Conference on Applied Statistics in Agriculture   2012 - 24th Annual Conference Proceedings

CORRECTING FOR AMPLIFICATION BIAS IN NEXT-GENERATION SEQUENCING DATA

Douglas Baumann

R. W. Doerge

Recommended Citation

CORRECTING FOR AMPLIFICATION BIAS
IN NEXT-GENERATION SEQUENCING DATA

Douglas Baumann and R.W. Doerge*
Department of Statistics, Purdue University

*Corresonding Author:
R.W. Doerge
Department of Statistics
Purdue University
250 N. University St.
West Lafayette, IN 47907
email: doerge@purdue.edu
phone: 765-494-6030
fax: 765-494-0558

ABSTRACT: Next-generation sequencing (NGS) technologies have opened the door to a wealth of knowledge and information about biological systems, particularly in genomics and epigenomics. These tools, although useful, carry with them additional technological and statistical challenges that need to be understood and addressed. One such issue is amplification bias. Specifically, the majority of 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, and thus sequenced read counts can be extremely biased. Unfortunately, current amplification bias controlling procedures introduce a dependence of gene expression on gene length, which effectively masks the effects of short genes with high transcription rates. In this work we present a novel procedure to account for amplification bias and demonstrate its effectiveness in estimating true gene expression independent of gene length.

1 Introduction

A number of biotechnological advances have been developed recently that assist scientists in associating genes with various biological outcomes. One such advance is called Next-Generation Sequencing (NGS). NGS technologies (Margulies et al., 2005; Mardis, 2008; Bennet, 2004) have allowed researchers to perform genome-wide studies at extremely high resolution on a variety of heritable biological phenomena, including gene expression via RNA-seq (Marioni et al., 2008), DNA methylation via MethylC-seq (Lister et al., 2008), and histone modifications via ChIP-seq (Mikkelsen et al., 2007). While these tools are useful, there are nontrivial technological and statistical issues that need to be understood and addressed in order to analyze these data properly.

1.1 RNA-seq Workflow

There are three primary technologies used in NGS applications: the 454 Genome Sequencer FLX by Roche (Margulies et al., 2005), SOLiD by Applied Biosystems (Mardis, 2008), and the Genome Analyzer (also called the "Solexa" sequencer) by Illumina (Bennet, 2004). In this work, we focus on data originating from the Illumina platform, which has been used in a variety of NGS applications (Marioni et al., 2008; Lister et al., 2008; Mikkelsen et al., 2007), though the methods presented can be easily adapted for the other technologies. RNA sequencing (RNA-seq) is an approach for quantitatively assessing gene expression levels using high-throughput sequencing. In gene expression applications, mRNA is isolated, randomly fragmented using high-frequency sound waves (referred to as sonication), and converted to complimentary DNA (cDNA). Fragments are then selected based on size (approximately 200 bases in length), and amplified through polymerase chain reaction (PCR) (Saiki et al., 1988). The first 36-50 bases of each fragment are actually sequenced (resulting in segments known as reads), since the quality of base calls degrades over the length of the fragment. In cases where a reference genome (i.e., reliable genomic sequence) exists, these reads are then aligned to the reference genome, providing the start and stop locations of the reads (Figure 1).

The mRNA transcripts, and subsequently the resulting reads, are assumed to be generated exclusively from gene regions that can produce upward of tens of thousands of transcripts per gene. After alignment to a reference genome, the number of reads that mapped to each gene is counted and serve as a representation of each gene's transcription level. Several statistical issues arise in the analysis of these discrete data. Since the cost of NGS studies is often prohibitively high, most studies employ relatively few biological replicates; in fact, many studies are unreplicated (i.e., one sample per treatment or condition) entirely. This presents non-trivial issues when estimating model parameters and deviations from the specified model. Secondly, sources of technical variation further complicate the statistical approaches needed to correctly analyze these data.

