Electronic Supplementary Information

Simultaneous RNA purification and size selection using an on-chip isotachophoresis with ionic spacer

Crystal M. Han[1,2], David Catoe[2], Sarah A. Munro[2,3], Ruba Khnouf[4,5], Juan G. Santiago[5], Michael P. Snyder[6], Marc L. Salit[2^], Can Cenik[6,7^]

^co-corresponding authors

Affiliations:

[1] Department of Mechanical Engineering, San Jose State University, San Jose, CA 95192, USA

[2] Joint Initiative for Metrology in Biology, National Institute of Standards and Technology, Stanford, CA 94305, USA

[3] Minnesota Supercomputing Institute, University of Minnesota, MN, 55455 USA

[4] Department of Biomedical Engineering, Jordan University of Science and Technology, Irbid, Jordan

[5] Department of Mechanical Engineering, Stanford University, Stanford, CA 94305, USA

[6] Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA

[7] Department of Molecular Biosciences, University of Texas at Austin, Austin, TX 78705, USA

Contents

Figure S10 Log ratio M values from Figure S9 plotted as a function of transcript length.
Figure S11 The standard deviation versus mean of $\log_2$ read counts from three replicates of ITP and gel electrophoresis

Table S1 Outlier gene IDs (included as supplementary .xlsx file)

Video S1 Visualization of ITP experiment (included as supplementary .mp4 file)

Section S1 ITP extraction with MNase-digested cell lysate

Chronic myelogenous leukemia cells (K562) were grown at 37 ˚C and 5% CO_2 in RPMI-1640. 10% fetal bovine serum (FBS) and 1% Penicillin-Streptomycin-Glutamine (PSG) were added to the culture medium. Approximately 0.33 million cells were pelleted by centrifugation at 1000 g for 10 min and washed with 1X phosphate buffer saline (PBS) solution. Finally, cells were pelleted once more and PBS was aspirated. 25 µl cell lysis buffer was added to the pellet and homogenized repeatedly by pipetting. Lysis buffer consisted of 5 mM HCl, 35 mM Bis-Tris, 5 mM KCl, 5 mM $MgCl_2$, 5 mM $CaCl_2$,1 mM Dithiothreitol (DTT), 0.5% polyvinylpyrrolidone (PVP), 1% (v/v) Triton X-100, 25 U/ml TurboDNase. $CaCl_2$ is required for activity of micrococcal nuclease (MNase). After incubation in lysis buffer for 5 min, the lysate was centrifuged at 1300 g for 10 min and the supernatant was recovered. 100 U of MNase was added to the supernatant, and the sample was incubated at 37 ˚C for 30 min following protocol suggested by Reid et al.[1]At the end of incubation, Alexa Fluor 488 and DyLight 488 were added such that the final concentrations are 250 nM and 750 nM respectively. In addition, 0.019 g of Urea was dissolved in the lysate to achieve 8 M Urea concentration in total 40 μl of digested cell lysate. The sample was incubated for 10 min at 95˚C, and immediately cooled on ice at the end of incubation. The prepared digested cell lysate was kept on ice until loading the chip (within 30 min). RNA from the sample was extracted following the same ITP protocol described in Materials and Methods. The Fraction 2 of each ITP experiment was collected and stored in -200 ˚C until quantified using bioanalyzer with Agilent 2100 Bioanalyzer System small RNA Analysis Kit. The RPMI-1640 and TurboDNase were purchased from Thermo Fisher Scientific Inc., and SUPERase In RNase inhibitor was purchased from Invitrogen.

Section S2 Sequencing read length analyses

Here we include analyses and discussions on sequencing read lengths for the sequencing results presented in Figure 5. In Figure S6, we present size distribution of all sequencing reads after the removal of the adapter sequence. Both methods had specific peaks (predominantly at 23nt and 35-36 nts) that corresponded to rRNA contaminants as is typical in other ribosome profiling data. ITP samples had a higher proportion of reads that were shorter than 20 nt, which were removed bioinformatically from further analyses. In contrast, the conventional gel electrophoresis method had a higher fraction of reads corresponding to a specific 35-nucleotide rRNA fragment.

