Gallego Romero et al. BMC Biology 2014, 12:42
http://www.biomedcentral.com/1741-7007/12/42

BMC Biology

RESEARCH ARTICLE Open Access

RNA-seq: impact of RNA degradation on transcript quantification

Irene Gallego Romero[1], Athma A Pai[1,2], Jenny Tung[1,3] and Yoav Gilad[1*]

Abstract

Background: The use of low quality RNA samples in whole-genome gene expression profiling remains controversial. It is unclear if transcript degradation in low quality RNA samples occurs uniformly, in which case the effects of degradation can be corrected via data normalization, or whether different transcripts are degraded at different rates, potentially biasing measurements of expression levels. This concern has rendered the use of low quality RNA samples in whole-genome expression profiling problematic. Yet, low quality samples (for example, samples collected in the course of fieldwork) are at times the sole means of addressing specific questions.

Results: We sought to quantify the impact of variation in RNA quality on estimates of gene expression levels based on RNA-seq data. To do so, we collected expression data from tissue samples that were allowed to decay for varying amounts of time prior to RNA extraction. The RNA samples we collected spanned the entire range of RNA Integrity Number (RIN) values (a metric commonly used to assess RNA quality). We observed widespread effects of RNA quality on measurements of gene expression levels, as well as a slight but significant loss of library complexity in more degraded samples.

Conclusions: While standard normalizations failed to account for the effects of degradation, we found that by explicitly controlling for the effects of RIN using a linear model framework we can correct for the majority of these effects. We conclude that in instances in which RIN and the effect of interest are not associated, this approach can help recover biologically meaningful signals in data from degraded RNA samples.

Keywords: RNA degradation, RIN, degradation, RNA, RNA-seq

Background

Degradation of RNA transcripts by the cellular machinery is a complex and highly regulated process. In live cells and tissues, the abundance of mRNA is tightly regulated, and transcripts are degraded at different rates by various mechanisms [1], partially in relation to their biological function [2-5]. In contrast, the fates of RNA transcripts in dying tissue, and the decay of isolated RNA are not part of normal cellular physiology and, therefore, are less likely to be tightly regulated. It remains largely unclear whether most transcript types decay at similar rates under such conditions or whether rates of RNA decay in dying tissues are associated with transcript-specific properties.

These questions are of great importance for studies that rely on sample collection in the field or in clinical settings (both from human populations as well as from other species), in which tissue samples often cannot immediately be stored in conditions that prevent RNA degradation. In these settings, extracted RNA is often partly degraded and may not faithfully represent *in vivo* gene expression levels. Sample storage in stabilizers like RNA-Later lessens this problem [6] but is not always feasible. Differences in RNA quality and sample handling could, therefore, confound subsequent analyses, especially if samples subjected to different amounts of degradation are naïvely compared against each other. The degree to which this confounder affects estimates of gene expression levels is not well understood.

There is also no consensus on the level of RNA decay that renders a sample unusable or on approaches to control for the effect of *ex vivo* processes in the analysis of gene expression data. Thus, while standardized RNA

* Correspondence: gilad@uchicago.edu
[1]Department of Human Genetics, University of Chicago, 920 E 58th St, CLSC 317, Chicago, IL 60637, USA
Full list of author information is available at the end of the article

quality metrics such as the Degradometer [7] or the RNA Integrity Number (RIN; [8]), provide well-defined empirical methods to assess and compare sample quality, there is no widely accepted criterion for sample inclusion. For example, proposed thresholds for sample inclusion have varied between RIN values as high as 8 [9] and as low as 3.95 [10]. The recent Genotype-Tissue Expression (GTEx) project [11], for instance, reports both the number of total RNA samples they collected as well as the number of RNA samples with RIN scores higher than 6, presumably as a measure of the number of high quality samples in the study.

Broadly speaking, three approaches can be adopted to deal with RNA samples of variable quality. First, RNA samples with evidence of substantial degradation can be excluded from further study; this approach relies on establishing a cut-off value for 'high quality' versus 'low quality' samples and suffers from the current lack of consensus on what this cut-off should be. It also could exclude the possibility of utilizing unique and difficult to collect samples from remote locations or historical collections. Second, if investigators are willing to assume that all transcript types decay at a similar rate, variation in gene expression estimates due to differences in RNA integrity could be accounted for by applying standard normalization procedures. Third, if different transcripts decay at different rates, and if these rates are consistent across samples for a given level of RNA degradation – for example, a given RIN value – a model that explicitly incorporates measured, sample-specific, degradation levels could be applied to gene expression data to correct for the confounding effects of degradation.

To date, most studies apply a combination of the first two approaches: an application of an arbitrary RNA quality cutoff (typically based on RIN score), followed by standard normalization of the data, which assumes that RNA samples at any RIN value higher than the chosen cutoff are not subjected to transcript-specific decay rates. However, current work on the effects of RNA decay has not yet provided clear guidelines with respect to these approaches. In addition, nearly all published work that focuses on RNA stability in tissues following cell death and/or sample isolation predates, or does not employ, high throughput sequencing technologies. These studies broadly suggest that both the quantity and quality of recovered RNA from tissues can be affected by acute pre-mortem stressors, such as pyrexia or prolonged hypoxia [12-14], and by the time to sample preservation and RNA extraction. The quantity and quality of recovered RNA are strongly dependent on the type of tissue studied [15], even when sampling from the same individual [16,17]. These differences in yield across tissues have resulted in a wide range of recommendations for an acceptable *post-mortem* interval for extracting usable, high-quality RNA, ranging from as little as 10 minutes [18] to upwards of 48 hours [19], depending on tissue source and preservation conditions.

Similarly, studies examining changes in the relative abundance of specific transcripts as a result of *ex vivo* RNA decay have reached somewhat contradictory recommendations. Some of this conflict may be attributable to methodological differences. Studies that focused on small numbers of genes assayed through quantitative PCR consistently report little to no effect of variation in RNA quality on gene expression estimates [6,19-22]. Conversely, microarray-based studies have repeatedly reported significant effects of variation of RNA quality on gene expression estimates, even after applying standard normalization approaches. Increasing the time from tissue harvesting to RNA extraction or cryopreservation from 0 to only 40 or 60 minutes, for example, significantly affected expression profiles in roughly 70% of surveyed genes in an experiment on human colon cancer tissues [20]. Likewise, a substantial fraction of genes in peripheral blood mononuclear cells (PBMCs) appears to be sensitive to *ex vivo* incubation [21]. Other microarray-based studies have reached similar conclusions, both in samples from humans [15,16,22,23] and other organisms [24], and have urged caution when analyzing RNA samples with medium or low RIN scores, although the definition of an acceptable RNA quality threshold remains elusive.

To examine the effects of RNA degradation in a setting relevant to field study sample collection, we sequenced RNA extracted from PBMC samples that were stored unprocessed at room temperature for different time periods, up to 84 hours. We collected RNA decay time-course data spanning almost the entire RIN quality scale and examined relative gene-specific degradation rates through RNA sequencing. Due to the high sensitivity and resolution of high-throughput RNA sequencing, our data provide an unprecedentedly detailed picture of the dynamics of RNA degradation in stressed, *ex vivo* cells. Based on our results, we develop specific recommendations for accounting for these effects in gene expression studies.

