Digital RNA Sequencing Minimizes Sequence-Dependent Bias and Amplification Noise with Optimized Single-Molecule Barcodes

Classification:
Biological Science
Systems Biology

Digital RNA Sequencing Minimizes Sequence-Dependent Bias and Amplification Noise with Optimized Single Molecule Barcodes

Katsuyuki Shiroguchi, Tony Z. Jia, Peter A. Sims, X. Sunney Xie*

Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford St., Cambridge, Massachusetts 02138.

***Corresponding Author:**
X. Sunney Xie
Department of Chemistry and Chemical Biology, Harvard University
12 Oxford St., Cambridge, Massachusetts 02138.
Tel: 617-496-9925
Email: xie@chemistry.harvard.edu

Abstract

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 *E. coli* with more accurate and reproducible quantification than conventional RNA-Seq.

The central goal of transcriptome profiling is to accurately quantify the abundance of RNA transcripts in a sample. While hybridization-based approaches like DNA microarrays can provide only a relative, analog measure of transcript abundance, sequencing-based approaches such as 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), though 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 prior to 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 prior to 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 1) much larger than the copy number of the most abundant target sequence and 2) 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 non-random (18), unless the barcodes are carefully 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 base-pairs in length (Table 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*, *SI Materials and Methods*). If both ends of a target sequence sample all of the barcodes randomly, the target sequence will have access to 145x145 = 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 which 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 *E. 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, pre-determined 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 due to sequencing errors. This causes barcodes to interconvert, generating quantification artifacts which 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 due to sequencing errors. The effect of sequencing error on both random and optimized barcode counting is clearly shown by simulation (Fig. S1, *SI Materials and Methods*).

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 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 Table 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 (TU). Not surprisingly, the read density is considerably less uniform across the TU than the number of digital counts, presumably due to 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 10-40 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 TUs 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 TUs were, by-and-large, intact RNA molecules following RNA synthesis, the cDNA fragments that map to one region of a given TU should have the same abundance as fragments that map to a different region of the same TU. We histogrammed the ratio between the variation in conventional counting ν_C and variation in digital counting ν_D for TUs 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 TU. For a TU where ν_C/ν_D exceeds one, conventional counting measures abundance less consistently along the TU 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 standard deviation) and 1.2 (s=0.5) for the complete set of analyzed TUs, indicating that conventional

counting is less consistent than digital counting across an average TU. Furthermore, the mean value of ν_C/ν_D increases with decreasing copy number and its distribution becomes broader (Fig. 3*D*). For TUs 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 TUs.

While 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 TU in Fig. 4*A*, which shows 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 TUs and genes for both conventional and digital counting in Fig. 4*B* and Fig. 4*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 prior to amplification (10), obviating the need for ligation.

One disadvantage of our technique is that it requires higher sequencing coverage than conventional RNA-Seq. This 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 downsampled 10-fold (*SI Materials and Methods*) without perturbing the correlation between abundance measured by digital PCR and digital barcode counting (Fig. S3). This 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 (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 chromatin immunoprecipitation 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 Material and Methods, (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 (1) the Illumina paired-end sequencing primers (33), (2) the Illumina PCR primers PE 1.0 and 2.0, (3) the 3' end of the Illumina PCR primers PE 1.0 and 2.0, (4) the whole *E. coli* genome [K-12 MG1655 strain (U00096.2)], and (5) 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 (Table 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, Dataset S4*).

***E. coli* Transcriptome 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 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 four base-pairs and if the difference in length was less than 9 base-pairs (Fig. S5 and Fig. 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*).

ACKNOWLEDGEMENTS. This work was supported by the United States National Institutes of Health National Human Genome Research Institute Grant (HG005097-01) to X.S.X. and the United States National Institutes of Health National Human Genome Research Institute Recovery Act Grand Opportunities Grant (1RC2HG005613-01) to X.S.X. 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. Illumina sequencing was performed at the Harvard FAS 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 (IDDRC) with the help of Hal Schneider and Ta-Wei Lin. Base-calling analysis was performed at the Harvard FAS Research Computing Group by Jiangwen Zhang.

REFERENCES

1. Mortazavi A, et al. (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, et al. (2008) Substantial biases in 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, et al. (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, et al. (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, et al. (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* 460:479-486.
16. Casbon JA, et al. (2011) A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Res* 39:e81.
17. Kinde I, et al. (2011) Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci USA* 108:9530-9535.
18. Fu G.K, et al. (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* online publication (doi:10.1038/nmeth.1778).
20. Hamady M, et al. (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Methods* 5:235-237.
21. Wang Z, et al. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10:57-63.
22. Au KF, et al. (2010) Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res* 38:4570-4578.
23. Lau NC, et al. (2001) An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* 294:858-862.
24. Axtell MJ, et al. (2006) A two-hit trigger for siRNA biogenesis in plants. *Cell* 127:565-577.
25. Core LJ, et al. (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, et al. (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, et al. (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:D583-D590.

Figure Legends

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.

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 due to 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 leftmost 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 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*). The inset shows 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.

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 which 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 TU as a function of TU 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 standard deviation of the digital counts among 99-base bins in a TU, respectively) associated with digital counting and the variation $\nu_C = s_C/\mu_C$ associated with conventional counting within each TU 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 TUs in different abundance ranges for each replicate. TU abundance is the sum of all digital counts for each fragment in the TU.

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 TU. As expected, the ratio fluctuates over a broader range for conventional counting than digital counting along the length of the TU. (*B*) Correlation between replicate sequencing runs for digital and conventional counting of TUs. 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.

FOOTNOTES

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

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.

All supplementary tables (S1-S7) may be found at the following location: http://bernstein.harvard.edu/digitalRNA-Seq/

The authors declare that Harvard University has filed a provisional patent application based on this work.