analytical chemistry

Letter

pubs.acs.org/ac

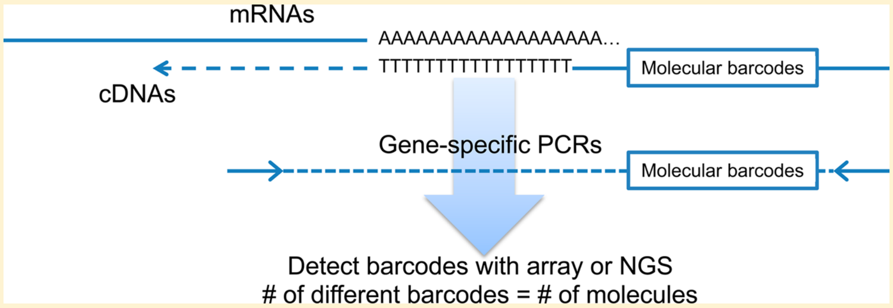
Terms of Use

Digital Encoding of Cellular mRNAs Enabling Precise and Absolute Gene Expression Measurement by Single-Molecule Counting

Glenn K. Fu,* Julie Wilhelmy, David Stern, H. Christina Fan, and Stephen P. A. Fodor*

Cellular Research Inc., 3183 Porter Drive, Palo Alto, California 94304, United States

Supporting Information



ABSTRACT: We present a new approach for the sensitive detection and accurate quantitation of messenger ribonucleic acid (mRNA) gene transcripts in single cells. First, the entire population of mRNAs is encoded with molecular barcodes during reverse transcription. After amplification of the gene targets of interest, molecular barcodes are counted by sequencing or scored on a simple hybridization detector to reveal the number of molecules in the starting sample. Since absolute quantities are measured, calibration to standards is unnecessary, and many of the relative quantitation challenges such as polymerase chain reaction (PCR) bias are avoided. We apply the method to gene expression analysis of minute sample quantities and demonstrate precise measurements with sensitivity down to sub single-cell levels. The method is an easy, single-tube, end point assay utilizing standard thermal cyclers and PCR reagents. Accurate and precise measurements are obtained without any need for cycle-to-cycle intensity-based real-time monitoring or physical partitioning into multiple reactions (e.g., digital PCR). Further, since all mRNA molecules are encoded with molecular barcodes, amplification can be used to generate more material for multiple measurements and technical replicates can be carried out on limited samples. The method is particularly useful for small sample quantities, such as single-cell experiments. Digital encoding of cellular content preserves true abundance levels and overcomes distortions introduced by amplification.

Single cell gene expression studies have increased our understanding of the function of individual cells in normal development and disease. Due to the limited messenger ribonucleic acid (mRNA) amount in a cell, ultrasensitive methods are necessary for reliable detection. Single cell RNA sequencing (RNA-seq) has been applied for profiling entire transcriptomes.[1−7] Many methods use Moloney murine leukemia virus (MMLV) template switching (TS) to incorporate a universal primer site during oligo-dT primed complementary DNA (cDNA) synthesis when reverse transcription (RT) reaches mRNA 5′ ends.[8] Universal polymerase chain reaction (PCR) is subsequently applied to amplify cDNAs for sequencing. TS has become increasingly popular because of fewer protocol steps, high overall efficiency, and the ability to represent full-length transcripts. However, questions regarding the sensitivity, accuracy, and technical reproducibility remain.

While RNA-seq is a powerful hypothesis-free global sampling tool for single cell experiments, directed approaches for validation and absolute quantitation of individual gene targets in single cells are lacking. Quantitative polymerase chain reaction (qPCR) has been the gold standard, but the small amount of single cell material limits measurements across multiple genes. To overcome, a "pre-amplification" step is frequently performed where several genes are first coamplified by multiplex PCR to generate adequate material for downstream qPCR analysis of individual genes. Unfortunately, preamplification introduces unpredictable bias in the relative amplitude of different genes, prohibiting absolute quantitation of the original mRNAs.[9,10]

