BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 5 2014, pages 601–605
doi:10.1093/bioinformatics/btt575

Genome analysis

Advance Access publication October 9, 2013

Robust adjustment of sequence tag abundance

Douglas D. Baumann† and Rebecca W. Doerge*

Department of Statistics, Purdue University, West Lafayette, IN 47907, USA

Associate Editor: Michael Brudno

ABSTRACT

Motivation: The majority of next-generation sequencing technologies effectively sample small amounts of DNA or RNA that are amplified (i.e. copied) before 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 dataset resulting in disparate gene ontologies arising from gene set enrichment analyses.

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

Contact: doerge@purdue.edu

Received on June 4, 2013; revised on September 19, 2013; accepted on September 27, 2013

1 INTRODUCTION

One cause of technical variation in next-generation sequencing (NGS) studies is amplification bias. Fragmented complementary DNA is subjected to amplification via polymerase chain reaction in all NGS applications (Bennet, 2004; Mardis, 2008; Margulies *et al*., 2005). 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 one read. This censoring procedure, herein referred to as 'Censoring', ignores the possibility of natural read duplication (multiple copies of the same read that 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 ~30% of the observed reads. Although approximately only 1% of the bases exhibited >10 duplicated reads, the number of reads starting at these bases comprised ~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.

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 randomly fractionates the messenger RNA (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, as 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 (ZTP) 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.

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 that 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.

RASTA approaches this goal in two steps: hierarchical clustering and distributional approximation. Hierarchical clustering, using complete linkage and Canberra distance (Lance and Williams, 1966), is used to cluster the read counts into two distinct groups. Because NGS gene expression studies produce discrete read counts, clustering is 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.

*To whom correspondence should be addressed.
†Present address: Department of Mathematics, University of Wisconsin - La Crosse, La Crosse, WI 54601.

The Canberra distance for two unique read counts (ξ_i^g, ξ_j^g) is defined as follows:

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

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

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 be the read counts for the n legitimate bases with observed reads for a given gene g. Because 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 via the vector generalized linear and additive models 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 $>T_g^*$ 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

$$y_i^g = min(x_i^g, T_g^*) \quad (2)$$

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

$$DGE_g = \sum_i y_i^g. \quad (3)$$

3 SIMULATION

3.1 Simulation design

A simulation study was conducted to evaluate and compare the performance of RASTA with 'Censoring'. For 1000 genes, gene counts were simulated following Auer and Doerge (2011) with the following modifications: amplification bias was incorporated by setting the prevalence of bias to $\pi_g^{bias} = .05$ (or 1 of every 1000 bases) and the bias DGE count to

$$\lambda_g^{bias} \sim exp(Poisson(\lambda = 4)) \quad (4)$$

for each of the 1000 genes. The value of π_g^{bias} and the distribution of λ_g^{bias} are based on the *Arabidopsis* samples sequenced in Lister *et al.* 2008. Gene lengths were also simulated based on *Arabidopsis* with

$$L_g \sim exp(Normal(\mu = 7, \sigma = 2)). \quad (5)$$

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

$$Poisson(\gamma_g = \frac{\lambda_g}{L_g}) \quad (6)$$

and the counts arising from amplification bias follow

$$Poisson(\pi_g^{bias} \frac{\lambda_g^{bias}}{L^g}). \quad (7)$$

For each gene, these counts were preprocessed by either truncating all counts to one (the current Censoring practice) or via RASTA, in addition to investigating the uncorrected data. These counts were then summed, giving rise to three separate DGE values for each gene. This process was repeated 500 times to account for simulation-to-simulation (sampling) variability.

For the 1000 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 (6) and (7)] with the following modifications: for differentially expressed genes, means were sampled separately from (6), 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). In addition, bias prevalence was set at $\pi_g^{bias(T_1)} = .05$ and $\pi_g^{bias(T_2)} = .02$ for the two treatments. For each simulated dataset, 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, 2008). *P*-values were adjusted using the Benjamini–Hochberg false discovery rate (FDR) procedure in edgeR (Benjamini and Hochberg, 1995).

3.2 Simulation results

Genewise log fold changes, as estimated by edgeR, were compared against simulated true log fold changes across the 1000 genes in each simulation. RASTA yields the greatest correlation between estimated and true values ($r = 0.52$), compared with the correlations yielded by uncorrected ($r = 0.32$) and censored ($r = 0.29$) data. To assess the effects of gene length on differential expression estimates, adjusted *P*-values and gene length were compared using 2D histograms (Fig. 1) for each of the three read count adjustment methods (Censoring, RASTA, No Correction). 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.

