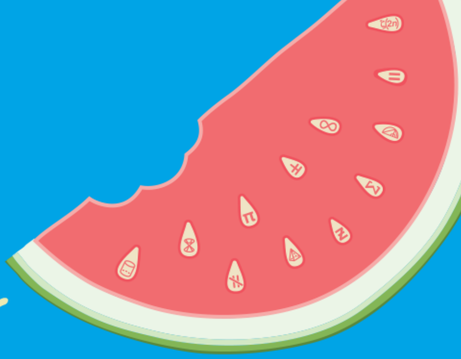
AMSI VACATION RESEARCH SCHOLARSHIPS 2021–22

Get a taste for Research this Summer



Multi-scale Poisson process approaches for analysis of paired-end high-throughput sequencing data

Yuxin Yan

Supervised by Heejung Shim

The University of Melbourne

AMSI AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE

Abstract

Multi-scale Poisson method has been proposed to detect differences between samples of sequence of count data with a binary group indicator. Previous methods used the idea of Wavelet transformation and factorize the likelihood of independent data $\sim$ an inhomogeneous Poisson process into a product of Poisson and Binomial. This report extend the 1-dimensional method into 2-dimensional using the idea of 2D Wavelet transformation, By doing so, the multi-scale method can be applied in more real-life scenarios.

1 Introduction

Understanding and detecting differences between one-dimensional count data has lots of real-life application scenarios. The basic setting for this problem is that the samples of the data are generated from an inhomogeneous Poisson process and each sample has a group indicator. To be more precise, each sample of sequence of count data has a set of intensity parameters captures different means in each location. The problem of interest is to detect differences between samples of sequence count data. In biology context, this context can be extended as a way of understanding the molecular basis of gene regulation. One of the most essential part of understanding the molecular basis of gene regulation is to identify differences in molecular phenotypes (e.g., gene expression, chromatin accessibility). These phenotypes are commonly measured using high-throughput sequencing assays (e.g., RNA-seq, ATAC-seq), which provide high-resolution measurements that reflect how the phenotypes vary along the genome in each sample. These assays, in particular, provide the number of sequences that arose from each location in the genome, where the magnitude of the count represents the intensity of the underlying phenotype at that location. Within this biology context, the problem of interest is to detect the difference between the true cellular-level traits of multiple groups.

Previous popular methods to detect the difference between multiple different sequence data are called window-based approach [2][4][6]. This method involves choosing a region size which could bring more challenge of the task. To fully detect the signals from the data, several previous multi-scale methods have been proposed, aiming to make better use of the high-resolution measurements from the data. However, the fundamental disadvantage of these methods is that they approximate read counts using normal distribution, which works well when counts are sufficiently large or sample sizes are sufficiently big, but performs badly when sample sizes are small or counts are low [6]. To fully exploit the high-resolution measurements of the cellular-level traits of multiple groups of samples, multiseq method [7] has been proposed. It is developed to detect and estimate differences in the intensity among samples along the genome, taking account of both the high-resolution and the count nature of the data. Specifically, it assumes that each sample's count data is generated by an inhomogeneous Poisson process with a spatially organized underlying intensity function. By extending this intensity function from existing multi-scale models for inhomogeneous Poisson processes, it further estimates and tests for the differences in the underlying intensity among samples.

Based on the multiseq method [7], 2-dimensional wavelet based multi-scale Poisson process method can be proposed. The underlying problem of interest is to detect differences between samples of image count data.

To be more specific, instead of each sample has a curve-shaped data, each sample has an image type data consisting of counts. Each sample has a 2-dimensional data, that can be visualized using an image. This report proposes a 2-dimensional multi-scale Poisson process method based on previous existing multi-scale models for inhomogeneous Poisson processes and multiseq method [7] by extending 1-dimensional wavelet transformation to 2-dimensional. This report will first review the multiseq methods and then introduce the proposed 2-dimensional method.

