Robust Adjustment of Sequence Tag Abundance

By

Douglas D. Baumann and R.W. Doerge

Technical Report #13-03

Department of Statistics
Purdue University

June, 2013

BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–4

Robust Adjustment of Sequence Tag Abundance

Douglas D. Baumann [1], R.W. Doerge [1] *

[1]Department of Statistics, Purdue University, West Lafayette, Indiana, 47907, USA

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: The majority of Next-Generation Sequencing (NGS) technologies effectively sample small amounts of DNA or RNA that are amplified (i.e., copied) prior to sequencing. The amplification process is not perfect, leading to extreme bias in sequenced read counts. We present a novel procedure to account for amplification bias and demonstrate its effectiveness in mitigating gene length dependence when estimating true gene expression.

Results: We tested the proposed method on simulated and real data. Simulations indicated that our method captures true gene expression more effectively than classic censoring-based approaches and leads to power gains in differential expression testing, particularly for shorter genes with high transcription rates. We applied our method to an unreplicated *Arabidopsis* RNA-seq data set resulting in disparate gene ontologies arising from gene set enrichment analyses.

Availability: R code to perform the RASTA procedures is freely available on the web at www.stat.purdue.edu/∼doerge/

Contact: doerge@purdue.edu

1 INTRODUCTION

One cause of technical variation in Next-Generation Sequencing (NGS) studies is amplification bias. Fragmented cDNA is subjected to amplification via polymerase chain reaction (PCR; Saiki *et al*. (1988)) in all NGS applications (Margulies *et al*., 2005; Mardis, 2008; Bennet, 2004). The amplification process is not perfect, and reads can suffer from amplification bias (Chepelev *et al*., 2009). This means that there may be extra copies of certain reads, perhaps tens of thousands of extra copies. The typical statistical procedure to correct for this bias is to ignore any duplicate reads by limiting the number of reads starting at the same base to be 1 read. This censoring procedure, herein referred to as "censoring," ignores the possibility of natural read duplication (multiple copies of the same read which is not due to amplification bias), and thus underestimates true read count. For example, in the human liver samples analyzed by (Marioni *et al*., 2008), 10-15% of the genic bases exhibited duplication, accounting for approximately 30% of the observed reads. While approximately only 1% of the bases exhibited more than 10 duplicated reads, the number of reads starting at these bases comprised approximately 10% of the total reads. The prevalence of duplicated reads in these samples illustrates the need for statistical methods that are able to correct for amplification bias without needlessly censoring natural duplication.

*to whom correspondence should be addressed

The effects of censoring on gene expression depend primarily on gene length and rate of transcription. Under censoring, at most only one read is considered to originate from each nucleotide in a gene. This artificially limits the estimate of gene expression to values less than or equal to gene length. Assuming that the sonication process truly randomly fractionates the mRNA, the expected occurrence of natural read duplication decreases as gene length increases for a given level of gene expression. Thus, the effects of censoring decrease as gene length increases. Conversely, for a given gene, the effects of censoring are more pronounced when gene transcription increases or when the total number of reads increases. In these cases, the sensitivity to detect differences between genes of short length is typically lower than that for longer genes when reads are censored. This length bias can be dramatically reduced when natural read duplication is allowed since the dependence on gene length is mitigated.

We present a novel approach to correct for amplification bias while allowing for natural duplication. The proposed method, Robust Adjustment of Sequence Tag Abundance (RASTA), accurately estimates true tag abundance by separating legitimate reads from incorrectly amplified reads through a novel application of hierarchical clustering. Further, it sets appropriate thresholds for the amplified reads through a novel application of the zero-truncated Poisson distribution. The impact of properly accounting for amplification bias using RASTA when testing for differential gene expression testing, both in terms of power and ranking of results, are investigated. While RASTA was developed and investigated for gene expression, the method is general enough to be applied to DNA methylation and histone modification studies as well.

