Comparative analysis of RNA sequencing methods for degraded or low-input samples

Xian Adiconis[1,4], Diego Borges-Rivera[1,4], Rahul Satija[1], David S DeLuca[1], Michele A Busby[1], Aaron M Berlin[1], Andrey Sivachenko[1], Dawn Anne Thompson[1], Alec Wysoker[1], Timothy Fennell[1], Andreas Gnirke[1], Nathalie Pochet[1], Aviv Regev[1–3] & Joshua Z Levin[1]

RNA-seq is an effective method for studying the transcriptome, but it can be difficult to apply to scarce or degraded RNA from fixed clinical samples, rare cell populations or cadavers. Recent studies have proposed several methods for RNA-seq of low-quality and/or low-quantity samples, but the relative merits of these methods have not been systematically analyzed. Here we compare five such methods using metrics relevant to transcriptome annotation, transcript discovery and gene expression. Using a single human RNA sample, we constructed and sequenced ten libraries with these methods and compared them against two control libraries. We found that the RNase H method performed best for chemically fragmented, low-quality RNA, and we confirmed this through analysis of actual degraded samples. RNase H can even effectively replace oligo(dT)-based methods for standard RNA-seq. SMART and NuGEN had distinct strengths for measuring low-quantity RNA. Our analysis allows biologists to select the most suitable methods and provides a benchmark for future method development.

RNA-seq allows us to comprehensively characterize the transcripts present in a biological sample. Although RNA-seq can, in principle, be used to measure transcripts in any sample, it has been challenging to apply standard protocols to samples with either low-quantity or low-quality (partially degraded) input RNA. First, most standard protocols in eukaryotic cells rely on oligo(dT) to isolate poly(A)$^+$ RNA[1] and thereby deplete the highly abundant ribosomal RNA (rRNA). Although this is a powerful technique, it fails to include many other poly(A)$^-$ transcripts for analysis[2]. In addition, for RNA that is not intact, oligo(dT) selection will isolate only the 3′-most portion of each transcript; and oligo(dT) selection is not practical with small amounts of RNA. Moreover, low RNA input can lead to low complexity and suboptimal results and thus often requires an additional amplification step.

Overcoming these challenges is critical to leveraging the power of RNA-seq for many biomedical applications. For example, total RNA-seq of low-quality samples is important for formalin-fixed, paraffin-embedded (FFPE) samples and for clinical samples available only from cadavers, such as those in the large-scale Genotype-Tissue Expression project (http://www.broadinstitute.org/gtex/). Low-quantity total RNA-seq paves the way for research with rare cell populations, minute tissue samples in cancer and even single cells[3–5].

Several methods have recently been proposed to overcome the challenges of low-quality and/or low-quantity RNA samples, including RNase H (also known as selective depletion of abundant RNA, or SDRNA)[6,7], Ribo-Zero[8] (Epicentre), duplex-specific nuclease[9] with light normalization (DSN-lite), the Ovation RNA-seq system (v.1 (ref. 10); v.2 tested here, hereafter referred to as 'NuGEN') and switching mechanism at the 5′ end of the RNA template (SMART; Clontech)[4] (**Fig. 1**). However, identifying the relative merits of each method as compared to a standard high-input, high-quality control—and determining the suitability of each for a particular project—requires careful comparison by multiple metrics[11,12]. To our knowledge, such a comparison had not been conducted to date.

Here we compared these five RNA-seq methods for low-quality and/or low-quantity samples using a comprehensive set of metrics. Starting from one sample of total RNA from a human cell line, we constructed a set of libraries for each method, as well as two control libraries, and sequenced them to deep coverage. For low-quality RNA, the RNase H method performed best. We confirmed these results by analysis of actual degraded samples. For low-quantity RNA, the SMART and NuGEN methods each had distinct strengths.

Whereas some metrics, such as percentage of exonic or rRNA reads, are important for all studies, other features may be critical in only some settings. For expression-profiling projects, metrics related to accuracy and biases in gene expression measurements are critical. Moreover, for samples with low– or high–GC content transcripts, researchers need to give even greater weight to GC bias metrics. For polymorphism detection projects—such as those for somatic mutations in cancer, RNA editing or allele-specific expression—evenness of coverage, 5′-to-3′ coverage bias, and complexity are important. For transcriptome annotation

[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [2]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [3]Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [4]These authors contributed equally to this work. Correspondence should be addressed to J.Z.L. (jlevin@broadinstitute.org) or A.R. (aregev@broad.mit.edu).

RECEIVED 18 FEBRUARY; ACCEPTED 15 APRIL; PUBLISHED ONLINE 19 MAY 2013; CORRECTED AFTER PRINT 2 DECEMBER 2013; DOI:10.1038/NMETH.2483

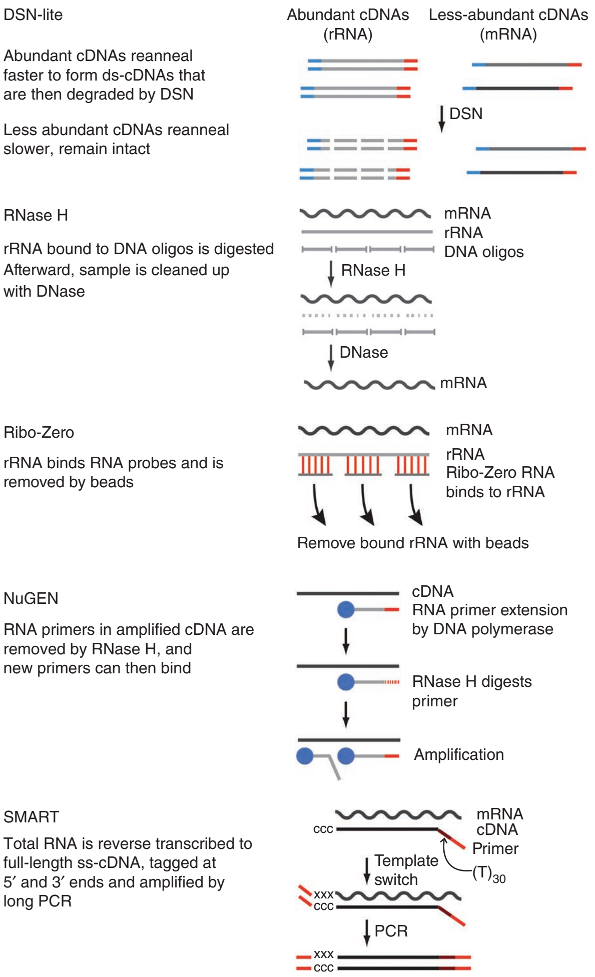
DSN-lite
Abundant cDNAs reanneal faster to form ds-cDNAs that are then degraded by DSN
Less abundant cDNAs reanneal slower, remain intact
Abundant cDNAs (rRNA)
Less-abundant cDNAs (mRNA)
DSN
RNase H
rRNA bound to DNA oligos is digested Afterward, sample is cleaned up with DNase
mRNA
rRNA
DNA oligos
RNase H
DNase
mRNA
Ribo-Zero
rRNA binds RNA probes and is removed by beads
mRNA
rRNA
Ribo-Zero RNA binds to rRNA
Remove bound rRNA with beads
NuGEN
RNA primers in amplified cDNA are removed by RNase H, and new primers can then bind
cDNA
RNA primer extension by DNA polymerase
RNase H digests primer
Amplification
SMART
Total RNA is reverse transcribed to full-length ss-cDNA, tagged at 5′ and 3′ ends and amplified by long PCR
CCC
mRNA
cDNA
Primer
(T)30
Template switch
XXX
CCC
PCR
XXX
CCC

Figure 1 | Methods for total RNA-seq. Salient details for five protocols for total RNA-seq are shown. DSN-lite, RNase H and Ribo-Zero were tested for low-quality samples. SMART was tested for low-quantity samples. NuGEN, which generates double-stranded (ds)-cDNA that is amplified with Ribo-SPIA (Single Primer Isothermal Amplification), was tested for both types of samples. RNA and matching cDNA are black, adaptors and primers are colored and rRNA is gray. ss-cDNA, single-stranded cDNA. $(T)_{30}$ is composed of 30 T bases.

projects—such as studies of large intergenic noncoding RNAs (lincRNAs), alternative splicing or fusion transcripts in cancer cells—evenness of coverage, 5′-to-3′ coverage bias, and coverage of 5′ and 3′ ends are most relevant. In all cases, ease of use and cost are important factors.

