Received: 29 July 2018 | Revised: 19 October 2018 | Accepted: 8 November 2018

DOI: 10.1111/1755-0998.12965

RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

Nonrandom RNAseq gene expression associated with RNAlater and flash freezing storage methods

Courtney N. Passow[1,*] | Thomas J. Y. Kono[2,*] | Bethany A. Stahl[3] | James B. Jaggard[3] | Alex C. Keene[3] | Suzanne E. McGaugh[1]

[1]Ecology, Evolution, and Behavior, University of Minnesota, Saint Paul, Minnesota

[2]Minnesota Supercomputing Institute, University of Minnesota, Minneapolis, Minnesota

[3]Department of Biological Sciences, Florida Atlantic University, Jupiter, Florida

Correspondence
Suzanne E. McGaugh, Ecology, Evolution, and Behavior, University of Minnesota, Saint Paul, Minnesota.
Email: smcgaugh@umn.edu

Funding information
National Institute of General Medical Sciences, Grant/Award Number: 1R01GM127872-01 ; University of Minnesota, Grant/Award Number: A15-32; Florida Atlantic University

Abstract

RNA sequencing is a popular next-generation sequencing technique for assaying genome-wide gene expression profiles. Nonetheless, it is susceptible to biases that are introduced by sample handling prior gene expression measurements. Two of the most common methods for preserving samples in both field-based and laboratory conditions are submersion in RNAlater and flash freezing in liquid nitrogen. Flash freezing in liquid nitrogen can be impractical, particularly for field collections. RNAlater is a solution for stabilizing tissue for longer-term storage as it rapidly permeates tissue to protect cellular RNA. In this study, we assessed genome-wide expression patterns in 30-day-old fry collected from the same brood at the same time point that were flash-frozen in liquid nitrogen and stored at −80°C or submerged and stored in RNAlater at room temperature, simulating conditions of fieldwork. We show that sample storage is a significant factor influencing observed differential gene expression. In particular, genes with elevated GC content exhibit higher observed expression levels in liquid nitrogen flash-freezing relative to RNAlater storage. Further, genes with higher expression in RNAlater relative to liquid nitrogen experience disproportionate enrichment for functional categories, many of which are involved in RNA processing. This suggests that RNAlater may elicit a physiological response that has the potential to bias biological interpretations of expression studies. The biases introduced to observed gene expression arising from mimicking many field-based studies are substantial and should not be ignored.

KEYWORDS

GC proportion, gene expression, gene length, liquid nitrogen, RNAlater, technical variation

1 | INTRODUCTION

High-throughput sequencing technologies, such as RNA sequencing methods, have revolutionized the quantification of genome-wide expression patterns across a broad range of fields in biological sciences (López-Maury, Marguerat, & Bähler, 2008; Wang, Gerstein, & Snyder, 2009). However, storage and RNA extraction methods prior to RNA-seq library preparation exert substantial impacts on biological studies and often account for the majority of variation in a data set if conditions and protocols are not identical across all samples (Todd, Black, & Gemmell, 2016). With the rise of RNAlater (Ambion, Invitrogen) as a popular storage method in field-based studies (De Smet et al., 2017; Wille et al., 2018), it is important to quantify if there are systematic biases in gene expression when samples are preserved in RNAlater vs. flash-frozen in liquid nitrogen. In our literature review, however, we could find few direct comparisons of

*These Authors contributed equally to this work.

RNAseq data obtained from the most common field preservation method RNAlater and the "gold standard" of flash freezing samples in liquid nitrogen (Alvarez, Schrey, & Richards, 2015; Wolf, 2013; but see Cheviron, Carling, & Brumfield, 2011; Choi, Ray, Lai, Alwood, & Globus, 2016). Further, few studies examine whether a systematic bias due to gene characteristics exists for samples preserved in RNAlater (Bray et al., 2010).

Currently, two of the most common methods for RNA preservation and storage are flash freezing in liquid nitrogen and preservation in aqueous sulphate salt solutions, such as commercially available RNAlater. Flash freezing, usually through the use of immersing the sample in dry ice or liquid nitrogen, is the most preferred means of stabilizing tissue samples for downstream analysis (Wolf, 2013). While preferred, it can often be difficult to access and transport dry ice or liquid nitrogen, particularly in field conditions (Mutter et al., 2004). Hence, in the past decade, it has become common practice, especially in field environments, to store RNAseq-destined samples in RNAlater, which minimizes the need to readily process samples or chill the tissue. RNAlater can rapidly permeate tissue to stabilize and protect RNA (Chowdary et al., 2006; Florell et al., 2001). Likewise, RNAlater-immersed samples can be stored safely at room temperature for a week and longer when stored at colder temperatures. Though, common practice in field conditions is to store samples in RNAlater for much longer than a week (Camacho-Sanchez, Burraco, Gomez-Mestre, & Leonard, 2013; Gorokhova, 2005). While the exact ingredients of commercial RNAlater are proprietary, the Material Safety Data Sheet lists inorganic salt as the major component and the homemade versions contain ammonium sulphate, sodium citrate, ethylenediaminetetraacetic acid (EDTA) and adjustment of pH using sulphuric acid.

In this study, we quantified the effects of storage condition on gene expression and examined differentially expressed genes for specific characteristics to assay for systematic bias. Individual, Mexican tetra fry (*Astyanax mexicanus*), were collected from the same brood and stored immediately in liquid nitrogen ($N = 6$) or RNAlater ($N = 5$). We specifically asked (a) Does storage condition affect patterns of differential gene expression and if so, (b) Are these effects on gene expression nonrandom, such that genes with certain features are differentially affected by storage condition? We found that a majority of the variation in gene expression was explained by storage condition. Likewise, we found that genes with higher GC content exhibited higher expression values in liquid nitrogen than RNAlater. Based on these findings, we conclude that RNAlater storage at room temperature for extended periods of time may potentially bias biological conclusions of RNAseq experiments.

