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

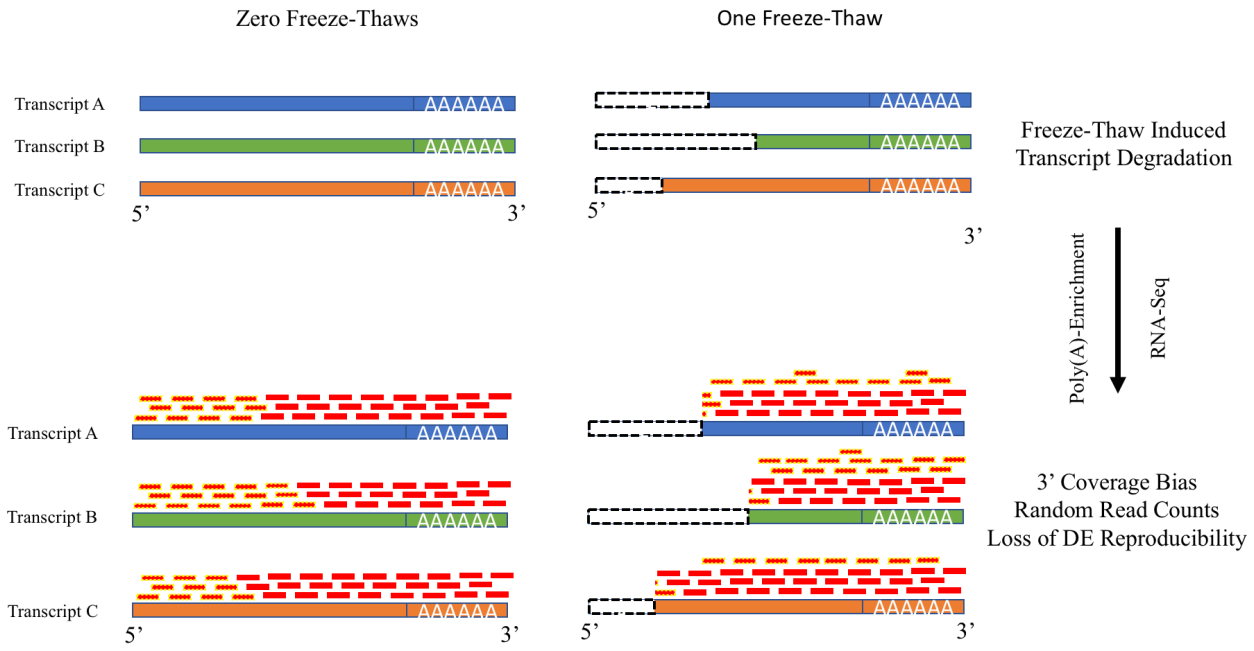
[6] Department of Bioengineering, University of California San Diego

[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



Abstract

RNA-Seq is ubiquitous, but depending on the biological question or the sample source, some

RNA-Seq collection methods require variable or sub-optimal sample handling. Rare tissue may need to be fixed or frozen prior to library preparation, or samples may remain at room temperature while other samples or reagents are handled. These challenges create uncharacterized impacts on sample quality. Understanding the relevant experimental factors that impact sample quality is crucial to ensuring reproducible RNA-Seq results. 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 quantify the noise between technical replicates. Each additional freeze-thaw cycle increased the random counts between technical replicates by approximately 4%. Additionally, after three freeze-thaw cycles, differential expression reproducibility decreased substantially. The use of poly(A)-enrichment for RNA sequencing is prevalent 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 cell death and affects sample quality. Sample storage conditions (e.g. temperature and the use of stabilizing reagents) affect sample quality via RNA degradation[14,15]. Degradation introduces variability in signal and can be impacted by sample handling. Non-uniformity in degradation across genes and samples causes inaccurate normalization and transcript quantification[16]. 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[17]. Of particular interest, freeze-thaw can induce 20% degradation of spike-in standards per cycle, a factor that may be generalizable to mRNA transcripts[18]. Freeze-thaw cycles increase RNA degradation by disrupting lysosomes which store RNases, freeing the enzymes to promiscuously catalyze nuclease activity[19]. Furthermore, partially defrosted crystals create uneven cleaving pressure on mRNA strands[20,21]. 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[22]. RIN-based quality control approaches rely on a heuristic threshold to assess sufficient quality[23,24]. RIN-based metrics have known confounders such as transcript level[25], 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[25,29,30]. 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[16,31].

Given the limited exploration of the relationship between freeze-thaw and RNA-seq data, especially in a manner that does not rely on RIN for assessment, a comprehensive quantification of the effect of freeze-thaw cycles is necessary. Here we highlight contexts in which freeze-thaw has the largest impact on downstream analyses and is a key first step to controlling for the

resultant technical variation between samples. To do so, we estimate the relative noise between technical replicates independently of RIN. We apply this approach to infer 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, and that freeze-thaw is detrimental to sample quality and differential expression (DE) reproducibility. Complementarily, we find that poly(A)-enrichment has an inherent 3' bias in gene coverage as compared to ribosome depletion, and that this bias is exacerbated by freeze-thaw cycles.

Results

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

Due to the scarcity in quantification of loss in sample quality due to freeze-thaw, we wanted to measure the introduction of noise—the randomness in read counts between technical replicates—to samples by freeze-thaw. As such, we simulated this randomness to reflect the dissimilarity between technical replicates. Since noise does not rely on RIN, we could compare freeze-thaw and RIN-effects independently.