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) and were compared with uncorrected read counts. In this study, *met1-3* mutants (deficient in methylation) were compared with wild-type (*Col-0*) controls. Gene start and stop locations were used to define 22 266 annotated genomic regions and

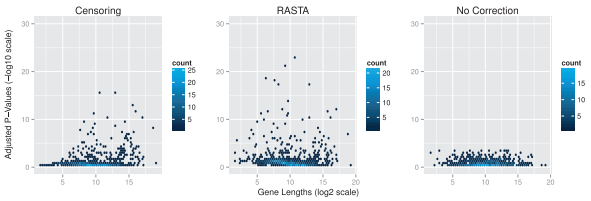

Fig. 1. Gene length bias simulation results. FDR-adjusted *P*-values are compared against corresponding gene lengths for each of the three read count adjustment methods (Censoring, RASTA, No Correction). The Censoring approach displays a bias of significant *P*-values toward longer genes. This bias is not evident in the RASTA and No Correction *P*-values, though RASTA allowed for more significant test results

Table 1. Distribution of read duplication for the unreplicated *met1-3* and *Col-0 Arabidopsis* lines in Lister *et al.* (2008)

Read Count Information	met1-3	Col-0
Total reads	5 997 689	6 283 230
Unique reads	2 991 256	1 264 135
Single bases with ≥ 5 reads	139 972	285 610
Single bases with ≥ 10 reads	38 718	72 227
Single bases with ≥ 100 reads	232	849
Maximum number of reads at a single base	5525	17 063

Note: The Col-0 wild-type sample displays considerably more duplication than the met1-3 mutants at each of the levels presented.

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, 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 (Berg *et al.*, 2009; Du *et al.*, 2010). The agriGO toolkit performs GSEA based on a hypergeometric distribution to assess the over- or underrepresentation of gene ontologies in the lists of significant genes when compared with 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 is 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.

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 and No Correction approaches yielded many more statistically significant differentially expressed genes than the Censoring method (8912, 9366 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 that reduces the degree of gene regulation (Lister *et al.*, 2008). The differentially expressed genes vary between the three approaches, though nearly all of significant genes yielded by the Censoring approach were also found in the other two approaches. In addition, the RASTA and uncorrected

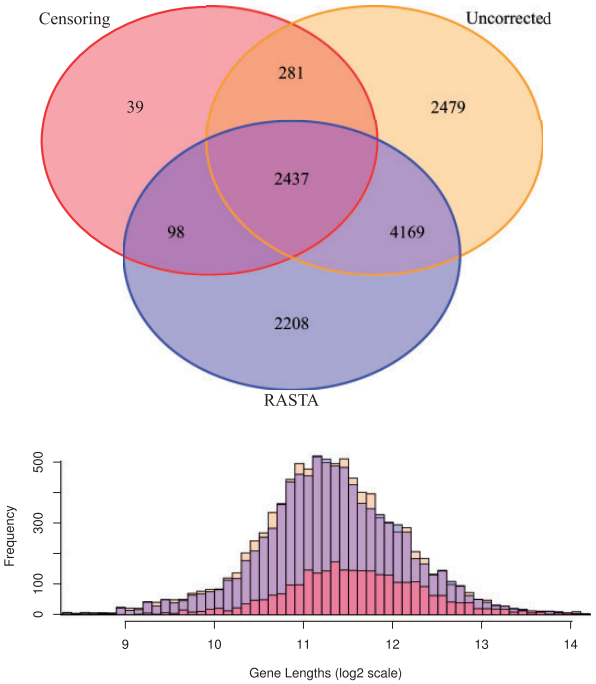

Fig. 2. Distribution of significantly differentially expressed genes under Censoring (red), RASTA (blue) and No Correction (orange) amplification bias controls procedures for the unreplicated *met1-3* and *Col-0 Arabidopsis* lines in Lister *et al.* (2008). The RASTA and uncorrected approaches both yielded many more significant results than the Censoring method, and these significant genes were shorter on average than the Censoring results

approaches yielded more significant genes with shorter length than did the Censoring approach (Fig. 2). The agriGO GSEA results based on the three 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 produced by each approach, only two are similar between the Censoring and RASTA, and only six are similar between RASTA and No Correction.

