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

Heejung Shim[1], Zhengrong Xing[2], Ester Pantaleo[2], Francesca Luca[3], Roger Pique-Regi[4] and Matthew Stephens[5]

[1] *School of Mathematics and Statistics and Melbourne Integrative Genomics, University of Melbourne, Melbourne, 3010, Australia, e-mail:* heejung.shim@unimelb.edu.au

[2] *Department of Statistics, University of Chicago, Chicago, IL 60637, USA, e-mail:* [*]xiamanoobix@gmail.com; [**]esterpantaleo@gmail.com

[3] *Department of Obstetrics and Gynecology and Center for Molecular Medicine and Genetics, Wayne State University, Detroit, MI 48201, USA, e-mail:* fluca@wayne.edu

[4] *Center for Molecular Medicine and Genetics, Wayne State University, Detroit, MI 48201, USA, e-mail:* rpique@wayne.edu

[5] *Department of Statistics and Human Genetics, University of Chicago, Chicago, IL 60637, USA, e-mail:* mstephens@uchicago.edu

Abstract: Estimating and testing for differences in molecular phenotypes (e.g. gene expression, chromatin accessibility, transcription factor binding) across conditions is an important part of understanding the molecular basis of gene regulation. These phenotypes are commonly measured using high-throughput sequencing assays (e.g., RNA-seq, ATAC-seq, ChIP-seq), which provide high-resolution count data that reflect how the phenotypes vary along the genome. Multiple methods have been proposed to help exploit these high-resolution measurements for differential expression analysis. However, they ignore the count nature of the data, instead using normal approximations that work well only for data with large sample sizes or high counts. Here we develop count-based methods to address this problem. We model the data for each sample using an inhomogeneous Poisson process with spatially structured underlying intensity function, and then, building on multi-scale models for the Poisson process, estimate and test for differences in the underlying intensity function across samples (or groups of samples). Using both simulation and real ATAC-seq data we show that our method outperforms previous normal-based methods, especially in situations with small sample sizes or low counts.

Keywords and phrases: Multiscale Poisson processes, Wavelets, Differential expression analysis, High-throughput sequencing assays, Multiscale, Bayesian inference, Count data, RNA-seq, ATAC-seq, Chromatin accessibility.

1. Introduction

To understand the molecular basis of gene regulation, scientists often study the way that molecular phenotypes—such as gene expression (Marioni et al., 2008; Moyerbrailean et al., 2015), chromatin accessibility (Buenrostro et al., 2013), and transcription factor binding (Luca et al., 2013; Buenrostro et al., 2013) —vary among samples or treatment conditions (e.g., different environments, developmental stages, or tissues). These phenotypes are usually measured using sequencing-based assays—such as RNA-seq (Mortazavi et al., 2008; Wang et al., 2008; Marioni et al., 2008), ChIP-seq (Johnson et al., 2007; Barski et al., 2007; Mikkelsen et al., 2007), DNase-seq (Boyle et al., 2008; Hesselberth et al., 2009), and ATAC-seq (Buenrostro et al., 2013)—which are cheap and high throughput. Collectively we refer to these assays as $*$-seq. These $*$-seq assays provide high-resolution measurements across the whole genome that reflect how the molecular phenotypes vary along the genome. Specifically, for each location in the genome, these assays provide a "count" of the number of sequences that arose from that location, where the size of the count reflects the intensity of the underlying phenotype at that location. In this paper we develop methods to detect and estimate *differences* in this intensity among samples along the genome, taking account of both the high-resolution and the count nature of the data.

The simplest approach to this problem is to divide the genome into regions, add up the counts in each sample in each region, and then test for differences in these total counts using one of a wide range of analysis methods available for this task (e.g., DESeq2 (Love, Huber and Anders, 2014), edgeR (Robinson, McCarthy and Smyth, 2010), limma+voom (Law et al., 2014)). We refer to these methods as "overall expression methods" because their most common usage is in gene expression studies, where the regions usually correspond to genes, and the methods aim to detect differences in the overall expression of each gene. The main limitation of this approach is the difficulty of selecting an appropriate region size: if regions are too small then the methods will have no power due to low counts; if regions are too big then one loses the sensitivity of the inference, and indeed one risks missing

signals that affect smaller subregions. In other words these approaches, while simple, do not fully exploit the high-resolution measurements in ∗-seq data.

To address this problem several authors have developed methods that aim to make better use of the high-resolution measurements (Shim and Stephens, 2015; Lee and Morris, 2016; Frazee et al., 2014; Collado-Torres et al., 2017). Shim and Stephens (2015) and Lee and Morris (2016) proposed wavelet-based methods which test for differences at multiple resolutions simultaneously, combining information across resolutions to avoid the problems of selecting a single resolution or region size. Frazee et al. (2014) and Collado-Torres et al. (2017) presented approaches which compute base-resolution test statistics along the genome, identify regions of consecutive bases showing a common differential expression signature, and test for each region by combining information across bases in the region. The main limitation of these methods is that they use a normal distribution to model the read counts, which can work well if counts are sufficiently high, or sample sizes are sufficiently large, but performs poorly for small sample sizes or low counts (Shim and Stephens, 2015). Finally, Ma and Soriano (2018) introduced a multi-scale method for analysis of distributional variance which uses hierarchical non-parametric models. However, this method focuses on modelling differences in relative frequency at multiple resolutions, ignoring differences in overall expression in regions, so it is not directly applicable to typical differential expression analysis.

In this paper, we introduce a method, multiseq, that 1) better exploits high-resolution measurements in the ∗-seq data, as well as 2) directly models the count nature of the data (see Supplementary Material Table 1). Specifically, we assume that the count data for each sample are generated by an inhomogeneous Poisson process with a spatially structured underlying intensity function, and estimate this intensity function using extensions of existing multi-scale models for inhomogeneous Poisson processes (Kolaczyk, 1999; Timmermann and Nowak, 1999; Xing, Carbonetto and Stephens, 2021). A key innovation of our method is that it estimates and tests for the *differences* in the underlying intensity among samples, taking account of the fact that these difference will be spatially structured. We illustrate the benefits of multiseq— particularly for small sample sizes or low counts—using both simulation and real data analyses. We apply mutliseq to ATAC-seq data of 3 copper treated and 3 control samples for a large-scale differential chromatin accessibility analysis involving hundreds of thousands of tests. We find that multiseq identifies 1083 differences at the false discovery rate

0.05, which is 2.6 and 1.6 times the number of differences identified by a normal-based method and a simpler overall expression method, respectively. Our methods are implemented in the R package multiseq, available at https://github.com/heejungshim/multiseq.

