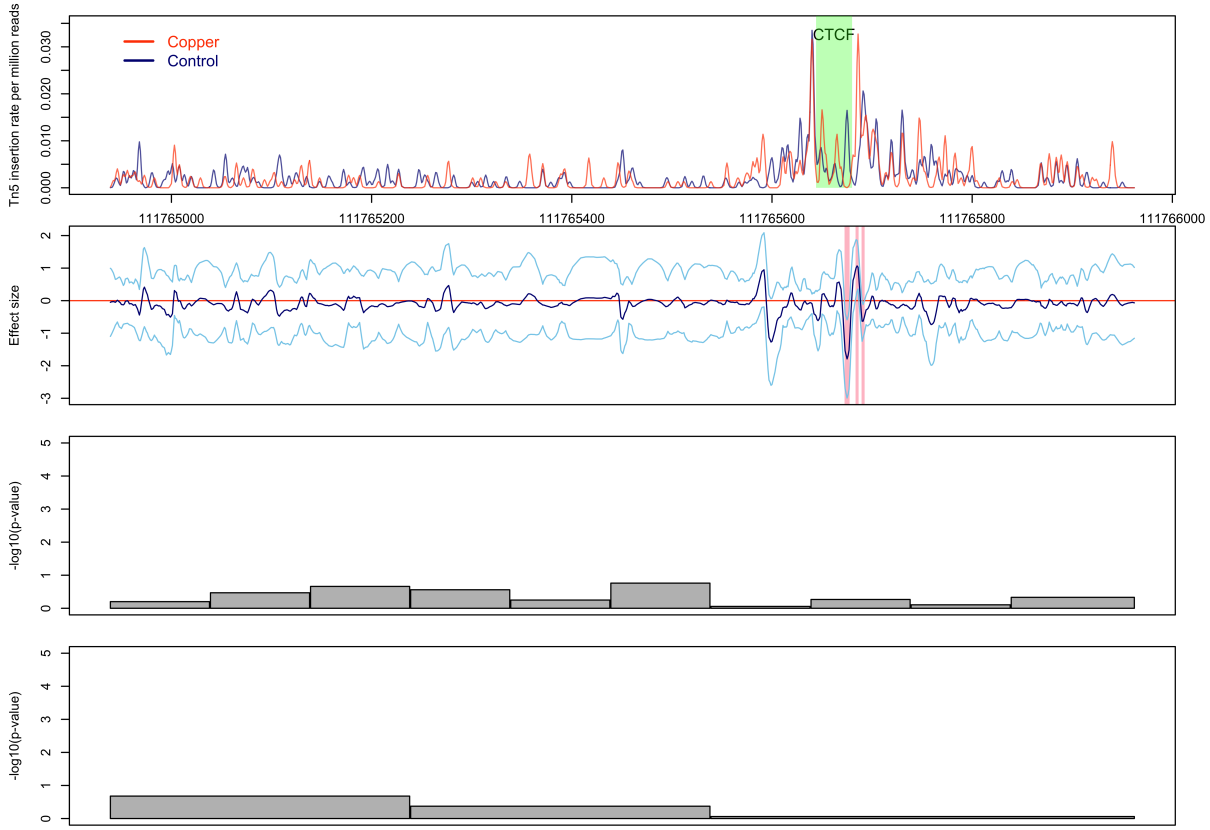

Figure 3: **Example of DERs identified by multiseq, but not by DESeq2 (chr1:111764939-111765962)** The top two panels are the same as Figure 4 of the main text. Labels and colors are as in Supplementary Material Fig 2. For multiseq, $\log \Lambda = 10.49$; p-value ≈ 0.000019. For DESeq2, p-values from the empirical distribution are ≈ 0.23 for 1024 bin, ≈ 0.14 for 300 bin, and ≈ 0.32 for 100 bin.

5 Modelling the effect of the covariate X on the overall expression

The effect of the covariate X on the overall expression (intensity) λ_{tot} ($:= \lambda_{1:B}$) can be modelled in multiple ways. Here, we illustrate two approaches that are currently provided in our software. Both approaches address the issue of different sequencing depths across samples. We use the index $s = 0$, $l = 1$ to indicate the model for the overall expression.

5.1 Binomial regression

Let y^i_{tot} represent the total read count over the region for sample i (i.e., $y^i_{\text{tot}} = y^i_{1:B}$). The high-throughput sequencing data can have different sequencing depths across samples. Thus, the observed difference in y_{tot} across samples can be due to the difference in the sequencing depth, rather than the difference in overall expression of the molecular phenotype. To take into account the difference in the sequencing depth, we consider the following binomial regression:

$$y^i_{tot}|s^i, X^i \sim \text{Bin}(s^i, p^i_{01}) \tag{1}$$

$$\log(p^i_{01}/(1-p^i_{01})) = \alpha^i_{01} = \mu_{01} + \beta_{01}X^i + u^i_{01}, \tag{2}$$

where u^i_{01} denotes a zero-mean individual-specific random effect to model both over-dispersion and biological variability among samples. Here, we introduce s^i to take into account sample-specific sequencing depths. One way of estimating it is to use the total number of reads mapped to the entire genome (see Robinson and Oshlack (2010); Anders and Huber (2010) for other ways). This binomial regression model is the same as the multi-scale models at other scales in equations (17) and (18) in the main text. Thus, we use the same priors and inference procedure, e.g., the normal approximation to likelihood.

5.2 Poisson regression

The multi-scale models for inhomogeneous Poisson processes assume $y^i_{tot} \sim \text{Pois}(\lambda^i_{\text{tot}})$; see equation (10) in the main text. A natural extension to model the effect of the covariate X on the overall expression is a Poisson regression:

$$y^i_{tot} \sim \text{Pois}(\lambda^i_{\text{tot}}), \tag{3}$$

$$\log(\lambda^i_{\text{tot}}) = \mu_{01} + \beta_{01}X^i + u^i_{01}, \tag{4}$$

where u^i_{01} denotes an individual-specific random effect. With a gamma distribution on $\exp(u^i_{01})$, it is the same as generalized linear models of negative binomial family with log-link which have been adapted by widely-used overall expression methods for differential analysis, such as DESeq2 (Love *et al.*, 2014) and edgeR (Robinson *et al.*, 2010). Those

methods have additional advancements such as 1) including *normalization constants* in their models for taking into account the difference in the sequencing depth and 2) using the shrinkage of estimated sample variances to pooled estimates for more stable inference at small sample sizes (Love *et al.*, 2014; Robinson *et al.*, 2010). Thus, it is desirable to incorporate outputs from the existing methods into multiseq for the overall expression. Indeed, our software provides an option to incorporate their outputs for the Poisson regression approach. For example, in the analysis of the paper, we run DESeq2 using total read counts over each region, obtain a p-value for each region from the output of DESeq2, and convert it to a likelihood ratio under the assumption that the p-value is computed using the likelihood ratio test with degrees of freedom = 1 (while DESeq2 uses the Wald test). Then, multiseq uses the resulting likelihood ratio in replacement of a likelihood ratio corresponding to $s = 0, l = 1$ in equation (20) in the main text. Moreover, multiseq can use the estimates from the existing methods to estimate the effect size on the overall expression, $\hat{\beta}_{01}$.

6 Reparameterization of μ and β to obtain (asymptotically) factorized likelihood

The likelihoods of μ_{sl} and β_{sl} under the generalized (binomial) linear mixed model are not factorized; see equations (17) and (18) in the main text. To ease computation, we reparameterize, following Wakefield (2009), from μ_{sl}, β_{sl} to μ^*_{sl}, β_{sl} whose likelihood (asymptotically) factorizes into two independent terms. With the factorized likelihoods, setting independent priors on μ^*_{sl} and β_{sl} allows for separate computations of the posteriors and marginal likelihoods (see the Supplementary Material section 7 for details).