Figure S7 shows size distribution of the reads that aligned to the transcriptome for the two methods. We found that sequencing libraries from neither method contained appreciable number of reads longer than 35 nt (<3% for ITP; <2% for gel electrophoresis). While our analyses utilized the entire range of footprints, the removal of these sequences >35 nt would not affect the current analyses and conclusions.

Section S3 Identification of transcripts with the largest deviations between the two methods

To systematically identify the transcripts with the largest deviations between the quantifications from the two methods, namely outliers from Figure 5b, we generated an MA-plot as shown in Figure S9. We grouped transcripts into 50 bins by their A-values. Following the typical method for outlier detection, we calculated for each bin 1.5 times the inter-quartile range of M-values. We then used a local polynomial regression of the calculated values against the mean A-value for each of the 50 bins. These lines are shown in red in Figure S9. We then defined all transcripts

that fall outside of these lines as "outliers" or simply the transcripts with the largest deviation between the two methods. In total, there were 230 such transcripts. A list of these transcripts is provided in Table S1

Next, we analyzed the functional annotations to characterize whether these outliers had any systematic differences. While transcripts that had higher counts in the conventional method were not enriched in any functional category, we found that chromatin associated transcripts such as histones had higher signal in the ITP-based methods (Table S1). Interestingly, these histone transcripts are known to be both highly translated and have short open reading frames. The most likely explanation for this difference is due to the capture of the 21-nt ribosome footprint fragments in ITP but not in the conventional method. Recent studies have revealed the presence of ~21 nt ribosome footprints representing mRNA fragments protected by a different conformation of the ribosome.[2,3] Importantly, these shorter ribosome footprints were missed by the typical gel extraction method with a pre-defined size limit.[2,3] The more permissive size range in our ITP experiment was able to capture footprints from both conformations of the ribosome as shown with the bimodal distribution of transcriptome mapping reads in ITP libraries (Figure S7). We realize that different size markers can be used in gel electrophoresis to capture these fragments as well. However, gel electrophoresis method used in the current paper reflects the predominant ribosome profiling protocol in the field.[4]

Furthermore, we analyzed whether there is a systematic difference between the two methods in ribosome occupancy measurement as a function of transcript length (Figure S10). We found that there is a statistically significant yet very weak relationship suggesting that ITP-based approach had higher quantifications for shorter transcripts (Spearman correlation $\rho = 0.12$; p-value < 0.01).

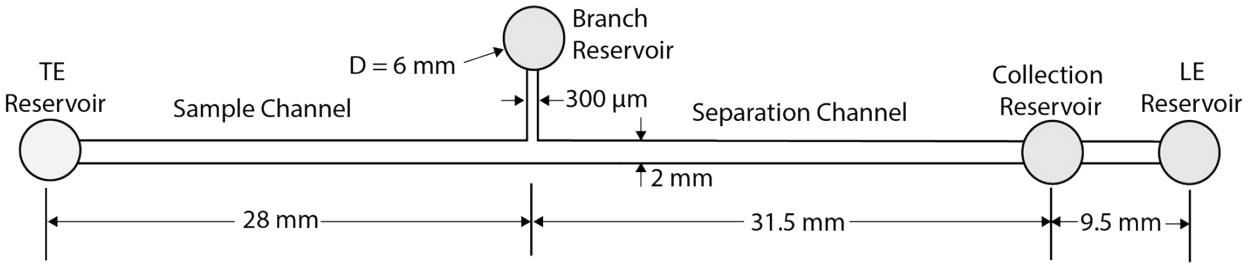


Figure S1 Design and specifications of the chip. Length and width dimensions of channel and reservoir diameters are shown. The height of the channel was 300 µm. The sample channel section allowed loading of 17 µl volume of sample. To prevent loss of RNA into branch channel, the branch channel width was designed to be >6 fold smaller than the main channel.