Results

We extracted RNA from 32 aliquots of PBMC samples from four individuals. The PBMC samples were stored at room temperature for 0 hours, 12 hours, 24 hours, 36 hours, 48 hours, 60 hours, 72 hours and 84 hours prior to RNA extraction. As expected, time to extraction significantly affected the RNA quality ($P < 10^{-11}$), with mean RIN = 9.3 at 0 hours and 3.8 at 84 hours [see Additional file 1: Table S1]. Based on the RIN values we chose to focus on 20 samples from five time points (0 hours, 12 hours, 24 hours, 48 hours and 84 hours) that spanned the entire scale of RNA quality. We generated poly-A-enriched RNA

sequencing libraries from the 20 samples using a standard RNA sequencing library preparation protocol (see [25]). We added a spike-in of non-human control RNA to each sample, which allowed us to confirm the effects of RNA degradation on the RNA sequencing results (see Methods for more details). Following sequencing, we randomly subsampled all libraries to a depth of 12,129,475 reads, the lowest number of reads/library observed in the data. We used BWA 0.6.3 to map reads, calculated reads per kilobase transcript per million (RPKM), and normalized the data using a standard quantile normalization approach (for example, as in [26]). We observed that sample RIN is associated with both the number of uniquely mapped reads (analysis of variance (ANOVA) $P < 10^{-3}$) and the number of reads mapped to genes ($P < 10^{-3}$; Additional file 2: Figure S1), with high RIN samples having greater numbers of both. Furthermore, the proportion of exogenous spike-in reads increases significantly as RIN decreases ($P < 10^{-10}$), as expected given degradation-driven loss of intact human transcripts in poor quality samples. Sequence reads from individual 2 were poorly mapped, especially in the later time-points (see Methods and Additional file 2: Figure S1); we thus excluded the data from all samples from this individual in subsequent analysis.

The effect of RNA degradation on RNAseq output

Principal component analysis of our data demonstrates that much of the variation (28.9%) in gene expression levels in our study is strongly associated with RNA sample RIN scores (Figure 1A; principal component 1 (PC1) associated with RIN scores $P < 10^{-7}$; no other PCs are significantly associated with either sample storage time or RIN score; Additional file 3: Table S2). We also observed a residual presence of inter-individual variation in the data, in spite of variable RNA quality (PCs 4 and 5; Additional file 4: Figure S2 and Additional file 3: Table S2). A correlation matrix based on the gene expression data (Figure 1B) indicates that while samples of relatively high quality RNA cluster by individual, data from RNA samples that experienced high yet similar degradation levels are more correlated than data from samples from the same individual across the time-points. This pattern contrasts with the naïve expectation that gene expression differences between individuals should be the strongest signal in the normalized data. Instead, inter-individual differences only predominate in the early stages of degradation, at the early time-points of 0 hours (mean RIN = 9.3) and 12 hours (mean RIN = 7.9). These observations are robust with respect to the approach used to estimate gene expression levels and – importantly – are not explained by unequal rates of degradation occurring at different distances from the 3′ poly-A tail. For example, we found nearly identical patterns when we estimated expression levels based only on reads that map to the 1,000 bp at the 3′ end of each gene (Additional file 5: Figure S3). Similarly, these effects are robust to the choice of mapping algorithm. Because BWA does not map reads across exon splice junctions, we also remapped our data (excluding individual 2) using TopHat 2.0.8 [27]. As expected, we found a high correlation between RPKM estimates based on

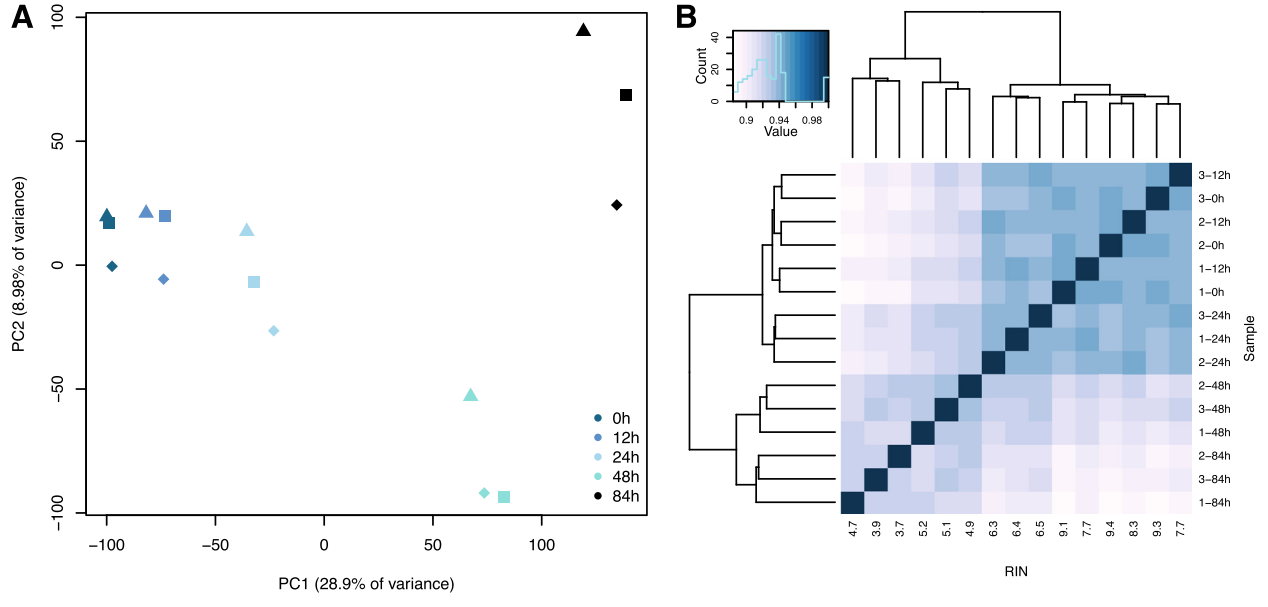

Figure 1 Broad effects of RNA degradation. A) PCA plot of the 15 samples included in the study based on data from 29,156 genes with at least one mapped read in a single individual. Different colors identify different time-points, while each shape indicates a particular individual in the data set. **B)** Spearman correlation plot of the 15 samples in the study. PCA, principal component analysis.

alignments with both approaches (Spearman's $\rho = 0.82$ when we consider all genes with at least one observation of RPKM > =0.3 in the entire data set; Spearman's $\rho = 0.85$ when we only consider genes with at least one observation of RPKM > =0.3 using data mapped with BWA, Additional file 6: Figure S4 and Additional file 7: Figure S5). Finally, we found that the global effects of RNA degradation on estimated gene expression levels could not be eliminated by globally regressing out RIN scores [see Additional file 8: Figure S6].

The possibility of reduced sequencing library complexity is often cited as a reason to exclude RNA samples of low quality. This concern is primarily based on the observation that sequencing RNA samples of lower RNA quality results in relatively decreased proportions of mappable reads, an observation corroborated in our study [see Additional file 2: Figure S1]. Yet, it is unclear to what extent this property affects the ability to estimate gene expression levels in RNA samples of low quality. To assess the effects of RIN on sample complexity, we plotted the distribution of RPKM values within individuals at different time points. Our data indicate that mean RPKM increases as sample RIN decreases ($P < 10^{-5}$ Additional file 9: Figure S7). This seems counterintuitive, but can be explained by the presence of a few highly expressed genes in the samples of low RNA quality. Indeed, relative to 0 hours, low RIN samples at 48 hours and 84 hours have an excess of low RPKM genes and a deficit of high RPKM genes, shifting the median RPKM downwards ($P < 10^{-4}$; Figure 2). We further found a positive association between the number of genes with at least one observation of RPKM ≥0.3 and RIN ($P < 10^{-4}$). Even when we subsampled all samples to the same number of sequencing reads, we still observed a high proportion of genes with low RPKM values in RNA samples of lower quality ($P < 10^{-4}$; Additional file 10: Figure S8). This suggests that a non-uniform effect of RNA degradation on gene expression levels results in somewhat lower complexity of the sequencing library (Figure 2, Additional file 10: Figure S8). On the other hand, both within a single individual and across the whole dataset, we found that