2 | METHODS

2.1 | Sample collection

Samples for the transcriptome analyses were collected from a surface population of *Astyanax mexicanus* (total of eight parents) that had been reared in the Keene laboratory at Florida Atlantic University for multiple generations. Parental fish were derived from wild-caught Río Choy stocks originally collected by William Jeffery. To minimize variation outside of storage methods, all individuals were collected from the same clutch (fertilized on 08-December-2016). Fish were raised in standard conditions, and three days prior to experiment, fish were transferred into dishes with 12–21 fish per dish in a 14:10 light–dark cycle. These fish were a part of a larger experiment, so fish were kept in total darkness for 24 hr prior to sampling, and sampled at 16:00 h (10 p.m.). Five individuals were sampled with forceps and stored in RNAlater, and six individuals were flash-frozen in liquid nitrogen and stored at −80°C. Fry at 30 days postfertilization (dpf) were <5 mm long, transparent and highly permeable. To mimic field conditions, RNAlater individuals were stored at room temperature for 17 days (Camacho-Sanchez et al., 2013; Kono, Nakamura, Ito, Tomita, & Arakawa, 2016). Procedures for all experiments performed were approved by the Institutional Animal Care and Use Committee at Florida Atlantic University (Protocol #A15-32).

2.2 | RNA extraction, library preparation and sequencing

For RNA isolation, all individuals were processed within a week of each other (between 19-January-2017 and 24-January-2017), and RNAlater-stored individuals were processed 17 days after initial storage (24-January-2017; Supporting Information Table S1) with the same researcher performing all extractions. Whole organisms (<30 mg of tissue) were homogenized using Fisherbrand pellet pestles and cordless motor (Fisher Scientific) in the lysate buffer RLT plus. Total RNA was extracted using the Qiagen RNAeasy Plus Mini Kit (Qiagen) and quantified using NanoDrop Spectrophotometer (Thermo Fisher Scientific), Ribogreen assay (Thermo Fisher Scientific) and Bioanalyzer RNA 6000 Nano assay (Agilent) to obtain RNA integrity numbers (RIN). All cDNA libraries were constructed at the University of Minnesota Genomics Center on the same day in the same batch. In brief, a total of 400 ng of RNA was used to isolate mRNA via oligo-dT purification. dsDNA was constructed from the mRNA by random-primed reverse transcription and second-strand cDNA synthesis. Strand-specific cDNA libraries were then constructed using TruSeq Nano Stranded RNA kit (Illumina), following manufacturer protocol. Library quality was assessed using Agilent DNA 1000 assay on a Bioanalyzer. To minimize batch effects, barcoded libraries were then pooled and sequenced across multiple lanes of an Illumina HiSeq 2500 to produce 125-bp paired-end reads at University of Minnesota Genomics Center (Supporting Information Table S1). All sequence data were deposited in the short read archive (Study Accession ID: RNAlater: SRX3446133, SRX3446136, SRX3446135, SRX3446155, SRX3446156; liquid nitrogen: SRS2736519, SRS2736520, SRS2736523, SRS2736524, SRS2736525, SRS2736526).

2.3 | RNAseq quality check

The raw RNAseq reads were quality checked using Fastqc (Andrews, 2014) and trimmed to removed adapters using the program Trimmomatic version 0.33 (Bolger, Lohse, & Usadel, 2014). Trimmed reads were mapped to the *Astyanax mexicanus* reference genome (version 1.0.2; GenBank Accession Number: GCA_000372685.1; McGaugh et al., 2014). Mapping was conducted using the splice-aware mapper STAR (Dobin et al., 2013), because it yielded the higher alignment percentage and quality compared to a similar mapping programme (HISAT2, results not shown; Kim, Langmead, & Salzberg, 2015). We used Stringtie (version 1.3.3d; Pertea, Kim, Pertea, Leek, & Salzberg, 2016; Pertea et al., 2015) to quantify number of reads mapped to each gene in the reference annotation set of the *A. mexicanus* genome, and used the python script provided with Stringtie (prepDE.py) to generate a gene counts matrix (Pertea et al., 2016). R (Team RC, 2014) was used to compare RIN between liquid nitrogen and RNAlater treatments using a nonparametric Kruskal–Wallis test.

2.4 | Variation in gene expression

To visualize changes in observed gene expression, we performed principal components analysis on a gene counts matrix. Genes with less than 100 counts across all samples were removed from the matrix because genes with low counts bias the differential expression tests (Love, Huber, & Anders, 2014). The resulting counts were decomposed into a reduced dimensionality data set with the *prcomp ()* function in R (Team RC, 2014). To understand the extent storage method affected the ability to detect inter-individual variation, we calculated the coefficient of variation in gene expression for each gene under both storage conditions.

To identify genes that showed the largest difference in observed gene expression between storage conditions, we performed a differential expression analysis between samples flash-frozen in liquid nitrogen (N = 6) and samples stored in RNAlater (N = 5) using DESeq2 (Love et al., 2014). DESeq2 normalizes expression counts for each sample and then fits a negative binomial model for counts for each gene. Samples with the same storage condition were treated as replicates (i.e., the variation due to storage was assumed to be greater than variation among biological samples). This was confirmed in the PCA plot (Figure 1), where PC1 linearly separated samples based on their treatments. p-values for differential expression were adjusted based on the Benjamini–Hochberg algorithm, using a default false discovery rate of at most 0.05 (Love et al., 2014). Genes were labelled as differentially expressed if the Benjamini–Hochberg adjusted p-value was <0.05. Log2(RNAlater/liquid nitrogen) values were calculated with DESeq2 and exported for further analysis.