2. Materials and methods

Assume that we have $*$-seq data from n samples across a genomic region divided into B equal bins. Specifically, the data consist of the "read counts" in each sample that map to each bin. Here, "map to a bin" means that the *start* of the read falls in the bin, so each read maps to exactly one bin. (The choice of bin-width constitutes a trade-off between computation and resolution. For example, the highest possible resolution of analysis would be at the single base level: each bin would be of length 1bp. However, computation could be reduced – typically by a factor of >10 – by using bins of length 10bp.) Let y_b^i denote the read count for sample i in bin b, and let $\boldsymbol{y}^i$ denote the vector $(y_1^i, \ldots, y_B^i)$.

We model the read counts as arising from an inhomogeneous Poisson process:

$$y_b^i \sim \text{Pois}(\lambda_b^i). \tag{1}$$

At a high level our goal is to estimate λ_b^i, and in particular to identify which bins b show differences in λ_b^i among samples i (or between groups of samples). Our methods are designed to address the following challenge. The information available for each bin b is very limited. For example, in the extreme case where each bin contains a single base, the typical number of reads in each bin could be 0 or 1. In this case, analyzing the data bin-by-bin will be hopeless. One solution would be to make the bins big enough to make such bin-by-bin analysis tractable. However, if the bins are made too big, this approach risks missing fine-resolution signals. To avoid this, we instead take a "multi-scale" approach to combine information across bins, exploiting the idea that nearby bins will often (though not always) tend to have similar effects, to smooth estimates across bins.

To implement this in practice we generalize previous multi-scale smoothing methods for inhomogeneous Poisson processes (Kolaczyk, 1999; Timmermann and Nowak, 1999; Xing, Carbonetto and Stephens, 2021). These methods essentially fit (1) for a *single* sample i, combining information across b. We review these methods in the next subsection before describing our extension to multiple samples.

2.1. *Multi-scale models for inhomogeneous Poisson processes: single sample*

Consider the model (1) for a single sample $i = 1$. To lighten notation we drop the superscript i, and so the observed data are $\boldsymbol{y} = (y_1, \ldots, y_B)$ with

$$y_b \sim \text{Pois}(\lambda_b). \tag{2}$$

Consider estimating $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_B)$ under the assumption that $\boldsymbol{\lambda}$ is "spatially structured": that is, where $|\lambda_b - \lambda_{b+1}|$ is small for most (but not necessarily all) b. For simplicity we assume that B is a power of 2, so $B = 2^J$ for some J.

The multi-scale approach to this problem involves two steps:

1. Reparameterize the Poisson model (2) using a 1-1 "multi-scale" transformation (Kolaczyk, 1999; Timmermann and Nowak, 1999), $\boldsymbol{\lambda} = f_{\text{ms}}(\boldsymbol{\alpha}, \lambda_{\text{tot}})$ defined below. The key property of this reparameterization is that each element of $\boldsymbol{\alpha}$ captures the spatial variation in $\boldsymbol{\lambda}$ at a particular scale (resolution) and location. Consequently one can capture the idea that $\boldsymbol{\lambda}$ is spatially structured by modelling sparsity of $\boldsymbol{\alpha}$, which is relatively straightforward. This is similar to the key idea of wavelets (Donoho and Johnstone, 1995), which are widely used in this way for Gaussian processes.
2. Perform shrinkage-based estimation of $\boldsymbol{\alpha}$. Xing, Carbonetto and Stephens (2021) do this by making a normal approximation to the likelihood for $\boldsymbol{\alpha}$, and then using the Empirical Bayes (EB) shrinkage methods from Stephens (2016).

Here we briefly summarize each step; see Xing, Carbonetto and Stephens (2021) for further details.

At scale $s = 1, \ldots, \log_2(B)$ define 2^{s-1} "locations" l by dividing the indices $\{1, \ldots, B\}$ into 2^{s-1} equal groups of consecutive indices, and let I_{sl} denote the indices of location l at scale s so formed. For example, at scale 3 there are $2^2 = 4$ locations, and:

$$I_{31} = [1, B/4]; \tag{3}$$
$$I_{32} = [B/4 + 1, B/2]; \tag{4}$$
$$I_{33} = [B/2 + 1, 3B/4]; \tag{5}$$
$$I_{34} = [3B/4 + 1, B]. \tag{6}$$

where $[a, b]$ denotes the indices $(a, a+1, \ldots, b)$. Further let I^-_{sl}, I^+_{sl} respectively denote the first and second halves of the indices in I_{sl}. So, for example, $I^-_{31} = [1, B/8]$ and $I^+_{31} = [B/8 + 1, B/4]$.

Now let $\lambda^-_{sl}, \lambda^+_{sl}$ respectively denote the sums of the values of λ across indices I^-_{sl} and I^+_{sl}. So,

$$\lambda^-_{sl} := \sum_{i \in I^-_{sl}} \lambda_i; \qquad \lambda^+_{sl} := \sum_{i \in I^+_{sl}} \lambda_i. \tag{7}$$

Finally define the multi-scale parameters $\boldsymbol{\alpha}$ by

$$\alpha_{sl} := \log[\lambda^-_{sl}/\lambda^+_{sl}]. \tag{8}$$

The intuition is that α_{sl} captures the change in (log) intensity between the first and second half of location I_{sl}. In particular, if λ is constant in I_{sl} then $\alpha_{sl} = 0$. When combined with the total intensity, $\lambda_{\text{tot}} := \sum_{i=1}^{b} \lambda_i$, the $\boldsymbol{\alpha}$ represent a 1-1 "multi-scale" reparameterization of $\boldsymbol{\lambda}$, and we write $\boldsymbol{\lambda} = f_{\text{ms}}(\boldsymbol{\alpha}, \lambda_{\text{tot}})$.

