Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes

Katsuyuki Shiroguchi, Tony Z. Jia, Peter A. Sims, and X. Sunney Xie[1]

Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138

Edited by Jonathan S. Weissman, University of California, San Francisco, CA, and approved December 14, 2011 (received for review November 1, 2011)

RNA sequencing (RNA-Seq) is a powerful tool for transcriptome profiling, but is hampered by sequence-dependent bias and inaccuracy at low copy numbers intrinsic to exponential PCR amplification. We developed a simple strategy for mitigating these complications, allowing truly digital RNA-Seq. Following reverse transcription, a large set of barcode sequences is added in excess, and nearly every cDNA molecule is uniquely labeled by random attachment of barcode sequences to both ends. After PCR, we applied paired-end deep sequencing to read the two barcodes and cDNA sequences. Rather than counting the number of reads, RNA abundance is measured based on the number of unique barcode sequences observed for a given cDNA sequence. We optimized the barcodes to be unambiguously identifiable, even in the presence of multiple sequencing errors. This method allows counting with single-copy resolution despite sequence-dependent bias and PCR-amplification noise, and is analogous to digital PCR but amendable to quantifying a whole transcriptome. We demonstrated transcriptome profiling of *Escherichia coli* with more accurate and reproducible quantification than conventional RNA-Seq.

systems biology | digital counting | next generation sequencing | gene expression | genomics

The central goal of transcriptome profiling is to accurately quantify the abundance of RNA transcripts in a sample. Although hybridization-based approaches like DNA microarrays can provide only a relative, analog measure of transcript abundance, sequencing-based approaches, such as RNA sequencing (RNA-Seq), have the advantage of removing hybridization bias among genes (1, 2) and offer the promise of true digital quantification.

The interpretation of conventional RNA-Seq is complicated by sequence-dependent bias and amplification noise from reverse transcription, adapter ligation, library amplification by PCR, solid-phase clonal amplification, and sequencing (3–5). NanoString technology mitigates these complications by eliminating enzymatic reactions and hybridizing color-coded probes directly to RNA for single-molecule detection (6), although it requires many specific probes. Other methods reduce bias in RNA-Seq by eliminating PCR and directly sequencing single molecules of RNA (7), or sequencing single molecules (8) or clonal populations (9) of cDNA. However, library amplification is desirable for sequencing small samples or single cells (10).

Conventional library amplification is based on PCR, but the exponential amplification afforded by PCR introduces noise, especially at low copy numbers (11). Digital PCR was introduced to circumvent this problem by distributing DNA molecules into many containers, each receiving zero or one molecules, which are amplified and detected by PCR (12). This technique has been successfully applied to RNA counting (13); however, it requires specific primers for each gene, which hinders high-throughput measurements.

Here we report a system-wide method for bias and noise reduction in RNA-Seq that allows the use of PCR to amplify a cDNA library before sequencing, providing accurate digital quantification of the transcriptome. In our approach, each cDNA molecule is attached to a unique barcode sequence from a large pool of barcodes before amplification (Fig. 1*A*) (14). Deep sequencing then allows quantification of the number of cDNA molecules in the original sample by counting the number of unique barcode sequences associated with a given cDNA sequence. This concept has been applied recently for studying protein-RNA interactions (15), to improve the sensitivity of DNA mutation detection (16, 17) and accuracy of DNA copy-number measurements for individual genes by threshold detection (18), and to perform karyotyping and mRNA profiling (19). However, barcode identification in these studies was not immune to errors incurred during library preparation, amplification, and sequencing, which can convert one barcode into another. Hence, a substantial fraction of reads contained misidentified barcodes (16–19), which in some cases were discarded using an artificial threshold (17–19). To avoid this complication, we designed optimized barcodes that can be ligated and amplified with minimal bias and distinguished from one another despite the accumulation of PCR mutations and sequencing errors.

Results

Barcoding Strategy for Digital RNA-Seq. Fig. 1*A* depicts the general concept of digital counting by random labeling of all target nucleic acid molecules in a sample with unique barcode sequences. To achieve unique barcoding of as many target sequences as possible, the set of barcode sequences introduced to the sample must be (*i*) much larger than the copy number of the most abundant target sequence and (*ii*) sampled randomly by the target sequences. If these two criteria are satisfied, then digital quantification of the target molecules by this method is limited only by sequencing depth and accuracy. Unlike conventional sequencing-based approaches to nucleic-acid quantification, the digital counting technique is no longer limited by intrinsic amplification noise and bias in downstream sample preparation and sequencing (Fig. 1*A*).

Implementation of the scheme in Fig. 1*A* for digital RNA-Seq requires several critical considerations. As noted above, if the barcode sequences are random, then a sequencing error at one position in a barcode will cause that barcode to be misidentified. This error-induced interconversion will occur even if the barcode sequences are nonrandom (18), unless the barcodes are carefully

Author contributions: K.S., T.Z.J., P.A.S., and X.S.X. designed research; K.S. and T.Z.J. performed research; K.S., T.Z.J., and P.A.S. analyzed the data; K.S., T.Z.J., P.A.S., and X.S.X. wrote the manuscript.

Conflict of interest statement: Harvard University has filed a provisional patent application based on this work.

This article is a PNAS Direct Submission.

Data deposition: The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession no. GSE34449).

[1]To whom correspondence should be addressed. E-mail: xie@chemistry.harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1118018109/-/DCSupplemental.

SYSTEMS BIOLOGY

IDT Exhibit 1030.0001

Downloaded from https://www.pnas.org by 216.59.127.254 on June 10, 2024 from IP address 216.59.127.254.

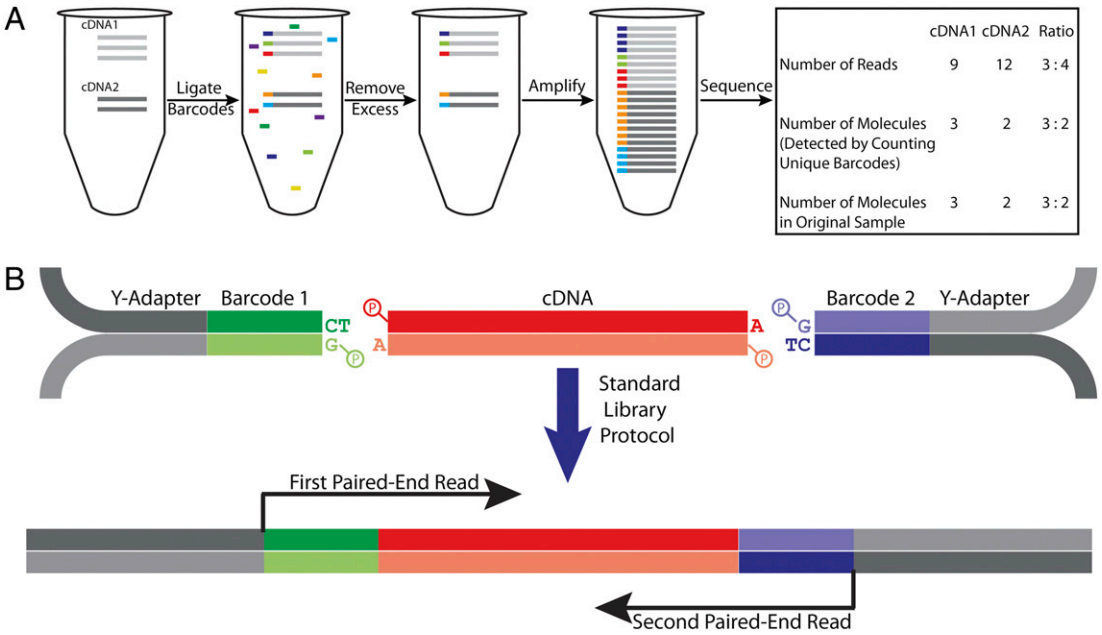


Fig. 1. Our scheme of digital RNA-Seq. (*A*) General principle of digital RNA-Seq. Assume the original sample contains two cDNA sequences, one with three copies and another with two copies. An overwhelming number of unique barcode sequences are added to the sample in excess, and five are randomly ligated to the cDNA molecules. Ideally, each cDNA molecule in the sample receives a unique barcode sequence. After removing the excess barcodes, the barcoded cDNA molecules are amplified by PCR. Because of intrinsic noise and sequence-dependent bias, the barcoded cDNA molecules are amplified unevenly. Consequently, after the amplicons are sequenced, it appears that there are three copies of cDNA1 for every four copies of cDNA2 based on the relative number of reads for each sequence. However, the ratio in the original sample was 3:2, which is accurately reflected in the relative number of unique barcodes associated with each cDNA sequence. (*B*) In our implementation of *A*, we found it advantageous to randomly ligate both ends of each phosphorylated cDNA fragment to a barcoded phosphorylated Illumina Y-shaped adapter. Note that the single T and A overhangs present on the barcodes and cDNA, respectively, are to enhance ligation efficiency. After this step, the sample is amplified by PCR and prepared for sequencing using the standard Illumina library protocol. For each amplicon, both barcode sequences and both strands of the cDNA sequence are read using paired-end deep sequencing.

designed so that multiple substitution errors and indels do not obscure their identities (20). Because DNA secondary structure can reduce amplification efficiency, the barcodes should not have significant sequence overlap or complementarity with each other, the adapter and primer sequences used in library preparation and sequencing, or the transcriptome-of-interest. Ideally, the barcode set will not contain sequence motifs that are known to be problematic for sequencing chemistries, such as long homopolymers and regions with high or low GC-content.

We used a computer program to generate a set of 145 barcode sequences 20-bp in length (Dataset S1) that satisfies the above criteria (*Materials and Methods* and *SI Materials and Methods*). The barcode sequences can sustain up to four substitution errors and remain unambiguously identifiable. In addition, a barcode that incurs up to nine substitution errors or the combination of one indel and five substitution errors will not take on the sequence of another barcode.