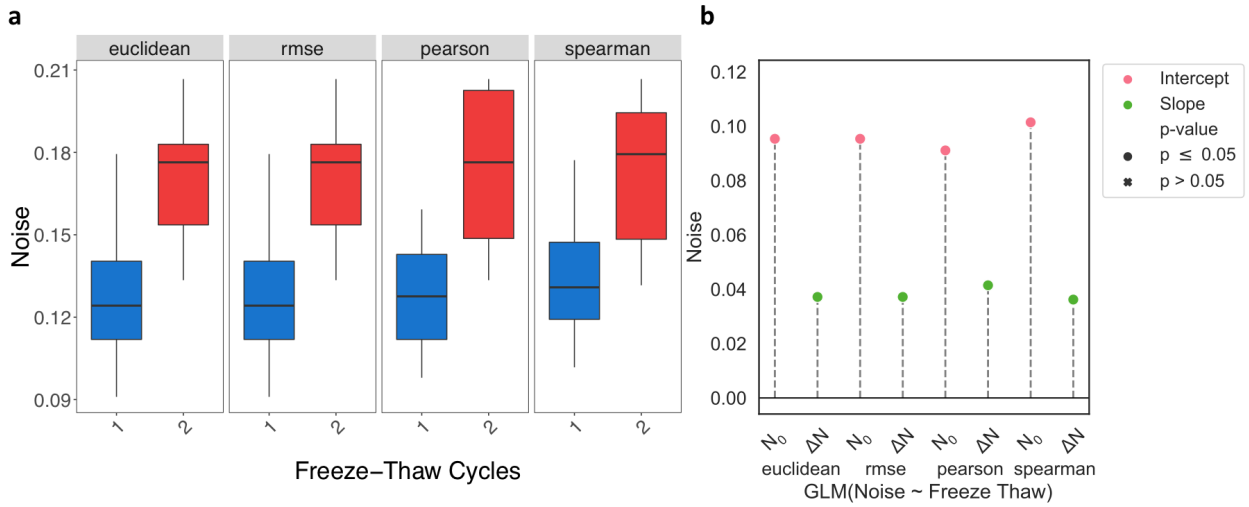


Figure 1: **Higher Noise in Samples with More Freeze-Thaw Cycles.** *From left to right, noise—the randomness in read counts between technical replicates—is estimated using Euclidean distance, RMSE, Pearson correlation,and Spearman correlation. (a) Box plots of noise for samples that underwent either one or two freeze-thaws. (b) A generalized linear model was used to determine the expected noise 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. 1a**). Noise between technical replicates when samples have only undergone one freeze-thaw was estimated to be 9.11-10.15% (Wald test, $p \leq 5.77e-7$). The expected increase in noise per additional freeze-thaw cycle was estimated to be 3.6-4.1 percentage points (Wald test, $p \leq 8.12e-3$) (**Fig. 1b**).

RIN Does Not Predict Noise After One Freeze-Thaw Cycles

Next, we asked whether RIN can also reflect the differences in sample quality as measured by noise.

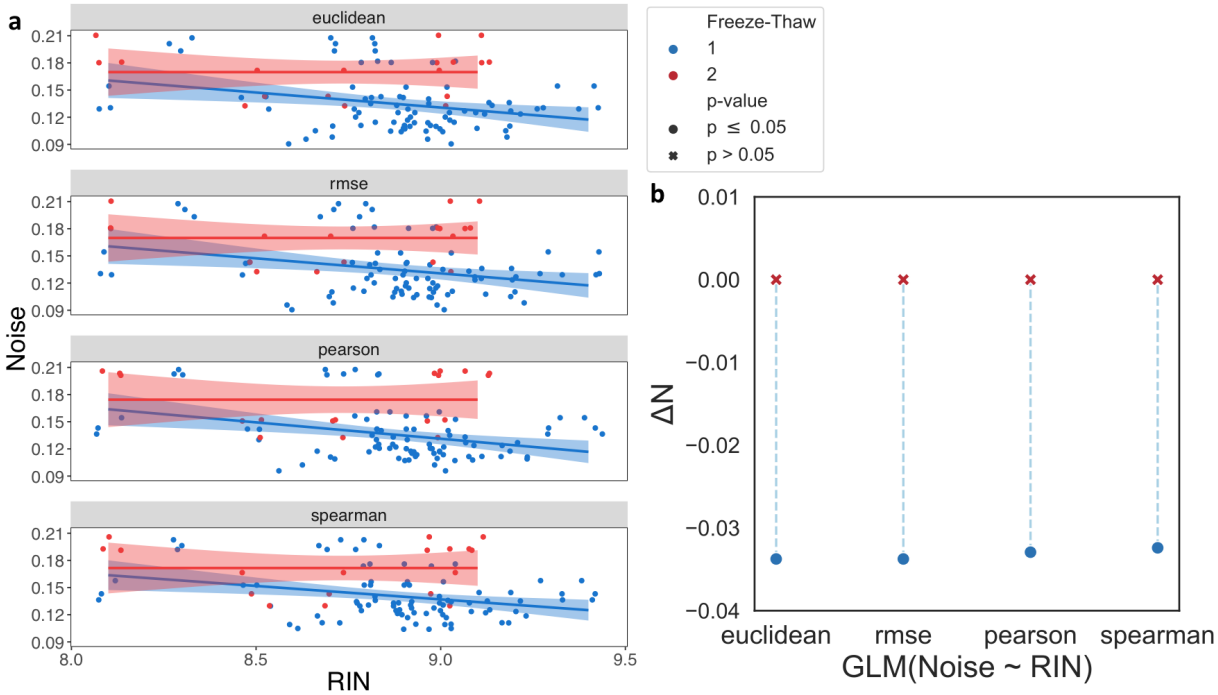


Figure 2: **Discrepancy in the Relationship Between Noise and RIN due to additional Freeze-Thaw.** *Examining the relationship between noise, calculated by Euclidean distance, RMSE, pearson, and spearman correlation, for samples that underwent either one (blue) or two (red) freeze-thaw cycles. (a) Scatter 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 ≤ 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 ≤ 6.3e-3) (Fig. **2a-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 may 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.