Statement of Authorship

Under the direction of my supervisor, I extended the 1-dimensional multiscale Poisson process methods to 2-dimensional case and wrote this report. My supervisor assisted with the work throughout, answered questions, and proofread this report.

2 Backgrounds

2.1 Window-based methods

Based on the characteristics of count nature and high-resolution measurements of the sequence data, previous methods [2][3][4][6][7] failed to detect the full information of the samples. The simplest method is to divide the sequences of data into sub-sections and add the total counts of each region followed by testing for differences in these total counts using analysis methods available [2][4][6]. However, one limitation of this method is that the size of the region is difficult to select: one needs balance between generating opposite conclusions or missing signals from the data. To be more specific, if the region size is too big, then the method loses the sensitivity of the inference, and it misses signals that affect smaller sub-regions of the data; if the region size is too small, then one risks missing power of the method because each sub-regions will have low counts. Generally, window-based approaches do not fully exploit the high-resolution information from the data.

Meanwhile, several methods have been proposed to solve this problem by making use of the multi-resolution signals from the data [3][7]. Wavelet-based multi-scale methods can test for differences by taking multiple resolutions into consideration, thus effectively avoid the problem of selecting a single resolution or region size. In this way, it can capture more signals compared to window-based methods. The main limitation of these methods is that they ignore the count nature of the data by approximating the read counts using a normal distribution which performs poorly for relatively small sample sizes or low counts.

2.2 Wavelet transformation

Wavelet representations are effective tools from signal and image processing applications [5]. In this report, Haar Discrete Wavelet Transformation (DWT) function has been used to applied to the multi-scale Poisson model, and this section provides a brief description of the Haar Wavelet Transform.

Consider a 1-dimensional count sequence of data. The 1-dimensional wavelet transformation involves reparameterizing it into different wavelet coeffficients of each scale. Consider a sequence of count data of length T with base locations from 1 to T. At scale $s = 1, ..., log_2(T)$, wavelet transformation splits the sequence into 2^s sub-sections with equal lengths. The wavelet transformation consists of two types of transforms of data: plus and minus which produce 2^s transformed data of the orginal space at scale s.

Specifically, starting from the last scale ($s = log_2(T)$), after splitting the data into 2^s sub-sections, the "plus" and "minus" transformations are applied separately on each pair of the sub-sections (i.e. the data point at each location) which results in T transformed data. Then before going into the next resolution, the "plus" and "minus" transformations are only applied on the previous resolution's "plus" transformed data. In this way, there are T/2 number of the transformed data involving in the next transformation where the same rule would be applied by adding and subtracting each pair of the previous transformed data. The DWT decomposes the data into "wavelet coefficients" (WCs), each of which captures the difference between intensity of the data in different locations for each scale. At the "zeroth scale", there is only one single WC, which is calculated as the total sum of the elements of the data. At the first scale, there is also one WC which captures the difference between the first half and the second half. At the second scale, there are two WCs, the first contrasts the first quarter and the second quarter of the original sequence; and the second contrasts the third quarter and the fourth quarter of the original sequence. This process continues reaching the final scale (i.e. $s = log_2(T)$), and at scale s there are 2^{s-1} WCs capturing contrasts between sub-sections of the original data of length 2^{T-s}.

To give a more detailed explanation, consider an example of sequence of count data of length 8, data at each base location is indexed as $x_1, x_2, ...x_8$. According to the wavelet transformation, at the "zeroth scale" there is only one total summation WC (θ_{01}) which simply adds all the count data. At the first scale, the WC (θ_{11}) contrasts the first half and the second half of the data. That is, $\theta_{11} := \sum_{i=1}^{4} x_i - \sum_{i=5}^{8} x_i$. The second scale is reparameterized by two WCs (θ_{21}, θ_{22}) respectively with the first calculated as $\theta_{21} := \sum_{i=1}^{2} x_i - \sum_{i=3}^{4} x_i$, the second calculated as $\theta_{22} := \sum_{i=5}^{6} x_i - \sum_{i=7}^{8} x_i$. And the last scale (i.e. s = 3) contrasts each pair of data by producing 4 WCs ($\theta_{31}, \theta_{32}, \theta_{33}, \theta_{34}$), where $\theta_{31} = x_1 - x_2, \theta_{32} = x_3 - x_4, \theta_{33} = x_5 - x_6, \theta_{34} = x_7 - x_8$.

