qSVA framework for RNA quality correction in differential expression analysis

Andrew E. Jaffe[a,b,c,d,1], Ran Tao[a], Alexis L. Norris[e,f], Marc Kealhofer[a,g], Abhinav Nellore[c,d,h], Joo Heon Shin[a], Dewey Kim[a], Yankai Jia[a], Thomas M. Hyde[a,i,j], Joel E. Kleinman[a,j], Richard E. Straub[a], Jeffrey T. Leek[c,d], and Daniel R. Weinberger[a,e,j,k]

[a]Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Baltimore, MD 21205; [b]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205; [c]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205; [d]Center for Computational Biology, Johns Hopkins University, Baltimore, MD 21205; [e]Department of Neuroscience, Johns Hopkins School of Medicine, Baltimore, MD 21205; [f]Department of Neurology, Kennedy Krieger Institute, Baltimore, MD 21205; [g]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205; [h]Department of Computer Science, Johns Hopkins University, Baltimore, MD 21205; [i]Department of Neurology, Johns Hopkins School of Medicine, Baltimore, MD 21205; [j]Department of Psychiatry and Behavioral Sciences, Johns Hopkins School of Medicine, Baltimore, MD 21205; and [k]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, MD 21205

Edited by Pasko Rakic, Yale University, New Haven, CT, and approved May 19, 2017 (received for review October 27, 2016)

RNA sequencing (RNA-seq) is a powerful approach for measuring gene expression levels in cells and tissues, but it relies on high-quality RNA. We demonstrate here that statistical adjustment using existing quality measures largely fails to remove the effects of RNA degradation when RNA quality associates with the outcome of interest. Using RNA-seq data from molecular degradation experiments of human primary tissues, we introduce a method—quality surrogate variable analysis (qSVA)—as a framework for estimating and removing the confounding effect of RNA quality in differential expression analysis. We show that this approach results in greatly improved replication rates (>3×) across two large independent postmortem human brain studies of schizophrenia and also removes potential RNA quality biases in earlier published work that compared expression levels of different brain regions and other diagnostic groups. Our approach can therefore improve the interpretation of differential expression analysis of transcriptomic data from human tissue.

RNA sequencing | differential expression analysis | statistical modeling | RNA quality

Microarrays and RNA sequencing (RNA-seq) can measure gene expression levels across hundreds of samples in a single experiment. As gene expression levels are measured with error, normalization procedures have been implemented for both microarray (1) and RNA sequencing (2) data to reduce technical variability, including controlling for variability associated with how and when the samples are run, so-called "batch" effects (3). Recent research has further characterized this expression variability in RNA-seq data (4–6), including demonstrating variability associated with technical factors involved in the preparation, sequencing, and analysis of samples. Variability in gene expression is particularly influenced by RNA quality (7) because accurately measuring gene expression levels strongly depends on the quality of the input RNA. This suggests that a portion of traditionally measured latent "batch" effects could actually be attributed to the underlying quality of the input RNA.

Postmortem studies typically extract RNA from tissue that has been susceptible to a wide variety of antemortem and postmortem factors. Several approaches exist for quantifying the quality of the input RNA before sequencing library construction, including UV absorption ratios of 280 nm to 260 nm and RNA integrity numbers (RINs). RIN is a machine learning-derived measurement resulting from placing RNA on a Bioanalyzer and obtaining a tracing of fragment sizes per sample. RIN ranges from 10 (very high quality RNA) to 0 (completely degraded RNA), and the apparent intactness of ribosomal RNAs (which are two large peaks in the fragment size tracing) is one of the most discriminating factors that distinguishes very high quality from moderate quality RNA (8). Recommended RIN thresholds for sample exclusion before data generation have been suggested as low as 5.0 for PCR (7) and 7.0 for RNA-seq. (9). However, even high quality samples (RIN > 8) demonstrate evidence of degradation, as transcriptome-wide gene expression levels strongly associate with RIN even among samples with high RINs, for example, in lymphoblastoid cell lines (6). Furthermore, the recent introduction of ribosomal depletion approaches for library construction, such as the Illumina Ribo-Zero technique, have permitted the sequencing of lower quality samples compared with previous polyadenylation section-based approaches (polyA+), including samples with RINs less than 3 (10).

Proposed measures of RNA quality can also be derived from the resulting RNA sequencing data, for example, by calculating the 5′ to 3′ read coverage bias (particularly in polyA+ data); transcript integrity numbers (11); various read mapping rates, including to autosomes, ribosomal RNAs, and mitochondrial RNAs (chrM); and gene/exon assignment rates (7). Although many of these approaches appear to capture the largest global effects on expression, for example, through positively correlating factors of expression data with the above-mentioned quality measures, the

Significance

Many studies use measurements of gene expression in human postmortem and ex vivo tissues like brain and blood to characterize genomic correlates of illness. However, molecular analyses of these tissues can be susceptible to a wide range of confounders that may be difficult to measure and remove. In this article, we describe an analysis framework for identifying and removing previously uncharacterized quality biases in measurements of RNA. Our paper critically highlights the shortcomings of standard RNA quality correction approaches, such as statistically adjusting for RNA integrity numbers. We show that the our framework removes residual confounding by RNA quality and greatly improves replication of significant differentially expressed genes across independent datasets by more than threefold compared with previous approaches.