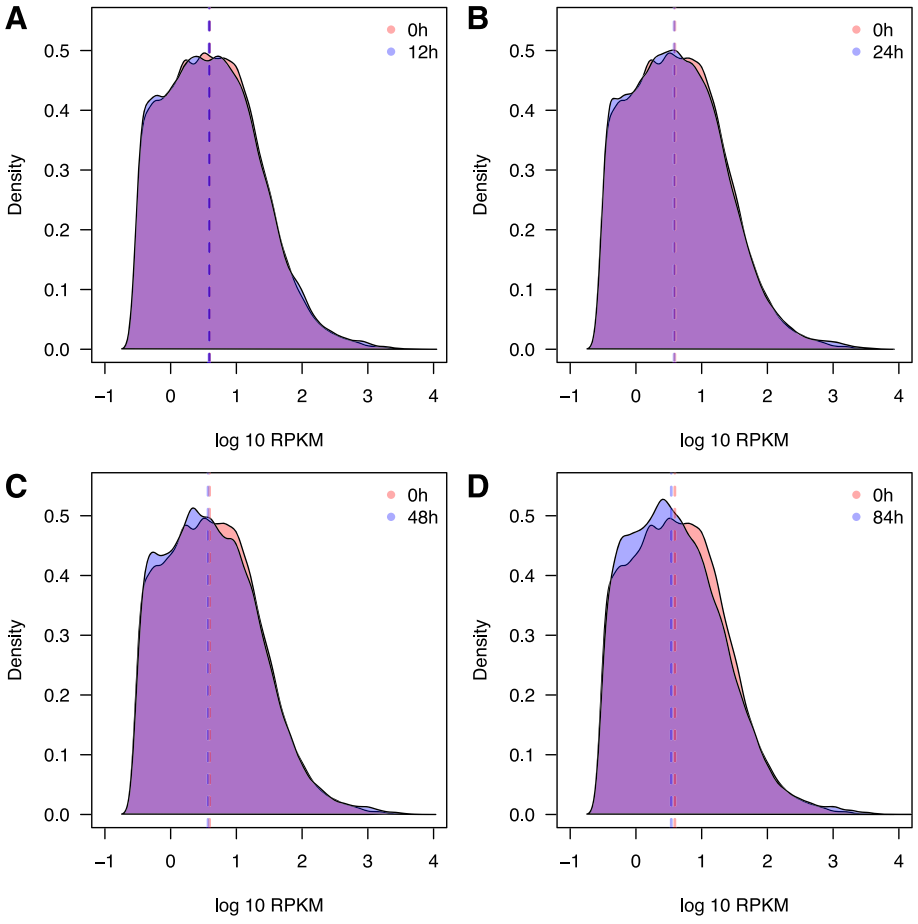

Figure 2 Changes in library complexity over time. Dashed lines indicate median RPKM at each time-point. **A)** Density plots of RPKM values among all three individuals at 0 hours and 12 hours. **B)** as A, but 0 hours and 24 hours. **C)** as A, but 0 hours and 48 hours. **D)** as A, but 0 hours and 84 hours. RPKM, reads per kilobase transcript per million.

nearly all genes whose expression could be measured at 0 hours are also detected as expressed throughout the entire time-course experiment. Only a few genes (Table 1) present in all individuals up until a given time point were completely absent from the data at later time points.

Different transcripts are degraded at different rates

We sought to understand better the nature of transcript degradation in the RNA samples of lower quality. Given our time course study design, we were able to estimate degradation rates for all genes detected as expressed at all five time-points. To do so, we fit a log-normal transform of a simple exponential decay function (see Methods) to quantile-normalized RPKM values for each gene that was detected as expressed in all individuals at all time-points. We considered the slope of this function, *k*, to be a proxy for the decay rate of the gene. We then compared this slope to the mean transcript degradation rate across all genes, which, as a result of our quantile normalization approach, is equal to 0 (thus, a value of 0 indicates no change in the relative rank of that transcript's expression level across time points). If all genes decay at the same rate, then no slopes should significantly differ from the mean value. However, at a false discovery rate (FDR) threshold of 1%, we found that 7,267 of the 11,923 genes tested (60.95%; see Methods) were associated with degradation rates that were significantly different from the mean (Figure 3; Additional file 11: Table S3). Of these genes, 3,522 had a negative slope (that is, they were degraded significantly faster than the mean degradation rate) and 3,745 had a positive slope (that is, these transcripts were degraded significantly slower than the mean degradation rate).

Although we might expect RNA degradation in decaying cells to be a random process, gene ontology (GO) analysis identified 118 and 293 significantly overrepresented categories among slowly and rapidly degraded genes, respectively (FDR = 5%; Additional file 12: Tables S4 and Additional file 13: Table S5). We present the functional enrichment results only as an indication that the rate of transcript decay is not random. These observations are robust to different normalization approaches, to the inclusion of RIN as a covariate in our linear model, and to fitting slopes using RIN instead of time-points. Limiting our analyses to the 1,000 bp closest to the 3′ end of transcripts also yields similar results.

We asked what properties, beyond GO functional categories, might be associated with the observed variation in transcript degradation rates. We found that the coding DNA sequence (CDS) length ($P < 10^{-12}$), %GC content ($P < 10^{-4}$), and 3′UTR length ($P < 10^{-15}$) are all significantly correlated with estimated transcript degradation rate (Figure 4A-C), with higher %GC content and increased length of both the 3′ UTR and CDS all associated with faster degradation. However, we found that total transcript length (5′ UTR + CDS + 3′ UTR) is not significantly correlated with degradation rates; instead, targets of both fast and slow degradation have longer transcripts than those that are degraded at an average rate (Figure 4D). The correlation between %GC content and CDS length is high ($\rho = -0.19$, $P < 10^{-16}$), but even when we control for the effects of either variable, the individual effects remain significant predictors of degradation rates ($P < 10^{-7}$). Our data thus suggest that both CDS length and %GC content affect degradation rate, and that observed degradation rates result from complex interactions between multiple forces. We again present these results as evidence for the non-random nature of the transcript degradation rate (yet, we do not presume at this time to offer mechanistic explanations for these correlations).

We also sought to investigate whether targets of slow, fast, or average degradation differ meaningfully in terms of broad biological categories. As expected given our poly-A enrichment strategy, most transcripts in our data originate from intact protein-coding genes, but we also observed four other biotypes represented by more than 100 distinct transcripts. The distribution of these biotypes across rapidly and slowly degraded transcripts is not random, with a significant enrichment of pseudogenes among transcripts that degrade slowly ($P = 0.015$), and an enrichment of intact protein-coding genes among the rapidly degraded transcripts ($P < 10^{-16}$, Figure 4E).

Controlling for the effect of RNA degradation in analyses of differential expression

Ultimately, the goal of most RNA sequencing studies is to estimate variation in gene expression levels or to identify genes that are differentially expressed between conditions, individuals or states. We thus considered the effects of RNA quality on measures of relative gene expression levels between time-points and on overall estimates of inter-individual variation in gene expression.

Table 1 Genes observed in all individuals until or after a particular time point

Seen until	0 hours	12 hours	24 hours	48 hours	84 hours
#genes	14	9	72	52	11,923
Mean RPKM when seen	0.68	0.679	1.29	1.09	32.689
Unseen before	**0 hours**	**12 hours**	**24 hours**	**48 hours**	**84 hours**
#genes	n/a	4	2	19	35
Mean RPKM when seen	n/a	1.078	2.212	2.769	3.034