RESULTS

A comparison of RNA-seq methods starting from total RNA

We evaluated five methods for preparing RNA-seq libraries from samples of low quality and/or low quantity (**Fig. 1**). For low-quality samples, we tested four methods (DSN-lite, RNase H, Ribo-Zero and NuGEN; **Fig. 1**) with highly fragmented RNA (Online Methods), preparing six libraries in all. All of these libraries started with 1 μg of total RNA, except for 'NuGEN 100f', which began with 100 ng of fragmented RNA. For the RNase H method, we created a second library by the same protocol with no additional spike-in RNA ('NS' library; Online Methods), and for DSN-lite, we created a second library by a protocol with PCR before DSN treatment (Online Methods). Because these second libraries performed very similarly to their counterparts, we report their results only in **Supplementary Tables 1–3**. For low-quantity samples, we tested two methods for 1 ng of intact total RNA ('NuGEN 1i' and SMART; **Fig. 1**) and an oligo(dT) selection method ('TruSeq'). Finally, we tested NuGEN with 1 ng of fragmented RNA ('NuGEN 1f'), which represented both low-quality and low-quantity input. For controls, we prepared two standard libraries from abundant high-quality RNA: one with oligo(dT) selection of poly(A)$^+$ RNA (PolyA) and the other from total RNA with no manipulation to remove rRNA (Total).

For each method, we prepared a cDNA library for Illumina sequencing, starting with total RNA from the human chronic myeloid leukemia cell line K-562. Using paired-end sequencing, we generated >75 million reads for each library (**Supplementary Table 1**).

Efficiency of rRNA depletion

We first assessed the fraction of reads aligning to rRNA (**Fig. 2a**). Because rRNA reads are not informative for most RNA-seq experiments, depletion of rRNA maximizes the coverage of the other transcripts present in a sample. Among the low-quality RNA libraries, RNase H had the lowest fraction of rRNA-aligning reads (0.1%), whereas Ribo-Zero (11.3%) and NuGEN 100f (23.2%) were substantially less efficient. Among the low-quantity RNA libraries, SMART had a much lower fraction of rRNA-aligning reads (5.5%) than TruSeq (17.4%) or NuGEN 1i (28.7%). NuGEN 1f performed slightly less well than NuGEN 1i. Most of the rRNA-aligning reads in the NuGEN 100i, 1i and 1f libraries (91%, 73% and 52%, respectively) were derived from mitochondrial rather than cytoplasmic RNA.

Library complexity

To directly compare the libraries by a comprehensive set of metrics, we sampled equivalent sequence data sets for each library (**Supplementary Fig. 1** and **Supplementary Note 1**). We first examined library complexity as measured by the duplication rate. A higher-complexity library provides a better sampling of the RNA present in a sample[11]. Among the low-quality RNA libraries, all had a duplication rate below 20%, an acceptable rate of complexity at this depth of sequencing (**Fig. 2b**). The low duplication rate of NuGEN 100f (4.4%) cannot be directly compared to those of the other libraries (**Supplementary Note 2**). As expected, DSN-lite with PCR before DSN treatment had a higher duplication rate than DSN-lite (22.0% versus 5.5%; **Supplementary Table 2**).

Complexity is especially important for low-quantity libraries (**Supplementary Note 2**). For an alternative measure of their complexity, we compared the number of genes detected in each library (Online Methods). NuGEN 1i had slightly more genes with coverage (transcripts per million (TPM) > 0.1) than SMART (14,149 versus 13,843 genes, respectively; **Supplementary Fig. 2** and **Supplementary Table 2**). The TruSeq method performed poorly on the basis of its extremely high duplication rate (90.7%) (**Fig. 2b**). NuGEN 1f did not perform as well as NuGEN 1i as determined from the number of genes covered (**Supplementary Fig. 2** and **Supplementary Table 2**).

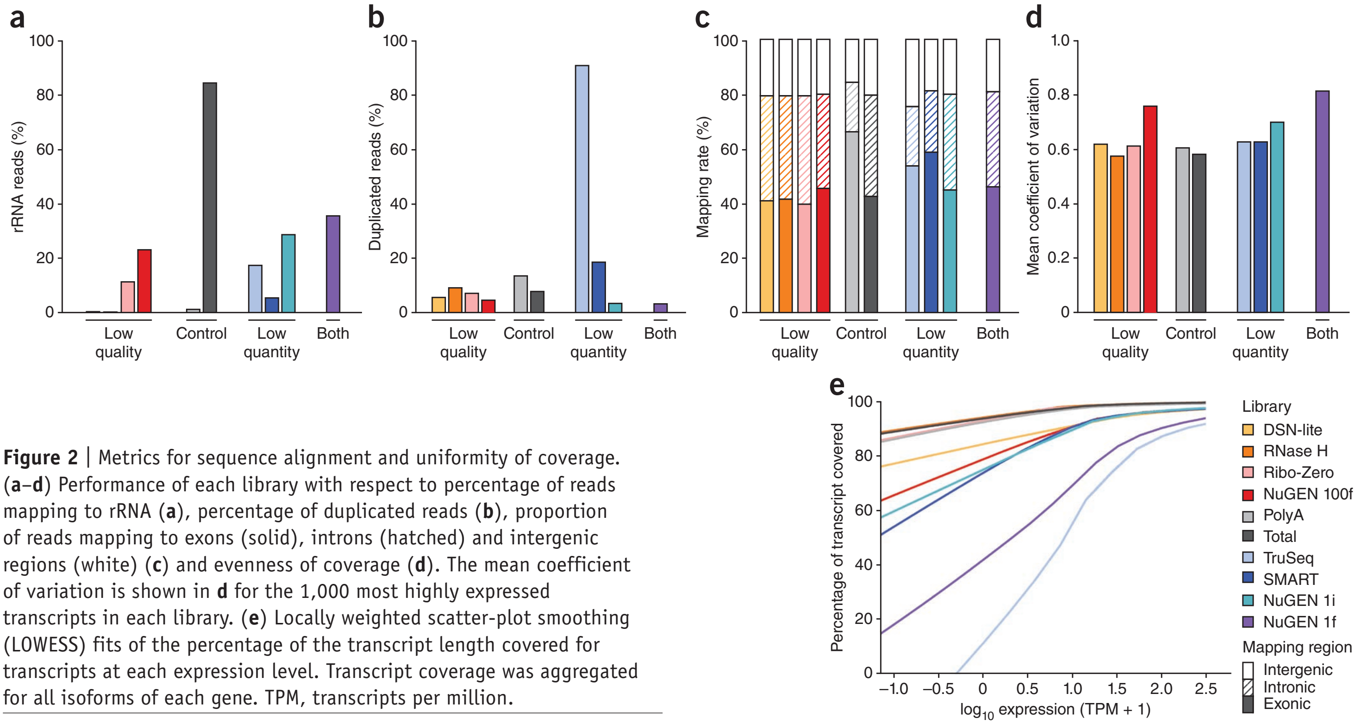

Figure 2 | Metrics for sequence alignment and uniformity of coverage. (**a**–**d**) Performance of each library with respect to percentage of reads mapping to rRNA (**a**), percentage of duplicated reads (**b**), proportion of reads mapping to exons (solid), introns (hatched) and intergenic regions (white) (**c**) and evenness of coverage (**d**). The mean coefficient of variation is shown in **d** for the 1,000 most highly expressed transcripts in each library. (**e**) Locally weighted scatter-plot smoothing (LOWESS) fits of the percentage of the transcript length covered for transcripts at each expression level. Transcript coverage was aggregated for all isoforms of each gene. TPM, transcripts per million.