Instead of using a single barcode sequence to identify each target molecule in our sample (15–19), we attached a barcode sequence to both ends of each target molecule (Fig. 1*B* and *SI Materials and Methods*). If both ends of a target sequence sample all of the barcodes randomly, the target sequence will have access to $145 \times 145 = 21{,}025$ unique labels. The two barcode sequences along with the target molecule sequence were then read out by paired-end sequencing (Fig. 1*B*). This paired-end strategy dramatically reduces the number of barcodes that must be designed and synthesized, is compatible with conventional paired-end library protocols, and provides long-range sequence information that improves mapping accuracy (21, 22). In addition, attaching barcodes to both ends increases the overall randomness of barcode sampling because the two ends of a target molecule are unlikely to have a similar degree of bias. We tested and characterized this method on a set of quantified DNA spike-in sequences and a cDNA library derived from the transcriptome of *Escherichia coli*.

Quantification of Spike-in Sequences and Barcode Sampling Bias. To calibrate our digital RNA-Seq system, we measured the concentrations of five synthetic DNA spike-in sequences using the Fluidigm digital PCR platform and used them as internal standards. The spike-in samples were barcoded, added to the barcoded *E. coli* cDNA library, and quantified using the sequencing-based digital counting strategy described above. Fig. 2*A* shows that the number of digital counts (i.e., unique barcodes) observed in deep sequencing is well-correlated with the digital PCR calibration of the spike-in sequences.

To evaluate the difference between using random barcode sequences and our optimized barcode sequences, we conducted two experiments. In one experiment, we labeled the spike-in molecules with random barcode sequences (*SI Materials and Methods*), and in the second experiment we used our optimized, predetermined barcode set. We constructed the histograms of the number of reads for all barcodes observed from the most abundant spike-in sequence (Fig. 2*B*). When using random barcodes (red histogram in Fig. 2*B* and *SI Materials and Methods*), the left-most bin exhibits a large peak because a substantial fraction of barcodes are infrequently read because of sequencing errors; this causes barcodes to interconvert, generating quantification artifacts that were also evident in previous reports (16–19). In stark contrast, the left-most bin, when using optimized barcodes (green histogram in Fig. 2*B*), has no such peak because our optimized barcode sequences avoid misidentification because of sequencing errors. The effect of sequencing error on both random and optimized barcode counting is clearly shown by simulation (Fig. S1 and *SI Materials and Methods*).

Downloaded from https://www.pnas.org by 216.59.127.254 on June 10, 2024 from IP address 216.59.127.254.

IDT Exhibit 1030.0002

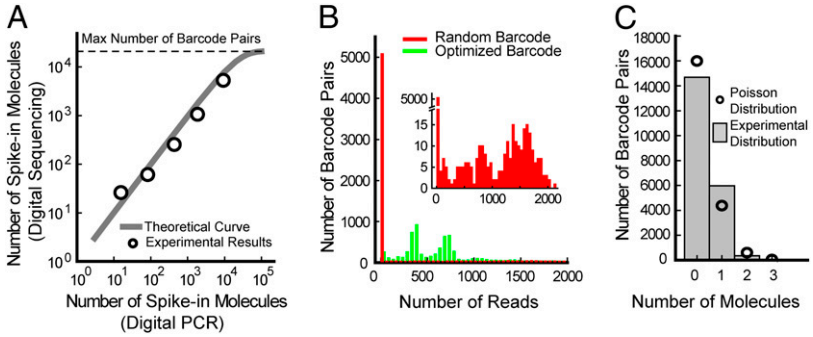


Fig. 2. Spike-in sequence quantification. (*A*) Correlation between the number of spike-in molecules for five different spike-in sequences as measured by digital PCR and digital counting of unique barcodes. The theoretical curve, which saturates because of the finite number of barcode pairs (21,025), is calculated based on the Poisson distribution (18). (*B*) Histograms of the number of reads corresponding to each observed barcode attached to the most abundant spike-in sequence for two experiments. The red histogram corresponds to a spike-in sequence labeled with random barcode sequences, and the green histogram corresponds to a spike-in sequence labeled with our optimized barcodes. Note the left-most bin in the red histogram is >10-times larger than that of the green histogram and contains a large number of unique barcodes with a low number of reads. This discrepancy is caused by various sequencing and PCR amplification errors, which generate new artifactual unique barcodes not present in the original sample and result in a large number of falsely identified unique barcodes (*SI Materials and Methods*). (*Inset*) The red histogram in greater detail. (*C*) Histogram of the number of times a barcode pair was observed with all five spike-in sequences (i.e., the number of spike-in molecules attached to a given barcode pair). Because the spike-in sequences sample the barcode pairs randomly with very little bias, the histogram follows a Poisson distribution.

We note that the green histogram in Fig. 2*B* is the distribution of the number of reads for the 5,311 uniquely barcoded molecules from a particular spike-in (*SI Materials and Methods*). Assuming each barcoded spike-in molecule is identical, the green histogram in Fig. 2*B* is essentially the probability distribution of the number of reads for a single molecule, which spans three orders-of-magnitude. This broad distribution arises primarily from intrinsic PCR amplification noise (11) in sample preparation. Given this broad single-molecule distribution, for low copy molecules in the original sample, counting the total number of reads (conventional RNA-Seq) would be catastrophic. On the other hand, this problem can be circumvented if one counts the number of different barcodes (integrated area of the histogram) using our digital RNA-Seq approach, yielding accurate quantification with single-copy resolution. The two counting schemes give the same results only when the copy number in the original sample is high, assuming there is no sequence-dependent bias.

Random sampling of the barcode sequences by each target sequence is essential for accurate digital counting. Fig. 2*C* shows that the distribution of observed molecule counts is in excellent agreement with Poisson statistics. Therefore, the five spike-in sequences sample the 21,025 barcode pairs without bias.

Digital Quantification of the *E. coli* Transcriptome. We obtained 26–32 million reads from our barcoded cDNA libraries that uniquely mapped to the *E. coli* genome (*Materials and Methods* and Dataset S2) in two replicate experiments. Fig. 3*A* shows the number of conventional and digital counts (unique barcodes) as a function of nucleotide position for the fumA transcription unit. Not surprisingly, the read density is considerably less uniform across the transcription unit than the number of digital counts, presumably because of intrinsic noise and bias in fragment amplification.

It is crucial for transcripts across the *E. coli* transcriptome to sample all barcodes evenly. Fig. 3*B* shows this distribution, which is close to Poisson but is somewhat overdispersed. Such biased sampling reduces the effective number of barcode sequences N_{eff} available. However, in our *E. coli* transcriptome sample, the copy number of the most abundant cDNA ranges from 10 to 40 copies for both counting methods. Based on Poisson statistics, even for the most abundant cDNA fragments in our sample, the required N_{eff} is ~100–400 for 95% unique labeling of all molecules (18). Because there are 21,025 barcode pairs available, on average the degree of randomness observed in Fig. 3*B* is sufficient.

The conventional method counts the number of amplicons, a quantity that is subject to bias and intrinsic amplification noise (11), rather than the number of molecules in the original sample. Conversely, in our digital counting scheme, unique barcode sequences distinguish each molecule in the sample, and so the effects of intrinsic noise are minimized. Fig. 3*C* shows how drastically different digital counting can be from conventional counting at low copy numbers, implying that digital counting of unique barcodes is advantageous, particularly for quantifying low copy fragments. We note that the correlation is stronger for high copy fragments and the same phenomenon is also observed for whole transcription units and genes (Fig. S2).

To demonstrate the superior accuracy of digital counting, we examined the uniformity of our abundance measurements within individual transcripts. Because individual transcription units were, by-and-large, intact RNA molecules following RNA synthesis, the cDNA fragments that map to one region of a given transcription unit should have the same abundance as fragments that map to a different region of the same transcription unit. We histogrammed the ratio between the variation in conventional counting ν_C and variation in digital counting ν_D for transcription units in different abundance ranges (Fig. 3*D*). A variation ratio of $\nu_C/\nu_D = 1$ indicates that both conventional and digital counting give similarly uniform abundances along the length of a transcription unit. For a transcription unit where ν_C/ν_D exceeds one, conventional counting measures abundance less consistently along the transcription unit than digital counting. The mean values of ν_C/ν_D in the two replicates are 1.4 ($s = 1.5$, where s is sample SD) and 1.2 ($s = 0.5$) for the complete set of analyzed transcription units, indicating that conventional counting is less consistent than digital counting across an average transcription unit. Furthermore, the mean value of ν_C/ν_D increases with decreasing copy number and its distribution becomes broader (Fig. 3*D*). For transcription units in the lowest abundance regime, the mean values of ν_C/ν_D are 1.9 ($s = 2.4$) and 1.3 ($s = 0.9$) for the two replicates. We conclude that, on average, digital counting outperforms conventional counting in terms of accuracy, and its performance advantage is most pronounced for low abundance transcription units.

Although Fig. 3 demonstrates that digital counting is less noisy and more accurate than conventional counting, Fig. 4 shows that digital counting is also more reproducible. We demonstrate this on the level of a single transcription unit in Fig. 4*A*, which shows

SYSTEMS BIOLOGY

Downloaded from https://www.pnas.org by 216.59.127.254 on June 10, 2024 from IP address 216.59.127.254.