To ascertain more precisely gene abundance levels, we have developed a molecular indexing (MI) approach for sensitive quantitation of mRNAs across multiple genes in single cells. Individual mRNA molecules are labeled at random from a pool of 960 sequence barcoded oligo-dT primers during RT, after which gene(s) of interest are amplified by PCR (Figure 1A). When tested on control poly-A RNA templates, the cDNA synthesis yields using these barcode-tailed oligo-dT primers are indistinguishable from standard oligo-dT primers. For

Received: February 1, 2014
Accepted: March 3, 2014
Published: March 3, 2014

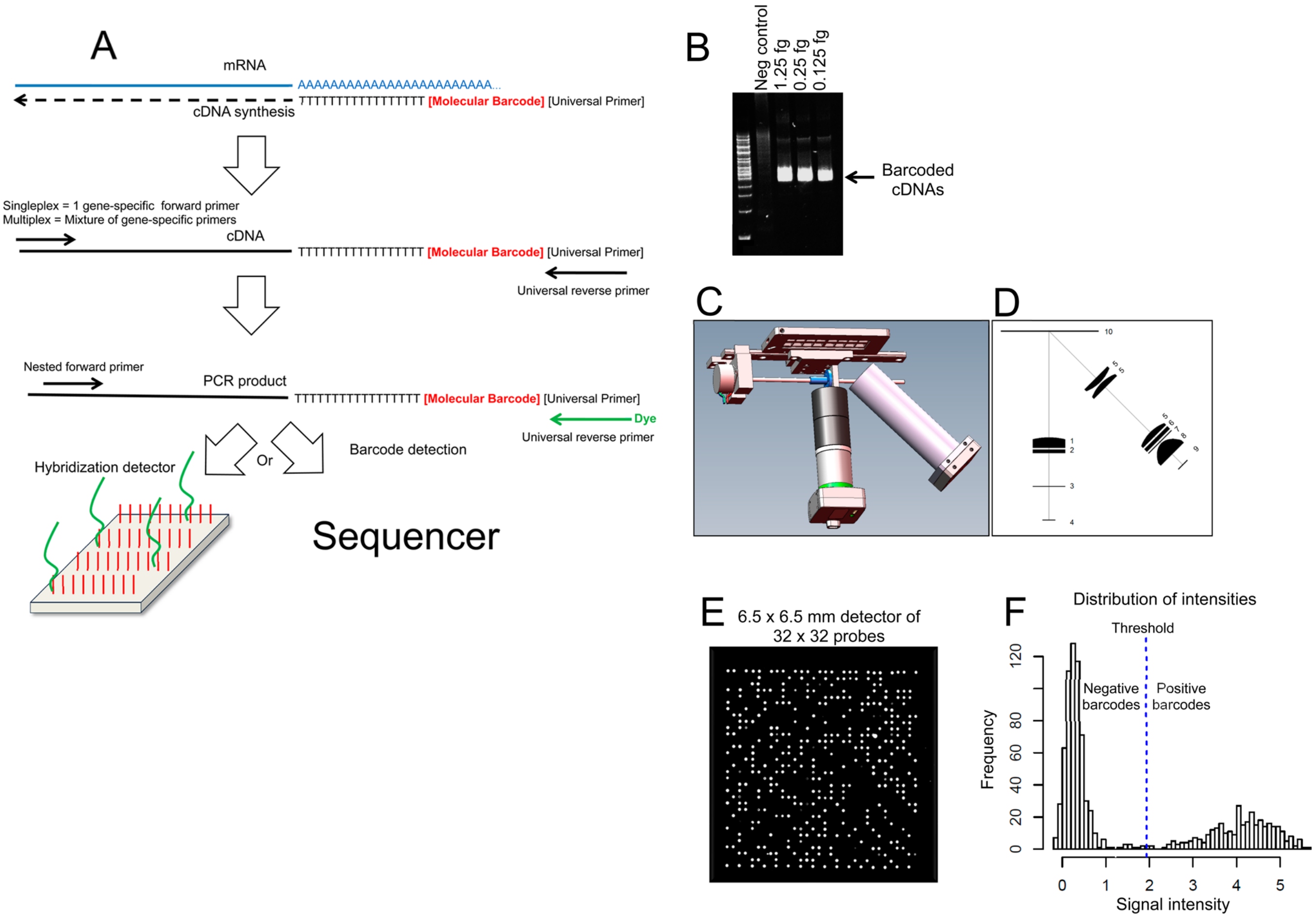

A
mRNA
AAAAAAAAAAAAAAAAAAAAAAA...
cDNA synthesis
TTTTTTTTTTTTTTTTT [Molecular Barcode] [Universal Primer]
Singleplex = 1 gene-specific forward primer
Multiplex = Mixture of gene-specific primers
cDNA
TTTTTTTTTTTTTTTTT [Molecular Barcode] [Universal Primer]
Universal reverse primer
Nested forward primer
PCR product
TTTTTTTTTTTTTTTTT [Molecular Barcode] [Universal Primer]
Dye
Universal reverse primer
Or
Barcode detection
Hybridization detector
Sequencer
B
Neg control
1.25 fg
0.25 fg
0.125 fg
Barcoded cDNAs
C
D
E
6.5 x 6.5 mm detector of 32 x 32 probes
F
Distribution of intensities
Threshold
Negative barcodes
Positive barcodes
Frequency
0
20
40
60
80
120
0
1
2
3
4
5
Signal intensity

Figure 1. Transcript counting using MI. (A) mRNAs are labeled using barcoded oligo-dT primers. cDNAs of interest are amplified with gene-specific and universal PCR primers (B). Dye labeled products generated by nested PCR are hybridized to a barcode detector and imaged (E) and counted (F) on a custom-built fluorescent CCD instrument. (C) Perspective view of the imaging and illumination optical train, and translation stage. (D) Optical components: achromatic cemented doublet lens (1), emission bandpass filter (2), camera lens (3), CCD (4), plano-convex lenses (5), excitation bandpass filter (6), rectangular aperture (7), aspheric lens (8), LED (9), and barcode detector (10).

