Supplementary Material For:

cDNA normalization by hydroxyapatite chromatography to enrich transcriptome diversity in RNA-seq applications

Victoria A. VanderNoot[1], Stanley A. Langevin[2], Owen D. Solberg [2], Pamela D. Lane [2], Deanna J. Curtis [2], Zachary W. Bent[2], Kelly P. Williams[2], Kamlesh D. Patel[3], Joseph S. Schoeniger[2], Steven S. Branda[1], and Todd W. Lane[2]

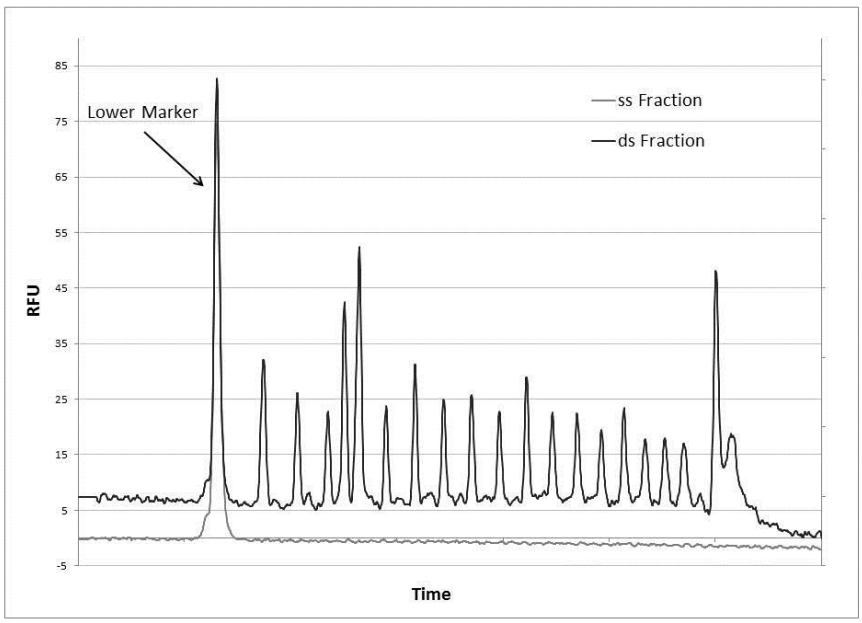
[1]Biotechnology and Bioengineering Department, Sandia National Laboratories, Livermore CA, USA, [2]Systems Biology Department, Sandia National Laboratories, Livermore CA, USA, and [3]Advanced Systems Engineering & Deployment Department, Sandia National Laboratories, Livermore CA, USA

Quality filter

Raw reads were processed by Qfilter, a custom pre-processing Perl script. Low-quality tails were trimmed from reads in the manner of bwa (1), library primer sequences were then trimmed from reads, and low-complexity sequence was masked with DUST (2). Reads were then rejected if length was < 30 nt, if ≥ 3 ambiguous bases were present, or if average positional quality score was < 30. Low-complexity sequence was then unmasked for the retained reads.

Genomic alignment

Following end-trimming and quality filtering, Bowtie2 (version 2.0.0-beta7) was run in sensitive mode with default settings to produce the alignment BAM files that were subsequently used for counting reads mapping to rRNA and other genomic compartments, for plotting coverage depth across the *E. coli* genome, and for plotting the average coverage of the 5' and 3' ends of genes. We used Tophat (version 2.0.4) to generate the transcript mappings and Cufflinks (version 2.0.2) to generate transcript abundance estimates that were subsequently used to make gene expression scatterplots and log-fold enrichment plots. Both TopHat and Cufflinks were run with default settings.