A key property of this reparameterization is that the likelihood factorizes into independent terms (Kolaczyk, 1999). Specifically:

$$p(\boldsymbol{y}; \boldsymbol{\alpha}, \lambda_{\text{tot}}) = \text{Pois}(\sum_b y_b; \lambda_{\text{tot}}) \prod_{sl} \text{Bin}\left(y^-_{sl}; y^-_{sl} + y^+_{sl}, \exp(\alpha_{sl})/(1 + \exp(\alpha_{sl}))\right) \tag{9}$$

where $\text{Pois}(\cdot; \lambda)$ denotes the probability mass function (pmf) of the Poisson distribution with parameter λ, $\text{Bin}(\cdot; n, p)$ denotes the pmf of the Binomial distribution with n trials and success probability p, and y^-_{sl}, y^+_{sl} denote, respectively, the sum of y_b over the indices I^-_{sl}, I^+_{sl}. This result follows from the elementary distributional result: if y_1, y_2 are independent, with $y_j \sim \text{Pois}(\lambda_j)$ then

$$p(y_1, y_2 | \lambda_1, \lambda_2) = \text{Pois}(y_1 + y_2; \lambda_1 + \lambda_2)\text{Bin}(y_1; y_1 + y_2, \lambda_1/(\lambda_1 + \lambda_2). \tag{10}$$

From (9) we note that the information in the data about α_{sl} is exactly that contained in a single binomial observation, which can be written

$$y^-_{sl} \sim \text{Bin}(y^-_{sl} + y^+_{sl}, p_{sl}) \tag{11}$$

$$\alpha_{sl} = \log(p_{sl}/(1 - p_{sl})). \tag{12}$$

With this reparameterization in place, the second step of the multi-scale approach is to perform shrinkage estimation of $\boldsymbol{\alpha}$. Xing, Carbonetto and

Stephens (2021) perform this shrinkage estimation by introducing normal approximations to the binomial likelihoods in (9), giving an (approximate) likelihood $L(\boldsymbol{\alpha}; \boldsymbol{y})$ of the form:

$$L(\boldsymbol{\alpha}; \boldsymbol{y}) = \prod_{sl} N(\hat{\alpha}_{sl}; \alpha_{sl}, s_{sl}^2), \tag{13}$$

where $N(\cdot; \mu, \sigma^2)$ denotes the density of a normal distribution with mean μ and variance σ^2, and $\hat{\alpha}_{sl}, s_{sl}^2$ are respectively estimates for α_{sl} and its standard error based on a binomial observation. This allows them to exploit the flexible EB shrinkage methods from Stephens (2016) to perform the shrinkage estimation. Specifically these methods perform shrinkage by estimating a prior distribution, g, for the elements of $\boldsymbol{\alpha}$ of the form:

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

where $\delta_0(\cdot)$ denotes a point mass at 0 and $\sigma_1, \ldots, \sigma_K$ dente 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 estimated by maximum likelihood. The posterior on the elements of $\boldsymbol{\alpha}$ given this estimated prior then has a closed form. See Stephens (2016) for details.

Xing, Carbonetto and Stephens (2021) also describe other details to improve performance. In particular, to allow for different amounts of shrinkage at different scales, they use scale-specific mixing proportions $\hat{\boldsymbol{\pi}}^s$ in (14), and estimate these using a "translation-invariant" transform (Coifman and Donoho, 1995) that efficiently averages results over all translations of the data. This averaging across translations also helps avoid artifactually large changes in estimated $\boldsymbol{\lambda}$ that can otherwise occur (for example) exactly halfway across the region.

2.2. *multiseq: multi-scale models for inhomogeneous Poisson processes from multiple groups of samples*

The main methodological contribution of our paper is to extend the ideas above to the multi-sample case. Specifically, we consider the model:

$$y_b^i | \boldsymbol{\alpha}^i, \lambda_{\text{tot}}^i \sim \text{Pois}(\lambda_b^i) \quad i = 1, \ldots, n, \tag{15}$$

where λ_b^i represents the b-th component of the multi-scale transformation $\boldsymbol{\lambda}^i = f_{\text{ms}}(\boldsymbol{\alpha}^i, \lambda_{\text{tot}}^i)$ described above.

We assume that a covariate X^i is measured on each sample i; for concreteness we assume $X^i \in \{0, 1\}$ is a binary group membership, but our methods apply equally to a continuous covariate. To model the effect of X^i on the intensity $\boldsymbol{\lambda}^i$, we introduce a linear model for the multi-scale parameters $\boldsymbol{\alpha}^i$:

$$\alpha_{sl}^i = \mu_{sl} + \beta_{sl} X^i + u_{sl}^i, \tag{16}$$

where μ_{sl} denotes the mean of α_{sl} for samples with $X^i = 0$, β_{sl} denotes the effect of X^i on α_{sl}, and u_{sl}^i denotes a zero-mean individual-specific random effect to model both over-dispersion and biological variability among samples.

Our goal now is to perform shrinkage-based estimation of the multi-scale parameters μ_{sl}, which captures the mean Poisson intensity across samples, and—more importantly for our purposes—β_{sl}, which captures the difference in intensity between groups. By performing shrinkage estimation we capture that both the underlying intensity and the difference in intensity between groups are expected to be spatially structured.

As in the single-sample case, the likelihood for the multi-scale parameters factorizes across s, l. In particular, analogous to (11), the information in μ_{sl}, β_{sl} is contained in n binomial observations:

$$y_{sl}^{i,-} \sim \text{Bin}(y_{sl}^{i,-} + y_{sl}^{i,+}, p_{sl}^i), \tag{17}$$

$$\log(p_{sl}^i/(1 - p_{sl}^i)) = \alpha_{sl}^i = \mu_{sl} + \beta_{sl} X^i + u_{sl}^i \qquad i = 1, \ldots, n. \tag{18}$$

This is a generalized (binomial) linear mixed model, and so we can use standard methods to obtain estimates $\hat{\mu}_{sl}, \hat{\beta}_{sl}$ and their corresponding standard errors $s_{\mu_{sl}}, s_{\beta_{sl}}$. Similar to Xing, Carbonetto and Stephens (2021) we then use these estimates and standard errors to form a normal approximation to the likelihood,

$$L(\boldsymbol{\mu}, \boldsymbol{\beta}; Y) = \prod_{sl} N(\hat{\mu}_{sl}; \mu_{sl}, s_{\mu_{sl}}^2) N(\hat{\beta}_{sl}; \beta_{sl}, s_{\beta_{sl}}^2), \tag{19}$$

and apply the methods from Stephens (2016) to produce shrinkage estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$.

This shrinkage estimation approach leads to smooth estimates of both the mean log-intensity in each group and the difference in log-intensity between the two groups. The methods also provide posterior distributions for $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$ and the marginal likelihood integrating out $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$. Details are given in Supplementary Material. In addition, Supplementary Material describes:

- A standard approach to fit the generalized linear mixed model provides estimates and standard errors for $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$, but we modified this procedure slightly to improve performance in practice.
- The likelihood for μ_{sl}, β_{sl} in (17)-(18) does not in general factorize into a term for μ_{sl} and a term for β_{sl}. Thus, to improve the approximation in (19), we reparameterize, following Wakefield (2009), from μ_{sl}, β_{sl} to μ^*_{sl}, β_{sl}, whose likelihoods asymptotically factorize into two independent terms.
- The effect of X on the overall expression (intensity) λ_{tot} can be modelled in multiple ways. Our software provides two approaches that are based on generalized linear models with random effects: Poisson regression and binomial regression. Both approaches address the issue of different sequencing depths across samples. The former has been adapted by widely-used overall expression methods for differential analysis (e.g., DESeq2 (Love, Huber and Anders, 2014) and edgeR (Robinson, McCarthy and Smyth, 2010)). Thus, our software provides an option to incorporate outputs from the existing methods for the Poisson regression approach. We used the output from DESeq2 in the analysis of this paper.