1.2 RNA-seq Gene Expression Data Analysis Methods

A wide variety of open-source software, primarily in R, exists to perform analyses of NGS gene expression data. In particular, edgeR (Robinson and Smyth, 2008) and DEseq (Anders and Huber, 2010) are the most commonly used packages, and both have a similar modeling strategy (we will focus on edgeR for clarity). In R, edgeR typically models differential gene expression using a Negative Binomial model. Suppose n_{gjk} is the observed number of reads mapping to gene g in
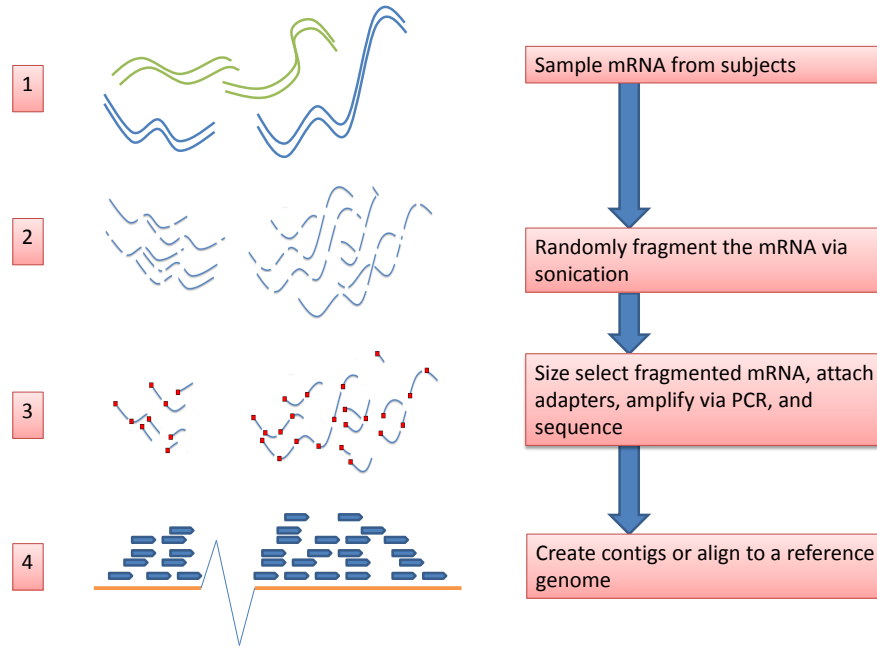
Figure 1: General overview of the RNA-seq process. In RNA-seq applications, mRNA is isolated and fragmented, then amplified, sequenced, and aligned to a known reference genome. Each of these steps has potential to introduce biases in the resulting data. In particular, the amplification process can potentially cause an over-representation of certain fragments, resulting in biased gene expression estimates.

treatment j and sample k. Then the Negative Binomial probability of observing n_{gjk} is defined as

$$P[N_{gjk} = n_{gjk}|\mu_{gj}, \phi_{gj}] = \frac{\Gamma(n_{gjk} + \phi_{gj}^{-1})}{\Gamma(\phi^{-1}) + \Gamma(n_{gjk} + 1)} \left(\frac{1}{1 + \mu_{gj}\phi_{gj}} \right)^{\phi_{gj}^{-1}} \left(\frac{\mu_{gj}}{\phi_{gj}^{-1} + \mu_{gj}} \right)^{n_{gjk}}. \qquad (1)$$

Under this model, $E[N_{gjk}] = \mu_{gj}$ and $V[N_{gjk}] = \mu_{gj} + \phi_{gj} \ mu_{gj}^2$. The parameter ϕ_{gj} represents the overdispersion, or extra variation relative to a Poisson assumption, present in the data.

In edgeR, ϕ_{gj} can be calculated in several ways. It can be set to be the same for all genes ($\phi_{gj} = \phi_0$), different for each gene ($\phi_{gj} = \phi_g$), or as a weighted average of ϕ_0 and ϕ_g. The weighted average represents a form of information-borrowing among genes, and often improves inference in settings with low replication (Robinson and Smyth, 2008). In many exploratory research settings, RNA-seq experiments employ only single replicates per treatment group. In these cases, ϕ_{gj} is set to 0, effectively reducing the Negative Binomial test to an Exact Poisson test, which assumes $E[N_{gj}] = \mu_{gj} = V[N_{gj}]$.