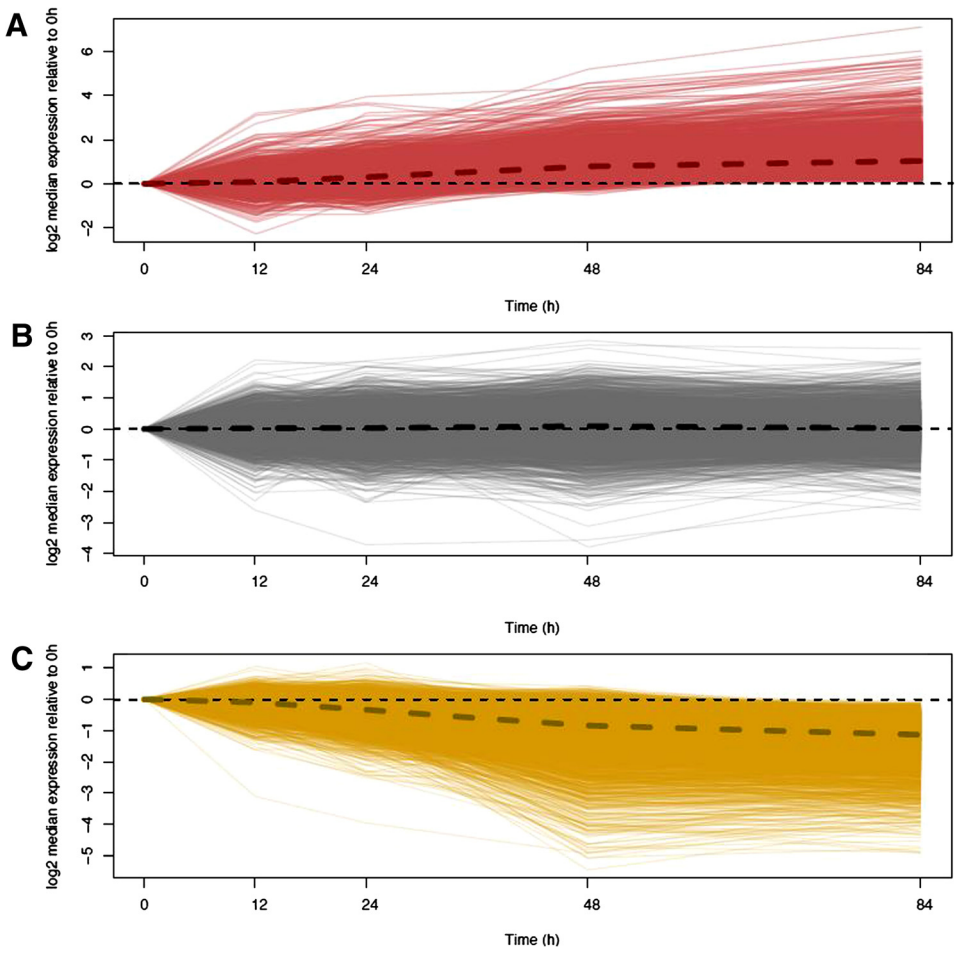

Figure 3 Log_{10} median abundance of genes across all three individuals relative to 0 hours. Plots are separated by slope. **A)** Transcripts with significantly slow rates of degradation relative to the mean rate (identified at 1% FDR, n = 3,745). **B)** Transcripts that are degraded at a rate close to the mean cellular rate (n = 4,656). **C)** Transcripts with significantly fast rates of degradation relative to the mean rate (identified at 1% FDR, n = 3,522). In all plots, the thick dashed line indicates the median degradation rate for all genes in that group, whereas the thin dashed line denotes no change in degradation rate relative to 0 hours. FDR, false discovery rate.

As a first step we analyzed the normalized expression data using a generalized linear model (GLM) approach (see Methods) to classify genes as differentially expressed between 0 hours and any other time-point. We only considered genes with at least one mapped read in all individuals at all time-points (n = 14,094). At an FDR of 5%, we identified 608 (4%) genes as differentially expressed by 12 hours. Both the number of differentially expressed genes and the magnitude of expression changes increase drastically along the time-course experiment (Table 2). By 84 hours, 9,998 genes (71%) are differentially expressed (FDR = 5%). Roughly half of these genes appear to be more highly expressed in the later time-points than at 0 hours. This may seem counterintuitive given that the change in expression is most likely the result of RNA degradation, yet this apparent increase in expression is due to our normalization approach (all transcripts in our experiment experience some level of degradation throughout the time course). Post normalization of the data, an apparent elevated expression level in the later time points, therefore, indicates slow degradation relative to the genome-wide mean rate of RNA decay.

As expected, when we include RIN as a covariate in the model the number of differentially expressed genes across time-points is drastically reduced (fewer than 50 genes are classified as differentially expressed between 0 hours and any other time-point; Table 2). These observations confirm that RIN is a robust indicator of degradation levels. Without accounting for RIN, the effect of variation in RNA quality on our data is overwhelming. To understand these effects better, we explored whether accounting for variation in RIN increased the power to detect other sources of (biologically relevant) variation in RNA-seq data, such as the variation in gene expression between individuals. We also investigated several alternative approaches for controlling for variation in RNA quality.

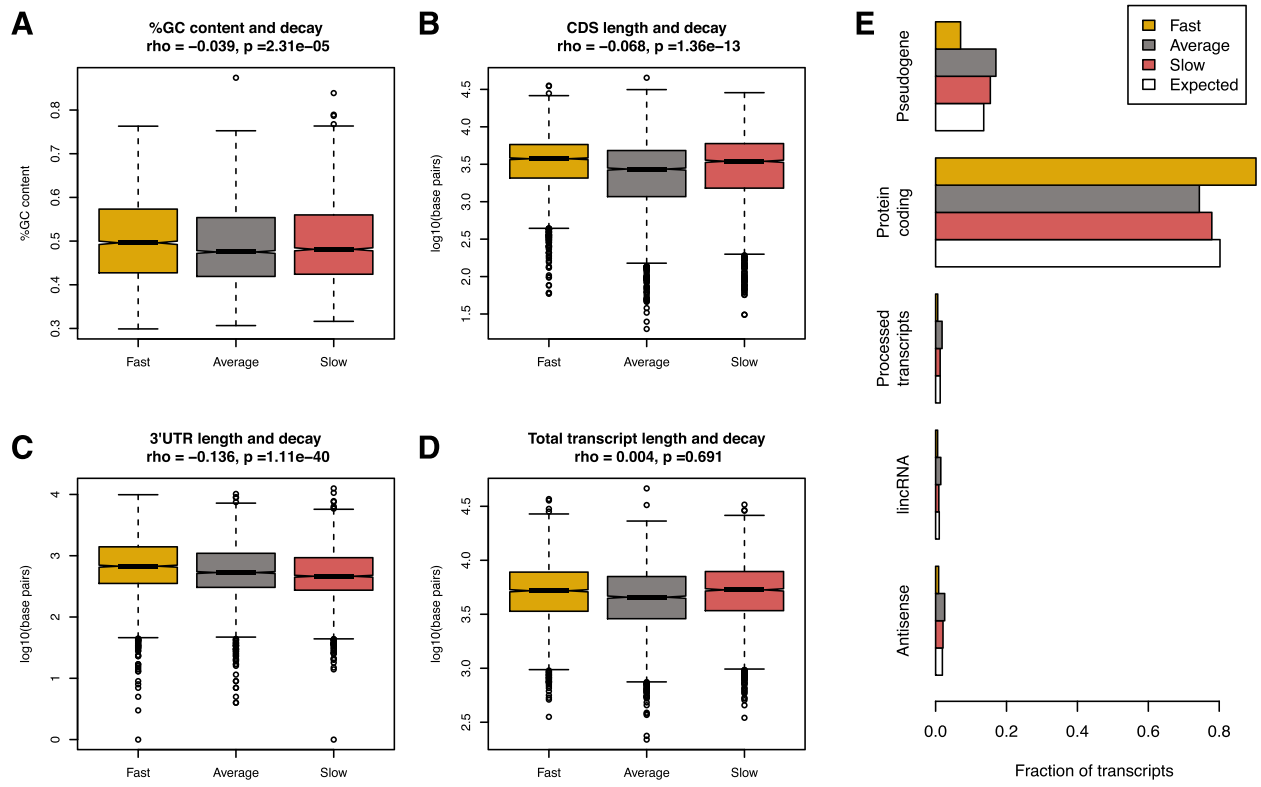

Figure 4 Characteristics of rapidly and slowly degraded transcripts. In all plots, rapidly degraded transcripts are plotted in gold, transcripts degraded at an average rate are plotted in grey and slowly degraded transcripts are in red. **A)** By transcript %GC content. **B)** By coding region length. **C)** By 3'UTR length. **D)** By complete transcript length. **E)** By ENSEMBL biotype.