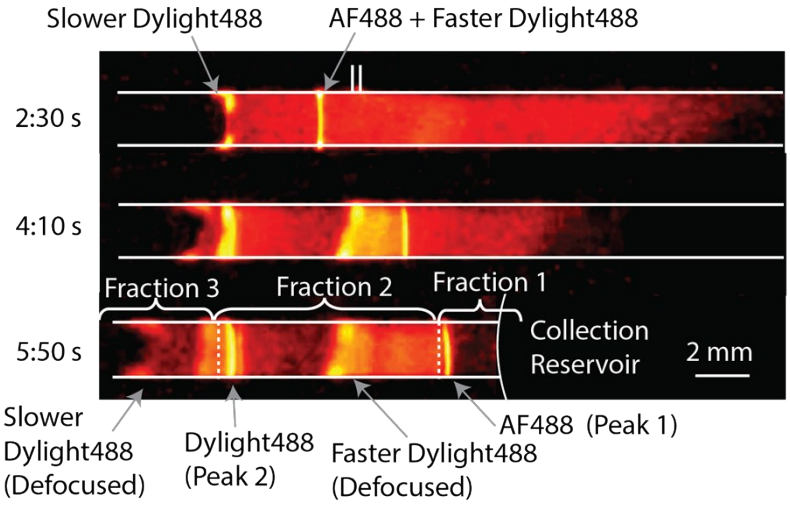


Figure S2 Fluorescence visualization of the ITP-based size-selection process at three times using AF488 and DyLight 488. Snapshots were taken at designated times from the supplementary video (Video S1). Initially, a sample containing two fluorescence dyes was loaded in the sample channel. Upon applying electric field at $t = 0$, ITP zones started migrating from the inlet of the sample channel, collecting and focusing dye molecules and RNAs. At $t = 2{:}30$ s, we observed two distinct ITP peaks separated by a growing spacer zone migrating towards the LE reservoir. The first peak focused AF488 and the faster fraction of DyLight488, while the second ITP peak contained the slower fraction of DyLight488. In the separation channel, faster fraction of DyLight488 defocused from peak 1 and migrated towards peak 2. Collection of sample fraction 1, 2, and 3 was determined by locations of peak 1 and peak 2. Downstream analysis of the three collected fractions showed that single nucleotides were co-focused with the AF488 in peak 1 (collected in Fraction 1), and RNAs in the size range of 2 - 35 nt were located in between the two peaks and at peak 2 (collected in Fraction 2). Longer RNAs were mostly included in Fraction 3.

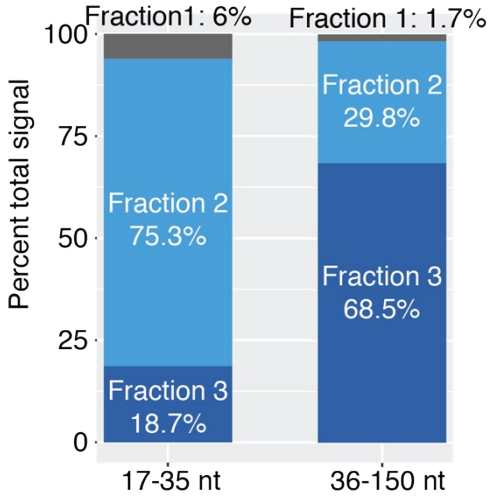


Figure S3 Average percentage of RNA included in each fraction for two size ranges; 17-35 nt and 36-150 nt, calculated from bioanalyzer data of three replicates of ITP size range selection experiments (the bioanalyzer data shown in Figure 3). The sum of signal from three fractions constitutes 100% in each size range.