Supplementary Figure 1. Bioanalyzer analysis of ssDNA and dsDNA fractions separated by micro-column hydroxyapatite chromatography. A sample of 25bp dsDNA ladder was injected onto the HAC column and eluted first with 100 mM sodium phosphate pH 7.0, 0.005% (w/v) SDS (ssDNA; lower trace) and then 320 mM sodium phosphate pH 7.0, 0.005% (w/v) SDS (dsDNA; upper trace, offset for clarity). (NB: the Lower Marker is a migration standard added during Bioanalyzer analysis and does not represent DNA present in either fraction)

Genomic annotations

To count hits to different genomic regions, all reads were first aligned to a Bowtie index containing only rRNA sequences; whole-genome alignments were carried out on those reads which did not align to the rRNA index. For each of the other genomic regions, we created a set of BED files with the "merge" and "subtract" functions of BEDTools. To transform the

Supplementary Table 1. Reproducibility and carry over data for micro-columns.

Cartridge Volume (µL)	Elution Reproducibility	Carryover		
		ΔCt First Cycle	ΔCt Second Cycle	ΔCt Third Cycle
30	10.2	7.5	9.9	10.8
6	12.5	9.7	11.8	12.1
2.5	9.6	11.0	12.5	13.0

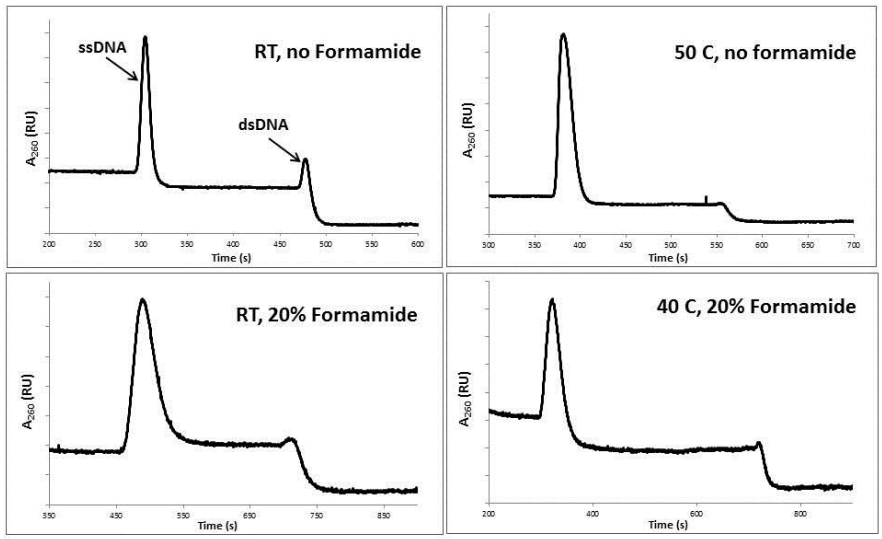
Reproducibility and carry over data for large columns (30 µL), medium (6 µL) and small columns (2.5 µL). Reproducibility was assessed using replicate samples (n=4) of ssDNA (56 mer) and concentration in the eluted fraction was quantified using OliGreen. In order to assess carry over, samples of PBMC cDNA (5-10 ng/µL) were applied to the columns and eluted normally. The columns were then flushed sequentially with cycles of wash buffer and dsDNA elution buffer. The elution buffer was collected and ds-cDNA was quantified using qPCR.

UCSC genome annotations into a set of dove-tailed, non-overlapping BED files for mitochondria, miRNA, tRNA, lincRNA, exon, intron, and intergenic, in that order of priority (i.e. a region corresponding to an intronic microRNA was assigned only to the microRNA category).

