CORRECTING FOR AMPLIFICATION BIAS IN NEXT-GENERATION SEQUENCING DATA

Douglas Baumann

R. W. Doerge

Follow this and additional works at: https://newprairiepress.org/agstatconference

Part of the Agriculture Commons, and the Applied Statistics Commons

Recommended Citation


This work is licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License.
CORRECTING FOR AMPLIFICATION BIAS IN NEXT-GENERATION SEQUENCING DATA

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

*Corresponding 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...
Sample mRNA from subjects
Randomly fragment the mRNA via sonication
Size select fragmented mRNA, attach adapters, amplify via PCR, and sequence
Create contigs or align to a reference genome

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

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

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

In edgeR, $\phi_{gj}$ 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 ($\phi_{gj}$), or as a weighted average of $\phi_0$ and $\phi_g$. 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, $\phi_{gj}$ 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.

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

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

be the read counts for the $n$ bases with observed reads for a given gene $g$, where $\lambda_g$ and $L_g$ are the overall transcription rate and length for gene $g$, respectively. Typically $\gamma_g$ is estimated as

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

but as $x_i^g$ are restricted to be positive only, $\hat{\gamma}_g$ provides a biased estimate of $\gamma_g$. 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

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

be the estimate of $\gamma_g$ 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

$$y_i^g = \min(x_i^g, T_g^*) \quad Pr(X_{\text{ZTP}} \leq T_g^*) = 0.95,$$

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

$$DGE_g = \sum_i y_i^g.$$
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

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

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 $\gamma_g$ 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 $\gamma_g$ 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 $\text{DGE} = \sum_i y_i^g$ 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 $\gamma_g^*$ 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

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

(8)
We incorporate amplification bias by setting the prevalence of bias to be $\pi^\text{bias} = 0.001$ (or 1 out of every 1000 bases), and the bias DGE count to be $\lambda^\text{bias} = \text{Uniform}(10, 1000)$ for each gene $g = 1, \ldots, 1000$. (9)

The value of $\pi^\text{bias}$ and the upper bound on $\lambda^\text{bias}$ 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 $L_g \sim \text{exp}(\text{Normal}(\mu = 8, \sigma = 2))$ for each gene $g = 1, \ldots, 1000$. (10)

For a given gene with parameters $\lambda_g$ and $\lambda^\text{bias}_g$, the legitimate reads follow $\text{Poisson}(\gamma_g = \frac{\lambda_g}{L_g})$ for each gene $g = 1, \ldots, 1000$, (11)

and the counts arising from amplification bias follow $\text{Poisson}(\frac{\pi^\text{bias}_g \lambda^\text{bias}_g}{L_g})$ for each gene $g = 1, \ldots, 1000$. (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 $\lambda^T_1_g$ and $\lambda^T_2_g$ for treatments $T_1$ and $T_2$ respectively; for non-differentially expressed genes, the means were sampled together ($\lambda_g$). 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 $\gamma_g$ values. In cases with lower $\gamma_g$ 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 ($\lambda$) and gene length ($L$), adjusted digital gene expression (DGE) values were scaled by length and plotted against the true $\gamma = \frac{\gamma}{L}$ values. As the true $\gamma$ values increase, RASTA more accurately adjusts the observed DGE counts relative to the current “censoring” procedure, which is bound above at $\gamma = 1$ (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 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].
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.

<table>
<thead>
<tr>
<th></th>
<th>met1-3</th>
<th>Col-0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Reads</td>
<td>5997689</td>
<td>6283230</td>
</tr>
<tr>
<td>Unique Reads</td>
<td>2991256</td>
<td>1264135</td>
</tr>
<tr>
<td>Single bases with ≥ 5 reads</td>
<td>139972</td>
<td>285610</td>
</tr>
<tr>
<td>Single bases with ≥ 10 reads</td>
<td>38718</td>
<td>72227</td>
</tr>
<tr>
<td>Single bases with ≥ 100 reads</td>
<td>232</td>
<td>849</td>
</tr>
<tr>
<td>Max number of reads at a single base</td>
<td>5525</td>
<td>17063</td>
</tr>
</tbody>
</table>
### 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 Hochberg, 1995). 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