Relative coverage of annotated exons and introns

To examine the composition of transcripts selected by each method, we assessed the proportion of reads mapping to annotated exons, introns and intergenic regions (**Fig. 2c**). Methods that are not based on the presence of a poly(A) tail also sample partially spliced (immature) RNA molecules because RNA splicing precedes polyadenylation[13]. As expected, PolyA, TruSeq and SMART, in which oligo(dT) was used to select for poly(A)$^+$ RNA or to prime cDNA synthesis, had a greater fraction of reads aligning to exons. All the non-oligo(dT) methods had similarly lower proportions of exonic reads. The fraction of intergenic reads did not vary much among all the libraries (**Fig. 2c**).

Evenness and continuity of transcript coverage

We next compared the evenness and continuity of transcript coverage among our libraries. For evenness of coverage, we calculated the mean coefficient of variation (CV) for each of the 1,000 most highly expressed transcripts in each library. A lower value, indicating less variation, signals better performance. Among the low-quality RNA libraries, RNase H and NuGEN 100f had the lowest and highest mean CVs, respectively (**Fig. 2d**). Among the low-quantity libraries, SMART outperformed NuGEN 1i (**Fig. 2d**). NuGEN 1f had the highest mean CV of any library (**Fig. 2d**).

To assess continuity of coverage, we considered the proportion of each gene's length covered by reads and the number of uncovered gaps in each transcript. Because expression levels are expected to affect both measures, we compared the former to the expression levels (**Fig. 2e** and **Supplementary Fig. 3**) and calculated the latter for the 1,000 most highly expressed transcripts in each library (**Supplementary Fig. 4**). Among the low-quality RNA libraries, RNase H and Ribo-Zero performed better than DSN-lite and NuGEN 100f for both measures. Among the low-quantity libraries, SMART and NuGEN 1i were comparable, with TruSeq showing the poorest performance. NuGEN 1f had less continuity of coverage than NuGEN 1i.

Coverage variation relative to 5′ and 3′ ends

Other measures of transcript coverage are the variation in coverage along each transcript from 5′ to 3′ (**Fig. 3a**) and the number of genes with covered 5′ or 3′ ends (**Fig. 3b**,**c**). Among the low-quality RNA libraries, RNase H and Ribo-Zero had the best coverage from 5′ to 3′ as well as at the ends (**Fig. 3**). All NuGEN libraries were biased toward increased coverage at the 3′ end, possibly owing to the use of oligo(dT) primers in addition to random primers for first-strand cDNA synthesis. Among the low-quantity libraries, NuGEN 1i had more even coverage from 5′ to 3′ than SMART, but SMART had slightly higher coverage at the ends. NuGEN 1f performed similarly to NuGEN 1i with respect to 5′-to-3′ bias but with fewer ends covered, though this could be partially due to the lower number of paired aligned reads for NuGEN 1f.

In some cases, the performance for these measures was affected by transcript length (**Supplementary Fig. 5**). Among the low-quality libraries, 5′-end coverage in transcripts longer than 1,000 bases was substantially lower in NuGEN 100f and DSN-lite, whereas RNase H and Ribo-Zero performed well even for longer transcripts (**Supplementary Fig. 5a**). There was a stronger 3′ bias in coverage in NuGEN 100f for transcripts shorter than 5,000 bases but not in the other libraries (**Supplementary Fig. 5b**). Among the low-quantity libraries, all of the libraries had better 5′-end coverage for transcripts shorter than 1,000 bases than for longer transcripts, with NuGEN 1i performing similarly to SMART for transcripts longer than 5,000 bases. Similarly, the 3′-end coverage was lower for longer transcript lengths in all libraries, but this result was most pronounced for TruSeq for transcripts longer than 5,000 bases. Furthermore, in SMART for longer transcripts, there was more 3′ bias that was also present,

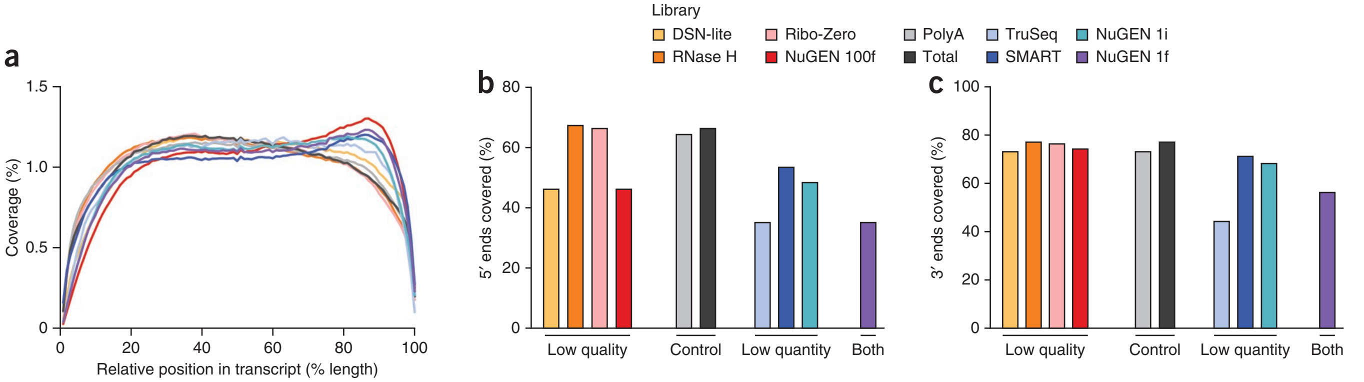

Figure 3 | 5′-to-3′ sequence coverage. (**a**) Normalized coverage by position. For each library, the average relative coverage is shown at each relative position along the transcripts' length. (**b**,**c**) Percentage of annotated 5′ (**b**) and 3′ ends (**c**) covered by reads.

to a lesser extent, in NuGEN 1i. The 3′ bias shown by SMART for longer transcripts has been observed previously[4] and is likely due to the oligo(dT) priming of cDNA synthesis. NuGEN 1f performed slightly less well than NuGEN 1i with respect to length biases (**Supplementary Fig. 5**).

Expression-level performance

As many RNA-seq experiments are focused on expression-level measurements, examining the performance of each library in this area is particularly important. For a gold-standard control, we sequenced a library from total RNA deeply to nearly 1 billion reads (**Supplementary Table 1**). After computationally removing all rRNA mapped reads, we created a 'truth' data set for libraries in which we removed rRNA experimentally, and we looked for those libraries with the best correlation to this control. Spike-in RNA may also be used to assess expression performance (**Supplementary Note 3**).

Among the low-quality libraries, RNase H and Ribo-Zero performed best, on the basis of their high correlations with this control library (Pearson correlation coefficient $r = 0.962$ and 0.955, respectively) as well as other measures of consistency at different expression levels, such as quantile-quantile (Q-Q) and *MA* plots[14]—the latter of which compares, for each gene, the difference in expression between two libraries against the mean expression of that gene in the two libraries (**Fig. 4** and **Supplementary Fig. 6**). Notably, RNase H NS had a correlation of 0.972 with RNase H, which indicates that this method is reproducible (**Supplementary Table 3**).

Among the low-quantity libraries, NuGEN 1i ($r = 0.861$) performed essentially the same as SMART ($r = 0.860$), which was followed by TruSeq ($r = 0.844$; **Fig. 4** and **Supplementary Fig. 6**). NuGEN 1f had the lowest correlation with the control library ($r = 0.787$; **Fig. 4**) but was better correlated with NuGEN 100f ($r = 0.877$), which indicates that libraries made from fragmented RNA by NuGEN may perform relatively consistently (**Supplementary Table 3**).

