A NON-PARAMETRIC EMPIRICAL BAYES APPROACH FOR ESTIMATING TRANSCRIPT ABUNDANCE IN UN-REPLICATED NEXT-GENERATION SEQUENCING DATA

Sanvesh Srivastava

R. W. Doerge

Recommended Citation

A NON-PARAMETRIC EMPIRICAL BAYES APPROACH FOR ESTIMATING TRANSCRIPT ABUNDANCE IN UN-REPLICATED NEXT-GENERATION SEQUENCING DATA

Sanvesh Srivastava and R.W. Doerge*
Department of Statistics, Purdue University, West Lafayette, IN 47907

* **Corresponding Author:**
R.W. Doerge
Department of Statistics
Purdue University
250 N. University St.
West Lafayette, IN 47907

e-mail: doerge@purdue.edu
phone: 765-494-6030
fax: 765-494-0558

Abstract

Empirical Bayes approaches have been widely used to analyze data from high throughput sequencing devices. These approaches rely on borrowing information available for all the genes across samples to get better estimates of gene level expression. To date, transcript abundance in data from next generation sequencing (NGS) technologies has been estimated using parametric approaches for analyzing count data, namely – gamma-Poisson model, negative binomial model, and over-dispersed logistic model. One serious limitation of these approaches is they cannot be applied in absence of replication.

The high cost of NGS technologies imposes a serious restriction on the number of biological replicates that can be assessed. In this work, a simple non–parametric empirical Bayes modeling approach is suggested for the estimation of transcript abundances in un-replicated NGS data. The empirical Bayes analysis of NGS data follows naturally from the empirical Bayes analysis of microarray data by modifying the distributional assumption on the observations. The analysis is

presented for transcript abundance estimation for two treatment groups in an un-replicated experiment, but it is easily extended for more treatment groups and replicated experiments.

Keywords: Empirical Bayes, Microarrays, Next-Generation Sequencing, Poisson distribution, Differential Gene Expression.

1. Introduction

NGS technologies have emerged as a promising alternative to previous technologies such as microarrays and Serial Analysis of Gene Expression (SAGE). Researchers have shown that results from NGS technologies are highly replicable with little technical variation (Marioni et al., 2008). Other studies have shown that NGS technologies have an important role to play in future genome related research (Shendure, 2008). RNA-Sequencing is an attractive area of application of NGS technologies (Cloonan et al., 2009). One of the important issues in RNA-Sequencing experiments is the estimation of transcript abundances.

Presently, very little attention has been paid to the estimation of transcript abundances in un-replicated experiments. The main reason for this is lack of reliable statistical inference in un-replicated experiments. But many un-replicated experiments are conducted by biologists for the purpose of surveying an organism, for preliminary analysis, or because of the high cost of NGS technologies. This paper presents an empirical Bayes method for the estimation of transcript abundances in un-replicated experiments.

The transcript abundance in NGS data have been estimated using a classical parametric model – over-dispersed logistic regression model (Baggerly et al., 2004) and also through Bayes and empirical Bayes approaches which model information from all the genes, namely – a Bayesian beta-binomial model (Vêncio et al., 2004) and an empirical Bayes gamma-Poisson model (Thygesen and Zwinderman, 2006). A conditional maximum likelihood approach based on a negative binomial model (Robinson and Smyth, 2007, 2008) has also been used to estimate transcript abundances in NGS data. While these approaches model within group variation to improve the estimation of transcript abundance, this information is missing in un-replicated experiments.

Our method takes advantage of the parallel structure of the NGS data at transcript level to compensate for the missing information about within group variation. It combines information available about transcript abundances from counts at the transcript level as well as counts available for all the transcripts to get better estimates of transcript abundances in an un-replicated

NGS experiment. This paper addresses the issue of estimating transcript abundances, but it is a matter of choice of semantics. The theoretical and practical details remain the same for estimating gene abundances or tag abundances.

2. Non–parametric empirical Bayes model framework

We use the empirical Bayes approach developed in Robbins (1956) to obtain the estimates of transcript abundance. The main difference between existing parametric empirical Bayes approaches and our non–parametric empirical Bayes model is the non–parametric prior distribution imposed on the transcript abundances. Our approach is minimally restrictive in prior assumptions and facilitates flexible and robust estimation of transcript abundances, specifically in absence of replication, when there is a limited availability of data, and when distributional assumptions are hard to verify. The hierarchical Poisson model, shown later (equation 2.2), guarantees the transcript abundance estimates can be calculated from the observed data easily and efficiently.