For convenience, we drop the scale and location subscripts for the following description of the reparameterization. Let $\hat{\mu}, \hat{\beta}$, s_μ, s_β be the maximum likelihood estimators (MLEs) and their standard errors from the generalized linear mixed model. We then have (asymptotically)

$$\begin{pmatrix} \hat{\mu} \\ \hat{\beta} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \beta \end{pmatrix}, \begin{pmatrix} I_{00} & I_{01} \\ I_{01} & I_{11} \end{pmatrix}^{-1} \right) \tag{5}$$

where $\mathbf{I} = \begin{pmatrix} I_{00} & I_{01} \\ I_{01} & I_{11} \end{pmatrix}$ is the Fisher's information matrix for the MLEs. Hence

$$s_\mu = se(\hat{\mu}) = \sqrt{\frac{I_{11}}{I_{00}I_{11} - I_{01}^2}} \tag{6}$$

$$s_\beta = se(\hat{\beta}) = \sqrt{\frac{I_{00}}{I_{00}I_{11} - I_{01}^2}} \tag{7}$$

$$cov(\hat{\mu}, \hat{\beta}) = -\frac{I_{01}}{I_{00}I_{11} - I_{01}^2}. \tag{8}$$

Now, by defining

$$\mu^* = \mu + \frac{I_{01}}{I_{00}}\beta = \mu - \frac{cov(\hat{\mu}, \hat{\beta})}{s_\beta^2}\beta \tag{9}$$

$$\hat{\mu}^* = \hat{\mu} + \frac{I_{01}}{I_{00}}\hat{\beta} = \hat{\mu} - \frac{cov(\hat{\mu}, \hat{\beta})}{s_\beta^2}\hat{\beta}, \tag{10}$$

we have

$$cov(\hat{\mu}^*, \hat{\beta}) = 0. \tag{11}$$

Therefore, we can write $\mathsf{P}(\hat{\mu}^*, \hat{\beta}|\mu^*, \beta) = \mathsf{P}(\hat{\mu}^*|\mu^*) \times \mathsf{P}(\hat{\beta}|\beta)$ (at least asymptotically). We can then set independent priors for μ^* and β to ease computation, resulting in independent posteriors for the two parameters; see the Supplementary Material section 7 for details. The posterior for μ can be easily obtained by using

$$\mu = \mu^* - \frac{I_{01}}{I_{00}}\beta. \tag{12}$$

7 Approximate factorized likelihood, prior, posterior, and marginal likelihood for the reparameterized multi-scale parameters

Application of the reparameterization (Supplementary Material section 6) and the normal approximation to the likelihood (section 2.2 in the main text) leads to the approximate factorized likelihood:

$$L(\boldsymbol{\mu}^*, \boldsymbol{\beta}; Y) = \prod_{s,l} N(\hat{\mu}^*_{sl}; \mu^*_{sl}, s^2_{\mu^*_{sl}}) N(\hat{\beta}_{sl}; \beta_{sl}, s^2_{\beta_{sl}}), \tag{13}$$

where $N(\cdot; \mu, \sigma^2)$ denotes the density of a normal distribution with mean μ and variance σ^2, and $s_{\mu^*_{sl}}$ denotes the standard error of $\hat{\mu}^*_{sl}$. See Supplementary Material section 8 for the computation of $\hat{\mu}^*_{sl}$, $\hat{\beta}_{sl}$, $s_{\mu^*_{sl}}$, and $s_{\beta_{sl}}$.

We impose independent and identically distributed priors on μ^*_{sl} and β_{sl}, leading to the same form for the posteriors and marginal likelihoods for μ^*_{sl} and β_{sl}. Thus, in this section we describe prior specification, posterior computation, and hyper-parameters estimation for β_{sl}. A similar approach can be applied to μ^*_{sl}. Following Xing *et al.* (2021), we use the methods from Stephens (2016) for the prior, posterior, and hyper-parameters estimation. This section provide a brief summary of the methods; see Stephens (2016) for details.

Now we drop the scale and location subscripts for notational convenience.

7.1 Prior

Following Xing *et al.* (2021) in using the methods from Stephens (2016), we use a unimodal distribution g with a mode at 0 as a shrinkage prior for β. Specifically,

$$g(\cdot; \boldsymbol{\pi}) = \pi_0 \delta_0(\cdot) + \sum_{k=1}^{K} \pi_k N(\cdot; 0, \sigma_k^2), \tag{14}$$

where $\delta_0(\cdot)$ denotes a point mass at 0. Here we take $\sigma_1, \ldots, \sigma_K$ to be a large and dense grid of fixed positive numbers spanning a range from very small to very big. The mixture proportions $\boldsymbol{\pi} = (\pi_0, \ldots, \pi_K)$, which are non-negative and sum to one, are hyper-parameters to be estimated. See Supplementary Material section 7.3 for the empirical Bayes methods to estimate the hyper-parameters.

7.2 Posterior

As derived in S2 in Supplementary Material of Stephens (2016), the posterior distribution of β is a mixture of $K+1$ components:

$$\mathsf{P}(\beta|\hat{\beta}, s_\beta^2, \boldsymbol{\pi}) = w_0 \delta_0(\beta) + \sum_{k=1}^{K} w_k N(\beta; \frac{\hat{\beta}\sigma_k^2}{s_\beta^2 + \sigma_k^2}, \frac{s_\beta^2 \sigma_k^2}{s_\beta^2 + \sigma_k^2}). \tag{15}$$

The posterior weights w_k for $k = 0, \ldots, K$ are

$$w_k = \frac{\pi_k N(\hat{\beta}; 0, s_\beta^2 + \sigma_k^2)}{\sum_{k'=0}^{K} \pi_{k'} N(\hat{\beta}; 0, s_\beta^2 + \sigma_{k'}^2)}, \tag{16}$$

where $\sigma_0^2 = 0$.

7.3 Empirical Bayes method for estimation of the hyper-parameters $\boldsymbol{\pi}$

The hyper-parameters $\boldsymbol{\pi}$ can be estimated by maximizing the likelihood $L(\boldsymbol{\pi}) = \prod_{sl} \mathsf{P}(\hat{\beta}_{sl}|s^2_{\hat{\beta}_{sl}}, \boldsymbol{\pi})$, where $\mathsf{P}(\hat{\beta}_{sl}|s^2_{\hat{\beta}_{sl}}, \boldsymbol{\pi})$ is the marginal likelihood integrating out β_{sl}, leading to he maximum likelihood estimate $\hat{\boldsymbol{\pi}} := \arg\max L(\boldsymbol{\pi})$. This is easily computed using an EM algorithm; see Stephens (2016). Note that the scale and location subscripts have been reintroduced in the marginal likelihood. As derived in S2 in Supplementary Material of Stephens (2016), the marginal likelihood is computed by

$$\mathsf{P}(\hat{\beta}_{sl}|s^2_{\hat{\beta}_{sl}}, \boldsymbol{\pi}) = \pi_0 N(\hat{\beta}_{sl}; 0, s^2_{\hat{\beta}_{sl}}) + \sum_{k=1}^{K} \pi_k N(\hat{\beta}_{sl}; 0, s^2_{\hat{\beta}_{sl}} + \sigma^2_k). \tag{17}$$

In practice, to allow for different amounts of shrinkage across scales, we consider scale-specific mixing proportions $\hat{\boldsymbol{\pi}}^s$ (i.e., scale-specific prior g_s) and estimate them using $\hat{\beta}_{sl}$, $s^2_{\hat{\beta}_{sl}}$ at the same scale after the "translation-invariant" transform (Xing *et al.*, 2021).

8 Estimates and standard errors for the reparameterized multi-scale parameters

In this section, we describe how to find the estimates for μ^*_{sl} and β_{sl} ($\hat{\mu}^*_{sl}$, $\hat{\beta}_{sl}$), and their standard errors ($s_{\hat{\mu}^*_{sl}}$, and $s_{\hat{\beta}_{sl}}$) which are required to obtain the approximate factorized likelihood of the reparameterized multi-scale parameters in (13). For notational convenience, we drop the scale and location subscripts in this section.

From (10),

$$\hat{\mu}^* = \hat{\mu} + \frac{I_{01}}{I_{00}}\hat{\beta} = \hat{\mu} - \frac{cov(\hat{\mu}, \hat{\beta})}{s^2_{\hat{\beta}}}\hat{\beta}, \tag{18}$$

resulting in

$$s_{\hat{\mu}^*} = \sqrt{s^2_{\hat{\mu}} + \left[\frac{cov(\hat{\mu}, \hat{\beta})}{s^2_{\hat{\beta}}}\right]^2 s^2_{\hat{\beta}} - 2\frac{cov(\hat{\mu}, \hat{\beta})}{s^2_{\hat{\beta}}} cov(\hat{\mu}, \hat{\beta})} \tag{19}$$

$$= \sqrt{s^2_{\hat{\mu}} - \frac{cov(\hat{\mu}, \hat{\beta})^2}{s^2_{\hat{\beta}}}}. \tag{20}$$

Therefore, once we obtain the estimates for μ and β ($\hat{\mu}$, $\hat{\beta}$), and their standard errors ($s_{\hat{\mu}}$, $s_{\hat{\beta}}$) and covariance ($cov(\hat{\mu}, \hat{\beta})$), we can find $\hat{\mu}^*$, $\hat{\beta}$, $s_{\hat{\mu}^*}$, and $s_{\hat{\beta}}$.