Author contributions: A.E.J., J.T.L., and D.R.W. designed research; A.E.J., R.T., A.L.N., J.H.S., D.K., Y.J., T.M.H., J.E.K., R.E.S., J.T.L., and D.R.W. performed research; A.E.J. and A.N. contributed new reagents/analytic tools; A.E.J., A.L.N., and M.K. analyzed data; and A.E.J., J.T.L., and D.R.W. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The sequences reported in this paper have been deposited with the National Center for Biotechnology Information (NCBI BioProject number PRJNA389171 and NCBI SRA project SRP108559).

[1]To whom correspondence should be addressed. Email: andrew.jaffe@libd.org.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1617384114/-/DCSupplemental.

presence and role of more subtle and gene/transcript-specific biases in RNA quality on measures of gene expression and resulting differential expression analysis is unclear. Furthermore, application of existing statistical methods to model latent RNA quality risks retaining false-positive associations in supervised approaches such as surrogate variable analysis (SVA) (12) or removing an outcome-associated biological signal in unsupervised approaches such as principal component analysis (PCA). Here we describe a general analytic framework to estimate and remove RNA quality confounding in differential expression analysis that first identifies transcript features most susceptible to RNA degradation using tissue degradation experiments and subsequently corrects independent datasets using the expression levels of these transcript features. We show that this framework, called quality surrogate variable analysis (qSVA), better identifies and removes confounding related to RNA quality in differential expression analysis than do observed measures of RNA quality alone.

Results

Degradation Experiments to Model Changes in RNA Quality. We hypothesized that examining RNA degradation in human tissue from experimental approaches consisting of leaving tissue at room temperature would identify metrics useful for quantifying RNA quality. We therefore examined the transcriptional landscape of degradation in dorsolateral prefrontal cortex (DLPFC) tissue (in a degradation experiment that we performed) and blood (specifically, peripheral blood mononuclear cells—PBMCs—that were publicly available). Briefly, we left DLPFC tissue from five brains at room temperature (off of ice) for 0, 15, 30, and 60 min; extracted RNA; measured RINs; and then constructed and sequenced both polyA+ and RiboZero libraries (*Materials and Methods* and *SI Appendix*, Table S1). The PBMC degradation experiment was a similar design over a longer time period, ranging from 12 h to 84 h (13), and the resulting RNAs were sequenced with polyA+ libraries. The RNA-seq reads from both experiments were processed identically (*SI Appendix*, *Full Methods and Materials*). Many technical covariates were strongly associated with degradation time in both blood and brain (*SI Appendix*, Table S2), and PCA suggested that degradation time was the strongest explanatory variable (PC1) of the transcriptome across each library type, explaining 47.5% and 39.0% of normalized gene counts in polyA+ and RiboZero DLPFC libraries, respectively, and 54.4% in blood (*SI Appendix*, Fig. S1). These first PCs more independently associated with degradation time than RIN in multivariate regression analysis (*SI Appendix*, Table S2), suggesting that these degradation experiments induce widespread changes in RNA quality that are not fully recognized by RIN.

Different mRNAs Degrade at Different Rates in Human Tissues. Many genes were highly susceptible to the effects of RNA degradation, including 12,324 genes at a false discovery rate (FDR) < 5% significance in the DLPFC polyA+ dataset (n = 2,303 at p_{bonf} < 5%), 10,981 genes in the DLPFC RiboZero dataset (n = 2,017 at p_{bonf} < 5%), and 11,170 genes in blood polyA+ data (n = 2,833 at p_{bonf} < 5%, Dataset S1). Regardless of tissue or library type, increased susceptibility to RNA degradation (e.g., a more negative degradation t-statistic) was associated with increased gene length and increased coding lengths, increased transcript expression, decreased guanine–cytosine (GC) content, and increased number of annotated transcripts (all but one P value < 2.2×10^{-16}, *SI Appendix*, Table S3). Enrichment analyses among predefined gene sets suggested dysregulation of a wide variety of cellular processes associated with increased degradation susceptibility (Dataset S2).

Because RNAs from different cell types may degrade at different rates, and both blood and DLPFC are mixtures of diverse cell types, we explored the role of cell-type–specific signal on RNA degradation. We estimated the relative proportions of 22 different blood cell types using existing reference data in the PBMCs (14) and found significant changes in the relative cellular composition comparing degraded to intact PBMC samples. Increased degradation time decreased the relative proportion of monocytes ($P = 1.82 \times 10^{-5}$) and increased the relative proportion of macrophages ($P = 8.63 \times 10^{-5}$), regulatory T cells ($P = 5.47 \times 10^{-3}$), and activated mast cells ($P = 1.07 \times 10^{-6}$, *SI Appendix*, Fig. S2 and Dataset S3). In DLPFC, because such a reference profile of brain cell types does not exist, we derived cell-type–specific candidate gene lists using available single-cell RNA-seq data (15). We found significant enrichment of these candidate genes among our degradation statistics overall ($P < 2.2 \times 10^{-16}$, *SI Appendix*, Fig. S3) as well as differential degradation effects by cell type (*SI Appendix*, Table S4 and *Materials and Methods*). These enrichment analyses indeed suggest that RNAs from different cell types may be differentially susceptible to degradation, which is captured uniquely by different RNA-seq library preparation methods.