detection, a dye labeled primer is used for PCR (Figure 1B) and barcodes are identified by fluorescence imaging after hybridization to a printed array of complementary probes (Figure 1C–E). For each starting molecule, PCR generates nanomolar concentrations of amplified DNA per barcode, which is approximately 100–1000× higher than the detection sensitivity of array hybridization. Alternatively, barcodes can be detected by sequencing. Although different genes amplify to variable degrees, the number of distinct barcodes present is largely unaffected. Counting the different barcodes reveals the absolute transcript copies (Figure 1F).[11]

Accurate quantitation by MI was established using serial dilutions of a synthetic control RNA spiked into a large background of *E. coli* RNA (Figure 2A), and absolute copy number measurements closely correlating with input were obtained (Pearson R-square = 0.9982). *GAPDH* was measured in serial dilutions of human liver total RNA and found to be indistinguishable within the experimental errors to digital PCR (Pearson R-square = 0.9997) (Figure 2B). Primarily designed for sensitive single cell measurements, the set of 960 barcodes affords a practical measurement dynamic range of approximately 10 to 4000 copies. However, highly concentrated RNA targets can also be tested by prediluting to this range prior to measurement.

Next, we tested absolute transcript quantitation for a set of high, medium, or low abundance genes in 12.5, 750, or 12 500 pg of liver total RNA, respectively (Supporting Information, Figure S1A–C). Each gene was measured alone in singleplex assays and in multiplex format (coamplification of all genes, followed by individual gene detection). Measurements agreed well between the two formats, demonstrating that multiple genes may be coassayed without compromising quantitation accuracy and the ability to obtain absolute counts.

Technical reproducibility is a key requisite of single cell analysis. Significant amplification and manipulation steps are often necessary when working with limited amounts of cell material, resulting in increased technical noise that masks detection of small biological variations among cells. To test MI directly on single cells, we selected individual K562 cells and lysed them in a PCR tube. The lysate volume was divided equally into 2 tubes, and *GAPDH* was measured. *GAPDH* counts varied from cell to cell as expected,[12] but technical replicates from the same cell were remarkably similar (Figure 2C).