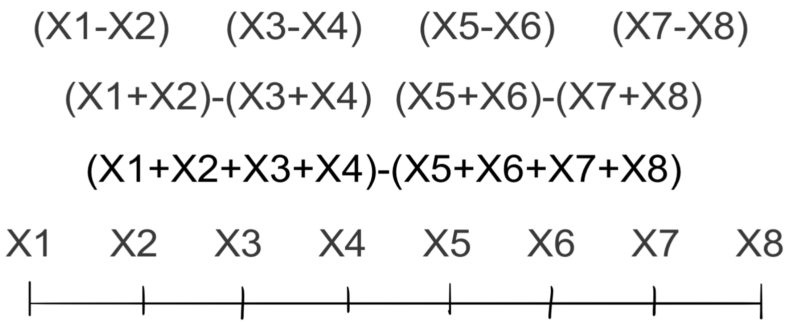


Figure 1: 1D Wavelet transformation

2.3 Multi-scale models for inhomogeneous Poisson processes

The multi-scale model for inhomogeneous Poisson process is build based on wavelet transformation and it reparameterizes the Poisson model using a 1-1 multi-scale transformation [1]. To begin with, first consider the model for sample size n=1. Suppose the observed data are $\mathbf{y} = (y_1, ..., y_B)$ with

$$y_b \sim Poisson(\lambda_b) \tag{1}$$

where B is a power of 2, so $B = 2^J$ for some J. The model splits the data into different number of sub-sections at each scale which results in a 1-1 multi-scale reparameterization of $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_B)$ [1]. At scale $s = 1, ..., log_2(B)$ define 2^{s-1} "locations" by dividing the indices $1, ..., B$ into 2^{s-1}equal groups of consecutive indices, and let l_{sl} denote the indices of the location at scale s so formed. And the assumption is that data is spatially structured which $\mid \lambda_b$-$\lambda_{b+1} \mid$ is small for most b. Here is a brief summary of the multi-scale model from previous work [8][9]:

At each scale $s = 1, ..., log_2 B$ define 2^{s-1} parameters capture the log difference between sub-sections by dividing the data $\mathbf{y}$ into 2^s equal length groups of different "locations" parameterized by I_{sl}, that is, s stands for scale and l stands for location. For example, at scale 3 there are $2^{3-1} = 4$ locations, that is:

$$I_{31} = [1, B/4], I_{32} = [B/4 + 1, B/2], I_{33} = [B/2 + 1, 3B/4], I_{34} = [3B/4 + 1, B]$$

where [a, b] denotes the base indices from a to b. Further define I^-_{sl}, I^+_{sl} as the first and second halves of the indices in I_{sl}. Then to define the α which captures the log differences of intensity between the first and the second halves of indices in I_{sl}, denote the sum of λ over the first half as λ^+_{sl}, the sum of across the second half as λ^-_{sl}. At last, define the multi-scale parameter of each location and scale by

$$\alpha_{sl} := log(\lambda^+_{sl}/\lambda^-_{sl})$$

The intuition is the same as the wavelet transformation which captures the difference between sub-regions along the sequence of the data. Specifically, if λ remains invariant in I_{sl} then $\alpha_{sl} = 0$. This is the essential reason that multi-scale model is applied to address such problem: it produces sparsity structure in the transformed space which makes it easier to solve compared to the original data space.