IDT Exhibit 1030.0003

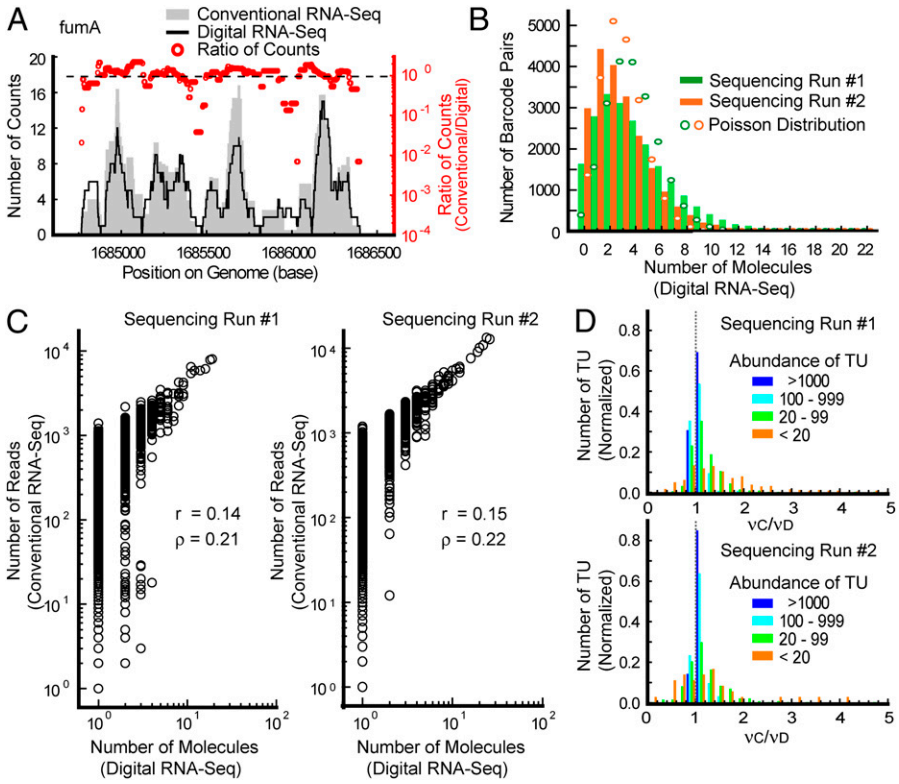


Fig. 3. Digital quantification of the *E. coli* transcriptome. (*A*) Conventional and digital counting results for the fumA transcription unit (TU) as a function of genome position. The conventional counts were calculated by using a conventional calibration curve that allows regression of the number of reads against the number of input molecules for all spike-in molecules (Fig. 2*A*). The digital counts were obtained by counting the number of unique barcodes associated with each fragment. The red dots are the ratios of these two numbers for each base. (*B*) Histograms of the number of times a barcode pair was observed with the *E. coli* cDNA sequences (i.e., the number of cDNA molecules attached to a given barcode pair) in the two replicates. Barcode sampling is more biased on average for *E. coli* cDNA fragments, but is still in reasonably good agreement with Poisson statistics. (*C*) Correlation between the number of reads (conventional counting) and the number of molecules obtained from digital counting of unique barcodes for every mapped fragment in the two replicates. For low copy molecules, the conventional counts are distributed over three orders-of-magnitude; this is because the conventional method counts amplicons, which are subject to intrinsic noise (11), rather than directly counting molecules in the original samples like the digital counting method. We note that higher copy fragments are less affected by intrinsic noise (11), as the number of molecules sequenced is greater; this effectively allows averaging over the read counts of many molecules in conventional RNA-Seq, decreasing the variance of counting in the process. (*D*) Uniformity of conventional vs. digital counting along the length of each transcription unit as a function of transcription unit abundance across the whole *E. coli* transcriptome for both replicates. We calculated the variation $\nu_D = s_D/\mu_D$ (where μ_D and s_D are the mean and sample SD of the digital counts among 99-base bins in a transcription unit, respectively) associated with digital counting and the variation $\nu_C = s_C/\mu_C$ associated with conventional counting within each transcription unit for which at least three bins contained on average at least one read. We then created the histogram of the ratio between conventional and digital counting variation (ν_C/ν_D) for transcription units in different abundance ranges for each replicate. Transcription unit abundance is the sum of all digital counts for each fragment in the transcription unit.

the ratio of counts between the two replicates for both conventional and digital counting along the fumA transcript. This ratio is consistently close to one for digital counting, but fluctuates over three orders-of-magnitude for conventional counting. We analyzed the global reproducibility of the whole transcriptome for quantification of transcription units and genes for both conventional and digital counting in Fig. 4 *B* and *C*, respectively. In both cases, the correlation between replicates is noticeably better for digital counting than conventional counting, particularly for low copy transcripts.

Discussion

Unlike previously reported methods of eliminating bias and noise from RNA-Seq (7–9), our strategy allows amplification by PCR and uses standard commercial protocols for sample preparation. However, the implementation described above also leaves considerable room for improvement. For example, one could ligate barcoded adapters directly to RNA (23, 24), reducing the bias that occurs during reverse transcription. Alternatively, a recently described protocol for processing mature mRNA from single mammalian cells could be modified to include barcoded primers for reverse transcription and second-strand synthesis before amplification (10), obviating the need for ligation.

One disadvantage of our technique is that it requires higher sequencing coverage than conventional RNA-Seq. This requirement is because both the transcriptome and the barcode set must be evenly sampled for accurate counting. However, the cost-per-base of deep sequencing continues to decrease rapidly. In our experiment, the mean number of reads per fragment was ~400. However, the spike-in sequencing reads can be randomly down-sampled 10-fold (*SI Materials and Methods*) without perturbing the correlation between abundance measured by digital PCR and digital barcode counting (Fig. S3), which implies that significantly lower coverage will suffice in many cases.

For applications where many cycles of PCR are required for sensitive detection, bias and noise reduction are crucial for accurate quantification. Although we demonstrated our technique on the *E. coli* transcriptome, we note that the maximum copy number for polyadenylated mRNA in a single mouse blastomere was found to be ~2,400 (10). With 155 optimized barcode sequences (10 more than were used in this study), one could uniquely label nearly every identical molecule in this system

Downloaded from https://www.pnas.org by 216.59.127.254 on June 10, 2024 from IP address 216.59.127.254.

IDT Exhibit 1030.0004

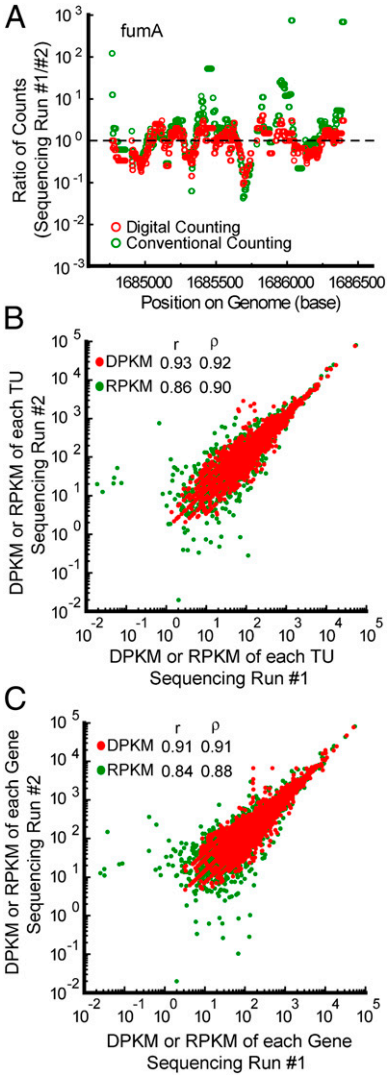


Fig. 4. Reproducibility of digital and conventional quantification of the *E. coli* transcriptome. (*A*) Ratio of counts between two replicate sequencing runs normalized by total uniquely mapped reads for digital counting plotted along with the ratio of counts between the two replicates for conventional counting of the fumA transcription unit. As expected, the ratio fluctuates over a broader range for conventional counting than digital counting along the length of the transcription unit. (*B*) Correlation between replicate sequencing runs for digital and conventional counting of transcription units. DPKM represents the uniquely mapped digital counts per kilobase per million total uniquely mapped molecules. RPKM represents the uniquely mapped reads per kilobase per million total uniquely mapped reads. (*C*) Correlation between replicate sequencing runs for digital and conventional counting of genes. Taken together, *B* and *C* demonstrate that digital counting is globally more reproducible than conventional counting.

(with 95% unique labeling for even the most abundant transcript). Hence, we expect this technique will be readily applicable to eukaryotic systems without substantial modification. In addition, we analyze the performance of digital and conventional counting in a simulation of differential expression analysis, a key application of RNA-Seq (Fig. S4). Our simulation, which accounts for experimentally measured copy number, barcode sampling bias and amplification noise distributions, shows that digital counting of unique barcodes outperforms conventional counting for differential expression analysis (*SI Materials and Methods*). Although it is always more difficult to reject the null hypothesis for low-abundance transcripts, we expect our digital counting scheme to be nonetheless more accurate than conventional counting for differential expression analysis at low copy numbers.

In addition to single-cell applications, we expect this technique to be particularly useful for nascent transcript sequencing by run-on (25) or RNA polymerase capture (26), ribosome profiling (27), and profiling of miRNA and other regulatory RNAs, which typically exist at low copy numbers. Significant recent progress has been made in minimizing bias induced by sample barcodes for multiplexed miRNA-Seq (28), and we expect that this technique could be applied to the introduction of barcodes for digital counting in any RNA-Seq experiment. In addition, one could use our approach to improve DNA sequencing experiments, such as ChIP sequencing (ChIP-Seq) (29), which is procedurally related to RNA-Seq and exposed to similar sources of bias and noise (30).

RNA-Seq holds substantial promise for basic research in biomedicine and may ultimately impact clinical diagnostics (21, 31, 32). However, challenges ranging from bias in sample preparation to limited sensitivity and remain significant. Digital RNA-Seq, along with continued improvements to sequencing technology, will lead to new applications and allow RNA-Seq to reach its full potential.

Materials and Methods

Generation and Optimization of Barcodes. We generated 2,358 random 20-base barcode candidates using a computer, such that even if a barcode accumulated nine mutations, it would not take the sequence of any other generated barcode sequences [unlike the random barcode case (*SI Materials and Methods* and Dataset S3)]. Barcode candidates containing homopolymers longer than four bases or GC-content less than 40% or greater than 60% were discarded. Barcode candidates were also discarded if each exceeded a certain degree of complementarity or sequence identity (total matches and maximum consecutive matches) with (*i*) the Illumina paired-end sequencing primers (33), (*ii*) the Illumina PCR primers PE 1.0 and 2.0, (*iii*) the 3′ end of the Illumina PCR primers PE 1.0 and 2.0, (*iv*) the whole *E. coli* genome [K-12 MG1655 strain (U00096.2)], and (*v*) all other generated barcode candidates (*SI Materials and Methods*). Any barcode candidate for which an indel mutation would place it within five point mutations of another barcode candidate was also discarded. The final population consisted of 150 barcodes, of which 145 were randomly chosen and used (Dataset S1).

***E. coli* RNA Preparation and cDNA Generation.** The cDNA library of *E. coli* [K-12 MG1655 strain (U00096.2)] was generated by a standard method (*SI Materials and Methods*).