Effects of transcript length and GC content on expression

Finally, we tested whether any of the methods introduced particular biases in estimating expression levels of specific subsets of genes. In particular, length and GC biases[15,16] have been reported previously in Illumina sequencing data.

For length bias, we compared library performance for short, medium and long genes. There was not much variation in

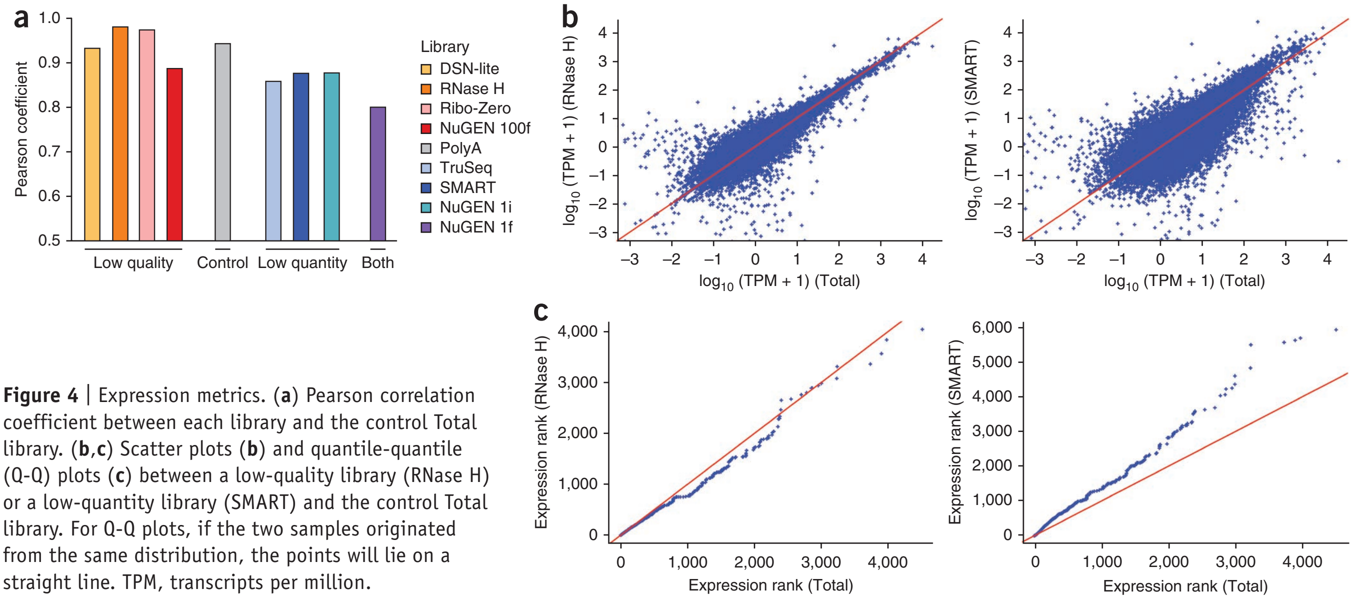

Figure 4 | Expression metrics. (**a**) Pearson correlation coefficient between each library and the control Total library. (**b**,**c**) Scatter plots (**b**) and quantile-quantile (Q-Q) plots (**c**) between a low-quality library (RNase H) or a low-quantity library (SMART) and the control Total library. For Q-Q plots, if the two samples originated from the same distribution, the points will lie on a straight line. TPM, transcripts per million.

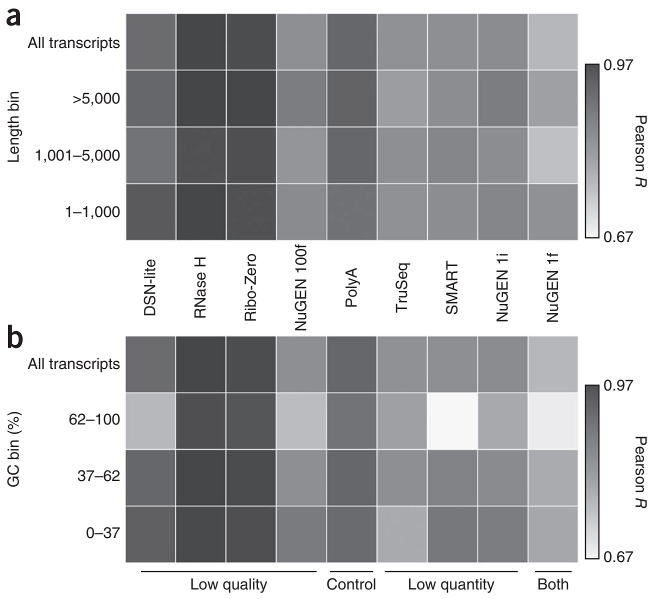

Figure 5 | Length and GC biases in expression metrics. (**a**,**b**) Pearson correlation coefficient between each library (columns) and the control Total library for either all transcripts (top row in **a**,**b**), transcripts with different lengths (**a**) or transcripts with different GC content (**b**). The numbers of transcripts expressed in the control Total library in the 1–1,000 bin, 1,001–5,000 bin and >5,000 bin for transcript length were 3,716, 38,088 and 7,050, respectively. The numbers of transcripts expressed in the control Total library in the ≤37% bin, 37–62% bin and >62% bin for GC content were 2,358, 42,660 and 3,836, respectively.

the expression metrics for genes of different lengths in any of the libraries (**Fig. 5a** and **Supplementary Fig. 7a**), except for NuGEN 1f, which performed slightly less well than NuGEN 1i with respect to length biases (**Fig. 5a**).

To detect GC bias, we examined expression metrics for subsets of genes with low, medium or high GC content. Among the low-quality libraries, RNase H and Ribo-Zero performed well, but NuGEN 100f and DSN-lite showed lower correlations to the control Total library at high GC content (**Fig. 5b** and **Supplementary Fig. 7b**). The poorer performance of DSN-lite is likely due to the hybridization proceeding more quickly for higher-GC cDNA inserts, as previously reported with DSN normalization[17]. Among the low-quantity libraries, although correlations between the Total library and either NuGEN 1i or SMART were lower for genes with high GC content than for all genes (r = 0.800 versus 0.861 for NuGEN 1i; r = 0.676 versus 0.860 for SMART), SMART performed particularly poorly, perhaps owing to GC-dependent differences in amplification in the first round of PCR (**Fig. 5b**). NuGEN 1f performed similarly to NuGEN 1i.

Evaluation of methods with actual degraded samples

To further test the applicability of methods for low-quality RNA in realistic clinical and biological settings, we applied the two most promising methods, RNase H and Ribo-Zero, to two RNA samples with degradation similar to what might be encountered in actual RNA-seq experiments. One sample—from kidney—was degraded because of FFPE fixation, and the other—from pancreas—was degraded during isolation, as it is difficult to isolate intact RNA from this organ[18]. We also prepared a Total library from total RNA from each sample as a control.