1.3   Amplification Bias, Natural Read Duplication, and Censoring

One cause of technical variation is amplification bias. As previously stated, fragmented cDNA is subjected to amplification via PCR in each of the NGS applications (Figure 1). The amplification

process is not perfect, and reads can suffer from amplification bias (Chepelev et al., 2009). This means that extra copies of certain reads may exist, perhaps tens of thousands of extra copies. The typical procedure to correct for this type of bias is to ignore any duplicate reads by limiting the number of reads starting at the same base to be 1 read (Figure 2). 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.
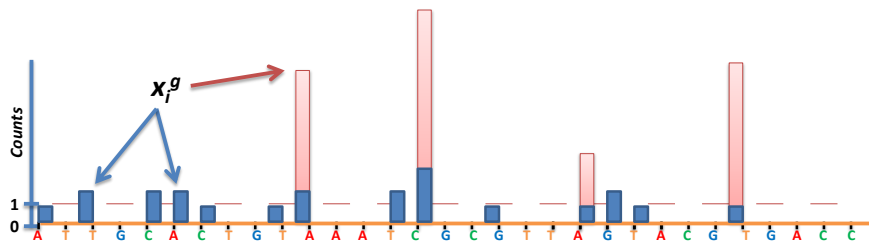

Figure 2: Representation of a single gene expression profile generated using RNA-seq. The blue bars represent legitimate reads, while the red bars represent amplification bias. x_i^g represents the number of reads at the ith genomic base. The "censoring" procedure truncates the number of reads at each base at 1, as depicted by the dashed red line.

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 from each nucleotide in a gene. This artificially limits the estimate of gene expression to values less than or equal to gene length. However, for a given level of gene expression, the expected occurrence of natural read duplication decreases as gene length increases when reads are assumed to be distributed uniformly across the gene. As such, the effects of censoring decrease as gene length increases. Conversely, for a given gene (with length L_g, for example), 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. This length bias can be dramatically reduced when natural read duplication is allowed, simply because the dependence on gene length is mitigated.

Perhaps most important is the ranking of genes based on statistical significance when testing for differential expression. Typically, genome-wide studies are used by researchers as *hypothesis generating studies*, from which interesting results are further investigated through confirmatory experiments. As such, the genes which exhibit the greatest statistical significance are prioritized in downstream, confirmatory studies. Correctly estimating gene read abundance, and consequently reducing the effects of length bias, has the potential to drastically reorder the gene rankings when compared to the censoring procedure.

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; Algorithm 1),

accurately estimates true tag abundance by separating legitimate reads from incorrectly amplified reads through a novel application of hierarchical clustering, and sets appropriate thresholds for the amplified reads through a novel application of the zero-truncated Poisson distribution. The effects of RASTA on differential gene expression testing, both in terms of power and ranking of results, are investigated. While we motivate the procedure through gene expression simulations, the method is general enough to be applied to DNA methylation and histone modification studies as well.

2   Robust Adjustment of Sequence Tag Abundance (RASTA)

RNA-seq reads can be assumed to be generated by two distinct processes: legitimate reads (including natural duplication) and amplification bias. As the number of reads increases for a given gene, either through increased transcription or through increased sequencing depth, the number of legitimately duplicated reads is expected to increase. For a given mapped read, we define "read count" as the number of observed mapped reads starting at the same base in the genome. Let

$$x_i^g \sim Poisson(\gamma_g = \frac{\lambda_g}{L_g}) \qquad\qquad i = 1, \ldots, n; \;\; x_i^g \geq 1 \qquad\qquad (2)$$

be the read counts for the n 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. Typically γ_g is estimated as

$$\hat{\gamma}_g = \frac{\sum_{i=1}^n x_i^g}{n}, \qquad\qquad (3)$$

but as x_i^g are restricted to be positive only, $\hat{\gamma}_g$ provides a biased estimate of γ_g. Instead, we model the legitimate base counts for a given gene using a zero-truncated $Poisson(\gamma_g^*)$ (ZTP) distribution (Yee and Wild, 1996). Let

$$\hat{\gamma_g^*} = E(X \cdot I(X \geq 1)) = \sum_{x=1}^{\infty} \frac{(\gamma_g)^x}{(x-1)!} e^{-\gamma_g} = \gamma_g \cdot Pr(Y \geq 1) \qquad\qquad (4)$$

be the estimate of γ_g obtained from the ZTP distribution. This value is readily estimated via the VGAM package (Yee, 2010) in R.