Sample-Adapter Ligation, Sequencing Sample Preparation, and Sequencing. The cDNA library was ligated to the barcode adapter mixture, and the sequencing sample was prepared by the standard Illumina protocol with some modifications along with an internal standard (*SI Materials and Methods* and Dataset S4).

***E. coli* Transcriptome Analysis.** From the raw sequencing data, we isolated reads that contained barcode sequences that corresponded to our original list of 145 barcodes in both forward and reverse reads for each sequencing cluster that had at most one mismatch. We then aligned the first 28 bases (26 bases for the second sequencing run) of the targeted sequence of both the forward and reverse reads of each cluster to the *E. coli* genome and kept the sequences that uniquely align fewer than three mismatches and where the two reads did not map to the same sense or antisense strand of the genome. The remaining sequences were mapped to transcription units (34) and sorted by starting and ending position, as well as forward and reverse barcodes (unique tag). Mapped sequence fragments with a length of at least 1,000 bases were discarded. All sequences within the same transcription unit that had the same unique tag were analyzed further. We determined that more than one sequence with the same unique tag were identical if the distance between their center positions was less than 4 bp and if the difference in length was less than 9 bp (Figs. S5 and S6). Thus, the read counts for sequences deemed identical were summed and the sequence with more read counts was deemed as the actual correct sequence. Then, for each unique sequence, we counted the number of unique barcode tags that appeared to determine the copy number of each sequence. The genome-wide expression profile by digital counting and conventional counting are visualized by Integrated Genome Browser (*SI Materials and Methods*; see Datasets S5, S6, and S7 for specific values for datapoints shown in Fig. S2).

SYSTEMS BIOLOGY

Downloaded from https://www.pnas.org by 216.59.127.254 on June 10, 2024 from IP address 216.59.127.254.

IDT Exhibit 1030.0005

ACKNOWLEDGMENTS. Illumina sequencing was performed at the Harvard Faculty of Arts and Sciences Center for Systems Biology, with the help of Christian Daly, and at the Tufts University School of Medicine Genomics Core Facility, with the help of Kip Bodi and James Schiemer; digital PCR was performed by the Fluidigm Genetic Analysis Facility at the Molecular Genetics Core Facility of the Children's Hospital Boston Intellectual and Developmental Disabilities Research Center, with the help of Hal Schneider and Ta-Wei Lin; and base-calling analysis was performed at the Harvard Faculty of Arts and Sciences Research Computing Group by Jiangwen Zhang. This work was supported by National Institutes of Health (NIH) National Human Genome Research Institute Grant HG005097-01 (to X.S.X.) and NIH National Human Genome Research Institute Recovery Act Grand Opportunities Grant 1RC2HG005613-01 (to X.S.X.); and K.S. was supported by a Postdoctoral Fellowship for Research Abroad from the Japanese Society for the Promotion of Science and a fellowship from The Uehara Memorial Foundation.

1. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5:621–628.
2. Nagalakshmi U, et al. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320:1344–1349.
3. Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* 36:e105.
4. Aird D, et al. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* 12:R18.
5. Zheng W, Chung LM, Zhao H (2011) Bias detection and correction in RNA-Sequencing data. *BMC Bioinformatics* 12:290.
6. Geiss GK, et al. (2008) Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nat Biotechnol* 26:317–325.
7. Ozsolak F, et al. (2009) Direct RNA sequencing. *Nature* 461:814–818.
8. Lipson D, et al. (2009) Quantification of the yeast transcriptome by single-molecule sequencing. *Nat Biotechnol* 27:652–658.
9. Mamanova L, et al. (2010) FRT-seq: amplification-free, strand-specific transcriptome sequencing. *Nat Methods* 7:130–132.
10. Tang F, et al. (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* 6:377–382.
11. Peccoud J, Jacob C (1996) Theoretical uncertainty of measurements using quantitative polymerase chain reaction. *Biophys J* 71:101–108.
12. Vogelstein B, Kinzler KW (1999) Digital PCR. *Proc Natl Acad Sci USA* 96:9236–9241.
13. Ottesen EA, Hong JW, Quake SR, Leadbetter JR (2006) Microfluidic digital PCR enables multigene analysis of individual environmental bacteria. *Science* 314:1464–1467.
14. Hug H, Schuler R (2003) Measurement of the number of molecules of a single mRNA species in a complex mRNA preparation. *J Theor Biol* 221:615–624.
15. Chi SW, Zang JB, Mele A, Darnell RB (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* 460:479–486.
16. Casbon JA, Osborne RJ, Brenner S, Lichtenstein CP (2011) A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Res* 39:e81.
17. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B (2011) Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci USA* 108:9530–9535.
18. Fu GK, Hu J, Wang PH, Fodor SP (2011) Counting individual DNA molecules by the stochastic attachment of diverse labels. *Proc Natl Acad Sci USA* 108:9026–9031.
19. Kivioja T, et al. (2011) Counting absolute numbers of molecules using unique molecular identifiers. *Nat Methods*, 10.1038/nmeth.1778.
20. Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Methods* 5:235–237.
21. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: A revolutionary tool for transcriptomics. *Nat Rev Genet* 10:57–63.
22. Au KF, Jiang H, Lin L, Xing Y, Wong WH (2010) Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res* 38:4570–4578.
23. Lau NC, Lim LP, Weinstein EG, Bartel DP (2001) An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* 294:858–862.
24. Axtell MJ, Jan C, Rajagopalan R, Bartel DP (2006) A two-hit trigger for siRNA biogenesis in plants. *Cell* 127:565–577.
25. Core LJ, Waterfall JJ, Lis JT (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* 322:1845–1848.
26. Churchman LS, Weissman JS (2011) Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* 469:368–373.
27. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS (2009) Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324:218–223.
28. Alon S, et al. (2011) Barcoding bias in high-throughput multiplex sequencing of miRNA. *Genome Res* 21:1506–1511.
29. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316:1497–1502.
30. Nix DA, et al. (2008) Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. *BMC Bioinformatics* 9:523.
31. Ozsolak F, et al. (2010) Amplification-free digital gene expression profiling from minute cell quantities. *Nat Methods* 7:619–621.
32. Shah SP, et al. (2009) Mutation of *FOXL2* in granulosa-cell tumors of the ovary. *N Engl J Med* 360:2719–2729.
33. Bentley DR, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456:53–59.
34. Keseler IM, et al. (2011) EcoCyc: A comprehensive database of *Escherichia coli* biology. *Nucleic Acids Res* 39(Database issue):D583–D590.

Downloaded from https://www.pnas.org by 216.59.127.254 on June 10, 2024 from IP address 216.59.127.254.

IDT Exhibit 1030.0006

Supporting Information

Shiroguchi et al. 10.1073/pnas.1118018109

SI Materials and Methods

Evaluation of Digital Counting for Synthesized DNAs Using Random Barcodes by Deep Sequencing. To evaluate the capability of random barcodes to digitally count DNA molecules, we attempted digital counting of synthesized DNA on the Illumina platform using single read sequencing.

Design of single-stranded DNA with random barcodes. We designed and purchased (Integrated DNA Technologies) nine single-stranded DNA molecules, each with adapter sequences at both ends that are compatible with the Illumina platform (Dataset S3). Ten, 16, or 20 random bases were positioned near the beginning of sequencing reads. We measured the concentration of each DNA by absorbance at 260 nm using the extinction coefficient provided by Integrated DNA Technologies.

Sequencing sample preparation and sequencing. We mixed all nine single-stranded DNA templates for a final concentration of 0.17 aM ssDNA-1, 0.67 aM ssDNA-2, 3.3 aM ssDNA-3, 33 aM ssDNA-4, 330 aM ssDNA-5, 660 aM ssDNA-6, 660 aM ssDNA-7, 660 aM ssDNA-8, and 660 aM ssDNA-9 in HF buffer [New England Biolabs(NEB)] with 0.4 mM dNTP, 1 μM of each amplification primer (Integrated DNA Technologies) (Dataset S3), and 1 U Phusion DNA polymerase (NEB) in a final volume of 50 μL. These DNA molecules were amplified by PCR (1 cycle of 98 °C for 30 s, 20 cycles of both 98 °C for 10 s and 60 °C for 30 s, and 1 cycle of 72 °C for 10 min). After column purification (Zymo Research), the concentration was measured by quantitative PCR (qPCR) (Fast SYBR Master Mix; Applied Biosystems) using conventional Illumina qPCR primers against a reference PCR product amplified from ssDNA-2 using the amplification primers mentioned above, the concentration of which was measured by NanoDrop (LMS). The samples were sequenced on an Illumina Genome Analyzer II in one lane with 36-base single reads. We obtained 26,566,253 reads resulting from 26,566,253 sequencing clusters, of which 25,738,380 mapped to the designed template sequences (Dataset S3).

Data analysis. We sorted the raw reads by the known sequence for each template (Dataset S3) and allowed up to two single-base mismatches and one insertion or deletion of a single base before the barcode region. We counted the number of molecules for each template by counting the number of uniquely observed barcodes.

Results. For all templates, the results were completely inconsistent with the estimated number of molecules added in the PCR mixture. For example, the measured number of molecules of ssDNA-8 [shown in Fig. 2*B* (red)] was 15.5-fold higher than the estimated number of the DNA molecules added to the original PCR mixture; we identified 155,161 unique barcodes compared with an estimated input of 10,000 molecules. The histogram of the distribution of the number of reads for the random barcodes [for example, Fig. 2*B* (red)] showed that all ssDNA templates have a remarkably high peak at a very low number of reads. When we assumed that two molecules with up to two mismatches in the barcode region were originally from same DNA molecule, the peak significantly decreased [for example, the peak in Fig. 2*B* (red) decreased about 30-fold]. This results in 23,890 remaining barcodes and a roughly 2.4-fold higher output than the estimated input (two other templates, ssDNA-1 and ssDNA-3 also result in a 5.8-fold and a 27-fold higher estimation, respectively, after the same process). Such error correction may reduce the number of potential false-positives, but it is still inaccurate and results in uncertainty; the maximum number of allowable nucleotide differences used for correction greatly affects the final results. This finding suggests that newly generated sequences at the barcode position, from sequencing or PCR amplification errors that did not exist in the original PCR mixture, may have become new artifactual unique barcodes, which resulted in a significant number of false-positives. The data for the red histogram shown in Fig. 2*B* was generated by down-sampling reads such that the maximum number of reads for any given molecule, and the number of total molecules depicted in the figure corresponds to the parallel optimized barcode experiment described in the main text.