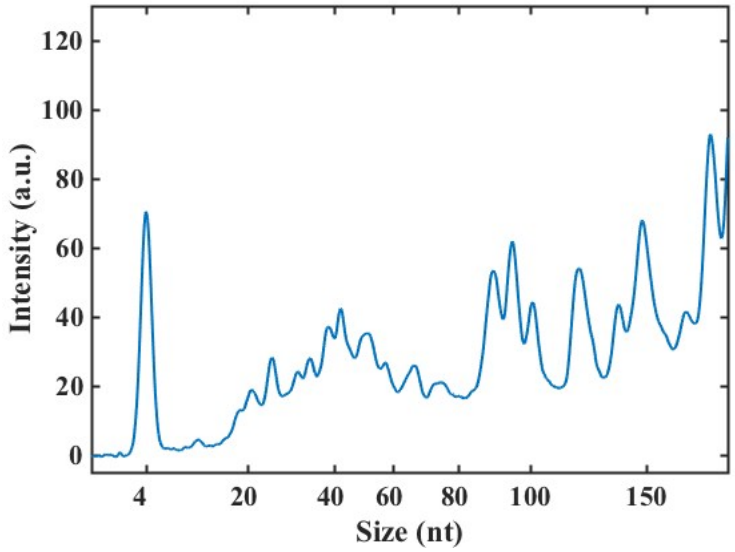


Figure S4 Bioanalyzer electropherogram of LCL RNA sample after miRNesasy kit and before ITP. The peak at 4 nt corresponds to a marker RNA associated with the bioanalyzer process. The Bioanalyzer trace shows the sample included all sizes of smRNAs bigger than ~18 nt, which makes it appropriate to test the longer RNA exclusion (shown in Figure 3 of main text).

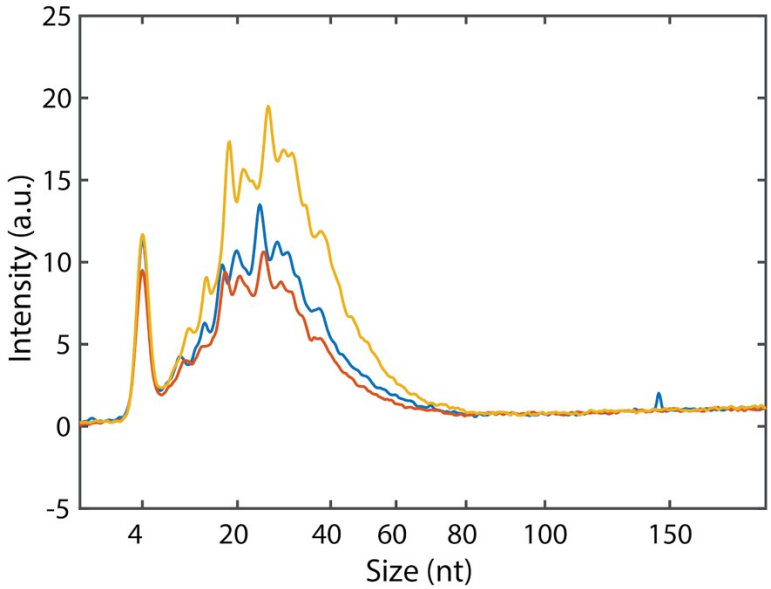
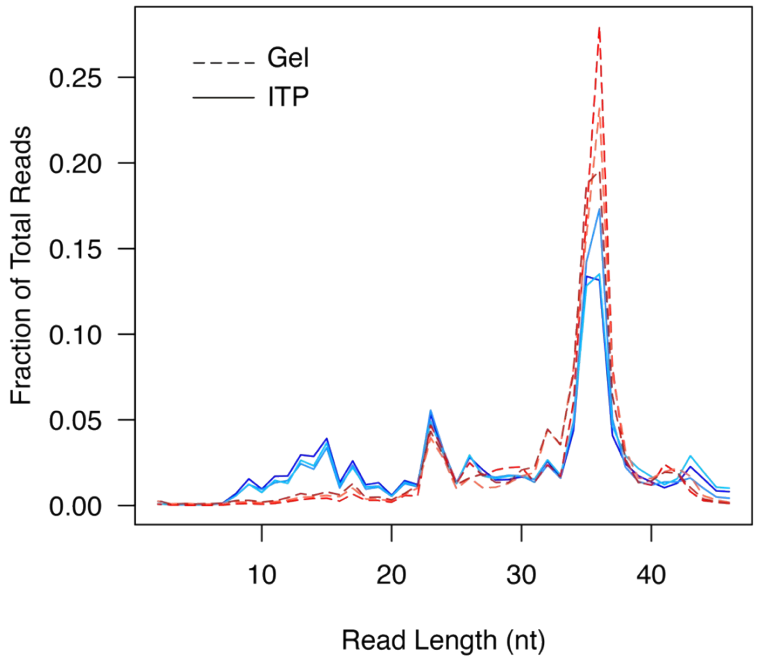