Our software provides three approaches to estimating $\hat{\mu}$, $\hat{\beta}$, s_μ, s_β, and $cov(\hat{\mu}, \hat{\beta})$. The first approach is a standard approach for the generalized (binomial) linear mixed model (see (17) and (18) in the main text). Fitting a mixed effects model is in general much slower than a fixed effects model, and even more so for binomial regression. As such, we make an approximation by fitting a quasibinomial model instead (see McCullagh and Nelder (1989)), capturing the random effect by an overdispersion factor. This approach is implemented in the glm.fit() function in R. However, due to the multiple weighted least squares steps involved, this approach can be computationally expensive given the number of parameters that we have. Thus, our software also provides two alternative approaches: one when X is a two-group categorical variable and the other when X is a quantitative variable. The following sections describe the details of these two approaches.

8.1 Case 1: X is a two-group categorical variable

We derive analytic forms for $\hat{\mu}$, $\hat{\beta}$, s_μ, s_β, and $cov(\hat{\mu}, \hat{\beta})$ for this case. Specifically, assume that we have groups labelled 0 and 1. WLOG, set group 0 to be the baseline, so that $\mu \equiv \alpha^0$ and $\beta \equiv \alpha^1 - \alpha^0$. We first obtain estimates and standard errors for α^0 and α^1 by modifying the procedure outlined in Appendix B.1 of Xing *et al.* (2021) that assumes a group has a single sample. Our case assumes each group has multiple samples. By assuming that the samples are independent, it is straightforward to extend the procedure in Appendix B.1 of Xing *et al.* (2021) by pooling the data across all samples, specifically the success and failure counts across all the samples when we consider data in the multi-scale space. The only modification we add is to allow for the presence of a sample-specific random effect. Instead of modelling the random effect directly, we make use of the form of overdispersion as described in (4.22) and (4.23) of McCullagh and Nelder (1989), and inflate $Var(\hat{\alpha}^k)$ by the overdispersion parameter. This procedure is appealing because this form of overdispersion is likely to be computationally efficient, and avoids making explicit modelling assumptions (see McCullagh and Nelder (1989)).

Once we obtain the estimates and standard errors for α^0 and α^1, denoted by $\hat{\alpha}^0$, $\hat{\alpha}^1$, s_{α^0}, and s_{α^1}, respectively, we can compute

$$\hat{\mu} = \hat{\alpha}^0, \tag{21}$$

$$\hat{\beta} = \hat{\alpha}^1 - \hat{\alpha}^0, \tag{22}$$

$$s_\mu = s_{\alpha^0}, \tag{23}$$

$$s_\beta = \sqrt{s^2_{\alpha^1} + s^2_{\alpha^0}}, \tag{24}$$

$$cov(\hat{\mu}, \hat{\beta}) = -s^2_{\alpha^0}. \tag{25}$$

8.2 Case 2: X is a quantitative variable

For this case, we compute $\hat{\mu}$, $\hat{\beta}$, s_μ, s_β, and $cov(\hat{\mu}, \hat{\beta})$ by using an approximate model instead. Specifically, instead of fitting the generalized (binomial) linear mixed model in (17) and (18) of the main text, we approximate the likelihoods of α^i using normal likelihoods (for each i). To this end, we will need unbiased estimates for α^i. We use the estimate and standard error for α^i obtained by the procedure outlined in Appendix B.1 of Xing *et al.* (2021) as an (approximately) unbiased estimate for α^i and its standard error. By using normal likelihoods instead of binomial likelihoods, the generalized (binomial) linear mixed model becomes

$$\hat{\alpha}^i = \mu + \beta X^i + u^i + \epsilon^i \tag{26}$$

where $\hat{\alpha}^i$ are considered as the observations for $i = 1, ..., n$, μ is the intercept, β is the coefficient for covariate X, u^i are iid $N(0, (\sigma^{(u)})^2)$ across i, and ϵ^i are independent $N(0, (\sigma^i)^2)$ across i, independent of u^i. Since we have standard errors for $\hat{\alpha}^i$, we can simply substitute σ^i by $\hat{\sigma}^i \equiv$ standard error of $\hat{\alpha}^i$. This model can then be reformulated as

$$\hat{\alpha}^i = \mu + \beta X^i + \tilde{\epsilon}^i \tag{27}$$

where $\hat{\alpha}^i$, μ and β are as above, and $\tilde{\epsilon}^i$ are independent $N(0, (\sigma^{(u)})^2 + (\hat{\sigma}^i)^2)$ across i. As such, we can fit (27) using a weighted least squares (WLS) approach with the form of heteroskedasticity known. The only unknown parameter in the weights is $\sigma^{(u)}$, which will have to be estimated in order to choose the right weights for (27).

The standard procedure for fitting a binomial regression model is to use an iterative weighted least squares approach. Although we have motivated (27) from anther point of view, fitting the model (27) using a standard WLS approach is actually (almost) equivalent to truncating the iterative procedure after the first step in the case of a binomial regression model. Thus, this presents another link between the two models, the generalized (binomial) linear mixed model and (27), and provides good justification for approximating the former model using the latter.

Although the alternative model (27) was motivated by the goal of computational efficiency, we note that it has another conceptual advantage over the generalized (binomial) linear mixed model, in the case where sequencing depths might be different for each sample. In our quasibinomial implementation, the success (and failure) counts for samples in the same group will be pooled together (since each Bernoulli count is assumed to be i.i.d.). This could then lead to the information from that sample overwhelming the information from all the rest of the samples, which might bias the parameter estimates if that sample happens to be different from the rest. The WLS approach (27) weighs each sample according to its associated uncertainty instead, which is a better alternative.

To actually fit (27) and obtain estimates and standard errors for μ and β, we seek to

minimize the weighted sum of square differences:

$$\sum_i w_i^2(\hat{\alpha}^i - \mu - \beta X^i)^2 \tag{28}$$

For notational convenience, the index i for sample has been moved to be the subscript for the following derivations. Here, ideally it should be the case that $w_i = (\sigma_u^2 + \hat{\sigma}_i^2)^{-\frac{1}{2}}$. By taking partial derivatives w.r.t. μ and β and setting them to 0, we have:

$$\sum_i w_i^2(\hat{\alpha}_i - \mu - \beta X_i) = 0 \tag{29}$$

$$\sum_i w_i^2 X_i(\hat{\alpha}_i - \mu - \beta X_i) = 0 \tag{30}$$

After some algebra, we obtain the following WLS estimates:

$$\hat{\beta} = \hat{\beta}_{WLS} = \frac{\sum_i w_i^2 \hat{\alpha}_i(X_i - \bar{X}_W)}{\sum_i X_i w_i^2(X_i - \bar{X}_W)} \tag{31}$$

$$\hat{\mu} = \hat{\mu}_{WLS} = \frac{\sum_i w_i^2(\hat{\alpha}_i - \hat{\beta}_{WLS} X_i)}{\sum_i w_i^2}, \tag{32}$$

where $\bar{X}_W = \frac{\sum_i w_i^2 X_i}{\sum_i w_i^2}$. From these we can also derive the standard errors for the estimates:

$$s_\beta = se(\hat{\beta}_{WLS}) = \sqrt{\frac{Var(w_i\hat{\alpha}_i)}{\sum_i X_i w_i^2(X_i - \bar{X}_W)}} \tag{33}$$

$$s_\mu = se(\hat{\mu}_{WLS}) = \sqrt{\sum_i \left(\frac{w_i}{\sum_j w_j^2} - \frac{\bar{X}_W w_i(X_i - \bar{X}_W)}{\sum_j X_j w_j^2(X_j - \bar{X}_W)}\right)^2 Var(w_i\hat{\alpha}_i)} \tag{34}$$

and also

$$cov(\hat{\mu}, \hat{\beta}) = cov(\hat{\mu}_{WLS}, \hat{\beta}_{WLS}) = \frac{\sum_i w_i^2(X_i - \bar{X}_W)Var(w_i\hat{\alpha}_i)}{\sum_i w_i^2 \sum_i w_i^2(X_i - \bar{X}_W)^2} - \bar{X}_W se^2(\hat{\beta}_{WLS}). \tag{35}$$

Further define

$$RSE = \sqrt{\frac{1}{n-2}\sum_i w_i^2(\hat{\alpha}_i - \hat{\mu}_{WLS} - \hat{\beta}_{WLS} g_i)^2} \tag{36}$$

as the residual standard error. Note that if w_i was exactly $(\sigma_u^2 + \hat{\sigma}_i^2)^{-\frac{1}{2}}$, the value of RSE would be 1. However, σ_u^2 is unknown, and hence must be estimated in some way.