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.

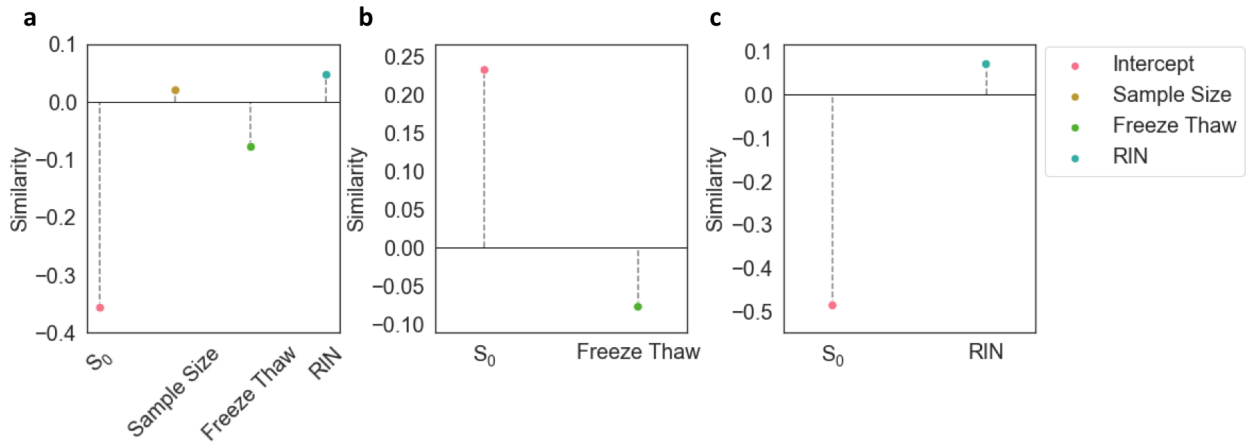


Figure 3: **Freeze-Thaw and RIN Both Confer Expected Effects on Similarity.** *Generalized linear models were used to quantify the change in similarity per unit increase in (a) sample size, number of freeze-thaws and RIN combined, (b) only the number of freeze-thaws, and (c) only RIN. S_0 represents the intercept estimate and sample size, freeze-thaw, and RIN represent coefficient estimates. All estimates are significant.*

As expected, similarity increases with subset size. This is reflected in the upward shift in the similarity distribution with increasing subset size (**Fig. S9**) and the estimated 0.02 (Wald test, p = 2.2e-5) increase in similarity per additional sample (**Fig. 3a)**; 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**.

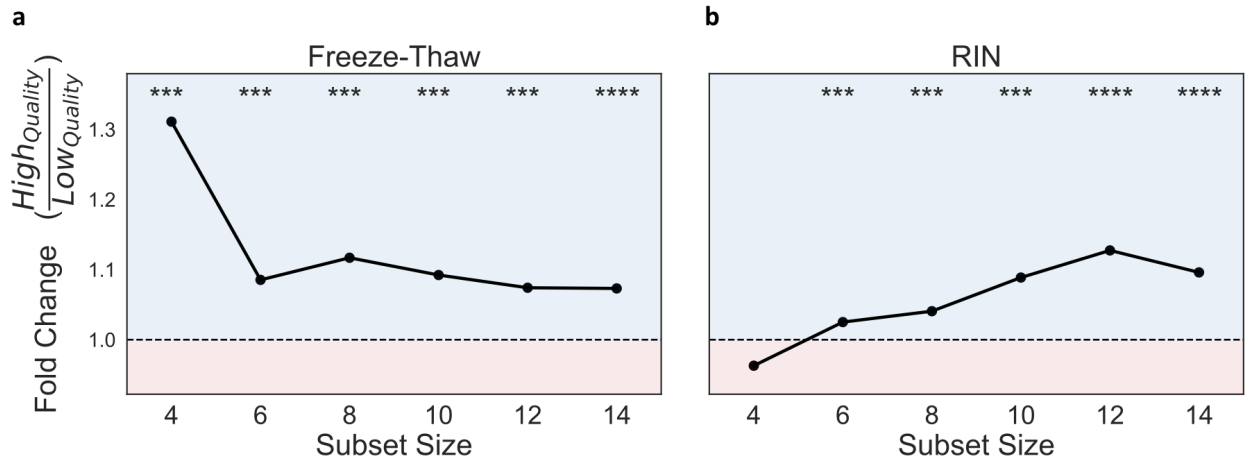


Figure 4: Higher Quality Samples have Higher Similarity at Most Subset Sizes. *Fold-change in median similarity of high quality subsets with respect to low quality subsets at each subset size. The region shaded in blue (fold-change > 1) indicates instances where the median similarity for high quality 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 (a) freeze-thaw or (b) 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. 4a-b**). 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. 3a-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. 3b)**. Even when accounting for subset size and the effects of RIN, the estimated decrease in similarity from freeze-thaw is nearly the same--0.078 (Wald test, $p = 8.77e\text{-}4$); 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**. **S10, 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 sizes (**Fig. 5a**).