Moreover, by applying a fundamental distributional result: if y_1 and y_2 are independent, with $y_i \sim Pois(\lambda_i)$, then

$$p(y_1, y_2|\lambda_1, \lambda_2) = Pois(y_1 + y_2; \lambda_1 + \lambda_2)Bin(y_1; y_1 + y_2, \lambda_1/(\lambda_1 + \lambda_2). \tag{2}$$

By applying (2), the likelihood of parameters can be factorized into independent terms [1]:

$$p(\boldsymbol{y}; \boldsymbol{\alpha}, \lambda_{tot}) = Pois(\sum_b y_b; \lambda_{tot}) \prod_{sl} Bin(y^-_{sl}; y^-_{sl} + y^+_{sl}, exp(\alpha_{sl})/(1 + exp(\alpha_{sl})), \tag{3}$$

where $Pois(; \lambda)$ denotes the probability mass function of Poisson distribution given parameter λ, while $Bin(; n, p)$ denotes the probability mass function of inhomogeneous Poisson given parameters n and p. Moreover, y^-_{sl}, y^+_{sl}

denote the sum of y_b over the indices I_{sl}^-, I_{sl}^+. In this way, the data in each location can be modelled as binomial distribution, where

$$y_{sl}^- \sim Bin(y_{sl}^- + y_{sl}^+, p_{sl}) \tag{4}$$

$$\alpha_{sl} = log(p_{sl}/(1 - p_{sl}) \tag{5}$$

3 Review of Multiseq

This section involves the review of previous 1-dimensional method called Multiseq [8].

3.1 Multi-scale models for inhomogeneous Poisson processes from multiple groups of samples

Now consider there is a group of samples of data with samples size n. To estimate $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_B)$ under the assumption that λ is spatially structured, the model is

$$y_b^i | \alpha_i, \lambda_{tot}^i \sim Pois(\lambda_b^i) i = 1, ..., n, \tag{6}$$

where λ_b^i represents the b^{th} component of the 1-1 multi-scale transformation. The multi-scale models transfer parameters of original space into the wavelet transformed space consists of WCs described above. The group indicator is modelled as covariate X_i measured on each sample, Multiseq assumes that $X_i \in \{0, 1\}$ which is a binary group indicator. The effect of the group indicator on the intensity is modelled using a linear model. The model involves two regression parts for the "zeroth" scale and other scales.

3.1.1 Poisson regression

Let y_{tot}^i denotes the total count over the region for sample i, since the multi-scale model assumes the total count follows a Poisson distribution with intensity λ_{tot}^i; see equation (6). Poisson regression is used to model the effect of the group indicator X_i:

$$y_{tot}^i \sim Pois(\lambda_{tot}^i) \tag{7}$$

$$log(\lambda_{tot}^i) = \mu_{01} + \beta_{01} X_i + u_{01}^i \tag{8}$$

where u_{01}^i models the random effect of individual samples to handle the problem of overdisperssion, μ_{01} models the average effect of covariate on the intensity when $X_i = 0$, β_{01} captures the difference in intensity between groups. The model for "zeroth" scale is equivalent to generalised linear model of Poisson distribution with a log link.

3.1.2 Binomial regression

For other scales, consider the single sample case, the likelihood factorizes into independent terms (4) (5). The same rule is applied here, the information in μ_{sl}, β_{sl} is contained in n binomial observations:

$$y_{sl}^{i,-} \sim Bin(y_{sl}^{i,-} + y_{sl}^{i,+}, p_{sl}^i) \tag{9}$$

$$log(p^i_{sl}/(1-p^i_{sl})) = \alpha^i_{sl} = \mu_{sl} + \beta_{sl}X_i + u^i_{sl} \tag{10}$$

The model is a generalized (Binomial) linear model with a logit link function. β_{sl} reveals how much response variable change when X_i changes one unit which is the effect of X_i on α_{sl}.
Moreover, mixture modelling is applied on the prior of β_{sl}:

$$\beta_{sl} \sim \gamma_{sl}N(0,\tau^2_{sl}) + (1-\tau_{sl})\delta_0 \tag{11}$$

$$\gamma_{sl} \sim Bernoulli(\pi_s) \tag{12}$$

where γ_{sl} is the mixing parameter, δ_0 is the zero point mass. Then with probability π_s, $\beta_{sl} \sim N(0,\tau^2_{sl})$; with probability $1-\pi_s, \beta_{sl} = \delta_0$.