2.5 | Linear model to determine factors influencing differential expression

To identify the factors that contribute to the variability in gene expression between preservation methods, we fit a linear model of observed gene expression of all genes as a function of various genomic characteristics. We tested the contributions of mean expression level, annotated coding gene length, exon number, GC content, presence or absence of simple sequence repeats and presence or absence of a homopolymer tract to differences in observed gene expression between preservation methods. We used the log2(RNAlater/liquid nitrogen) values from DESeq2 as the measure of change in observed gene expression and the mean of normalized counts across all samples as the mean expression level. The annotated gene length was calculated as the length of the coding region of the longest transcript from each gene. A simple sequence repeat was defined as two or more nucleotides repeated at least three times in tandem, and a homopolymer tract was defined as a single nucleotide repeated at least six times in tandem in the reference genome. Repeat presence or absence was based only on the reference genome sequence and were not scored to be polymorphic in the sample. Reference data were downloaded from Ensembl BioMart (Durinck et al., 2005; Durinck, Spellman, Birney, & Huber, 2009), and custom Python scripts were used to extract exon number and calculate coding length and GC content. The presence/absence of a simple sequence repeat and the presence/absence of a homopolymer repeat were scored with a custom Python script. All scripts used for analysis are available on our GitHub repository. Notably, the reference genome is a Pachón cavefish, and it is conceivable that some homopolymers and sequence repeats may not be identical in the surface fish.

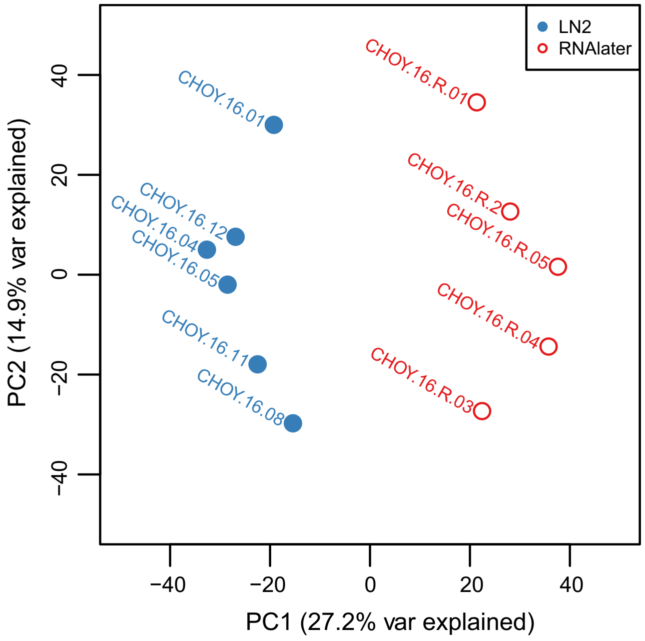

FIGURE 1 Principal components analysis plot showing PC1 and PC2 for each sample. RNAlater samples (red, open circles) are linearly separated from liquid nitrogen samples (blue, closed circles) by PC1 [Colour figure can be viewed at wileyonlinelibrary.com]

We performed model selection on a series of linear models using likelihood ratio tests of nested models. The “full model” was as follows:

$$\begin{aligned} Y =& \alpha + \beta_0 M + \beta_1 G + \beta_2 L + \beta_3 E + \beta_4 S + \beta_5 H \\ &+ \beta_6 (G \times S) + \beta_7 (G \times H) + \varepsilon, \end{aligned} \quad ()$$

where Y is log2(RNAlater/liquid nitrogen) of expression between treatments, M is the normalized mean expression value across all samples, G is GC content, L is coding gene length, E is the total number of exons in the gene, S is simple sequence repeats (SSR) presence/absence and H is homopolymer presence/absence. GC content, coding length of the gene and exon number were treated as continuous variables, and SSR presence and homopolymer presence were treated as categorical variables. Model selection proceeded by testing the contributions of the interaction terms to the variance explained and removing them if not significant. We tested the terms with the lowest nonsignificant t-values in the regression and removed them if they did not significantly improve model fit.

2.6 | Annotation of differentially expressed genes

Because we expected most of the variation was going to be explained by a technical variable (i.e., preservation and storage), we did not expect biologically meaningful functional annotations. However, we conducted annotation analyses using differentially expressed genes at the 0.05 false discovery rate. Zebrafish (*Danio rerio*) genes that were one-to-one orthologs with *Astyanax* were used for a gene ontology (GO) term enrichment analysis. PANTHER analysis (Mi et al., 2016; http://pantherdb.org/tools/compareToRefList.jsp) was run using only 1:1 orthologs between zebrafish and *Astyanax* with database current as of 30-April-2018. Within the PANTHER suite, we used PANTHER v13.1 overrepresentation tests (i.e., Fisher's exact tests with FDR multiple test correction) with the Reactome v58, PANTHER proteins, GoSLIM, GO and PANTHER Pathways. The target list was the zebrafish genes that were orthologous to differentially expressed *Astyanax* genes, and the background list was all zebrafish genes genome-wide. We confirmed these results by performing GO term enrichment with the GOrilla webserver (Eden, Navon, Steinfeld, Lipson, & Yakhini, 2009) (http://cbl-gorilla.cs.technion.ac.il/), with a database current as of 06-October-2018. The target list was the zebrafish genes that were orthologous to differentially expressed *Astyanax* genes, and the background list was all one-to-one orthologs between zebrafish and cavefish in our expression data set.

2.7 | Script availability

Scripts to perform all data QC and processing are available at https://github.com/TomJKono/CaveFish_RNAlater.

3 | RESULTS

3.1 | Mapping statistics and annotation

RNA sequencing from whole, 30 days postfertilization individuals yielded a total of 108,874,500 reads for individuals stored in liquid nitrogen (mean = 18,145,750 ± SD 1,938,410 per individual; N = 6) and 82,448,455 reads for individuals stored in RNAlater (mean = 16,489,691 ± SD 1,890,519 per individual; N = 5; Table 1). While all RIN scores from the extracted total RNA passed the threshold (>7; Supporting Information Table S1), to proceed into library preparation, RIN scores were significantly different between RNAlater and liquid nitrogen treatments (Kruskal–Wallis chi-squared = 7.6744, df = 1, P-value = 0.0056; RNAlater mean RIN = 8.60, liquid nitrogen mean RIN = 9.83).