Efficacy of Digital Counting Strategy. The correlation plot in Fig. 2*A* demonstrates the efficacy of our digital counting strategy for the spike-in sequences. In addition, the histogram in Fig. 2*B* (green) for the optimized barcode also highlights an important difference between the conventional approach of counting the number of reads corresponding to a target sequence and digital counting based on unique barcode labeling. We know that the abundance of the spike-in molecule was accurately quantified by our digital counting scheme, and so the vast majority of unique barcodes observed for the spike-in molecule correspond to individual molecules of the spike-in in our original sample. Hence, the correct number of digital counts for most of those molecules is one. Because Fig. 2*B* is a histogram of the number of reads observed for each of the spike-in barcodes, it is equivalent to the conventional counting distribution for individual copies of the spike-in in the original sample. Therefore, according to the conventional counting strategy, the abundance of each individual copy of the spike-in in our original sample varies over three orders-of-magnitude (Fig. 2*B*). This discrepancy is most likely because of intrinsic noise in sample preparation (arising from amplification and sequencing), which in this case is catastrophic for the conventional method but does not affect the accuracy of our digital counting strategy.

Tenfold Down-Sampling of Spike-in Reads. We randomly down-sampled the spike-in reads in the replicate experiment by a factor of 10. For each read of each of the five spike-ins, there was a 10% chance that it was kept and counted, whereas the other 90% was discarded. Fig. S3 shows that there is little dropout between these two conditions (the data show the average dropout rate for the five spike-ins to be 1.6%) and that for the spike-in with the highest number of molecules, the change in the single-molecule coverage histogram is minimal.

Simulation Comparison of Performance of Optimized Barcoding Scheme vs. Random Barcoding Scheme. We assumed a theoretical library of 6,000 identical template molecules and barcoded each molecule *in silico* under two separate conditions: (*i*) with 6,000 optimized barcodes and (*ii*) with 6,000 random barcodes. The optimized barcodes mirror the technique described in the main text (Paired-end reads, 20 nucleotides), and the random barcodes mirror the technique described above (single-end reads, 16 nucleotides). We note that previous uses of the random barcode procedure have only involved single-end reads (1, 2). Each of the barcoded templates in both conditions were "amplified" without sequence-dependent bias and noise for simplicity to a copy number of 400. Then, using the averaged sequencing error profile for the Illumina platform characterized by Nakamura et al. (3), we simulated sequencing errors of the first 53 bases. Mapping dropout was also included in the simulation (we allowed at most two sequencing errors per transcript for mapping). If we allow one mismatch in both barcode regions simulating

IDT Exhibit 1030.0007

actual experimental conditions for the optimized barcode simulation, we observe that there are no artificially created barcodes (i.e., false-positives) (Fig. S1). However, for the random barcode simulation, there is a substantial peak for low copy barcodes, which is because of sequencing errors generating new barcodes not originally present in the original sample, as we see that the total number of original molecules estimated is almost two orders-of-magnitude greater than the actual value (Fig. S1). From the *in silico* sequencing results in the random case, we cannot determine which barcodes were actually attached to the template molecules and which barcodes were generated by sequencing error because all possible random sequences can be potentially attached to the template molecule. As a result, sequencing error generates many false-positives. In the case of optimized barcodes, sequencing errors occurred as well. However, we can identify the sequences that were not in the set of used barcodes, and so sequencing errors in the barcode region can be identified and removed, which does not result in false-positives. If we assume resulting random barcode sequences that differ by one nucleotide are identical, the estimated output is 58,556 and still contains many apparent false-positives (the two nucleotide case results in 9,597 estimated output molecules). Although error correction reduces the number of false-positives, it is clearly not as accurate as our optimized barcode pair method and still results in uncertainty; the maximum number of allowable nucleotide differences used for correction greatly affects the final results.

Simulations Demonstrating the Utility of Digital RNA Sequencing for Differential Expression Analysis. We performed computer simulations to demonstrate that digital counting using optimized barcodes can substantially outperform conventional counting in differential expression analysis experiments, where expression profiles are compared under various conditions. We simulated differential expression analysis for three cases: a system with the same copy number distribution as the *Escherichia coli* fragment library described in Fig. 3, a system with approximately the same copy number distribution as the *E. coli* transcription unit library described in Fig. S2, and a system with the copy number distribution that was experimentally measured for microRNA in human stomach tissue (4). In each case, two gene expression profiles, each having the same copy number distribution, were generated randomly and compared with each other. For the *E. coli* fragment, transcription unit, and human microRNA libraries, each molecule in the simulation (83,863 and 352,088 total molecules for the *E. coli* fragments and human microRNA libraries, respectively, and 83,500 total molecules for ∼4,000 transcription units for the transcription unit library) was assigned a barcode from a large pool (21,025 barcode pairs for the *E. coli* fragment and transcription unit libraries corresponding to 145 barcode sequences and 189,225 barcode pairs for the microRNA library, corresponding to 435 barcode sequences). However, the barcodes were not assigned randomly. Instead, they were assigned based on the barcode sampling bias distribution, which we experimentally measured for the *E. coli* transcriptome (Fig. 3*B*). After assigning barcodes, we simulated PCR amplification of each molecule. We know from our conventional counting results for the *E. coli* transcriptome that a significant amount of noise is incurred during amplification because not every PCR cycle for every molecule is 100% efficient (Fig. 3*C*). Imperfect efficiency and exponential amplification conspire to produce the majority of the variability in conventional counting of low copy transcripts observed in Fig. 3*C*. Hence, in our simulation, we assumed a PCR cycle efficiency (70%) that reproduces the experimentally measured amplification variance for single molecules observed in Fig. 3*C*. We note that the PCR simulation does not account for sequence-dependent or cycle-dependent amplification bias. Just as in our *E. coli* transcriptome RNA sequencing (RNA-Seq) experiment, we applied 18 cycles of PCR to every molecule in each library. At the end of the simulation, we counted the number of unique barcodes assigned to each sequence (digital counts) and the number of PCR amplicons for each sequence (conventional counts). After running this simulation for both libraries in the two cases, we obtained a simulated fold-change for each gene and compared it to the original input fold-change for each gene. Fig. S4 shows the ratio of the simulated to input fold-changes for each fragment in the *E. coli* fragment library, each transcription unit in the *E. coli* transcription unit library, and each microRNA in the microRNA library for both digital and conventional counting as a function of the lower of the two input copy numbers for each gene. Although the digital counting result is imperfect because the barcode sampling is slightly biased and the set of barcode sequences is finite, digital counting vastly outperforms conventional counting in terms of accurately quantifying fold-changes between two gene expression profiles in both cases. We note that the copy-number distributions in the two cases differ significantly. Copy numbers fall between 1 and 20 for the *E. coli* fragment library and between 1 and ∼20,000 for the microRNA library, but digital counting is advantageous in both cases. In addition, the performance of conventional counting vs. digital counting is significantly improved for the *E. coli* transcription unit simulation compared with the fragment library because the PCR noise for each transcription unit is averaged over several fragments. This finding is consistent with the improved correlation of conventional vs. digital counting depicted in Fig. S2 compared with the fragment correlation shown in Fig. 3*C*. The overall performance discrepancy between conventional and digital counting and the increase in conventional counting accuracy as a function of copy number are expected from the amplification noise that is evident in Fig. 3*C*. We conclude that our digital counting scheme will be advantageous for differential expression analysis, particularly for applications involving small fragments as in the microRNA case simulated here.

In-Depth Design of Barcode. Each of the 2,358 barcodes was analyzed for sequence characteristics that would contribute to either amplification or sequencing errors.

Initial filtering. All barcodes with less than 40% or greater than 60% GC-content or containing homopolymers greater than length four were deleted. All barcode sequences were compared with the PE 1.0 and 2.0 Illumina PCR primer sequences and discarded if there were more than 10 total base matches or more than five consecutive base matches in any possible alignment with either primer sequence (sense and antisense). Each barcode was compared with the final four, five, and six bases closest to the 3′ end of the PCR primer sequence (sense and antisense), respectively. The final six bases for both the PE 1.0 and PE 2.0 are identical. If any of these regions contained more than three consecutive base matches with a given barcode in any possible alignment, that barcode was discarded. All barcode sequences were compared with all other barcode sequences (including cases of offset sequences), and a barcode was discarded if there were more than 15 total base matches or 10 consecutive base matches to any other barcode sequence (sense and antisense) in any possible alignment. The total number of hydrogen bonds in the longest consecutive matching region of each barcode to any other barcode sequence was calculated, and barcodes with greater than 26 total hydrogen bonds in that region were discarded. Each barcode was aligned with the entire *E. coli* genome (sense and antisense). If at any position in the genome a barcode contained more than 16 total matching bases, 12 consecutive matching bases, or 32 hydrogen bonds present in any given consecutively matching region, that barcode was discarded. Finally, all possible indels were generated for each barcode and the resulting sequence was compared with each original barcode sequence. If the resulting indel sequence could incur fewer than five point mutations and result in the exact sequence of any

IDT Exhibit 1030.0008

barcode, the barcode sequence that generated the given indel was deleted.