3.2 Testing for differences between multiple groups of samples

To test/detect the difference between groups of samples is equivalent to test for non-zero effects over the region. The null hypothesis is $H_0 : \beta_{sl} = 0 \forall s, l$ which is equivalent to testing $\pi_s = 0 \forall s$ in the prior for β. To be more specific, if $\pi_s = 0$ then β_{sl} is always equal to δ_0.
Then likelihood ratio test is used to detect the non-zero effects over the sample regions [7], the test statistic is:

$$\hat{\Lambda} = \Pi_{sl}\frac{P(\hat{\beta_{sl}}|s^2_{sl},\hat{\boldsymbol{\pi}})}{P(\hat{\beta_{sl}}|s^2_{sl},\pi_0 = 1}) \tag{13}$$

where $\hat{\boldsymbol{\pi}}$ denotes the maximum likelihood estimator, that is $\hat{\boldsymbol{\pi}} := \text{argmax}\Pi_{sl}P(\hat{\beta_{sl}}|s^2_{sl})$.

3.3 Effect size estimation

To provide more interpretable estimates of the effect of X, the posteriors in the transformed multi-scale space is transformed into the log-intensity $log\boldsymbol{\lambda}$ into the original observation space. We define the effect on base location index in the original space as $\beta^0_b := log(\lambda^{(1)}_b/\lambda^{(0)}_b)$ where $\boldsymbol{\lambda^{(0)}}, \boldsymbol{\lambda^{(1)}}$denote the values for $\boldsymbol{\lambda}$ for individual sample in group 0 or 1 respectively.
The group 0 is set to be the baseline, therefore, $\boldsymbol{\mu} \equiv \boldsymbol{\alpha_0}$ and $\boldsymbol{\mu} \equiv \boldsymbol{\alpha_1} - \boldsymbol{\alpha_0}$. The relationship between $\boldsymbol{\lambda^i}$ and $\boldsymbol{\alpha^i}$ for $i = 0, 1$ is explored under the assumption that the random effects $u = 0$. From the elementary properties of the Poisson distribution, the intensity parameter of Poisson at base b λ^i_b can be written as a product of the total intensity λ^i_{tot} and the binomial probability of success and failure $p^i_{sl}, q^i_{sl} := 1 - p^i_{sl}$, where

$$p^i_{sl} = \frac{e^{\alpha^i_{sl}}}{1+e^{\alpha^i_{sl}}}, q^i_{sl} = \frac{1}{1+e^{\alpha^i_{sl}}} \tag{14}$$

To give a more specific example, the intensity at the two leftmost positions (i.e. $b = 1, 2$) can be written as

$$\lambda^i_1 = \lambda^i_{tot}[\Pi^{J-1}_{s=1}p^i_{sl}]p^i_{J1}, \lambda^i_2 = \lambda^i_{tot}[\Pi^{J-1}_{s=1}p^i_{sl}]q^i_{J1} \tag{15}$$

where p^i_{sl} and q^i_{sl} can be considered as the probabilities of assigning the count to the left half and the right half of the sub-regions at the scale s and location l for a region.