To estimate σ_u^2, we start off by fitting the model with no random effect (i.e. $\sigma_u^2 = 0$). Note that we can simply substitute $Var(w_i\hat{\alpha}_i)$ in (33)-(35) with its estimate, RSE^2, when

fitting the model. In this case, the value of RSE would be an indication of the presence of the random effect. *If* the additional sample specific random effect were absent, then the value of RSE would be close to 1. Assuming that our assumptions hold however, RSE would be different from 1, and would provide some information on the amount of "overdispersion" present. Hence, consider the "dispersion factor" defined by $\hat{\lambda} = \min\{1, RSE^2\}$. This is because $\sigma_u^2 + \hat{\sigma}_i^2 \geq \hat{\sigma}_i^2$, so we do not allow for "underdispersion".

To avoid heavy computation, we only seek to find a crude estimate for σ_u^2. For the model with no random effect, the "dispersion factor" $\hat{\lambda}$ is a multiplier on ${\hat{\sigma}_i}^2$, whereas the model in (27) captures over-dispersion using an additive effect. Nevertheless, the total amount of variation in the data is the same regardless of the model we fit, so we set

$$\sum_i (\hat{\sigma}_i^2 + \sigma_u^2) = \sum_i \hat{\lambda}\hat{\sigma}_i^2 \tag{37}$$

to solve for σ_u^2. This results in

$$\hat{\sigma}_u^2 = \frac{\hat{\lambda} - 1}{n} \sum_i \hat{\sigma}_i^2. \tag{38}$$

With the estimate in hand, we can then fit (27) with weights $w_i = (\hat{\sigma}_i^2 + \hat{\sigma}_u^2)^{-\frac{1}{2}}$. Provided that we have a moderate number of observations as well as a moderate number of binomial counts in each observation, the value of RSE we get from this fit should be close to 1, indicating that we have a decent fit. Given $\hat{\mu}$, $\hat{\beta}$, s_μ, s_β, and $cov(\hat{\mu}, \hat{\beta})$ in (31), (32), (33), (34), and (35), we can compute the approximate factorized likelihood of the reparameterized multi-scale parameters.

9 Effect size estimation

The above (approximated) factorized likelihoods and independent mixture priors for $\boldsymbol{\beta}$ yield independent posterior distributions on $\boldsymbol{\beta}$, each being a mixture of a point mass at zero and normal distributions in (15). To provide more interpretable estimates of the effect of X, we wish to convert posterior on $\boldsymbol{\beta}$ to posterior on $\boldsymbol{\beta}^o$ that represent the effect on the underlying log-intensity function $\log \boldsymbol{\lambda}$. The posterior on $\boldsymbol{\beta}^o$ does not have a simple analytic form. However, we can approximate the pointwise posterior mean and variance of $\boldsymbol{\beta}^o$ using the relationship between $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$ and Taylor approximations. Alternatively, posterior inference could be performed by sampling from the posterior for $\boldsymbol{\beta}^o$. It can be achieved by first simulating samples from the posteriors on $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$, and transforming them to posterior samples for $\boldsymbol{\beta}^o$ using the relationship with $\boldsymbol{\lambda}$. In the following sections, we first describe the Taylor approximation methods for two cases: one when X is a two-group categorical variable (Supplementary material section 9.1) and the other when X is a quantitative variable (Supplementary material section 9.2), and then describe the sampling based methods (Supplementary material section 9.3).

9.1 Taylor approximations for case 1: X is a two-group categorical variable

We assume that we have groups labelled 0 and 1. WLOG, set group 0 to be the baseline, so that $\boldsymbol{\mu} \equiv \boldsymbol{\alpha}^0$ and $\boldsymbol{\beta} \equiv \boldsymbol{\alpha}^1 - \boldsymbol{\alpha}^0$. We first describe a relationship between $\boldsymbol{\lambda}^i$ and $\boldsymbol{\alpha}^i$ for $i = 0, 1$ when random effects u in (16) of the main text are set to 0. Following elementary properties of the Poisson distribution, we can write λ_b^i for $b = 1, \ldots, B$ as a product of λ_{tot}^i, and the binomial success or failure probabilities p_{sl}^i, $q_{sl}^i := 1 - p_{sl}^i$, where

$$p_{sl}^i = \frac{e^{\alpha_{sl}^i}}{1 + e^{\alpha_{sl}^i}}, \quad q_{sl}^i = \frac{1}{1 + e^{\alpha_{sl}^i}}. \tag{39}$$

For example, the intensity at the two leftmost positions (i.e., $b = 1, 2$) can be written as

$$\lambda_1^i = \lambda_{\text{tot}}^i \big[\prod_{s=1}^{J-1} p_{s1}^i \big] p_{J1}^i, \quad \lambda_2^i = \lambda_{\text{tot}}^i \big[\prod_{s=1}^{J-1} p_{s1}^i \big] q_{J1}^i, \tag{40}$$

and the intensity at the two rightmost positions (i.e., $b = B - 1, B$) are

$$\lambda_{B-1}^i = \lambda_{\text{tot}}^i \big[\prod_{s=1}^{J-1} q_{s2^{s-1}}^i \big] p_{J2^{J-1}}^i, \quad \lambda_B^i = \lambda_{\text{tot}}^i \big[\prod_{s=1}^{J-1} q_{s2^{s-1}}^i \big] q_{J2^{J-1}}^i, \tag{41}$$

where p_{sl}^i and q_{sl}^i can be considered as the probabilities of assigning a read to the left half and right half of the area at the scale s and location l of a region. Let $\{c_1, ..., c_J\}$ be the binary representation of $b - 1$, and $d_m = \sum_{j=1}^m c_j 2^{m-j} + 1$ for $m = 1, ..., J - 1$. Then, for $b = 1, \ldots, B$, λ_b^i can be written as

$$\lambda_b^i = \lambda_{\text{tot}}^i [p_{11}^i]^{1-c_1} [p_{2,d_1}^i]^{1-c_2} \ldots [p_{J,d_{J-1}}^i]^{1-c_J} [q_{11}^i]^{c_1} [q_{2,d_1}^i]^{c_2} \ldots [q_{J,d_{J-1}}^i]^{c_J}. \tag{42}$$

Using this relationship, now we can represent $\beta_b^o = \log \lambda_b^1 - \log \lambda_b^0$ as a sum of $\log \left(\frac{\lambda_{\text{tot}}^1}{\lambda_{\text{tot}}^0} \right)$ and $\log \left(\frac{p_{sl}^1}{p_{sl}^0} \right)$ or $\log \left(\frac{q_{sl}^1}{q_{sl}^0} \right)$, where

$$\log \left(\frac{p_{sl}^1}{p_{sl}^0} \right) = \log \left(\frac{1 + e^{-\mu_{sl}}}{1 + e^{-(\mu_{sl} + \beta_{sl})}} \right) \tag{43}$$

$$= \log \left(\frac{1 + e^{-(\mu_{sl}^* + \gamma_{sl}\beta_{sl})}}{1 + e^{-(\mu_{sl}^* + \gamma_{sl}\beta_{sl} + \beta_{sl})}} \right), \tag{44}$$

$$\log \left(\frac{q_{sl}^1}{q_{sl}^0} \right) = \log \left(\frac{1 + e^{\mu_{sl}}}{1 + e^{\mu_{sl} + \beta_{sl}}} \right) \tag{45}$$

$$= \log \left(\frac{1 + e^{\mu_{sl}^* + \gamma_{sl}\beta_{sl}}}{1 + e^{\mu_{sl}^* + \gamma_{sl}\beta_{sl} + \beta_{sl}}} \right), \tag{46}$$

$$\gamma_{sl} = \frac{cov(\hat{\mu}_{sl}, \hat{\beta}_{sl})}{s_{\beta_{sl}}^2}. \tag{47}$$

Specifically,

$$\begin{aligned}\beta_b^o &= \log\left(\frac{\lambda_{\text{tot}}^1}{\lambda_{\text{tot}}^0}\right) + (1-c_1)\log\left(\frac{p_{11}^1}{p_{11}^0}\right) + (1-c_2)\log\left(\frac{p_{2,d_1}^1}{p_{2,d_1}^0}\right)\ldots + (1-c_J)\log\left(\frac{p_{J,d_{J-1}}^1}{p_{J,d_{J-1}}^0}\right) \\ &\quad + c_1\log\left(\frac{q_{11}^1}{q_{11}^0}\right) + c_2\log\left(\frac{q_{2,d_1}^1}{q_{2,d_1}^0}\right)\ldots + c_J\log\left(\frac{q_{J,d_{J-1}}^1}{q_{J,d_{J-1}}^0}\right). \qquad (48)\end{aligned}$$

For example, the effects on the log intensity at the two leftmost positions (i.e., $b = 1, 2$) can be represented as

$$\beta_1^o = \log\left(\frac{\lambda_{\text{tot}}^1}{\lambda_{\text{tot}}^0}\right) + \sum_{s=1}^{J-1}\left[\log\left(\frac{p_{s1}^1}{p_{s1}^0}\right)\right] + \log\left(\frac{p_{J1}^1}{p_{J1}^0}\right), \qquad (49)$$

$$\beta_2^o = \log\left(\frac{\lambda_{\text{tot}}^1}{\lambda_{\text{tot}}^0}\right) + \sum_{s=1}^{J-1}\left[\log\left(\frac{p_{s1}^1}{p_{s1}^0}\right)\right] + \log\left(\frac{q_{J1}^1}{q_{J1}^0}\right). \qquad (50)$$

The terms in (48) are independent, so once we find the posterior means and variances of $\log\left(\frac{\lambda_{\text{tot}}^1}{\lambda_{\text{tot}}^0}\right)$, $\log\left(\frac{p_{sl}^1}{p_{sl}^0}\right)$ and $\log\left(\frac{q_{sl}^1}{q_{sl}^0}\right)$, the posterior means and variances of β_b^o for $b = 1, \ldots, B$ can be easily computed.