To detect gene expression changes from a biological response, we induced K562 cells with hemin to increase fetal hemoglobin (*HBG2*) synthesis.[13] The absolute quantity of *GAPDH* and *HBG2* was measured for each of 10 induced or untreated cells. A large variation in levels of *HBG2* was

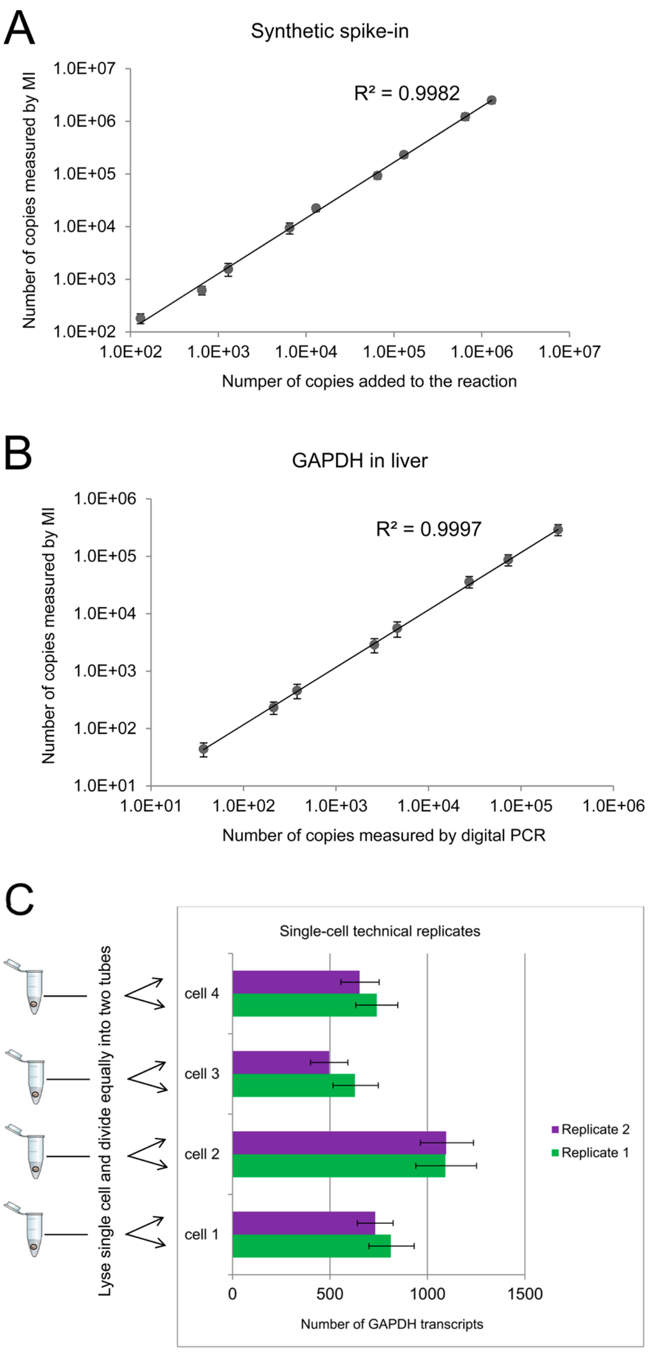


Figure 2. MI measurement accuracy. (A) Measured vs input copies of serial dilutions of a synthetic spike-in RNA. (B) *GAPDH* mRNA measurements in serial dilutions of liver total RNA vs digital PCR. (C) Technical replicate measurements of *GAPDH* directly in single K562 cell lysates. Absolute counts for each half cell volume are shown. Error bars show 95% measurement confidence intervals.