2.2.1. Testing for differences in molecular phenotype over the region between multiple groups of samples

To test for non-zero differences (i.e., effects) over the region, we test the null hypothesis $H_0 : \beta_{sl} = 0 \ \forall s, l$. This is equivalent to testing $\pi_0 = 1$ in the prior (14) for β, which, following Shim and Stephens (2015), we test using the likelihood ratio test statistic

$$\Lambda = \prod_{sl} \frac{\mathsf{P}(\hat{\beta}_{sl} | s^2_{\beta_{sl}}, \hat{\boldsymbol{\pi}})}{\mathsf{P}(\hat{\beta}_{sl} | s^2_{\beta_{sl}}, \pi_0 = 1)}, \tag{20}$$

where $\hat{\boldsymbol{\pi}}$ denotes the maximum likelihood estimate $\hat{\boldsymbol{\pi}} := \arg\max \prod_{sl} \mathsf{P}(\hat{\beta}_{sl} | s^2_{\beta_{sl}}, \boldsymbol{\pi})$; see Supplementary Material for details of the marginal likelihood $\mathsf{P}(\hat{\beta}_{sl} | s^2_{\beta_{sl}}, \boldsymbol{\pi})$. We assess the significance of the likelihood ratio test statistic using its empirical distribution under H_0. The empirical distribution can be obtained by permutation (i.e., permutation of sample labels for X; see Shim and Stephens (2015)) or by using a data set that is expected to have no difference in phenotype among samples (e.g., by comparing controls vs controls; see our analysis in section 3).

2.2.2. Effect size estimation

The factorized normal likelihood and mixture priors on $\boldsymbol{\beta}$ yield closed forms for the posterior distributions; see Supplementary Material. To provide more interpretable estimates of the effect of X, we wish to convert these posteriors in the multi-scale space to posteriors for the underlying log-intensity function $\log \boldsymbol{\lambda}$ in the original observation space. We define the effect on base b in the observation space as $\beta_b^o := \log(\lambda_b^{(1)}/\lambda_b^{(0)})$, where $\boldsymbol{\lambda}^{(0)}, \boldsymbol{\lambda}^{(1)}$ denote the values for $\boldsymbol{\lambda}$ for an individual in group 0 and 1 respectively, whose random effects u in (16) are set to 0. The posterior on $\boldsymbol{\beta}^o$ does not have a simple analytic form. However, we can approximate the pointwise posterior mean and variance using Taylor series approximations (see Supplementary Material). Other types of posterior inference could be performed by sampling from the posterior for $\boldsymbol{\beta}^o$, which 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 $\boldsymbol{\lambda}$ (see Supplementary Material).

3. Results

In this section, we assess the performance of multiseq against existing methods on both simulated and real data sets. We first illustrate the advantages of multiseq compared to WaveQTL (Shim and Stephens, 2015), which fails to directly model the count nature of $*$-seq data, in a simulation study. Then, we show the application of multiseq to real data with small sample size, and compare multiseq with WaveQTL and an overall expression method, DESeq2 (Love, Huber and Anders, 2014).

3.1. Simulation study

To demonstrate the benefits of multiseq in the analysis of data sets with small sample sizes or low read counts, we compared multiseq with WaveQTL on simulated data. (Supplementary Material presents results for the overall expression method DESeq2 (Love, Huber and Anders, 2014), which performs consistently less well than multiseq here because it does not exploit the high-resolution information in the data.)

Following Shim and Stephens (2015), we simulated realistic data by subsampling reads from a real DNase-seq data set from Degner et al. (2012) which measures chromatin accessibility (Boyle et al., 2008) along the genome.

These data consist of ≈ 2.75 B reads for 70 samples (≈ 39 M reads per sample). We performed simulations using the 544 regions of length 1024bp that were reported as associated with genetic variants in Shim and Stephens (2015). For each region, we simulated two data sets for varying sample sizes (6, 10, and 70) and expected library read depths (39×0.1 M, 39×0.5 M, 39 M, 39×2 M, 39× 4 M), each of which has two groups with an equal number of samples in each group. One of the two data sets is null (i.e., no effect) and the other is non-null, with the effect given by the estimated true effects from Shim and Stephens (2015). See Supplementary Material for the detailed simulation procedure. This procedure results in 1088 data sets (544 null and 544 non-null) for each sample size and expected library depth.

We applied multiseq and WaveQTL to each simulated data set, obtained the test statistic for each method (the likelihood ratio test statistic in (20) for multiseq and a similar likelihood ratio statistic for WaveQTL; see Shim and Stephens (2015) for details), and assessed the performance of each method based on area under the receiver operating characteristic curve (AUC).

3.1.1. multiseq has a potential to better maintain power at small sample sizes or for low read counts compared to WaveQTL

We first present results for varying sample sizes (6, 10, and 70) at a sinlge expected library read depth (39 M). Compared with WaveQTL, multiseq achieved higher AUC at all three sample sizes, with larger difference in performance at smaller sample sizes (Fig 1A).

Next we present results for varying library read depth (39 M to 39×0.5 M, 39×0.1 M) at fixed sample size (70). The results in Fig 1B show that multiseq achieved higher AUC than WaveQTL at all three library read depths. In particular, for the lowest expected read depth of 39×0.1 M, WaveQTL has almost no power to detect signals (AUC=0.56), while multiseq still shows a moderate power (AUC=0.68).

In summary, these simulations illustrate the potential of multiseq, which directly models the count nature of ∗-seq data, to better maintain power at small sample sizes or for low read counts compared to WaveQTL.