Let n_{gt} be the observed count of transcript g in the sample with treatment t and θ_{gt} is the expectation of n_{gt}. The library size of a particular treatment group t is defined as the total number of transcripts in the treatment group (and, may not be known *apriori*) and is denoted as $n_{.t}$. The transcript counts depend on the library size, as large library size implies high transcript counts. Our aim is to estimate the transcript abundance, λ_{gt}, which is independent of library size. We normalize θ_{gt} by dividing it by $n_{.t}$ to obtain the transcript abundance, λ_{gt}. Equation 2.1 shows the relationship between λ_{gt}, $n_{.t}$ and θ_{gt}.

$$(2.1) \qquad n_{.t}\lambda_{gt} = \theta_{gt} \qquad \text{where } g = 1\ldots G \text{ and } t = 1,2$$

The statistical model assumes: n_{gt} given θ_{gt} are independent for different transcripts g in a particular treatment group t, n_{gt} given θ_{gt} follows a Poisson distribution with mean parameters θ_{gt}, respectively, and θ_{gt} in a particular treatment group t are assumed to follow a non–parametric distribution $\pi_t(\theta)$, *apriori*. Essentially, we consider the hierarchical Poisson model specification as follows.

$$n_{gt} \mid \theta_{gt} \overset{independent}{\sim} Poisson\left(\theta_{gt}\right)$$

(2.2)
$$\theta_{gt} \overset{i.i.d}{\sim} \pi_t\left(\theta\right)$$

$$g = 1 \ldots G \qquad t = 1,2$$

The empirical Bayes estimate, $\hat{\theta}_{gt}$, of θ_{gt} is obtained in equation 2.3 (Carlin and Louis, 2008, Section 5.3.2, equation 5.30).

(2.3)
$$\hat{\theta}_{gt} = \left(n_{gt} + 1\right) \times \frac{\#\left\{n_{kt} : n_{kt} = n_{gt} + 1, k = 1 \ldots G\right\}}{\#\left\{n_{kt} : n_{kt} = n_{gt}, k = 1 \ldots G\right\}} \qquad g = 1 \ldots G \qquad t = 1,2$$

The hierarchical structure of the model equation 2.2 ensures $\hat{\theta}_{gt}$ is easily and efficiently estimable from the observed data. Equation 2.3 illustrates information borrowing, as $\hat{\theta}_{gt}$ borrows information from all the transcript counts available in the treatment group t in addition to the count of transcript g.

A naïve estimate of the library size $n_{.t}$ for a particular treatment t is

(2.4)
$$\hat{n}_{.t} = \sum_{g=1}^{G} n_{gt}$$

(2.5)
$$\hat{n}_{.t} = \sum_{g=1}^{G} n_{gt} + n_{(G)t} \times G,$$

where $n_{(G)t}$ is the maximum of all the available transcript counts in the sample with treatment t.

Because the estimate in equation 2.4 will underestimate the true library size; instead we will use the estimate in equation 2.5. This is similar to Efron's idea of offsetting the naïve estimate by a quartile of its gene wise value (Efron et al., 2001). In our case, we modify the naïve estimate in equation 2.4 by adding the maximum (100th quartile, $n_{(G)t}$) of all the transcript counts in the sample with treatment t times the number of transcripts (G). We need to multiply $n_{(G)t}$ by G to make the transcript count comparable to the order of library size. The offset corrects for the negative bias of equation 2.4 and improves the estimate $\hat{\lambda}_{gt}$ (equation 2.6).

The estimate of transcript abundance, $\hat{\lambda}_{gt}$, is obtained by normalizing $\hat{\theta}_{gt}$ by the estimate of corresponding library size $\hat{n}_{\cdot t}$ (equation 2.6).

(2.6)
$$\hat{\lambda}_{gt} = \frac{\hat{\theta}_{gt}}{\hat{n}_{\cdot t}}$$

The estimate $\hat{\lambda}_{gt}$ is based on $\hat{\theta}_{gt}$, so it also borrows information from all the transcript counts in treatment t to estimate the abundance of transcript g. Therefore, even if there is a single observation for a transcript in any treatment, the estimate in equation 2.6 uses all the transcript counts available in a particular treatment to calculate $\hat{\lambda}_{gt}$. In addition, because the estimate is obtained using a non–parametric prior, it is more flexible and robust. This is a desirable feature as the available data is limited in un-replicated experiments; in addition, $\hat{\lambda}_{gt}$ is obtained with minimal assumptions and uses all the available data in treatment t. The parametric empirical Bayes estimates may not be robust to deviation of data from prior distributional assumptions. Also, any prior distributional assumptions may be hard to verify in un-replicated experiments.

3. Simulations and Results

The empirical Bayes estimate $\hat{\lambda}_{gt}$ (equation 2.6) gains its robustness and flexibility by borrowing information available along the parallel structure of transcripts for treatment t. This implies the increase in number of transcripts in a sample results in an increase in the reliability of the estimate (due to increased information borrowing). This fact is demonstrated via a simulation study in which the number of transcripts (G) available in a sample increases as 200, 2000, and 20000. Typically, the total number of transcripts in a particular treatment in NGS data is of the order of at least thousands, depending on the diversity of the expressed mRNAs and sequencing time (Robinson and Smyth, 2008). The situation totaling 200 transcripts is included as an extreme scenario, and to assess the performance of empirical Bayes estimates under the limitation of available information to share (for details about information sharing in empirical Bayes methods, please see: Efron, 2003). There are two treatment levels: 1 and 2; treatment 1 is the base-line case.