TABLE 1 Reported are the number of reads (after adapter trimming) used as input for the mapping software (STAR), number of reads that uniquely mapped to the reference genome, and the per cent of reads that mapped to the reference genome. "Liquid N2" stands for liquid nitrogen

Sample name	Treatment	Input reads	Uniquely mapped reads	% Mapped
CHOY-16-01	Liquid N2	20,162,412	18,125,738	89.90%
CHOY-16-04	Liquid N2	15,760,631	13,812,190	87.64%
CHOY-16-05	Liquid N2	18,025,208	16,015,383	88.85%
CHOY-16-08	Liquid N2	16,368,007	14,584,314	89.10%
CHOY-16-11	Liquid N2	17,997,036	15,126,300	89.61%
CHOY-16-12	Liquid N2	20,561,206	18,221,558	88.62%
CHOY-16-R-01	RNAlater	17,984,846	15,643,479	86.98%
CHOY-16-R-03	RNAlater	17,064,911	14,913,653	87.39%
CHOY-16-R-04	RNAlater	13,585,649	11,809,525	86.93%
CHOY-16-R-05	RNAlater	15,692,250	13,716,160	87.41%
CHOY-16-R-2	RNAlater	18,120,799	15,851,038	87.47%

Total yield of reads and number of uniquely mapping reads were not significantly different between treatments (t = 1.4301; p = 0.1875). On average, samples mapped 88.17% of the reads to the *Astyanax mexicanus* genome (range: 86.93%-89.90%), with liquid nitrogen samples mapping on average 88.95% and RNAlater mapping 87.24%.

Filtering of the gene counts matrix to include only genes with ≥100 reads resulted in 15,515 genes being used for both clustering and differential expression analysis. Annotations were extracted from the *Astyanax mexicanus* annotation file (Astyanax_mexicanus.AstMex102.91.gtf). Distributions of raw and filtered gene expression counts are given in Supporting Information Figure S1.

The coefficients of variation between liquid nitrogen and RNAlater-preserved samples show a positive correlation (Supporting Information Figure S2, Kendall's Tau, τ = 0.267, p < 2e-16), suggesting that the genes that are highly variable in the liquid nitrogen treatment are also highly variable in RNAlater storage. Thus, we do not expect that the storage methods significantly impact the ability to detect variation among individuals. However, there are slightly more genes with higher coefficients of variation in liquid nitrogen than in RNAlater (9,043 genes) than vice versa (6,472 genes), suggesting that RNAlater may reduce variation among individuals.

3.2 | PCA and differentially expressed genes

Principal components analysis showed that the major axis of differentiation among the samples was treatment (Figure 1). This corresponds to the first principal component and explains 27.2% of the variation. Beyond the first principal component, the samples do not cluster into further discernible subgroups, suggesting that the main axis of differentiation among these samples is their storage conditions (Figure 1).

A total of 2,708 (17.5%) genes were significantly differentially expressed between treatments at the 0.05 significance level (Figure 2). Of these, 1,635 exhibited significantly lower observed expression in RNAlater than liquid nitrogen, and 1,073 exhibited significantly higher observed expression in RNAlater than in liquid nitrogen.

3.3 | Genomic characters contributing to differential expression

We identified four characteristics that contribute significantly to differential gene expression between treatments. Mean expression across samples, GC content, exon number, and interaction between GC content and SSR presence/absence were significant terms in the model (Table 2, Figure 3, Supporting Information Figure S3). GC content exhibited the largest coefficient. The coefficient for GC content is negative, suggesting that genes with higher GC content have a higher relative expression in liquid nitrogen than RNAlater (Supporting Information Figure S4). SSR presence also exhibited a nonsignificant association which resulted in higher relative expression in liquid nitrogen than RNAlater. Mean expression and exon number were significant, such that they exhibited a positive relationship with genes showing higher expression values in RNAlater (i.e., greater mean expression and more exons both related to higher expression in RNAlater). The small regression coefficients of these variables imply, however, that these factors have negligible impacts on differential gene expression observed between preservation methods. The interaction term between GC proportion and SSR presence/absence was also significant which we interpret to mean that SSR presence with high GC content is associated with higher expression in RNAlater. Despite the SSR term not being significant in the analysis of variance (Table 2), removing the term significantly impacted model fit.

3.4 | Annotation of differentially expressed genes

We expected little GO term enrichment as differences in gene expression would likely be due to differences in preservation techniques, not biological variation. The PANTHER suite annotation for genes that were significantly lower expressed in RNAlater compared to liquid nitrogen exhibited very few enriched functional categories (Supporting Information). However, many categories were significantly enriched for genes that were more highly expressed in RNAlater than liquid nitrogen. The most enriched categories in Reactome pathways are involved in gene expression and processing of mRNA. Likewise, enriched PANTHER protein classes include RNA binding proteins, mRNA processing and splicing factors, and transcription