Figure S5 Bioanalyzer electropherograms of RNA extracted from digested cell lysate using ITP size-range selection method. Each line indicates each replicate of size-selection experiments. The peaks at 4 nt indicates a marker added from the bioanalyzer process.



Figures S6 Distribution of read lengths in sequencing libraries prepared by ITP (blue solid lines) and gel electrophoresis (red dashed lines) methods. Different colors indicate different replicates. We sequenced reads using a single end 50 nucleotide chemistry. The ITP samples are shown in shades of blue with solid lines. The gel electrophoresis samples are in shades of red and dashed lines. The fraction of total reads after adapter removal was plotted as a function of read length.

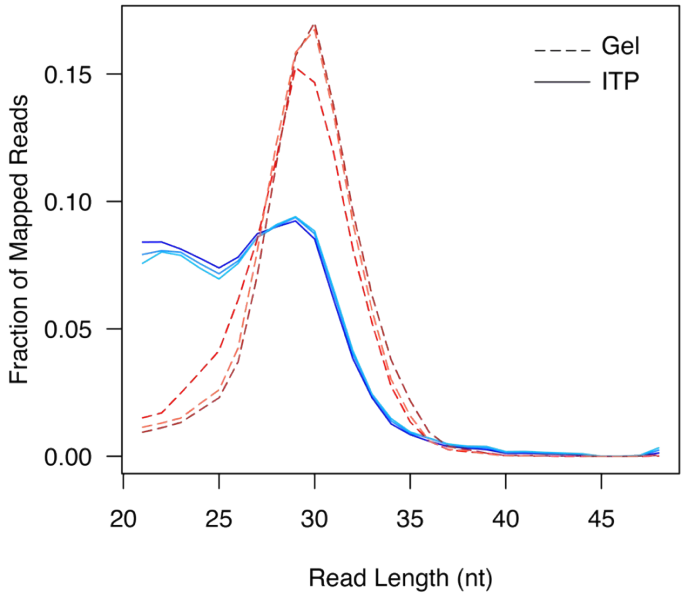


Figure S7 Distribution of read lengths in transcriptome-mapping reads from ITP-extracted RNA (solid lines) and gel electrophoresis-extracted RNA (dotted lines). Different colors indicate different replicates. The fraction of transcriptome-mapping reads after adapter removal was plotted as a function of read length.