Biological and Technical Specificity of RNA Degradation Transcriptome Associations. Given the strong influence of RNA degradation on the transcriptome, we examined whether these degradation effects were brain- and degradation-method–specific. We directly compared the DLPFC polyA+ and PBMC degradation datasets to determine tissue specificity. The rate of degradation, as measured by RIN, was more rapid in our brain samples, as PBMCs still had high quality RNA after 12 h at room temperature (all RINs > 7.7), compared with DLPFC samples having RINs less than 6.6 after just 1 h. We found only a weak global correlation between the gene degradation susceptibility statistics (*SI Appendix*, Fig. S4*A*) and much smaller degradation rates of individual genes (median: 33.6% versus 0.44%; 90th percentile: 213.8% versus 1.4% change per hour) between PBMCs and DLPFC, suggesting global differences in the transcriptome changes resulting from degradation. However, we processed public Association of Biomolecular Resource Facilities Next Generation Sequencing (ABRF-NGS) data (that were sequenced with a RiboZero protocol) that compared three brain reference RNA samples treated with RNase-A to nine untreated samples (10). In this RNA (rather than tissue) degradation experiment, there were 13,553 genes significantly associated with RNase treatment (at FDR < 5%). There was significant global overlap between degradation induced by our experiment at the tissue level (using DLPFC RiboZero data) compared with the RNA levels: 7,700 (65.7%) genes were significantly differentially expressed in both experiments (odds ratio: 6.28, $P < 2.2 \times 10^{-16}$) and there was significant global correlation of degradation susceptibility statistics ($P < 2.2 \times 10^{-16}$, *SI Appendix*, Fig. S4*B*). Therefore, the strongest RNA degradation effects appear tissue-specific, but within a tissue, RNase A-like activity is likely a major factor contributing to the RNA degradation.

Strong Bias in Differential Expression Analysis in Confounded Designs. Based on the preceding results, we thus reasoned that many prior findings in differential expression analyses of postmortem brain datasets may have been susceptible to RNA degradation confounding. For example, many studies comparing different diagnostic groups typically have significant group differences in measures of RNA quality (e.g., RINs). We therefore used two large RNA-seq datasets from the prefrontal cortex comparing patients with schizophrenia to adult controls: Lieber Institute for Brain Development (LIBD, "discovery" data, polyA+ protocol, n = 351) and CommonMind Consortium (CMC, "replication" data, RiboZero protocol, n = 331) (14). Both studies indeed had significantly lower RINs in the control versus schizophrenia groups: LIBD—$P = 4.4 \times 10^{-5}$ (mean RIN: 8.4 versus 8.1) and CMC—$P = 7.6 \times 10^{-8}$ (mean RIN: 7.8 versus 7.4). We first created a new diagnostic plot to compare differential expression statistics for outcome to the degradation statistics from RNA degradation experiments (fold change in expression per minute or its corresponding t-statistic). This approach, which we call the "differential expression quality" (DEqual) plot, can illustrate

transcriptome-wide RNA degradation bias in a given dataset. We observed strong positive correlation between univariate differential expression statistics for diagnosis and experimental degradation in the LIBD dataset (Fig. 1*A* and *SI Appendix*, Fig. S5*A*). Here the directionality of change associated with diagnosis at a particular gene can be predicted almost entirely by its relationship with degradation and the difference in RNA quality between outcome groups. Among the 24,122 genes with reads per kilobase per million mapped (RPKM) > 0.1, we found that 11,408 (47.3%) genes were significantly differentially expressed at FDR < 5% in the discovery dataset, further suggesting confounding by RNA quality. We posit that removing the correlation between degradation-associated and diagnosis-associated statistics illustrated in the DEqual plot will show that RNA quality has been properly adjusted for in the differential expression analysis.

Statistically Adjusting for RIN Fails to Remove Degradation Bias. Given the DEqual plots from the univariate analysis, the significant difference in RIN between the schizophrenia and control groups, and the large number of differentially expressed genes, we expected that adjusting the differential expression analysis for RINs would reduce the degree of degradation bias. However, RIN adjustment only partially reduced the correlation between diagnosis and degradation statistics (Figs. 1*B* and *SI Appendix*, Fig. S5*B*) from Pearson correlation, $r = 0.464$ to $r = 0.358$ and only reduced the number of FDR-significant differentially expressed genes from 11,408 to 6,622 in the discovery dataset. The degree of RNA degradation bias was practically identical when further modeling RIN nonlinearly, e.g., further adjusting for RIN and RIN^2 (*SI Appendix*, Fig. S6). We further adjusted the differential expression analysis for other observed variables, including clinical and technical covariates ("observed" model: age, sex, ethnicity, chrM map rate, gene assignment rate, and RIN), which again only partially reduced both the correlation between diagnosis and degradation statistics (to $r = 0.291$, Fig. 1*C*) and the number of genes that were significantly differentially expressed ($n = 2{,}215$).

We also used the PBMC degradation dataset to show that RIN adjustment fails to account for the differences in RNA degradation between outcome groups. Here, we modeled differences in expression between individuals 1 and 2 after inducing confounding by degradation time by removing $T = 0$ for individual 1 and $T = 84$ for individual 2 (*Materials and Methods*). As expected, univariate analysis showed a strong correlation between the individual effect and the degradation effect (*SI Appendix*, Fig. S7*A*, $r = 0.495$). Again, statistical adjustment for RIN in this confounded design does not remove the strong degradation bias (*SI Appendix*, Fig. S7*B*, $r = 0.307$). Here, in this experimental dataset, unlike the schizophrenia case-control datasets described above, we have a gold standard surrogate of RNA degradation—the time at room temperature—and show that adjusting for this measure completely removes the RNA degradation bias (*SI Appendix*, Fig. S7*C*, $r = -0.09$). These results suggest that RIN or other observed quality variables may be a poor surrogate for total RNA quality and that adjusting for RIN in differential expression analysis is insufficient to remove potential RNA degradation confounding.