By many metrics, our results with these libraries were similar to those with the fragmented K-562 libraries (**Fig. 6** and **Supplementary Fig. 8**). The fraction of rRNA reads was low for RNase H (0.1–0.3%) but was substantially higher for Ribo-Zero (19.3–25.1%; **Fig. 6a**). The unexpectedly low fraction of reads aligning to the genome and transcriptome reference sequences for kidney Ribo-Zero was due to a large fraction of reads (46%) aligning to an ~300-base-pair region in the 45S rRNA transcript adjacent to the 3′ end of the 28S rRNA on an unplaced contig not included in our genome or transcriptome reference sequences (**Supplementary Table 1**). This region also had reads aligning to it in other libraries, though not as many. The complexities of RNase H and Ribo-Zero as measured by duplication rates were generally comparable and less than 20% (**Supplementary Table 2**). Libraries with better duplication rates had more genes detected. Furthermore, these tissue-derived libraries had more genes detected than did the libraries from K-562 cells. We observed a greater fraction of reads aligning to intron sequences in the FFPE libraries than in other libraries (**Fig. 6b**), as has been previously reported[7]. The evenness of coverage was slightly better for RNase H than Ribo-Zero as measured by the CV (**Fig. 6c**), proportion of gene-length coverage (**Supplementary Fig. 8a**) and number of gaps in coverage of the 1,000 most highly expressed transcripts (**Supplementary Table 2**). Transcript coverage from 5′ to 3′ (**Supplementary Fig. 8b**), at 5′ and 3′ ends (**Supplementary Table 2**)

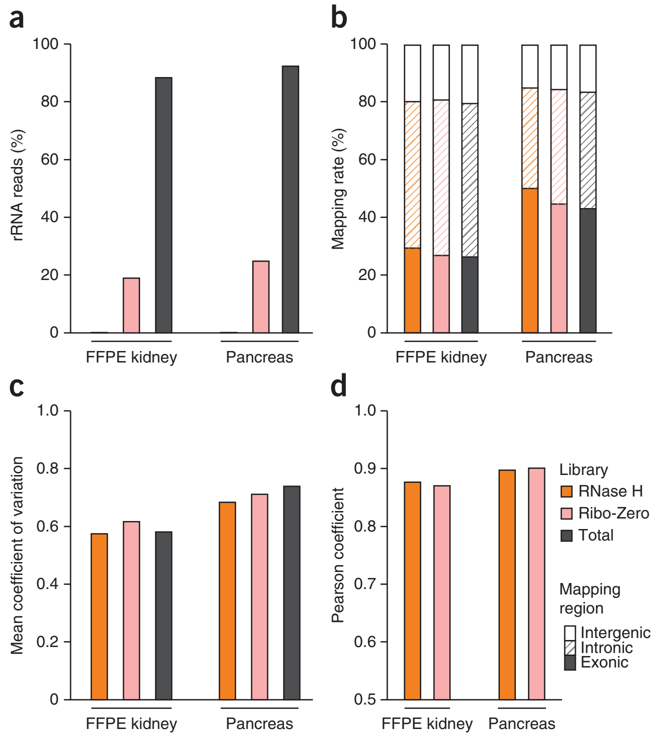

Figure 6 | Performance for actual degraded samples. Key metrics for RNase H, Ribo-Zero and Total libraries from pancreas and formalin-fixed, paraffin-embedded (FFPE) kidney RNA. (**a**) Percentage of reads mapping to rRNA. (**b**) Proportion of reads mapping to exons (solid), introns (hatched) and intergenic regions (white). (**c**) Mean coefficient of variation for the 1,000 most highly expressed transcripts in each library. (**d**) Pearson correlation coefficient between each library and a control Total library.

and for transcripts of different lengths (**Supplementary Fig. 8c**,**d**) were all comparable between the two types of libraries, with the possible exception of slightly lower coverage of the 3′ ends for RNase H in pancreas, especially for the longest transcripts. There was essentially no difference in expression levels for these rRNA-depleted libraries as determined from their correlation to expression in corresponding Total libraries (**Fig. 6d**) and other expression plots (**Supplementary Fig. 8e–g**). These correlations ($r = 0.867–0.894$) were lower than those for the K-562 libraries from low-quality RNA ($r = 0.955–0.962$) (**Figs. 4a** and **6d**). As with the results for RNase H and Ribo-Zero with K-562 RNA, we did not detect any strong biases in expression correlation for these libraries with respect to varying length and GC content (**Supplementary Fig. 8h**,**i**).

DISCUSSION

We compared five distinct methods for RNA-seq with low-quantity and/or low-quality input RNA by a comprehensive set of quality measures. These methods also vary in the associated time and cost of materials (**Supplementary Table 4**). The per-sample costs of the commercial kits (for Ribo-Zero, NuGEN and SMART) are substantially higher than those of the other methods (DSN-lite and RNase H). The amount of time per library for each method is similar, except that DSN-lite requires about one additional day.

Overall, the RNase H method performed best for low-quality RNA by most measures (**Supplementary Table 5**). Ribo-Zero performed similarly to RNase H by many metrics; as such, Ribo-Zero might be acceptable for researchers who prefer to use a kit or have only a few samples. We note that this option has a higher cost per sample and requires deeper sequencing to compensate for higher rRNA levels (**Figs. 2a** and **6a**) and other reads aligning in the 45S rRNA transcript adjacent to the 3′ end of the 28S rRNA (**Supplementary Table 1**). The similarity between our results from actual degraded (pancreas and FFPE kidney; **Fig. 6**, **Supplementary Tables 1** and **2** and **Supplementary Fig. 8**) and chemically fragmented (K-562) RNA supports the use of the latter as a model for the former. SMART and NuGEN each had specific advantages for low-quantity samples (**Supplementary Note 4**). We excluded some methods from our comparisons because of their technical limitations or performance issues (**Supplementary Note 5**). While this manuscript was under review, two RNA-seq methods for low-quantity samples were published[19], but these became available too late for us to include them in our comparisons.

Only some aspects of the biases associated with length or GC content can be corrected computationally *post hoc*. For example, the *ab initio* RNA-seq assembler Cufflinks[20] controls for 5′-to-3′ and GC biases when estimating expression levels. However, if a method undersamples 5′ ends, the undersampling will inherently limit the method's ability to annotate and correctly assemble novel transcripts. Similarly, for identification of single-nucleotide polymorphisms with RNA-seq, a bias in coverage cannot be computationally corrected and will prevent identification of many such polymorphisms or will require deeper sequencing and higher costs.

One of the major distinctions between methods is whether they focus on poly(A)$^+$ transcripts (through their use of oligo(dT)) or capture both poly(A)$^+$ and poly(A)$^-$ RNAs. The former can measure mRNAs at a lower sequencing cost than the latter and might simplify transcript assembly because a higher fraction of reads align to exons (**Fig. 2c**). The latter would help study poly(A)$^-$ mRNAs as well as immature transcripts. Oligo(dT) selection did not perform well for our low-quantity sample (TruSeq; **Figs. 2**–**5**) and is not appropriate for low-quality RNA-seq either, as discussed above. We note that all of the non–poly(A)-based methods for low-quality RNA can be extended to prokaryotic samples that do not have RNA with poly(A) tails, as we have recently shown for DSN-lite and Ribo-Zero[17], which broadens their utility. Furthermore, expression levels estimated with the PolyA library were not identical to Total library expression levels (**Supplementary Table 3**). Thus, an argument can be made that researchers performing 'standard' RNA-seq (with high-quantity and high-quality RNA) should nevertheless use an rRNA depletion method, such as RNase H, rather than oligo(dT). In this way, we can capture a more complete view of the transcriptome and facilitate direct comparisons between high- and low-quality RNA samples.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession code. Gene Expression Omnibus: GSE40705 (sequence data).

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank D. Sinicropi for alerting us to his RNase H method, Broad Genomics Platform for sequencing work, C. Nusbaum, R. Johnson, M. Yassour and V. Savova for helpful discussions, and L. Gaffney for assistance in drafting figures. Work was supported by US National Institutes of Health Pioneer Award DP1-OD003958-01, US National Human Genome Research Institute (NHGRI) 1P01HG005062-01, NHGRI Center of Excellence in Genome Science Award 1P50HG006193-01, the Howard Hughes Medical Institute, the Merkin Foundation for Stem Cell Research and the Klarman Cell Observatory at the Broad Institute (to A.R.), and by NHGRI grant HG03067. N.P. was supported by a postdoctoral research fellowship of the Fund for Scientific Research–Flanders (FWO Vlaanderen).

