bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Multiple freeze-thaw cycles lead to a loss of consistency in poly(A)-enriched RNA sequencing

Benjamin P. Kellman[1,2,*], Hratch M. Baghdassarian[1,2,*], Tiziano Pramparo[3], Isaac Shamie[1,2], Vahid Gazestani[1,3], Arjana Begzati[4], Shengzhong Li[1,6], Srinivasa Nalabolu[3], Sarah Murray[5], Linda Lopez[3], Karen Pierce[3], Eric Courchesne[3], Nathan E. Lewis[1,6,7]

[1] Department of Pediatrics, University of California, San Diego

[2] Bioinformatics and Systems Biology Program, University of California San Diego

[3] Autism Center of Excellence, Department of Neuroscience, University of California San Diego

[4] Department of Medicine, University of California San Diego

[5] Department of Pathology, University of California San Diego

[6] Department of Bioengineering, University of California San Diego

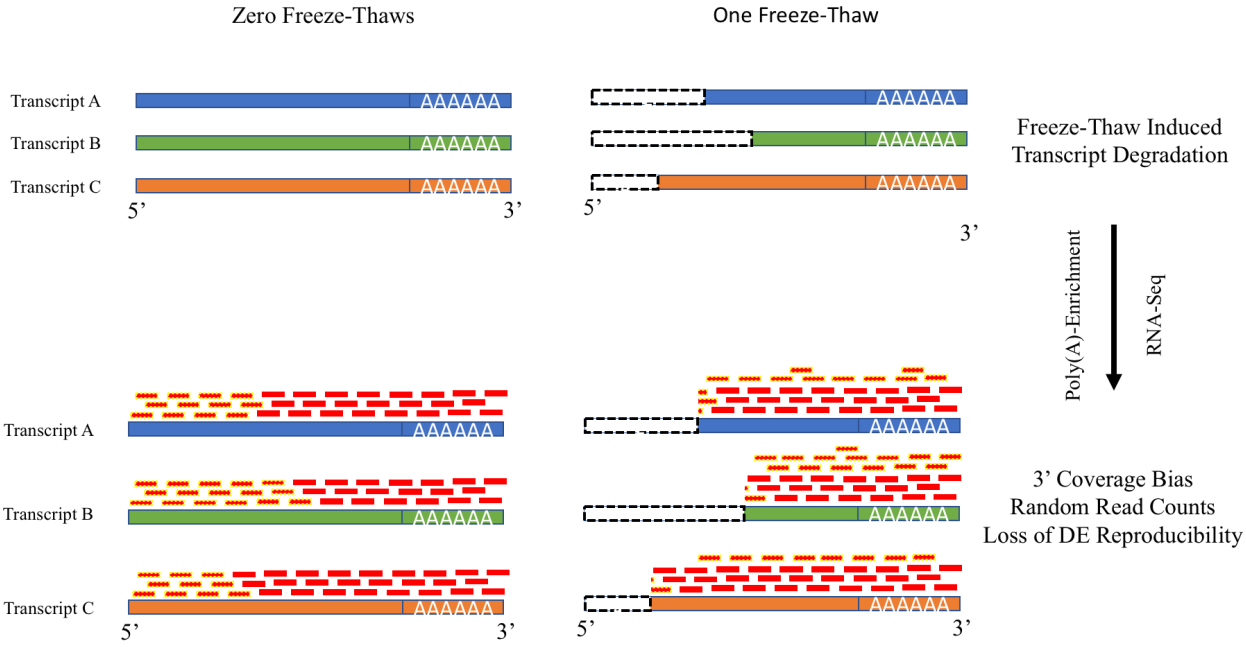
[7] Novo Nordisk Foundation Center for Biosustainability, University of California San Diego

[*] These authors contributed equally

Correspondence to: Nathan E. Lewis, nlewisres@ucsd.edu

Keywords: RNA-Seq; quality control; freeze-thaw; sample preparation; differential expression

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



Abstract

RNA-Seq is ubiquitous, but depending on the study, sub-optimal sample handling may be required, resulting in repeated freeze-thaw cycles. However, little is known about how each cycle impacts downstream analyses, due to a lack of study and known limitations in common RNA quality metrics, e.g., RIN, at quantifying RNA degradation following repeated freeze-thaws. Here we quantify the impact of repeated freeze-thaw on the reliability of downstream RNA-Seq analysis. To do so, we developed a method to estimate the relative noise between technical replicates independently of RIN. Using this approach we inferred the effect of both RIN and the number of freeze-thaw cycles on sample noise. We find that RIN is unable to fully account for the change in sample noise due to freeze-thaw cycles. Additionally, freeze-thaw is detrimental to sample quality and differential expression (DE) reproducibility, approaching zero after three cycles for poly(A)-enriched samples, wherein the inherent 3' bias in read coverage is more

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

exacerbated by freeze-thaw cycles, while ribosome-depleted samples are less affected by freeze-thaws. The use of poly(A)-enrichment for RNA sequencing is pervasive in library preparation of frozen tissue, and thus, it is important during experimental design and data analysis to consider the impact of repeated freeze-thaw cycles on reproducibility.

Introduction

RNA sequencing (RNA-Seq) is a ubiquitous technology, used to answer a wide range of biological questions. Methods for aligning, quantifying, normalizing and analyzing expression data are available through popular packages such as Tophat, STAR, cufflinks, SVA, RUV, Combat, DESeq2, edgeR, Kallisto, Salmon, BWA-MEM, and many others[1–11]. Each method aims to accommodate and mitigate the unique challenges presented by RNA-Seq data. Some approaches attempt to account for characterized variability in RNA-Seq measurements due to factors such as sequencing depth, gene length, and transcripts' physical characteristics (e.g., GC content). Others account for "unwanted variance" due to technical, batch, or experimental variation. Yet, the influence of sample processing, such as tissue lysis and processing time[12–14], is not sufficiently characterized such that it can be explicitly controlled. It is important to adequately characterize noise introduced to RNA-Seq measurements by sample processing steps to optimize sample quality, account for transcript degradation, and improve the accuracy and reproducibility of sequencing.

Transcript degradation continues after sample acquisition and affects data quality. Sample storage conditions (e.g. temperature and the use of stabilizing reagents) affect sample quality via RNA degradation[14,15]. Yet, varying sources of degradation can impact RNA-Seq in different manners[16]. Degradation introduces variability in signal and can be impacted by sample handling.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Non-uniformity in degradation across genes and samples causes inaccurate normalization and transcript quantification[17]. Poly(A)-enrichment methods are commonly used to separate mRNA from other highly abundant RNA molecules (e.g., rRNA, tRNA, snoRNAs, etc.), but variable degradation directly impacts read counts by causing non-uniform transcript coverage[18]. Of particular interest, freeze-thaw can induce 20% degradation of spike-in standards per cycle, a factor that may be generalizable to mRNA transcripts[19]. Freeze-thaw cycles increase RNA degradation by disrupting lysosomes which store RNases, freeing the enzymes to promiscuously catalyze nuclease activity[20]. Furthermore, partially defrosted crystals create uneven cleaving pressure on mRNA strands[21,22]. Despite these observations, the extent to which freeze-thaw negatively impacts count and differential expression in RNA-Seq analyses has not been comprehensively characterized.

Standard sample quality control often relies on RNA integrity number (RIN), which quantifies the 28S to 18S rRNA ratio[23]. RIN-based quality control approaches rely on a heuristic threshold to assess sufficient quality[24,25]. RIN-based metrics have known confounders such as transcript level, and thus have been called into question as an appropriate quality metric[26]. For example, RIN failed to indicate a decrease in sample quality in lung cancer tissue samples that underwent five freeze-thaw cycles[27] and, in statistical analyses, failed to correct for the effects of degradation[28]. Despite this, many studies rely on RIN to correct for and assess sample quality confounders[16,29,30]. This is especially problematic in the case of transcript degradation because RIN scores are based on entire samples, while degradation effects can be transcript-specific[17,3132]. Furthermore, existing studies on degradation are not simply generalizable to freeze-thaw, which has distinct and independent effects on sample quality and must be fully explored as such[16,33].

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Here, we tested the susceptibility of poly(A)-enriched RNA-Seq results after multiple freeze-thaw cycles. We assessed sample quality independently of RIN by simulating read count variability to capture the noise between technical replicates. We found that each additional freeze-thaw cycle increased the random counts between technical replicates by approximately 4%. Subsequently, differential expression reproducibility approached zero after three freeze-thaw cycles. These effects are not captured by RIN. We find that these effects are reflected in increasing 3' bias in read coverage when combining poly(A)-extraction with freeze-thaw, a phenomena that appears to be generalizable to publically available datasets.