Without accounting for RIN, we classified few genes (48 to 100; Table 2) as differentially expressed between pairs of individuals. This property of the data is captured by a heat map of sample pairwise correlation calculated using only the top 10% (1,410) most variable genes across individuals at 0 hours. As can be seen in Figure 5A, while at the early time-points inter-individual differences are the predominant source of variation in the data, degradation overwhelms these differences in the low quality (low RIN) RNA samples from 48 hours and 84 hours. Hence, inclusion of these time points in our GLM, which considers samples from the same individual but different time points as 'technical replicates', obscures much of the true signal of inter-individual variability.

To recover this signal, we tested two approaches for explicitly accounting for RIN when estimating differential gene expression across individuals: (1) incorporating RIN as a covariate in our GLM; and (2) analyzing the residuals of gene expression levels after first regressing out

Table 2 Number of identified DE genes

	GLM: reads approximate time point		
Time point	**GLM**	**GLM + RIN**	**Regress RIN, GLM**
0 h versus 12 h	608	5	26
0 h versus 24 h	3,704	5	203
0 h versus 48 h	8,756	47	5
0 h versus 84 h	9,998	42	0
	GLM: reads approximate individual b		
Individuals	**GLM**	**GLM + RIN**	**Regress RIN, GLM**
Ind 1 versus Ind 3	69	553	268
Ind 1 versus Ind 4	48	401	190
Ind 3 versus Ind 4	100	573	299

b, Treating time as technical replicates; DE, Differentially expressed; GLM, generalized linear model; H, hours; RIN, RNA integrity number.

RIN from the normalized gene expression data (Table 2). Both approaches result in the identification of many more genes as differentially expressed between individuals (401 to 573 when incorporating RIN directly into our GLM, 190 to 299 when testing for differential expression using residuals; Table 2). We also repeated the pairwise correlation analysis using the same 1,410 most variable genes identified above, but this time we used the residuals after regressing the effect of RIN from the data. The residuals cluster well by individual throughout the entire time course experiment, regardless of RNA quality (Figure 5B).

Finally, we examined the overlap between the subset of the 10% of most variable genes across individuals at 0 hours (the 1,410 genes used to generate Figure 5) and those identified as differentially expressed across individuals as described above (Table 3). Of the two approaches we employed to account for the effect of RIN, testing for differential expression after removing the effects of RIN on the data (method 2) yielded higher concordance between DE genes and those with high inter-individual variance at 0 hours, suggesting it may be a better approach than simply including RIN as a covariate.

Discussion

Our observations indicate that the effects of RNA degradation following death or tissue isolation are pervasive and can rapidly obscure inter-individual differences in gene expression. Yet, we also found that by using RNA-seq nearly all genes observed at our first time-point could still be detected even in severely degraded RNA samples, although the estimated relative expression levels were drastically affected by degradation. Although post-mortem RNA degradation is considered a non-regulated process, some of the traditional predictors of regulated RNA decay rates in the cell are also associated with variation in RNA quality in our data. For example, longer protein coding regions and 3′ UTRs are correlated with more rapid degradation, similar to previously reported trends [5,28,29]. Total transcript length, however, which is a significant predictor of regulated RNA decay in the cell, is not associated with variation in degradation rates in our data.

The effect of RNA degradation can be accounted for

We confirmed previous observations of decreasing data quality as time from tissue extraction to RNA isolation increased [see Additional file 2: Figure S1], both with respect to the number of high quality reads we were able to generate from our sequencing libraries and library complexity. While increased time to RNA extraction did not generally result in the complete loss of transcripts (less than 8% of transcripts are lost), the relative expression levels of many transcripts were drastically altered over the time-course experiment, with 61% of genes classified as differentially expressed between 0 hours (mean RIN of 9.3) and 84 hours (mean RIN of 3.78). This proportion of differentially expressed genes is in line with previous reports of the effects of warm ischemia on human gene expression in tumor biopsies, as assessed using microarrays [20,22]. The potential of RNA degradation to skew measurements of

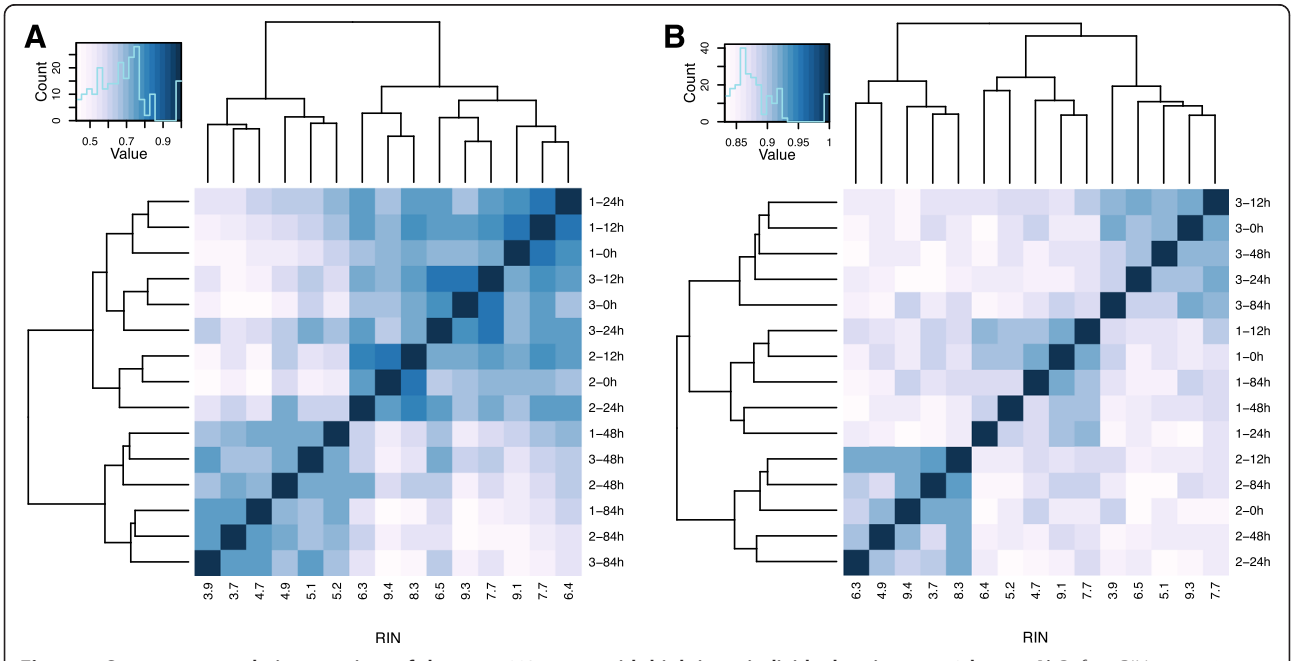

Figure 5 Spearman correlation matrices of the top 10% genes with high inter-individual variance at 0 hours. A) Before RIN correction. **B)** After regressing the effects of RIN. RIN, RNA integrity number.

Table 3 DE genes across pairs of individuals and overlap with top 10% most variable genes at 0 hours

	GLM individual		GLM, individual + RIN		Regress RIN, GLM individual	
Test	Number DE genes	% overlap	Number DE genes	% overlap	Number DE genes	% overlap
Ind 1 vs Ind 3	69	86.96%	553	45.39%	268	75.00%
Ind 1 vs Ind 4	48	89.58%	401	50.12%	190	78.95%
Ind 3 vs Ind 4	100	87.00%	573	49.21%	299	73.91%
All individuals	160	85.00%	1053	42.64%	521	71.98%

DE, Differentially expressed; GLM, generalized linear model; RIN, RNA integrity number.

gene expression levels and obscure biologically meaningful signals is, therefore, apparent. If there are systematic differences in RNA quality between two classes of samples being compared, we predict that the effect of RNA quality on relative estimates of gene expression levels would be responsible for much of the signal in the data. Furthermore, as degradation rate is to some degree associated with biological function [see Additional file 11: Tables S3 and Additional file 12: Table S4], it has the potential to confound naïve comparisons of functional annotations as well.

