Artifact of Detecting Biomarkers Associated with Sequencing Depth in RNA-Seq

by

RuoFei Yin

BE, Tianjin Medical University, 2021

Submitted to the Graduate Faculty of the

School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2023

UNIVERSITY OF PITTSBURGH

SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

RuoFei Yin

It was defended on

April 21, 2023

and approved by

George C Tseng, ScD, Professor, Department of Biostatistics, Department of Human Genetics, Department of Computational & Systems Biology

Jenna C Carlson, PhD, Assistant Professor, Department of Biostatistics, Department of Human Genetics

Kang-Hsien Fan, PhD, Research Assistant Professor, Department of Human Genetics

Thesis Advisor: George C Tseng, ScD, Professor, Department of Biostatistics, Department of Human Genetics, Department of Computational & Systems Biology

Artifact of detecting biomarkers associated with sequencing depth in RNA-Seq

RuoFei Yin, MS

University of Pittsburgh, 2023

RNA-Seq is a highly sensitive and accurate sequencing technique that uses next-generation sequencing (NGS) technology to reveal the presence and quantity of RNA in a biological sample at a given moment, which is useful for studying the behavior of genes under different biological conditions.[1,2] An essential step in an RNA-Seq study is normalization, in which raw data are adjusted to account for systematic technical biases such as library size and transcript length.[3] Multiple popular normalization methods have been proposed and widely used, including counts per million (CPM), transcripts per million (TPM) and reads per kilobase million (RPKM). Although systematic experimental bias and technical variation are expected to be eliminated after normalization, we surprisingly found a large proportion of genes associated with library size in human post-mortem striatum normalized RNA-seq data. In this thesis, we confirmed the universal existence of this problem by systematically examining 159 Gene Expression Omnibus (GEO) datasets and 24 of The Cancer Genome Atlas (TCGA) datasets. We conducted a simulation study to rule out potential causes from count data quantification and examined a potential solution to correct the artifact based on a Poisson model with variable rates for different nucleotide patterns from a previous publication. We reproduced the results of this paper and applied this published model to these data to see if the library size affected the regression. We performed linear regression analysis on the model coefficients and library size, which did not show evidence of an association. Thus, for a future direction, we plan to replace this Poisson model with a negative binomial model which may improve the model fitting and develop as a solution to correct the artifact. If successful, the new normalization will improve association analysis and biomarker detection in basic and clinical studies of diseases.

Public health significance: Limited amount of research has been focused on the artifact of the biomarkers associated with sequencing depth in normalized RNA-Seq datasets, which should be corrected to improve accuracy in downstream translation research. This paper tries to figure out this artifact.

Table of Contents

List of Tables

List of Figures

Preface

I am heartily thankful to my thesis advisor, Dr. George C Tseng and my academic advisor, Dr. Jeanine M Buchanich, for all the encouragement, guidance and support. They are wonderful advisors through my master's years at University of Pittsburgh. I also owe a deep sense of gratitude to my committee members, Dr. Jenna C Carlson and Dr. Kang-Hsien Fan, for agreeing to serve on my thesis committee and for their generous help and support on this thesis. I would like to thank Dr. Colleen McClung for the permission to use the dataset for this thesis. I would also like to express my gratitude to all faculty members, staff and students from the Department of Biostatistics of Pitt Public Health, who gave me tremendous help and support along my way of academic work and seeking further studies. I am thankful to Pitt for giving me this opportunity to join this warm family and have a wonderful time.

1.0 Introduction

RNA-Seq has become a widely used technology for transcriptome analysis and has gradually replaced traditional microarray, due to its low cost and the ability to provide a comprehensive and quantitative view of gene expression.[2] In a typical RNA-seq experiment, mRNA samples are prepared, fragmented, and reverse transcribed to cDNA, which is then sequenced using high-throughput sequencing technologies. The resulting reads are mapped to a reference genome or transcriptome to provide a quantitative measure of gene expression levels. This powerful tool for studying gene expression offers a higher dynamic range than microarrays, making it suitable for the detection of low-abundance transcripts. [4,5] Furthermore, RNA-seq does not depend on genome annotation for prior probe selection, so non-model or novel organisms can also be sequenced without having a reference genome.[6] These advantages have contributed to the growing popularity of RNA-seq and have led to a reduction in its overall cost, making it an increasingly attractive option over microarrays.

RNA-seq data can be affected by several sources of technical variability, such as sequencing depth (i.e., library size) and transcript length. Inappropriate handling of those variabilities could lead to potential biases that can impact downstream analysis. Thus, an essential step in an RNA-Seq study is normalization, in which raw count data are adjusted to account for factors that prevent direct comparison of expression measures.[7] Several normalization methods have been developed to mitigate these technical biases and to enable accurate comparisons of gene expression levels across samples. Counts per million (CPM) is a simple method for normalizing gene expression data by accounting for differences in sequencing depth (library size) under the assumption that total mRNA is consistent across samples.[8] It scales the read counts for each gene by the total number of reads in the sample, and then multiplies by a million to enable comparisons across samples. There are also several more advanced normalization methods such as transcripts per million (TPM), reads per kilobase million (RPKM) and trimmed mean of M-values (TMM). TPM accounts for differences in both library size and gene length. First, it divides the read counts by the length of each gene in kilobases to obtain the reads per kilobase (RPK) values, and then the "per million" scaling factor is calculated by summing up all the RPK values in a sample and dividing this number by a million. Finally, we calculate TPM by dividing the RPK values by the "per million" scaling factor. RPKM is very similar to TPM, the only difference is the order of operations. TMM uses a weighted trimmed mean of the log expression ratios between samples, which is a method for normalizing gene expression data that accounts for differences in library size and RNA composition. The above are common normalization methods, allowing for more accurate comparisons of gene expression levels between samples. Thus, we assume that after normalization, the unwanted technical effect will be eliminated from the RNA-seq data.