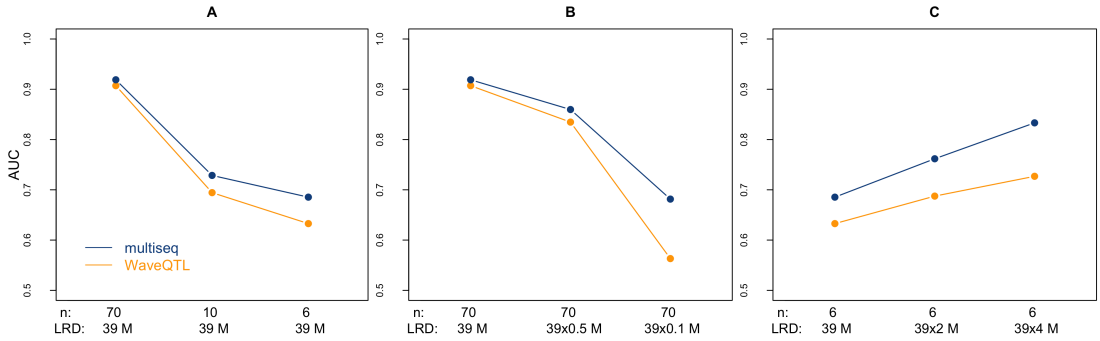

Fig 1: **multiseq has a potential to better maintain power at small sample sizes or for low read counts compared to WaveQTL.** Performance of multiseq (in blue) and WaveQTL (in orange) is assessed at different sample sizes (n) or expected library read depths (LRD) by using area under the receiver operating characteristic curve (AUC). Panel A shows the performance at the expected library read depth of 39 M as sample size varies over 70, 10, and 6. Panel B shows the performance at sample size 70 as the expected library read depth decreases from 39 M to 39×0.5 M, 39×0.1 M. Panel C shows the performance at the sample size 6 as the expected library read depth increases to 39×2 M and 39×4 M.

3.1.2. For small sample sizes, increasing library read depths leads to better performance of multiseq

Previous studies have shown that power of overall expression methods to detect effects in lowly-expressed regions can be increased by increasing library read depth (Tarazona et al., 2011; Robinson and Storey, 2014). Fig 1B shows that power of both multiseq and WaveQTL increases with library read depth at the relatively large sample size 70. Now we assess the performance improvement at the smaller sample size 6. The results (Fig 1C) confirm that increased read depths lead to better performance for multiseq. For example, the AUC for multiseq at the sample size 6 with 39×2 M and 39×4 M read depths (0.76 and 0.83, respectively) are higher than the AUC of multiseq at the larger sample size 10 with 39 M library read depth (0.72). However, consistent with previous studies of overall expression methods (Liu, Zhou and White, 2014; Busby et al., 2013), there is an upper bound in the performance improvement as read depth increases (Supplementary Material

Figure 1). While WaveQTL also shows improved performance for increased read depths, the rate of increase is relatively small compared to multiseq (Fig 1C).

3.2. ATAC-seq data analysis

3.2.1. Data

To understand regulatory mechanisms underlying cellular response to environmental perturbations, we collected ATAC-seq data to measure chromatin accessibility in samples treated with copper, as well as two types of controls (control 1 and control 2). We measured three samples for each condition, giving nine samples in all, and generated a total of $\approx$ 627 M 38bp paired-end reads. See Supplementary Material for a detailed description of the data and preprocessing steps. Our goal is to detect differences in chromatin accessibility in copper-treated samples vs control. Since no difference in chromatin accessibility is expected between the two controls, we use these to construct an empirical null distribution of the test statistic for each method, and then compare the copper-treated sample to one of the controls.

3.2.2. Analysis

Accessible chromatin regions, which tend to contain functional elements of the genome, have a median length of about 300bp (Degner et al., 2012). Thus, we analysed regions of length 1024bp, which are large enough to cover potential differences due to functional elements. (See Shim and Stephens (2015) for a discussion of robustness of multi-scale methods to choice of region size, and trade-offs between power and computation.) We focused our differential analysis on 242,714 regions that are the top 5% of regions with the highest chromatin accessibility (see Supplementary Material for details).

For each region, we applied multiseq, WaveQTL, and DESeq2 to the copper treated vs. control 1 samples, and computed the test statistic for each method for H_0 : no difference in chromatin accessibility between the conditions for the entire region. We used the likelihood ratio test statistic in (20) for multiseq and a similar likelihood ratio statistic for WaveQTL (see Shim and Stephens (2015) for details). We applied DESeq2 using the total read count over each region as input (bin size = 1024) and used a p-value for each region as a test statistic. We then assessed the significance of the test statistic for each

method using its empirical distribution under H_0, constructed by applying the method to the two controls (control 1 vs. control 2). By this procedure, we obtained for each method a p-value testing H_0 for each region which we converted to a q-value using the qvalue package (Storey et al., 2020). We compared the methods by the number of differentially expressed regions (DERs) detected at a given q-value threshold (more DERs being better).

3.2.3. multiseq outperforms WaveQTL

Fig 2A shows the number of DERs for each method as the False Discovery Rate (FDR) varies from 0.001 to 0.1. multiseq detects considerably more DERs than WaveQTL at all values of the FDR (the blue line for multiseq is above the orange line for WaveQTL). For example, at FDR = 0.05, multiseq identifies 1083 DERs, which is 2.64 times the 411 DERs identified by WaveQTL. In addition, multiseq identifies most DERs found by WaveQTL (Fig 2B).

3.2.4. The advantages of multiseq are greatest for low-count regions

We hypothesized that the advantages of multiseq will be greatest in regions of lower read counts. To assess this we compute the total read count across 6 samples for each region, and compared the distributions of these total read counts for two sets of regions: 1) 271 DERs detected by multiseq and WaveQTL (FDR = 0.05 for both methods, blue points in Fig 2B) and 2) 186 DERs detected by multiseq (FDR = 0.05), but not WaveQTL (FDR = 0.5, red points in Fig 2B). The results (Fig 2C) show that the DERs identified only by multiseq (set 2) have much lower read counts than those detected by both methods (median total read count: 1660 and 344 for the set 1 and set 2, respectively). This confirms that multiseq is better powered for regions with lower read counts compared to WaveQTL.

3.2.5. multiseq increases power using high-resolution information in the data, in addition to total read counts over the regions

A potential advantage of multiseq over overall expression methods (e.g., DESeq2 (Love, Huber and Anders, 2014), edgeR (Robinson, McCarthy and Smyth, 2010)) is that it can exploit the high-resolution information in the measurements. To assess the contribution of this feature to this analysis, we