Now we use Taylor approximations to approximate the posterior means and variances of $\log\left(\frac{p_{sl}^1}{p_{sl}^0}\right)$ and $\log\left(\frac{q_{sl}^1}{q_{sl}^0}\right)$ as follows. We define

$$f(x) = \log(1 + e^x) \qquad (51)$$

and consider the second order Taylor approximations to $f(x)$ around $-\mu^*$ and μ^*:

$$f(x) \approx f(-\mu^*) + f'(-\mu^*)(x + \mu^*) + \frac{f''(-\mu^*)}{2}(x + \mu^*)^2 \qquad (52)$$

$$f(x) \approx f(\mu^*) + f'(\mu^*)(x - \mu^*) + \frac{f''(\mu^*)}{2}(x - \mu^*)^2, \qquad (53)$$

where

$$f'(x) = \frac{e^x}{1 + e^x} \qquad (54)$$

$$f''(x) = \frac{e^x}{(1 + e^x)^2}. \qquad (55)$$

Replacing the dummy variable x above by the corresponding terms for the different cases, $-(\mu^* + \gamma\beta), -(\mu^* + (1 + \gamma)\beta), (\mu^* + \gamma\beta)$ or $(\mu + (1 + \gamma)\beta)$ gives us

$$\log\left(\frac{p_{sl}^1}{p_{sl}^0}\right) \approx f'(-\mu_{sl}^*)\beta_{sl} - \frac{f''(-\mu_{sl}^*)}{2}\beta_{sl}^2((1 + \gamma_{sl})^2 - \gamma_{sl}^2) \qquad (56)$$

$$\log\left(\frac{q_{sl}^1}{q_{sl}^0}\right) \approx -f'(\mu_{sl}^*)\beta_{sl} - \frac{f''(\mu_{sl}^*)}{2}\beta_{sl}^2((1 + \gamma_{sl})^2 - \gamma_{sl}^2). \qquad (57)$$

Using the independence of μ^* and β, we have

$$\begin{aligned}
E\left(\log\left(\frac{p^1_{sl}}{p^0_{sl}}\right)\right) \approx\ & E(f'(-\mu^*_{sl}))E(\beta_{sl}) \\
& -\frac{((1+\gamma_{sl})^2-\gamma^2_{sl})}{2}E(f''(-\mu^*_{sl}))E(\beta^2_{sl}) \qquad (58)\\
E\left(\log\left(\frac{q^1_{sl}}{q^0_{sl}}\right)\right) \approx\ & -E(f'(\mu^*_{sl}))E(\beta_{sl}) \\
& -\frac{((1+\gamma_{sl})^2-\gamma^2_{sl})}{2}E(f''(\mu^*_{sl}))E(\beta^2_{sl}) \qquad (59)\\
E\left(\log^2\left(\frac{p^1_{sl}}{p^0_{sl}}\right)\right) \approx\ & E(\{f'(-\mu^*_{sl})\}^2)E(\beta^2_{sl}) \\
& +\frac{((1+\gamma_{sl})^2-\gamma^2_{sl})^2}{4}E(\{f''(-\mu^*_{sl})\}^2)E(\beta^4_{sl}) \\
& -((1+\gamma_{sl})^2-\gamma^2_{sl})E(f'(-\mu^*_{sl})f''(-\mu^*_{sl}))E(\beta^3_{sl}) \qquad (60)\\
E\left(\log^2\left(\frac{q^1_{sl}}{q^0_{sl}}\right)\right) \approx\ & E(\{f'(\mu^*_{sl})\}^2)E(\beta^2_{sl}) \\
& +\frac{((1+\gamma_{sl})^2-\gamma^2_{sl})^2}{4}E(\{f''(\mu^*_{sl})\}^2)E(\beta^4_{sl}) \\
& +((1+\gamma_{sl})^2-\gamma^2_{sl})E(f'(\mu^*_{sl})f''(\mu^*_{sl}))E(\beta^3_{sl}). \qquad (61)
\end{aligned}$$

By noting that the posterior distributions for β_{sl} is a mixture of normal distributions (see (15)), we can easily compute $E(\beta^k_{sl}), k = 1, 2, 3, 4$. Specifically, suppose that

$$\beta \sim \sum_i \pi_i N(\theta_i, \sigma^2_i). \qquad (62)$$

Then

$$E(\beta^h) = \sum_i \pi_i E(N^h_i) \qquad (63)$$

for any nonnegative integer h, where $N_i \sim N(\theta_i, \sigma^2_i)$ and

$$\begin{aligned}
E(N^3_i) &= \theta^3_i + 3\theta_i\sigma^2_i \qquad (64)\\
E(N^4_i) &= \theta^4_i + 6\theta^2_i\sigma^2_i + 3\sigma^4_i. \qquad (65)
\end{aligned}$$

We can easily compute $E(\mu^*_{sl})$ and $E((\mu^*_{sl})^2)$ in a similar way. Then, we approximate the quantities $E(f'(-\mu^*_{sl}))$, $E(f'(\mu^*_{sl}))$, $E(f''(-\mu^*_{sl}))$, $E(f''(\mu^*_{sl}))$, $E(\{f'(-\mu^*_{sl})\}^2)$, $E(\{f'(\mu^*_{sl})\}^2)$, $E(\{f''(-\mu^*_{sl})\}^2)$, $E(\{f''(\mu^*_{sl})\}^2)$, $E(f'(-\mu^*_{sl})f''(-\mu^*_{sl}))$ and $E(f'(\mu^*_{sl})f''(\mu^*_{sl}))$ by using $E(\mu^*_{sl})$, $E((\mu^*_{sl})^2)$ and the following Taylor approximations. We first consider a random variable Z and an arbitrary real-valued function $g()$. To compute $E(g(Z))$, note that

$$g(Z) \approx g(E(Z)) + g'(E(Z))(Z - E(Z)) + \frac{g''(E(Z))}{2}(Z - E(Z))^2 \qquad (66)$$

so that

$$E(g(Z)) \approx g(E(Z)) + \frac{g''(E(Z))}{2} Var(Z). \tag{67}$$

Combining all these allows us to (approximately) compute the quantities in (58)-(61), and hence the posterior means and variances of $\log\left(\frac{p^1_{sl}}{p^0_{sl}}\right)$ and $\log\left(\frac{q^1_{sl}}{q^0_{sl}}\right)$.

The posterior mean and variance of $\log\left(\frac{\lambda^1_{\text{tot}}}{\lambda^0_{\text{tot}}}\right)$ depend on the choice of models for the effect on the overall expression (see Supplementary Material section 5). For example, $\log\left(\frac{\lambda^1_{\text{tot}}}{\lambda^0_{\text{tot}}}\right) = \log\left(\frac{p^1_{01}}{p^0_{01}}\right)$ for the binomial regression, and the posterior mean and variance of $\log\left(\frac{p^1_{01}}{p^0_{01}}\right)$ can be approximated by the methods described above. For the Poisson regression, $\log\left(\frac{\lambda^1_{\text{tot}}}{\lambda^0_{\text{tot}}}\right) = \beta_{01}$, and the mean and variance of β_{01} can be obtained by 1) fitting the Poisson regression with random effects in (3) and (4) or 2) using the output from existing methods.