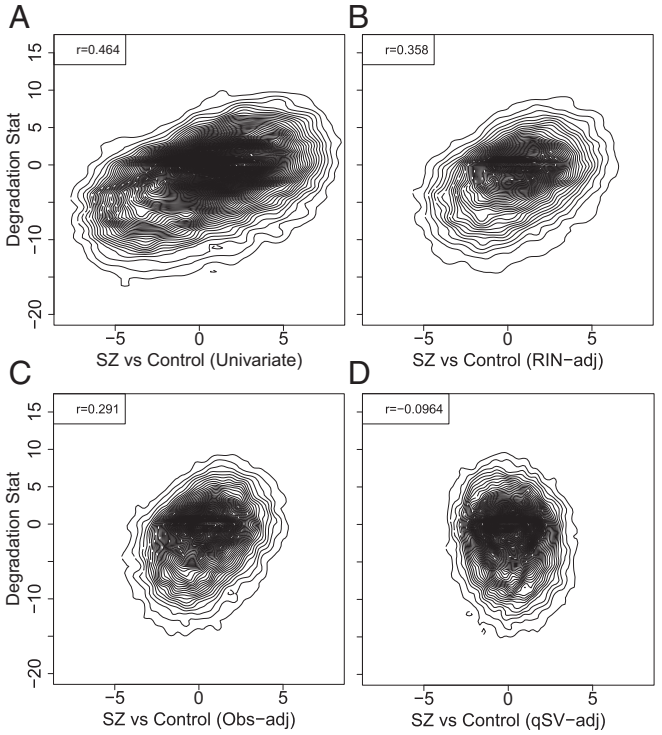

Fig. 1. Differential expression quality (DEqual) plots for schizophrenia-control expression differences. Each DEqual plot compares the effect of RNA degradation from an independent degradation experiment on the *y* axis to the effect of the outcome of interest, here schizophrenia (SZ) compared with controls. Each point is a gene, and effects here are shown as *T*-statistics for each effect. (*A*) DEqual plot for univariate case-control analysis shows strong correlation between degradation and diagnosis effects. (*B*) DEqual plot for RIN-adjusted case-control differences largely fails to remove degradation bias. (*C*) DEqual plot when adjusting for observed clinical and technical covariates, including age, sex, ethnicity, chrM mapping rate, gene assignment rate, and RIN, also fails to remove degradation bias. (*D*) DEqual plot demonstrating that the qSVA framework successfully removes positive correlation between degradation and SZ effects.

qSVA to Correct for RNA Degradation Bias. We hypothesized that we could leverage the experimental degradation datasets to better estimate factors related to RNA quality in RNA-sequencing datasets. This approach relies on estimating the transcript features most susceptible to RNA degradation and using these features as "negative control" features akin to approaches such as remove unwanted variation (RUV) (2) or SVA (12). The broad concept of the algorithm is to identify transcript features that are especially sensitive to degradation in the tissue of interest and then to quantify these same features in the experimental differential expression dataset and create a set of factors that are used to control for RNA quality bias (see *SI Appendix*, *Full Methods and Materials* for details). We defined those features that were Bonferroni-significantly associated with degradation in each dataset: the top 1,000 features in the DLPFC and PBMC polyA+ datasets (among thousands that were significant) and the 515 features in the DLPFC RiboZero data (step #1, see *SI Appendix*, *Full Methods and Materials* and Datasets S4 and S5). Interestingly, the transcript features in DLPFC across these two library types were completely non-overlapping, suggesting that the features most susceptible to degradation likely differ by library type. Within polyA+ data, there were only four degradation-susceptible features overlapping between DLPFC and PBMCs (within genes: *PNKD*, *MBOAT7*, *ENG*, and *SULF2*). These features can then be quantified in new user-provided samples for step #2 from BAM or BigWig files (*SI Appendix*, *Full Methods and Materials*), resulting in coverage estimates for each feature and new sample. Then, for step #3, factor analysis on the log-transformed degradation matrix of coverage estimates generates quality surrogate variables (qSVs). In step #4, the qSVs are then included as adjustment variables in differential expression analysis. The qSVA approach is available in the SVA Bioconductor package (https://bioconductor.org/packages/sva) (16), and the example code to run the statistical framework is described in *SI Appendix*, *Full Methods and Materials*.