AUTHOR CONTRIBUTIONS

J.Z.L., X.A. and A.R. conceived the research. X.A. prepared the cDNA libraries. D.B.-R., R.S., N.P., M.A.B. and A.R. developed and performed computational analysis. D.S.D. contributed code. D.S.D., A.M.B., A.S., A.W. and T.F. helped with computational analysis. D.A.T., N.P., A.R. and J.Z.L. supervised the research. J.Z.L., X.A., D.B.-R. and A.R. wrote the paper. R.S., A.G. and D.S.D. assisted in editing the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Aviv, H. & Leder, P. Purification of biologically active globin messenger RNA by chromatography on oligothymidylic acid-cellulose. *Proc. Natl. Acad. Sci. USA* **69**, 1408–1412 (1972).
2. Yang, L., Duff, M.O., Graveley, B.R., Carmichael, G.G. & Chen, L.L. Genomewide characterization of non-polyadenylated RNAs. *Genome Biol.* **12**, R16 (2011).
3. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
4. Ramsköld, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).
5. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* **21**, 1160–1167 (2011).
6. Sinicropi, D. & Morlan, J. Methods for depleting RNA from nucleic acid samples. US patent application 20110111409 (2011).

7. Morlan, J.D., Qu, K. & Sinicropi, D.V. Selective depletion of rRNA enables whole transcriptome profiling of archival fixed tissue. *PLoS ONE* **7**, e42882 (2012).
8. Huang, R. *et al.* An RNA-Seq strategy to detect the complete coding and non-coding transcriptome including full-length imprinted macro ncRNAs. *PLoS ONE* **6**, e27288 (2011).
9. Yi, H. *et al.* Duplex-specific nuclease efficiently removes rRNA for prokaryotic RNA-seq. *Nucleic Acids Res.* **39**, e140 (2011).
10. Tariq, M.A., Kim, H.J., Jejelowo, O. & Pourmand, N. Whole-transcriptome RNAseq analysis from minute amount of total RNA. *Nucleic Acids Res.* **39**, e120 (2011).
11. Levin, J.Z. *et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods* **7**, 709–715 (2010).
12. DeLuca, D.S. *et al.* RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28**, 1530–1532 (2012).
13. Beyer, A.L. & Osheim, Y.N. Splice site selection, rate of splicing, and alternative splicing on nascent transcripts. *Genes Dev.* **2**, 754–765 (1988).
14. Yang, Y.H. *et al.* Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30**, e15 (2002).
15. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12**, R18 (2011).
16. Rosenkranz, R., Borodina, T., Lehrach, H. & Himmelbauer, H. Characterizing the mouse ES cell transcriptome with Illumina sequencing. *Genomics* **92**, 187–194 (2008).
17. Giannoukos, G. *et al.* Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Biol.* **13**, R23 (2012).
18. Griffin, M., Abu-El-Haija, M., Abu-El-Haija, M., Rokhlina, T. & Uc, A. Simplified and versatile method for isolation of high-quality RNA from pancreas. *Biotechniques* **52**, 332–334 (2012).
19. Pan, X. *et al.* Two methods for full-length RNA sequencing for low quantities of cells and single cells. *Proc. Natl. Acad. Sci. USA* **110**, 594–599 (2013).
20. Roberts, A., Trapnell, C., Donaghey, J., Rinn, J.L. & Pachter, L. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol.* **12**, R22 (2011).

ONLINE METHODS

Low-quality RNA samples. To prepare fragmented RNA from high-quality human K-562 RNA (RNA Integrity Number (RIN) 9.0; Ambion), we mixed 10 μg of K-562 total RNA (Ambion) with 2 μl of ERCC RNA Spike-In Control Mix 1 (1:10, Ambion). We heated this mixture in 1× fragmentation buffer (Affymetrix) at 85 °C for 3 min, quickly chilled it on ice and purified the RNA with 2.2× RNAClean SPRI beads (Beckman Coulter Genomics).

For human kidney FFPE tissue sections (Cybrdi), we isolated total RNA with the MasterPure RNA Purification Kit (Epicentre) according to the manufacturer's instructions with the following modifications. In step C, we incubated the sample at 37 °C for 30 min to remove contaminating DNA. In addition to the original protocol, we used phenol:chloroform:isoamyl alcohol (25:24:1; Invitrogen) extraction and ethanol precipitation to minimize possible carryover of organic solvent and proteins. We also did a purification with 1.0× RNAClean SPRI beads to remove RNA fragments that could be too small for standard Illumina sequencing and might therefore affect the overall library quality.

For a partially degraded human pancreas total RNA sample (Zyagen), we did two rounds of purification using 2.2× RNAClean SPRI beads. We then removed contaminating DNA using TURBO DNase (Ambion) rigorous treatment and purified the DNase-treated RNA with 2.2× RNAClean SPRI beads.

We assessed the extent of RNA degradation in each sample (kidney FFPE, pancreas and K-562) using a BioAnalyzer (Agilent; **Supplementary Fig. 9**). No fragmentation was necessary for the kidney FFPE and pancreas RNA. Although these particular samples were degraded to the point that no additional fragmentation was necessary to prepare libraries for RNA-seq, this choice could be adjusted for other samples based on a check of their RNA integrity.

RNase H libraries. Our RNase H protocol is similar to the published method[6,7], with some minor differences including oligonucleotide composition and lengths and DNase treatment (see below). To prepare rRNA oligonucleotide pools, we designed 195 50-base DNA oligonucleotides covering the reverse complement of the entire length of each rRNA (**Supplementary Table 6**). We then pooled together equal molar amounts of each these oligonucleotides (Eurofins MWG Operon).

To deplete rRNA, we added 1,000 ng rRNA pooled oligonucleotides to 1,000 ng of fragmented RNA, incubated the mixture in 1× hybridization buffer (200 mM NaCl, 100 mM Tris-HCl, pH 7.4) in a final volume of 5 μl at 95 °C for 2 min and then slowly ramped the temperature (−0.1 °C/s) to 45 °C. We added 5 μl preheated RNase H reaction mix that contains 10 U of Hybridase Thermostable RNase H (Epicentre), 0.5 μmol Tris-HCl, pH 7.5, 1 μmol NaCl and 0.2 μmol $MgCl_2$ to the RNA and DNA oligo mix, incubated this mixture at 45 °C for 30 min and then placed it on ice. We purified the RNase H–treated RNA with 2.2× RNAClean SPRI beads. We removed the oligonucleotides using TURBO DNase rigorous treatment and purified the DNase-treated RNA with 2.2× RNAClean SPRI beads.

For the RNase H NS library, we also prepared an RNase H rRNA-depleted sample by using 1,000 ng of fragmented K-562 total RNA without ERCC RNA Spike-In control RNA following the protocol described above.

For the pancreas and FFPE kidney RNase H libraries, we prepared rRNA-depleted samples starting from 1,000 ng RNA following this protocol but omitting the fragmentation step.

For all RNase H libraries, we synthesized double-stranded cDNA from RNase H–treated RNA using the "control (non–strand-specific)" protocol as described[11], except that we purified the double-stranded cDNA with 1.8× AMPure XP SPRI beads (Beckman Coulter Genomics). We prepared indexed paired-end libraries for Illumina sequencing as described[11], using Phusion High-Fidelity DNA polymerase with GC buffer (New England BioLabs) and 2 M betaine for the final PCR amplification step, with the following modifications. First, we used forked adaptors containing unique 8-base index sequences to enable pooling of libraries in the same flow-cell lane. Second, we adjusted adaptor input proportionally to the cDNA input with 2 μl of 15 μM adaptor for each 1 μg cDNA but with no less than 1.2 μl. Third, we size-selected the ligation product by using two rounds of 0.7× AMPure XP SPRI beads cleanup after raising the volume of the ligation reaction to 100 μl. Fourth, we used 55 °C as the annealing temperature in PCR with the universal indexing primers (forward primer, 5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGAC-3′; reverse primer, 5′-CAAGCAGAAGACGGCATACGAGAT-3′). Fifth, we performed 10 cycles of PCR for the K-562 and pancreas samples and 12 cycles for the FFPE sample. Sixth, we removed PCR primers using 1.0× AMPure XP SPRI beads.