For a given value of $\hat{\gamma_g^*}$, 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 we define T_g^* as the 95th percentile of the $Poisson(\hat{\gamma_g^*})$ distribution. Then, for each x_i^g, define

$$y_i^g = min(x_i^g, T_g^*) \qquad\qquad Pr(X^{ZTP} \leq T_g^*) = 0.95, \qquad\qquad (5)$$

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

$$DGE_g = \sum_i y_i^g. \qquad\qquad (6)$$

In order to estimate γ_g from observed data, we first partition the read counts into legitimate reads and amplification biased reads. To achieve this, we apply hierarchical clustering to the unique

values of $x_1^g, x_2^g, \ldots, x_n^g$ for a given gene g. Unique x_i^g are used here to decrease computation time. For a pair of read counts (x_i^g, x_j^g) from gene g, we define the distance (or dissimilarity) to be

$$d_{ij}^g = \frac{|x_i^g - x_j^g|}{|x_i^g + x_j^g|}, \tag{7}$$

known as the Canberra distance (Lance and Williams, 1966, 1967). The Canberra distance is less sensitive to large values when compared to the Manhattan or Euclidean distances (Krause, 1987) and is appropriate when detecting deviations from normal observations (Krebs, 1989; Emran and Ye, 2001). Hierarchical clustering is employed with complete linkage based on the distance matrix D^g. Under complete linkage, clusters with the smallest maximum pairwise distance are merged in each step. After clustering, the resulting hierarchical tree is divided into two clusters that represent the legitimate reads and the biased reads. Read counts assigned to the cluster with the lowest mean are then used to estimate γ_g as described previously, under the assumption that amplification bias produces larger counts than expected under the Poisson generating process for legitimate reads.

Algorithm 1. Robust Adjustment of Sequence Tag Abundance (RASTA) Algorithm

0. Summarize read information as the number of reads starting at each base $(x_1^g, x_2^g, \ldots, x_n^g)$
    for each gene, ignoring bases for which no reads were observed (Figure 2).

1. Estimate pairwise distances between read counts (x_i^g, x_j^g) for gene g as $d_{ij}^g = \frac{|x_i^g - x_j^g|}{|x_i^g + x_j^g|}$.

2. (a) Perform hierarchical clustering on the unique base count values for each gene using complete linkage.
    (b) Define legitimate base counts as those assigned to the cluster with the smallest mean.

3. Estimate the mean legitimate base count γ_g via a zero-truncated Poisson distribution.

4. Set the threshold $T_g = max(1, T_g^*)$, where T_g^* is defined as the 95th percentile
    of the $Poisson(\gamma_g^*)$ distribution.

5. Set the adjusted read count as $y_i^g = min(x_i^g, T_g)$.

6. Set digital gene expression as $DGE = \sum_i y_i^g$ for each gene.

There are issues that arise from this process when genes experience no sign of amplification bias. In cases of low digital gene expression (DGE) relative to gene length, the clustering algorithm generally assigns the greatest read counts to the "bias" cluster. In these cases, RASTA may effectively ignore the duplicate reads (i.e., $T_g^* = 1$), which is the same as censoring; this results in a lower estimate of γ_i^g since naturally duplicated reads are not used in the estimation process. However, RASTA typically allows for larger DGE values when compared to ad hoc methods for genes with high DGE relative to gene length or those with amplification bias.

3 Simulation Study

3.1 Simulation Settings

A simulation study was conducted to evaluate and compare the performance of RASTA to "censoring." We simulated RNA-seq data by fixing the total number of genes at 1,000. The frequency of DGE counts per gene approximately follows a power-law (Balwierz et al., 2009), and as such, we simulated the true DGE rate for gene g as

$$\lambda_g \sim exp(Pareto(location = 3.5, scale = 7)) \qquad g = 1, \ldots, 1000 \tag{8}$$

(Auer and Doerge, 2011). We incorporate amplification bias by setting the prevalence of bias to be $\pi_g^{bias} = .001$ (or 1 out of every 1000 bases), and the bias DGE count to be

$$\lambda_g^{bias} = Uniform(10, 1000) \qquad\qquad g = 1, \ldots, 1000. \qquad\qquad (9)$$