observed in untreated cells (336 to 4336 copies, Supporting Information, Figure S2). Although *GAPDH* levels also varied, the range was consistent with cell to cell variation at 456−1328 copies. After hemin treatment, there was no significant change to the *GAPDH* expression (176−1536 copies), but large increases in *HBG2* were detected (up to 14 952 copies). Measured absolute *GAPDH* counts and *HBG2* induction levels agree well with previous reports.[14,15]

As an alternative to hybridization, we have coupled MI with sequencing to test a larger number of genes. To determine accuracy, we performed standard global RNA-seq analysis[16] on 500 ng of a bulk lymphocyte RNA sample and calculated RPKM values from 2.9 × 10^6 mapped reads. The sample was diluted, and oligo-dT barcoded cDNA from 10 pg of RNA was used as template for multiplex PCR of 96 human genes (Supporting Information, Table S1). At this point, individual genes may be tested from the PCR product by scoring barcodes on the hybridization detector for one-off measurements or for presequencing QC. We sequenced the PCR product, and barcodes were counted to determine absolute copy numbers using 3.99 × 10^6 mapped reads. The numbers of reads (Figure 3A) or molecules (Figure 3B) from the 10 pg sample were compared with RPKM values from the 500 ng bulk sample. To avoid stochastic losses of rare transcripts from sample dilution, only higher abundance genes (RPKM > 30) were included for comparison. Although RPKM is a relative measurement subject to PCR bias, its correlation to the 10 pg sample is significantly higher for molecules (R-square = 0.7529) than reads (R-square = 0.4566), demonstrating increased measurement precision and decreased sensitivity to amplification noise when counting molecules instead of reads.[4,11,17−19] In addition, high quantitation precision for rare transcripts was established by

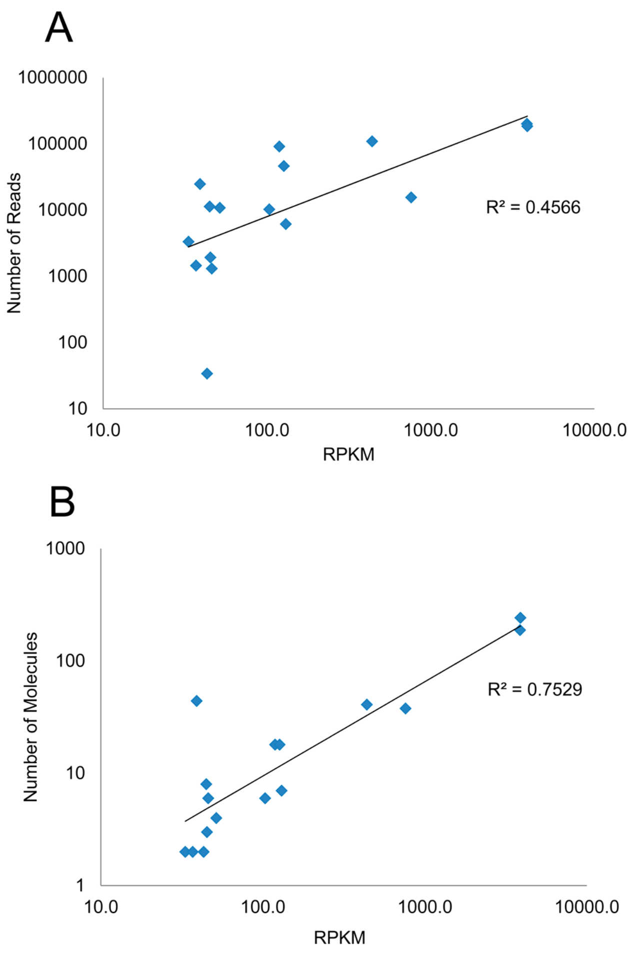


Figure 3. Gene expression analysis by MI and sequencing. Ten pg (~1 cell equivalent) of lymphocyte total RNA was barcoded during RT, and 96 genes were amplified by multiplex PCR and sequenced. Numbers of mapped reads (A) or molecule counts (B) are compared with RPKM values from conventional RNA-Seq of 500 ng (~50 000 cells) of the same sample (for genes >30 RPKM).

accurate measurements of small numbers of spike-in control RNAs in MI assays (Supporting Information, Figure S3).