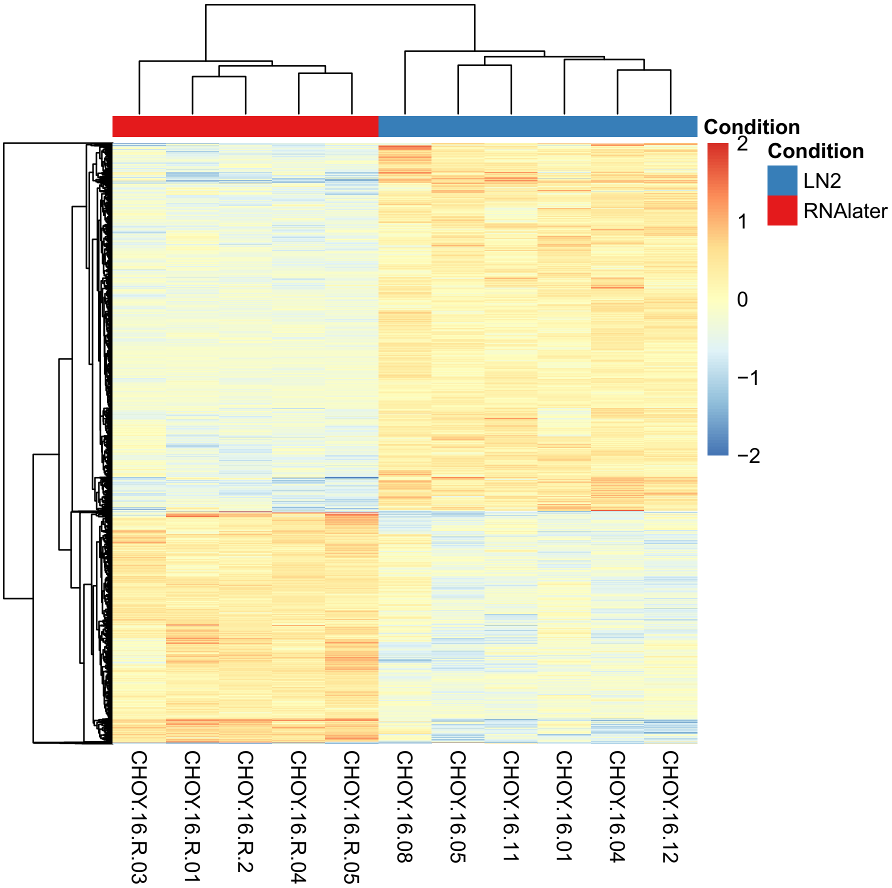

FIGURE 2 Clustering heat map showing genes that are differentially expressed among RNAlater samples and liquid nitrogen samples. Gene expression values have been normalized by sample and then centred about 0 for each gene. This heat map contains differentially expressed genes (after FDR correct with $p < 0.05$) including 1,073 genes that with higher expression values in the RNAlater treatment relative to the liquid nitrogen treatment, and 1,635 genes that exhibited lower expression values the RNAlater treatment

TABLE 2 Terms in the linear model that explain differences in expression between RNAlater storage and liquid nitrogen flash freezing and −80°C storage

Term	Sum Sq	Df	*F*-value	Estimate (*SE*)	*p*-value
Mean expression	1,088.8	1	496.2719	0.155547 (0.007308)	<2e-16
GC proportion	134.5	1	61.3069	−5.277778 (1.358944)	5.452e-15
Exon number	584.9	1	266.6218	0.026825 (0.001670)	<2.2e-16
SSR presence	0.2	1	0.0938	−1.620474 (0.703584)	0.75935
GC proportion:SSR presence	12.2	1	5.5619	3.269607 (1.386380)	0.01838

factors. Enriched GO terms included RNA binding and RNA processing. Similar results were obtained with the GOrilla analyses (Supporting Information Figures S5–S10). This consistent elevation of enrichment of functional categories for genes that are more abundant after an RNAlater treatment suggests that this treatment may be altering the physiology of the sample.

4 | DISCUSSION

Many sources contribute to variation in observed gene expression. Of these, most researchers are interested in assaying the variation that is due to a biological factor, such as genetic or physiological differences between samples. However, variation due to technical factors, such as noise in hybridization efficacy in microarray studies (Altman, 2005) or noise in the number of reads that map to a particular gene in RNAseq studies are large sources of variability in observed gene expression, and can substantially influence results (Bryant, Smyth, Robins-Browne, & Curtis, 2011; Marioni, Mason, Mane, Stephens, & Gilad, 2008). For RNA sequencing studies, the sources of technical variation are still being discovered, but can include many aspects of sample handling prior to actual measurement (McIntyre et al., 2011). Previous microarray studies have compared the sample handling procedures that were tested in our study and have found no difference downstream, particularly in differential gene expression patterns (Dekairelle, Vorst, Tombal, & Gala, 2007; Mutter et al., 2004). These studies, however, may not apply to the variance profile of RNA sequencing studies (Romero, Ruvinsky, & Gilad, 2012).

Our results suggest that sample handling is an important factor in variation of observed gene expression. While the total percentages of reads mapped were generally similar between the two treatments, the treatments we tested had a significant impact on RNA quality. Our results suggest that preservation in RNAlater for extended periods of time, as opposed to flash freezing, nonrandomly impacts gene expression values of over 20% of the transcriptome. Notably, other studies have found substantial RNA degradation for samples stored in RNAlater over extended periods, even when samples were stored at 4°C (Jones & Kennedy, 2015) or −80°C (Riesgo, Pérez-Porro, Carmona, Leys, & Giribet, 2012b). In our study, samples that were stored in RNAlater exhibited lower average RIN scores than samples that were flash-frozen in liquid nitrogen (Supporting Information Table S1), so our findings may be related to RNA degradation. Despite this, our RINs would be considered as acceptable for downstream applications, such as RNA sequencing library preparation (Imbeaud et al., 2005).

Our results suggest genes with higher GC content, fewer exons and lower expression are better preserved in liquid nitrogen. Conversely, our results suggest that genes with higher GC content, fewer exons or lower mean expression may not be as well preserved with RNAlater (De Wit et al., 2012). The functional enrichment for genes exhibiting significantly higher observed expression in RNAlater than liquid nitrogen indicates that RNAlater may be substantially altering the physiology of the samples during fixation or that RNAlater preserves certain functional categories of genes better than liquid nitrogen. The latter seems unlikely as it is difficult to hypothesize a mechanism, and upregulation of genes associated with RNA metabolism and translation has been observed in other studies comparing RNAlater to liquid nitrogen preservation (Bray et al., 2010). Further, the converse does not appear to have extensive enrichment for certain functional categories (i.e., genes that experience presumably better preservation in liquid nitrogen than RNAlater often do not fall in particular functional categories).

Based on our results, we recommend that researchers use caution when comparing gene expression values derived from RNAseq data sets that may have variable storage conditions. This is especially important with the growth of genomics technologies and accessibility of public data in repositories such as the NCBI Sequence Read Archive. Many entries in these databases do not routinely report metadata such as storage conditions, posing a serious challenge for data utilization. Further, future work could expand on examination of storage in TRIzol (Fisher Scientific, Hampton, NH) as recent work indicates expression patterns might be substantially different from liquid nitrogen (Kono et al., 2016). Likewise, various taxonomic groups may be more susceptible to variation in storage conditions due to differences in tissue permeability or presence of secondary compounds (Riesgo, Perez-Porro, Carmona, Leys, & Giribet, 2012a).