2 METHODS

Observed RNA-seq reads are assumed to be generated by two distinct processes: legitimate reads (including natural duplication) and amplification bias. For a given mapped read, we define "read count" as the number of observed mapped reads which start at the same base in the genome. Let x_i^g be the read counts for base $i = 1, \ldots, n$ for a given gene g, where n is the number of bases with observed reads in gene g. Given that the x_i^g are generated by two distinct processes, the goal in correctly accounting for amplification bias is to accurately classify each x_i^g into legitimate and erroneous clusters.

Hierarchical clustering, using complete linkage (Sorensen, 1948) and Canberra distance (Lance and Williams, 1966), was used to cluster the read counts into two distinct groups. Since NGS gene expression studies produce discrete read counts, clustering was performed on the unique read count values. Let $(\xi_1^g, \ldots, \xi_m^g)$, where $m \leq n$, be the unique read counts values corresponding to $(x_1^g, \ldots, x_n^g)$ for gene g. The Canberra distance for two

unique read counts (ξ_i^g, ξ_j^g) is defined as

$$d_{ij}^g = \frac{|\xi_i^g - \xi_j^g|}{|\xi_i^g + \xi_j^g|}. \quad (1)$$

In practical settings and simulations $m << n$, thus providing a marked computational time improvement over traditional clustering algorithms based on all read counts.

In order to estimate the distribution of the legitimate reads for each gene g, we assume that the sonication and selection process (Bennet, 2004) randomly fragments the mRNA. Given this random fragmentation process, let

$$x_i^g \sim Poisson(\gamma_g = \frac{\lambda_g}{L_g}) \quad (2)$$

be the distribution of read counts for the n legitimate bases with observed reads for a given gene g, where λ_g and L_g are the overall transcription rate and length for gene g, respectively. Since x_i^g are restricted to be positive only, the legitimate base counts for a given gene are modeled using a zero-truncated $Poisson(\gamma_g^*)$ (ZTP) distribution (Yee, 1996) via the VGAM package (Yee, 2010) in R (R Core Development Team, 2011).

For an estimated value of $\widehat{\gamma_g^*}$, a threshold T_g^* can be defined such that any counts greater than T_g^* at a given base location can be considered to be a result of amplification bias. Here, T_g^* is defined as the 95^{th} percentile of the $ZTP(\gamma_g^*)$ distribution. Then, for each x_i^g, define

$$y_i^g = min(x_i^g, T_g^*) \quad (3)$$

and the digital gene expression (DGE) estimate for gene g is defined as

$$DGE_g = \sum_i y_i^g. \quad (4)$$

3 SIMULATION

3.1 Simulation Design

A simulation study was conducted to evaluate and compare the performance of RASTA to "censoring". For 1,000 genes, gene counts were simulated following Auer & Doerge (2011) with the following modifications: Amplification bias was incorporated by setting the prevalence of bias to $\pi_g^{bias} = .001$ (or 1 out of every 1000 bases), and the bias DGE count to

$$\lambda_g^{bias} \sim Uniform(10, 1000) \quad (5)$$

for each of the 1,000 genes. The value of π_g^{bias} and the upper bound on λ_g^{bias} are relatively conservative, as the prevalence of amplification bias in real data often exceeds 1%, and the erroneously amplified read counts can exist in tens of thousands (Marioni *et al.*, 2008; Lister *et al.*, 2008). Gene lengths were simulated based on the *Mus musculis* and *Drosophila melanogaster* annotation databases from Ensembl (Flicek *et al.*, 2011) with

$$L_g \sim exp(Normal(\mu = 8, \sigma = 2)). \quad (6)$$

For a given gene with parameters λ_g and λ_g^{bias}, the legitimate reads follow

$$Poisson(\gamma_g = \frac{\lambda_g}{L_g}) \quad (7)$$

and the counts arising from amplification bias follow

$$Poisson(\pi_g^{bias} \frac{\lambda_g^{bias}}{L^g}). \quad (8)$$

For each gene, these counts were preprocessed by either truncating all counts to 1 (the current censoring practice) or via RASTA. These modified counts were then summed, giving rise to an adjusted DGE value for each gene. This process was repeated 500 times to account for simulation-to-simulation (sampling) variability.