MI and sample indexing can be used together to determine sensitivity and reproducibility across a plate of 96 samples. We incorporated sample barcodes on the oligo-dT primers (Supporting Information, Figure S4, Table S2) and pipetted into each well equal amounts of 5 synthetic bacterial control RNAs mixed with single cell amounts of various human RNAs. After RT, samples were pooled for multiplex PCR and sequencing. On average, 76.3% of the control molecules added were detected (Supporting Information, Figure S5), which is 25 to 28 percentage points or about 50% higher than the detection efficiency reported for TS methods.[4,7] The large increase in detection efficiency is not surprising because gene-specific PCR circumvents losses arising from TS (inefficient strand-switching or RT not reaching 5′ mRNA ends). Technical reproducibility (measured by standard deviations, Supporting Information, Figure S5) is about an order of magnitude better than TS for ~100 input RNA molecules.[2] MI also provided a slightly better coefficient of variation (23.5%) than RT-qPCR.[20]

Several single cell RNA-seq improvements have been described recently. Smart-seq2[5] includes modifications to increase cDNA yields, and high detection efficiencies have been reported in nanoliter volumes.[4,7] In addition, 5 random base barcodes on the TS oligo were demonstrated as a useful tool to correct PCR bias and to obtain absolute quantitation.[4] Establishing absolute yields and efficiencies for each transcript in a global approach is a difficult if not unachievable task. It is therefore very useful to have an independent means to establish absolute efficiencies and numbers. In the method described here, we use gene-specific primers for cDNA synthesis to avoid the inefficiencies of TS. In our hands, TS efficiency measurements with barcoded synthetic RNAs showed that of the cDNAs synthesized by RT, only ~14−23% were strand-switched (Supporting Information, Figures S6 and S7). Furthermore, it has been shown that changes in amplification conditions significantly alter expression measurements and create large numbers of artifacts when using TS.[5,21] As an independent method, we directly barcode cDNAs using oligo-dT priming so that truncated cDNAs are still detectable (in contrast, TS requires full-length cDNA synthesis). Finally, we have employed well selected 21 nucleotide molecular barcodes that are unambiguous in sequencing, which can also be counted by hybridization.

The use of both global RNA-seq and gene-targeted methods can be an effective way to interrogate single cells. Global RNA-seq provides a whole transcriptome view but at lower sensitivity and accuracy, and oversampling of high abundance genes can obscure detection of rare transcripts. Gene targeting provides efficient focus of sequencing bandwidth on desired transcripts, including those that are rare, and enables absolute quantitation with the highest sensitivity and accuracy when combined with MI. Future improvements are directed at optimizations to further improve RT yields (for example, by elevating reaction temperature to reduce secondary structure) and at increasing multiplexing to hundreds or thousands of genes, with sample indexing for more cells.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Additional information as noted in text. This information is available free of charge via the Internet at http://pubs.acs.org/

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: gfu@cellular-research.com. Tel: (+1) 650 7526144.
*E-mail: sfodor@cellular-research.com. Tel: (+1) 650 7526144.

Notes

The authors declare the following competing financial interest(s): The authors are employed by Cellular Research Inc., a for-profit entity.

■ ACKNOWLEDGMENTS

We thank R. Rudoff for mechanical engineering; G. Wu for algorithm development; H. Greenman and V. Bressler for imaging software; W. Xu for sequencing analysis; M. Pieprzyk and S. Quake for helpful discussions. This work was supported by the National Human Genome Research Institute of the National Institutes of Health Grants R43HG007129 and R43HG007130 (to G.K.F.).

■ REFERENCES

(1) Tang, F.; Barbacioru, C.; Wang, Y.; Nordman, E.; Lee, C.; Xu, N.; Wang, X.; Bodeau, J.; Tuch, B. B.; Siddiqui, A.; Lao, K.; Surani, M. A. *Nat. Methods* **2009**, *6*, 377.