Several caveats are important in interpreting our study. While technical variation from storage condition is the dominant contributor to variation in our study, we acknowledge that biological variation also contributes to our observations. The samples in each storage condition are separate, whole individuals from the same clutch of fish. Fry at 30 dpf are too small to divide tissues equally into preservation treatments and obtain sufficient RNA quantity for RNAseq. Yet, even if a larger tissue sample was cut and divided, one

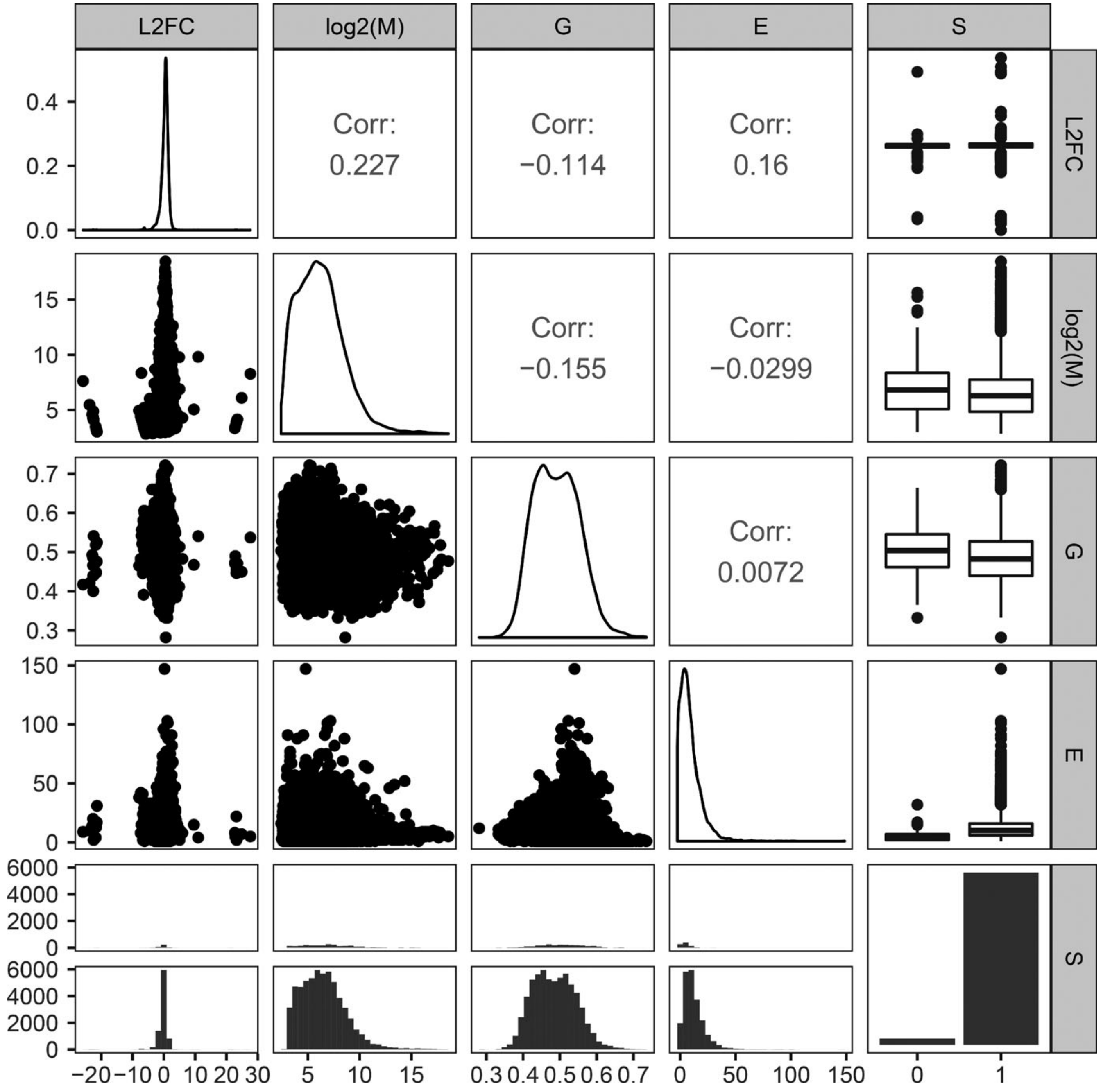

FIGURE 3 Relationships among the dependent variables retained in the best-fitting generalized linear model to explain the log2(RNAlater/liquid nitrogen) for each gene. L2FC: Log2(RNAlater/liquid nitrogen); log2(m): log2(mean expression across all samples); G: GC content; E: exon number; S: SSR presence (1) or absence (0). The panels along the diagonal show distributions of the individual explanatory variables with continuous variables displayed as density curves and categorical variables displayed as bar plots. Joint distributions or correlation coefficients are shown in the off-diagonal panels. Two continuous variables are shown as correlation coefficients and scatter plots. A continuous and categorical variables are shown as split box plots and split histograms

might expect biological variation due to different cell populations. Additionally, juvenile fish tissue may interact with the RNAlater buffer in different ways from other organisms. However, other studies have demonstrated similar effects between RNAlater and flash freezing. For instance, between preservation methods over 5,000 differentially regulated genes have been obtained from *Arabidopsis thaliana* tissue (c.f. Kruse, Basu, Luesse, & Wyatt, 2017). Though the *Arabidopsis* study did not assay systematic biases of particular gene attributes to preservation methods, many differentially regulated genes were related to osmotic stress, indicating a strong transcriptional response to RNAlater.

We also acknowledge that extraction batch was confounded with storage treatment. RNAlater samples were extracted in the same batch, while liquid nitrogen samples were extracted over several different batches (Supporting Information Table S1). The samples were part of a larger study, with 20 total RNA extraction batches of 169 liquid nitrogen samples and 1 extraction batch of the five RNAlater samples. Among the169 liquid nitrogen samples, lane of sequencing (which was randomized for RNAlater and liquid nitrogen samples in this study, Supporting Information Table S1) and RNA extraction batch accounts for very little variation (Supporting Information Figure S11, Table S2). Though we cannot discount that the RNA extraction of the RNAlater-stored samples was different in some way and our results could potentially be due to RNA extraction batch, we view this as unlikely because the identical research, equipment, and reagents were used over a short window of time (e.g., 24 of the 169 liquid nitrogen samples were extracted on the same day as the RNAlater samples).