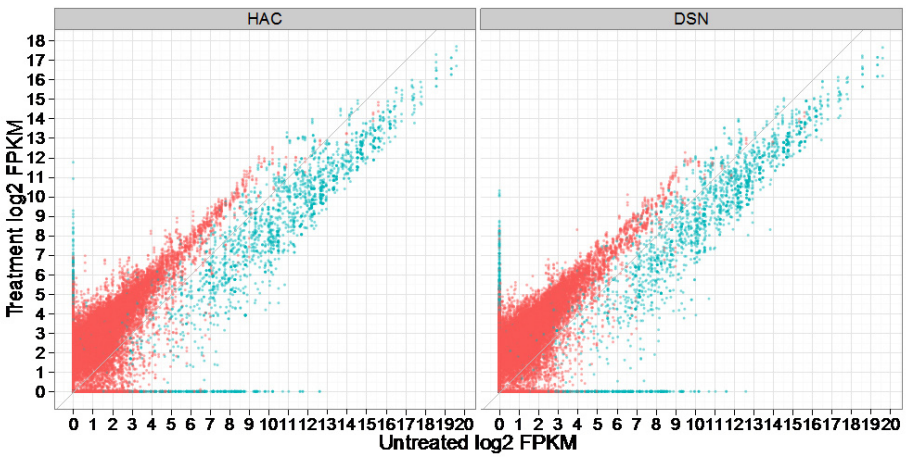
The GTF files used by TopHat/Cufflinks were provided by the UCSC table browser (refSeq GTF file for *E. coli* K12 genome) and by Ensemble FTP site (Homo_sapiens.GRCh37.68.gtf for the human genome).

References

1. **Li, H. and Durbin, R.** 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*:1754-1760.
2. **Morgulis, A., Gertz, E.M., Schaffer, A.A., and Agarwala, R.** 2006. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. Journal of computational biology : a journal of computational molecular cell biology *13*:1028-1040.



Supplementary Figure 2. Minimization of secondary structure formation in microcolumn HAC. Top left: HAC of 193-mer ssDNA Ultramer carried out at room temperature; Top right: HAC of 193-mer ssDNA Ultramer carried out at 50°C; Lower left: HAC of 193-mer ssDNA Ultramer carried out at room temperature with addition of 20% formamide in the sample and buffers; Lower right: HAC of 193-mer ssDNA Ultramer carried out at 40°C with addition of 20% formamide in the sample and buffers.



Supplementary Figure 3. Scatterplot analysis of human PBMC small RNA populations following HAC and DSN normalization. Normalized RNA transcript abundance (FPKM) was estimated by Cufflinks. Each point represents one transcript, plotted on a log2 scale. The y-axis shows suppressed FPKM (y-axis) and the x-axis shows the untreated FPKM. Red plots represent large RNA transcripts (>200bps) and blue plots represent small RNA transcripts (<200bps). Normalized RNA transcripts plotted above the diagonal line are enriched and transcripts below the line are suppressed when compared to the mean FPKM values of the untreated controls.

Supplementary Table 2. Summary of sequencing data generated from *E. coli* and PBMC RNA-Seq libraries. Table 2A represents qfilter and transcriptome mapping results from untreated and treated *E. coli* K12 RNA-Seq libraries followed by Table 2B which represents bioinformatic analysis of the untreated and treated human PBMC libraries.

A. *E. coli* K-12

Sample	Raw Reads	Rejected Reads	*E. coli* rRNA	*E. coli* genome & transcriptome	Unmapped
Untreated 1	14,137,037	4,570,001	8,005,693	1,353,935	207,408
Untreated 2	13,970,325	4,519,719	7,941,159	1,308,979	200,468
Untreated 3	14,461,694	4,831,840	8,075,140	1,346,148	208,566
HAC 1	13,227,536	2,450,550	1,370,368	8,725,309	681,309
HAC 2	13,628,118	2,420,848	1,352,615	9147099	707,556
HAC 3	15,494,992	2,989,655	1,973,805	9,787,730	743,802
Ribo-Zero 1	10,825,105	5,544,222	1,024,867	3,557,054	698,962
Ribo-Zero 2	11,952,941	6,331,090	1,079,733	3,777,355	764,763
Ribo-Zero 3	15,465,465	8,353,732	1,362,988	4,733,690	1,015,055