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 read counts can exist in tens of thousands (Marioni et al., 2008; Lister et al., 2008). These conservative estimates were chosen to incorporate any recent (and future) improvements to the biological protocols developed for NGS technologies. Gene lengths were simulated based on the mouse and *Drosophila melanogaster* annotation databases from Ensembl (Flicek et al., 2011) with

$$L_g \sim exp(Normal(\mu = 8, \sigma = 2)) \qquad\qquad g = 1, \ldots, 1000. \qquad\qquad (10)$$

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

$$Poisson(\gamma_g = \frac{\lambda_g}{L_g}) \qquad\qquad g = 1, \ldots, 1000, \qquad\qquad (11)$$

and the counts arising from amplification bias follow

$$Poisson(\pi_g^{bias} \frac{\lambda_g^{bias}}{L^g}) \qquad\qquad g = 1, \ldots, 1000. \qquad\qquad (12)$$

For each gene, the counts were preprocessed by either truncating all counts to 1 (i.e., the current censoring practice) or assessed via RASTA. These modified counts were then added, and resulted in 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 differentially expressed (500) and non-differentially expressed (500) genes were generated for three biological replicates in two treatments. DGE rates for each gene were generated as follows: for differentially expressed genes, means were sampled separately from (8), yielding $\lambda_g^{T_1}$ and $\lambda_g^{T_2}$ for treatments T_1 and T_2 respectively; for non-differentially expressed genes, the means were sampled together (λ_g). For each simulated data set, both RASTA and "censoring" were applied to the observed base counts. Resulting 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 (Benjamini and Hochberg, 1995) in edgeR.

3.2   Comparing Adjusted DGE to True DGE

Figure 3 illustrates the effectiveness of RASTA to maintain accurate DGE estimates. In general, RASTA captures the true transcription rate more accurately than the censoring method, particularly for genes with larger γ_g values. In cases with lower γ_g values, RASTA often returns the same values as the "censoring" approach. In other words, RASTA generally does *no worse* than the "censoring" approach, and in many cases drastically out-performs its counterpart.
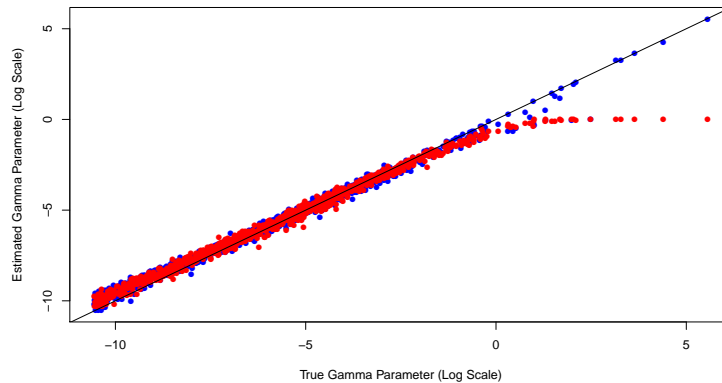
Figure 3: Simulation results for amplification bias correction methods. The RASTA method is displayed in blue, while the censoring method is displayed in red. Since natural read duplication is a function of both transcription rate (λ) and gene length (L), adjusted digital gene expression (DGE) values were scaled by length and plotted against the true $\gamma = \frac{\lambda}{L}$ values. As the true γ values increase, RASTA more accurately adjusts the observed DGE counts relative to the current "censoring" procedure, which is bound above at $\hat{\gamma} = 1$ (or zero in log scale as displayed above).

3.3   Effects of RASTA on Differential Expression Analyses

For the simulations previously described, statistical power and false discovery rates (FDR) were estimated by calculating averages power and FDR ($q = 0.05$) across the simulations. RASTA yields similar statistical 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 realized FDRs were similar, summaries comparing true and estimated log fold changes showed greater accuracy under the RASTA method. This is expected when considered in conjunction with the results in Figure 3. To illustrate this, estimated log fold changes were regressed against true log fold changes (Figure 4) for the purpose of demonstrating accuracy (i.e., the relative closeness of the RASTA and "censoring" approaches to the identity line). The regression slope for RASTA was considerably closer to 1.0 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.