Results

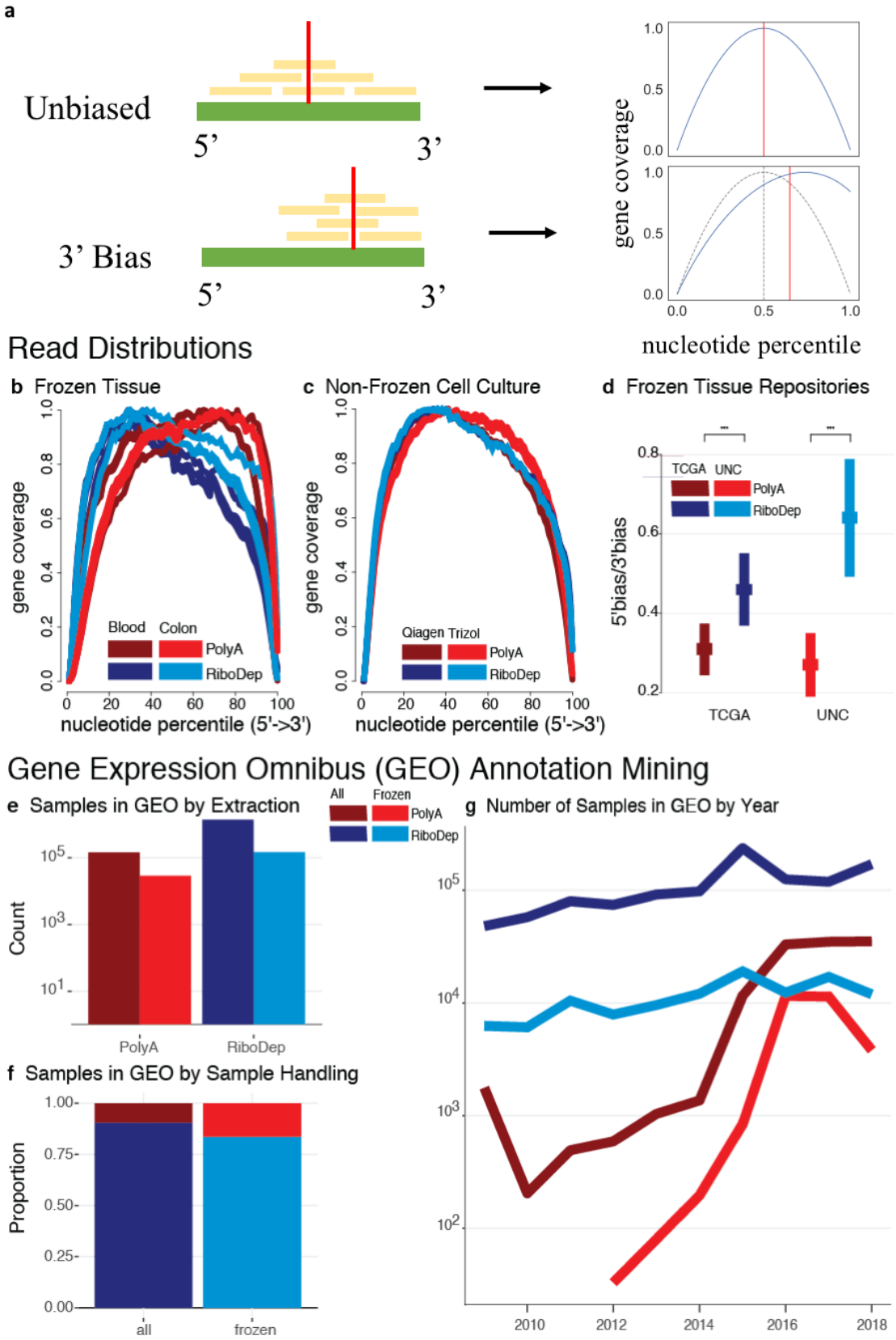
3' bias in read coverage of public datasets is associated to poly(A)-extraction and freeze-thaw

To examine the prevalence of bias from repeated freeze-thaw cycles, we search for the presence of RNA-seq distortion in publically available datasets. Specifically, we analyze the gene-coverage distribution in samples prepared with either poly(A)-extraction or ribosomal depletion. Since freeze-thaw enhances transcript degradation and poly(A)-enriched samples select mRNA by hybridization to the poly(A)-tail, we expect increased read coverage on the 3' end of transcripts--3' bias--when these two factors are combined. To test this expectation, we compared gene body coverage from the 5' to 3' end between poly(A)-enrichment and ribosomal depletion prepared samples with and without freezing. Specifically, we examined the median coverage percentile, the percentile-normalized nucleotide at which median cumulative coverage for a given sample is achieved (**Fig. 1a**).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

We analyzed the gene-coverage distribution in three studies: (1) RNA extraction in previously frozen solid and liquid tissues, (2) RNA extracted immediately after lysis of cultured cells without freezing, and (3) RNA extraction in previously frozen tissue from important public tissue resources.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figure 1: **3' Bias is Exacerbated in Frozen, Poly(A)-extracted Samples Across Multiple Studies:** *(a) Demonstration for determining median coverage percentile (red vertical line). When coverage is unbiased, reads (yellow) are distributed throughout the entire body of the transcript (green). In the absence of read bias and observing coverage as a function of the nucleotide percentile, we see that cumulative coverage along the transcript reaches 50% half-way through the gene body, at the 50th percentile nucleotide. In contrast, given a 3' read bias, there is a shift in the distribution of reads and cumulative coverage reaches 50% at, for example, the 60th percentile nucleotide. This results in a rightward shift in median coverage percentile towards the 3' end of the transcript. In the middle row, gene coverage (y-axis) at the ith nucleotide percentile from 5' to 3' (x-axis) is displayed for samples that were extracted using either poly(A)-enrichment or ribosomal depletion. Gene body coverage distributions were calculated for (b) for tissue samples that underwent an unspecified number of freeze-thaw cycles and (c) cell-culture samples samples that underwent no freeze-thaw cycles. (d) Comparison of 5' to 3' bias ratio (y-axis) of samples from the TCGA and UNC tissue repositories (x-axis) between extraction methods (two-sample t-test). Quantifying human RNA samples listed in GEO from 2008-2018, and stratifying by those annotated as "frozen", we observe (d) the number of samples prepared with poly(A)-extraction or ribosomal depletion (x-axis) (gray), (f) the proportion of samples extracted using either method, and (g) the change in the number of samples over time.*

In the first study[34], a comparison of the performance of poly(A) and ribosomal depletion in liquid and solid frozen tissue, we see a significant (one-sided Wilcoxon test, p = 0.01591) shift in the median gene coverage percentile towards the 3' end in the poly(A)-extracted samples (**Fig. 1b)**. Using generalized linear regression, we found that the median coverage percentile of poly(A)-extracted samples was 3.88% higher (Wald p = 0.011) than comparable ribosomal depletion extracted samples. In the second study[35], cells were not frozen before extraction and there was a small and insignificant difference in the 3' bias associated with library preparation (one-sided Wilcoxon test, p = 0.13, **Fig. 1c**). While the first two studies extract RNA from tissue and cell-

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

126 culture respectively--tissue extraction is typically lower quality, they are internally controlled
127 and therefore comparable. Finally, the third study[36] examines the impact of RNA extraction in
128 frozen tissue from the UNC and TCGA tumor tissue repositories. We found a significant (two-
129 sample t-test, p<1e16) decrease in the 5'-to-3' coverage ratio in poly(A)-extracted samples
130 compared to ribosomal depletion (**Fig. 1d**), indicating an increase in 3' bias. 3' bias was
131 consistent across RNA extracted from tissue frozen at either repository.

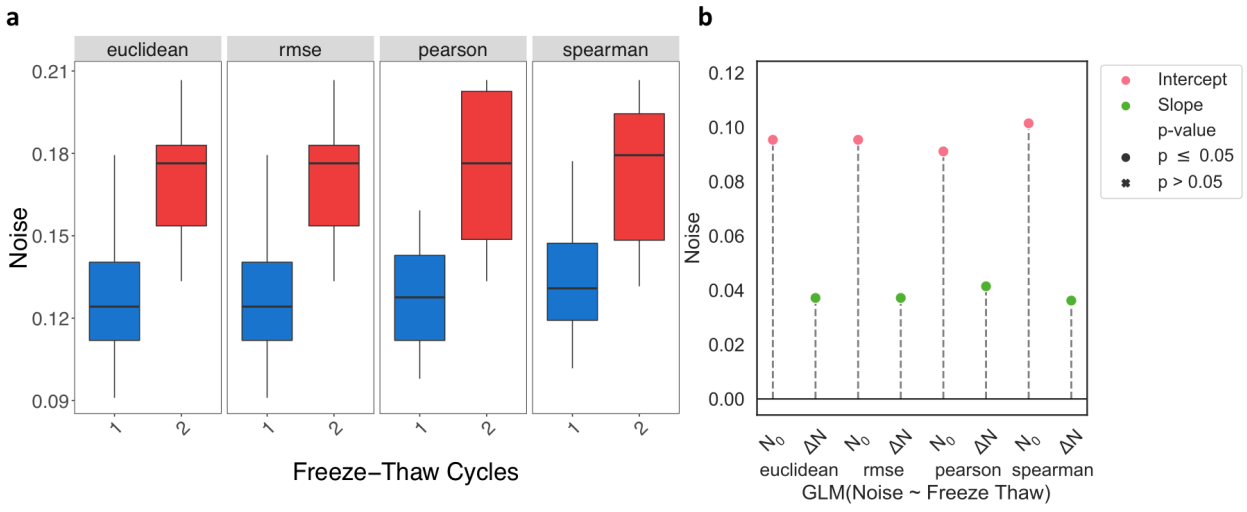
132 Next, we explored how widespread the impact of these observations may be by quantifying the
133 prevalence of poly(A)-extraction from frozen tissue by examining metadata in the Gene
134 Expression Omnibus (GEO). With GEOmetadb [37], we queried all human RNA samples between
135 2008 and 2018 using either poly(A)-extraction or ribosomal depletion. There are tens to hundred
136 of thousands of samples annotated as "frozen" for both total and poly(A)-extraction methods
137 (**Fig. 1e**). We note that this is an order of magnitude less than the total query, a value potentially
138 diminished by the complexity of the metadata. In samples annotated as "frozen", the frequency
139 of poly(A) extraction increases from less than 10% to over 25% **(Fig. 1f)** suggesting that the
140 problematic combination is prevalent and apparently preferred. Finally, stratifying this trend over
141 time, we see that poly(A)-extraction, as well as the relative proportion of poly(A)-extracted
142 frozen samples, is increasing in popularity relative to total RNA extraction, where usage has
143 remained fairly consistent (**Fig. 1g**).Taken together, these results indicate a potential, widespread
144 distortion in RNA-seq associated with a deleterious interaction between poly(A)-extraction and
145 freeze-thaw. To explore this potential more formally, the remainder of our analyses focus on a
146 specific experiment to address this question. Specifically we subjected whole-blood extracted
147 leukocyte samples--with technical replicates--from autistic or typically developing toddlers to a
148 varying number of freeze-thaw cycles, which we record along with other sample quality metrics