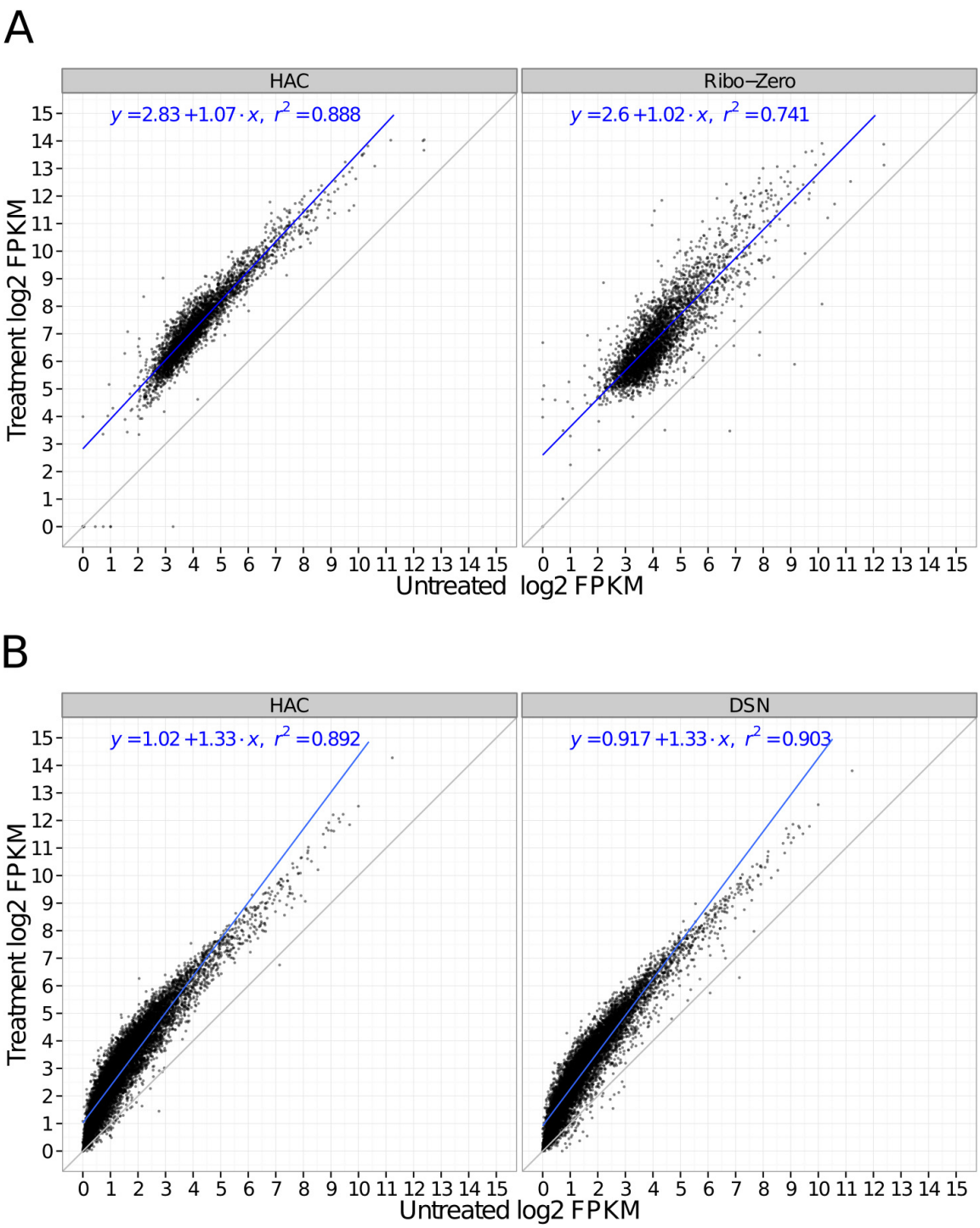
B. Human PBMCs

Sample	Raw Reads	Rejected Reads	Human rRNA	Human genome & transcriptome	Unmapped
Untreated 1	12,776,856	5,020,607	6,416,512	1,113,830	225,907
Untreated 2	9,005,971	3,577,432	4,445,705	806,455	176,379
Untreated 3	12,177,174	4,679,298	5,903,241	1,358,643	235,992
HAC 1	11,430,776	2,638,654	1,577,365	6,617,938	596,819