For the 1,000 simulated genes, both non-differentially expressed (500) and differentially expressed (500) genes were generated for three replicates in two treatments. DGE rates for each gene were generated (Equations 7 - 8) with the following modifications: for differentially expressed genes, means were sampled separately from (7), yielding $\lambda_g^{T_1}$ and $\lambda_g^{T_2}$ for treatments T_1 and T_2; for non-differentially expressed genes, the means were sampled together (λ_g). For each simulated data set, we applied RASTA and "censoring" to the observed base counts. The adjusted gene counts were analyzed for differential expression using the exact negative binomial model in edgeR under a common dispersion assumption (Robinson and Smyth, 2007, 2008). P-values were adjusted using the Benjamini-Hochberg procedure in edgeR (Benjamini and Hochberg, 1995).

3.2 Simulation Results

Statistical power and false discovery rates (FDR) were estimated by taking the averages of true positive and false positive rates ($\alpha = 0.05$) across the simulations. RASTA yields similar effective power and FDR in simulations when compared to the censoring procedure (power: 0.655 vs. 0.602, FDR: 0.23 vs 0.14, respectively). Although the power and FDR rates were similar, summaries comparing true and estimated log fold changes showed greater accuracy under the RASTA method. To illustrate this, estimated log fold changes were regressed against true log fold changes (Figure 1; the relative closeness of the RASTA and "censoring" approaches to the identity line). The regression slope for RASTA was considerably closer to 1 than the censoring method (0.95 and 0.83, respectively), indicating an increase in accuracy when estimating true log fold change between the two treatments.

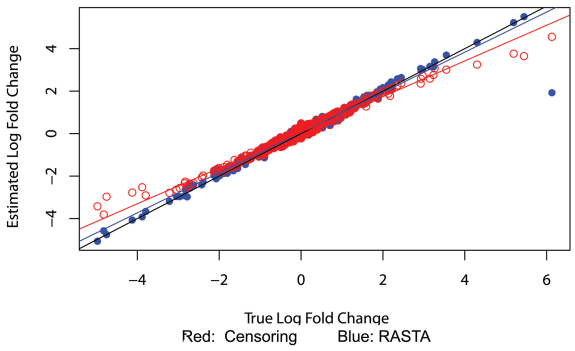

Fig. 1. Simulation results for true vs. estimated log fold change when comparing RASTA versus "censoring." As the true log fold change values increase (in absolute value), RASTA (blue) more accurately estimates the log fold change relative to the censoring (red) procedure (regression slopes: 0.95 vs 0.83, respectively).

In order to assess the relationship between adjusted p-values and gene length, loess smoothing (Cleveland, 1979) was applied to the results from the edgeR analyses (Figure 2). In addition to the simulated DGE levels representative of those typically observed in current RNA-seq studies (displayed in solid lines), Figure 2 also displays results from simulations in which these DGE levels were doubled on average (displayed in dashed lines). By more accurately estimating DGE using RASTA, especially for shorter genes with high DGE, RASTA is able to all but eliminate length bias in these simulations as average DGE levels increase.

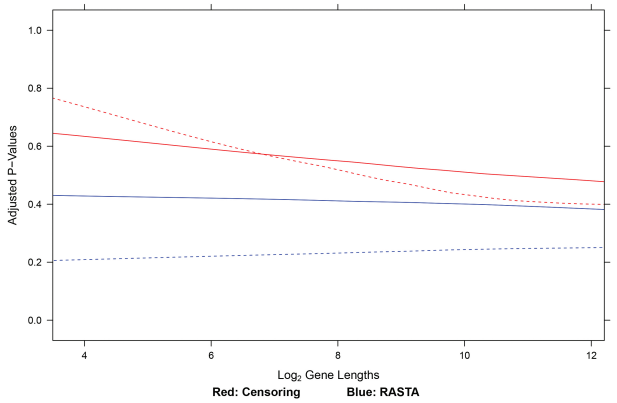

Fig. 2. Gene length bias simulation results. The censoring method is presented in red, while the RASTA method is presented in blue. The solid lines represent simulated gene expression levels based on (Auer and Doerge, 2011). The dashed lines represent a doubling, on average, of DGE levels. For the original simulation settings, RASTA provided a marginal improvement over the censoring procedure. When average DGE was increased, RASTA showed little evidence of length bias, while the censoring procedure's bias became much more pronounced.