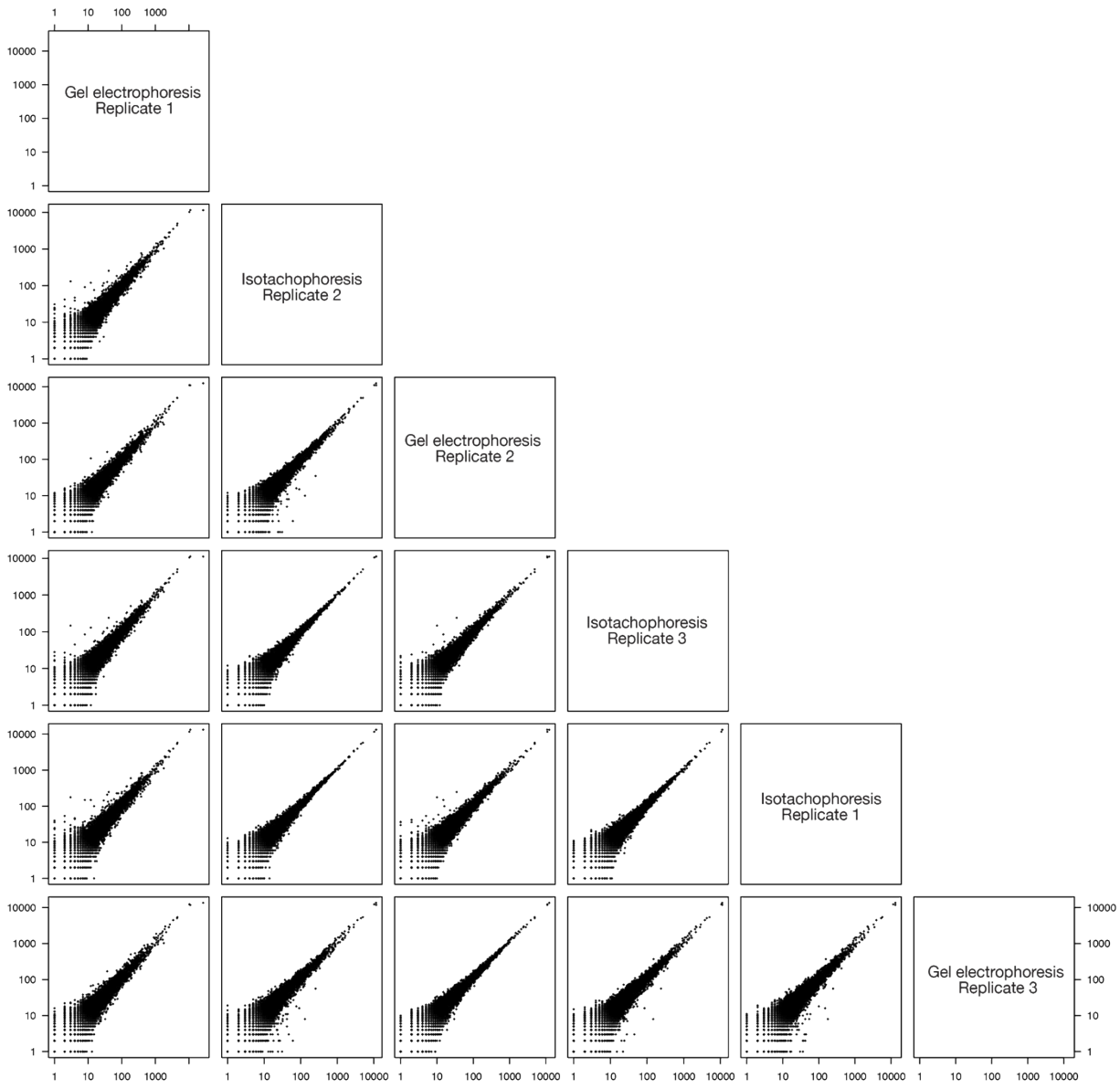


Figure S8 Pair-wise comparison of read counts per million reads (cpm). Three replicates were generated from RNAs extracted by ITP or gel-electrophoresis methods. The number of sequencing reads for each transcript is calculated and compared for all pairs of experiments. Spearman rank correlations ranged from 0.93 to 0.95.