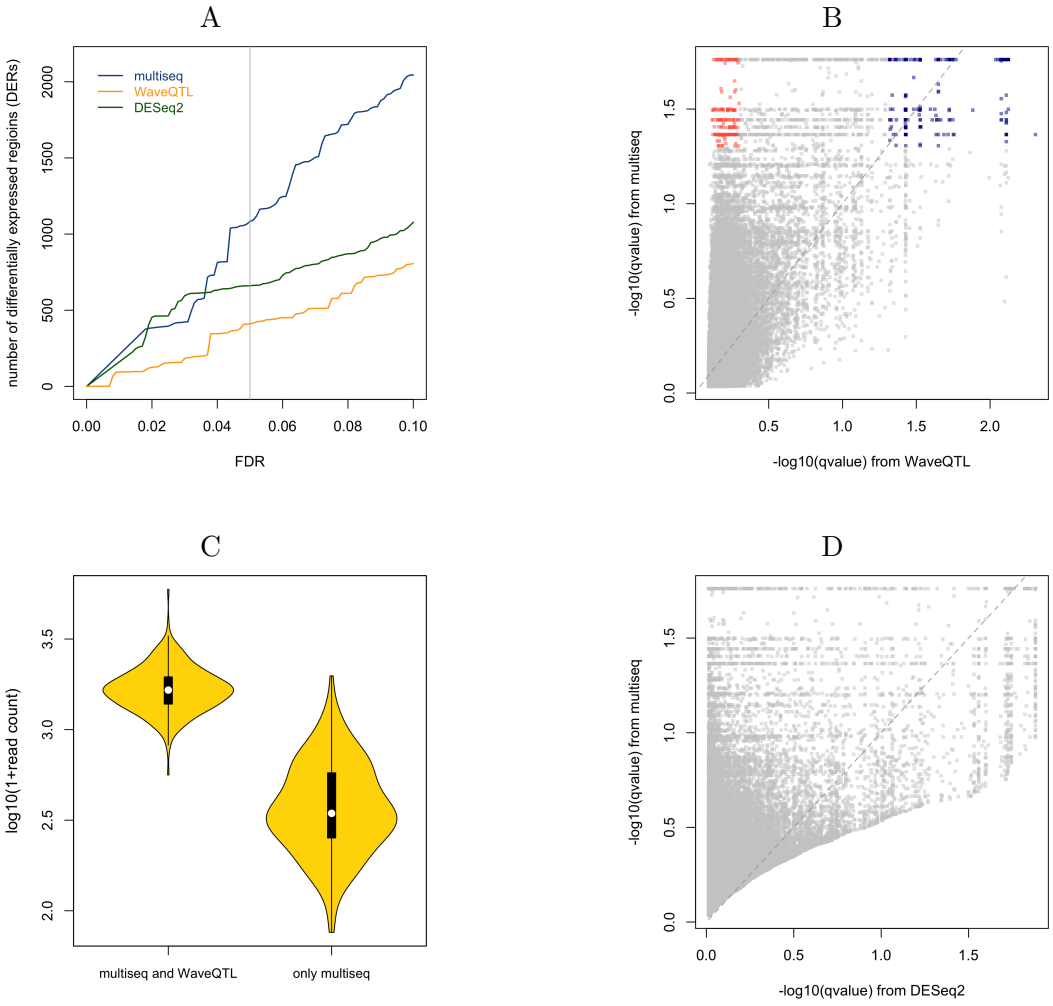

Fig 2: **Comparison of multiseq with WaveQTL and DESeq2.** Panel A shows the number of DERs identified by each method at a given FDR. The grey line indicates FDR = 0.05. Panel B shows a scatter plot of the q-values from multiseq versus the q-values from WaveQTL. The q-values are computed by the qvalue package (Storey et al., 2020). The dashed line indicates the $y = x$ line. The 271 DERs detected by both methods at FDR = 0.05 are colored in blue. The 186 DERs detected by multiseq at FDR = 0.05, but missed by WaveQTL at FDR = 0.5, are colored in red. Most points are above or around the $y = x$ line, reflecting that most DERs identified by WaveQTL are detected by multiseq. Panel C shows the distributions of the total read count for two sets of regions: 1) DERs detected by both multiseq and WaveQTL (blue points in panel A) and 2) DERs detected by multiseq, but not WaveQTL (red points in panel A). For better visualization, regions with very large read counts (the 7 regions in set 1 with total read count $>$ 6000) are excluded from the distributions. Panel D shows a scatter plot of the q-values from multiseq versus the q-values from DESeq2. Most points are above or around the $y = x$ line, indicating that multiseq identifies most DERs detected by DESeq2.

compared multiseq to DESeq2. Since multiseq uses the output from DESeq2 to compute a likelihood ratio measuring the support for the difference in overall expression (see Supplementary Material) and then combines it with the higher-resolution information from its multi-scale model, any difference in performance between multiseq and DESeq2 is due to the additional information in the higher-resolution information. Fig 2A shows that multiseq increases power compared to DESeq2 for most values of the FDR, with slightly decreased power at FDR $= 0.02 \sim 0.036$. The increased power indicates the contribution of the high-resolution information to the identification of regions with differential chromatin accessibility. Moreover, multiseq identifies the majority of DERs detected by DESeq2 (Fig 2D). See Supplementary Material for a comparison to DESeq2 with different bin sizes, where multiseq consistently outperforms them.

It is natural to ask what kinds of patterns of effects are better suited to multi-scale analysis. Shim and Stephens (2015) showed that overall expression methods are well powered to identify effects appearing on most regions in the same direction, but that multi-scale methods have an advantage for effects over much shorter scales or in opposite directions. We observed consistent patterns in this study. Fig 3 and Fig 4 show two example of DERs identified by multiseq, but not DESeq2, which we now discuss in turn.

The DER in Fig 3 shows strong effects in multiple areas. The effect in the first pink area is consistent in direction over about 50bp and the effects in the other pink areas are in opposite directions over about 100bp. DESeq2 missed this signal because the effects in opposite directions partially cancel each other out, leading to a weak overall signal in the 1024bp region. multiseq makes better use of the whole signal and easily captured it. Furthermore, we applied DESeq2 with smaller bin sizes (300bp and 100bp), and only DESeq2 with 100bp bin successfully detects the signal (Supplementary Material Figure 2).

The DER in Fig 4 has effects in a relatively narrow area (the pink areas span $\approx$ 20bp). While multiseq captured this signal, DESeq2 failed to detect it partly because the signal affects the area much smaller than DESeq2 bin size (20bp $\ll$ 1024 bp), and also because the effects in opposite directions partially cancel each other out. DESeq2 with smaller bin sizes still missed this signal (Supplementary Material Figure 3).

See Shim and Stephens (2015) for more extensive comparisons and discussions of the benefits of multi-scale approaches vs overall expression methods.