However, we found the library size effect on the gene expression is not completely removed after CPM normalization in most datasets we verified. This problem first came to light from analysis of multiple in-house postmortem brain tissue data and cancer datasets, where we performed gene-by-gene association analysis with clinical and technical variables after CPM normalization and surprisingly found through linear regression that thousands of biomarkers remained significantly correlated with library size. To investigate further, we expanded our analysis to publicly available datasets, including 159 Gene Expression Omnibus (GEO) and 24 The Cancer Genome Atlas (TCGA) datasets and systematically confirmed the universal existence of this problem. One hypothesized reason for this problem is that CPM assumes most genes are not differentially expressed between samples, which is not proper for some datasets. So, we further investigated other advanced normalization methods, such as TPM, RPKM and TMM, which take into account the distribution of expression levels across genes and adjust for differences in library size and other sources of variation simultaneously. However, this problem existed with all these normalization methods. Thus, we conclude that this problem is universal no matter what datasets and what normalization methods we use. The paper is structured as follows. In Section 2.1, we will briefly describe our initial finding of this problem from a motivating dataset, followed by the empirical evaluation among publicly available datasets, including 159 GEO and 24 TCGA datasets in Section 2.2. Section 3 presents simulation studies to examine a potential cause using count data quantification, which lets us exclude this as a potential cause. Section 4 evaluates and applies a Poisson model in RNA-seq to seek potential solutions to correct the artifact as well as its applications. A conclusion and future directions are included in Section 5. Although the current Poisson model does not provide a solution to correct the artifact, we discuss a future direction of negative binomial modeling to possibly correct for the bias.

2.0 Empirical Evaluation of the Artifact

In this section we describe the dataset that motivated us to investigate the problem. In addition, we also verify the universal existence of this problem among publicly available RNA-seq datasets.

2.1 Initial Finding: A Motivating Dataset

We first encourtered this problem in a homo sapiens post-mortem striatum RNA-seq dataset, which contains 3 brain regions: caudate, putamen, and nucleus accumbens (NAc). After a standard preprocessing pipeline (fastQC-Hisat2-HTSeq), we obtained a count matrix of 30338 genes and 116 samples in each brain region. We then normalized the count to log2 continuous counts per million (CPM), where the effect of library size should have been normalized. However, unexpectedly, we found that library sizes were still significantly correlated with normalized gene expression with 13,925 genes (45.9%) having q-value < 0.05. This motivated us to investigate if any other of our in-house datasets had similar issues, and we found that in 8 out of 12 in-house postmortem brain tissue and cancer datasets from collaborators, the proportion of genes significantly related to library size exceeded 20%, as shown in Table 1. Q-value was calculated by Benjamini-Hochberg correction.

Table 1. Significant Gene Results of 12 Datasets from Collaborators.

Dataset	Sample size	Number of significant genes (q value < 0.01)	Number of significant genes (q value < 0.05)	Biomarkers proportions (q-value < 0.05)
NAc_mouse(Darius)	53	3619	6921	30.04%
Human_OUD	80	5938	10091	33.26%
Jian_BRCA	54	4	45	0.10%
Kyle_Caudate	114	12141	15343	50.57%
Kyle_NAc	113	4559	7825	25.79%
Kyle_Putamen	114	10056	13869	45.71%
Lauren_D1D2	96	16313	17724	76.93%
Lauren_STAR	96	16092	17521	76.67%
Lauren_PFC_NAc	53	240	579	2.51%
Mouse_FCG_NAc	48	0	0	0.00%
Mouse_FCG_PFC	48	0	0	0.00%
Mouse_morphine	62	4224	6292	27.31%

2.2 Large-Scale Empirical Evaluation of Public Datasets

In order to systematically confirm whether this issue also exists pervasively in publicly available RNA-seq datasets, we checked GEO and TCGA datasets.