Score filtering. In addition to these universal thresholds, a more in-depth analysis of barcode-barcode and barcode-*E. coli* genome hybridization was performed, particularly regarding barcode hybridization melting temperatures with respect to PCR amplification. Although for a given barcode sequence, when comparing complementarity to a large set of reference sequences through sequence alignment, there will be an alignment condition that results in a region in the barcode sequence where the absolute maximum number of consecutive base matches is achieved (as described above), there are other possible conditions when a barcode contains a region where the number of consecutive bases in the region of maximum consecutive base matches does not reach the absolute maximum value, as described above. We shall define this value for any given region as the score. When comparing a barcode to the sense and antisense sequences of all other barcodes, we determined all alignment conditions in the cases where the score of the region that contained the highest number of consecutive base matches (first score) were 10, 9, 8, and 7 bases, respectively. For each of these four first scores, the condition where the maximum score of the region where the second-highest number of consecutive base matches (second score) occurred was determined. For example, we shall denote a condition where the first score is 10 and the second score is 3 as "10-3" and define this condition as the duplex. If the sum of the first score and the second score compared with all other barcode sequences for any barcode was greater than 12, that barcode was discarded. If the distance between these two regions (maximum and second-most consecutive matches) was one base, and the sum of the first score and the second score was greater than 11, the barcode was discarded. The maximum value of the number of consecutive base matches for a region under all alignment conditions that contained the third-highest number consecutive base matches (third score) was also determined for all barcodes. Given the maximum third score, the respective maximum first score was determined; the maximum second score given both of these conditions was also determined. We shall define this condition as the triplex and denote it, for example, as 7-4-3. We manually deleted the barcodes with the following triplexes: 7-3-3, 6-5-4, 6-5-3, 6-4-4, 5-5-4, and 5-4-4. We also manually deleted barcodes with a triplex of 6-4-3, where both the distances between adjacent regions corresponding to the scores was one base.

The same analysis was done for all barcodes aligned against the entire *E. coli* genome (sense and antisense). Barcodes with a first score and second score sum of greater than 15 were discarded. Barcodes where the first score region and the second score region were separated by one nucleotide and had a first score and second score sum of more than 14 were discarded as well. Barcodes with the following triplexes were deleted: 8-4-4, 7-5-4, 6-6-4, 6-5-5, and 5-5-5. After filtering, a total of 150 barcodes remained.

In-Depth Design and Preparation of Adapter. ***Adapter design.*** To avoid sequencing errors resulting from cluster overlap (i.e., low sequence complexity) and to reduce potential ligation bias, an additional two- to five-bp extension—CT, ACT, GACT, or TGACT—was added to the 3′-end of each barcode. These sequences mimic the T-overhang in the conventional Illumina paired-end adapter and conserve the sequence of the last two bases. For each of the 150 final barcodes, we attached these four different adapter extensions to the 3′ end of the barcode. We then obtained the same values as used in the initial filtering step (see above) for each of the four adapter candidates for all 150 barcodes. We determined the following four parameters of analysis: PCR primer matching (PC), 3′ end of PCR primer matching (TP), barcode-barcode matching (BB), and barcode *E. coli* genome matching (EC), and calculated the complementarity score of each category for all barcode-adapter candidates as follows:

PC = {Sum of [maximum total base matches to the PE 1.0 and PE 2.0 PCR primers (sense and antisense for a total of four terms)]} + 2·{Sum of [maximum consecutive base matches to the PE 1.0 and PE 2.0 PCR primers (sense and antisense for a total of four terms)]^2}

TP = {Sum of [maximum total base matches to the final four bases of the PCR primer sequence (sense and antisense for a total of two terms)]^2} + 1.5·{Sum of [maximum total base matches to the final five bases of the PCR primer sequence (sense and antisense for a total of two terms)]^2} + 2·{Sum of [maximum total base matches to the final six bases of the PCR primer sequence (sense and antisense for a total of two terms)]^2}

BB = {Sum of [maximum total base matches to all other barcode candidates (sense and antisense for a total of two terms)]^2} + 2·{Sum of [maximum consecutive base matches to all other barcode candidates (sense and antisense for a total of two terms)]^2}

EC = Maximum total base matches to entire *E. coli* genome (sense only) + 2· [maximum consecutive base matches to the entire *E. coli* genome (sense only)]^2

The total complementarity score (TC) for each barcode candidate was calculated as follows:

TC = 3·PC + 15·TP + BB + EC

The TC value gives us a metric to determine the expected efficacy of each barcode candidate during PCR amplification. A low TC represents a lower chance of amplification errors caused by unwanted hybridization between barcodes and adapters, primers, or the sample. For each barcode, we selected the barcode-adapter candidate that had either the lowest or second-lowest TC among the four: this resulted in 150 final barcode-adapter sequences, of which 145 were randomly chosen and used.

Thirty-seven CT extensions and 36 of each of the other three extensions were used. Adapters were then designed in the same Y-shaped construct as the conventional Illumina paired-end adapter with a 22- to 25-bp extension that contained the barcode and a T-overhang (Fig. 1*B*). Both strands (A and B) of the adapter were ordered from Integrated DNA Technologies (Dataset S1).

Adapter generation. The 5′-end of strand B was phosphorylated in T4 DNA Ligase Reaction Buffer (NEB) containing 40-μM strand B and 20 U T4 polynucleotide kinase (NEB) at 37 °C for 60 min in a 20 μL volume, followed by a 25-min incubation at 70 °C, and a 5-min incubation at 90 °C for enzyme inactivation. The phosphorylated strand B was annealed to each respective strand A in NEB Buffer 2 (NEB). Each solution contained 20-μM strand A and 20-μM strand B in a total volume of 20 μL. The solutions were first raised to 90 °C and cooled to 25 °C at a rate of 5 °C per minute (annealing temperature condition). Finally, equal volumes of all 145 annealed adapters were mixed.

Design and preparation of spike-in and "normalization" DNA. Fifteen-thousand random 30-bp sequences were generated such that even if a sequence accumulated 15 mutations, it would still be identifiable and distinguishable from all other generated sequences. Spike-in and normalization candidates with a maximum homopolymer length of greater than 3, or a GC-content less than 11 or greater than 19 were discarded. Spike-in and normalization candidates were also discarded if they exceeded a certain degree of complementarity or sequence identity (total matches and maximum consecutive matches) with (*i*) the Illumina paired-end sequencing primers. (*ii*) the 3′-end of the sequencing primers, (*iii*) the whole *E. coli* genome [K-12 MG1655 strain (U00096.2)], and (*iv*) all other generated spike-in candidates in the same fashion as barcode design. The final population consisted of 40 spike-in and normalization DNA candidates, of which three were chosen at random seven times (without replacement) and con-

IDT Exhibit 1030.0009

catenated, with one deletion at the 60th base of strand A (corresponding to the 31st base of strand B) and an addition of a single A to the end of the sequence to form seven 90-base spike-in DNA sequences and one normalization DNA sequence. Both strands of the 5′-end phosphorylated DNA oligos were ordered from Integrated DNA Technologies (Dataset S4) and were annealed in 0.3× NEB Buffer 2 (NEB) with 50 μM of each strand using the annealing temperature gradient. All seven spike-in sequences were ligated to the barcoded adapter mixture in NEB Next Quick Ligation Reaction Buffer (NEB) with 6.7 μM annealed spike-in, 6.7-μM barcoded adapter, and 6 μM Quick Ligase (NEB) by incubating at 25 °C for 30 min. The product was run on a 5% polyacrylamide gel (Bio-Rad), and the targeted band (at ∼270 bp) was removed from the gel. The gel slice was cut into small pieces and the embedded DNA was extracted into diffusion buffer (10 mM Tris-Cl pH 8.0, 50 mM NaCl, 0.1 mM EDTA) by overnight incubation at room temperature. Then, the extracted spike-ins were purified on a column (Zymo Research). Sequence analysis (GeneWiz) confirmed that the band contained the expected ligation product. The concentration of each spike-in was estimated by qPCR (Fast SYBR Master Mix; Applied Biosystems) using sequence-specific qPCR primers (Dataset S4) against a known-concentration Y-shaped Reference DNA (below). The concentrations of spike-ins for the second deepsequencing run were measured in parallel by digital PCR (Fluidigm) at the Molecular Genetics Core Facility of Children's Hospital Boston Intellectual and Developmental Disabilities Research Center. Each spike-in was measured a total of ten times on two separate chips (48.770).

Design and Preparation of Y-Shaped Reference DNA. From the original list of 150 barcode candidates, we chose two barcodes that were not present in the final list of 145 used. Then, we concatenated the two barcodes with the Y-shaped adapter sequences and a 90-bp targeted sequence mimic such that the targeted sequence mimic was between the barcodes, which were between the adapters (Dataset S4). The 90-bp targeted-sequence mimic was designed the same way as the spike-in and normalization DNAs. Both strands of the DNA oligos were ordered from Integrated DNA Technologies and their concentrations were measured by absorbance at 260 nm using the extinction coefficient provided by Integrated DNA Technologies. The DNA oligos were annealed in water with 5 μM of each strand using the annealing temperature gradient.