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

149 such as RIN.

150 An Additional Freeze-Thaw Cycle Increases Random Read Counts 1.4-Fold

151 To address the scarcity in quantification of loss in sample quality due to freeze-thaw, we
152 compare changes in sample quality between technical replicates. We first note that neither RIN
153 nor TIN capture significant (one-sided Wilcoxon test) decreases in sample quality due to
154 increased freeze-thaw (**Fig. S1**). Given previous indications that these metrics may not
155 sufficiently address transcript degradation, we instead measure the introduction of noise—the
156 randomness in read counts between technical replicates—to samples by freeze-thaw. We
157 simulated this randomness to reflect the dissimilarity between technical replicates. Since noise
158 does not rely on RIN, we could compare freeze-thaw and RIN-effects independently.



159

160 **Figure 2**: **Higher Noise in Samples with More Freeze-Thaw Cycles.** *From left to right, noise—*
161 *the randomness in read counts between technical replicates—is estimated using Euclidean distance,*
162 *RMSE, Pearson correlation,and Spearman correlation. (a) Box plots of noise for samples that underwent*
163 *either one or two freeze-thaws. (b) A generalized linear model was used to determine the expected noise*

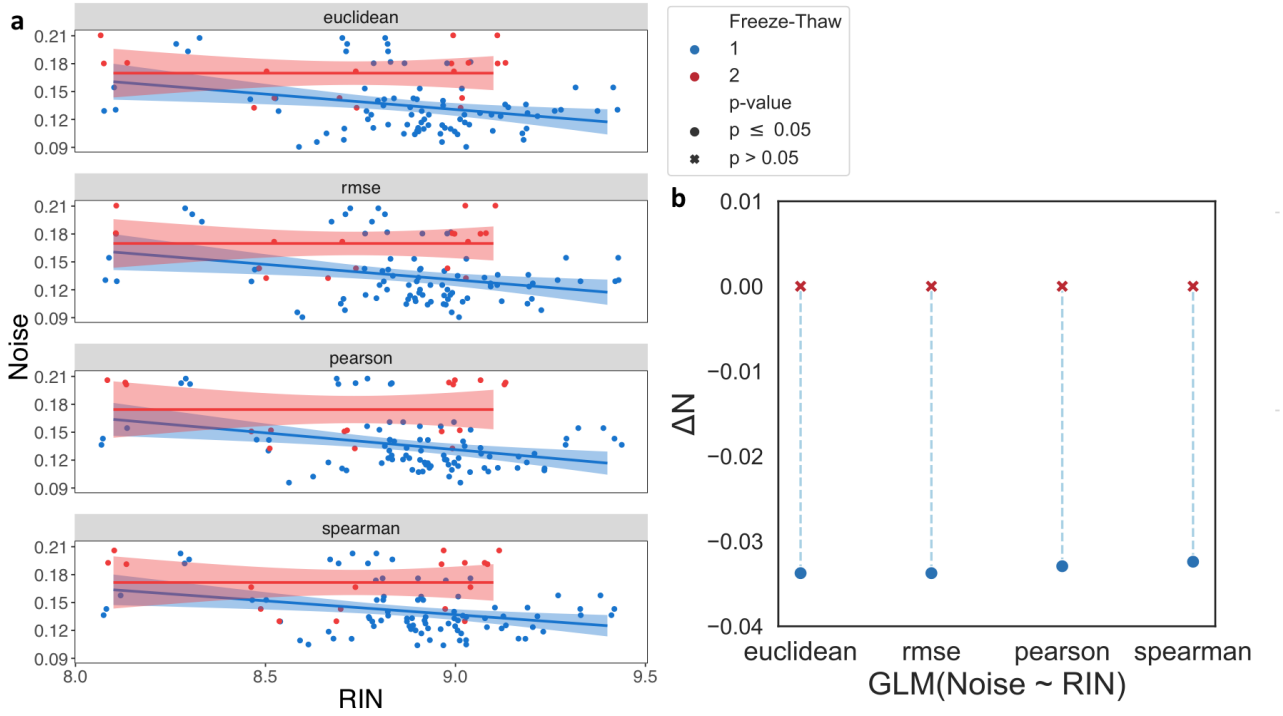
bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

at one freeze-thaw (N_0, pink) and the expected change in noise with each additional freeze-thaw (ΔN, green). All estimates are significant ($p \leq 0.05$).

Median noise increased 1.4-fold from one to two freeze-thaw cycles (one-sided Mann-Whitney U test, $p \leq 0.007$) on average across all measures (**Fig. 2a**). Noise between technical replicates when samples have only undergone one freeze-thaw was estimated to be 9.11-10.15% (Wald test, $p \leq 5.77e\text{-}7$). The expected increase in noise per additional freeze-thaw cycle was estimated to be 3.6-4.1 percentage points (Wald test, $p \leq 8.12e\text{-}3$) (**Fig. 2b**).

RIN Does Not Predict Additional Noise After One Freeze-Thaw Cycle

To follow up on our observations that RIN does not sufficiently capture changes in sample quality due to freeze-thaw, we asked whether RIN can reflect the differences in sample quality as measured by noise.



bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figure 3: **Discrepancy in the Relationship Between Noise and RIN due to additional Freeze-Thaw.** *Examining the relationship between noise, calculated by Euclidean distance, RMSE, pearson, and spearman correlation, for samples that underwent either one (blue) or two (red) freeze-thaw cycles. (a) Scatter plots comparing noise (y-axis) to RIN (x-axis). The solid lines show a linear regression fit and the shaded regions is the 95% confidence interval for this fit. (b) The expected change in noise due to a one point increase in RIN (ΔN, y-axis) estimated by a generalized linear model. Significant estimates ($p \leq 0.05$) are marked by a circle and insignificant estimates are marked by a cross.*

When only considering samples that underwent one freeze-thaw, each unit increase in RIN decreases noise by 3.24-3.38 percentage points for all metrics (Wald test, $p \leq 6.3e\text{-}3$) (Fig. **3a-b**). Yet, when only accounting for samples that underwent two freeze-thaw cycles, noise does not significantly change as RIN increases. Taken together, these results indicate that while RIN can be a good measure of noise for samples that underwent one freeze-thaw, it does not capture the loss in sample quality induced by two freeze-thaw cycles.

Differential Expression Similarity Increases 10.3% in High Quality Samples

Next, we investigated how the introduction of technical variation, noise, impacts downstream RNA-Seq analysis, specifically, differential expression (DE) analysis. As such, we assessed the reproducibility of DE results on combinations of samples with varying sample quality. We compared DE results between subsets of various sizes (4 - 14 samples). We measure reproducibility using similarity or discordance, based on correlation and dispersion, respectively. Higher similarity and lower discordance each represent higher reproducibility. We use these measures to assess differences that arise between subsets consisting of high quality (low freeze-thaw or high RIN) and low quality (high freeze-thaw or low RIN) samples.

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

We held two expectations regarding the effect of sample quality on DE reproducibility in the context of similarity: 1) the reproducibility between subsets with high quality samples should be higher than those with low quality samples, and 2) both subset size and sample quality should interact to increase the reproducibility of DE analysis; the increase in stability is reflected by a higher rate of increase in reproducibility with respect to subset size.