Improved Replication for Schizophrenia Differential Expression Using qSVA. We applied the qSVA algorithm to the LIBD polyA+ RNA-seq data with the observed model (consisting of observed clinical and technical confounders) described above. Here, adjustment completely attenuated degradation bias (Fig. 1*D*, $r = -0.09$ using T-statistics and $r = -0.037$ using $\log_2$ fold changes). Following this adjustment, there were only 183 genes differentially expressed at FDR < 5%, further suggesting a reduction of RNA degradation bias in differential expression analysis of schizophrenia patients versus controls. The qSVs themselves were strongly associated with observed variables including chrM alignment rate, RIN, total gene assignment rate, overall alignment rate, age, and postmortem interval (*SI Appendix*, Fig. S8). Similarly, in the CMC dataset, the qSVs, calculated using the DLPFC RiboZero-based degradation features, were strongly associated with RIN, total gene assignment rate, institute where the sample was collected, and sequencing and flow cell batches (*SI Appendix*, Fig. S9). These results suggest that enriching for degradation signal via the independent tissue degradation experiment can identify more robust measures of RNA quality directly from RNA-seq experiments than relying on single observable measures.

Although the qSVA approach appears to remove RNA degradation bias in brain differential expression analysis as illustrated in the DEqual plot, we further observed that adjusting for transcriptome-wide PCs also removes the degradation effects (*SI Appendix*, Fig. S10, $r = -0.02$). This suggests that factor-based approaches—including qSVA but also more generally PCA—can identify and subsequently remove latent measures of RNA quality. However, unsupervised approaches like PCA run the risk of removing true biological difference. Moreover, "supervised" factor-based approaches, such as SVA that rely on residualizing around a provided statistical model, largely preserved RNA degradation bias (*SI Appendix*, Fig. S11). We therefore used replication signal across these independent datasets—LIBD and CMC—to more fully contrast the value of the different degradation adjustment approaches. For a given adjustment approach, we calculated replication rates of differentially expressed genes discovered in the LIBD dataset at different significance thresholds in the CMC dataset. We found the lowest replication rates (<20%) regardless of significance threshold when adjusting only for observed clinical and technical variables including RIN, as well as SVA residualizing on only diagnosis (Fig. 2). Although we had high replication rates among marginally significant genes ($P < 0.001$) using SVA residualizing on the observed variables described above, we found strong inflation of the test statistics among both the LIBD (9,033 genes at FDR < 5%) and CMC (6,924 genes at FDR < 5%) datasets. Among those genes significantly differentially expressed ($P < 10^{-4}$), we found the highest replication rates using qSVA, as well as relatively linear improvements in the replication rate as the discovery P values threshold dropped. Importantly, the qSVs calculated in the LIBD and CMC datasets were based on different degradation features, as the CMC data were RiboZero and the LIBD data were polyA+. These results therefore show that qSVA leads to greatly improved replication in postmortem brain transcriptomic studies.

Applicability of qSVA to Other Tissues and Brain Regions. We next examined the generalizability of the qSVA framework to other tissues and brain regions. We tested the first step of degradation feature selection in the PBMC dataset (resulting in degradation features (Dataset S6) and the ABRF RNaseA dataset using DLPFC RiboZero-specific degradation features (Dataset S5). In both datasets, the top estimated qSV was strongly associated with the experimental degradation condition (PBMC: $P = 4.56 \times 10^{-13}$, *SI Appendix*, Fig. S12*A*; ABRF: $P = 3.57 \times 10^{-7}$, *SI Appendix*, Fig. S12*B*). In the confounded individual example from

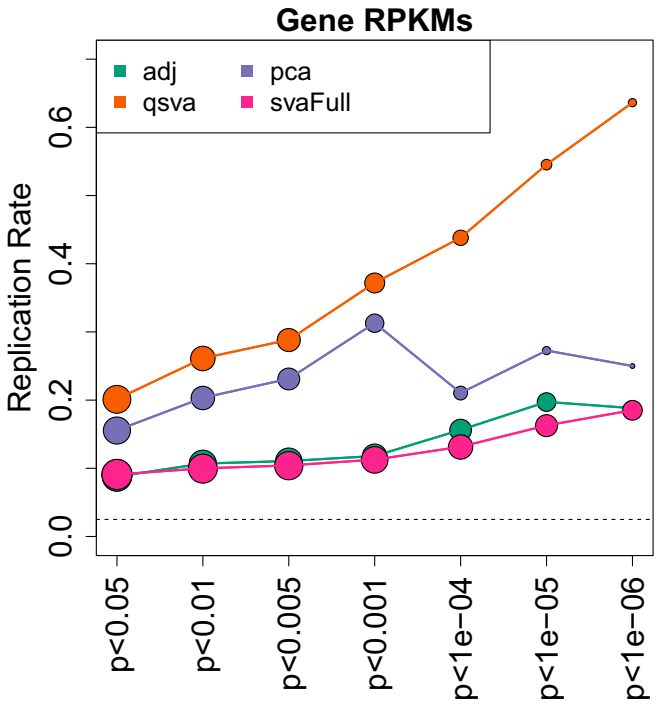

Fig. 2. qSVA improves replication across independent datasets. We modeled SZ-control expression differences using four statistical models in the LIBD (discovery) and CMC (replication) datasets. For a given significance threshold in the discovery dataset, we computed the replication rate (same fold-change direction for case status and $P < 0.05$) in the replication dataset. The qSVA approach had the highest replication rate, and the covariate-adjusted and SVA approaches had the lowest replication rates.

the PBMC dataset, we successfully removed degradation bias selecting degradation-susceptible features from the PBMC degradation data (*SI Appendix*, Fig. S13*A*). Here the qSV adjustment resulted in less statistically biased effect estimates (i.e., $\log_2$ fold changes for the effect of "individual") compared with the statistical model adjusting for observed degradation time (*SI Appendix*, Fig. S13*B*). Conversely, the statistical bias in differential expression signal from the RIN-adjusted model for the effect of individual relative to the degradation time-adjusted model was much larger (*SI Appendix*, Fig. S13*C*). These results suggest this general framework can work well in other tissues.