Table 1. Distribution of read duplication for the unreplicated *met1-3* and *Col-0 Arabidopsis* lines in Lister et al. (2008). The *Col-0* wild-type sample displays considerably more duplication than the *met1-3* mutants at each of the levels presented.

	met1-3	*Col-0*
Total Reads	5997689	6283230
Unique Reads	2991256	1264135
Single bases with $\geq$ 5 reads	139972	285610
Single bases with $\geq$ 10 reads	38718	72227
Single bases with $\geq$ 100 reads	232	849
Max number of reads at a single base	5525	17063

4 APPLICATION TO ARABIDOPSIS

4.1 Materials and Methods

The censoring and RASTA approaches were used to preprocess the unreplicated *Arabidopsis* RNA-seq data from Lister et al. (2008). In this study, *met1-3* mutants (deficient in methylation) were compared to wild-type (*Col-0*) controls. Gene start and stop locations were used to define 22,266 annotated genomic regions, and were based on the Columbia reference genome gained from The Arabidopsis Information Resource (TAIR, Swarbreck *et al.* (2008)). Although the total number of mapped reads for the *met1-3* and *Col-0* samples were approximately equal (5,997,689 and 6,283,230, respectively), the occurrence of read duplication, either from natural duplication or amplification bias, was dramatically different between the two samples (Table 1).

Gene counts under each of the control procedures were analyzed using the exact negative binomial model in edgeR (Robinson and Smyth, 2007, 2008). P-values were adjusted using the Benjamini-Hochberg FDR procedure (Benjamini and Hochberg, 1995), and the nominal significance threshold was set at $\alpha = 0.01$. Gene set enrichment analysis (GSEA) was performed on the resulting lists of significant genes using agriGO (Du *et al.*, 2010; Berg *et al.*, 2009). The agriGO toolkit performs GSEA based on a hypergeometric distribution to assess the over- or under-representation of gene ontologies in the lists of significant genes when compared to all genes with annotated ontologies, and corrects for multiple testing using FDR under dependence assumptions (Benjamini and Yekutieli, 2001). The collection of gene ontologies for each differentially expressed gene are collated, and if the proportion of a particular ontology in the differentially expressed genes is significantly different (higher or lower) than the corresponding proportion in the entire gene set, that function is reported in agriGO.

Table 2. Gene Set Enrichment Analysis results (top five ontologies) from the agriGO toolkit under censoring and RASTA amplification bias control procedures for the unreplicated *met1-3* and *Col-0 Arabidopsis* lines in Lister *et al.* (2008). The "GO Term" and "Description" columns represent the gene ontologies enriched in the significant gene lists when compared to all *Arabidopsis* gene ontologies. The p-values are based on the hypergeometric distribution, and are adjusted via FDR under dependence (Benjamini and Yekutieli, 2001). The resulting enriched ontologies for the censoring and RASTA approaches are quite disparate, indicating that the control procedure is highly influential in downstream analyses.

RASTA		
GO Term	Ontology Description	Adj. p-value
GO:0009791	Post-embryonic development	4.2e-76
GO:0034641	Cellular nitrogen compound metabolic process	5.7e-33
GO:0032501	Multicellular organismal process	2.4e-24
GO:0009987	Cellular process	5.9e-24
GO:0007275	Multicellular organismal development	1.4e-23

Censoring		
GO Term	Ontology Description	Adj. p-value
GO:0009628	Response to abiotic stimulus	2.2e-19
GO:0050896	Response to stimulus	8.2e-17
GO:0009791	Post-embryonic development	1.6e-16
GO:0006950	Response to stress	3e-16
GO:0044262	Cellular carbohydrate metabolic process	3.3e-16