As expected, similarity increases with subset size. This is reflected in the upward shift in the similarity distribution with increasing subset size (**Fig. S10**) and the estimated 0.02 (Wald test, p = 2.2e-5) increase in similarity per additional sample (**Fig. 4a)**; thus, expected similarity would increase by 0.20 in a subset with 14 samples over a subset with 4 samples. Regression results for each model predicting similarity are reported in **Supplementary Table 4**.

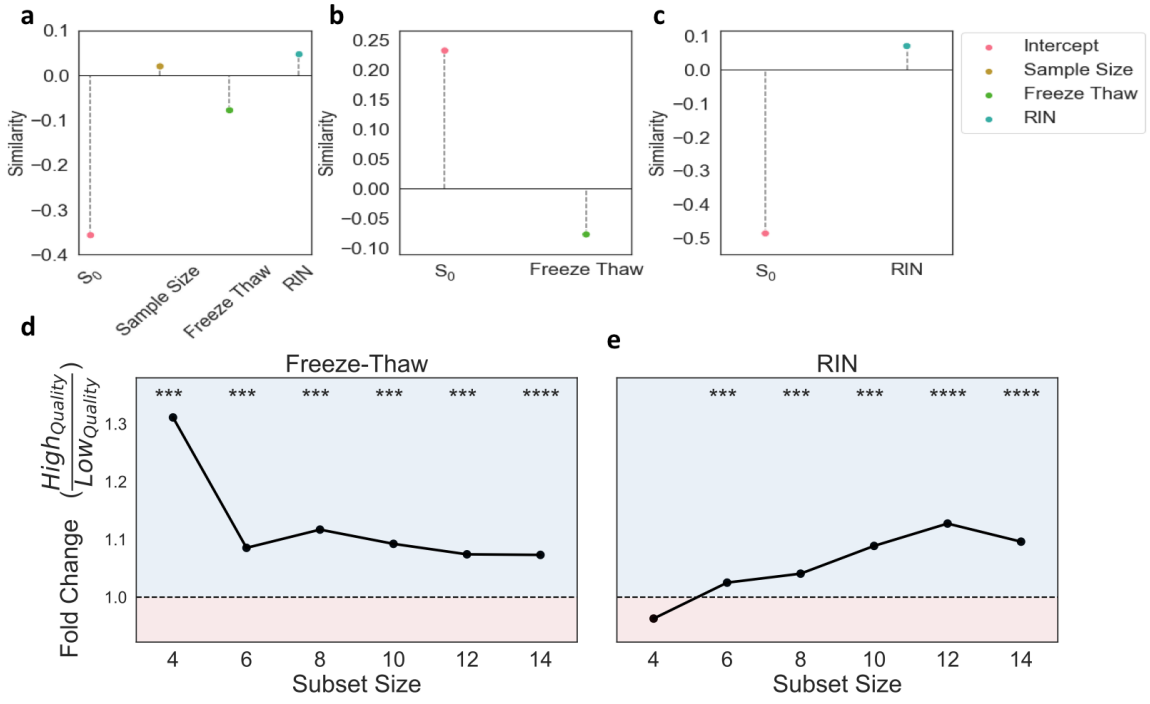


Figure 4: **Freeze-Thaw and RIN Both Demonstrate Higher Similarity with Increased**

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Quality. *Top panels summarize generalized linear models used to quantify the change in similarity per unit increase in (a) sample size, number of freeze-thaws and RIN combined, (b) only the number of freeze-thaws, and (c) only RIN. S_0 represents the intercept estimate and sample size, freeze-thaw, and RIN represent coefficient estimates. All estimates are significant. Bottom panels demonstrate fold-change in median similarity of high quality subsets with respect to low quality subsets at each subset size. The region shaded in blue (fold-change > 1) indicates instances where the median similarity for high quality is larger than that of low quality. The region shaded in red (fold-change < 1) indicates instances where the median similarity for low quality is larger than that of high quality. Average (d) freeze-thaw or (e) RIN are used to place subset pairs into high or low quality sample bins. Significance (one-sided Mann-Whitney U test) of comparisons in similarity distributions between high and low quality subset pairs are displayed above each subset size.*

We tested our first expectation by placing subset pairs into high and low sample quality bins, defined by either RIN or freeze-thaw, for each subset size and comparing their similarity values. Regardless of sample quality, DE similarity increases with subset size. Yet, for nearly all subset sizes, higher quality bins have significantly (one-sided Mann-Whitney U test, $p \leq 2.8e\text{-}17$) higher similarity than low quality bins (**Fig. 4e-f**). Across subset sizes, we observed an average 1.13-fold and 1.06-fold increase in similarity from low to high quality samples for freeze-thaw and RIN, respectively.

Similarity significantly (Wald test, $p \leq 9.2e\text{-}3$) decreases with the number of freeze-thaws and increases with RIN when accounting for the effects of sample size (**Fig. 4a-c**), validating our second expectation. Similarity decreases by 0.077 per additional freeze-thaw cycle (Wald test, $p = 8.77e\text{-}4$). Given the estimated similarity of 0.23 for samples that have not undergone freeze-thaw, this implies that DE reproducibility will approach zero after approximately three freeze-thaw cycles (**Fig. 4b**). Even when accounting for subset size and the effects of RIN, the

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

estimated decrease in similarity from freeze-thaw is nearly the same--0.078 (Wald test, p = 8.77e-4); this further corroborated that RIN alone cannot capture the changes in sample quality due to freeze-thaw. Taken together, these results indicate that higher sample quality increases DE reproducibility as measured by similarity.

Discordance Decreases ~5-fold In High Quality Samples

We further investigated the relationship between DE reproducibility and sample quality using an effect size sensitive measure of discordance. Specifically, we explored how sample quality affects the relationship between discordance and the mean-variance standardized effect at each subset size. In this context, we expected 1) discordance at any given effect size to be lower in high-quality subsets and 2) the rate of increase in discordance to be lower in high quality subsets relative to low quality subsets.

Corresponding to the regression models used for this analysis, we label our expected discordance when effect size is zero as D_0 and the change in discordance per unit increase in effect size as ΔD. We observed a significant (Wald test, $p \leq 9.45e\text{-}141$) decreasing trend in both D_0 and ΔD with increasing subset size (**Fig**. **S11, Supplementary Table 5)**.

As expected, independent of sample quality, ΔD demonstrates an overall decreasing trend with respect to subset size for both RIN and freeze-thaw. Given freeze-thaw, at a subset size of 6, there is a 1.1-fold decrease in the value of ΔD from low quality subsets to high quality subsets. The disparity in ΔD between high and low sample quality ($\Delta m = \Delta D_{\text{Low Quality}} / \Delta D_{\text{High Quality}}$) increases nearly monotonically through to the subset size of 14, at which point there is a 3.2-fold decrease; the monotonicity is an indication of the stability of this relation between discordance and sample quality. This causes notable differences in discordance values, even at low effect

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

sizes (**Fig. 5a**).