(3.1)
$$n_{gt} \sim Poisson\left(\theta_{gt}\right)$$

$$\theta_{g1} = L_1 \times z_{g1}$$

$$\theta_{g2} = \begin{cases} L_2 \times z_{g2} & \text{for } g = d+1\ldots G \\ 10 + L_2 \times z_{g2} & \text{for } g = 1\ldots d \end{cases} \quad \text{(differentially expressed transcripts)}$$

(3.2) $\quad z_{gt} \overset{independent}{\sim} \begin{cases} Uniform(a,b) \\ Gamma(\alpha, \beta) \end{cases}$

for $g = 1\ldots G$ and $t = 1, 2$. G varies as 200, 2000, and 20000.

(3.3)

$$z \sim Uniform(a,b) \Rightarrow f(z) = \begin{cases} \dfrac{1}{b-a} & \text{for } a \leq z \leq b \\ 0 & \text{otherwise} \end{cases}$$

$$z \sim Gamma(\alpha, \beta) \Rightarrow f(z) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} z^{\alpha-1} e^{-\beta z}$$

Equations 3.1–3.3 give details about the overall simulation setting. The transcript counts n_{gt} are generated from Poisson distribution with mean parameters θ_{gt} (equation 3.1). We further assume that 10% of the transcripts are differentially expressed to make the simulation study close to reality (even if the aim of the paper is not to study differential expression of transcripts). This is accomplished by making the *first* 10% of the transcripts in treatment 2 to have θ_{g2} higher than the θ_{g1} by 10 unit (equation 3.2). The difference of 10 units is arbitrary; if the difference is increased, the pattern of the estimated abundances agrees more with that of the true abundances. Other details about the effects of choosing the difference between θ_{g2} and θ_{g1} are in the Appendix. In addition to describing the estimation of $\hat{\lambda}_{gt}$ (equation 2.6), we also demonstrate the robustness and flexibility of estimate $\hat{\lambda}_{gt}$ by using two prior distributions on θ_{gt}; one prior is a uniform distribution and the other is a gamma distribution prior. A heuristic exploratory data analysis follows for detecting differentially expressed transcripts in un-replicated experiments.

3.1 Simulation using uniform and gamma prior on θ_{gt}

We simulate the mean parameters (θ_{gt}) of the transcript counts (n_{gt}) from uniform and gamma distribution by first generating z_{gt} from uniform and gamma distribution, respectively and then

scaling z_{gt} by the appropriate library size (equations 3.2–3.3). The distribution specific parameters (a, b, α, and β in equation 3.3) and library sizes (L_1, L_2 for treatment 1 and 2, respectively in equation 3.2) are chosen to increase the number of transcripts with low counts. This was done to make the simulation closer to real life scenario where a large number of transcript counts in NGS data are low; low counts are typically less than or equal to 5. The parameters α, and β of gamma prior were chosen such that the mean and variance of the uniform and gamma prior match. Table 3.1 contains the values and details of the parameters a, b, α, β, L_1, L_2, and d depending on the total number of transcripts (G) in the sample (equations 3.1-3.3). The parameters a, b, α, and β determine the mean and variance of z_{gt} depending on the uniform or gamma prior. L_1, and L_2 control the scaling of z_{gt} depending on the total number of transcripts. These parameters control the value of θ_{gt} generated, which is further used to generate n_{gt}. The first 10% of the transcripts are differentially expressed, which is denoted by the parameter d. The results of the simulations did not change noticeably on choosing different parameters.

Transcripts (G)	(a, b)	(α, β)	L_1	L_2	d
200	(0.0001, 0.001)	(4.5, 8223)	4×10^4	5×10^4	20
2000	(0.00001, 0.0001)	(4.5, 82236)	4×10^5	5×10^5	200
20000	(0.000001, 0.00001)	(4.5, 822367)	4×10^6	5×10^6	2000

Table 3.1: The parameter settings in the simulation depending on the total number of transcripts in the sample (G). The parameters a and b are the range parameters of the uniform distribution of z_{gt} in equation 3.2. The parameters α and β are the shape and rate parameters of the gamma distribution of z_{gt} in equation 3.2. L_1 and L_2 are the library sizes for treatments 1 and 2, respectively. The parameter d denotes the number of differentially expressed transcripts in the sample.