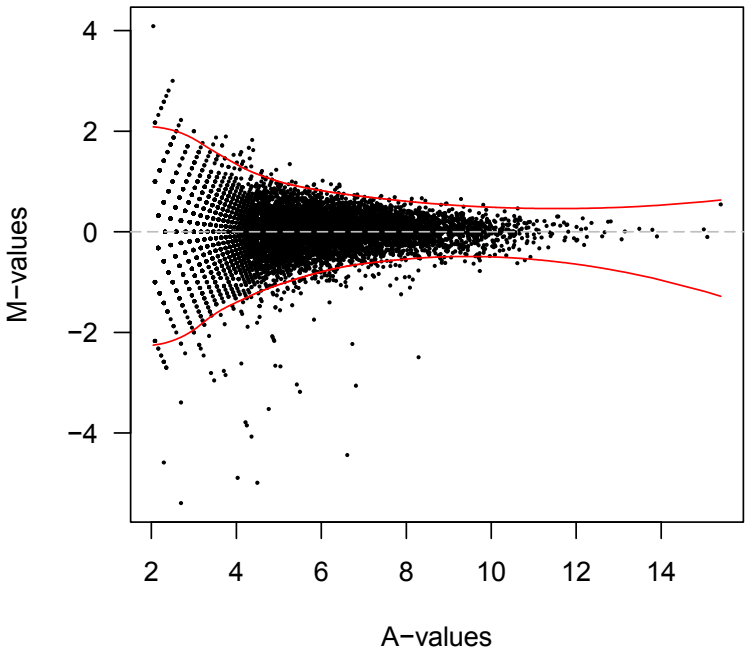


Figure S9 MA-plot generated with reads mapped to transcripts from both methods. Log ratio M is defined by log2 of ratio of gel electrophoresis to ITP methods. A-values represent the log2 average of gel electrophoresis and ITP methods. Red lines represent polynomial regression of the 1.5 times inter-quartile range of M-values against the mean A-values. Total 230 transcripts that fell outside of these lines are identified as outliers and summarized in Table S1.

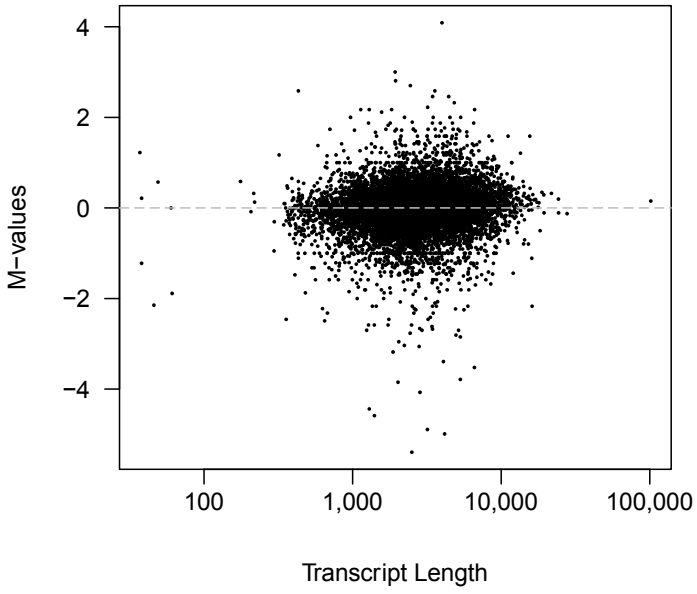


Figure S10 Log ratio M values plotted as a function of transcript length. Very week correlation between the M-value and the transcript length is observed (Spearman correlation $\rho = 0.12$; p-value < 0.01)