HAC 2	15,668,551	4,123,264	3,041,321	7,778,734	725,232
HAC 3	11,070,190	3,048,456	2,536,710	5,006,262	478,762
DSN 1	16,111,749	4,455,326	4,683,780	6,338,706	633,937
DSN 2	12,092,140	3,267,228	3,519,024	4,822,690	483,198
DSN 3	14,204,529	3,370,205	2,840,942	7,277,512	715,870

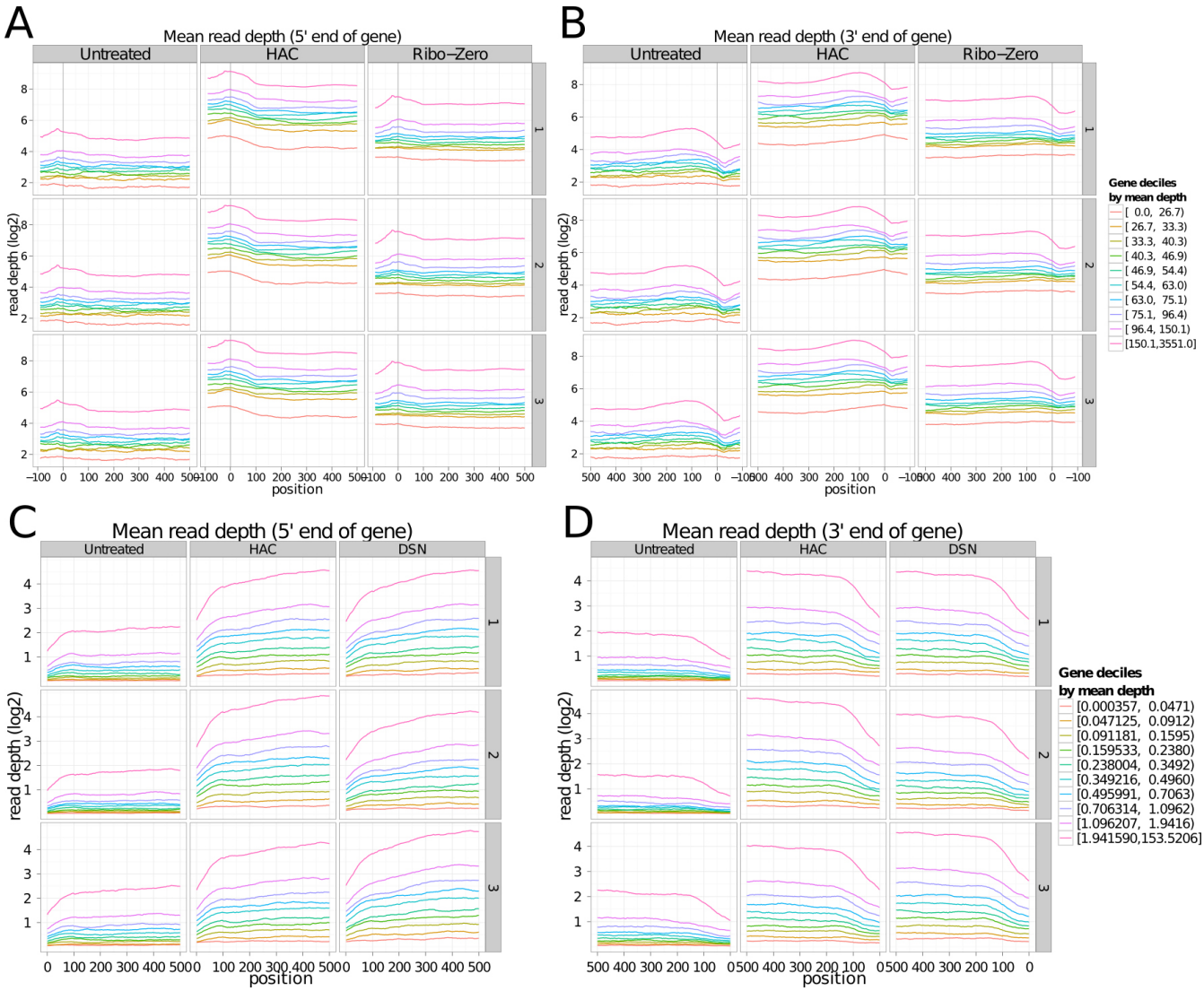
Supplementary Table S3. Analysis of tRNA and rRNA reads.