Figures 3.1.1 and 3.1.2 summarize the results of simulation using uniform and gamma priors on θ_{gt} by a smoothed color density representation of the scatter plot of $\log_2$ estimated abundance of transcripts vs $\log_2$ true abundance of transcripts. The smooth scatter plot is obtained through a kernel density estimate (R Development Core Team, 2010). The overall pattern of the scatter plot is captured by the loess curve (in red color) and the line of slope 1 (in black color) denotes the ideal case – when the estimated transcript abundances equal the true transcript abundances. Due to the parameter setting of the simulation study (Table 3.1), the true transcript abundances are high when the total number of transcripts is 200 and the true transcript abundances are low when the total number of transcripts is 20000. This pattern is also observed in Figures 3.1.1 and 3.1.2.

Figures 3.1.1 and 3.1.2 show two key features of the simulation results irrespective of the prior on θ_{gt}. First, it is clear the loess curve is almost parallel to the line of slope 1 for all the three transcript numbers (except one case in treatment 1 when the total number of transcripts is 200 and prior is gamma distribution). This implies that the overall pattern of estimated transcript abundances agrees with the pattern of true transcript abundances, but in most cases the empirical Bayes method overestimates the transcript abundances. Second, we observe the increase in total number of transcripts in the sample results in a precise estimation of transcript abundances and an increase in the agreement with the pattern of their true value. This is demonstrated by the shrinkage of the width of the blue band around the loess curve as the total number of transcripts increase from 200 to 20000 in both the treatments. Due to a small mean and skewness of the gamma distribution of z_{gt}, the loess curve in figure 3.1.2 is not parallel to the line of slope 1 as it is in figure 3.1.1. We also observe the pattern of scatter plots in Figures 3.1.1 and 3.1.2 is diffuse when the total number of transcripts is 200. This is due to limited information sharing between transcripts when the total number of transcripts in the sample is 200.
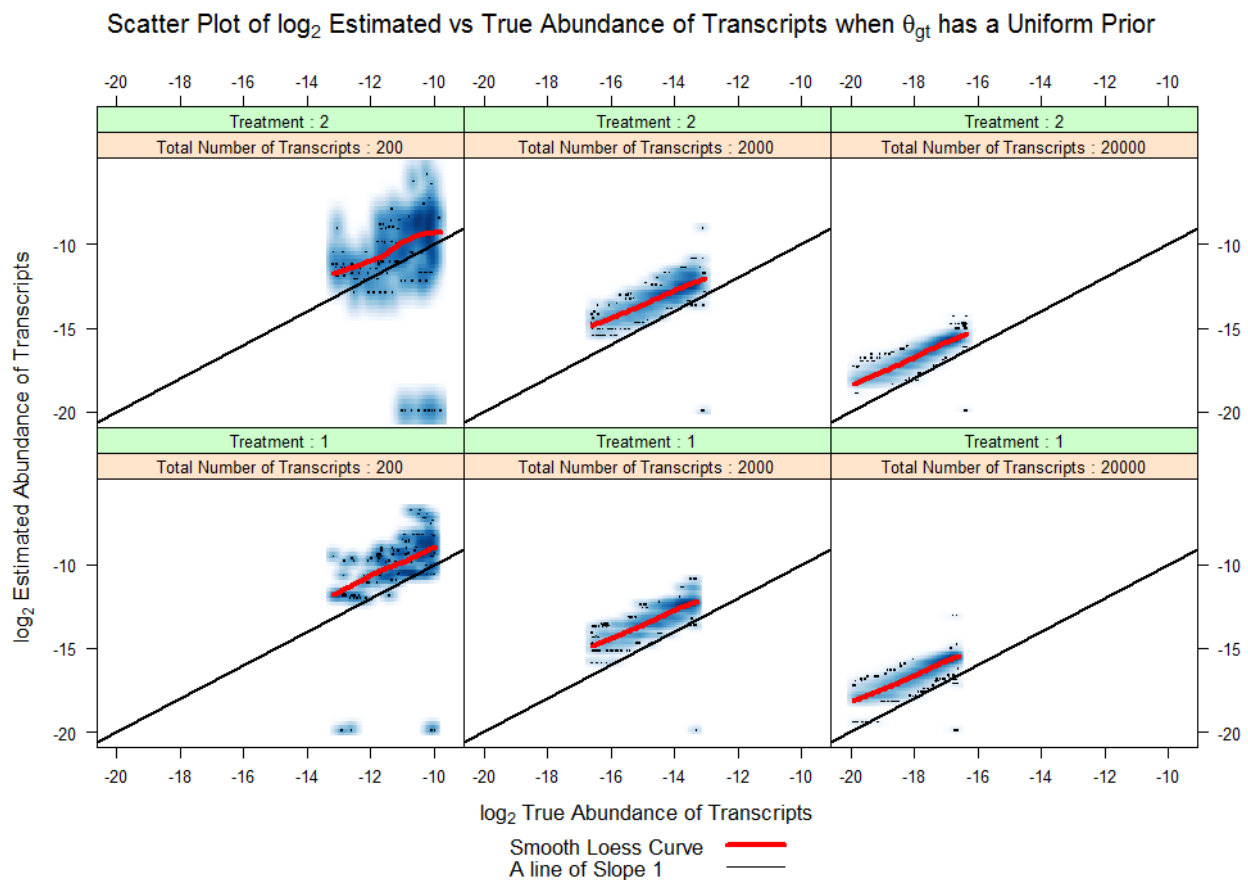