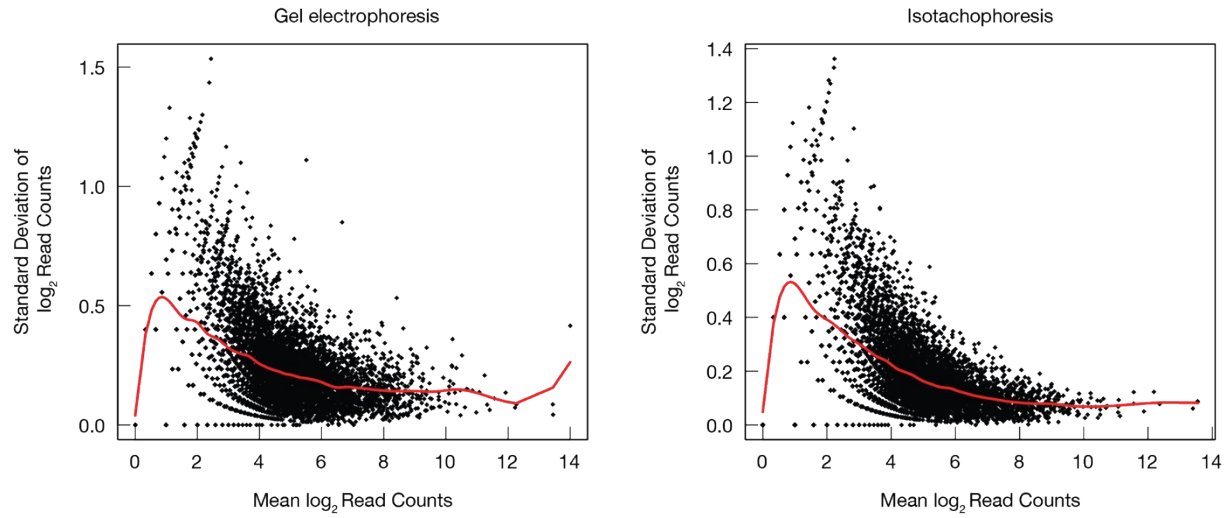


Figure S11 The standard deviation and mean of $\log_2$ read counts per each transcript was calculated and from three replicates of sequencing libraries generated from size-selected RNAs using ITP-based and electrophoresis-based extraction. Each dot represents one transcript and the red lines correspond to the best-fit cubic spline to the data.

Table S1 Outlier gene IDs (included as supplementary .xlsx file)
On Sheet 1 of the file, we present all gene IDs of outliers that are categorized due to high expression in gel method. On Sheet 2 of the file, we include all gene IDs of outliers identified due to high expression in ITP method. On Sheet 3 of the file, we include Gene ontology functional term annotations for transcripts that are deemed outliers in the ITP method. The enrichment analysis was done using FuncAssociate.[5]

Video S1 Visualization of ITP experiment (included as supplementary .mp4 file)
Under electric field, ITP zones migrated from the inlet of the sample channel, collecting and focusing dye molecules and RNAs. In the sample channel, two distinct ITP peaks formed and were separated by a spacer zone growing in length. Note that RNA was not separated by the target size range at this stage. Once ITP zones entered into the separation channel containing the sieving matrix, we observed re-distribution of dyes (visualized by fluorescence signal) and RNAs (verified by bioanalyzer) because mobilities of the dyes and RNAs were significantly decreased by the sieving matrix. We empirically found out that DyLight488 species included molecules with a wide range of mobilities, which we attributed to partially degraded NHS esters conjugated to the dyes. As a result, a faster DyLight488 fraction that was previously co-focused with AF488 in the first peak defocused in the separation channel and slowly migrated towards the second peak. We also observed defocusing of a slower DyLight488 fraction from the second peak. The dyes served as markers for deciding the collection time of Fraction 1 (single nucleotides) and Fraction 2 (~2-35 nt RNA fragments) from the collection reservoir.

Supplementary References

1 D. W. Reid, S. Shenolikar and C. V. Nicchitta, *Methods*, 2015, **91**, 69–74.
2 L. F. Lareau, D. H. Hite, G. J. Hogan and P. O. Brown, *Elife*, 2014, **3**, e01257.
3 C. C.-C. Wu, B. Zinshteyn, K. A. Wehner and R. Green, *Molecular Cell*, 2019, 73, 959–970.e5.
4 N. T. Ingolia, *Cell*, 2016, **165**, 22–33.
5 G. F. Berriz, J. E. Beaver, C. Cenik, M. Tasan and F. P. Roth, *Bioinformatics*, 2009, **25**, 3043–3044.