	Untreated	DSN	HAC
Reads (millions)	6.99±1.29	10.48±1.46	9.50±1.86
tRNA/read (%)	0.319±0.031	0.095±0.028	0.122±0.028
Sense/tRNA (%)	99.784±0.024	99.445±0.104	99.371±0.116
Mature/sense tRNA (%)	89.531±0.336	89.238±0.517	89.707±0.363
Mature/antisense tRNA (%)	81.0±13.4	37.7±10.9	39.8±8.0
rRNA/read (%)	77.1±2.1	33.3±7.5	23.8±6.5
Sense/rRNA (%)	99.734±0.008	99.814±0.008	99.767±0.038

Genomic coordinates for 511 nucleus-encoded human mature tRNA sequences were taken from the UCSC genome browser (excluding those marked as pseudogenes or undetermined identity) and all reads (post-quality-filtering) that mapped with any overlap to these regions were collected. Mapping was performed without accounting for tRNA splicing, which occurs occurs for 31 of these genes. tRNA hits were segregated into sense vs. antisense categories, and the fraction of mature reads (where both read ends fall completely within the tRNA mature coordinates) was measured for both sense and antisense fractions. For rRNA hits the fraction in the sense orientation was measured.



Supplementary Figure 4. Scatterplot analysis of *E. coli* and human PBMC gene-coding RNA transcripts following HAC, DSN normalization, or RiboZero treatments. The points represent gene-coding RNA transcripts of length > 200 bp and with FPKM > 0 (as estimated by CuffLinks) in both the untreated and treated RNA-Seq libraries. FPKM values are plotted in log2 scale. (A) represents *E. coli* RNA seq libraries treated by HAC normalization or RiboZero rRNA depletion methods compared to untreated control based on mean abundance (FPKM). (B) represents human PBMC RNA-Seq libraries following HAC or DSN normalization.



Supplementary Figure 5. Coverage evenness at the 5′ and 3′ terminal ends of gene coding RNA transcripts from *E. coli* and PBMC RNA-Seq libraries. Mean coverage, plotted on a log(2) scale, at the 5′ and 3′ ends of gene transcripts from *E. coli* (A and B) and human PBMC (C and D) RNA-Seq libraries. To make this plot, protein-coding transcripts between 1 and 5 kb were selected from the UCSC K12 annotation (*E. coli*) or the Ensembl GRCh37.68 annotation (human PBMC). For genes with multiple transcripts, only the most highly expressed transcript was selected. Selected transcripts were sorted into deciles based on expression level in the untreated samples. SAMtools mpileup was used to tabulate the depth at each nucleotide position of each selected transcript. Then, for each of the first 500 nucleotide positions at the 5′ end the mean depth was calculated. The same calculation was made at the 3′ end after right-aligning all transcripts. For the *E. coli* plots (A and B), the analysis included an extra 100 bp beyond the annotated end of the transcript, to better capture the decay in coverage.