Figure 3.1.1: Smoothed color density representation of the scatter plot of $\log_2$ estimated abundance of transcripts vs $\log_2$ true abundance of transcripts, obtained through a kernel density

estimate. θ_{gt} has a uniform prior. The treatments 2 and 1 are in the first and second row, respectively. The number of transcripts in a sample for a particular treatment vary across the columns as 200, 2000, and 20000. The intensity of the blue color is proportional to the number of points in the region. The superimposed red line in the scatter plot corresponds to the loess curve and the black line corresponds to a line of slope 1. The loess curve is almost parallel the line of slope 1 (except in treatment 2, when the total number of transcripts is 200), implying an overall agreement between the estimated abundance of transcripts and the true abundance of transcripts with a positive bias in most of the cases.
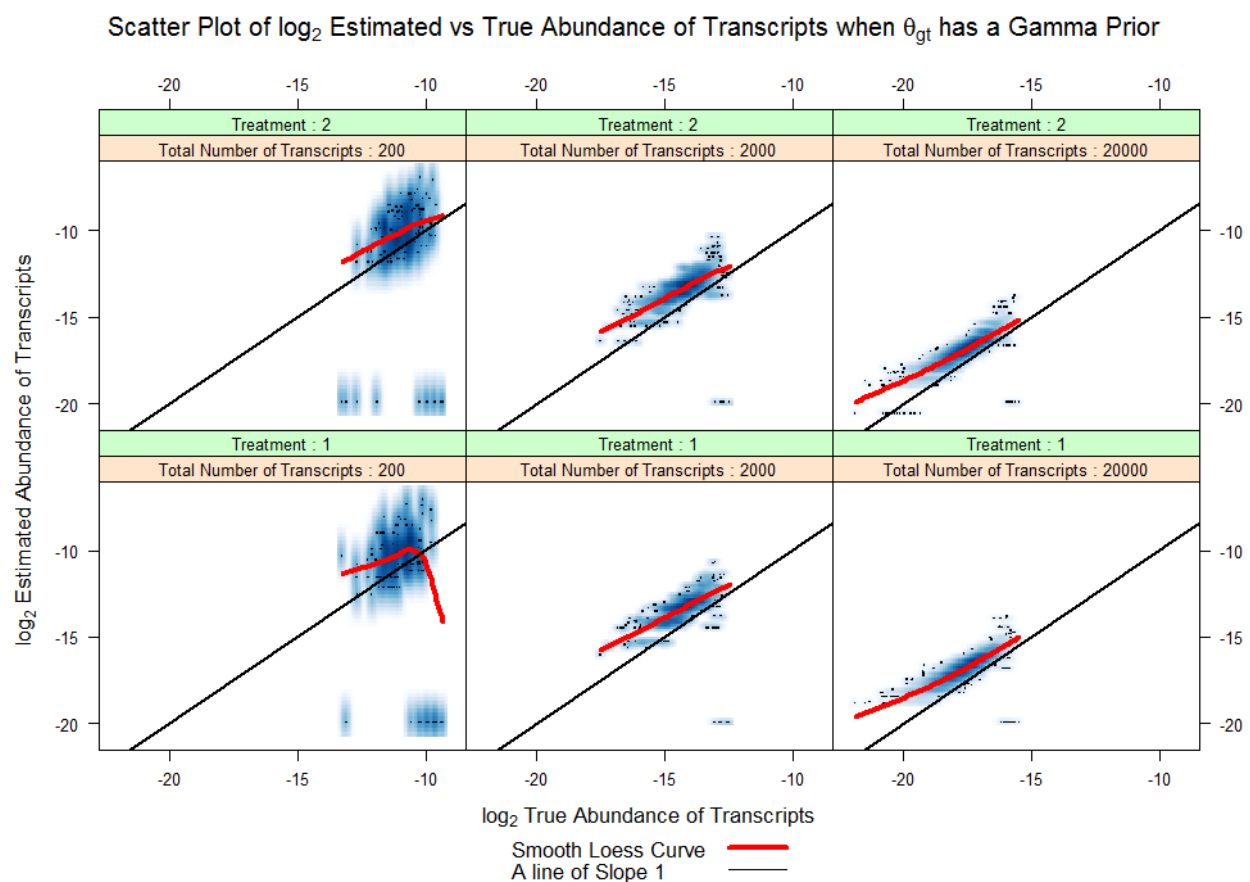

Figure 3.1.2: Smoothed color density representation of the scatter plot of $\log_2$ estimated abundance of transcripts vs $\log_2$ true abundance of transcripts, obtained through a kernel density estimate. θ_{gt} has a gamma prior. The treatments 2 and 1 are in the first and second row, respectively. The number of transcripts in a sample for a particular treatment vary across the columns as 200, 2000, and 20000. The intensity of the blue color is proportional to the number of points in the region. The superimposed red line in the scatter plot corresponds to the loess curve and the black line corresponds to a line of slope 1. The loess curve is almost parallel the line of slope 1 (except in treatment 1, when the total number of transcripts is 200), implying an overall

agreement between the estimated abundance of transcripts and the true abundance of transcripts with a positive bias in most of the cases.