For GEO, we established specific inclusion criteria to select GEO datasets whose samples were from “homo sapiens” (organism), “expression profiling by high throughput sequencing” (type), “Illumina” (platform), and with the number of samples ranging from 100 to 300. In Step 1, we searched datasets on the National Library of Medicine Gene Expression Omnibus website (https://www.ncbi.nlm.nih.gov/geo/) and conducted preliminary screening by activating the following filters existed on the website: “Homo sapiens”, “Expression profiling by high throughput sequencing”, “Illumina”, and sample count from 100 to 300. Through this screening step, we obtained 1,286 GEO datasets. We, however, found that not all of them perfectly meet our criteria. In Step 2, we performed a further screening to remove series that had more than one organism type and for which the type was expression profiling by array; we removed these datasets from consideration, retaining a total of 549 series. In Step 3, we conducted the final filtering manually. Out of these 549 GEO series, we found 198 of them only provided normalized datasets instead of raw count datasets, 141 had decimals in their raw count data, 38 did not provide sufficient supplementary data files, and 13 datasets where raw count data was obtained from multiple platforms. After filtering out these datasets, we finally narrowed down to 159 datasets for our empirical evaluation. Out of these datasets, 97 of them detected more than 20% genes with significant association with library size under q-value<0.05.

With regards to TCGA, we obtained the raw gene count dataset for 24 cancer types from GSM1536837, with a total of 9,264 tumor samples. We excluded 4 datasets with a sample size less than 100, leaving 20 cancer remaining. Of these 20 datasets, 14 of them detected more than 20% of biomarkers with significant association with library size (q-value<0.05). To investigate why some of the studies did not have many significant associations with library size, we examined the genome-wide correlation of expression levels of every pair of samples within each study. Box plots of the pairwise correlations for each TCGA cancer type (y-axis) are shown in Figure 1 and cancer types are ordered by the percent of significantly associated genes. The top 10 cancer types are ordered by the percent of associated genes (55-65%; labeled in red) also had the highest pairwise correlation. In contrast, the other 10 cancer types had significantly lower pairwise correlation among samples (p-value=0.000173), showing larger heterogeneity (biological variation) that may have reduced the power to detect association with library size. In conclusion, after examining RNA-seq datasets in these public repositories, we were able to confirm the universal existence of the problem.

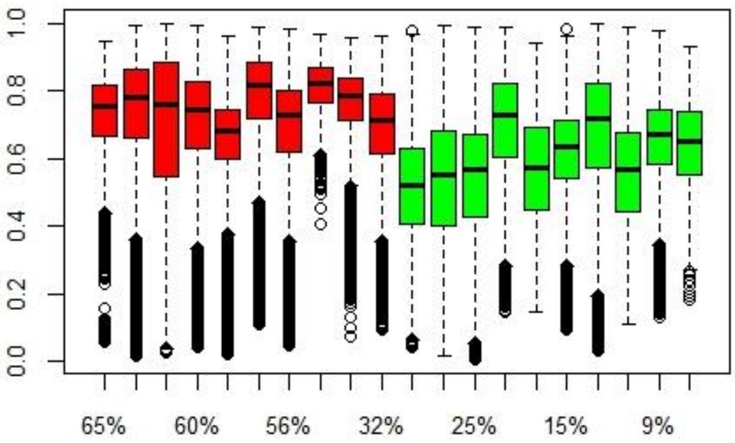

Figure 1. Genome-Wide Correlation Across TCGA Cancer Dataset.

Red: the cancer types with the highest percent of associated genes; Green: the cancer types with the lowest percent of associated genes.

3.0 Using Simulation to Examine Potential Causes from Count Data Quantification

In this section, we design a simulation from real TCGA data and describe its rationale, simulation settings and results.

3.1 Rationale and Settings of Simulation

One possible source of the observed artifact of biomarkers being associated with library size is the expression quantification of count data. Literature has shown that the quantification process from count data to a continuous measurement can lose data information, especially for low expression genes.[9] As each sample may have different genome-wide expression distribution in read data, we randomly simulated a count data matrix for samples based on the expression distribution of one selected sample (i.e., the one with the highest or the lowest library size). If the artifact association did not exist in the simulated data, we could conclude that this problem is not caused by count data quantification.

To perform the simulation, we followed these steps:

1) For each cancer type, choose the 2 samples that have the largest and the smallest library sizes.
2) Calculate the proportions of expression counts of each gene (i.e., raw count of a given gene divided by the total counts of all genes in a sample).
3) Using the vector of proportions, simulate a vector of counts for genes in a sample by multinomial distribution (using bootstrapping to resample the original library size). Repeat

this to simulate 100 samples with the same distribution (perfect sequencing) but varying the library size.

4) Perform simple linear regression between the simulated counts and their library sizes to identify the artifact (genes associated with library size) in the simulated datasets.

3.2 Results

For all 20 of the TCGA cancer datasets, the proportion of genes that was significantly associated with library size was less than 0.5%. The simulation result shows that the artifact association is unlikely to be caused by count data quantification. Thus, we next explore the possibility that the sequencing bias associated with library size is caused by the sequence content of each gene.

4.0 Poisson Model to Correct the Artifact

In this section, we reproduce the results of a reference paper that attempted to reduce the biases in gene expression estimates due to the non-uniformity of read rates by building Poisson model to produce the sequencing preference parameters. Then, we applied this Poisson model to the in-house datasets to see if the coefficients of the models varied based on library size and identified potential methods to correct the problem.