Finally, long-term storage temperature is confounded with liquid nitrogen and RNAlater treatments in our study and long-term storage temperature is known to drive RNA integrity (Gayral et al., 2013; Kono et al., 2016). Our goal was to replicate typical field experiments, where reliable refrigeration is not available for substantial amounts of time, and RNAlater is used as the predominant preservation method. Despite these caveats, our work demonstrates that differing preservation methods and storage conditions nonrandomly impact gene expression, which may bias interpretation of results of RNA sequencing experiments. We look forward to future work that more thoroughly quantifies the impact on interpretation of biological signal derived solely from preservation methods (e.g., Bray et al., 2010).

ACKNOWLEDGEMENTS

We thank the University of Minnesota Genomics Center for their guidance and performing the cDNA library preparations and Illumina HiSeq 2500 sequencing. The authors acknowledge the Minnesota Supercomputing Institute (MSI) at the University of Minnesota for providing resources that contributed to the research results reported within this paper. URL: http://www.msi.umn.edu. Funding was supported by NIH (1R01GM127872-01 to SEM and ACK). CNP was supported by Grand Challenges in Biology Postdoctoral Program at University of Minnesota College of Biological Sciences. Institutional Animal Care and Use Committee at Florida Atlantic University (Protocol #A15-32).

AUTHOR CONTRIBUTIONS

All authors designed and conducted the experiment. C.N.P., T.J.Y.K. and S.E.M. performed the analyses and wrote the paper. B.A.S., J.B.J. and A.C.K. edited the manuscript

DATA ACCESSIBILITY

All reads are available in NCBI short read archive under accession numbers SRX3446133, SRX3446136, SRX3446135, SRX3446155, SRX3446156, SRS2736519, SRS2736520, SRS2736523, SRS2736524, SRS2736525 and SRS2736526. Scripts to perform all data handling and analysis tasks are available in a GitHub repository at https://github.com/TomJKono/CaveFish_RNAlater.

ORCID

Courtney N. Passow https://orcid.org/0000-0001-9873-8983
Suzanne E. McGaugh https://orcid.org/0000-0003-3163-3436

REFERENCES

Altman, N. (2005). Replication, variation and normalisation in microarray experiments. *Applied Bioinformatics*, *4*, 33–44. https://doi.org/10.2165/00822942-200504010-00004

Alvarez, M., Schrey, A. W., & Richards, C. L. (2015). Ten years of transcriptomics in wild populations: What have we learned about their ecology and evolution? *Molecular Ecology*, *24*, 710–725. https://doi.org/10.1111/mec.13055

Andrews, S. (2014). FastQC: A quality control tool for high throughput sequence data. Version 0.11. 2. *Babraham Institute*, *Cambridge*, *UK*. Retrieved from http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bray, S. E., Paulin, F. E., Fong, S. C., Baker, L., Carey, F. A., Levison, D. A., … Kernohan, N. M. (2010). Gene expression in colorectal neoplasia: Modifications induced by tissue ischaemic time and tissue handling protocol. *Histopathology*, *56*, 240–250. https://doi.org/10.1111/j.1365-2559.2009.03470.x

Bryant, P. A., Smyth, G. K., Robins-Browne, R., & Curtis, N. (2011). Technical variability is greater than biological variability in a microarray experiment but both are outweighed by changes induced by stimulation. *PLoS ONE*, *6*, e19556. https://doi.org/10.1371/journal.pone.0019556

Camacho-Sanchez, M., Burraco, P., Gomez-Mestre, I., & Leonard, J. A. (2013). Preservation of RNA and DNA from mammal samples under field conditions. *Molecular Ecology Resources*, *13*, 663–673. https://doi.org/10.1111/1755-0998.12108

Cheviron, Z. A., Carling, M. D., & Brumfield, R. T. (2011). Effects of postmortem interval and preservation method on RNA isolated from field-preserved avian tissues. *The Condor*, *113*, 483–489. https://doi.org/10.1525/cond.2011.100201

Choi, S., Ray, H. E., Lai, S.-H., Alwood, J. S., & Globus, R. K. (2016). Preservation of multiple mammalian tissues to maximize science return from ground based and spaceflight experiments. *PLoS ONE*, *11*, e0167391. https://doi.org/10.1371/journal.pone.0167391

Chowdary, D., Lathrop, J., Skelton, J., Curtin, K., Briggs, T., Zhang, Y., … Mazumder, A. (2006). Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative. *The Journal of Molecular Diagnostics*, *8*, 31–39. https://doi.org/10.2353/jmoldx.2006.050056

De Smet, L., Hatjina, F., Ioannidis, P., Hamamtzoglou, A., Schoonvaere, K., Francis, F., … de Graaf, D. C. (2017). Stress indicator gene expression profiles, colony dynamics and tissue development of honey bees exposed to sub-lethal doses of imidacloprid in laboratory and field experiments. *PLoS ONE*, *12*, e0171529. https://doi.org/10.1371/journal.pone.0171529

De Wit, P., Pespeni, M. H., Ladner, J. T., Barshis, D. J., Seneca, F., Jaris, H., … Palumbi, S. R. (2012). The simple fool's guide to population genomics via RNA-Seq: An introduction to high-throughput sequencing data analysis. *Molecular Ecology Resources*, *12*, 1058–1067. https://doi.org/10.1111/1755-0998.12003

Dekairelle, A.-F., Van der Vorst, S., Tombal, B., & Gala, J.-L. (2007). Preservation of RNA for functional analysis of separated alleles in yeast: Comparison of snap-frozen and RNALater® solid tissue storage methods. *Clinical Chemical Laboratory Medicine*, *45*, 1283–1287.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., & Huber, W. (2005). BioMart and Bioconductor: A powerful link between biological databases and microarray data analysis. *Bioinformatics*, *21*, 3439–3440. https://doi.org/10.1093/bioinformatics/bti525