The only major concern in our simulations is the positive bias in the non-parametric empirical Bayes estimates of transcript abundances. Given the limitation of the available data in un-replicated experiments, it is hard to decrease the bias. However, this can be accomplished by better estimation of the library sizes.

The positive bias in empirical Bayes estimates is not a big issue, as biologists are mostly interested in comparisons (contrasts) between two treatments rather than individual treatment effects. Figures 3.1.1 and 3.1.2 show that the positive bias in estimates is almost same in treatments 1 and 2. The theoretical justification for this observation is the majority of the transcript abundances are for non-differentially expressed transcripts and they are estimated from the same prior distribution on θ_{gt}. Therefore, a possible solution to the problem of positive bias is to subtract the $\log_2$ abundances of transcripts in treatment 1 from the corresponding $\log_2$ abundances of transcripts in treatment 2. This eliminates most of the positive bias and the resulting quantity is the $\log_2$ fold change of transcript abundances of treatment 2 with respect to treatment 1. The fold change is a familiar scale for the biologists and hence easier to work with. This idea is further expanded in the next section to perform an exploratory data analysis for the differential expression of transcripts.

3.2 Exploratory data analysis for detecting differential expression

As pointed out before, lack of replication results in unreliable statistical inference as related to differential expression. However, after estimating transcript abundance we an exploratory data analysis can be performed to detect differentially expressed genes in treatment 2 with respect to treatment 1. Figure 3.2.1 illustrates differentially expression when the total number of transcripts in the sample is 20000 and θ_{gt} has a gamma prior. The scatter plot shows estimated $\log_2$ fold change of transcript abundances from treatment 1 to treatment 2 versus transcript number. In the simulation, the first 2000 transcripts are differentially expressed with a positive $\log_2$ fold change. This can be seen in the scatter plot as a band of green and red (extreme left) shifted above the rest of the green band in scatter plot. The differentially expressed transcripts (with higher means) in treatment 2 have higher fold change compared to transcripts which are not differentially expressed. Thus, the differentially expressed transcripts separate out from the un-expressed transcripts by an upward (or downward) shift in general. It is also seen that there are transcripts whose estimated fold change is higher than 2 or lower than 0.5, but they are not differentially

expressed. These are the false positives of the heuristic based analysis of differentially expressed transcripts.
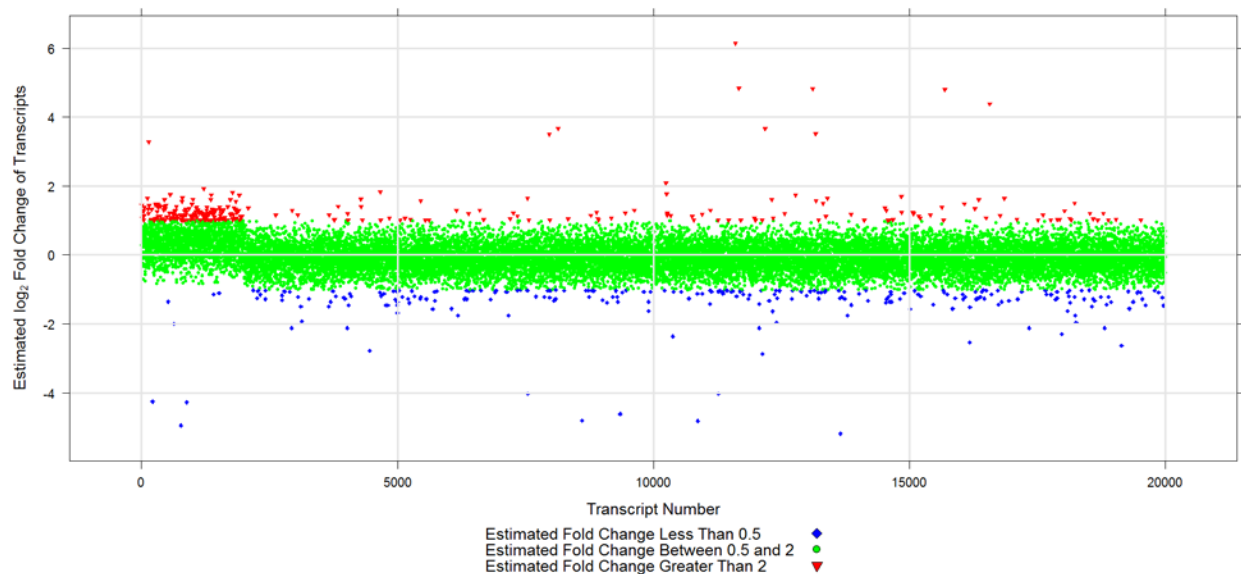