4.1 Reproduce Poisson Model in a Reference Paper

Each position/nucleotide j in gene i has a count, which is the sum of the reads in which mapping starts at that nucleotide. These counts are obtained by summing the reads in which mapping starts at that specific nucleotide. However, to make efficient use of this data, it is essential to have a suitable model in statistics that accurately reflects the underlying biological processes. Previous analysis methods have relied on a Poisson model with constant rate, which assumption is that counts from a particular isoform are sampled independently from a Poisson distribution with a rate that is proportional to gene expression level. This assumption may not hold in all cases, and a more sophisticated statistical model may be necessary to fully capture the complexity of the underlying biological processes. [10,11]

However, Li et al. (2010) found that this model was inadequate in fitting real data, They developed a more sophisticated model that considers a Poisson model with variable rates to improve the modeling of the counts; that is, the Poisson model with different rate(mean value) models the counts from an isoform.[12] And they also found that the rate of Poisson not only depends on the gene expression level, but also the nucleotide of the read by examining the similarities among counts of different tissues. Hence, they designed the model which rate is the product of gene expression level and the ‘sequencing preference’ of reads that start at that position. This sequencing preference is a factor that indicates the likelihood for a read being generated at a particular position. We took inspiration from this Poisson model in this paper and tried to apply this model to our in-house data to see if the library size of each sample will affect the coefficients of this Poisson model intuitively. Thus, we first reproduced the results of this paper and figured out the whole pipeline.

4.1.1 Datasets and Overdispersion

This paper refers to three datasets. Table 2 lists the basic information of these datasets. Each dataset includes sub-datasets, and there are 8 sub-datasets in total: three (tissues) for Wold data, three (groups) for Burge data, and two (cell lines) for Grimmond data. In this paper, the sub-datasets mentioned above were analyzed separately.

Table 2. Basic Information of the Datasets.

Dataset	Tissues/cell lines	Data Sources	Organism	Platform
Wold	Brain(w1)	SRA001030	mouse	Illumina’s Solexa
	Liver(w2)			
	Skeletal muscle(w3)			
Burge	adipose, brain, and breast(b1)	GSE12946	human	Illumina’s Solexa
	Colon, heart, and liver(b2)			

	Lymph node, skeletal muscle, and testes(b3)			
Grimmond	Embryoid bodies (g1)	GSE10518	mouse	ABI's SOLiD
	Embryoid stem (g2)			

First, the count data was obtained from the original datasets. We downloaded the annotation and sequences of RefSeq genes (mouse mm9 and human hg38) from the UCSC genome browser website. We got the sequence and the exact position for each gene based on the gene gtf file and whole genome Fasta file. Then, we mapped the nucleotide reads to every position of every gene and counted the number of reads whose mapping starts at each position of genes, allowing 2 mismatches. In addition, this model only involved the top 100 genes exhibiting the highest expression levels, which all other genes are not considered. These counts of the selected gene were exclusively employed when building the model. Because a significant proportion can be accounted for by the selected genes, they can provide ample information for determining sequencing preferences. We used the Wold liver data as an example. In the top 100 genes, the average sequencing depth is 44, that is, each position has 44 reads whose mappings start with it. We also randomly sampled the reads so that the sequencing depth becomes 20, 10, 5, 1 or 0.1, and then applied the Poisson model to them to evaluate the impact of sequencing depth. Figure 2 shows that while the R^2 increases as the sequencing becomes deeper, it changes little when sequencing depth is over 10. Thus, the high R^2 in the top 100 genes shows that the information in the top 100 genes is enough to train a good model, and we do not need to include more genes, which would lengthen the computational time.

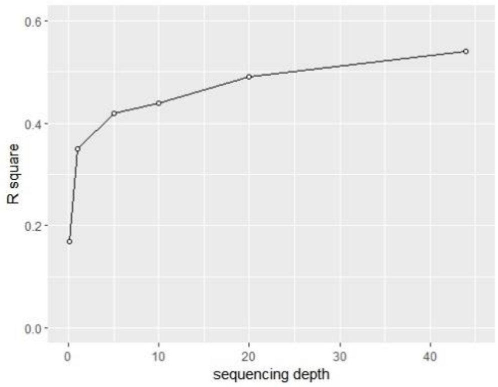

Figure 2. R^2 Corresponds to Different Sequencing Depths.

There are two clear indications that the counts do not adhere to the constant-rate Poisson model. Firstly, the data exhibits a significant degree of over-dispersion. Table 3 presents the relevant values of the variance-to-mean ratios in the top 100 genes of each sub-dataset, which should equal 1 if there is no overdispersion. Second, the count pattern of counts are consistent across various sub-datasets within the same datasets. Figure 3(A) shows the counts in the gene *APOE* of all three tissues of the Wold data, which is the original result in the paper, Figure 3(B) shows the count in the gene *FTH1* which was reproduced by us. This observation also applies to other genes in the Wold dataset as well as the genes in the other 2 main datasets. Hence, this provides compelling evidence that the counts for different positions within the same gene are not sampled from an identical distribution.