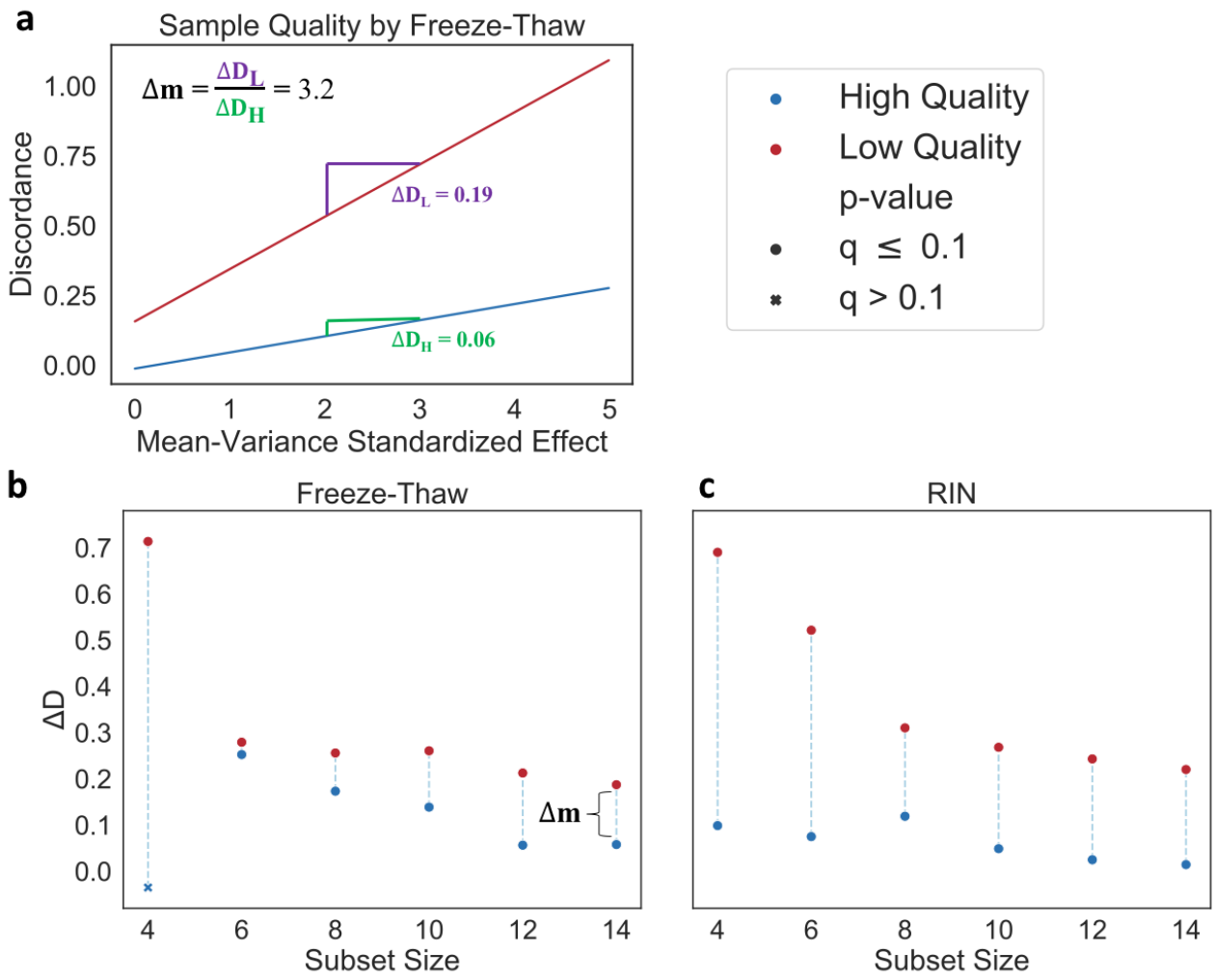


Figure 5: Higher Sample Quality has Lower Discordance at Each Subset Size. *GLM estimates of discordance predicted from effect size and subset size at high and low quality subsets at each subset size. High sample quality (blue) is compared to low sample quality (red). ΔD values represent the change in discordance per unit increase in effect size. (a) The predicted discordance with respect to the mean-variance standardized effect at a subset size of 14; sample quality is assessed by freeze-thaw. The disparity (Δm) between the change in discordance per unit increase in effect size for high (ΔD_H) and low (ΔD_H) quality subsets is also displayed. Summary of results for each subset size (x-axis) for sample quality represented by either (b) freeze-thaw or (c) RIN. Significant estimates (Wald test, Benjamini-*

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Hochberg FDR correction, $q \leq 0.1$) are marked by a circle and insignificant estimates are marked by a cross. For freeze-thaw, Δm corresponding to panel A is also displayed.

We estimate discordance with respect to effect size at each subset size and for subsets of either high and low quality (**Fig. 5a**). Consistent with our expectations, D_0 and ΔD are consistently lower for high quality subsets as compared to low quality subsets for both freeze-thaw and RIN across all subset sizes (**Fig. 5b-c**). Nearly all estimates are significant after multiple test correction (Wald test, Benjamini-Hochberg FDR correction, $q \leq 0.07$), with the exception of those for the smallest subset size for freeze-thaw.

Taken together, these results indicate that higher sample quality increases DE reproducibility as measured by discordance.

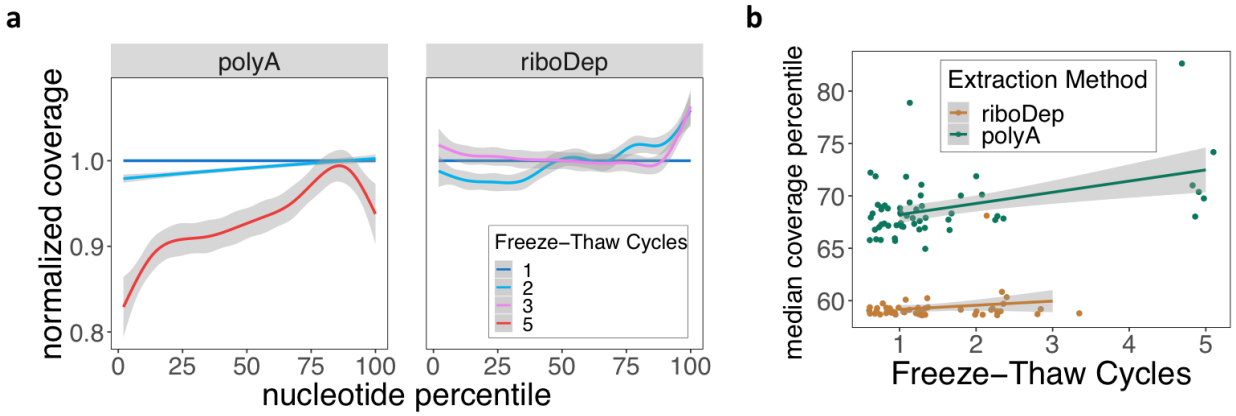
Additional freeze-thaw cycles show increased 3' bias in poly(A)-enriched but not ribosomal depletion samples

Finally, we asked whether repeated freeze-thaw cycles can induce a 3' bias, consistent with the induction of random reads and the loss of DE reproducibility as well as our initial observation in the public datasets.

Using the median coverage percentile, we found a shift in mRNA coverage towards the 3' end in the poly(A)-enriched samples relative to ribosomal depletion (**Fig. S13a**). Specifically, the median coverage percentile for poly(A)-enriched samples is significantly (one-sided Wilcoxon test, $p < 2.2e\text{-}16$) larger than that of ribosome depletion (**Fig. S13b**). Samples prepared with poly(A)-extraction have more 3' coverage bias compared to ribosomal depletion in both one (one-sided Wilcoxon test, $p = 3.5e\text{-}15$) and two (one-sided Wilcoxon test, $p = 1.4e\text{-}4$) freeze-thaw cycles **(Fig. S13d)**. Altogether, this indicates an overall 3' bias of poly(A)-enriched

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

290 samples, even independently of freeze-thaw **(Fig. S13b**).



291

292 **Figure 6**: **Freeze-Thaw Cycles Exacerbate 3' Bias in poly(A)-enriched samples.** *(a) Gene coverage (y-*
293 *axis) at the i^{th} nucleotide percentile (x-axis) for samples that underwent 1-5 freeze-thaws and were*
294 *extracted using either poly(A)-enrichment or ribosome depletion. Coverage is normalized to samples that*
295 *underwent one freeze-thaw. For each sample, coverage is averaged across all genes; samples are*
296 *aggregated using generalized additive model smoothing, with shaded regions representing 95%*
297 *confidence intervals. (b) Linear model fits comparing the change in median coverage percentile to the*
298 *number of freeze-thaw cycles for ribosome depletion (orange) or poly(A)-extraction (green).*

299 Crucially, this 3' bias is accentuated when samples are stratified by the number of freeze-thaw
300 cycles **(Fig. 6a).** We observe a significant increase (Wald test, p = 8.3e-4) in median coverage
301 percentile due to the number of freeze-thaw cycles in poly(A)-enrichment. The increase was not
302 maintained when was isolated using ribosome depletion (Wald test, p = 0.155) **(Fig. 6b,**
303 **Supplementary Table 7)**. For poly(A)-enriched samples, median coverage percentile increases
304 3.3 percentage points per root freeze-thaw cycle; the square-root of freeze-thaw was used to
305 stabilize variance. We further demonstrate a dependency of 3' bias on freeze-thaw cycles by
306 showing that median coverage percentile significantly increases with freeze-thaw in poly(A)-
307 enriched samples (Kruskal-Wallis test, p = 0.01). This 3' bias is particularly apparent after five

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

freeze-thaw cycles (one-sided Wilcoxon test, p = 2.4e-3). In contrast, we observe that median coverage percentile does not significantly change across freeze-thaw cycle counts in ribosomal depletion (Kruskal-Wallis test, p = 0.084) **(Fig. S13c**).

Taken together, these analyses indicate that poly(A)-enrichment inherently introduces a 3' bias in coverage as compared to ribosome depletion, and that this bias is exclusively exacerbated in poly(A)-enriched samples due to freeze-thaw cycles. Thus, 3' bias may indicate the severity of freeze-thaw induced signal degradation in poly(A)-extracted samples. If this 3' bias is the root cause of freeze-thaw induced instability in absolute and differential RNA-seq quantification, such instabilities may be subverted by substituting poly(A)-selection for ribosomal depletion during library preparation; such instabilities may be an unnecessary and avoidable inconvenience.

Discussion

Despite the utility and ubiquity of RNA-Seq, many of the confounding elements associated with the technology are still being characterized. In this work, we demonstrated how one such confounder--sample freeze-thaw--impacts sample quality and downstream analyses. We highlighted biases in publically available datasets, and observed an increased 3' bias when both freeze-thaw and poly(A)-extraction are features of a sample.

To gain a comprehensive understanding of this effect, we first simulated technical replication to measure the noise between technical replicates with different treatments. This allowed us to examine the impact of freeze-thaw cycles and the ability of RIN to capture those impacts. Next, we examined the impact of freeze-thaw cycles on the robustness and reproducibility of differential expression analysis. We found freeze-thaw cycles were substantially detrimental to

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

the stability of gene expression analysis. By our estimates and at these subset sizes, reproducibility of a differential expression signature approaches zero after three freeze-thaw cycles (**Supplementary Table 4**). Finally, we demonstrated that poly(A)-enriched samples demonstrate substantial 3' bias in read coverage with increased freeze-thaw cycles. Our results have implications with regards to technical variation due to sample handling, the sensitivity of differential gene expression analysis for frozen tissues and samples, and the utility of RIN.