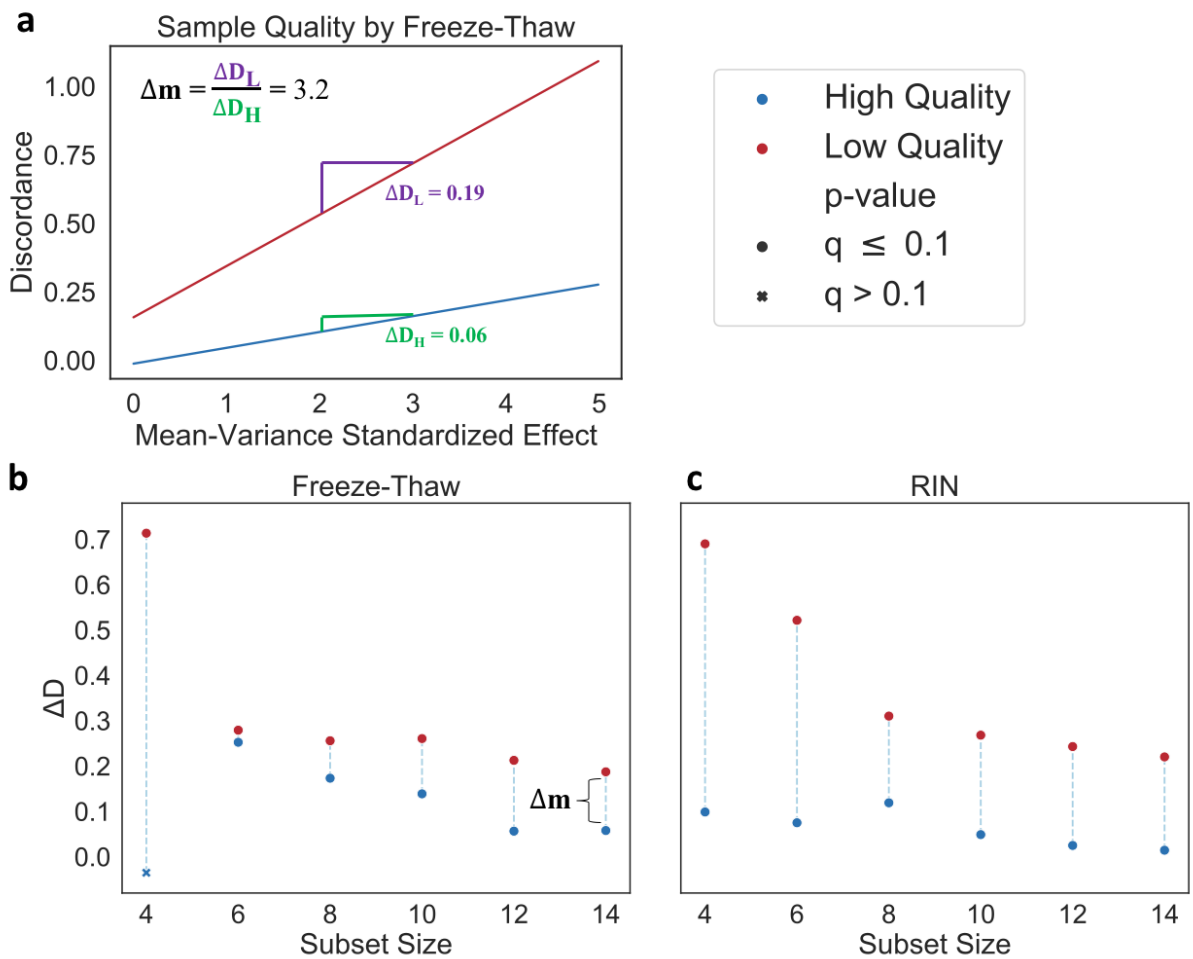


Figure 5: Higher Sample Quality has Lower Discordance at Each Subset Size. *GLM estimates of discordance predicted from effect size and subset size at high and low quality subsets at each subset size. High sample quality (blue) is compared to low sample quality (red). ΔD values represent the change in 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-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

202 across all subset sizes (**Fig. 5b-c**). Nearly all estimates are significant after multiple test
203 correction (Wald test, Benjamini-Hochberg FDR correction, $q \leq 0.07$), with the exception of
204 those for the smallest subset size for freeze-thaw.

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

207 Additional freeze-thaw cycles show increased 3' bias in poly(A)-enriched but 208 not ribosomal depletion samples

209 To identify one potential source in the distortion of RNA-Seq data and results, we asked if
210 repeated freeze-thaw cycles introduce a bias in read coverage of poly(A)-enriched samples. If
211 freeze-thaw enhances transcript degradation, we expect to see more reads on the 3' end of
212 poly(A)-enriched samples; the 3' end must be present due to the extraction method but, due to
213 degradation and cleavage between the 3' and 5' ends, there may be fewer reads covering the 5'
214 end. Thus, we analyzed the median read coverage with respect to the 0th percentile nucleotide
215 (5' end), the 50th percentile (uniform coverage over the gene), or the 100th percentile nucleotide
216 (3' end).

217 We compared cumulative coverage from the 5' to 3' end between poly(A)-enrichment or
218 ribosome depletion extraction methods. Specifically, we assess the median coverage percentile,
219 the transcript-wide average nucleotide percentile at which median cumulative coverage for a
220 given sample is achieved (**Fig. 6a**). We found a shift in mRNA coverage towards the 3' end in
221 the poly(A)-enriched samples . However, mRNA samples prepared using ribosome depletion did
222 not demonstrate this 3'-bias (**Fig. S12a**). We note that the median coverage percentile for
223 poly(A)-enriched samples is significantly (one-sided Wilcoxon test, $p < 2.2e-16$) larger than that
224 of ribosome depletion (**Fig. S12b**).