4.2 Results

The presence of DNA methylation typically serves as a transcriptional regulator in eukaryote species; when depleted, gene transcription typically increases (Riggs, 1975; Robertson, 2005; Shames *et al.*, 2007). The RASTA approach yielded many more statistically significant differentially expressed genes than the censoring method (8912 and 2855 genes, respectively). This increase is in concordance with the biological knowledge that when comparing the two *Arabidopsis* lines, *met1-3* is deficient in methylation maintenance which reduces the degree of gene regulation (Lister *et al.*, 2008). The agriGO GSEA results based on the two gene lists (Table 2) display a stark contrast in enriched gene ontologies, indicating that appropriate amplification bias control is important for discovery and downstream confirmation studies. In fact, of the top ten significant ontologies (top five shown in Table 2) produced by RASTA and censoring, only two are similar between the two lists.

5 DISCUSSION

Accurately estimating digital gene expression, and subsequently differential gene expression, is a primary challenge in Next-Generation RNA sequencing studies. One of the key sources for technical variation between samples, and between or within treatments, is amplification bias. Controlling for this bias not only improves the accuracy of DGE estimates (Figure 1), it dramatically changes downstream analyses. Since confirmatory studies often target the most statistically significant differentially expressed genes (i.e., the genes with the lowest p-values), the ordering of results plays an important role in downstream analyses.

As the costs for sequencing decrease, we anticipate that researchers will want a greater number of sequenced reads in order to more accurately detect differences in expression levels between treatments. This scenario provides some cause for caution, as blindly seeking high read counts invites the possibility of over-amplification in order to achieve a particular observed sequencing depth or coverage. If sequenced reads are systematically over-amplified, as is the case in Shiroguchi et al. (2012), researchers are relegated to only two approaches: Digital RNA Sequencing (DRS, (Shiroguchi *et al.*, 2012)), when the additional amplification is expected before sequencing; and censoring, when the amplification is not planned. DRS is a promising biological approach to account for amplification bias, but its use comes at significant cost to the researcher. First, it requires greater sequencing depth than conventional RNA-seq studies in order to effectively sample read/barcode pairs. Secondly, DRS prohibits barcoding for efficient sequencing. Where several samples could be sequenced in the same lane using sample-specific barcodes normally, the DRS procedure requires separate lanes for each sample. Finally, at least in the *E. coli* data from Shiroguchi et al. (2012), the extra time and sequencing costs associated with DRS could be eliminated by just using the censoring approach. This would be true when reads are systematically over-amplified in general. However, the censoring approach is insensitive to natural read duplication, which in turn results in an underestimation of true DGE when reads are actually naturally duplicated.

Achieving greater sequencing depth can be done correctly, without limiting the choice in amplification bias control procedures, simply by using a larger sample of mRNA from subjects. As sequencing depth increases due to larger biological samples of mRNA, the occurrence of legitimately duplicated reads will increase. Assuming that reasonable amplification is employed prior to sequencing, the proposed RASTA approach is well-suited to account for amplification bias even in the context of increased natural read duplication. In these settings, the censoring approach will consistently underestimate the true DGE; on the other hand, the DRS approach is likely to produce similar results to RASTA, though with greater restrictions and increased sequencing cost. As a statistical procedure, RASTA costs very little to the researcher since it is computationally efficient and requires no additional sequencing or sequencing reagents. At the same time, the hierarchical clustering and zero-truncated Poisson estimation procedures used in RASTA are powerful and are able to accurately classify legitimate and erroneous reads when both exist for a given gene.

ACKNOWLEDGEMENT

We thank Andrea Schorn from the Martienssen lab at Cold Spring Harbor Laboratory, and Sanvesh Srivastava from the Doerge group in the Department of Statistics, Purdue University for helpful discussions.

Funding: This work is funded in part by a National Science Foundation (DBI-1025976) grant to RWD and her colleagues.

REFERENCES

Auer, P. and Doerge, R. (2011), A Two-Stage Poisson Model for Testing RNA-Seq Data, *Statistical Applications in Genetics and Molecular Biology* **10**, 26.

Benjamini, Y. and Hochberg, Y. (1995), Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society, Series B (Methodological)* **57**, 289-300.