In these simulations, the resulting order of of the ranked genes varies dramatically between censoring and RASTA. We investigated this phenomenon in several ways. First, we inspected the 100 genes found to be highly differentially expressed between the RASTA and censoring methods, and calculated the number of genes that appear in both lists. Over the 1000 simulations, the average number of matching results between the two lists was only approximately 14 (mean: 14.30, standard deviation: 3.84). Generalizing this, we compared the number of matching genes in lists of top n genes ($n = 1, \ldots, 1000$), and the results are displayed in Figure 5. For list lengths less than 200 genes, the number of matching elements between the two methods is quite low. As the list lengths exceed 400 genes, the rate of matching increases. We also compared the ranks between the two full lists using Friedman's Test (Friedman, 1937, 1940). The average p-value of Friedman's Test between RASTA and "censoring" is 0.03 (standard deviation: 0.12), which provides additional support that the order of results is quite different between the two methods. Finally, we compared
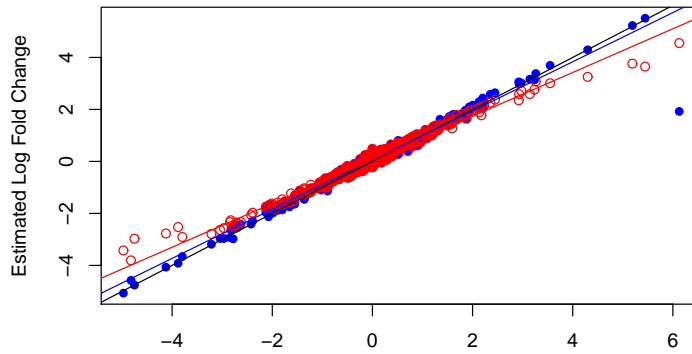
Figure 4: Simulation results for true vs. estimated log fold change when comparing RASTA versus "censoring." As the true log fold change values increase (in absolute value), RASTA (blue) more accurately estimates the log fold change relative to the censoring (red) procedure (regression slopes: 0.95 vs 0.83, respectively).
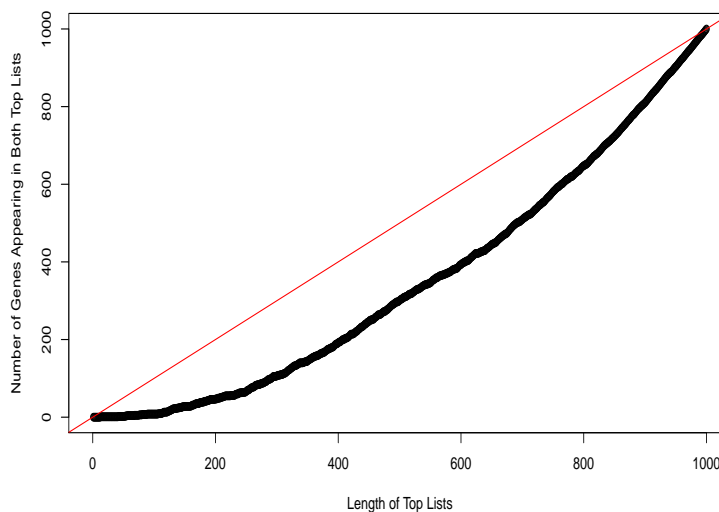