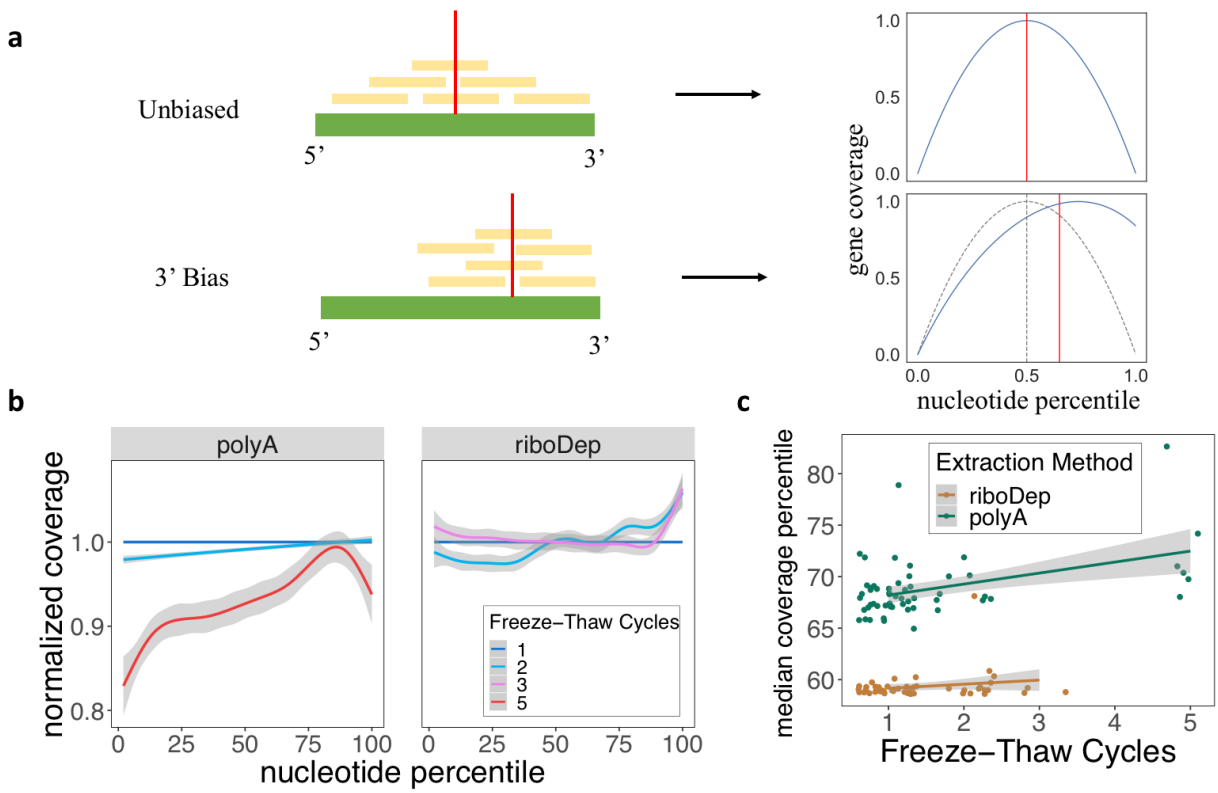


Figure 6: **Freeze-Thaw Cycles Exacerbate 3' Bias in poly(A)-enriched samples.** *(a) Demonstration for determining median coverage percentile (red vertical line). When coverage is unbiased, reads (yellow) are distributed throughout the entire body of the transcript (green). In the absence of read bias and observing coverage as a function of the nucleotide percentile, we see that cumulative coverage along the transcript reaches 50% at the 50th percentile nucleotide. In contrast, given a 3' read bias, there is a shift in the distribution of reads and cumulative coverage is 50% at the 60th percentile nucleotide. This results in a rightward shift in median coverage percentile towards the 3' end of the transcript. This is analogous to an increase in median coverage percentile from 50 to 60. (b) Gene coverage (y-axis) at the ith nucleotide percentile (x-axis) for samples that underwent 1-5 freeze-thaws and were extracted using either poly(A)-enrichment or ribosome depletion. Coverage is normalized to samples that underwent one freeze-thaw. For each sample, coverage is averaged across all genes; samples are aggregated using generalized additive model smoothing, with shaded regions representing 95% confidence intervals. (c) Linear model fits comparing the change in median coverage percentile to the number of freeze-thaw cycles for ribosome depletion (orange) or poly(A)-extraction (green). d. Shaded regions represent 95% confidence intervals.*

The 3' bias is accentuated when samples are stratified by the number of freeze-thaw cycles **(Fig. 6b).** Samples prepared with poly(A)-extraction have more 3' coverage bias compared to ribosome depletion in both one (one-sided Wilcoxon test, p = 3.5e-15) and two (one-sided Wilcoxon test, p = 1.4e-4) freeze-thaw cycles **(Fig. S12d)**, indicating an overall 3' bias of poly(A)-enriched samples, even independently of freeze-thaw **(Fig. S12b**).

We observe a significant increase (Wald test, p = 8.3e-4) in median coverage percentile due to the number of freeze-thaw cycles in poly(A)-enrichment. The increase was not maintained when

mRNA was isolated using ribosome depletion (Wald test, p = 0.155) **(Fig. 6c, Supplementary Table 7)**. For poly(A)-enriched samples, median coverage percentile increases 3.3 percentage points per root freeze-thaw cycle; the square-root of freeze-thaw was used to stabilize variance.

