Smart-seq2 for sensitive full-length transcriptome profiling in single cells

Simone Picelli[1], Åsa K Björklund[1,2], Omid R Faridani[1], Sven Sagasser[1,2], Gösta Winberg[1,2] & Rickard Sandberg[1,2]

Single-cell gene expression analyses hold promise for characterizing cellular heterogeneity, but current methods compromise on either the coverage, the sensitivity or the throughput. Here, we introduce Smart-seq2 with improved reverse transcription, template switching and preamplification to increase both yield and length of cDNA libraries generated from individual cells. Smart-seq2 transcriptome libraries have improved detection, coverage, bias and accuracy compared to Smart-seq libraries and are generated with off-the-shelf reagents at lower cost.

Several methods exist for constructing full-length cDNAs from large amounts of RNA, including cap-enrichment procedures[1–3], but it is still challenging to obtain full-length transcriptome coverage from single cells. Existing methods either use 3′-end poly(A) tailing of cDNA[4,5] or template switching[6,7], or they sacrifice full-length coverage altogether for multiplexing before cDNA amplification[8,9]. We recently showed that Smart-seq, which relies on template switching, provides more even read coverage across transcripts than poly(A)-tailing methods[7], consistent with the common use of template switching in applications designed to capture RNA 5′ ends[8,10]. Despite widespread use of single-cell transcriptome profiling methods, no systematic efforts have been made to improve cDNA library yield and average length from single-cell amounts.

We systematically evaluated a large number of variations in reverse transcription, template-switching oligonucleotides (TSOs) and PCR preamplification (for a total of 457 experiments) and compared the results to those from commercial Smart-Seq (hereafter called SMARTer) in terms of cDNA library yield and length from 1 ng of starting total RNA (**Supplementary Table 1**). In particular, exchanging only a single guanylate for a locked nucleic acid (LNA)[11] guanylate at the TSO 3′ end (rGrG+G) led to a two-fold increase in cDNA yield relative to that obtained with the SMARTer IIA oligo ($P = 7.2 \times 10^{-3}$, $n \geq 8$, Student's t-test; **Fig. 1a**, **Supplementary Table 2** and **Supplementary Fig. 1**). This is likely a consequence of the increased thermal stability of LNA:DNA base pairs (1–8 °C per LNA monomer). Additionally, we found that the presence of the methyl group donor betaine[12] in combination with higher $MgCl_2$ concentrations significantly increased yield (by two- to fourfold; $P \leq 1.3 \times 10^{-3}$, $n \geq 6$, Student's t-test,

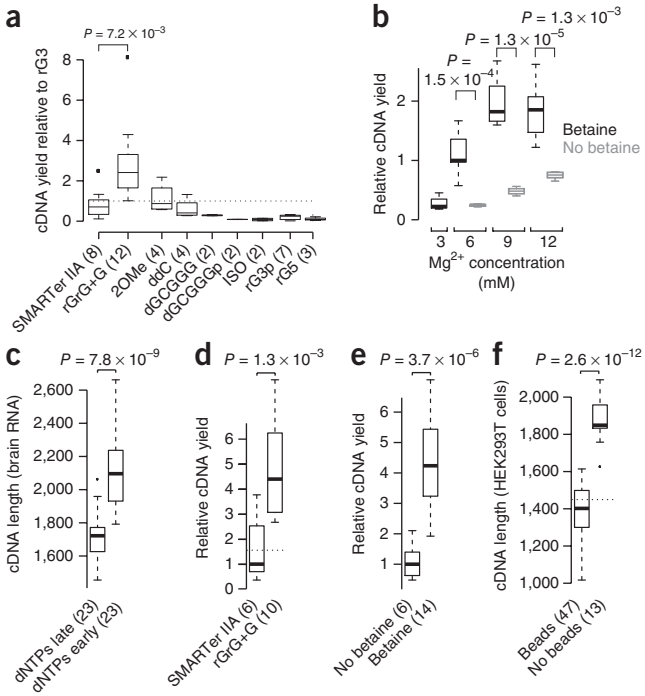

Figure 1 | Improvements in cDNA library yield and length. (**a**) Median yield of preamplified cDNA obtained using different TSOs, relative to those obtained using the rG3 oligo. All oligo sequences are found in **Supplementary Table 1**. (**b**) Median yield of preamplified cDNA in reactions with (black) or without betaine (gray) and as a function of increasing Mg^{2+} concentration, relative to cDNA yields obtained using SMARTer-like conditions. (**c**) Length of preamplified cDNA generated in reactions where dNTPs were added before RNA denaturation (early) or in the reverse transcription master mix (late). Experiments shown in **a**–**c** were based on 1 ng total RNA of mouse brain origin. (**d**) Median yield of preamplified cDNA from HEK293T cells using the LNA-G (rGrG+G) and SMARTer IIA template-switching oligos with the optimized protocol. Dotted horizontal line indicates median yield from commercial SMARTer reactions. (**e**) Median yield of preamplified cDNA from DG-75 cells in reactions with or without betaine. (**f**) Lengths of cDNA libraries generated from single HEK293T cells in reactions with or without bead extraction. Throughout figure, data are represented as box plots with numbers of replicates in parenthesis. Significant differences were determined using Student's t-test.