***E. coli* RNA Preparation and cDNA Generation.** *E. coli* [K-12 MG1655 strain (U00096.2)] was grown overnight at 30 °C in LB medium. The resulting culture was diluted 500-fold in fresh LB medium and grown at 30 °C for 3.5 h, such that the O.D. at 600 nm became 0.30–0.35. One milliliter of cells were quickly killed by addition of 0.1 mL stop solution [90% (vol/vol) ethanol and 10% (vol/vol) phenol]. The cells were collected by centrifugation (9,100 × *g*, 1.5 min, room temperature), suspended in 1 mL cooled PBS (Lonza), and centrifuged again (16,000 × *g*, 1.5 min, room temperature). The supernatant was removed and the cells were suspended in 0.1 mL of 1 mg/mL lysozyme in TE Buffer (pH 8.0) (Ambion). Next, 0.1 mL of lysis buffer (Genosys) was added and the mixture was vortexed for 5 s. After adding 0.2 mL of phenol chloroform pH 4.5 (Sigma) and vortexing three times for 5 s, the mixture was centrifuged (16,000 × *g*, 3 min, room temperature). The top layer of solution was taken and 0.15 mL of 100% 2-Propanol (Sigma) was added; the mixture was left on ice for 30 min. The solution was centrifuged (16,000 × *g*, 30 min, 4 °C) to precipitate the RNA. The RNA pellet was washed twice by centrifugation (16,000 × *g*, 5 min, 4 °C) with 0.75 mL of cold 70% (vol/vol) ethanol. After the second centrifugation, the supernatant was removed and the pellet was dried for 15 min at room temperature. Then, 88 μL of water was added and the mixture was incubated for 15 min at room temperature, followed by resuspension. The resulting solution was mixed with 0.04U/μL DNase I (NEB) in DNase I Reaction Buffer (NEB) for a total volume of 100 μL, and the mixture was incubated at 37 °C for 30 min followed by addition of EDTA (Sigma) to a final concentration of 5 mM. The mixture was incubated at 75 °C for 10 min to inactivate DNase I, followed by column purification. Ribosomal RNA was removed using Ribo-Zero rRNA Removal Kit (Gram-negative bacteria) (Epicentre, Illumina). From this point, we followed the conventional Illumina protocol for mRNA sequencing sample preparation with a few modifications. The purified RNA was fragmented in 0.5× fragmentation buffer (Ambion) with ∼500 ng RNA in a 100-μL reaction solution. The solution was incubated on ice for 1 min after the fragmentation buffer was added, followed by a 6-min incubation at 70 °C. The tube was placed on ice and incubated for 1 min followed by addition of 4-μL stop solution (Ambion). The fragmented RNA was purified with a column and eluted in 11.1 μL in water. One microliter of 50 μM Random Hexamer Primer (Applied Biosystems) was added to this solution and incubated at 65 °C for 5 min and then placed on ice. Four microliters 5× First Strand Buffer (Invitrogen), 2 μL 100 mM DTT (Invitrogen), 0.4 μL 25 mM dNTP Mix (Applied Biosystems), and 0.5 μL RNase inhibitor (Applied Biosystems) was added to the mixture. This mixture was then incubated at 25 °C for 2 min, followed by the addition of 1 μL SuperScript II (Invitrogen). The mixture was then incubated at 25 °C for 10 min, 42 °C for 50 min, and 70 °C for 15 min to synthesize the first strand of the cDNA and inactivate the enzyme, which was placed on ice and then purified on a column. The eluate from this column was used to generate the second strand of cDNA in NEB Next Second Strand Synthesis Reaction Buffer (NEB) with 0.3 U/μL DNA polymerase I (*E. coli*) (NEB), 1.25 U/μL *E.coli* DNA Ligase, and 0.25 U/μL RNase H in an 800 μL total volume solution at 16 °C for 2.5 h, followed by the column purification. The eluted double stranded cDNA was end-repaired in T4 DNA Ligase Buffer (NEB) with 0.4 mM Deoxynucleotide Solution Mix (NEB), 0.5 U/μL T4 DNA polymerase (NEB), 0.5 U/μL T4 Polynucleotide Kinase (NEB) in a 200 μL reaction solution by incubating at 20 °C for 30 min followed by column purification. The eluted end-repaired cDNA was dA-tailed in NEB 2 buffer (NEB) with NEB Next dA-tailing Reaction Buffer with 1mM dATP (NEB) and 0.3 U/μL Klenow Fragment (3′ → 5′ exo-) in a 50-μL solution by incubating at 37 °C for 30 min followed by column purification.

Sample-Adapter Ligation, Sequencing Sample Preparation, and Sequencing. The cDNA library was ligated to the barcoded adapter mixture and the conventional Illumina paired-end adapter (without phosphorothioate bond) (Integrated DNA Technologies) respectively in the NEB Next Quick Ligation Reaction Buffer (NEB) with 5.4 μL of cDNA produced above, 1.9 μM barcoded adapter (or conventional Illumina paired-end adapter), and 3.6 μM Quick Ligase (NEB), in a total volume of 10 μL by incubating at 25 °C for 15 min. The two solutions were separately run on a 5% polyacrylamide gel and the portion between 250 and 300 bp was cut. The gel slice was cut into small pieces and the embedded DNA was extracted into diffusion buffer (10 mM Tris-Cl pH 8.0, 50 mM NaCl, 0.1 mM EDTA) by overnight incubation at room temperature. Then, the extracted DNAs were column purified. The concentrations of the purified products were measured by qPCR (Fast Fast SYBR Master Mix; Applied Biosystems) against a known-concentration Y-shaped reference sequence (Dataset S4) using our designed qPCR primers (Dataset S4) purchased from Integrated DNA Technologies. The sample ligated to the barcoded adapter and the conventional Illumina Paired-end adapter were amplified by PCR (1 cycle of 98 °C for 1 min, 18 cycles of 98 °C for 1 s, 65 °C

IDT Exhibit 1030.0010

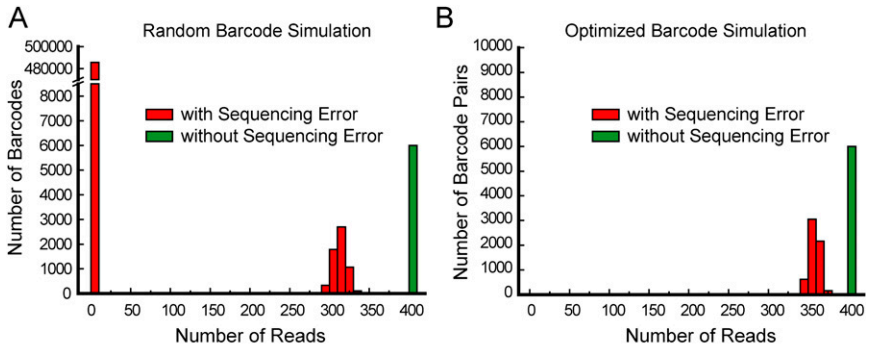
for 45 s, 72 °C for 40 s, and 1 cycle of 72 °C for 5 min) in HF buffer (NEB) with 0.63 mM dNTP, 0.5 mM of each amplification primer modified from the Illumina PCR primers PE 1.0 and 2.0 (Integrated DNA Technologies) (Dataset S4), 25 fM DNA sample, and Phusion DNA polymerase (NEB) in 20 μL, with spike-in DNAs (0.71 aM Spike-in 1, 1.0 aM Spike-in 2, 4.4 aM Spike-in 3, 18 aM Spike-in 4, 150 aM Spike-in 5, 480 aM Spike-in 6 for the first sequencing run, and 1.3 aM Spike-in 1, 6.9 aM Spike-in 3, 36 aM Spike-in 4, 150 aM Spike-in 5, 770 aM Spike-in 7 for the second sequencing run). Then, 10 pM Normalization DNA was added to both PCR products, and the DNA was purified twice on a column. The concentration of the purified product was measured by qPCR (Fast Fast SYBR Master Mix; Applied Biosystems) using the conventional Illumina qPCR primers (Integrated DNA Technologies) against a PCR product amplified from Y-shaped reference DNA using modified Illumina PCR primers whose concentration was measured by NanoDrop (LMS). The final concentration of each spike-in and normalization DNA in the purified products was measured by qPCR using sequence-specific qPCR primers described above and compared by normalization. The length distribution of the purified PCR product was measured by Bioanalyzer (Agilent). Samples with barcoded adapters were sequenced on an Illumina HiSeq 2000 with 2 × 100 (for the first sequencing run) and 2 × 50 (for the second) base paired-end reads in one lane. Sequence data has been deposited in the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/, accession no. GSE34449).

Genome Viewer Files. We have generated genome viewer files (.sam format) and index files (.sai) for the IGV software for fragments from both the digital counting method and the conventional counting method. This software allows visualization of the difference between both methods in a genome-wide manner. The supplementary files have been deposited in the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/, accession no. GSE34449).

Spike-in Analysis. From the raw sequencing data, we isolated reads which contained barcode sequences that corresponded to our original list of 145 barcodes in both forward and reverse reads for each sequencing cluster that had at most one mismatch. We then aligned the first 28 bases (26 bases for the second sequencing run) of the targeted sequence of both the forward and reverse reads of each cluster to each spike-in sequence, which is known. Sequences with more than two mismatches were discarded. We then counted the number of unique tags present in each spike-in to determine the number of copies of each spike-in.

Comparison of Noise in Conventional vs. Digital Counting for the *E. coli* Transcriptome. We summed the total number of reads and the total number of digital counts for each base in each of the mapped sequences. For each transcription unit, we created bins that were 99-bp long and summed the total number of reads and the total number of digital counts present in each bin. Bins that yielded an average number of digital counts per base of greater than or equal to 1 were selected, and from these bins we calculated the average and sample SD of the summed reads and summed digital counts, respectively. We define the noise to be the sample SD divided by the mean and calculated this value for both reads and digital counts; the ratio of the noise for reads to digital counts was computed for each transcription unit (Fig. 3*D*).

1. Casbon JA, Osborne RJ, Brenner S, Lichtenstein CP (2011) A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Res* 39:e81.
2. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B (2011) Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci USA* 108:9530–9535.
3. Nakamura K, et al. (2011) Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res* 39:e90.
4. Ribeiro-dos-Santos A, et al. (2010) Ultra-deep sequencing reveals the microRNA expression pattern of the human stomach. *PLoS ONE* 5:e13205.



C

	Number of Barcodes or Barcode Pairs		
	Before Sequencing (Original Solution)	After Sequencing (Simulation)	
		w/o Sequencing Error	w/ Sequencing Error
Random Barcodes	6000	6000	491532
Optimized Barcodes	6000	6000	6000

Fig. S1. Histogram of unique barcodes from parallel simulation of the theoretical library using optimized barcodes and random barcodes (*SI Materials and Methods*). (*A*) A library of 6,000 identical molecules was barcoded *in silico* by random barcodes (single-end 16 nucleotides). (*B*) Same as *A*, but with optimized barcodes (paired-end 20 nucleotides). Both conditions were then "amplified" 400-fold. After accounting for sequencing error in both the barcode region and the mapping region, the resulting number of unique barcodes and the respective number of reads were histogrammed for each case. (*C*) In the random barcode case, many artifactual barcode sequences were generated by sequencing error and thus have a low number of reads. Sequencing error occurred for the optimized barcode sequence as well. However, because the optimized barcode sequences were known (i.e., unused sequences are known) and designed to have minimally overlapping sequences, the sequencing errors were identified. Then, sequenced reads which contained errors were removed (*Materials and Methods*), and there is no population of artifactual barcode sequences with a small number of sequencing reads.