We further demonstrate a dependency of 3' bias on freeze-thaw cycles by showing that median coverage percentile significantly increases with freeze-thaw in poly(A)-enriched samples (Kruskal-Wallis test, p = 0.01). This 3' bias is particularly apparent after five 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 ribosome depletion (Kruskal-Wallis test, p = 0.084) **(Fig. S12c)**.

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. If this 3' bias is the root cause of freeze-thaw induced instability in absolute and differential RNAseq quantification, such instabilities may be subverted by substituting poly(A) selection for ribosome 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 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 the stability of gene expression analysis. By our estimates and at these subset sizes, reproducibility of a differential expression signature could approache zero after only three freeze-thaws (**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[32,33]. As such, accounting for technical variation has been a major research area of focus for the past decade[1,2,5,33,34]. 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, direct connection between freeze-thaw and sample quality has mainly been assessed via RIN[25,35,36].

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[37], 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. S14**). Thus, each freeze-thaw introduces a non-negligible level of noise to the quantification of gene expression and differential expression 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.)[25,38], 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[16]. 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[39]. 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[38]. 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-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[40–43] due to freeze-thaw sheds light on these challenges, since RIN cannot quantify quality at the individual gene level[22]. This is reflected in the fact that samples with RIN > 8 demonstrate degradation[44]. 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[25], 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[27,45].

328 As such, we re-examined the utility of RIN as a measure of sample quality in relation to our
329 noise estimation of random reads per sample. We found that while noise increases with both
330 decreasing RIN and increasing freeze-thaw, RIN may be an insufficient indicator of quality for
331 samples that have undergone two or more freeze-thaws. Given these results, RIN may not
332 always be a good metric to quantify the difference between technical replicates that have
333 undergone variable sample handling. We validate noise by confirming that it does not change
334 with input RNA concentration, excepting outliers **(Fig. S11**). Therefore, in the future, noise
335 could be a useful supplement to RIN when technical replicates are present.

336 To confirm our expectation that freeze-thaw decreases sample quality and to further characterize
337 the mechanism, we checked for the presence of a 3' bias in coverage. This builds on observations
338 that a lower percentage of poly(A)-enriched transcripts are covered[36]. We compared coverage to
339 ribosome depleted RNA-Seq data, which does not use 3' hybridization to retain transcripts. We
340 find that poly(A)-enrichment does in fact introduce a strong 3' bias in coverage as compared to
341 ribosome depletion. This bias is further exacerbated with additional freeze-thaw cycles in
342 poly(A)-enriched but not ribosome depleted samples. This implies that degradation due to
343 freeze-thaw may not impact RNA-sequencing of ribosome depleted samples to the extent that it
344 does in poly(A)-enriched samples.

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

351 Conclusion

352 Altogether, these results indicate that transcriptomics quality control steps cannot rely on RIN
353 alone for samples that have undergone poly(A)-enrichment and multiple freeze-thaws.
354 Furthermore, accounting for the effect of freeze-thaw on poly(A)-enriched RNA sequencing is
355 crucial. poly(A)-enrichment is prevalent for RNA-sequencing, and, in parallel, samples that
356 undergo multiple freeze-thaws are common in many protocols, especially rare tissues, e.g.,
357 postmortem neural tissue. Yet, there is no clear recommendation to avoid poly(A)-enrichment
358 following multiple freeze-thaws. Together, these results indicate that ribosome depletion could
359 be a better alternative when freeze-thaw is necessary.

360 **Author contributions**

361 B.P.K., N.E.L., T.P., S.M., and E.C. designed and planned the experiments. T.P., S.N., K.P.,
362 S.M. and L.L. collected the samples, managed diagnostics, conducted transcriptome assays and
363 managed the data. B.P.K, H.M.B. and V.G. planned and conducted analyses. I.S. performed
364 RNA-seq QC. A.B. performed functional enrichment analyses. S.L. wrote the RNA-Seq
365 processing pipeline. B.P.K., H.M.B, and N.E.L wrote the manuscript. N.E.L. supervised the

project.

Acknowledgements

This work was supported by NIMH R01-MH110558 (E.C., N.E.L., T.P., V.G., S.N., K.P.,L.L.), NIDCD R01-DC016385 (E.C., T.P.), T32GM008806 (H.M.B.), R35 GM119850 (B.P.K., I.S.), the Simons Foundation (E.C.), and generous funding from the Novo Nordisk Foundation through Center for Biosustainability at the Technical University of Denmark (NNF10CC1016517 N.E.L., S.L.).

Methods

Subject Recruitment

In this study, we performed transcriptomics analyses of blood samples drawn from 100 male toddlers with the age range of 1–4 years. Research procedures were approved by the Institutional Review Board of the University of California, San Diego. Parents of toddlers underwent informed consent procedures with a psychologist or study coordinator at the time of their child's enrollment. Additional details for the recruitment protocol are executed as described in Gazestani et al.[46]

Sample Collection and Storage

Blood samples were usually taken at the end of the clinical evaluation sessions. To monitor health status, the temperature of each toddler was monitored using an ear digital thermometer immediately preceding the blood draw. The blood draw was scheduled for a different day when the temperature was higher than 37 °C. Moreover, blood draw was not taken if a toddler had some illness (for example, cold or flu), as observed by us or stated by parents. We collected 4–6 ml blood into EDTA-coated tubes from each toddler. Blood leukocytes were captured using LeukoLOCK filters (Ambion). After rinsing the LeukoLOCK filters with PBS, the filters were flushed with RNAlater (Invitrogen) to stabilize 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). Ribosome depletion was used to prepare an additional 96 samples (from 84 individuals). Relevant metadata regarding polyA-enriched 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 [47]. 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) [48]. Finally, HTSeq using default settings (v0.6.1) was used to quantify reads [48,49].