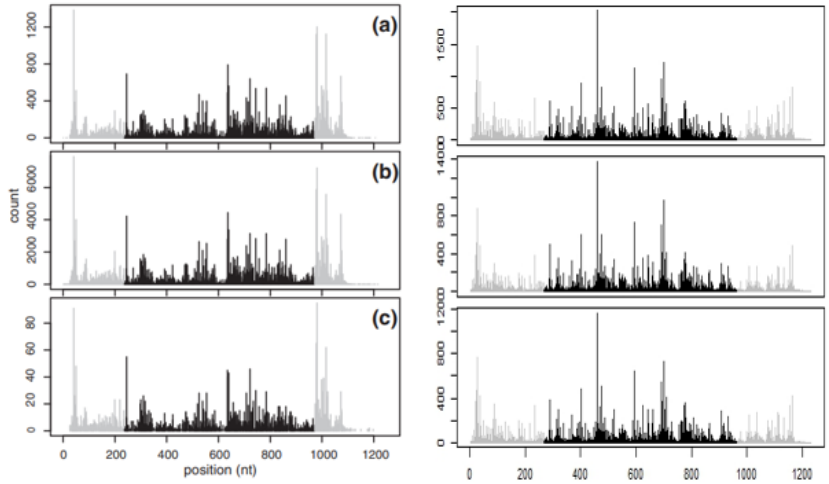

Figure 3. The Wold Data Includes Read Counts along a Gene in Various Tissues.

(Left panel) APOE gene (Right panel) FTH1 gene. The count of reads starting at each position is represented by every vertical line . Nt: nucleotides.

Table 3. Variance-to-Mean Ratios in Different Datasets.

The values in the parentheses are the results we reproduced, and the values in front of them are the original values in the paper.

Dataset	**Sub-dataset**	**Variance-to-mean ratios**		
		Maximum	**Median**	**Minimum**
Wold	w1	248(224)	36(33)	21(34)
	w2	1503(1633)	48(54)	19(32)
	w3	2088(2064)	34(33)	18(15)
Burge	b1	835(789)	78(60)	14(14)
	b2	1187(1233)	102(140)	28(20)
	b3	1593(1542)	112(134)	20(24)
Grimmond	g1	24385(24384)	806(800)	47(58)
	g2	9162(9162)	355(345)	22(21)

4.1.2 Poisson Linear Model and Performance

This paper developed a model for the distribution of the count of reads initiating at nucleotide j of gene I (treated as n_{ij}), which is dependent on the expression (treated as μ_i) and the nucleotide sequence around a particular nucleotide (with a length of K) is indicated as $b_{ij1}, b_{ij2}, \dots, b_{ijK})$. We assume n_{ij}~Poisson (μ_{ij}), where μ_{ij} is the mean(rate) of the Poisson distribution, and μ_{ij}=$\omega_{ij}\mu_i$, where ω_{ij} is the sequencing preference. This dependency on the neighboring sequence may help mitigate the bias in gene expression caused by the non-uniformity of the counts:

$$\log(\mu_{ij}) = v_i + \alpha + \sum_{k=1}^{K} \sum_{h\in\{A,C,G\}} \beta_{kh} I(b_{ijk} = h)$$

where v_i= log(μ_i), α is a constant term, I(b_{ijK}=h) equals 1 if the k^{th} nucleotide of the neighboring sequence is h, and 0 otherwise, and the coefficient of the impact of occurrence of the letter h in the k^{th} position, denoted as β_{kh}. We incorporated the 40 nucleotides preceding the first nucleotide of the reads, as well as the 40 nucleotides following them. Thus, this model uses 3*80=240 parameters (β_{kh}) to model the sequencing preferences (3 represents base A, C, and G; base T was treated as the reference). We followed these steps to build the model:

1) Initialize $\widehat{v_i}$ = log [$\sum_{j=1}^{L_i} n_{ij}/L_i$], where L_i is the length of gene i.

2) Assuming the offsets v_i =$\widehat{v_i}$ are known, the Poisson model can be fitted to get $\hat{\alpha}$ and $\widehat{\beta_{kh}}$.

3) Update $\widehat{v_i}$=log [$\sum_{j=1}^{L_i} n_{ij}/W_i$], where W_i is the sum of sequencing preferences of all nucleotides of gene *i*, that is, W_i=$\sum_{j=1}^{L_i}$ exp[$\hat{\alpha} + \sum_{k=1}^{K} \sum_{h\in\{A,C,G\}} \beta_{kh} I(b_{ijk} = h)$].

4) Repeat steps 2-3 until the deviance decreases by less than 1%.

We applied this model to each of the 8 sub-datasets and use R^2 to measure the goodness-of-fit. We define:

$$\boldsymbol{R^2 = 1\text{-}d/\,d_0}$$

Here, d represents the deviance of the Poisson model, while d_0 represents the deviance of the null model, which is the original naïve model assuming identical sequencing preference. In the Poisson model, deviance is used as a measure of the goodness-of-fit of the model, rather than the variance. This is because the Poisson distribution assumes that variance-to-mean ratio is 1, and the deviance takes this into account. The deviance here is a measure of the difference between the observed data and the fitted model. It is calculated as the difference between the log-likelihood of the fitted model and the log-likelihood of the saturated model which is a model that perfectly fits the data. The final R^2 values we achieved are listed in Table 4. Approximately 40 to 50% of the variance can be accounted for by this linear model, in broad terms.