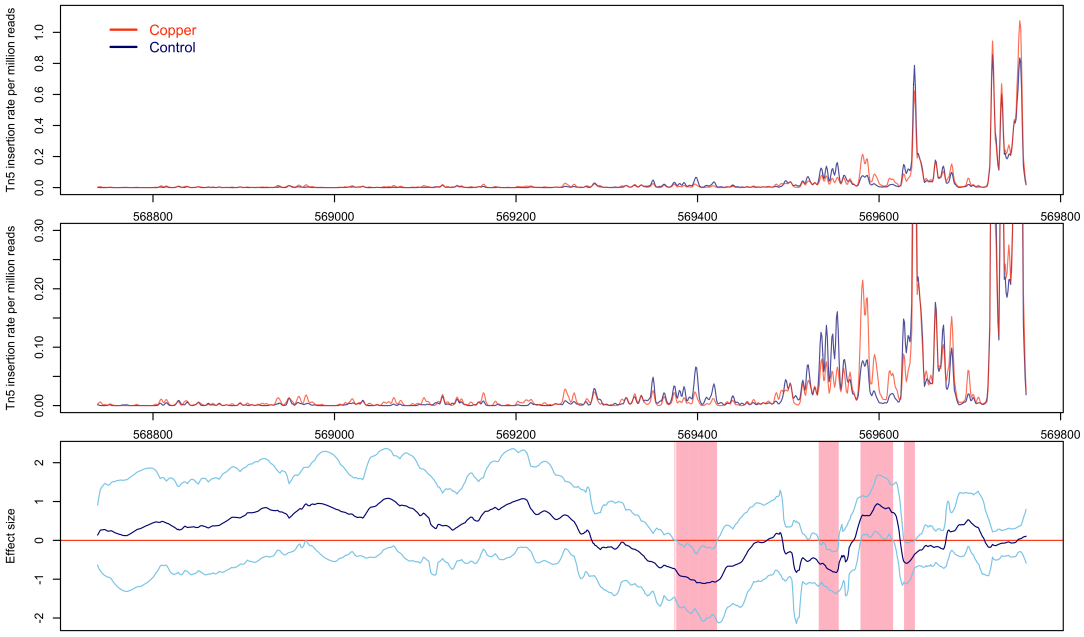

Fig 3: **Example of DERs identified by multiseq, but missed by DESeq2 (chr1:568739-569762).** The top two panels show the average Tn5 insertion rates along the region for each group with different limits of the y-axis. The bottom panel shows the posterior mean for effect size, i.e., the difference in log-intensity between the two groups (blue), and ± 2.6 posterior standard deviations (sky blue). Areas showing strong difference (zero is outside of the interval constructed by the two sky blue lines) are colored in pink. For multiseq, $\log \Lambda \approx 98.91$; p-value ≈ 0.0000034. For DESeq2, p-value ≈ 0.19.

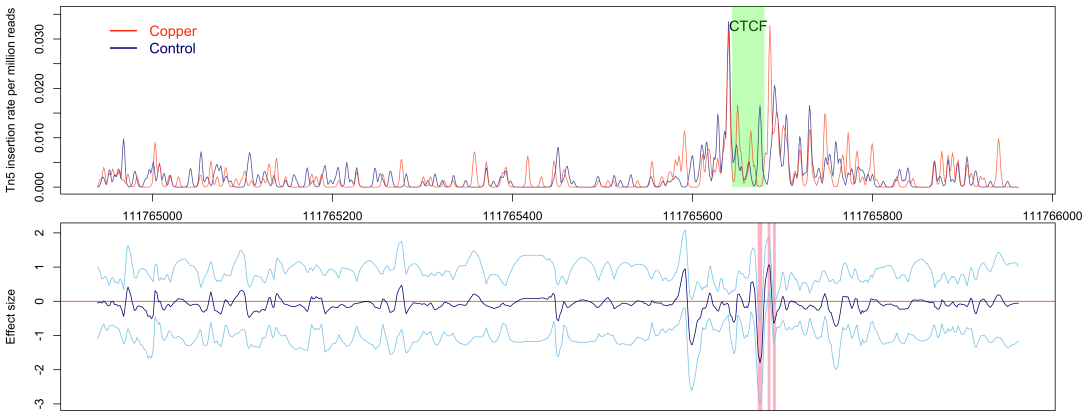

Fig 4: **Example of DERs identified by multiseq, but missed by DESeq2 (chr1:111764939-111765962).** Labels and colors are as in Fig 3. In the top panel, the green block indicates the putative CTCF binding site, identified by the software CENTIPEDE (Pique-Regi et al., 2011). The bottom panel shows the posterior mean for effect size and ±1.5 posterior standard deviations. For multiseq, $\log \Lambda = 10.49$; p-value ≈ 0.000019. For DESeq2, p-value ≈ 0.23.

3.2.6. Potential mechanism underlying the difference in chromatin accessibility

In Fig 4, the pink areas with effects occur near a binding site for CTCF (CCCTC binding factor). This suggests that the observed changes in chromatin accessibility in this region may be related to changes in CTCF binding. Indeed, independent analysis of copper-treated and control samples using the software CENTIPEDE (Pique-Regi et al., 2011) shows a change in CTCF binding strength between two conditions (posterior odds: 540 in copper-treatment and 405 in control samples). Moreover, using RNA-seq data from Moyerbrailean et al. (2015) measuring gene expression on the same samples, we tested for differential gene expression in nearby genes, and observed a 23% decrease of gene expression of CHI3L2, a gene within 1 Kb of this DER. It therefore seems plausible that the increase in CTCF binding in copper-treated samples insulates CHI3L2, causing the decrease in gene expression by insulating its promoter. (However, note that this discussion of potential mechanism is inevitably speculative given the limits of current data.)

3.2.7. Computation

Analysis of the entire data set (242,714 tests of copper treatment vs control 1 and 242,714 tests of two controls) took about 103 CPU hours (user + system). This consisted of 10 hours for the data preprocessing and 93 hours for running multiseq. Because the analysis of each region is independent, the entire analysis is massively parallelizable in a naive way (on average 0.76 sec CPU time for each region). Software and scripts implementing our methods and analyses are available at `https://github.com/heejungshim/multiseq` and `https://github.com/heejungshim/multiseq_Shim_et_al`.

4. Discussion

We have developed a novel multi-scale method, multiseq, to estimate and test for differences in molecular phenotypes between multiple groups of samples using high-throughput sequencing data. The method is built on multi-scale models for inhomogeneous Poisson processes which enable multiseq to have two features: 1) better use of the high-resolution information in the data, and 2) direct modeling of the count nature of the data. The first feature allows multiseq to effectively detect differences that vary in scales or directions.

Previous work (Shim and Stephens, 2015) demonstrated the advantages of multi-scale approaches over overall expression methods, and our experiments support these results. The second feature, which is a main contribution of our work, allows multiseq to better maintain power in the analysis of data with small sample sizes or low read counts compared to the previous normal-based methods. We demonstrate this advantage on the simulation study with data sets of different sample sizes and library read depths, and the analysis of ATAC-seq data with small sample size 6. Finally, multiseq is computationally tractable for a large-scale differential expression analysis involving hundreds of thousands of tests.