As the first step in our framework involves generating experimentally derived degradation expression profiles, which may be impractical for small laboratories or projects, we assessed the cross tissue and cell-type applicability of our PBMC- and DLPFC-derived degradation-susceptible features. First, we quantified DLPFC-derived (polyA+, Dataset S4) degradation features in the PBMC dataset; here the top qSV showed similar association with degradation time as above ($P = 7.93 \times 10^{-9}$) and also successfully removed correlation between confounded individual effects and the effect of degradation ($r = 0.015$). The estimated $\log_2$ fold changes for the quality-corrected individual effects were highly correlated using qSVs derived either from PBMC or DLPFC degradation data features ($r = 0.997$, *SI Appendix*, Fig. S13*D*). We next derived qSVs from the PBMC degradation-susceptible transcript features in the LIBD DLPFC schizophrenia-control data and evaluated the performance using DEqual plots and calculating the number of genes significantly differently expressed. Here, although the $\log_2$ fold changes when adjusting using blood versus brain degradation features were correlated (*SI Appendix*, Fig. S14*A*, $r = 0.6$), there was stronger negative correlation between degradation susceptibility in brain- and blood-adjusted case control

differences (*SI Appendix*, Fig. S14*B*, $r = -0.11$). However, using the blood degradation, qSVA yielded 1,057 genes significantly differentially expressed at FDR < 5%, approximately five times more than using the brain degradation-susceptible transcript features, suggesting that brain-specific degradation effects might not be captured using PBMC-susceptible features.

We further used the Genotype-Tissue Expression (GTEx) project RNA-seq expression data—$n = 9{,}502$ across 49 detailed tissues (17)—to characterize the generalizability of DLPFC-derived degradation features to other brain regions and tissue types. We ran differential expression analysis comparing each of 48 detailed tissues in GTEx to BA9 frontal cortex before and after qSVA correction. In the unadjusted analyses, we found a strong association between resulting correlations in DEqual plots and the difference in perceived RNA quality (in chrM mapping rates, *SI Appendix*, Fig. S15*A*, $r = 0.736$, $P = 2.44 \times 10^{-9}$). These quality associations were driven by the 12 other brain regions ($r = 0.88$, $P = 2.44 \times 10^{-9}$) as the nonbrain tissues showed no association ($r = 0.19$, $P = 0.26$). Here qSVA correction removed the overall quality effects across the detailed tissues, largely by removing the positive correlation in the brain samples (*SI Appendix*, Fig. S15*B*, $r = 0.0$, $P = 0.97$). These results suggest that using DLPFC-derived degradation features for qSVA correction may work well in other brain regions, but may not be appropriate for RNA degradation correction in other tissues in the body.

Degradation Bias Signal in Published Differential Expression Analyses. We finally compared the presence of RNA quality bias in published differential expression analyses in human brain for different disorders. As there are currently few additional large RNA-seq studies of postmortem human brain tissue in disease states, we used previously published large microarray datasets on differential expression in autism spectrum disorder (ASD) (18) and Alzheimer's disease (AD) (19) across multiple brain regions. In the ASD dataset, patients had significantly lower RINs than controls in the frontal ($P = 0.021$) but not temporal ($P = 0.70$) cortex, and, in the AD dataset, patients scored significantly lower than controls for the single RIN provided across the three brain regions ($P = 1.23 \times 10^{-10}$). To generate qSVs for these data, we mapped the probes on each microarray platform to the genome, extracted coverage from our RNA-seq data, selected those probe sequences that were significantly associated with degradation (*Materials and Methods*). In the ASD dataset, those probes most associated with degradation ($n = 1{,}129$ at $p_{bonf} < 1\%$) were almost uniformly more lowly expressed in patients compared with controls in the frontal cortex (*SI Appendix*, Fig. S16*A*, $P = 2.2 \times 10^{-49}$). The directionality of enrichment followed the diagnosis and degradation associations, given that almost all degradation-susceptible probes decreased in expression over time (98.5%) and that RINs were lower in patients compared with controls. In the temporal cortex, where RINs did not significantly differ between cases and controls, there was attenuated, but still significant, enrichment in the same negative direction ($P = 4.77 \times 10^{-6}$). Following the qSVA procedure (PCA on the 1,129 susceptible probes and the adjustment for the resulting qSVs), the association between degradation-susceptible probes and diagnosis was removed ($P = 0.496$, *SI Appendix*, Fig. S16*B*).

We found the same enrichment among differentially expressed probes for AD across all three brain regions and the 653 degradation-susceptible probes on this microarray, including in the prefrontal cortex ($P = 1.27 \times 10^{-48}$, *SI Appendix*, Fig. S16*C*), cerebellum ($P = 1.82 \times 10^{-33}$), and visual cortex ($P = 2.35 \times 10^{-35}$). Adjusting for the resulting qSVs again removed the association between diagnosis and degradation susceptibility in the prefrontal cortex ($P = 0.66$, *SI Appendix*, Fig. S16*D*) and cerebellum ($P = 0.49$) and greatly reduced the association in the visual cortex ($P = 6.11 \times 10^{-4}$). The qSVA correction also greatly reduced the magnitude of the differential expression test statistics across the entire platform (*SI Appendix*, Fig. S16 *C* versus *D*). These results further underscore the risk of potentially spurious findings based on uncorrected RNA quality confounding.

