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 \( x_g^i \) 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 \( x_g^i \) are generated by two distinct processes, the goal in correctly accounting for amplification bias is to accurately classify each \( x_g^i \) 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 \( (x_1^g, \ldots, x_m^g) \), 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 \( (\xi^1_g, \xi^2_g) \) is defined as follows:

\[
d_g = \frac{|\xi^1_g - \xi^2_g|}{|\xi^1_g + \xi^2_g|}
\]

(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 \( x^g \) be the read counts for the \( n \) legitimate bases with observed reads for a given gene \( g \). Because \( x^g \) are restricted to be positive only, the legitimate base counts for a given gene are modeled using a zero-truncated Poisson \( (y^g) \) (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 \( y^g \), 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 95th percentile of the ZTP\( (y^g) \) distribution. Then, for each \( x^g \), define

\[
y^g = \min(x^g, T^g)
\]

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

\[
DGE_g = \sum_i y^g_i.
\]

(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 \( \pi^b_g = .05 \) (or 1 of every 1000 bases) and the bias DGE count to

\[
\lambda^b_g \sim \exp(\text{Poisson}(\lambda = 4))
\]

(4)

for each of the 1000 genes. The value of \( \pi^b_g \) and the distribution of \( \lambda^b_g \) are based on the Arabidopsis samples sequenced in Lister et al. 2008. Gene lengths were also simulated based on Arabidopsis with

\[
L_g \sim \exp(\text{Normal}(\mu = 7, \sigma = 2)).
\]

(5)

For a given gene with parameters \( \lambda_g \) and \( \lambda^b_g \), the legitimate reads follow

\[
\text{Poisson}(y_g = \frac{\lambda_g}{L_g})
\]

(6)

and the counts arising from amplification bias follow

\[
\text{Poisson}(\pi^b_g \frac{\lambda^b_g}{L_g}).
\]

(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 \( \lambda^b_g \) and \( \lambda^b_g \) for treatments \( T_1 \) and \( T_2 \); for non-differentially expressed genes, the means were sampled together (\( \lambda_g \)). In addition, bias prevalence was set at \( \pi^b_g(T_1) = .05 \) and \( \pi^b_g(T_2) = .02 \) 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.](https://example.com/)
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 (8,912, 9,366 and 2,855 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 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
for each sample. Finally, at least in the barcodes normally, the DRS procedure requires separate lanes. Ples could be sequenced in the same lane using sample-specific procedures for the unreplicated met1-3 and Col-0 Arabidopsis lines in Lister et al. (2008)

<table>
<thead>
<tr>
<th>Gene Ontology term</th>
<th>Ontology description</th>
<th>Adjusted P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Censoring</td>
<td>Response to abiotic stimulus</td>
<td>2.2e-19</td>
</tr>
<tr>
<td>GO:0000628</td>
<td>Response to stimulus</td>
<td>8.2e-17</td>
</tr>
<tr>
<td>GO:0009791</td>
<td>Post-embryonic development</td>
<td>1.6e-16</td>
</tr>
<tr>
<td>GO:0006950</td>
<td>Response to stress</td>
<td>3e-16</td>
</tr>
<tr>
<td>GO:0044262</td>
<td>Cellular carbohydrate metabolic process</td>
<td>3.3e-16</td>
</tr>
<tr>
<td>RASTA</td>
<td>Post-embryonic development</td>
<td>4.2e-76</td>
</tr>
<tr>
<td>GO:0034641</td>
<td>Cellular nitrogen compound metabolic process</td>
<td>5.7e-33</td>
</tr>
<tr>
<td>GO:0032501</td>
<td>Multicellular organisinal process</td>
<td>2.4e-24</td>
</tr>
<tr>
<td>GO:0009987</td>
<td>Cellular process</td>
<td>5.9e-24</td>
</tr>
<tr>
<td>GO:0007275</td>
<td>Multicellular organisinal development</td>
<td>1.4e-23</td>
</tr>
<tr>
<td>No correction</td>
<td>Post-embryonic development</td>
<td>1.4e-80</td>
</tr>
<tr>
<td>GO:0034641</td>
<td>Cellular nitrogen compound metabolic process</td>
<td>1.6e-40</td>
</tr>
<tr>
<td>GO:010035</td>
<td>Response to inorganic substance</td>
<td>1.1e-30</td>
</tr>
<tr>
<td>GO:0009987</td>
<td>Cellular process</td>
<td>1.1e-29</td>
</tr>
<tr>
<td>GO:0006950</td>
<td>Response to stress</td>
<td>6.3e-23</td>
</tr>
</tbody>
</table>

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