Technical variation in RNA-Seq is substantial and can be attributed to a variety of factors, including read coverage, mRNA sampling fraction, library preparation batch, GC content, and sample handling[38,39]. As such, accounting for technical variation has been a major research area of focus for the past decade[1,2,5,39,40]. Degradation in combination with poly(A)-enrichment is a known source of variation in RNA-Seq. Yet, before technical variation can be accounted for, it must be characterized. While studies have looked into the effect of degradation on RNA-Seq, each mode of degradation impacts sample quality differently, and direct connections between freeze-thaw and sample quality has mainly been assessed via RIN[16,41,42].

Our simulation of technical replicates helps delineate technical variation due to sample handling--specifically, freeze-thaw. Furthermore, the resulting noise provides an estimate for the number of random read counts associated with a gene. For example, given an average 25 million reads sequenced per sample, our approximate 4 percentage points difference in noise between one and two freeze-thaw cycles gives an expected stochasticity in 1 million of those reads; approximating the number of protein-coding genes in the human genome to be 20-25 thousand[43], we can expect a difference of ~40-50 random counts per gene to exist between technical replicates due to a freeze-thaw cycle (**Supplementary Methods, Fig. S15**). Thus, each freeze-thaw introduces a non-negligible level of noise to the quantification of gene expression and differential expression

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

of such genes.

To check for the possibility that there is a signature which can help correct for freeze-thaw distortion of RNA-Seq, we attempt to find a group of common DE genes across various DE methods. We find no such signature (**Supplementary Results**). This is expected, given that a major source of reduced sample quality due to freeze-thaw is mRNA degradation, which occurs randomly for each transcript and sample. A possible path forward is to correct for sample degradation. Several methods have been proposed for this. While some of these methods rely on RIN or similar metrics (e.g. mRIN, TIN, etc.)[16,44], others have implemented statistical frameworks which account for gene-specific biases. DegNorm, for example, accounts for the gene-specific relative randomness in degradation in its correction approach[17]. Quality surrogate variable analysis (qSVA) specifically improves differential expression by identifying transcript features associated with RNA degradation for its correction[28]. Furthermore, there are recent methods which only assay the 3' end of a transcript and therefore claim robustness in degraded samples[45]. While these approaches could help account for noise introduced by RNA degradation during repeated freeze-thaw cycles, they cannot necessarily remedy the associated loss of signal.

The effect of freeze-thaw and resultant degradation on RNA-Seq is particularly concerning when considering differential gene expression analysis. It has been observed that RNA degradation can induce the apparent differential expression in as many as 56% of genes[44]. To this end, we quantified this loss of DE reproducibility by measuring similarity and discordance in the context of sample quality. We found a decrease in reproducibility with both decreasing RIN and increasing freeze-thaw. Interestingly, for most reproducibility assessments, we observed a monotonic or near monotonic increase in disparity between low and high quality subsets with respect to subset size. Similarity demonstrated a larger average magnitude of disparity for freeze-

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

thaw, whereas discordance demonstrated a larger average magnitude of disparity for RIN.

Based on our analysis, the utility of RIN in assessing quality when samples undergo freeze-thaw is questionable. The non-uniformity in mRNA degradation[46–49] due to freeze-thaw sheds light on these challenges, since RIN cannot quantify quality at the individual gene level[23]. This is reflected in the fact that samples with RIN > 8 demonstrate degradation[32]. Furthermore, results assessing the effect of freeze-thaw cycles on RIN are inconclusive. While some studies claim RIN can be used to account for degradation effects in RNA-Seq[16], others suggest it does not sufficiently capture the effects of degradation on sample quality[26,28]. When directly observing the effect of freeze-thaw on RIN, studies have found a negligible effect[12] or can only detect an effect after numerous cycles[27,50].

As such, we re-examined the utility of RIN as a measure of sample quality in relation to our noise estimation of random reads per sample[23]. We found that while noise increases with both decreasing RIN and increasing freeze-thaw, RIN may be an insufficient indicator of quality for samples that have undergone two or more freeze-thaws. Given these results, RIN may not always be a good metric to quantify the difference between technical replicates that have undergone variable sample handling[17,26–28]. We validate noise by confirming that it does not change with input RNA concentration, excepting outliers **(Fig. S12**). Therefore, in the future, noise could be a useful supplement to RIN when technical replicates are present.

The fact that our predicted decrease in similarity due to freeze-thaw does not change when incorporating RIN into our model further indicates that RIN alone cannot capture the changes in sample quality due to freeze-thaw. Despite this, RIN is a good indicator of sample quality, if not specifically for freeze-thaw. This is reflected in the fact that RIN validates our expectations for DE reproducibility analysis and the comparable range of noise, similarity, and discordance

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

399 values between freeze-thaw and RIN assessments.

400 Finally, to confirm our expectation that freeze-thaw decreases sample quality[18–22] and to further
401 characterize the underlying mechanism, we validated the presence of a 3' bias in coverage. This
402 builds on our and others' observations that a lower percentage of poly(A)-enriched transcripts
403 are covered[42]. We compared coverage to ribosome depleted RNA-Seq data, which does not use
404 3' hybridization to retain transcripts. We find that poly(A)-enrichment does in fact introduce a
405 strong 3' bias in coverage as compared to ribosome depletion. This bias is further exacerbated
406 with additional freeze-thaw cycles in poly(A)-enriched but not ribosome depleted samples. This
407 implies that degradation due to freeze-thaw does not impact RNA-sequencing of ribosome
408 depleted samples to the extent that it does in poly(A)-enriched samples. In light of our
409 demonstrations that 3' bias is associated with a substantial increase in noise and a decrease in DE
410 reproducibility, these findings suggest that RNA-seq from samples that have both been poly(A)-
411 extracted and undergone freeze-thaw cycles likely has unknown, diminished stability. While not
412 all studies have technical replicates to estimate noise, the presence of exaggerated 3' bias when
413 poly(A)-extraction is combined with freeze-thaw can be a simple indicator of RNA-seq
414 distortion.

415 Conclusion

416 Altogether, these results indicate that transcriptomics quality control steps cannot rely on RIN
417 alone for samples that have undergone poly(A)-enrichment and multiple freeze-thaws.
418 Furthermore, accounting for the effect of freeze-thaw on poly(A)-enriched RNA sequencing is
419 crucial. poly(A)-enrichment is prevalent for RNA-sequencing, and, in parallel, samples that
420 undergo multiple freeze-thaws are common in many protocols, especially rare tissues, e.g.,

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

421 postmortem neural tissue. Yet, there is no clear recommendation to avoid poly(A)-enrichment
422 following multiple freeze-thaws. Together, these results indicate that ribosomal depletion could
423 be a better alternative when freeze-thaw is necessary.

424 Declarations

425 Ethics Approval and Consent to Participate

426 In this study, we performed transcriptomics analyses of blood samples drawn from male toddlers
427 with the age range of 1–4 years. Research procedures were approved by the Institutional Review
428 Board of the University of California, San Diego. Parents of toddlers underwent informed
429 consent procedures with a psychologist or study coordinator at the time of their child's
430 enrollment. Additional details for the recruitment protocol are executed as described in Gazestani
431 et al.[51]

432 Author contributions

433 B.P.K., N.E.L., T.P., S.M., and E.C. designed and planned the experiments. T.P., S.N., K.P.,
434 S.M. and L.L. collected the samples, managed diagnostics, conducted transcriptome assays and
435 managed the data. B.P.K, H.M.B. and V.G. planned and conducted analyses. I.S. performed
436 RNA-seq QC. A.B. performed functional enrichment analyses. S.L. wrote the RNA-Seq
437 processing pipeline. B.P.K., H.M.B, and N.E.L wrote the manuscript. N.E.L. supervised the
438 project.

439 Acknowledgements and Funding

440 This work was supported by NIMH R01-MH110558 (E.C., N.E.L., T.P., V.G., S.N., K.P.,L.L.),

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

441 NIDCD R01-DC016385 (E.C., T.P.), T32GM008806 (H.M.B.), R35 GM119850 (B.P.K., I.S.),
442 the Simons Foundation (E.C.), and generous funding from the Novo Nordisk Foundation through
443 Center for Biosustainability at the Technical University of Denmark (NNF10CC1016517 S.L.).

444 Availability of Data and Materials

445 The datasets supporting the conclusions of this article are available in the NCBI Sequence Read
446 Archive repository (PRJNA627540,
447 https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA627540&o=acc_s%3Aa). Associated
448 metadata is available in Supplementary Tables 1-2.