IDT Exhibit 1030.0011

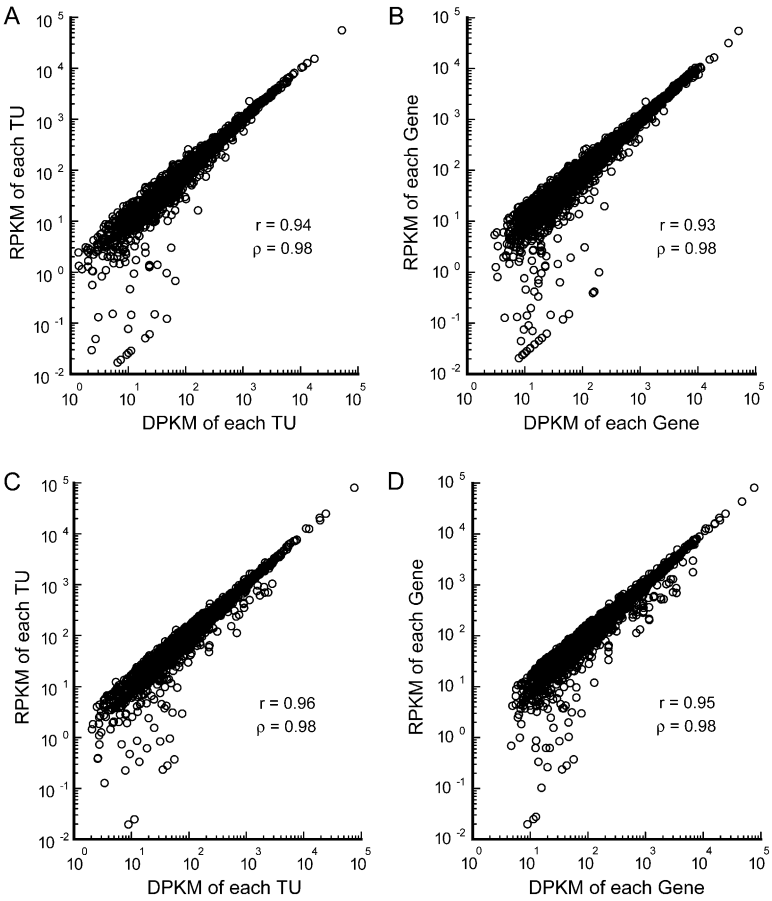


Fig. S2. (*A*) Comparison of uniquely mapped reads per kilobase of each transcription unit per million total uniquely mapped reads (RPKM) and uniquely mapped digital counts per kilobase of each transcription unit per million total uniquely mapped molecules (DPKM) for all detected transcription units. (*B*) The same plot as *A*, but for genes. (*C* and *D*) The same plots as *A* and *B*, respectively, from the second sequencing run. Genes and transcription units exhibit stronger correlations between RPKM and DPKM compared with individual fragment sequences (Fig. 3*C*) because of the averaging of reads and counts over long genes or transcription units. At higher copy numbers, analog reads and digital counts are well-correlated, but at lower copy numbers the data are not nearly as correlated (see Datasets S5, S6, and S7 for specific values for each datapoint).

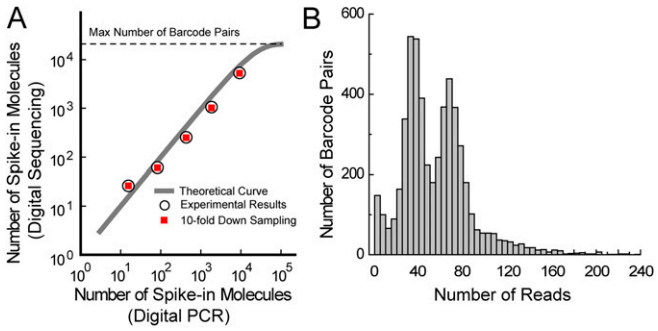


Fig. S3. Down-sampling of all spike-in data by a factor of 10 and the resulting digital counts obtained (*SI Materials and Methods*). (*A*) The correlation between the number of digital counts of all down-sampled spike-in molecules and the number of spike-in molecules measured by digital PCR. For each of the spike-in sequences, we still detected an average of 98% of the original number of unique barcodes. (*B*) The histogram of conventional reads for all unique barcodes in the down-sampled population of the highest copy spike-in (spike-in 7).

IDT Exhibit 1030.0012

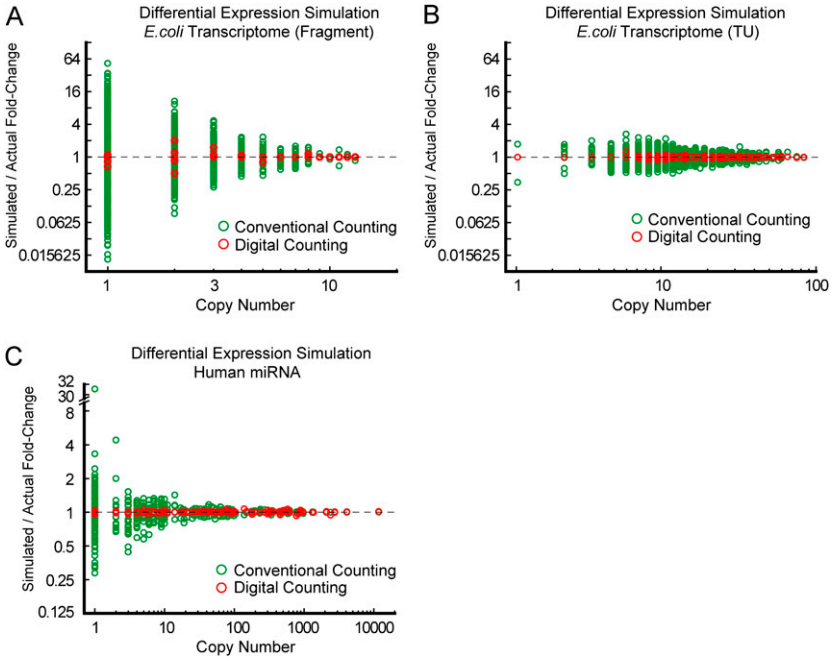


Fig. S4. Simulation demonstrating the superior performance of digital counting over conventional counting in differential expression analysis (*SI Materials and Methods*). RNA expression quantification was simulated using experimentally measured copy numbers, barcode sampling, and amplification noise distributions for two different libraries for each of three different systems [*E. coli* transcriptome fragments (*A*), *E. coli* transcription units (*B*), and human stomach microRNA (*C*)]. We plot the ratio of simulated to actual fold-change for each gene as a function of the lower of two copy numbers for the two compared libraries. Ideally, the value of this ratio is one for all genes. Because digital counting is almost completely immune to amplification noise, it gives consistently superior performance to conventional counting for differential expression, even at low copy numbers. We note that the discrepancy between conventional and digital counting is smaller for the *E. coli* transcription unit library in *B* than for the fragment library in *A* because amplification noise can be averaged over many fragments in the case of long transcription units.

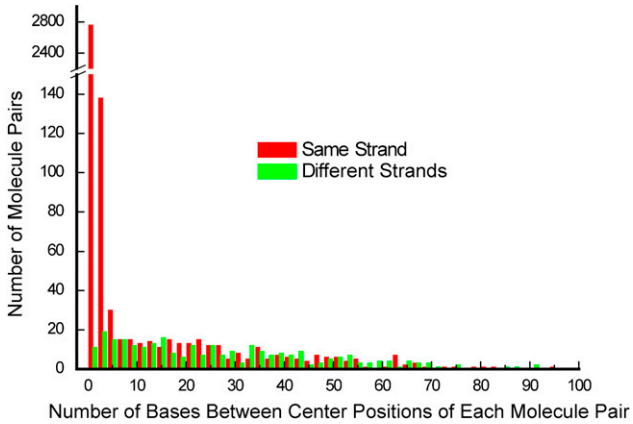


Fig. S5. Histogram of the number of bases between the center positions of all pairs of molecules mapped to the same transcription unit that contain the same barcode for pairs of molecules both mapped to the sense or antisense strand of the *E. coli* genome (red), and also for pairs of molecules mapping to different strands of the *E. coli* genome (green). We are able to distinguish the different strands of an original cDNA molecule because of the design of the paired-end sequencing adapters (Fig. 1*B*). The distribution for molecule pairs mapping to different strands is more uniform than the distribution for molecule pairs mapping to the same strand, which confirms that molecule pairs mapping to different strands on the *E. coli* genome are on average far apart from each other. From the distribution of molecule pairs mapping to the same strand, we determined that molecule pairs with fewer than four bases between center positions are identical molecules. This analysis allows us to keep 98% of all molecules but reduces the counting error rate.

IDT Exhibit 1030.0013

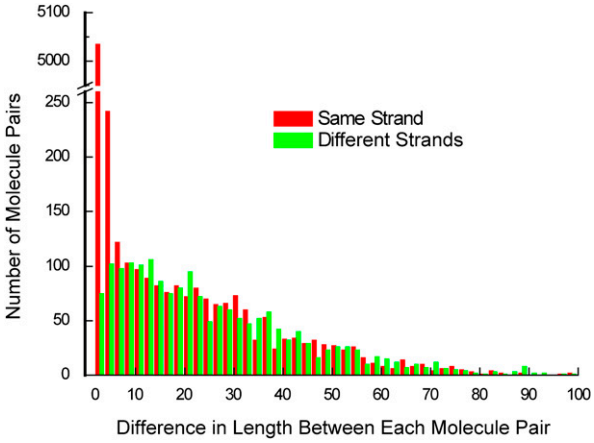


Fig. S6. Histogram of the difference in fragment length for all pairs of molecules mapped to the same transcription unit that contain the same barcode for pairs of molecules both mapped to the sense or antisense strand of the *E. coli* genome (red), and also for pairs of molecules mapping to different strands of the *E. coli* genome (green). The distribution for molecule pairs mapping to different strands is more uniform than the distribution for molecule pairs mapping to the same strand, which confirms that molecule pairs mapping to different strands on the *E. coli* genome are on average similar in size. From the distribution of molecule pairs mapping to the same strand, we determined that molecule pairs with a difference in fragment length of less than nine are identical molecules. This analysis allows us to keep 96% of all molecules but reduces the counting error rate.

Other Supporting Information Files

Dataset S1 (XLS)
Dataset S2 (XLS)
Dataset S3 (XLS)
Dataset S4 (XLS)
Dataset S5 (XLS)
Dataset S6 (XLS)
Dataset S7 (XLS)

IDT Exhibit 1030.0014