For group category variable (i.e. $X = 0$ or $X = 1$), the value of the covariate at the baseline is defined as X_0, and a unit increase from baseline is defined as $X_1 := X_0 + 1$. The the effect size β_b^0 is defined as:

$$\beta_b^0 = log\lambda_b^1 - log\lambda_b^0, \tag{16}$$

where λ_b^i denotes the intensity for X_i at position b. Using the relationship between $\boldsymbol{\lambda^i}$ and $\boldsymbol{\alpha^i}$, equation (16) can be rewritten as a sum of $log(\frac{\lambda_{tot}^1}{\lambda_{tot}^0})$ and $log(\frac{p_{tot}^1}{p_{tot}^0})$ or $log(\frac{q_{tot}^1}{q_{tot}^0})$.

4 2D Multi-scale models

4.1 2D Wavelet transformation

To extend the 1-dimensional model into 2-dimensional case, the problem of interest has now changed into detecting differences between a sequence of image type data. For the previous 1-dimensional case, each samples consists of s sequence of count numbers, while for 2-dimensional case, each sample has an image type data consists of counts. Let us denote the image size by $T \times T$, where $T = 2^J$ for some J. Each of the sample also has a group indicator g_i, $g_i \in \{0, 1\}$.

The idea of 2-dimensional Wavelet transformation is widely used in image compression [5], and the wavelet decomposition can be viewed as a decomposition using a set of independent frequency channels. The decomposition of 2-dimensional wavelet can be divided into 4 sub-transforms. Each transformation involves a sub-section of the original image space of 2×2 area, then the WCs are noted down before going into the next resolution. By applying the same transformations on "plus" transformed WCs, further finer resolution's wavelet transformations can be done.

To give a more specific example, consider an 8×8 image count data as bellow (Figure 2). There are 4 types of wavelet transforms in total which are shown in Figure 2, each of them involving transform an area of 2×2 count data from the sample. Take the first transformation as an example (Figure 3), each consecutive 2×2 square

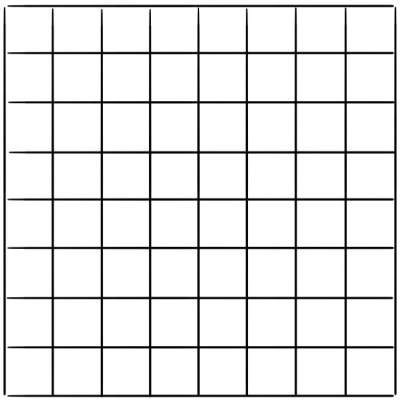


Figure 2: 8×8 image data example

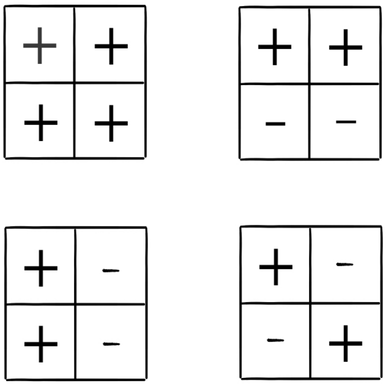


Figure 3: 4 2D Wavelet transformation types

area is transformed as a WC, and there are 16 WCs after the single transformation on the whole area over the data region. The same rule is applied for other 3 types of transformations. After the first scale's transformation, 64 WCs are produced before going into the next resolution.

The transformation of the next scale is then applied on 4×4 count image data which consists of WCs from the "plus" transformation of the previous scale [Figure 5]. Then the same rule is applied: before going into the next resolution, 16 WCs are produced.

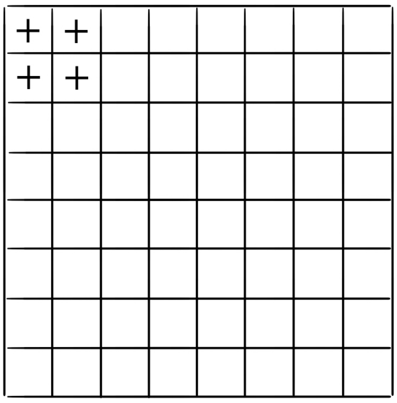


Figure 4: "Plus" transformation on image data

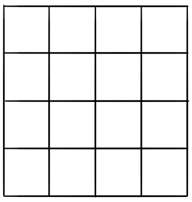


Figure 5: 4×4 image data for the next resolution

The finest resolution consists of a 4×4 image data producing only 4 WCs.