Although multiseq was motivated by differential high-throughput sequencing data analysis for molecular phenotypes, it could be more generally applied to association analysis between a sequence of counts and other covariate, either continuous or discrete. For example, it could be used to detect associations between a molecular phenotype, such as chromatin accessibility measured by the sequencing data, and a continuous covariate, such as gene expression. Or, it could be used to detect and estimate differences in the intensity of any two (nonhomogenous) Poisson processes unrelated to genomics, such as gamma-ray burst signals (Kolaczyk, 1999) in astronomy, between two conditions.

In our ATAC-seq data analysis, we constructed the empirical null distribution of the test statistic using a *null* data set that is expected to have no differences in chromatin accessibility between conditions. This distribution is required to obtain p-values that are used to compute q-values and estimate FDR. Alternatively, a null distribution could be generated by permuting sample labels (e.g., see Shim and Stephens (2015)), or a Bayesian FDR could be used to address the multiple testing issue, as it does not require p-values. For example, Morris et al. (2008) used wavelet-based functional mixed models to identify positions (or bins) with absolute values of effect sizes $> \delta$, and introduced an approach for controlling the Bayesian FDR that uses the posterior distributions of effect sizes. Ma and Soriano (2018) also controlled the Bayesian FDR to identify resolutions/locations with significant signal in their approach for analysis of distributional variance.

We have demonstrated that multiseq outperforms a wavelet-based approach, particularly in the analysis of data with small sample sizes or low read counts. However, there are still opportunities for potential improvements. First, the normal approximations to binomial likelihoods use the estimates of parameters and their standard errors, but with small sample sizes, the

estimated standard errors can be less stable. In genomics, this issue has been addressed by using the shrinkage of the estimated sample variances to pooled estimates which are more stable (Smyth, 2004). Incorporating the shrinkage estimates to multiseq could potentially further improve performance in small sample sizes. Second, our model assumes independent priors on effect sizes across scales and locations, but in practice the strengths of effects in multiscale models tend to have dependencies - they tend to propagate across adjacent locations and scales (Crouse, Nowak and Baraniuk, 1998). Priors that exploit the dependencies, such as the tree-like structure described in Crouse, Nowak and Baraniuk (1998), can effectively combine information across different scales and locations. Extending multiseq to impose such priors could further improve performance.

Supplementary material

Supplementary material is available at https://github.com/heejungshim/multiseq_Shim_et_al/blob/main/manuscript/multiseq_supp.pdf.

Acknowledgements

We thank Jack Degner for invaluable discussions and Yao-ban Chan for helpful comments on a draft manuscript. We also thank the members of the H. Shim, M. Stephens, and J. Pritchard labs for helpful discussions. This work was supported by NIH grant HG002585.

References

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell* **129** 823–37.

Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S. and Crawford, G. E. (2008). High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132** 311–22.

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.

Busby, M. A., Stewart, C., Miller, C. A., Grzeda, K. R. and Marth, G. T. (2013). Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. *Bioinformatics* **29** 656-657.

Coifman, R. R. and Donoho, D. L. (1995). Translation-invariant denoising. In *Wavelets and statistics* 125–150. Springer.

Collado-Torres, L., Nellore, A., Frazee, A. C., Wilks, C., Love, M. I., Langmead, B., Irizarry, R. A., Leek, J. T. and Jaffe, A. E. (2017). Flexible expressed region analysis for RNA-seq with derfinder. *Nucleic acids research* **45** e9–e9.

Crouse, M. S., Nowak, R. D. and Baraniuk, R. G. (1998). Wavelet-based statistical signal processing using hidden Markov models. *IEEE Transactions on Signal Processing* **46** 886–902.

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.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to Unknown Smoothness via Wavelet Shrinkage. *Journal of the American Statistical Association* **90** 1200–1224.

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* **15** 413–426.

Hesselberth, J. R., Chen, X., Zhang, Z., Sabo, P. J., Sandstrom, R., Reynolds, A. P., Thurman, R. E., Neph, S., Kuehn, M. S., Noble, W. S., Fields, S. and Stamatoyannopoulos, J. A. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature methods* **6** 283–9.

Johnson, D. S., Mortazavi, A., Myers, R. M. and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316** 1497–502.

Kolaczyk, E. D. (1999). Bayesian Multiscale Models for Poisson Processes. *Journal of the American Statistical Association* **94** 920–933.

Law, C. W., Chen, Y., Shi, W. and Smyth, G. K. (2014). Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome biology* **15** 1.

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

664–672.

Liu, Y., Zhou, J. and White, K. P. (2014). RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics* **30** 301–304.

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.

Luca, F., Maranville, J. C., Richards, A. L., Witonsky, D. B., Stephens, M. and Di Rienzo, A. (2013). Genetic, functional and molecular features of glucocorticoid receptor binding. *PloS one* **8** e61654.

Ma, L. and Soriano, J. (2018). Analysis of Distributional Variation Through Graphical Multi-Scale Beta-Binomial Models. *Journal of Computational and Graphical Statistics* **27** 529-541.

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research* **18** 1509–17.

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R. P., Lee, W., Mendenhall, E., O'Donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., Nusbaum, C., Lander, E. S. and Bernstein, B. E. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448** 553–60.

Morris, J. S., Brown, P. J., Herrick, R. C., Baggerly, K. A. and Coombes, K. R. (2008). Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models. *Biometrics* **64** 479–489.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* **5** 621–8.

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**.

Pique-Regi, R., Degner, J. F., Pai, A. A., Boyle, A. P., Song, L., Lee, B.-k., Gaffney, D. J., Gilad, Y. and Pritchard, J. K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome research* **21** 447–55.

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, D. G. and Storey, J. D. (2014). subSeq: determining appropriate sequencing depth through efficient read subsampling. *Bioinformatics* **30** 3424–3426.

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.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology* **3**.

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 R package version 2.20.0.

Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A. and Conesa, A. (2011). Differential expression in RNA-seq: a matter of depth. *Genome research* **21** 2213–2223.

Timmermann, K. E. and Nowak, R. D. (1999). Multiscale modeling and estimation of Poisson processes with application to photon-limited imaging. *IEEE Transactions on Information Theory* **45** 846–862.

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

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P. and Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* **456** 470–6.

Xing, Z., Carbonetto, P. and Stephens, M. (2021). Flexible Signal Denoising via Flexible Empirical Bayes Shrinkage. *Journal of Machine Learning Research* **22** 1-28.