9.2 Taylor approximations for case 2: X is a quantitative variable

Let X_0 denote the value of the covariate at the baseline, and define $X_1 := X_0 + 1$, which is a unit increase from X_0. Then, the effect size β^o_b can be written as

$$\beta^o_b = \log \lambda^1_b - \log \lambda^0_b, \tag{68}$$

where $\lambda^i_b, i = 0, 1$ corresponds to the intensity for X_i at position b. Again, this will involve a sum of $\log\left(\frac{p^1_{sl}}{p^0_{sl}}\right)$ and $\log\left(\frac{q^1_{sl}}{q^0_{sl}}\right)$ for the relevant s, l. Hence, we need to find the posterior means and variances of

$$\log\left(\frac{p^1_{sl}}{p^0_{sl}}\right) = \log\left(\frac{1+e^{-(\mu_{sl}+\beta_{sl}X_0)}}{1+e^{-(\mu_{sl}+\beta_{sl}X_1)}}\right) \tag{69}$$

$$= \log\left(\frac{1+e^{-(\mu_{sl}+\beta_{sl}X_0)}}{1+e^{-(\mu_{sl}+\beta_{sl}X_0+\beta_{sl})}}\right) \tag{70}$$

$$= \log\left(\frac{1+e^{-(\mu^*_{sl}+\gamma_{sl}\beta_{sl}+\beta_{sl}X_0)}}{1+e^{-(\mu^*_{sl}+\gamma_{sl}\beta_{sl}+\beta_{sl}X_0+\beta_{sl})}}\right) \tag{71}$$

$$\log\left(\frac{q^1_{sl}}{q^0_{sl}}\right) = \log\left(\frac{1+e^{(\mu_{sl}+\beta_{sl}X_0)}}{1+e^{(\mu_{sl}+\beta_{sl}X_1)}}\right) \tag{72}$$

$$= \log\left(\frac{1+e^{(\mu_{sl}+\beta_{sl}X_0)}}{1+e^{(\mu_{sl}+\beta_{sl}X_0+\beta_{sl})}}\right) \tag{73}$$

$$= \log\left(\frac{1+e^{(\mu^*_{sl}+\gamma_{sl}\beta_{sl}+\beta_{sl}X_0)}}{1+e^{(\mu^*_{sl}+\gamma_{sl}\beta_{sl}+\beta_{sl}X_0+\beta_{sl})}}\right) \tag{74}$$

The procedure for computing the posterior means and variances of these quantities is almost identical to that in Supplementary Material section 9.1, with only one exception:

the Taylor series expansions in this case involves only the first order terms, as the second order terms cannot be easily computed under the linearity assumption in our multiseq model (see (15), (16), (17) in the main text). Hence, the approximations are given by

$$f(x) \approx f(-\mu^*) + f'(-\mu^*)(x + \mu^*) \tag{75}$$
$$f(x) \approx f(\mu^*) + f'(\mu^*)(x - \mu^*) \tag{76}$$

where

$$f(x) = \log(1 + e^x), \quad f'(x) = \frac{e^x}{1 + e^x}. \tag{77}$$

Then, we obtain the approximations

$$\log\left(\frac{p^1_{sl}}{p^0_{sl}}\right) \approx f'(-\mu^*_{sl})\beta_{sl} \tag{78}$$
$$\log\left(\frac{q^1_{sl}}{q^0_{sl}}\right) \approx -f'(\mu^*_{sl})\beta_{sl} \tag{79}$$

and thus

$$E\left(\log\left(\frac{p_{sl}}{p_{sl}}\right)\right) \approx E(f'(-\mu^*_{sl}))E(\beta_{sl}) \tag{80}$$
$$E\left(\log\left(\frac{q_{sl}}{q_{sl}}\right)\right) \approx -E(f'(\mu^*_{sl}))E(\beta_{sl}) \tag{81}$$
$$E\left(\log^2\left(\frac{p_{sl}}{p_{sl}}\right)\right) \approx E(\{f'(-\mu^*_{sl})\}^2)E(\beta^2_{sl}) \tag{82}$$
$$E\left(\log^2\left(\frac{q_{sl}}{q_{sl}}\right)\right) \approx E(\{f'(\mu^*_{sl})\}^2)E(\beta^2_{sl}) \tag{83}$$

As with the case when X is a two-group categorical variable, all the necessary quantities in (80)-(83) can be approximated using (67), allowing us to compute the posterior means and variances of (71) and (74). The posterior mean and variance of $\log\left(\frac{\lambda^1_{\text{tot}}}{\lambda^0_{\text{tot}}}\right)$ can be computed by using the procedure in Supplementary Material section 9.1 with the modification described in this section. Then, computing posterior means and variances of β^o_b for $b = 1, \ldots, B$ is exactly the same as in Supplementary Material section 9.1.

9.3 Sampling based approach

To simulate samples from the posterior for $\boldsymbol{\beta}^o$, we first simulate samples from the posteriors on μ^*_{sl} and β_{sl} that have a mixture of normal distributions. Then, we can transform them to posterior samples for $\log\left(\frac{p^1_{sl}}{p^0_{sl}}\right)$ and $\log\left(\frac{q^1_{sl}}{q^0_{sl}}\right)$ using the relationships: (44) and (46) for a two-group categorical variable X, and (71) and (74) for a quantitative variable X (note that (71) and (74) include X_0, the value of the covariate at the baseline). Finally, using (48), we can obtain posterior samples for $\boldsymbol{\beta}^o$.

10 Simulation study

We based our simulations on the 544 dsQTLs (genetic variants associated with DNase-seq data) identified by either or both of WaveQTL or 100bp window approach at FDR=0.01 in Shim and Stephens (2015). Each dsQTL consists of DNase-seq count data over a region of length 1024bp and genotypes at the most strongly associated genetic variant on 70 individuals. When the most strongly associated genetic variants from WaveQTL and window-based approach were different, we used the one from the analysis with the stronger p-value. Note that the analysis in Shim and Stephens (2015) reports 578 such dsQTLs, but we used the 544 dsQTL for the simulations that have at least three heterozygotes. This restriction is required to obtain stable effect size estimates for simulations (see Supplementary Material section 10.2).

To simulate data sets with realistic effects, following Shim and Stephens (2015), we based on our simulations on estimated effects from the 544 dsQTLs. To avoid biasing the estimated effect sizes towards one approach or other we used another denoising method implemented in `bayesthr` function from the R package `wavethresh` to estimate the effect sizes for simulations (details in Supplementary Material section 10.2). Using the thinning procedure described in Supplementary Material section 10.1, we simulated two data sets for each dsQTL, each of which has two groups with an equal number of samples in each group. One of the two data sets has no difference (i.e., no effect) between the groups and the other has non-zero effect. We then applied each method (multiseq, WaveQTL, and DESeq2) to the 1088 data sets (544 with no effects and 544 with non-zero effects), obtained test statistic, and assessed the performance of each method based on area under an receiver operating characteristic curve (AUC). We performed this simulation for varying sample sizes (6, 10, and 70) and expected library read depths (39 M, 39×0.1 M, 39×0.5 M, 39×2 M, 39× 4 M). The main text presented results from comparison between WaveQTL and multiseq. Supplementary Material section 10.3 presents the details of DESeq2 application on the simulated data sets, and results from comparison of all methods, where multiseq consistently outperformed DESeq2

10.1 Simulating reads by thinning real data for each region

Here we describe a thinning procedure to simulate reads from DNase-seq data for 70 samples. Input to this procedure includes 1) the DNase-seq data for a given region of length 1024bp: c_b^j denotes read count at base b for sample j, $b = 1, \ldots, B = 1024$ and $j = 1, \ldots, 70$, 2) effect size for each base: $e_1, \ldots, e_B$, and 3) the number of samples to be simulated: n. Then, this procedure simulates a data set with two groups of samples (an equal number of samples in each group) where the ratio (on average) between the two groups is the effect size of $e_1, \ldots, e_B$. Note that $e_b = 1 \; \forall b$ results in no difference between the two groups.

Let X_b^i denote the number of reads in sample i at base b; we wish to simulate X_b^i for

$i = 1, \ldots, n$ and $b = 1, \ldots, B$. We perform this simulation by "thinning" the DNase-seq reads at base b, totaled across 70 samples ($c_b^+ := \sum_{j=1}^{70} c_b^j$) by a sample-specific and base-specific thinning probability p_b^i. That is,

$$X_b^i \sim \text{Binom}(c_b^+, p_b^i) \quad \text{for} \quad b = 1, \ldots, B, \quad \text{and} \quad i = 1, \ldots, n. \tag{84}$$

To introduce overdispersion and biological variability among samples into the simulation scheme, we further sample the thinning probability p_b^i from a Beta distribution with a group-specific mean, $p_b^{g_0}$ for group 0 and $p_b^{g_1}$ for group 1, and a shared dispersion parameter. Here, we used $p_b^{g_0} = \frac{2}{70(1+e_b)}$, and $p_b^{g_1} = \frac{2e_b}{70(1+e_b)}$. This ensures that (1) the ratio (on average) between the two groups is the effect size e_b and (2) $E[\sum_{i=1}^n X_b^i] = c_b^+ \sum_{i=1}^n E[p_b^i] = c_b^+ \frac{n}{70}$ $\forall b$ regardless of the effect size e_b.