However, the marked effects of RNA degradation on the relative expression level of most genes can be corrected, to a large degree, using relatively simple statistical methods. Indeed, we found that the inclusion of RIN in our model was sufficient to account for much of the effect of degradation and allowed us to identify a reasonable number of differentially expressed genes between pairs of individuals in our data. These were not spurious signals generated by our approach; they recapitulated observations made at 0 hours (when RNA quality was excellent), but were originally dwarfed by the magnitude of degradation-driven expression changes in the uncorrected data. A similar approach – taking into account variation in RIN – has been previously proposed for the analysis of RTq-PCR data abundance [30]. Nevertheless, our observations suggest that some of the effects of transcriptional degradation in *ex vivo* samples cannot be corrected. Library complexity decreases somewhat with lower RNA quality, and some genes (approximately 5%) can no longer be detected at the later time-points. Based on our data we conclude that these effects cannot be corrected by simply sequencing more degraded libraries to a greater depth.

In a study similar to our own, Opitz *et al.* [31] subjected extracted RNA samples from three advanced human rectal cancer biopsies to degradation through increasingly longer incubation at 60˚C and then considered the evidence of time-point/RIN–driven degradation using microarray data. The RIN values spanned by their data mirror values in ours, but the results do not. In contrast to the large RIN-associated effects we observed, Opitz *et al.* reported that of 41,000 tested probe-level 2data points only 275 demonstrated significant degradation effects, with inter-individual differences being the predominant signal in the data. Assuming that differences in the platforms used (microarrays and RNAseq) are not the reason for this discrepancy, one possible explanation for this stark difference between the studies is that lower RIN scores as a result of degradation of extracted RNA samples (Opitz *et al.)* may reflect substantially different properties than lower RIN scores that are the result of degradation of RNA in decaying cells (our study). Based on the observations of Opitz *et al.* we hypothesize that degradation rates of isolated RNA may be mostly linear and uniform; thus, the degradation effects can be accounted for by employing standard normalization approaches. In contrast, degradation rates of RNA in a dying tissue sample, a situation that mirrors more closely conditions likely to be faced by investigators in clinical or field settings, is not uniform across transcripts. Because these differences cannot be neglected in downstream analyses, knowledge of the context in which degradation occurs is, therefore, crucial.

Our observations suggest that actively mediated degradation of transcripts may occur during necrosis; namely, degradation of RNA in a dying tissue may not be a completely random process. Biologically mediated degradation, whether actively driven by the cell's decay machinery [1], or simply the consequence of the leakage of RNases into cells as membranes are disrupted, is different from the heat-driven degradation of naked RNA, which in turn is likely to be different from the degradation caused by continued freeze-thaw cycles [32]. It is likely that in a dying tissue, most degradation is initially biologically mediated and directed towards specific classes of transcripts, but as the cellular environment continues to deteriorate, the relative importance of stochastic degradation may increase such that at later time-points degradation becomes increasingly uncoupled from biological function.

Additionally, the increased resolution of RNA sequencing relative to other platforms used to assay gene expression levels [25] is both a hindrance and a boon in this situation, allowing for detection of subtler differences than ever before, but also warranting greater caution when analyzing samples of differing quality.

Recommendation regarding the inclusion of RNA samples in a study

Previous studies [8-10,23,32-34] have sought to provide an RNA degradation threshold below which in-depth analysis of RNA is not recommended. However, these studies have reached conflicting conclusions. Our data suggest that if a simple cut-off value is to be used, a conservative cut-off in the context of RNA degradation in dying tissue samples lies between 7.9 and 6.4, the mean RIN scores associated with 12 hours and 24 hours in our time course experiment, respectively. We observed few differences in measurements of gene expression between 0 hours and 12 hours, as evidenced by the low number of genes identified as differentially expressed between the two time-points. Thus, it may be tempting to conclude that so long as all samples in any particular study have roughly similar RINs explicit correction is not necessary. However, when we test for differential expression between other close time-points we identify 3,020 genes as differentially expressed between 48 hours and 84 hours (difference in mean RIN = 1.3), and 5,293 between 24 hours and 48 hours (difference in mean RIN = 1). It is clear that measurements of gene expression are extremely sensitive to starting sample quality.

Conclusions

Our observations indicate that useful data can be collected using RNA sequencing even from highly degraded samples. As long as RIN scores are not associated with the effect of interest in the study (namely, different classes of samples in the study are not associated with different distributions of RIN scores), accounting for RIN scores explicitly can be an effective approach. In our study, we were able to identify differently expressed genes between individuals even when RNA samples with RIN scores around 4 were included. Excluding the samples with RIN values lower than 6.4 in our study would have resulted in a less powerful design than including these samples and globally correcting for RIN values. Given these results, we believe that under most circumstances, the most effective approach may be to include all samples regardless of quality, and explicitly model a measure of RNA quality in the analysis.

Methods

RNA degradation

We obtained Buffy coat samples from four adult Caucasian males from Research Blood Components LLC (Boston, MA, USA) and separated PBMCs through a standard Ficoll gradient purification. Each sample was split into aliquots of four million live cells and resuspended in 200 uL of PBS. Cells were kept at room temperature and aliquots from each sample lysed every twelve hours by addition of 700 uL of RLT buffer (Qiagen, Valencia, CA, USA) with beta-mercaptoethanol (Sigma-Aldrich, St Louis, MO, USA) added at 10 uL BME/1 mL RLT according to the manufacturer's instructions. Lysed cells were immediately frozen and not thawed until RNA extraction.

Extraction and sequencing

RNA was extracted using the Qiagen RNeasy kit. Extracted RNA quality was assessed with a BioAnalyzer (Agilent Technologies, Wilmington, DE, USA). From these results we selected five time-points – 0 hours, 12 hours, 24 hours, 48 hours and 84 hours – that encompassed a large stretch of the degradation spectrum. We then prepared poly-A-enriched RNA sequencing libraries for all 20 individual/time-point combinations according to a previously published protocol [25], using 1.5 μg of total RNA per library in all instances. In all instances, we added 15 ng (1%) of an exogenous RNA spike-in during library preparation, composed of equal parts *Caenorhabditis elegans*, *Drosophila melanogaster* and *Danio rerio* total RNA. Samples were multiplexed and sequenced on four lanes (two per library preparation strategy) of an Illumina HiSeq2000 using standard protocols and reagents. Reads were 50 bp in length. All generated reads have been deposited into the Sequence Read Archive (SRA) under accession numbers SAMN02769865-SAMN02769884.

Data mapping and normalization

Data were combined across lanes and data for all libraries were randomly subsampled to the lowest observed number of reads, 12,129,475. Reads were independently mapped to the human (hg19), *D. rerio* (danRer7), *D. melanogaster* (dm3), and *C. elegans* (ce10) genomes using BWA 0.6.2 [35]. All reference genomes were obtained from the UCSC Genome Browser [36]. Only reads that mapped exclusively to a single site in the human genome with one or zero mismatches were retained for downstream analyses. Following mapping, we removed all reads that mapped to more than one genome. At this point we also discarded one individual – individual number 2 - due to low mappability and read quality in the later time points [see Additional file 2: Figure S1]. We also mapped all reads using TopHat 2.0.8 and the same quality thresholds and filtering steps.

We calculated RPKM [37] for all ENSEMBL v71 [38] human genes in our data. Genes with multiple transcripts were collapsed into a single transcript containing all exons of the gene; where multiple exons of different size overlapped the same genomic region, the entire region was kept. We discarded all exonic regions transcribed as part of more than one gene. Additionally, we quantile-normalized both RPKM and read count-level data across individuals using the lumiN function in the Bioconductor [39] package *lumi* [40], which controls for, and dampens, technical sampling variance in highly expressed genes. Read counts were log2 transformed prior to quantile normalization to

generate a normal distribution; analyses were carried out on subsequently untransformed counts.