Figure 4.1: Scatter plot of estimated $\log_2$ fold change of transcript abundances from treatment 1 to treatment 2 vs transcript number; the total number of transcripts in the sample is 20000 and θ_{gt} has a gamma prior. The inverted red triangles denote all the transcripts with estimated fold change greater than 2, the blue colored rhombus denote all the transcripts with estimated fold change less than 0.5, and the green colored circles represent all the transcripts with fold change between 0.5 and 2. In the simulation, the first 2000 transcripts are differentially expressed which can be seen in the scatter plot as a band of green and red (extreme left) shifted above the rest of the green band in scatter plot. This is a heuristic to detect differentially expressed transcripts in treatment 2 with respect to 1.

4. Discussion

Due to lack of sufficient information to share, empirical Bayes method did not perform well when the total transcripts is 200 compared to the case with 2000 and 20000 transcripts. Therefore, empirical Bayes methods are advantageous when the number of transcripts in the sample is high.

Statistically, the lack of replication imposes a serious restriction on the detection of differentially expressed transcripts. The non–parametric empirical Bayes method of estimating transcript

abundances can be extended to the detection of differentially expressed transcripts if the number of replicates in a particular treatment group is increased. This will lead to better estimation of within group variation and thus better overall inference about differential expression of transcripts.

NGS data with replicates can be analyzed by treating each replicate as an observation from an un-replicated experiment. We can estimate the transcript abundances using the methods of Section 2 and average the abundances for a particular transcript across replicates to obtain the overall transcript abundances for that treatment group. But, by doing this, we do not model the hierarchy available for the replicates. Statistical inference can be improved by modeling this hierarchy and this is the direction of our future research in this area. The limitation of the analysis to two treatment groups is for illustration purpose, but this method is equally applicable to the case of multiple treatment groups.

5. Summary

We have shown a simple non–parametric empirical Bayes modeling approach for estimating transcript abundances in un-replicated NGS experiments. The method is easy to implement and facilitates robust and flexible estimation of transcript abundances with limited assumptions. The methodology is readily extended to replicated experiments and multiple treatment groups. We have also presented a heuristic to detect differentially expressed transcripts in un-replicated experiments.

Appendix

Here, we analyze the effects of choosing the difference between θ_{g2} and θ_{g1} on the estimation of transcript abundances. These effects were mentioned briefly in Section 3. Figure A.1 summarizes the effects of choosing the difference between θ_{g2} and θ_{g1} as 10 and 100 on the estimation of transcript abundances. The prior on θ_{gt} is a gamma distribution. The number of total transcripts and treatments remain the same as in the simulations (Section 3). We observe as the difference between θ_{g2} and θ_{g1} increases from 10 to 100, the agreement between the loess curve and the slope of line 1 becomes almost close to the ideal scenario – when the estimated transcript abundances equal the true transcript abundances. This observation holds specifically for the differentially expressed transcripts which have higher transcript abundances. These patterns of observations also held for other values of differences between θ_{g2} and θ_{g1} and for uniform prior