Figure 5: Simulation results for the number of matching elements between the top lists of length n. The superimposed red line indicates perfect matching between the lists. The rate of matching remains low for the top ranked genes in lists of less than 200 genes.
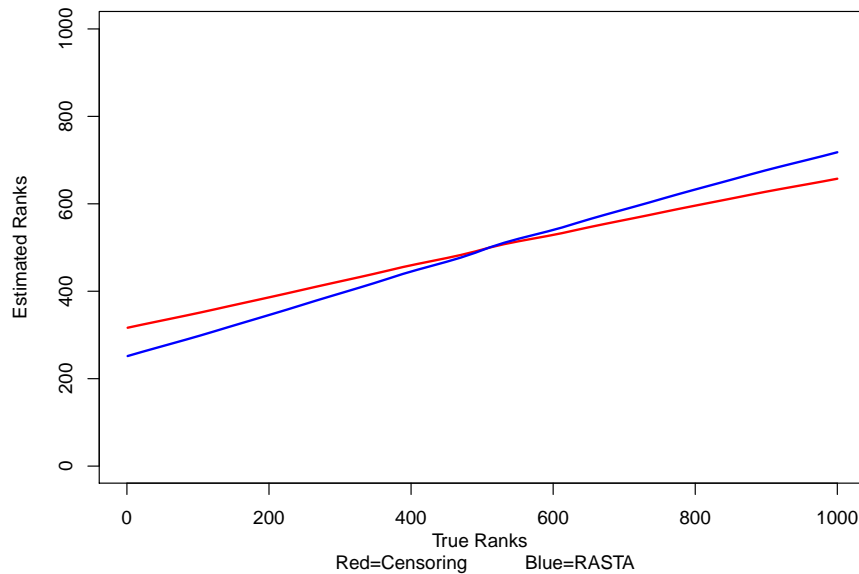
Figure 6: Simulation results comparing estimated vs. true differential gene expression ranks for RASTA and "censoring" approaches. True ranks are defined as the Canberra distance between the simulated rates of transcription. Estimated ranks are defined by the ordered p-values after differential expression tests are performed. The RASTA approach (blue) more accurately estimates true gene ranks than the censoring procedure (red).

the true rankings of genes to the estimated rankings produced by differential testing (Figure 6). On average across the simulations, RASTA more accurately captures the true ranks when compared to the censoring method. These results not only support the use of RASTA when controlling for amplification bias, but they also indicate a real need to address amplification bias as a problem that is currently affecting the results being reported in the literature.

3.4   Gene Length Bias

The prevalence of natural read duplication depends on quantification of gene expression (i.e., the DGE) and gene length. For a given level of DGE, natural read duplication is more likely to occur in shorter genes than in longer genes, and as such, shorter genes are generally more affected by the current censoring procedure. Because of this, there is a bias toward longer genes when testing for differential gene expression. By more accurately estimating DGE, especially for shorter genes with high DGE, we are able to all but eliminate length bias in our simulations as average DGE levels increase (Figure 7).
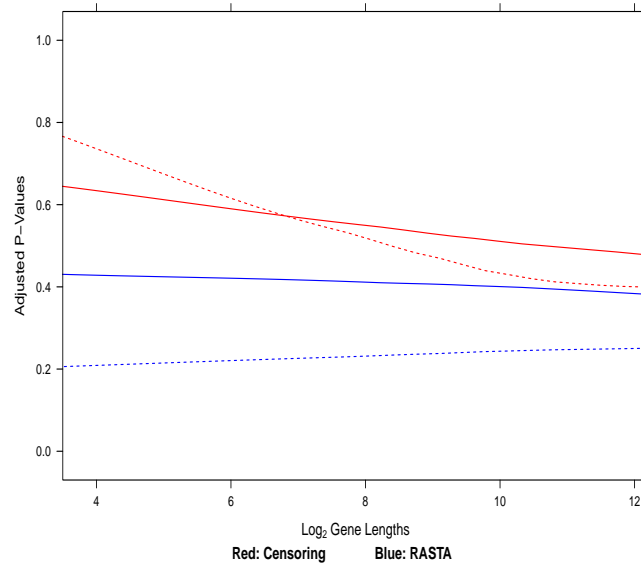
Figure 7: Gene length bias simulation results. The censoring method is presented in red, while the RASTA method is presented in blue. The solid lines represent simulated gene expression levels based on (Auer and Doerge, 2011). The dashed lines represent a doubling, on average, of DGE levels. For the 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.

4   Real Data Analysis

4.1   Data and Methods

The censoring and RASTA approaches were employed 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, 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 using a test based on the 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 (Benjamini and Yekutieli, 2001).

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

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.

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; Arand et al., 2012). The RASTA analysis yielded many more statistically significant differentially expressed genes than the censoring method (8912 and 2855 genes, respectively). This increase in number of differentially expressed gene results is in concordance with the biological knowledge of the two *Arabidopsis* lines (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.