4.2 Modelling

The idea of reparametrization of the independent likelihood shown in section 2.3 is also applied for the 2D case. From the elementary distributional results, the likelihood can be written as the product of Binomial and Poisson. In 2D case, the probability of success of the binomial distribution is defined as the total sum of "plus" divided by the total sum over the corresponding region. In this way, there are 3 p_{slj} for each scale and different location where j indicates the type of transformation.

For example, in the last resolution, there are 3 parameters of probability of success for binomial, that are: p_{sl1}, p_{sl2}, and p_{sl3}, each of them denotes the total sum of "plus" divided by the sum over the region: p_{sl1} is calculated as the top half of the data divided by the total sum, p_{sl2} is calculated as the left half of the data divided by the total sum, p_{sl3} is calculated as the cross elements (i.e. top left and bottom right) divided by the total sum [Figure 6].

5 Conclusion

Multi-scale methods [7][8] have been proposed to help exploit 1-dimensional high-resolution measurements for sequence of count data. One of the most essential application in biology is the identification of differences in

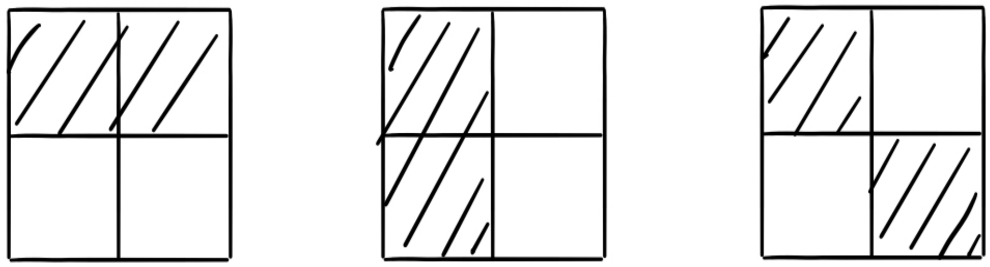


Figure 6: Probability of success in Binomial for 2D

molecular phenotypes using high-resolution measurements from some high-resolution assays. However, they are designed to analyse single-end read high-throughput sequencing data, so it is not directly applicable to paired-end read data. This project builds up on the multi-scale method proposed in [7] and develops and implements statistical methods that better exploit high-resolution measurements in the 2-dimensional image count data. Future work can be done by applying this proposed methods to identify differences using paired-end Hi-C data.

References

[1] Kolaczyk, E., 1999. Bayesian Multiscale Models for Poisson Processes. *Journal of the American Statistical Association,* 94(447), pp.920-933.

[2] Law, C., Chen, Y., Shi, W. and Smyth, G., 2014. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology,* 15(2), p.R29.

[3] Lee, W. and Morris, J., 2015. Identification of differentially methylated loci using wavelet-based functional mixed models. *Bioinformatics,* 32(5), pp.664-672.

[4] Love, M., Huber, W. and Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology,* 15(12).

[5] Mallat, S., 1989. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 11(7), pp.674-693.

[6] Robinson, M., McCarthy, D. and Smyth, G., 2009. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1), pp.139-140.*Bioinformatics* 26 139–140.

[7] Shim, H. and Stephens, M., 2015. Wavelet-based genetic association analysis of functional phenotypes arising from high-throughput sequencing assays. *The Annals of Applied Statistics,* 9(2).

[8] Shim et al., 2021. Multi-scale Poisson process approaches for differential expression analysis of high-throughput sequencing data. arXiv:2106.13634

[9] Xing, Z., Carbonetto, P. and Stephens, M., 2021. Flexible Signal De-noising via Flexible Empirical Bayes Shrinkage. *Journal of Machine Learning Research* 22 1-28.