All statistical analyses were carried out using R 2.15.2.

Calculation of decay rates

We estimated the decay rate of the 11,923 genes with an RPKM >0.3 in all individuals at all time-points by fitting a first order log-normal transform of the classical first-order decay equation:

$$\ln(y(t)) = B_0 - kt + \varepsilon$$

where *y(t)* is the mRNA abundance of a given gene at time *t* (in quantile-normalized RPKM), B_0 is the abundance at the initial time-point, and *k* the decay rate, with the variance term ε being normally distributed. Data from all three individuals were considered simultaneously; that is, we obtained a single decay constant for each gene across all three individuals. To control for the high FDR of expressed genes at low expression levels, all RPKM observations <0.3 were discarded for all subsequent analyses, as in [41].

Length and per-transcript %GC content were calculated using BEDTools (version 2.16.2 [42]), using the same gene models described above. Biotype as well as 5′ and 3′ UTR length were retrieved from ENSEMBL v71. In those instances where there are multiple UTRs associated with the same gene, we used the median UTR length for each gene in all calculations.

Differential expression and gene enrichment

Differentially expressed genes were identified using the R package *edgeR* [43], utilizing a GLM framework with time, individual ID and sample RIN as covariates, as described above. Only those genes with a minimum observation of one mapped read across all individuals at all time-points were included. Instead of quantile normalization as described above, all data were normalized using trimmed mean of M values normalization (TMM, [44]), which corrects for the observed differences in informative reads between sequencing libraries. Inter-individual variance estimates were generated after variance stabilization of read counts using the *predFC* function in edgeR.

Downstream gene enrichment analyses were carried out using the R package *topGO* [45], using the 'classic' algorithm and a minimum node size of five. All significance values given in the text have been corrected to an FDR of 5% or 1%, using the qvalue method of [46]. In all cases, the background data set included all 14,094 genes with complete observations.

Additional files

Additional file 1: Table S1. Relationship between RIN and time to RNA extraction.

Additional file 2: Figure S1. Fraction of reads mapped from generated libraries. All samples were randomly subset to the same depth prior to mapping.

Additional file 3: Table S2. Correlations between PCs and covariates.

Additional file 4: Figure S2. PCA plot of principal components 4 and 5, the only components significantly associated with inter-individual variation in the data. Different colors identify different time-points, while each shape indicates a particular individual in the data set.

Additional file 5: Figure S3. A) PCA plot of the 15 samples included in the study based on data from 27,856 genes with at least one mapped read to the 1,000-most 3′ base pairs in a single individual. Different colors identify different time-points, while each shape indicates a particular individual in the data set. **B)** Spearman correlation plot of the 15 samples in the study, using only data trimmed to the 1,000-most 3′ bp.

Additional file 6: Figure S4. Density plot of RPKM estimates per gene after mapping with BWA and TopHat. Only genes with an RPKM > =0.3 after mapping with BWA are shown.

Additional file 7: Figure S5. Spearman correlation plot as in Figure 1 using data mapped by TopHat. **A)** Correlations across 33,438 genes with at least one instance of one read mapped by TopHat. **B)** Correlations across 29,156 genes with at least one instance of one read mapped by BWA.

Additional file 8: Figure S6. A) PCA plot of the 15 samples included in the study based on data from 29,156 genes with at least one mapped read in a single individual, after correcting for the effects of RIN on the data. Different colors identify different time-points, while each shape indicates a particular individual in the data set. **B)** Spearman correlation plot of the 15 samples in the study, after correcting for the effects of RIN on the data.

Additional file 9: Figure S7. Mean RPKM as a function of time (h) to sample collection.

Additional file 10: Figure S8. Effects of sequencing depth on library complexity. Dashed red lines indicate median RPKM in each subset. **(A** to **D)** Density plots of RPKM values in the 0-hour data when subsampled to indicated depths. For comparison, the observed distribution of RPKM values in the 84-hour data is plotted in each figure in blue.

Additional file 11: Table S3. Estimated decay rates for 11,923 tested genes.

Additional file 12: Table S4. Significantly overrepresented GO terms among slowly degraded genes.

Additional file 13: Table S5. Significantly overrepresented GO terms among rapidly degraded genes.

Competing interests

The authors declare they have no competing interests.

Authors' contributions

IGR analyzed the data and wrote the manuscript with input from all authors. AAP conceived and designed the study, and performed experiments. JT performed experiments. YG conceived the study and helped write the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Rachel A. Myers as well as members of the Gilad lab for helpful comments and discussions. This work was supported by NIH grant HG006123 to YG. IGR was supported by a Sir Henry Wellcome Postdoctoral Fellowship. AAP was supported by an American Heart Foundation Predoctoral Fellowship. JT was supported by a Chicago Fellows Postdoctoral Fellowship.

Author details
[1]Department of Human Genetics, University of Chicago, 920 E 58th St, CLSC 317, Chicago, IL 60637, USA. [2]Present address: Department of Biology, Massachusetts Institute of Technology, 31 Ames Street, 68-271A, Cambridge, MA 02139-4307, USA. [3]Present address: Department of Evolutionary Anthropology and Duke Population Research Institute, Duke University, 125 Science Drive, Durham, NC 27708, USA.