5   Discussion

As the costs for sequencing decrease, researchers will require greater and greater sequencing depth simply due to the demand of accurate sequencing. As sequencing depth increases, the occurrence of legitimately duplicated reads will increase. As such, the manner in which amplification bias is controlled is likely to have a significant impact on any RNA-seq study. The choice of control procedures has the potential to affect the order and importance of significantly differentially expressed genes since the individual gene expression estimates may change considerably (Figure 2). Specifically, it is to this point that we believe RASTA will have the most effect. Since confirmatory studies often target the most differentially expressed genes (i.e., the genes with the lowest p-values), the ordering of results plays an important role in downstream analyses. In other words, while RASTA may not provide more statistical power to detect differences between two treatments in all settings, it may provide a vastly different ordering of significant results.

RASTA				
GO Term	Ontology Description	Input	Reference	Adj. p-value
GO:0009791	Post-embryonic development	382	705	4.2e-76
GO:0034641	Cellular nitrogen compound metabolic process	236	506	5.7e-33
GO:0032501	Multicellular organismal process	664	2094	2.4e-24
GO:0009987	Cellular process	3036	11684	5.9e-24
GO:0007275	Multicellular organismal development	640	2020	1.4e-23
GO:0010035	Response to inorganic substance	138	279	3.9e-22
GO:0033036	Macromolecule localization	194	462	1.6e-20
GO:0003006	Reproductive developmental process	341	978	2.1e-19
GO:0048856	Anatomical structure development	542	1726	2.8e-19
GO:0008152	Metabolic process	2720	10614	2.9e-19
Censoring				
GO Term	Ontology Description	Input	Reference	Adj. p-value
GO:0009628	Response to abiotic stimulus	209	1471	2.2e-19
GO:0050896	Response to stimulus	440	4057	8.2e-17
GO:0009791	Post-embryonic development	119	705	1.6e-16
GO:0006950	Response to stress	279	2320	3e-16
GO:0044262	Cellular carbohydrate metabolic process	84	417	3.3e-16
GO:0010876	Lipid localization	18	24	6.2e-14
GO:0010035	Response to inorganic substance	62	279	6.5e-14
GO:0009266	Response to temperature stimulus	84	485	2.2e-12
GO:0042221	Response to chemical stimulus	239	2085	5.6e-12
GO:0034641	Cellular nitrogen compound metabolic process	81	506	3.5e-10

Table 2: Gene Set Enrichment Analysis results (top ten 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 number of genes with each ontology in the significant gene lists and the *Arabidopsis* reference are listed in the "Input" and "Reference" columns, respectively. 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.

References

Anders, S. and W. Huber (2010). Differential expression analysis for sequence count data. *Genome Biology 11*.

Arand, J., D. Spieler, T. Karius, M. Branco, D. Meilinger, A. Meissner, T. Junewein, G. Xu, H. Leionhardt, V. Wolf, and J. Walter (2012). In vivo control of cpg and non-cpg dna methylation by dna methyltransferases. *PLoS Genetics 8*, e1002750.

Auer, P. and R. Doerge (2011). A two-stage poisson model for testing rna-seq data. *Statistical Applications in Genetics and Molecular Biology 10*, 26.

Balwierz, P., P. Carninci, C. Daub, J. Kawai, Y. Hayashizaki, W. Van Belle, C. Beisel, and E. van Nimwegen (2009). Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepcage data. *Genome Biology 10*, R79.

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

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

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

Emran, S. and N. Ye (2001). Robustness of canberra metric in computer intrusion detection. *Proceedings of the 2001 IEEE, Workshop on Information Assurance and Security 1*, 1.

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

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association 32*, 675–701.

Friedman, M. (1940). A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics 11*, 86–92.

Krause, E. (1987). *Taxicab Geometry*. Dover.

Krebs, C. (1989). *Ecological Methodology*. Harper-Collins, New York.

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

Lance, G. and W. Williams (1967). Mixed-data classificatiory programs, i.) agglomerative systems. *Australian Computer Journal 1*, 15–20.

Lister, R., R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, A. H. Millar, and J. R. Ecker (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., M. Egholm, W. Altman, S. Attiya, and J. Bader (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature 437*, 376–380.

Marioni, J., C. Mason, S. Mane, M. Stephens, and Y. Gilad (2008). Rna-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research 18*, 1509–1517.

Mikkelsen, T., M. Ku, D. Jaffe, B. Issac, and E. Lieberman (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature 448*, 553–560.

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.