Ribo-Zero libraries. To deplete rRNA, we used 1,000 ng of fragmented K-562 RNA prepared as described above. We used the Ribo-Zero rRNA Removal Kit (catalog no. RZH1046, Epicentre) and followed the manufacturer's instructions, except at the last purification step we used 1.8× RNAClean SPRI beads instead of ethanol precipitation.

We also prepared rRNA-depleted samples from 1,000 ng each of the pancreas and FFPE kidney RNA described above with the Ribo-Zero Magnetic Gold Kit (catalog no. MRZG126, Epicentre), following the manufacturer's instructions.

For the Ribo-Zero libraries, we synthesized double-stranded cDNA and prepared an indexed Illumina library as described for the RNase H libraries.

PolyA library. To isolate poly(A)$^+$ mRNA, we used 10 μg of intact K-562 total RNA with the Dynabeads mRNA Purification Kit (Invitrogen) and followed manufacturer's standard protocol, except that we performed an additional round of purification before the final elution.

We added 1 μl Spike-In Control to 10 ng of poly(A)$^+$ mRNA. We fragmented and purified RNA, synthesized double-stranded cDNA and prepared indexed Illumina libraries as described for the RNase H libraries, except that we performed 12 cycles of PCR.

DSN-lite libraries. We used 1,000 ng of fragmented RNA prepared as described above. We synthesized double-stranded cDNA and prepared an indexed Illumina library using the same protocol as for the RNase H libraries, except that we omitted PCR before DSN treatment. To deplete rRNA cDNAs, we mixed all of the ligation products from the library construction ligation step with 4.5 μl hybridization buffer (2 M NaCl, 200 mM HEPES, pH 7.3) in a total volume of 18 μl. We heated this hybridization mix at 98 °C for 2 min and then held it at 68 °C for 5 h. We added 20 μl preheated 2× DSN buffer (Axxora) to the hybridization mix and incubated it at 68 °C for 10 min before adding 2 U of DSN (Axxora). We then incubated the reaction mix at 68 °C for another

doi:10.1038/nmeth.2483

25 min, terminated the reaction by adding 20 µl of stop solution (Axxora), placed it on ice and purified the reaction mix with 1.8× AMPure XP SPRI beads. We performed nine cycles of PCR as described for the RNase H libraries.

We also constructed a second DSN-lite library with PCR before the DSN treatment. In this case, we performed six cycles of PCR after adaptor ligation and used 100 ng of PCR-enriched cDNA library for DSN treatment. We performed the final PCR as described for the RNase H libraries with 12 cycles, except that we used a different reverse primer (5′-CAAGCAGAAGACGGCAT ACGAGATxxxxxxxxGTGACTGGAGTTCAGACGTGT-3′, with "xxxxxxxx" matching the 8-base index for each library; sequences are derivatives of proprietary Illumina nucleotides).

NuGEN libraries. We synthesized and amplified cDNA using the Ovation RNA-seq System (version 2, NuGEN) from the following three RNA samples: 1 ng intact K-562 total RNA with 0.5 µl ERCC RNA Spike-In Control Mix 1 (1:25,000) (NuGEN 1i), 100 ng fragmented RNA (NuGEN 100f) and 1 ng fragmented RNA (NuGEN 1f)—with the latter two prepared as described above. We sheared all the cDNA under the following conditions: 3 min with 10% duty cycle, 5% intensity and 100 cycles per burst in the frequency-sweeping mode (Covaris S2 machine). We purified the sheared cDNA with 2.2× AMPure XP SPRI beads and then prepared indexed Illumina libraries as described for the RNase H libraries, except that we performed six cycles of PCR.

SMART library. We synthesized and amplified cDNA from 1 ng intact K-562 total RNA using the SMARTer Ultra Low RNA Kit (Clontech) according to the manufacturer's protocol. We sheared the cDNA under the following conditions: 5 min with 10% duty cycle, 5% intensity and 200 cycles per burst in the frequency-sweeping mode (Covaris S2 machine). We purified the sheared cDNA with 2.2× AMPure XP SPRI beads and then prepared a standard paired-end library for Illumina sequencing as described[11], except that we size-selected the ligation product as described above for the RNase H libraries and we used 12 cycles of PCR.

TruSeq library. We used 1 ng intact K-562 total RNA in the preparation of the TruSeq library, for which we used the TruSeq RNA Sample Preparation Kit (version 1, Illumina) according to the manufacturer's protocol, except for that we used SuperScript III in first-strand cDNA synthesis and incubated the reaction at 50 °C instead of 42 °C.

Total RNA libraries. We used 1,000 ng of fragmented K-562 RNA prepared as described above to prepare a control library. We used the same cDNA synthesis and indexed Illumina library construction protocol described for the RNase H libraries, except that we performed six cycles of PCR.

We also used 1,000 ng each of the purified and DNase-treated FFPE and pancreas RNAs described above to prepare the control libraries using the same cDNA synthesis and indexed Illumina library construction protocol described for the RNase H libraries, except that we performed nine cycles of PCR for the FFPE sample and seven cycles for the pancreas sample.

Sequencing. We sequenced each of the K-562 cDNA libraries with an Illumina HiSeq2000 (76 base-paired reads, except SMART, which was 101 base-paired reads). We sequenced the FFPE and pancreas libraries on an Illumina HiSeq2500 (76 base-paired reads). All sequencing used version 3 flow cells and cluster chemistry. For indexed libraries, a third read of 8 bases was done as well. Sequence reads were binned by index read before further analysis. We used only PF reads for our analysis.

Library preprocessing. For the SMART library, reads were trimmed to 76 bases to match the other reads in this study. The amplified, long cDNAs contained adaptor sequences on both ends, and these sequences were present in reads originating from sheared cDNAs derived from the cDNA ends. We further trimmed reads before aligning them to the human genome or transcriptome so that we removed the specific adaptor-derived sequences (5′-AAGCAGTGGTATCAACGCAGAGTACTTTTTT TTTTTTTTTTTTTTTTTTTTTTTTT-3′ and 5′-AAGCAGTGG TATCAACGCAGAGTACATGGG-3′) present at the beginning of approximately one-third of the second of the paired-end reads.

Identification of rRNA reads. To calculate the percentage of reads originating from rRNA for each library, we aligned all reads to human rRNA (NR_003286.1, NR_003287.1, V00589.1, NR_003285.2, gi|251831106:648-1,601, gi|251831106:1,671–3,229) using BWA[21] (version 0.5.9-r16) in paired-end mode and a maximum edit distance of 0.04 (default). We marked reads in which both mates aligned to rRNA as rRNA reads, and we discarded them from further analysis. We also used BWA to align reads not marked as rRNA to unassembled contig GL000220.1, which contains the 45S rRNA transcript comprising the 18S, 5.8S and 28S rRNAs and two internal and two external transcribed spacer sequences[22]. We quantified the number of reads overlapping the 3′ external transcribed spacer region of the 45S rRNA (position 118,417–118,780).

Library mapping to the genome. We mapped all libraries to the human genome (hg19 including only chromosomes 1–22, X and Y and mitochondria) using Tophat[23] (version 1.3.3) with default parameters and without gene annotations. We removed unmated reads, retaining only read pairs in which both reads aligned to the genome. For each K-562 library, we then sampled 42.5 million of the remaining reads for all subsequent analyses (except for NuGEN 1f, for which we retained all 26.3 million reads that mapped as pairs). For the pancreas and FFPE kidney libraries, we sampled 12.8 million and 23.7 million reads, respectively.