Table 4. Goodness-of-Fit of the Poisson Model Across Different Datasets.

[a]The lengths of the surrounding sequences we consider; The values in the parentheses are the results we reproduced, and the values in front of them are the original values in the paper.

Dataset	Sub-dataset	R^2(80 nucleotides[a])
Wold	w1	0.52(0.46)
	w2	0.51(0.49)
	w3	0.48(0.54)
Burge	b1	0.43(0.32)
	b2	0.37(0.32)

	b3	0.45(0.39)
Grimmond	g1	0.47(0.46)
	g2	0.45(0.45)

Figure 4 shows all original results of coefficients in the Poisson linear model in the reference paper. Figure 5 shows the results we reproduced. Each sub-dataset is plotted separately to aid in visualization. In general, the coefficients located in the central part possess larger absolute values compared to those on either side, where they tend towards zero. This demonstrates that the nucleotides in the vicinity of the first position of a read have a more substantial impact on the sequencing preference. We provide an example of how to interpret these coefficients. For instance, in the Wold brain data, the coefficient for C at the first position (represented by the blue rectangle at position 0 in panel a) is 0.81. This indicates that replacing the nucleotide T with C would increase the sequencing preference by a factor of $e^{0.81}$ is 2.25.

The coefficients exhibit remarkable similarity across each sub-dataset. This Poisson linear model shows that 32 to 54% of the non-uniformity can be explained by the sequence difference, which can give better estimators for the downstream analysis of RNA-Seq data.

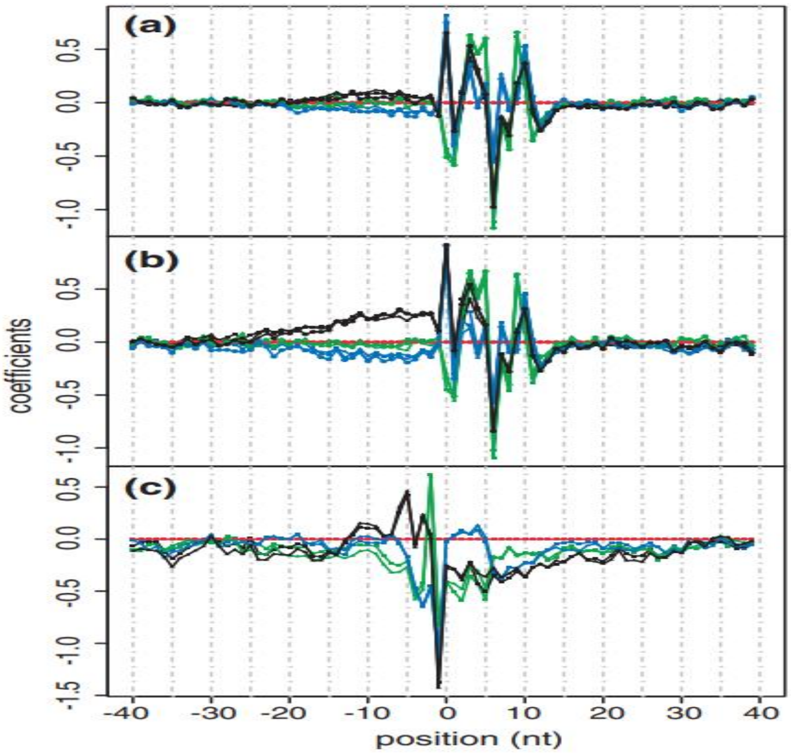

Figure 4. Coefficients of Poisson Linear Models in Different Datasets.

Color: red, T; green, A; blue, C; black, G. (a) Coefficients in the Wold data. (b) Coefficients in the Burge data. (c) Coefficients in the Grimmond data. Nt: nucleotides.