Using this thinning procedure, we simulate two data sets for each region: $\mathbf{X^A}$ with the effect $e_1, \ldots, e_B$, and $\mathbf{X^N}$ with no effect ($e_b = 1$ $\forall b$) between the two groups. The latter point (2) above guarantees that the expected read counts of the two data sets are the same, i.e., $E[\mathbf{X^A}] = E[\mathbf{X^N}]$.

10.1.1 Changing expected library read depths

The DNase-seq data from Degner *et al.* (2012) provide a total of ≈ 2.75 B reads for the 70 samples corresponding to ≈ 39 M reads for each sample. The thinning procedure simulates a data set with $E[\sum_{i=1}^n X_b^i] = c_b^+ \frac{n}{70}$ $\forall b$, so each sample in the data set has the expected library read depth (on average) ≈ 39 M. We can change the expected library read depths in the simulations by rescaling c_b^+. Specifically, we modify (84) as follows:

$$X_b^i \sim \text{Binom}(\lceil S \times c_b^+ \rceil, p_b^i) \quad \text{for} \quad b = 1, \ldots, B, \quad \text{and} \quad i = 1, \ldots, n, \tag{85}$$

where $\lceil S \times c_b^+ \rceil$ represents the smallest integer greater than or equal to $S \times c_b^+$, and $0 < S < 1$ (or $1 < S$) decreases (or increases) the read depths.

10.2 Estimating effect size $e_1, \ldots, e_B$ from the real data for each region

We estimate the effect size $e_1, \ldots, e_B$ using the DNase-seq data for each region (of length 1024bp; $B = 1024$) and genotypes at the most strongly associated genetic variant as follows. To estimate $e_1, \ldots, e_B$ as the ratio between two group means, we use only two groups of samples (major homozygotes and heterozygotes). For each group of samples, we calculate the average read count across samples at each position b, and smooth the average read counts over the region using `bayesthr` function from the R package `wavethresh`, yielding the smoothed mean curve. Let s_b^M and s_b^H for $\forall b$ denote the smoothed mean curves for major homozygotes and heterozygote, respectively. We obtain the effect size

$e_1, \ldots, e_B$ for the thinning procedure using the ratio of the two curves. Specifically,

$$e_b = \frac{\max\left(\frac{1}{70}, s_b^H\right)}{\max\left(\frac{1}{70}, s_b^N\right)} \quad \text{for} \quad b = 1, \ldots, B. \tag{86}$$

Here, we use $\frac{1}{70}$ as a minimum value for mean curves to ensure that $0 < e_b < \infty$.

10.3 Comparison of multiseq, WaveQTL, and DESeq2

We consider three bin sizes, 100, 300, and 1024bp, for DESeq2 analysis. We applied DESeq2 to the simulated data sets as follows. For each bin size, we divide each 1024bp region into multiple bins. For each bin, we compute a p-value testing for difference between groups using DESeq2. Then, we use the minimum p-value across all bins within each region as the test statistic for that region, with a smaller p-value indicating more evidence against the null.

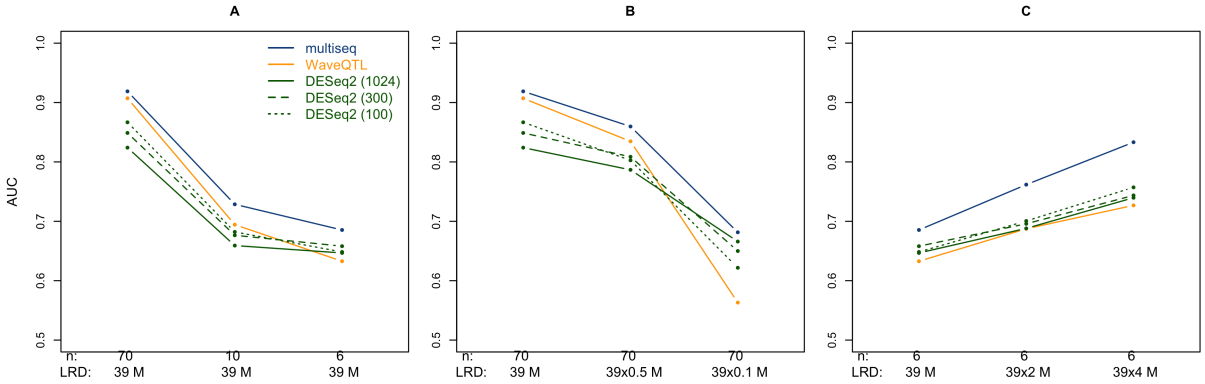

Figure 4: Labels and colors are as in Figure 1 of the main text. The AUC for DESeq2 with different bin sizes are colored in green.

Figure 4 shows results from DESeq2 analysis along with the results from multiseq and WaveQTL (see the main text). Compared to the overall expression method DESeq2, multiseq better exploits high-resolution information in the data, leading to the better performance for all cases, regardless of the choice for bin size in DESeq2. This observation is consistent to the results from the comparison between multi-scale methods and overall expression methods in Shim and Stephens (2015). See Shim and Stephens (2015) for more comprehensive study for the comparison. On the other hand, the performance of WaveQTL is worse than DESeq2 for the small sample size of 6 (even with the increased read depth 39×4 M in panel C) or for the read depth decreased up to 39×0.1 M (with the large sample size 70 in panel B). The reason is because WaveQTL is not suited to small sample sizes or low read counts in the first place.

11 Analysis of ATAC-seq data

11.1 Data

The lymphoblastoid cell line (LCL) GM18508 was purchased from Coriell Cell Repository. Cells were cultured in RPMI containing 5% charcoal-stripped FBS and treated for 6 hours with 6.00×10^{-5} M copper as described in Moyerbrailean *et al.* (2015). Cells were also cultured in parallel in the absence of any treatment (control 1) and with a vehicle control (control 2, ethanol 1/10,000 μL). We then followed the protocol by Buenrostro *et al.* (2013) to lyse 25,000-100,000 cells and prepare ATAC-seq libraries, with the exception that we used the Illumina Nextera Index Kit (Cat#15055290) in the PCR enrichment step. Individual libraries fragment distribution was assessed on the Agilent Bioanalyzer and pooling proportions were determined using the qPCR Kapa library quantification kit (KAPA Biosystems). Library pools were run on the Illumina NextSeq 500 Desktop sequencer in the Luca/Pique-Regi laboratory and on the Illumina HiSeq 2500 at the Michigan State University Genomics Core to obtain 38 bp paired-end reads. Libraries from three experimental replicates (9 samples total) were pooled and sequenced on multiple sequencing runs for a total of ∼233 M reads for the copper treatment, ∼203 M reads for the control 1 (media) and ∼191 M reads for the control 2 (ethanol).

11.2 Pre-processing the data

11.2.1 Combining paired-end reads to read counts for each base

The ATAC-seq data are paired-end reads. However, multiseq ignore the information on paired-end reads such as the length of DNA fragments, so we combined forward and reverse reads as follows. Each Tn5 transposase insertion generates two fragments that potentially produce two reads with different strands. For each base, we count the number of reads (regardless of their strands) that are produced by Tn5 transposase at that particular base. Specifically, we shift '+' strand reads 4 bp to the left and '−' strand reads 4 bp to the right, and then we count $5'$ ends of reads (regardless of their strands) for each base.

11.2.2 Removing reads that are potentially artifacts due to PCR amplification

For each base of a 1024bp region, we first identify samples whose reads are potentially artifacts due to PCR amplification as follows. Let y_b^i denote the read count for sample i at base b. If this count is much bigger than counts from nearby bases, specifically, in our ATAC-seq analysis

$$\frac{y_b^i}{\sum_{b'=b-50}^{b+50} y_{b'}^i} > 0.9, \tag{87}$$

we consider y_b^i reads from PCR amplification. If at least one sample has PCR artifacts reads at that base, we replace the read counts of all samples at that base with $\lceil \max(1, \tilde{y}_b) \rceil$, where $\tilde{y}_b$ represents the average read counts over samples without PCR artifacts reads, and $\lceil x \rceil$ represents the smallest integer greater than or equal to x. So read counts at this base won't affect different analysis. If all samples have PCR artifacts reads, we use $\tilde{y}_b = 1$.