Benjamini, Y. and Yekutieli, D. (2001), The control of the false discovery rate in multiple testing under dependency, *Annals of Statistics* **29**, 1165-1188.

Bennet, S. (2004), Solexa ltd., *Pharmacogenomics* **5**, 433-438.

Berg, B. and Thanthiriwatte, C. and Manda, P. and Bridges, S. (2009), Comparing gene annotation enrichment tools for functional modeling of agricultural microarray data, *BMC Bioinformatics* **10**, S9.

Chepelev, I. and Wei, G. and Tang, Q. and Zhao, K. (2009), Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq, *Nucleic Acids Research* **37**, e106.

Cleveland, W. (1979), Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association* **74**, 829-836.

Du, Z.; Zhou, X.; Ling, Y.; Zhang, Z. and Su, Z. (2010), agriGO: a GO analysis toolkit for the agricultural community, *Nucleic Acids Research* **38**, W64-W70.

Flicek, P.; Amode, M. R.; Barrell, D.; Beal, K.; Brent, S.; Chen, Y.; Clapham, P.; Coates, G.; Fairly, S.; Fitzgerald, S.; Gorgon, L.; Hendrix, M.; Hourlier, T.; Johnson, N. and Searle, S. (2011), Ensembl 2011, *Nucleic Acids Research* **39**, D800-D806.

Lance, G. and Williams, W. (1966), Computer programs for hierarchical polythetic classification ("similarity analysis")., *Computer Journal* **9**, 60-64.

Lister, R.; OMalley, R. C.; Tonti-Filippini, J.; Gregory, B. D.; Berry, C. C.; Millar, A. H. and Ecker, J. R. (2008), Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis, *Cell* **133**, 523-536.

Mardis, E. (2008), Next-generation DNA sequencing methods, *Annual Review of Genomics and Human Genetics* **9**, 387-402.

Margulies, M.; Egholm, M.; Altman, W.; Attiya, S. and Bader, J. (2005), Genome sequencing in microfabricated high-density picolitre reactors, *Nature* **437**, 376-380.

Marioni, J.; Mason, C.; Mane, S.; 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 Core Development Team (2011), R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Riggs, A. (1975), X inactivation, differentiation, and DNA methylation, *Cytogenics and Cell Genetics* **14**, 9-25.

Robertson, K. (2005), DNA methylation and human disease, *Nature Reviews Genetics* **6**, 597-610.

Robinson, M.; McCarthy, D. and Smyth, G. (2010), edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, *Bioinformatics* **26**, 139-140.

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

Robinson, M. and Smyth, G. (2007), Moderated statistical tests for assessing differences in tag abundance, *Bioinformatics* **23**, 2881-2887.

Saiki, R.; Gelfand, D. and Stoffel, S. and Scharf, S. and Higuchi, R. (1988), Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase, *Science* **239**, 487-491.

Shames, D.; Minna, J. and Gazdar, A. (2007), DNA Methylation in Health, Disease, and Cancer, *Current Molecular Medicine* **7**, 85-102.

Shiroguchi, K.; Jia, T.; Sims, P. and Xie, X. (2012), Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes, *Proceedings of the National Academy of Science* **109**, 1347-1352.

Sorensen, T. (1948), A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to the analyses of the vegetation on Danish commons, *Biologiske Skrifter* **5**, 1-34.

Swarbreck, D.; Wilks, C.; Lamesch, P.; Berardini, T.; Garcia-Hernandez, M.; Foerster, H.; Li, D.; Meyer, T.; Muller, R.; Ploetz, L.; Radenbaugh, A.; Singh, S.; Swing, V.; Tissier, C.; Zhang, P. and Huala, E. (2008), The Arabidopsis Information Resource (TAIR): gene structure and function annotation, *Nucleic Acids Research* **36**, 1009-1014.

Yee, T. and Wild, C. (1996), Vector Generalized Additive Models, *Journal of the Royal Statistical Society, Series B* **58**, 481-493.

Yee, T. (2010), The VGAM Package for Categorical Data Analysis, *Journal of Statistical Software* **32**, 1-34.