on θ_{gt}. The patterns, when the total number of transcripts is 200, are irregular due to limited information sharing among the transcripts.
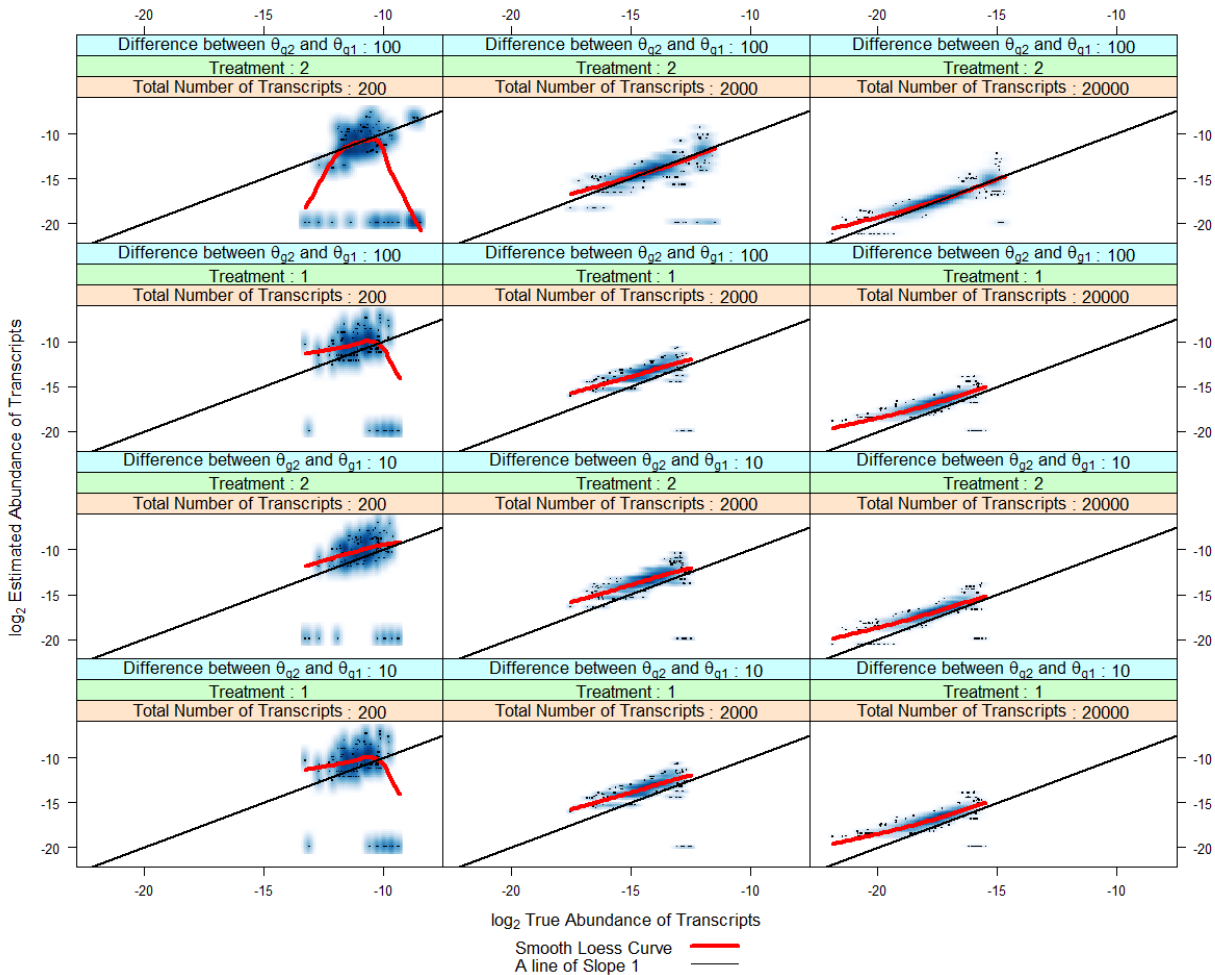

Figure A.1: Smoothed color density representation of the scatter plot of $\log_2$ estimated abundance of transcripts vs $\log_2$ true abundance of transcripts. The first two rows correspond to the difference between θ_{g2} and θ_{g1} of 100 and the next two rows correspond to the difference of 10. The treatments 2 and 1 are in the first and second row respectively for each value of the difference. The number of transcripts in a sample vary across the columns as 200, 2000, and 20000. θ_{gt} has a gamma prior. The intensity of the blue color is proportional to the number of points in the region. The superimposed red line in the scatter plot corresponds to the loess curve and the black line corresponds to a line of slope 1.

Acknowledgement

References

Baggerly, K. A., Deng, L., Morris, J. S., and Aldaz, C. M., (2004). Overdispersed logistic regression for SAGE: Modelling multiple groups and covariates. BMC Bioinformatics 5: 144.

Carlin, B. P., and Louis, T. A., (2008). Bayesian Methods for Data Analysis, Third Edition. Boca Raton, FL: Chapman and Hall/CRC. ISBN 1-584-88697-8.

Cloonan, N., Xu, Q., Faulkner, G. J., Taylor, D. F., Tang, T. P., et al., (2009). RNA-MATE: a recursive mapping strategy for high-throughput RNA-Sequencing data. Bioinformatics 25: 2615–2616.

Efron, B., Tibshirani, R., Storey, J., and Tusher, V., (2001). Empirical Bayes Analysis of a Microarray Experiment, Journal of the American Statistical Association 96, 1151-1160.

Efron, B., (2003). Robbins, empirical Bayes and microarrays. Annals of Statistics **31**, 366–378.

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: 1509–1517.

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Robbins, H., (1956). An Empirical Bayes Approach to Statistics. Proc. Third Berkeley Symp. 1, 152-163, Univ. of Calif. Press.

Robinson, M. D., and Smyth, G. K., (2007). Moderated statistical tests for assessing differences in tag abundance. Bioinformatics 23: 2881–2887.

Robinson, M. D., and Smyth, G. K., (2008). Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics 9: 321-332.

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: 3.

Shendure, J., (2008). The beginning of the end for microarrays? Nature Methods 5: 585–587.

Thygesen, H. H., and Zwinderman, A. H., (2006). Modeling Sage data with a truncated gamma-Poisson model. BMC Bioinformatics 7: 157.

Vêncio, R. Z., Brentani, H., Patrão, D. F., and Pereira, C. A., (2004). Bayesian model accounting for within-class biological variability in Serial Analysis of Gene Expression (SAGE). BMC Bioinformatics 5: 119–131.