Discussion

We describe a framework for quantifying and removing RNA quality biases in differential expression analysis. We first characterized aspects of the landscape of RNA degradation across the human DLPFC and PBMC transcriptomes and identified largely tissue-specific degradation signals. The cell types represented in bulk/mixed tissues like brain and PBMCs further showed differential susceptibility to RNA degradation. We used these experimental degradation datasets to identify the most degradation-susceptible transcript features in PBMC and DLPFC RNA-seq libraries and developed an approach called qSVA to use expression levels of these regions in new/user-provided samples to estimate and remove RNA degradation bias in differential expression analyses. We show that the qSVA approach results in better replication across independent studies and in various public tissue datasets than existing popular statistical models that model observed measures of RNA quality like RIN, chrM mapping rate, and gene assignment rate. Our qSVA approach has a potential advantage over general PCA or RUV adjustments— particularly, less risk of removing true signals along with the noise. Reanalysis of previously published microarray datasets for AD and ASD further suggests that probes differentially expressed for diagnosis were highly associated in a predictable directionality with RNA degradation susceptibility in both datasets.

We also demonstrated that adjusting for measures of RIN largely fails to remove RNA degradation bias and formally showed that RIN correction is more statistically biased at estimating fold changes than qSVA when the true degradation effect is known. The estimation of RIN itself is heavily driven by the intactness of ribosomal RNAs (8), which appears only weakly associated with the underlying quality of total or polyadenylated RNAs across different subjects or tissues. Variance components analysis of RIN values within the full GTEx dataset suggests that tissue source explains approximately three times more variance than individual identity (44.5% versus 14.7%). However, within only the GTEx brain samples, the predictor corresponding to individual explained more variability in RIN than did brain region (28.0% versus 18.7%). Finally, using the LIBD DLPFC dataset, we found no evidence that individual genotype predicted individual RIN; the smallest FDR for a genotype effect on RIN was 0.64 (*SI Appendix*). Indeed, total RNA quality may be more complex than a single number per sample, as the resulting qSVs in both the LIBD and CMC datasets associate with a variety of technical factors (*SI Appendix*, Figs. S7 and S8) that may each influence RNA quality in subtle ways. Therefore, although the RIN value may be a rough guide in determining whether or not to study a particular sample, we would argue that it is not a particularly accurate or useful gauge of RNA quality after data have already been generated.

The applicability of specific tissue-derived degradation-susceptible regions to other tissues or cell types is an important consideration in differential expression analysis, particularly when measured RNA quality associates with the outcome of interest. One practical recommendation for other brain regions would be to use the degradation data from DLPFC and PBMCs to create DEqual plots, quantify the potential RNA degradation bias from its correlation, and then evaluate how the DEqual plot changes when performing qSVA using the DLPFC and PBMC degradation regions. If this qSVA correction fails to remove strong correlation between differential expression effects of degradation and outcome, researchers probably need to generate their own reference degradation datasets and apply the qSVA algorithm.

Differences in latent RNA quality and the underlying cellular composition of homogenate tissue sources (20–22) are two of the strongest confounding factors in postmortem human studies. The qSVA approach here that uses quality-associated features is analogous to our previously proposed approach that uses cell-type–associated features to untangle the confounding effects of cellular composition (sparse PCA) (23). The current study does suggest a potential interaction between RNA quality and cellular composition (*SI Appendix*, Fig. S2 and Table S4), which may be more difficult to statistically isolate the two strong confounding effects, particularly in PBMCs, or when shifting cellular composition is involved in a disease process. Nevertheless, our degradation correction framework can improve the interpretation of differential expression analysis of transcriptomic data.

Materials and Methods

Tissue Degradation Experiment. DLPFC gray matter from five donors was dissected, pulverized, and mixed on dry ice. Approximately 100 mg of pulverized tissue was aliquoted four times for each subject on dry ice followed by tissue aliquots at room temperature except one aliquot of each subject that was kept on dry ice for the time 0 data point. RNA was extracted and sequenced using polyA+ and RiboZero protocols. Data were processed with TopHat (v2.0.13) using the reference transcriptome to initially guide alignment, based on known transcripts in the Illumina iGenomes version of University of California at Santa Cruz knownGene GTF file (using the "–G" argument in the software) (24). Gene counts were generated using the featureCounts tool (25) based on the more recent Ensembl v75, and counts were converted to RPKM values using the total number of aligned reads across the autosomal and sex chromosomes. All public datasets were processed with a similar protocol. All tissues were obtained with informed consent from the legal next of kin (protocol #12–24 approved by the Institutional Review Board of the Department of Health and Mental Hygiene of the State of Maryland).

Degradation Data Analysis. For the samples in each library and tissue type, we separately modeled expression as a function of degradation time, adjusting for the donor and using the limma R Bioconductor package (26). Gene set enrichment analyses were performed on the ordered degradation *T*-statistics from the polyA+ and RiboZero library types among those genes with Entrez Gene IDs using the gseGO and gseKEGG functions in the clusterProfiler R package (27). Cell-type–specific analyses were conducted with CIBERSORT with the default LM22 reference panel and 500 permutations (14) for the PBMC degradation datasets, and DLPFC enrichment was based on 285 cells from adult donors that were previously classified as astrocytes, endothelial cells, microglia, neurons, oligodendrocytes, and oligodendrocyte progenitor cells (15).