449 Competing Interests

450 There are no competing interests in this study.

451 Methods

452 Sample Collection and Storage

453 Blood samples drawn from male toddlers with the age range of 1-4 years were usually taken at
454 the end of the clinical evaluation sessions. To monitor health status, the temperature of each
455 toddler was monitored using an ear digital thermometer immediately preceding the blood draw.
456 The blood draw was scheduled for a different day when the temperature was higher than 37 °C.
457 Moreover, blood draw was not taken if a toddler had some illness (for example, cold or flu), as
458 observed by us or stated by parents. We collected 4–6 ml blood into EDTA-coated tubes from
459 each toddler. Blood leukocytes were captured using LeukoLOCK filters (Ambion). After rinsing
460 the LeukoLOCK filters with PBS, the filters were flushed with RNAlater (Invitrogen) to stabilize

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

RNA within the intact leukocytes. After RNA stabilization, the LeukoLOCK filters were immediately placed in a −20 °C freezer. Additional RNA standards were sourced from normal human peripheral leukocytes pooled from 39 Asian individuals, ages 18 to 47 (Takara/ClonTech: 636592). The RNA standards underwent 1-5 simulated freeze-thaw cycles; 24hrs frozen at −80 °C and 1hr defrosting on ice.

RNA Extraction, Sequencing and Quantification

For 47 samples (from 16 individuals), mRNA was extracted using polyA selection with the TruSeq Stranded mRNA library preparation kit (Illumina). Ribosomal depletion was used to prepare an additional 52 samples. Relevant metadata regarding polyA-enriched and ribosomal depleted samples can be found in **Supplementary Table 1-2**. Ribosome depletion prepared samples used the TruSeq Stranded Total RNA with RiboZero Gold library preparation kit (Illumina). RNA Integrity Numbers (RIN) were measured using a NanoDrop ND-1000 (ThermoFisher). Poly-A selected samples were sequenced using 50-base pair single end sequencing on a HiSeq4000 (Illumina) to a depth of 25M reads. The ribo-depletion prepared libraries were sequenced using 100-base pair paired end sequencing on a HiSeq4000 (Illumina) to a depth of 50M reads.

Fastq files for each sample underwent quality control using FastQC (v0.33). PolyA and adaptor-trimming were conducted using Trimmomatic [52]. Reads were aligned to the gencode annotated (v25) human reference genome (GRCh38) using STAR (v2.4.0) [7]. Alignments were processed to sorted SAM files using SAMtools (v1.7) [53]. Finally, HTSeq using default settings (v0.6.1) was used to quantify reads [53,54].

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Estimation of noise between technical replicates

To estimate the noise between technical replicates of the same individual blood samples, we simulate random loss and gain of reads **(Fig. S2)**. One technical replicate was chosen as the "reference" replicate, making the other technical replicate the "target" replicate. Specifically, we designated replicates that have undergone one freeze-thaw as the reference, and those that underwent two freeze-thaw cycles as the target. The dissimilarity between replicates is measured by one of four metrics (Euclidean distance, RMSE, Pearson correlation, and Spearman correlation). We iteratively add and remove random reads to the reference replicate until the dissimilarity between the simulated replicate and the reference replicate was equal to the dissimilarity between a target replicate and the reference replicate **(Fig. S3, Figure S2**). We define the noise between the reference and target replicate as the fraction of reads added or removed per total reads in the reference replicate to achieve the aforementioned level of dissimilarity. We represent this as a percentage, e.g. 5% noise between a reference and target replicate can be interpreted as 5% randomness between their reads. For additional details on noise simulation, see Supplementary Methods.

Measuring the Effect of Sample Quality on Noise

To measure the association between noise and sample quality metrics (number of freeze-thaw cycles, input RNA concentrations, and RNA integrity number), we used a generalized linear model (GLM). The significance of the model parameters is determined by the Wald test. All results are reported in **Supplementary Table 3**.

For each model, to mitigate the contribution of potential confounding variables, samples with input RNA concentrations in the top and bottom 5% ($|z| \geq 1.645$) were removed, decreasing the

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

total number of samples from 47 to 41. For noise prediction from concentration, samples with more than one freeze-thaw were also excluded, decreasing the total number of samples to 35. Noise prediction from RIN was run separately for samples that had undergone one freeze-thaw and samples that had undergone two freeze-thaws.

Differential Expression Analysis

We assess whether the observed sample qualities (freeze-thaw and RIN) have an impact on differential expression (DE) reproducibility using a bootstrapping approach. DE was run on random sample subsets of varying sizes **(Fig. S4**). Before subsetting, we filtered our expression matrix for genes with an average count $\leq$ 20 across all samples. This reduced the number of genes from 10,028 to 4,520. The total number of samples considered was 46 when disregarding samples that were industry standards, were not assigned to either an ASD or TD indication, or did not have a recorded sample quality (RIN or freeze-thaw) value.

We generated subsets containing N = 4-14 samples. For each subset size N, we generated 2,000 unique subsets. Each subset had an equal number of TD or ASD samples. Additionally, only one replicate from each blood sample could be included. These requirements limited our subset size to a maximum of 14 samples.

DE between ASD and TD subjects was conducted using DESeq2[1]. **Fig. S10** summarizes DE results for all subsets. To account for potential confounders, we used RUV[5]. Specifically, we use a set of "in-silico empirical" negative control genes, including all but the top 5,000 differentially expressed genes as described in section 2.4 of the documentation for RUVseq (http://bioconductor.org/packages/release/bioc/vignettes/RUVSeq/inst/doc/RUVSeq.pdf). We check that RUV produces consistent results with previous Autism leukocyte gene expression

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

526 signatures[51,55], see **Supplementary Results.**

527 Similarity to Assess Differential Expression Reproducibility

528 To assess DE reproducibility, we measure the similarity in log-fold-change (LFC) values
529 between DE runs. Similarity is calculated as the pairwise spearman correlation of LFC between
530 all subsets of the same size **(Fig. S6)**. Genes with a median base mean (the mean of counts of all
531 samples, normalizing for sequencing depth) or median LFC in the bottom 10th percentile across
532 all subsets were excluded, filtering for low magnitude effects **(Fig. S5)**.

533 Average RIN and freeze-thaw were measured for all subset pairs. Resulting distributions for all
534 collected values from similarity analyses are displayed in **Fig. S7**.

535 Next, subsets of each size were split into two quantile bins for each quality metric separately.
536 High sample quality bins (low freeze-thaw or high RIN) were compared to low sample quality
537 bins. High sample quality subsets were tested for higher similarity than low sample quality bins
538 using a one-sided Mann-Whitney U test.

539 Additionally, three generalized linear models (GLMs) were fit to quantify the contribution of
540 sample quality metrics to the change in similarity for DE results across subsets. We fit one model
541 to predict similarity from freeze-thaw and RIN, while also accounting for the improvement in
542 reproducibility due to increase in subset size (Similarity ~ Freeze-Thaw + RIN + Subset Size).
543 We also fit two models predicting similarity from freeze-thaw or RIN alone.

544 Discordance to Assess Differential Expression Reproducibility

545 We adapted a measure of concordance to measure discordance, or the lack of reproducibility,
546 between differential expression results[56]. Average RIN and freeze-thaw were calculated for each

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

subset **(Fig. S8-9)**. Subsets for each subset size were split into two quantile bins for either quality metric (RIN and freeze-thaw). Genes with a median base mean across all subsets in the bottom tenth percentile were excluded from the analysis **(Fig. S5)**.

We do not use the original concordance at the top (CAT) metric because we are not comparing our results to a gold standard dataset. Instead, we use gene-wise LFC standard deviation across subsets as a measure of discordance. Thus, the average LFC for each gene across DE runs is analogous to the gold standard, and the dispersion from this average indicates a lack of reproducibility. At each combination of subset size and quality bins, we calculate discordance and compare it to the gene-wise median effect size **(Fig. S8)**. We measure effect size as the mean-variance standardized effect[1]. This and two additional effect size metrics (Cohen's d and absolute median LFC) we use are further described in **Fig. S9**. Results for all three effect size metrics reflect similar trends and can be found in **Supplementary Tables 5-6.**

We used a GLM to predict discordance from effect size at each subset size. Additionally, in a separate GLM, we account for the interaction between effect size and sample quality (Discordance ~ Effect Size x Sample Quality) at each subset size. Here, sample quality is a dummy variable, assuming a value of 0 for low quality and 1 for high quality. We did not include a term for subset size because regressions were fit within each subset size.

Read Coverage Bias

The distribution of read coverage over each gene body was measured using *geneBody_coverage.py* from the *RSeQC* (v3.0.0) package [57]. We measure this coverage ranging from the 0th percentile (5' end) to the 100th percentile (3' end) nucleotide. The ith percentile nucleotide is calculated as $nucleotide_i \,/\, length_{gene}$. Coverage at the ith percentile nucleotide is

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

569 normalized across all genes within a sample.

570 For a given sample, the median coverage percentile is defined as the nucleotide percentile at
571 which median cumulative coverage is achieved; cumulative coverage is aggregated from the 5'
572 end to the 3' end. The larger the median coverage percentile value, the larger the 3' bias in
573 coverage. We include 9 industry standards to our analysis--six of which had undergone five
574 freeze-thaw cycles and three of which had undergone one freeze-thaw cycle--to explore the
575 impact at higher freeze-thaw counts. We also include ribosomal depletion extracted samples as a
576 negative control.