11.2.3 Matching sequencing depths

Widely-used overall expression methods for differential analysis, such as DESeq2 (Love *et al.*, 2014) and edgeR (Robinson *et al.*, 2010), combine the information from multiple tests (e.g., multiple genes or regions) to obtain more stable estimates for small sample sizes. Thus, their software require data for multiple tests as input. In our application of DESeq2 to the ATAC-seq data, we merge the data for 242,714 tests (i.e., regions) from copper treatment vs control 1 and that from control 1 vs control 2 into one data for 485,428 tests. Then, we use the merged data as input to the software to ensure that parameters for null models are shared between the two data sets (indeed, when we run the software separately for the two data sets, we observed very poor performance of DESEq2 in our comparison). Note that the output from DESeq2 has been also incorporated into multiseq to measure the support for difference in overall expression. To merge the data sets, we have to match sequencing depths between copper treatment and control 2 samples as they are treated as the same samples in the merged data. Thus, if two matched samples (between copper and control 2) have different sequencing depths, we match their depths by downsampling the reads of the sample with higher depth. This can be achieved by sampling each read with sample-specific probabilities computed as follows. Let l_i^c and l_i^e for $i = 1, 2, 3$ represent the total number of reads mapped to the genome for copper and control 2 samples, respectively. The sampling probabilities are computed by

$$p_i^c = \frac{\min(l_i^c, l_i^e)}{l_i^c}, \quad p_i^e = \frac{\min(l_i^c, l_i^e)}{l_i^e} \quad \text{for } i = 1, 2, 3. \tag{88}$$

We used the total number of reads mapped to the genome as scaling factors. Robinson and Oshlack (2010); Anders and Huber (2010) introduce other quantities that can be used as scaling factors.

We performed the comparison in the main text after downsampling reads to match sequencing depths.

11.3 Selection of 242,714 regions: the top 5% of 1024 bp regions with the highest chromatin accessibility

We identify areas of the genome that have many ATAC-seq reads mapped to them because they are most likely to contain functional regulatory elements. Specifically, we divide

the whole genome into non-overlapping 300bp bins (Degner *et al.* (2012) observed that functional regulatory elements are typically contained over roughly 200-300bp), compute the number of reads mapped to each bin across six samples from copper treatment and control 1, and take the top 5% of bins ranked according to those numbers. We merge those bins if they are adjacent to each other and if the length of the merged bin is less than 1024bp, leading to 242,714 bins. The 242,714 1024bp regions used in our analysis are centered at those bins.

11.4 Comparison with overall expression methods DESeq2

We consider three bin sizes, 100, 300, and 1024bp, for DESeq2 analysis. For each bin size, we divide each 1024bp region into multiple bins. For each bin, we compute a p-value testing for difference between groups using DESeq2. Then, we use the minimum p-value across all bins within each region as the test statistic for that region, with a smaller p-value indicating more evidence against the null. We assessed the significance of the test statistic for each method using its empirical distribution under H_0. We constructed the empirical null distribution by applying the method to two controls (control 1 vs control 2) for the 242,714 regions and pooling the resulting test statistics. Each method yields a p-value testing H_0 for each region. Using these p-values, we used the qvalue package (Storey *et al.*, 2020) to estimate the False Discovery Rate (FDR) for each method at a given p-value threshold. We then compare the methods by the number of differentially expressed regions (DERs) detected at a given FDR (more DERs being better).

Figure 5 shows results from DESeq2 analysis along with the results from multiseq and WaveQTL (see the main text for details of multiseq and WaveQTL analysis). Compared to the overall expression methods DESeq2, multiseq better exploits high-resolution information in the data, leading to the increased power for most values of the FDR, regardless of the choice for bin size in DESeq2 (DESeq2 with bin size 1024bp shows slightly better power compared to multiseq at FDR $= 0.02 \sim 0.036$). It is consistent to the results from our simulation study in Supplementary Material section 10.3 and the comparison between multi-scale methods vs overall expression methods in Shim and Stephens (2015). Figure 5 also shows that the performance of DESeq2 varies with the choice of their bin size. See Shim and Stephens (2015) for more comprehensive study for the comparison between multi-scale methods vs overall expression methods.

On the other hand, the performance of WaveQTL is worse than DESeq2 with 1024bp bin size and multiseq (Figure 5). The reason is because WaveQTL is not suited to small sample size 6 in the first place.

References

Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome biology*, **11**, 1.

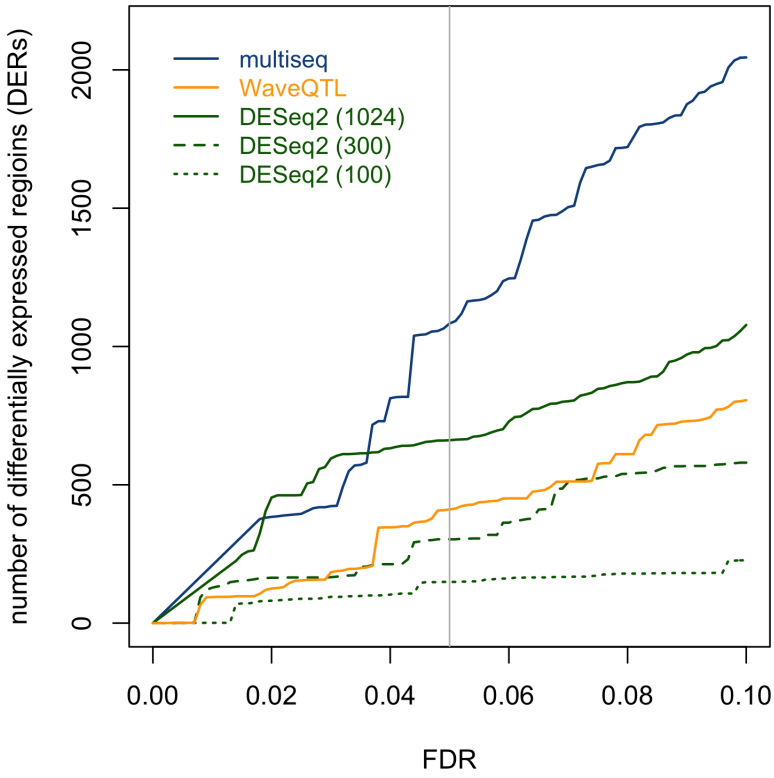

Figure 5: **Comparison of multiseq, WaveQTL, and DESeq2.** Each line shows the number of DERs identified by each method at a given FDR. Grey line indicates FDR = 0.05.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. and Greenleaf, W. J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nature methods*, **10**, 1213–1218.

Degner, J. F., Pai, A. a., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D. J., Pickrell, J. K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G. E., Stephens, M., Gilad, Y. and Pritchard, J. K. (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, **482**, 390–4.

Frazee, A. C., Sabunciyan, S., Hansen, K. D., Irizarry, R. A. and Leek, J. T. (2014) Differential expression analysis of rna-seq data at single-base resolution. *Biostatistics*, kxt053.

Lee, W. and Morris, J. S. (2016) Identification of differentially methylated loci using wavelet-based functional mixed models. *Bioinformatics*, **32**, 664–672.

Love, M. I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, **15**, 1–21.

McCullagh, P. and Nelder, J. (1989) *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall.

Moyerbrailean, G. A., Davis, G. O., Harvey, C. T., Watza, D., Wen, X., Pique-Regi, R. and Luca, F. (2015) A high-throughput rna-seq approach to profile transcriptional responses. *Scientific reports*, **5**.

Pickrell, J. K., Marioni, J. C., Pai, A. a., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y. and Pritchard, J. K. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, **464**, 768–72.

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010) edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Robinson, M. D. and Oshlack, A. (2010) A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, **11**, 1–9.

Shim, H. and Stephens, M. (2015) Wavelet-based genetic association analysis of functional phenotypes arising from high-throughput sequencing assays. *Ann. Appl. Stat.*, **9**, 665–686.

Stephens, M. (2016) False discovery rates: a new deal. *Biostatistics*, **18**, 275–294.

Storey, J. D., Bass, A. J., Dabney, A. and Robinson, D. (2020) *qvalue: Q-value estimation for false discovery rate control*. URL `http://github.com/jdstorey/qvalue`. R package version 2.20.0.

Wakefield, J. (2009) Bayes factors for genome-wide association studies: comparison with p-values. *Genetic epidemiology*, **33**, 79–86.

Xing, Z., Carbonetto, P. and Stephens, M. (2021) Flexible signal denoising via flexible empirical bayes shrinkage. *Journal of Machine Learning Research*, **22**, 1–28.