(2) Islam, S.; Kjallquist, U.; Moliner, A.; Zajac, P.; Fan, J. B.; Lonnerberg, P.; Linnarsson, S. *Genome Res.* **2011**, *21*, 1160.

(3) Ramskold, D.; Luo, S.; Wang, Y. C.; Li, R.; Deng, Q.; Faridani, O. R.; Daniels, G. A.; Khrebtukova, I.; Loring, J. F.; Laurent, L. C.; Schroth, G. P.; Sandberg, R. *Nat. Biotechnol.* **2012**, *30*, 777.

(4) Islam, S.; Zeisel, A.; Joost, S.; La Manno, G.; Zajac, P.; Kasper, M.; Lonnerberg, P.; Linnarsson, S. *Nat. Methods* **2014**, *11*, 163.

(5) Picelli, S.; Bjorklund, A. K.; Faridani, O. R.; Sagasser, S.; Winberg, G.; Sandberg, R. *Nat. Methods* **2013**, *10*, 1096.

(6) Shalek, A. K.; Satija, R.; Adiconis, X.; Gertner, R. S.; Gaublomme, J. T.; Raychowdhury, R.; Schwartz, S.; Yosef, N.; Malboeuf, C.; Lu, D.; Trombetta, J. J.; Gennert, D.; Gnirke, A.; Goren, A.; Hacohen, N.; Levin, J. Z.; Park, H.; Regev, A. *Nature* **2013**, *498*, 236.

(7) Wu, A. R.; Neff, N. F.; Kalisky, T.; Dalerba, P.; Treutlein, B.; Rothenberg, M. E.; Mburu, F. M.; Mantalas, G. L.; Sim, S.; Clarke, M. F.; Quake, S. R. *Nat. Methods* **2014**, *11*, 41.

(8) Zhu, Y. Y.; Machleder, E. M.; Chenchik, A.; Li, R.; Siebert, P. D. *BioTechniques* **2001**, *30*, 892.

(9) Sanders, R.; Huggett, J. F.; Bushell, C. A.; Cowen, S.; Scott, D. J.; Foy, C. A. *Anal. Chem.* **2011**, *83*, 6474.

(10) Whale, A. S.; Cowen, S.; Foy, C. A.; Huggett, J. F. *PloS One* **2013**, *8*, No. e58177.

(11) Fu, G. K.; Hu, J.; Wang, P. H.; Fodor, S. P. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 9026.

(12) Warren, L.; Bryder, D.; Weissman, I. L.; Quake, S. R. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 17807.

(13) Charnay, P.; Maniatis, T. *Science* **1983**, *220*, 1281.

(14) Smith, R. D.; Malley, J. D.; Schechter, A. N. *Nucleic Acids Res.* **2000**, *28*, 4998.

(15) White, A. K.; Heyries, K. A.; Doolin, C.; Vaninsberghe, M.; Hansen, C. L. *Anal. Chem.* **2013**, *85*, 7182.

(16) Mortazavi, A.; Williams, B. A.; McCue, K.; Schaeffer, L.; Wold, B. *Nat. Methods* **2008**, *5*, 621.

(17) Fu, G. K.; Xu, W.; Wilhelmy, J.; Mindrinos, M. N.; Davis, R. W.; Xiao, W.; Fodor, S. P. A. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, DOI: 10.1073/pnas.1323732111.

(18) Shiroguchi, K.; Jia, T. Z.; Sims, P. A.; Xie, X. S. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 1347.

(19) Jabara, C. B.; Jones, C. D.; Roach, J.; Anderson, J. A.; Swanstrom, R. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 20166.

(20) White, R. A., 3rd; Quake, S. R.; Curr, K. *J. Virol. Methods* **2012**, *179*, 45.

(21) Tang, D. T.; Plessy, C.; Salimullah, M.; Suzuki, A. M.; Calligaris, R.; Gustincich, S.; Carninci, P. *Nucleic Acids Res.* **2013**, *41*, No. e44.