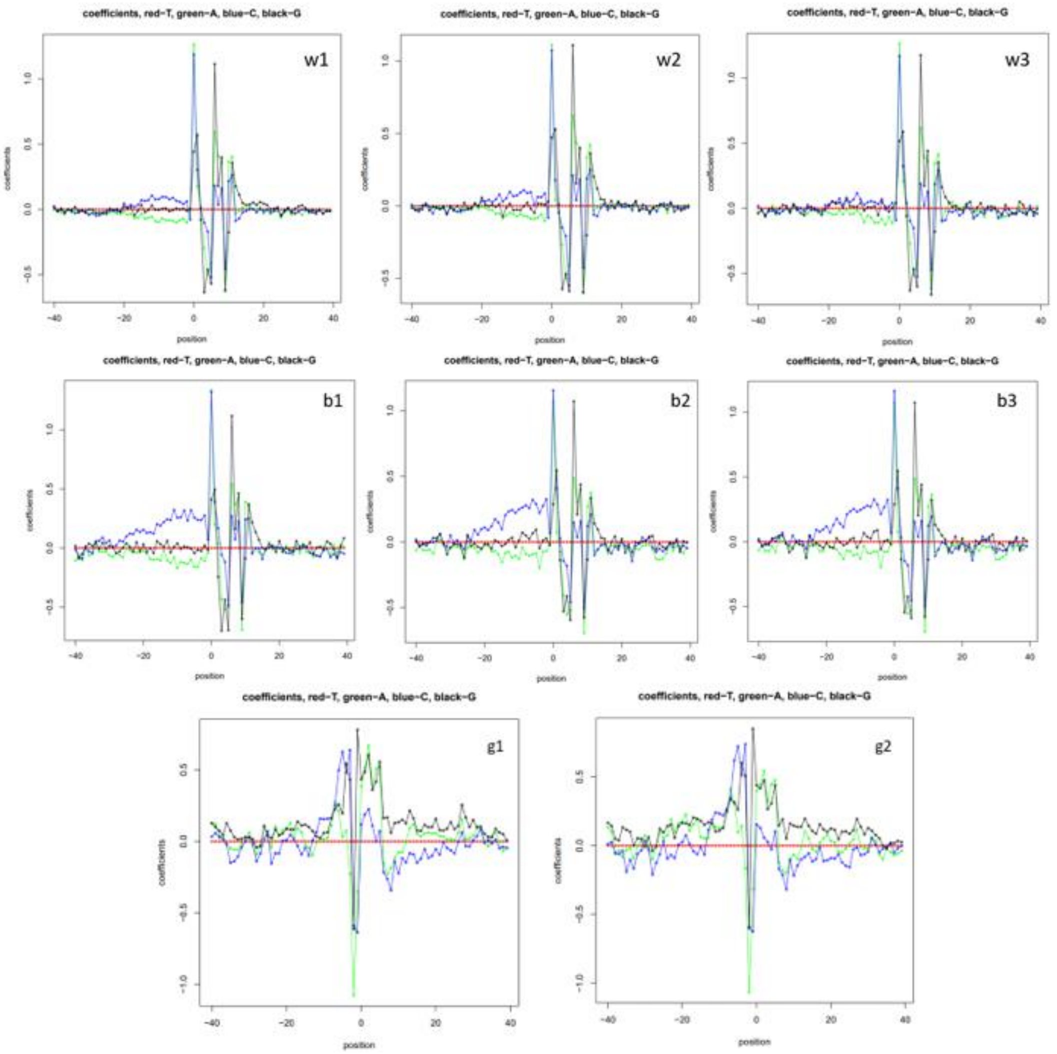

Figure 5. Reproduced Coefficients of the Poisson Models in Different Sub-Datasets.

4.2 Extended Poisson Model for Better Normalization

We reproduced the results of the paper and applied this model to our own data to see if the library size of each sample will affect the coefficients of this Poisson model intuitively.

We chose 40 bam files of mouse provided by Colleen A McClung (the 20 with the largest library size and 20 with the smallest library size), GTF file and FASTA file of mouse provided by

UCSC to get the count for each position in the top 100 genes with the highest gene expression level. Finally, we got 40 datasets for 40 samples. Each dataset has 4 columns with names: index, tag, seq and count. “index” is an index for the gene from where this count comes. “tag” is an integer value, 0 means to consider this count. In our datasets, -2 means the UTR part, and -1 means the further 100 bp. “seq” is the nucleotide of this position and it must be capital T or A or C or G. “count” is the count of reads starting at this position. After getting the datasets, we used these datasets to build 40 Poisson models and for each model, there are 3*80=240 coefficients such as pM40A, pM40C, pM40G, pM39A, ……, p0A, p0C, p0G, ……, p40A, p40C, p40G, where p means position, M means minus. Figure 6 shows the coefficients of the Poisson linear models across different samples.

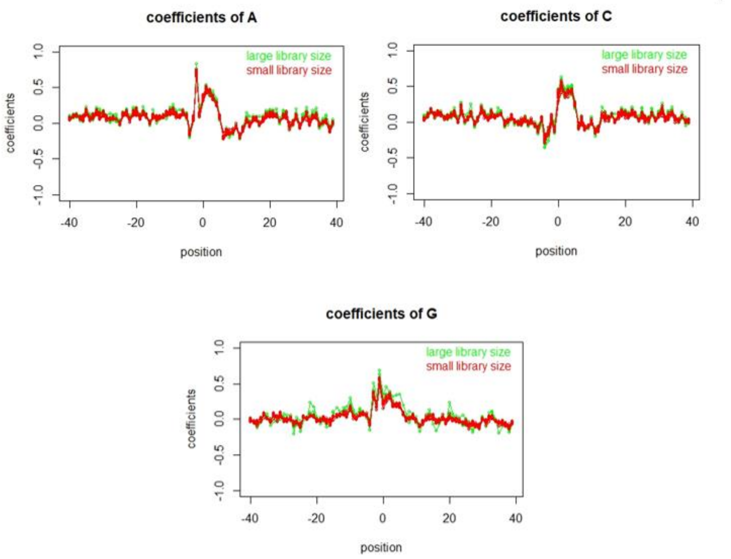

Figure 6. Coefficients of Poisson Models in 40 Datasets for Base A, C and G.

Color coding for 2 groups separated by library size: green, 20 samples with largest library size; red, 20 samples with smallest library sizes.