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. S1)**. 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. S2, Figure S1**). 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 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. S3**). Before subsetting, we filtered our expression matrix for genes with an average count ≤ 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

441 unique subsets. Each subset had an equal number of TD or ASD samples. Additionally, only one
442 replicate from each blood sample could be included. These requirements limited our subset size
443 to a maximum of 14 samples.

444 DE between ASD and TD subjects was conducted using DESeq2[1]. **Fig. S9** summarizes DE
445 results for all subsets. To account for potential confounders, we used RUV[5]. Specifically, we use
446 a set of "in-silico empirical" negative control genes, including all but the top 5,000 differentially
447 expressed genes as described in section 2.4 of the documentation for RUVseq
448 (http://bioconductor.org/packages/release/bioc/vignettes/RUVSeq/inst/doc/RUVSeq.pdf). We
449 check that RUV produces consistent results with previous Autism leukocyte gene expression
450 signatures[46,50], see **Supplementary Results.**

451 Similarity to Assess Differential Expression Reproducibility

452 To assess DE reproducibility, we measure the similarity in log-fold-change (LFC) values
453 between DE runs. Similarity is calculated as the pairwise spearman correlation of LFC between
454 all subsets of the same size **(Fig. S5)**. Genes with a median base mean (the mean of counts of all
455 samples, normalizing for sequencing depth) or median LFC in the bottom 10th percentile across
456 all subsets were excluded, filtering for low magnitude effects **(Fig. S4)**.

457 Average RIN and freeze-thaw were measured for all subset pairs. Resulting distributions for all
458 collected values from similarity analyses are displayed in **Fig. S6**.

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

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

468 Discordance to Assess Differential Expression Reproducibility

469 We adapted a measure of concordance to measure discordance, or the lack of reproducibility,
470 between differential expression results[51]. Average RIN and freeze-thaw were calculated for each
471 subset **(Fig. S7-8)**. Subsets for each subset size were split into two quantile bins for either quality
472 metric (RIN and freeze-thaw). Genes with a median base mean across all subsets in the bottom
473 tenth percentile were excluded from the analysis **(Fig. S4)**.

474 We do not use the original concordance at the top (CAT) metric because we are not comparing
475 our results to a gold standard dataset. Instead, we use gene-wise LFC standard deviation across
476 subsets as a measure of discordance. Thus, the average LFC for each gene across DE runs is
477 analogous to the gold standard, and the dispersion from this average indicates a lack of
478 reproducibility. At each combination of subset size and quality bins, we calculate discordance

and compare it to the gene-wise median effect size **(Fig. S7**). 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. S8**. 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 [52]. 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 normalized across all genes within a sample.

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

501 **References**

502 1. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion
503 for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
504 2. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for
505 differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139
506 (2010).
507 3. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq
508 experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
509 4. Leek, J. T. & Storey, J. D. Capturing Heterogeneity in Gene Expression Studies by
510 Surrogate Variable Analysis. *PLoS Genet.* **3**, e161 (2007).
511 5. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor
512 analysis of control genes or samples. *Nat. Biotechnol.* **32**, 896–902 (2014).
513 6. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression
514 data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
515 7. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15 (2013).
516 8. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17**, 1–
517 19 (2016).
518 9. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq
519 quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
520 10. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and
521 bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
522 11. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
523 (2013).
524 12. Jun, E. *et al.* Method Optimization for Extracting High-Quality RNA From the Human
525 Pancreas Tissue. *Transl. Oncol.* **11**, 800–807 (2018).
526 13. Passow, C. N. *et al.* RNAlater and flash freezing storage methods nonrandomly influence
527 observed gene expression in RNAseq experiments. *bioRxiv* 379834 (2018)
528 doi:10.1101/379834.
529 14. Micke, P. *et al.* Biobanking of fresh frozen tissue: RNA is stable in nonfixed surgical
530 specimens. *Lab. Invest.* **86**, 202–211 (2006).
531 15. Ohmomo, H. *et al.* Reduction of Systematic Bias in Transcriptome Data from Human
532 Peripheral Blood Mononuclear Cells for Transportation and Biobanking. *PLoS One* **9**,
533 (2014).
534 16. Xiong, B., Yang, Y., Fineis, F. R. & Wang, J. P. DegNorm: normalization of generalized
535 transcript degradation improves accuracy in RNA-seq analysis. *Genome Biol.* **20**, 75–75
536 (2019).
537 17. Li, S. *et al.* Multi-platform assessment of transcriptome profiling using RNA-seq in the
538 ABRF next-generation sequencing study. *Nat. Biotechnol.* **32**, 915–925 (2014).
539 18. Svensson, V. *et al.* Power analysis of single-cell RNA-sequencing experiments. *Nat.*
540 *Methods* **14**, 381–387 (2017).
541 19. Locksley E. McGann, Hongyou Yang, Michele Walterson. Manifestations of cell damage
542 after freezing and thawing. *Cryobiology* **25**, 178–185 (1988).
543 20. Röder, B., Frühwirth, K., Vogl, C., Wagner, M. & Rossmanith, P. Impact of long-term
544 storage on stability of standard DNA for nucleic acid-based methods. *J. Clin. Microbiol.* **48**,
545 4260–4262 (2010).
546 21. Shao, W., Khin, S. & Kopp, W. C. Characterization of effect of repeated freeze and thaw