LIBD Discovery Dataset Modeling. We used the LIBD DLPFC polyA+ RNA-seq on 155 schizophrenia cases and 196 controls (criteria: ages between 17 and 80, gene assignment rate > 0.5, mapping rate > 0.7, RIN > 6, not outlying on second ancestry PC, only self-reported Caucasians and African Americans) described in Jaffe et al. (28). We fit a series of statistical models at the gene level, modeling $\log_2$-transformed gene-level RPKM (*SI Appendix*) . We used the lmTest and ebayes functions in the limma Bioconductor package (26) to fit all of the statistical models to estimate $\log_2$ fold changes, moderated *T*-statistics, and corresponding *P* values.

CMC Replications Dataset Analysis. We performed differential expression analysis on 159 patients and 172 controls (selecting on total gene assignment rate > 0.3, alignment rate > 0.8, RIN > 6, ages between 18 and 80, non-outlying on genetic ancestry PCs 3 and 5, and keeping only reported Caucasians and African Americans). We similarly fit four of the statistical models at the gene level, modeling $\log_2$-transformed gene-level RPKM (with an offset of 1).

GTEx Analysis. We retained all GTEx samples that had RINs > 5 and belonged to subtissues (SMTSD metadata column) with at least 40 samples, resulting in data on 9,502 samples across 49 detailed tissues. We retained the 36,552 genes that had mean RPKM > 0.2 in at least one subtissue. We modeled differential expression of each of 48 subtissues compared with Brain-Frontal Cortex (BA9) and measured the Pearson correlation present in the resulting DEqual plots, e.g., between the subtissue-specific $\log_2$ fold changes to the DLPFC polyA+ degradation data $\log_2$ fold changes for degradation time.

Microarray Data Processing and Analysis of Published Studies. We extrapolated the expression levels of the probes for each microarray platform in our degradation RNA-seq dataset by aligning microarray probes to the genome and quantifying resulting coverage in the RNA-seq datasets.

See additional details in *SI Appendix, Full Methods and Materials*.

ACKNOWLEDGMENTS. A.E.J. was partially supported by NIH Grant R21MH109956 and J.T.L. was supported by NIH Grant R01GM105705. This work was also supported by the Lieber Institute for Brain Development. Corresponding acknowledgment statements for GTEx and CMC datasets are available in the *SI Appendix*.

1. Irizarry RA, et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4:249–264.
2. Risso D, Ngai J, Speed TP, Dudoit S (2014) Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol* 32:896–902.
3. Leek JT, et al. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet* 11:733–739.
4. Li S, et al. (2014) Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat Biotechnol* 32:888–895.
5. SEQC/MAQC-III Consortium (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat Biotechnol* 32:903–914.
6. 't Hoen PA, et al. (2013) Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat Biotechnol* 31:1015–1022.
7. Adiconis X, et al. (2013) Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nat Methods* 10:623–629.
8. Schroeder A, et al. (2006) The RIN: An RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol Biol* 7:3.
9. Consortium ER (2014) REMC standards and guidelines for RNA-sequencing. Available at www.roadmapepigenomics.org/files/protocols/data/rna-analysis/REMC_RNA-seqStandards_final.pdf. Accessed June 7, 2017.
10. Li S, et al. (2014) Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nat Biotechnol* 32:915–925.
11. Wang L, et al. (2016) Measure transcript integrity using RNA-seq data. *BMC Bioinformatics* 17:58.
12. Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* 3:1724–1735.
13. Gallego Romero I, Pai AA, Tung J, Gilad Y (2014) RNA-seq: Impact of RNA degradation on transcript quantification. *BMC Biol* 12:42.
14. Fromer M, et al. (2016) Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat Neurosci* 19:1442–1453.
15. Darmanis S, et al. (2015) A survey of human brain transcriptome diversity at the single cell level. *Proc Natl Acad Sci USA* 112:7285–7290.
16. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28:882–883.
17. Consortium GT; GTEx Consortium (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* 348:648–660.
18. Voineagu I, et al. (2011) Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* 474:380–384.
19. Zhang B, et al. (2013) Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* 153:707–720.
20. Jaffe AE (2016) Postmortem human brain genomics in neuropsychiatric disorders: How far can we go? *Curr Opin Neurobiol* 36:107–111.
21. Jaffe AE, et al. (2016) Mapping DNA methylation across development, genotype and schizophrenia in the human frontal cortex. *Nat Neurosci* 19:40–47.
22. Jaffe AE, et al. (2015) Developmental regulation of human cortex transcription and its clinical relevance at single base resolution. *Nat Neurosci* 18:154–161.
23. Jaffe AE, Irizarry RA (2014) Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol* 15:R31.
24. Kim D, et al. (2013) TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 14:R36.
25. Liao Y, Smyth GK, Shi W (2014) featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30:923–930.
26. Smyth GK (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3:Article3.
27. Yu G, Wang LG, Han Y, He QY (2012) clusterProfiler: An R package for comparing biological themes among gene clusters. *Omics* 16:284–287.
28. Jaffe AE, et al. (April 5, 2017) Developmental and genetic regulation of the human cortex transcriptome in schizophrenia, doi.org/10.1101/124321.