We can not tell if the coefficients are related to the library size directly from this figure directly. Thus, we performed the linear regression on these coefficients to see if it is affected by different sample library sizes. For base A, we have 40 coefficients for 40 samples in each position, so we constructed 80 simple linear regression models for the 80 positions. Table 5 shows the results of these linear regression models for bases A, C, and G separately (T was treated as the reference base).

Therefore, we can conclude that the coefficients are not related to the library size significantly (p value > 0.05) and we need to read more reference papers to find other potential solutions.

Table 5. The Results of Linear Regression Model.

Base	**Number(proportions) of models which p value of library size < 0.05**	**Mean of p value**	**Variance of p value**	**Median of p value**
A	3(3.75%)	0.34	0.07	0.40
C	1(1.25%)	0.49	0.11	0.56
G	0(0.00%)	0.51	0.06	0.42

5.0 Conclusions and Future Direction

In this paper, we found a problem that thousands of biomarkers remained significantly correlatied (q value < 0.05) with library size through linear regression models after normalization in our collaboration datasets. Then we verified the universal existence of this problem among the publicly available RNA-seq datasets such as GEO and TCGA datasets. We present simulation studies to examine cause from quantification, and then, we found a reference paper which designed a Poisson model to consider the sequencing preference parameters, that is, a factor showing how likely it is for a read to be generated at the position. In order to apply this model in our own dataset while considering the library size effect, we reproduced the results of this paper and cleared the pipeline, which obtained similar results to the reference paper. At the end, we used our own datasets to fit this Poisson model but found that the coefficients in the fitted model were not related to the library size.

In view of this, for the future direction, we will evaluate how much dispersion will be reduced by the fit residual after fitting the non-constant Poisson model. If over-dispersion still exists in this Poisson model, that is, the variance-to-mean ratio of fit residual much larger than 1, then we will consider replacing the Poisson model by negative binomial model, which can accommodate overdispersion by introducing an additional parameter that allows for more variability in the data.

Bibliography

[1] Chu Y, Corey DR. RNA sequencing: platform selection, experimental design, and data interpretation. Nucleic Acid Ther. 2012 Aug;22(4):271-4. doi: 10.1089/nat.2012.0367. Epub 2012 Jul 25. PMID: 22830413; PMCID: PMC3426205.

[2] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009 Jan;10(1):57-63. doi: 10.1038/nrg2484. PMID: 19015660; PMCID: PMC2949280.

[3] Evans C, Hardin J, Stoebel DM. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. Brief Bioinform. 2018 Sep 28;19(5):776-792. doi: 10.1093/bib/bbx008. PMID: 28334202; PMCID: PMC6171491.

[4] Rao MS, Van Vleet TR, Ciurlionis R, Buck WR, Mittelstadt SW, Blomme EAG, Liguori MJ. Comparison of RNA-Seq and Microarray Gene Expression Platforms for the Toxicogenomic Evaluation of Liver From Short-Term Rat Toxicity Studies. Front Genet. 2019 Jan 22;9:636. doi: 10.3389/fgene.2018.00636. PMID: 30723492; PMCID: PMC6349826.

[5] Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS One. 2014 Jan 16;9(1):e78644. doi: 10.1371/journal.pone.0078644. PMID: 24454679; PMCID: PMC3894192.

[6] van der Kloet FM, Buurmans J, Jonker MJ, Smilde AK, Westerhuis JA. Increased comparability between RNA-Seq and microarray data by utilization of gene sets. PLoS Comput Biol. 2020 Sep 30;16(9):e1008295. doi: 10.1371/journal.pcbi.1008295. PMID: 32997685; PMCID: PMC7549825.

[7] Evans C, Hardin J, Stoebel DM. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. Brief Bioinform. 2018 Sep 28;19(5):776-792. doi: 10.1093/bib/bbx008. PMID: 28334202; PMCID: PMC6171491.

[8] Tarazona S, Furió-Tarí P, Turrà D, Pietro AD, Nueda MJ, Ferrer A, Conesa A. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic Acids Res. 2015 Dec 2;43(21):e140. doi: 10.1093/nar/gkv711. Epub 2015 Jul 16. PMID: 26184878; PMCID: PMC4666377.

[9] Ma T, Liang F, Tseng G. Biomarker detection and categorization in ribonucleic acid sequencing meta-analysis using Bayesian hierarchical models. J R Stat Soc Ser C Appl Stat. 2017 Aug;66(4):847-867. doi: 10.1111/rssc.12199. Epub 2016 Dec 16. PMID: 28785119; PMCID: PMC5543999.

[10] Jiang H, Wong WH. Statistical inferences for isoform expression in RNA-Seq. Bioinformatics. 2009 Apr 15;25(8):1026-32. doi: 10.1093/bioinformatics/btp113. Epub 2009 Feb 25. PMID: 19244387; PMCID: PMC2666817.

[11] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008 Jul;5(7):621-8. doi: 10.1038/nmeth.1226. Epub 2008 May 30. PMID: 18516045.

[12] Li J, Jiang H, Wong WH. Modeling non-uniformity in short-read rates in RNA-Seq data. Genome Biol. 2010;11(5):R50. doi: 10.1186/gb-2010-11-5-r50. Epub 2010 May 11. PMID: 20459815; PMCID: PMC2898062.