Calculation of read-level metrics. After sampling equal numbers of genome-aligned reads, we generated alignment files for RNA-SeQC[12]. We sorted the alignment files by genomic coordinate, added read-group information and marked duplicate reads. We used RNA-SeQC (version 1.1.5) to calculate the following metrics: duplication rate, read alignment positions (exonic, intronic or intergenic), coefficient of variation, percentage of gene covered and coverage gaps (that is, ≥5 adjacent bases without coverage). The coverage gap definition is somewhat arbitrary but is consistent with our previous metrics[11]. We compiled the results in R (http://www.r-project.org/) for downstream visualization. We used the UCSC Genome Browser[24] knownGene transcript data set (version 05-Feb-2012) to annotate the genomic alignments.

Library complexity based on duplicate reads. We determined the number of duplicated read pairs using MarkDuplicates in Picard Tools (http://picard.sourceforge.net/). We defined duplicate reads as each having both mates aligned to the same position (primary alignments) in the genome. The duplication rate estimates the fraction of identical read pairs in a library due to the final PCR amplification step.

Continuity of coverage. We obtained from RNA-SeQC the percentage of each transcript that is not covered. We then converted from transcript to gene (locus) by aggregating each isoform's measures and weighing them equally (arithmetic mean). Next we plotted this mean coverage versus the mean expression level across all isoforms per gene and calculated in each plot the LOWESS fit for the data using R.

Library mapping to the transcriptome. To calculate metrics related to transcript expression levels, we aligned all non-rRNA reads directly to a transcriptome-based index. First we created a Bowtie alignment index based on the knownGene transcriptome annotations from the UCSC Genome Browser[24], adding 125-base poly(A) tails to the end of each known transcript sequence as previously described[25]. We then used Bowtie[26] (version 0.12.7) to align reads to this index, allowing at most two mismatches per read and using all other default parameters of RSEM[25]. We required each read to align with its mate pair to the same transcript. For each K-562 library, we then sampled 20 million of the aligned reads (except for NuGEN 1f, which had only 12 million of such mapped reads) and remapped those reads with RSEM using the same parameters as above but allowing an unlimited number of hits per read. We used this mapping in the subsequent analyses. These RSEM calculations took between 4 and 12 h of CPU time on a server using ten threads with four cores, with each library taking about 1–2 GB of RAM. Similarly, for the pancreas and FFPE kidney libraries, we sampled 6.6 million and 6.7 million aligned reads, respectively.

Coverage relative to 5′ and 3′ ends. Using the Bowtie alignments to the transcriptome, we examined read coverage at the ends and along the length of the transcripts. We used the knownGene annotation from the UCSC Genome Browser[24] to obtain transcription start and end sites. We defined a transcript start or end to be covered if at least one read mapped to the first or last annotated 100 bp of the transcript, respectively. We excluded transcripts not expressed in the control Total library. To avoid using reads mapping to the artificial poly(A) tail in the transcriptome index (see above), we removed all reads containing more than 40 consecutive A bases. After removing these reads, we found that SMART still contained a large number of reads mapping to the poly(A) tail, in part because some of the SMART reads derived from the 3′ end of transcripts, once trimmed of their adaptors, were only 21 bases long. Therefore, we also removed reads with at least 21 consecutive A bases from the SMART library. Although these reads comprised 3.4% of the remaining reads in the SMART library, they were <0.1% of the remaining reads in the other libraries. Finally, for transcripts that had more than 500 mapped reads, we divided each transcript into 100 equally sized length bins from 5′ to 3′ end. For each bin, we calculated the relative coverage compared to the coverage for the entire transcript. We calculated each metric for all transcripts as well as in transcript groups based on length.

Expression quantification and related metrics. We used the read alignments to the UCSC knownGene transcriptome as input to RSEM[25] (version 1.1.17). RSEM calculates expression estimates that are corrected for isoform expression. RSEM-produced tau values were multiplied by 1,000,000 and are analogous to fragments per kilobase per million sequenced reads (FPKM) estimates. We refer to these expression values as 'TPM' and used them for all subsequent analysis.

To calculate expression correlations, we first added 1 to each expression value and then used the $\log_{10}$ of this sum as input (to give more equal weight to genes with lower expression values) to calculate the Pearson correlation between each of the libraries using R (version 2.14.2) and ggplot2 (ref. 27).

For scatter, *MA* and Q-Q plots, we compared two sets of expression-level data points (D_1, D_2). Scatter plots show the $\log_{10}(D_1 + 1)$ versus $\log_{10}(D_2 + 1)$. *MA* plots show $\log_{10}(D_1) + \log_{10}(D_2)$ versus $\log_{10}(D_1) - \log_{10}(D_2)$. Q-Q plots show a quantile-quantile plot of D_1 versus D_2. We removed the top ten most highly expressed transcripts from each of our Q-Q plots, defined on a per-sample basis.

For spike-in RNA analysis, we aligned the reads for each library from a single lane of sequencing to the ERCC reference sequences (Ambion) with BWA[21]. We then used a custom Picard module to parse the aligned BAM file to calculate the coverage (reads out of total reads) for each spike-in sequence. For each library, we calculated the Pearson coefficient with the $\log_{10}$ (coverage + 1×10^{-8}) relative to the $\log_{10}$ (stated input amounts + 1×10^{-8}) (Ambion; **Supplementary Table 7**).

Number of genes covered. We estimated the number of genes detected by each library out of those genes (loci) that had at least one transcript detected at a TPM threshold of 0.1 in the control Total library. We calculated the number of genes detected at different TPM thresholds, showing that the relative ranking of methods is largely robust to the specific threshold.

Binning GC and length analysis. We aggregated the results from RNA-SeQC and the expression values produced by RSEM using R. We then calculated the mean of each tested measure (Pearson correlation, 5′-to-3′ coverage and ends and expression plots) for bins of specific GC content or length of the transcript/gene. The bins for GC content were 0% to ≤37%, >37% to ≤62% and >62 to 100%. The length bins were 1–999, 1,000–5,000 and >5,000 bases or 1–1,000, 1,001–5,000 and >5,000 bases, as described in each figure.

21. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
22. Maden, B.E. *et al.* Clones of human ribosomal DNA containing the complete 18 S-rRNA and 28 S-rRNA genes. Characterization, a detailed map of the human ribosomal transcription unit and diversity among clones. *Biochem. J.* **246**, 519–527 (1987).
23. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
24. Dreszer, T.R. *et al.* The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res.* **40**, D918–D923 (2012).
25. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
26. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
27. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis (Use R!)* (Springer, New York, 2009).

doi:10.1038/nmeth.2483

Corrigendum: Comparative analysis of RNA sequencing methods for degraded or low-input samples

Xian Adiconis, Diego Borges-Rivera, Rahul Satija, David S DeLuca, Michele A Busby, Aaron M Berlin, Andrey Sivachenko, Dawn Anne Thompson, Alec Wysoker, Timothy Fennell, Andreas Gnirke, Nathalie Pochet, Aviv Regev & Joshua Z Levin

***Nat. Methods* 10, 623–629 (2013); published online 19 May 2013; corrected after 2 December 2013**

In the version of this article initially published, in the Online Methods "RNase H libraries" section, the sentence beginning with "We added 5 µl preheated RNase H…." should have read "We added 5 µl preheated RNase H reaction mix that contains 10 U of Hybridase Thermostable RNase H (Epicentre), 0.5 µmol Tris-HCl, pH 7.5, 1 µmol NaCl and 0.2 µmol $MgCl_2$ to the RNA and DNA oligo mix, incubated this mixture at 45 °C for 30 min and then placed it on ice." The errors have been corrected in the HTML and PDF versions of this article.