577 We extended this read coverage bias to three additional datasets, all of which contained both
578 poly(A)-extracted and ribosomal depletion extracted samples. The first dataset (PRJNA427184)
579 contains samples for liquid and solid frozen tissue [34]. The second (PRJEB4197) contains
580 HEK293 samples that explicitly never underwent freeze-thaw cycles [35]. The third
581 (phs000676.v1.p1) contains frozen tissue samples from the UNC and TCG tumor tissue
582 repositories [36]. For the first two datasets, fastq files were aligned using STAR and gene body
583 coverage was calculated from alignments using RSeQC as previously described. We did not
584 directly analyze the raw files from TCGA or UNC, but instead reanalyzed the reported 5' to 3'
585 bias ratios. This ratio is similar to the inverse of the median coverage percentile: the smaller it is,
586 the larger the 3' bias in read coverage.

587

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

References

1. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
2. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139 (2010).
3. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
4. Leek, J. T. & Storey, J. D. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genet.* **3**, e161 (2007).
5. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* **32**, 896–902 (2014).
6. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
7. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15 (2013).
8. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 1–19 (2016).
9. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
10. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
11. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

12. Jun, E. *et al.* Method Optimization for Extracting High-Quality RNA From the Human Pancreas Tissue. *Transl. Oncol.* **11**, 800–807 (2018).

13. Passow, C. N. *et al.* RNAlater and flash freezing storage methods nonrandomly influence observed gene expression in RNAseq experiments. *bioRxiv* 379834 (2018) doi:10.1101/379834.

14. Micke, P. *et al.* Biobanking of fresh frozen tissue: RNA is stable in nonfixed surgical specimens. *Lab. Invest.* **86**, 202–211 (2006).

15. Ohmomo, H. *et al.* Reduction of Systematic Bias in Transcriptome Data from Human Peripheral Blood Mononuclear Cells for Transportation and Biobanking. *PLoS One* **9**, (2014).

16. Romero, I. G., Pai, A. A., Tung, J. & Gilad, Y. RNA-seq: impact of RNA degradation on transcript quantification. *BMC Biol.* **12**, 42 (2014).

17. Xiong, B., Yang, Y., Fineis, F. R. & Wang, J. P. DegNorm: normalization of generalized transcript degradation improves accuracy in RNA-seq analysis. *Genome Biol.* **20**, 75–75 (2019).

18. Li, S. *et al.* Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nat. Biotechnol.* **32**, 915–925 (2014).

19. Svensson, V. *et al.* Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* **14**, 381–387 (2017).

20. Locksley E. McGann, Hongyou Yang, Michele Walterson. Manifestations of cell damage after freezing and thawing. *Cryobiology* **25**, 178–185 (1988).

21. Röder, B., Frühwirth, K., Vogl, C., Wagner, M. & Rossmanith, P. Impact of long-term storage on stability of standard DNA for nucleic acid-based methods. *J. Clin. Microbiol.* **48**,

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

4260–4262 (2010).

22. Shao, W., Khin, S. & Kopp, W. C. Characterization of effect of repeated freeze and thaw cycles on stability of genomic DNA using pulsed field gel electrophoresis. *Biopreserv. Biobank.* **10**, 4–11 (2012).

23. Schroeder, A. *et al.* The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol. Biol.* **7**, 3 (2006).

24. Reiman, M., Laan, M., Rull, K. & Sõber, S. Effects of RNA integrity on transcript quantification by total RNA sequencing of clinically collected human placental samples. *FASEB J.* **31**, 3298–3308 (2017).

25. Shen, Y. *et al.* Impact of RNA integrity and blood sample storage conditions on the gene expression analysis. *Onco. Targets. Ther.* **11**, 3573 (2018).

26. Sonntag, K.-C. *et al.* Limited predictability of postmortem human brain tissue quality by RNA integrity numbers. *J. Neurochem.* **138**, 53 (2016).

27. Yu, K. *et al.* Effect of multiple cycles of freeze–thawing on the RNA quality of lung cancer tissues. *Cell and Tissue Banking* vol. 18 433–440 (2017).

28. Jaffe, A. E. *et al.* qSVA framework for RNA quality correction in differential expression analysis. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 7130 (2017).

29. Bao, W.-G. *et al.* Biobanking of Fresh-frozen Human Colon Tissues: Impact of Tissue Ex-vivo Ischemia Times and Storage Periods on RNA Quality. *Ann. Surg. Oncol.* **20**, 1737–1744 (2012).

30. Zeugner, S., Mayr, T., Zietz, C., Aust, D. E. & Baretton, G. B. RNA Quality in Fresh-Frozen Gastrointestinal Tumor Specimens—Experiences from the Tumor and Healthy Tissue Bank TU Dresden. in *Pre-Analytics of Pathological Specimens in Oncology* 85–93

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

(Springer, Cham, 2015).

31. Li, J., Jiang, H. & Wong, W. H. Modeling non-uniformity in short-read rates in RNA-Seq data. *Genome Biology* **11**, (2010).

32. Hoen, P. A. C. *et al.* Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat. Biotechnol.* **31**, 1015–1022 (2013).

33. Thompson, K. L., Scott Pine, P., Rosenzweig, B. A., Turpaz, Y. & Retief, J. Characterization of the effect of sample quality on high density oligonucleotide microarray data using progressively degraded rat liver RNA. *BMC Biotechnol.* **7**, 57 (2007).

34. Zhao, S., Zhang, Y., Gamini, R., Zhang, B. & von Schack, D. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Sci. Rep.* **8**, 1–12 (2018).

35. Sultan, M. *et al.* Influence of RNA extraction methods and library selection schemes on RNA-seq data. *BMC Genomics* **15**, 675 (2014).

36. Zhao, W. *et al.* Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genomics* **15**, (2014).

37. Zhu, Y., Davis, S., Stephens, R., Meltzer, P. S. & Chen, Y. GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus. *Bioinformatics* **24**, 2798–2800 (2008).

38. McIntyre, L. M. *et al.* RNA-seq: technical variability and sampling. *BMC Genomics* **12**, 1–13 (2011).

39. Hansen, K. D., Irizarry, R. A. & Wu, Z. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* **13**, 204–216 (2012).

40. Leek, J. T., Evan Johnson, W., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* vol. 28 882–883 (2012).

41. Copois, V. *et al.* Impact of RNA degradation on gene expression profiles: assessment of different methods to reliably determine RNA quality. *J. Biotechnol.* **127**, 549–559 (2007).
42. Adiconis, X. *et al.* Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nat. Methods* **10**, 623–629 (2013).
43. Pertea, M. *et al.* Thousands of large-scale RNA sequencing experiments yield a comprehensive new human gene list and reveal extensive transcriptional noise. *bioRxiv* 332825 (2018) doi:10.1101/332825.
44. Wang, L. *et al.* Measure transcript integrity using RNA-seq data. *BMC Bioinformatics* **17**, 58 (2016).
45. Foley, J. W. *et al.* Gene expression profiling of single cells from archival tissue with laser-capture microdissection and Smart-3SEQ. *Genome Research* vol. 29 1816–1825 (2019).
46. Wang, Y. *et al.* Precision and functional specificity in mRNA decay. *Proceedings of the National Academy of Sciences* vol. 99 5860–5865 (2002).
47. Yang, E. *et al.* Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res.* **13**, 1863–1872 (2003).
48. Narsai, R. *et al.* Genome-wide analysis of mRNA decay rates and their determinants in Arabidopsis thaliana. *Plant Cell* **19**, 3418–3436 (2007).
49. Feng, H., Zhang, X. & Zhang, C. mRIN for direct assessment of genome-wide and gene-specific mRNA integrity from large-scale RNA-sequencing data. *Nat. Commun.* **6**, 7816 (2015).
50. Wang, Y. *et al.* The Impact of Different Preservation Conditions and Freezing-Thawing

bioRxiv preprint doi: https://doi.org/10.1101/2020.04.01.020792; this version posted April 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

703 Cycles on Quality of RNA, DNA, and Proteins in Cancer Tissue. *Biopreserv. Biobank.* **13**,
704 335–347 (2015).
705 51. Gazestani, V. H. *et al.* A perturbed gene network containing PI3K/AKT, RAS/ERK,
706 WNT/β-catenin pathways in leukocytes is linked to ASD genetics and symptom severity.
707 *bioRxiv* 435917 (2019) doi:10.1101/435917.
708 52. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
709 sequence data. *Bioinformatics* **30**, 2114 (2014).
710 53. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–
711 2079 (2009).
712 54. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-
713 throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
714 55. Pramparo, T. *et al.* Prediction of autism by translation and immune/inflammation
715 coexpressed genes in toddlers from pediatric community practices. *JAMA Psychiatry* **72**,
716 386–394 (2015).
717 56. Collado-Torres, L. *et al.* Reproducible RNA-seq analysis using recount2. *Nat. Biotechnol.*
718 **35**, 319–321 (2017).
719 57. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments.
720 *Bioinformatics* **28**, 2184–2185 (2012).

721