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *NatureProtocols*, *4*, 1184. https://doi.org/10.1038/nprot.2009.97

Eden, E., Navon, R., Steinfeld, I., Lipson, D., & Yakhini, Z. (2009). GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, *10*, 48. https://doi.org/10.1186/1471-2105-10-48

Florell, S. R., Coffin, C. M., Holden, J. A., Zimmermann, J. W., Gerwels, J. W., Summers, B. K., … Leachman, S. A. (2001). Preservation of RNA for functional genomic studies: A multidisciplinary tumor bank protocol. *Modern Pathology*, *14*, 116. https://doi.org/10.1038/modpathol.3880267

Gayral, P., Melo-Ferreira, J., Glémin, S., Bierne, N., Carneiro, M., Nabholz, B., … Galtier, N. (2013). Reference-free population genomics from next-generation transcriptome data and the vertebrate–invertebrate gap. *PLoS Genetics*, *9*, e1003457. https://doi.org/10.1371/journal.pgen.1003457

Gorokhova, E. (2005). Effects of preservation and storage of microcrustaceans in RNAlater on RNA and DNA degradation. *Limnology and Oceanography: Methods*, *3*, 143–148.

Imbeaud, S., Graudens, E., Boulanger, V., Barlet, X., Zaborski, P., Eveno, E., … Auffray, C. (2005). Towards standardization of RNA quality assessment using user-independent classifiers of microcapillary electrophoresis traces. *Nucleic Acids Research*, *33*, e56. https://doi.org/10.1093/nar/gni054

Jones, S. P., & Kennedy, S. W. (2015). Feathers as a source of RNA for genomic studies in avian species. *Ecotoxicology*, *24*, 55–60. https://doi.org/10.1007/s10646-014-1354-z

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *NatureMethods*, *12*, 357. https://doi.org/10.1038/nmeth.3317

Kono, N., Nakamura, H., Ito, Y., Tomita, M., & Arakawa, K. (2016). Evaluation of the impact of RNA preservation methods of spiders for de novo transcriptome assembly. *Molecular Ecology Resources*, *16*, 662–672.

Kruse, C. P., Basu, P., Luesse, D. R., & Wyatt, S. E. (2017). Transcriptome and proteome responses in RNAlater preserved tissue of Arabidopsis thaliana. *PLoS ONE*, *12*, e0175943. https://doi.org/10.1371/journal.pone.0175943

López-Maury, L., Marguerat, S., & Bähler, J. (2008). Tuning gene expression to changing environments: From rapid responses to evolutionary adaptation. *Nature Reviews Genetics*, *9*, 583. https://doi.org/10.1038/nrg2398

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *GenomeBiology*, *15*, 550. https://doi.org/10.1186/s13059-014-0550-8

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., & Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *GenomeResearch*, *18*, 1509–1517. https://doi.org/10.1101/gr.079558.108

McGaugh, S. E., Gross, J. B., Aken, B., Blin, M., Borowsky, R., Chalopin, D., … Warren, W. C. (2014). The cavefish genome reveals candidate genes for eye loss. *Nature Communications*, *5*, 5307–5307. https://doi.org/10.1038/ncomms6307

McIntyre, L. M., Lopiano, K. K., Morse, A. M., Amin, V., Oberg, A. L., Young, L. J., & Nuzhdin, S. V. (2011). RNA-seq: Technical variability and sampling. *BMC Genomics*, *12*, 293. https://doi.org/10.1186/1471-2164-12-293

Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., & Thomas, P. D. (2016). PANTHER version 11: Expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research*, *45*, D183–D189.

Mutter, G. L., Zahrieh, D., Liu, C., Neuberg, D., Finkelstein, D., Baker, H. E., & Warrington, J. A. (2004). Comparison of frozen and RNALater solid tissue storage methods for use in RNA expression microarrays. *BMC Genomics*, *5*, 88.

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., & Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *NatureProtocols*, *11*, 1650. https://doi.org/10.1038/nprot.2016.095

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *NatureBiotechnology*, *33*, 290. https://doi.org/10.1038/nbt.3122

Riesgo, A., Perez-Porro, A. R., Carmona, S., Leys, S. P., & Giribet, G. (2012a). Optimization of preservation and storage time of sponge tissues to obtain quality mRNA for next-generation sequencing. *Molecular Ecology Resources*, *12*, 312–322.

Riesgo, A., Pérez-Porro, A. R., Carmona, S., Leys, S. P., & Giribet, G. (2012b). Optimization of preservation and storage time of sponge tissues to obtain quality mRNA for next-generation sequencing. *Molecular Ecology Resources*, *12*, 312–322.

Romero, I. G., Ruvinsky, I., & Gilad, Y. (2012). Comparative studies of gene expression and the evolution of gene regulation. *Nature Reviews Genetics*, *13*, 505. https://doi.org/10.1038/nrg3229

Team RC (2014). R: A language and environment for statistical computing. In *R foundation for statistical computing*.

Todd, E. V., Black, M. A., & Gemmell, N. J. (2016). The power and promise of RNA-seq in ecology and evolution. *Molecular Ecology*, *25*, 1224–1241.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*, 57. https://doi.org/10.1038/nrg2484

Wille, M., Yin, H., Lundkvist, Å., Xu, J., Muradrasoli, S., & Järhult, J. D. (2018). RNAlater® is a viable storage option for avian influenza sampling in logistically challenging conditions. *Journal of Virological Methods*, *252*, 32–36. https://doi.org/10.1016/j.jviromet.2017.11.004

Wolf, J. B. (2013). Principles of transcriptome analysis and gene expression quantification: An RNA-seq tutorial. *Molecular Ecology Resources*, *13*, 559–572. https://doi.org/10.1111/1755-0998.12109

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Passow CN, Kono TJY, Stahl BA, Jaggard JB, Keene AC, McGaugh SE. Nonrandom RNAseq gene expression associated with RNAlater and flash freezing storage methods. *Mol Ecol Resour*. 2019;19:456–464. https://doi.org/10.1111/1755-0998.12965