547 cycles on stability of genomic DNA using pulsed field gel electrophoresis. *Biopreserv.*
548 *Biobank.* **10**, 4–11 (2012).
549 22. Schroeder, A. *et al.* The RIN: an RNA integrity number for assigning integrity values to
550 RNA measurements. *BMC Mol. Biol.* **7**, 3 (2006).
551 23. Reiman, M., Laan, M., Rull, K. & Sõber, S. Effects of RNA integrity on transcript
552 quantification by total RNA sequencing of clinically collected human placental samples.
553 *FASEB J.* **31**, 3298–3308 (2017).
554 24. Shen, Y. *et al.* Impact of RNA integrity and blood sample storage conditions on the gene
555 expression analysis. *Onco. Targets. Ther.* **11**, 3573 (2018).
556 25. Romero, I. G., Pai, A. A., Tung, J. & Gilad, Y. RNA-seq: impact of RNA degradation on
557 transcript quantification. *BMC Biol.* **12**, 42 (2014).
558 26. Sonntag, K.-C. *et al.* Limited predictability of postmortem human brain tissue quality by
559 RNA integrity numbers. *J. Neurochem.* **138**, 53 (2016).
560 27. Yu, K. *et al.* Effect of multiple cycles of freeze–thawing on the RNA quality of lung cancer
561 tissues. *Cell and Tissue Banking* vol. 18 433–440 (2017).
562 28. Jaffe, A. E. *et al.* qSVA framework for RNA quality correction in differential expression
563 analysis. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 7130 (2017).
564 29. Bao, W.-G. *et al.* Biobanking of Fresh-frozen Human Colon Tissues: Impact of Tissue Ex-
565 vivo Ischemia Times and Storage Periods on RNA Quality. *Ann. Surg. Oncol.* **20**, 1737–
566 1744 (2012).
567 30. Zeugner, S., Mayr, T., Zietz, C., Aust, D. E. & Baretton, G. B. RNA Quality in Fresh-
568 Frozen Gastrointestinal Tumor Specimens—Experiences from the Tumor and Healthy
569 Tissue Bank TU Dresden. in *Pre-Analytics of Pathological Specimens in Oncology* 85–93
570 (Springer, Cham, 2015).
571 31. Li, J., Jiang, H. & Wong, W. H. Modeling non-uniformity in short-read rates in RNA-Seq
572 data. *Genome Biology* **11**, (2010).
573 32. McIntyre, L. M. *et al.* RNA-seq: technical variability and sampling. *BMC Genomics* **12**, 1–
574 13 (2011).
575 33. Hansen, K. D., Irizarry, R. A. & Wu, Z. Removing technical variability in RNA-seq data
576 using conditional quantile normalization. *Biostatistics* **13**, 204–216 (2012).
577 34. Leek, J. T., Evan Johnson, W., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package
578 for removing batch effects and other unwanted variation in high-throughput experiments.
579 *Bioinformatics* vol. 28 882–883 (2012).
580 35. Copois, V. *et al.* Impact of RNA degradation on gene expression profiles: assessment of
581 different methods to reliably determine RNA quality. *J. Biotechnol.* **127**, 549–559 (2007).
582 36. Adiconis, X. *et al.* Comparative analysis of RNA sequencing methods for degraded or low-
583 input samples. *Nat. Methods* **10**, 623–629 (2013).
584 37. Pertea, M. *et al.* Thousands of large-scale RNA sequencing experiments yield a
585 comprehensive new human gene list and reveal extensive transcriptional noise. *bioRxiv*
586 332825 (2018) doi:10.1101/332825.
587 38. Wang, L. *et al.* Measure transcript integrity using RNA-seq data. *BMC Bioinformatics* **17**,
588 58 (2016).
589 39. Foley, J. W. *et al.* Gene expression profiling of single cells from archival tissue with laser-
590 capture microdissection and Smart-3SEQ. *Genome Research* vol. 29 1816–1825 (2019).
591 40. Wang, Y. *et al.* Precision and functional specificity in mRNA decay. *Proceedings of the*
592 *National Academy of Sciences* vol. 99 5860–5865 (2002).

41. Yang, E. *et al.* Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res.* **13**, 1863–1872 (2003).
42. Narsai, R. *et al.* Genome-wide analysis of mRNA decay rates and their determinants in Arabidopsis thaliana. *Plant Cell* **19**, 3418–3436 (2007).
43. 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).
44. Hoen, P. A. C. *et al.* Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat. Biotechnol.* **31**, 1015–1022 (2013).
45. Wang, Y. *et al.* The Impact of Different Preservation Conditions and Freezing-Thawing Cycles on Quality of RNA, DNA, and Proteins in Cancer Tissue. *Biopreserv. Biobank.* **13**, 335–347 (2015).
46. Gazestani, V. H. *et al.* A perturbed gene network containing PI3K/AKT, RAS/ERK, WNT/β-catenin pathways in leukocytes is linked to ASD genetics and symptom severity. *bioRxiv* 435917 (2019) doi:10.1101/435917.
47. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114 (2014).
48. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
49. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
50. Pramparo, T. *et al.* Prediction of autism by translation and immune/inflammation coexpressed genes in toddlers from pediatric community practices. *JAMA Psychiatry* **72**, 386–394 (2015).
51. Collado-Torres, L. *et al.* Reproducible RNA-seq analysis using recount2. *Nat. Biotechnol.* **35**, 319–321 (2017).
52. Wang, L., Wang, S. & Li, W. RSeQC: quality control of RNA-seq experiments. *Bioinformatics* **28**, 2184–2185 (2012).