5 DISCUSSION

Accurately estimating DGE, 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, it dramatically changes downstream analyses. Because 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 to more accurately detect differences in expression levels between treatments. This scenario provides some cause for caution, as

Table 2. GSEA results (top five ontologies) from the agriGO toolkit under Censoring, RASTA and No Correction amplification bias control procedures for the unreplicated *met1-3* and *Col-0 Arabidopsis* lines in Lister *et al.* (2008)

Gene Ontology term	Ontology description	Adjusted *P*-value
Censoring		
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
RASTA		
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
No correction		
GO:0009791	Post-embryonic development	1.4e-80
GO:0034641	Cellular nitrogen compound metabolic process	1.6e-40
GO:0010035	Response to inorganic substance	1.1e-30
GO:0009987	Cellular process	1.1e-29
GO:0006950	Response to stress	6.3e-23

Note: The 'GO Term' and 'Description' columns represent the gene ontologies enriched in the significant gene lists when compared with 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 disparate, whereas the RASTA and No Correction approach several similar ontologies. These results indicate that the control procedure is highly influential in downstream analyses.

blindly seeking high read counts invites the possibility of over-amplification 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), the hierarchical clustering used in RASTA erroneously classifies many amplified reads as legitimate, therein overestimating true read counts. In these cases, 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 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 *Escherichia 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, as both approaches yield similar results in these data. 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 used before 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.

The clustering and distributional considerations made in RASTA assume that the mRNA fragmentation process is random, and the amplification process is unbiased to genomic content. These assumptions may be violated under GC amplification bias, differential isoform expression or genomic sonication bias. In these cases, the ZTP distribution used in RASTA could be replaced by a similarly formed zero-truncated negative binomial (ZTNB) distribution. When the ZTNB distribution was applied to simulated data and the Lister *et al.* (2008) *Arabidopsis* data, the resulting model fits were similar to the ZTP model fits, and the analyses took nearly three times longer when using the ZTNB parameterization. However, the negative binomial approach may be more applicable in other datasets and is a straightforward extension to RASTA. In addition, the hierarchical clustering and ZTP/ZTNB estimation procedures used in RASTA could serve as a precursor to subsequent isoform discovery and abundance estimation analyses proposed by Mezlini *et al.* (2013).

As a statistical procedure, RASTA costs little to the researcher, as it is computationally efficient and requires no additional sequencing or sequencing reagents. At the same time, the hierarchical clustering and ZTP estimation procedures used in RASTA are powerful and are able to accurately classify legitimate and erroneous reads when both exist for a given gene.

ACKNOWLEDGEMENTS

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

Funding: National Science Foundation (DBI-1025976) grant (to R.W.D. and colleagues).

Conflict of Interest: none declared.

REFERENCES

Auer,P. and Doerge,R. (2011) A two-stage poisson model for testing RNA-Seq data. *Stat. Appl. Genet. Mol. Biol.*, **10**, 26.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.*, **57**, 289–300.

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

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

Berg,B. *et al.* (2009) Comparing gene annotation enrichment tools for functional modeling of agricultural microarray data. *BMC Bioinformatics*, **10**, S9.

Chepelev,I. *et al.* (2009) Detection of single nucleotide variations in expressed exons of the human genome using RNA-Seq. *Nucleic Acids Res.*, **37**, e106.

Du,Z. *et al.* (2010) agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Res.*, **38**, W64–W70.

Lance,G. and Williams,W. (1966) Computer programs for hierarchical polythetic classification (''similarity analysis''). *Comput. J.*, **9**, 60–64.

Lister,R. *et al.* (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell*, **133**, 523–536.

Mardis,E. (2008) Next-generation DNA sequencing methods. *Ann. Rev. Genomics Hum. Genet.*, **9**, 387–402.

Margulies,M. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactor. *Nature*, **437**, 376–380.

Marioni,J. *et al.* (2008) RNA-Seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.

Mezlini,A.M. *et al.* (2013) iReckon: simultaneous isoform discovery and abundance estimation from RNA-seq data. *Genome Res.*, **23**, 519–529.

R Core Development Team. (2011) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Riggs,A. (1975) X inactivation, differentiation, and DNA methylation. *Cytogenet. Cell Genet.*, **14**, 9–25.

Robertson,K. (2005) DNA methylation and human disease. *Nat. Rev. Genet.*, **6**, 597–610.

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

Shames,D. *et al.* (2007) DNA methylation in health, disease, and cancer. *Curr. Mol. Med.*, **7**, 85–102.

Shiroguchi,K. *et al.* (2012) Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proc. Natl Acad. Sci.*, **109**, 1347–1352.

Swarbreck,D. *et al.* (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.*, **36**, 1009–1014.

Yee,T. (2010) The VGAM package for categorical data analysis. *J. Stat. Softw.*, **32**, 1–34.