Received: 23 May 2014 Accepted: 27 May 2014
Published: 30 May 2014

References

1. Garneau NL, Wilusz J, Wilusz CJ: **The highways and byways of mRNA decay.** *Nat Rev Mol Cell Biol* 2007, **8:**113–126.
2. Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, Chen W, Selbach M: **Global quantification of mammalian gene expression control.** *Nature* 2011, **473:**337–342.
3. Rabani M, Levin JZ, Fan L, Adiconis X, Raychowdhury R, Garber M, Gnirke A, Nusbaum C, Hacohen N, Friedman N, Amit I, Regev A: **Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells.** *Nat Biotechnol* 2011, **29:**436–442.
4. Tani H, Mizutani R, Salam KA, Tano K, Ijiri K, Wakamatsu A, Isogai T, Suzuki Y, Akimitsu N: **Genome-wide determination of RNA stability reveals hundreds of short-lived non-coding transcripts in mammals.** *Genome Res* 2012, **22:**947–956.
5. Yang E, van Nimwegen E, Zavolan M, Rajewsky N, Schroeder M, Magnasco M, Darnell JE Jr: **Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes.** *Genome Res* 2003, **13:**1863–1872.
6. Micke P, Ohshima M, Tahmasebpoor S, Ren ZP, Ostman A, Pontén F, Botling J: **Biobanking of fresh frozen tissue: RNA is stable in nonfixed surgical specimens.** *Lab Invest* 2006, **86:**202–211.
7. Auer H, Liyanarachchi S, Newsom D, Klisovic MI, Marcucci G, Kornacker K: **Chipping away at the chip bias: RNA degradation in microarray analysis.** *Nat Genet* 2003, **35:**292–293.
8. Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, Lightfoot S, Menzel W, Granzow M, Ragg T: **The RIN: an RNA integrity number for assigning integrity values to RNA measurements.** *BMC Mol Biol* 2006, **7:**3.
9. Imbeaud S, Graudens E, Boulanger V, Barlet X, Zaborski P, Eveno E, Mueller O, Schroeder A, Auffray C: **Towards standardization of RNA quality assessment using user-independent classifiers of microcapillary electrophoresis traces.** *Nucleic Acids Res* 2005, **33:**e56.
10. Weis S, Llenos IC, Dulay JR, Elashoff M, Martinez-Murillo F, Miller CL: **Quality control for microarray analysis of human brain samples: the impact of postmortem factors, RNA characteristics, and histopathology.** *J Neurosci Methods* 2007, **165:**198–209.
11. **Genotype-Tissue Expression Porta.** [http://www.gtexportal.org/home/]
12. Durrenberger PF, Fernando S, Kashefi SN, Ferrer I, Hauw JJ, Seilhean D, Smith C, Walker R, Al-Sarraj S, Troakes C, Palkovits M, Kasztner M, Huitinga I, Arzberger T, Dexter DT, Kretzschmar H, Reynolds R: **Effects of antemortem and postmortem variables on human brain mRNA quality: a BrainNet Europe study.** *J Neuropathol Exp Neurol* 2010, **69:**70–81.
13. Harrison PJ, Heath PR, Eastwood SL, Burnet PW, McDonald B, Pearson RC: **The relative importance of premortem acidosis and postmortem interval for human brain gene expression studies: selective mRNA vulnerability and comparison with their encoded proteins.** *Neurosci Lett* 1995, **200:**151–154.
14. Tomita H, Vawter MP, Walsh DM, Evans SJ, Choudary PV, Li J, Overman KM, Atz ME, Myers RM, Jones EG, Watson SJ, Akil H, Bunney WE Jr: **Effect of agonal and postmortem factors on gene expression profile: quality control in microarray analyses of postmortem human brain.** *Biol Psychiatry* 2004, **55:**346–352.
15. Miyatake Y, Ikeda H, Michimata R, Koizumi S, Ishizu A, Nishimura N, Yoshiki T: **Differential modulation of gene expression among rat tissues with warm ischemia.** *Exp Mol Pathol* 2004, **77:**222–230.
16. Lee J, Hever A, Willhite D, Zlotnik A, Hevezi P: **Effects of RNA degradation on gene expression analysis of human postmortem tissues.** *FASEB J* 2005, **19:**1356–1358.
17. Inoue H, Kimura A, Tuji T: **Degradation profile of mRNA in a dead rat body: basic semi-quantification study.** *Forensic Sci Int* 2002, **130:**127–132.
18. Hong SH, Baek HA, Jang KY, Chung MJ, Moon WS, Kang MJ, Lee DG, Park HS: **Effects of delay in the snap freezing of colorectal cancer tissues on the quality of DNA and RNA.** *J Kor Soc Coloproctol* 2010, **26:**316–323.
19. Heinrich M, Matt K, Lutz-Bonengel S, Schmidt U: **Successful RNA extraction from various human postmortem tissues.** *Int J Legal Med* 2007, **121:**136–142.
20. Huang J, Qi R, Quackenbush J, Dauway E, Lazaridis E, Yeatman T: **Effects of ischemia on gene expression.** *J Surg Res* 2001, **99:**222–227.
21. Baechler EC, Batliwalla FM, Karypis G, Gaffney PM, Moser K, Ortmann WA, Espe KJ, Balasubramanian S, Hughes KM, Chan JP, Begovich A, Chang SY, Gregersen PK, Behrens TW: **Expression levels for many genes in human peripheral blood cells are highly sensitive to ex vivo incubation.** *Genes Immun* 2004, **5:**347–353.
22. Bray SE, Paulin FE, Fong SC, Baker L, Carey F, Levison D, Steele RJ, Kernohan NM: **Gene expression in colorectal neoplasia: modifications induced by tissue ischaemic time and tissue handling protocol.** *Histopathology* 2010, **56:**240–250.
23. Ibberson D, Benes V, Muckenthaler MU, Castoldi M: **RNA degradation compromises the reliability of microRNA expression profiling.** *BMC Biotechnol* 2009, **9:**102.
24. Catts VS, Catts SV, Fernandez HR, Taylor JM, Coulson EJ, Lutze-Mann LH: **A microarray study of post-mortem mRNA degradation in mouse brain tissue.** *Mol Brain Res* 2005, **138:**164–177.
25. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: **RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.** *Genome Res* 2008, **18:**1509–1517.
26. Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, Robertson AJ, Perkins AC, Bruce SJ, Lee CC, Ranade SS, Peckham HE, Manning JM, McKernan KJ, Grimmond SM: **Stem cell transcriptome profiling via massive-scale mRNA sequencing.** *Nat Methods* 2008, **5:**613–619.
27. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.** *Genome Biol* 2011, **14:**R36.
28. Narsai R, Howell KA, Millar AH, O'Toole N, Small I, Whelan J: **Genome-wide analysis of mRNA decay rates and their determinants in Arabidopsis thaliana.** *Plant Cell* 2007, **19:**3418–3436.
29. Bernstein JA, Khodursky AB, Lin PH, Lin-Chao S, Cohen SN: **Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays.** *Proc Natl Acad Sci U S A* 2002, **99:**9697–9702.
30. Ho-Pun-Cheung A, Bascoul-Mollevi C, Assenat E, Boissière-Michot F, Bibeau F, Cellier D, Ychou M, Lopez-Crapez E: **Reverse transcription-quantitative polymerase chain reaction: description of a RIN-based algorithm for accurate data normalization.** *BMC Mol Biol* 2009, **10:**31.
31. Opitz L, Salinas-Riester G, Grade M, Jung K, Jo P, Emons G, Ghadimi BM, Beissbarth T, Gaedcke J: **Impact of RNA degradation on gene expression profiling.** *BMC Med Genet* 2010, **3:**36.
32. Thompson KL, Pine PS, Rosenzweig BA, Turpaz Y, Retief J: **Characterization of the effect of sample quality on high density oligonucleotide microarray data using progressively degraded rat liver RNA.** *BMC Biotechnol* 2007, **7:**57.
33. Strand C, Enell J, Hedenfalk I, Ferno M: **RNA quality in frozen breast cancer samples and the influence on gene expression analysis–a comparison of three evaluation methods using microcapillary electrophoresis traces.** *BMC Mol Biol* 2007, **8:**38.
34. Fleige S, Pfaffl MW: **RNA integrity and the effect on the real-time qRT-PCR performance.** *Mol Aspects Med* 2006, **27:**126–139.
35. Auerbach RK, Euskirchen G, Rozowsky J, Lamarre-Vincent N, Moqtaderi Z, Lefrançois P, Struhl K, Gerstein M, Snyder M: **Mapping accessible chromatin regions using Sono-Seq.** *Proc Natl Acad Sci U S A* 2009, **106:**14926–14931.
36. Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, Raney BJ, Pohl A, Malladi VS, Li CH, Lee BT, Learned K, Kirkup V, Hsu F, Heitner S, Harte RA, Haeussler M, Guruvadoo L, Goldman M, Giardine BM, Fujita PA, Dreszer TR, Diekhans M, Cline MS, Clawson H, Barber GP, *et al*: **The UCSC Genome Browser database: extensions and updates 2013.** *Nucleic Acids Res* 2013, **41:**D64–D69.
37. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5:**621–628.
38. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Garcia-Giron C, Gordon L, Hourlier T, Hunt

S, Juettemann T, Kahari AK, Keenan S, Komorowska M, Kulesha E, Longden I, Maurel T, McLaren WM, Muffato M, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, *et al*: **Ensembl 2013.** *Nucleic Acids Res* 2013, **41:**D48–D55.
39. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5:**R80.
40. Du P, Kibbe WA, Lin SM: **lumi: a pipeline for processing Illumina microarray.** *Bioinformatics* 2008, **24:**1547–1548.
41. Ramsköld D, Wang ET, Burge CB, Sandberg R: **An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data.** *PLoS Comput Biol* 2009, **5:**e1000598.
42. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 2010, **26:**841–842.
43. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26:**139–140.
44. Robinson MD, Oshlack A: **A scaling normalization method for differential expression analysis of RNA-seq data.** *Genome Biol* 2010, **11:**R25.
45. Alexa A, Rahnenfuhrer J, Lengauer T: **Improved scoring of functional groups from gene expression data by decorrelating GO graph structure.** *Bioinformatics* 2006, **22:**1600–1607.
46. Storey JD, Tibshirani R: **Statistical significance for genomewide studies.** *Proc Natl Acad Sci U S A* 2003, **100:**9440–9445.

doi:10.1186/1741-7007-12-42
Cite this article as: Gallego Romero *et al.*: RNA-seq: impact of RNA degradation on transcript quantification. *BMC Biology* 2014 **12**:42.