[1]Ludwig Institute for Cancer Research, Stockholm, Sweden. [2]Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden. Correspondence should be addressed to R.S. (Rickard.Sandberg@ki.se).

RECEIVED 6 MAY; ACCEPTED 17 JULY; PUBLISHED ONLINE 22 SEPTEMBER 2013; DOI:10.1038/NMETH.2639

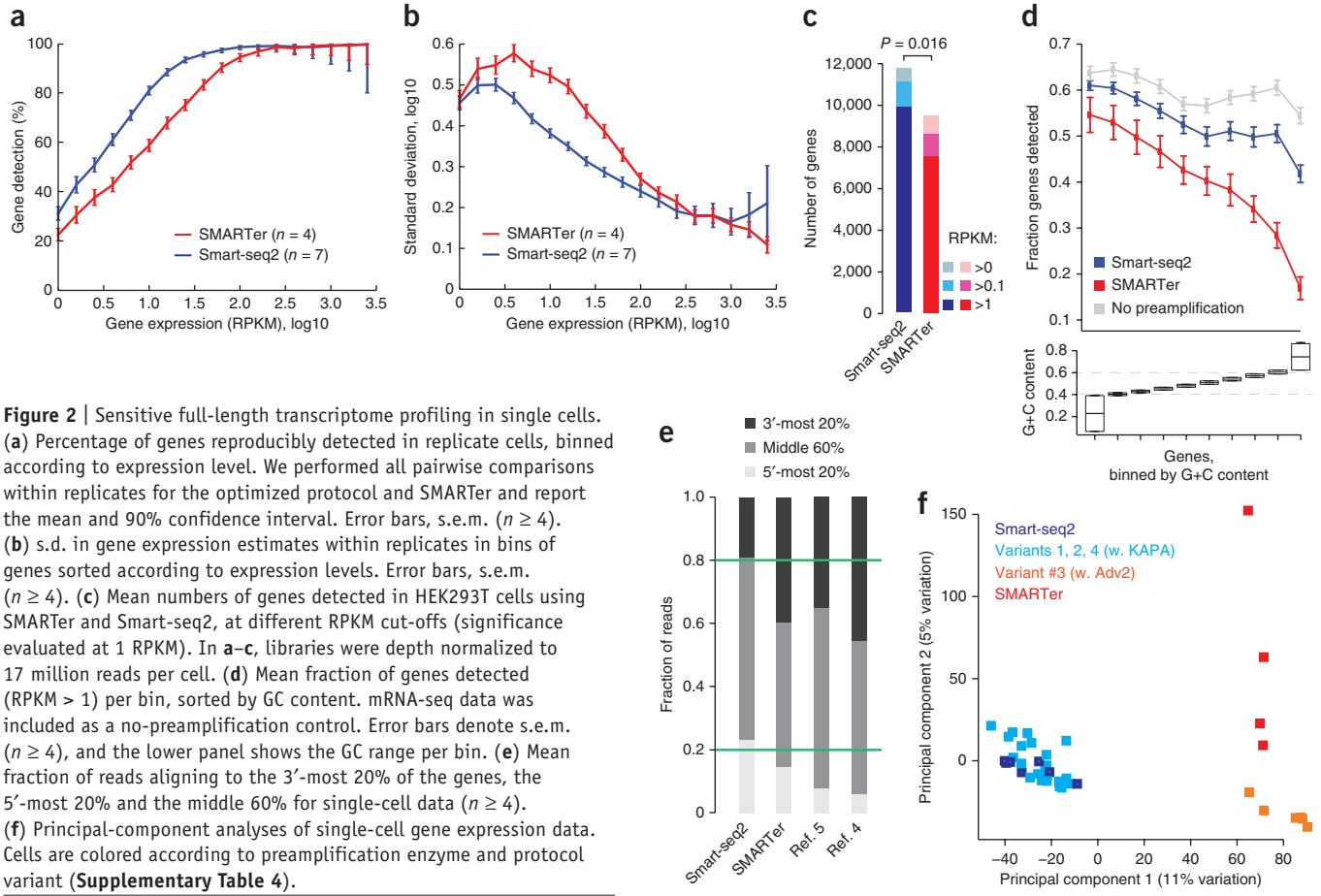

Figure 2 | Sensitive full-length transcriptome profiling in single cells. (**a**) Percentage of genes reproducibly detected in replicate cells, binned according to expression level. We performed all pairwise comparisons within replicates for the optimized protocol and SMARTer and report the mean and 90% confidence interval. Error bars, s.e.m. ($n \geq 4$). (**b**) s.d. in gene expression estimates within replicates in bins of genes sorted according to expression levels. Error bars, s.e.m. ($n \geq 4$). (**c**) Mean numbers of genes detected in HEK293T cells using SMARTer and Smart-seq2, at different RPKM cut-offs (significance evaluated at 1 RPKM). In **a**–**c**, libraries were depth normalized to 17 million reads per cell. (**d**) Mean fraction of genes detected (RPKM > 1) per bin, sorted by GC content. mRNA-seq data was included as a no-preamplification control. Error bars denote s.e.m. ($n \geq 4$), and the lower panel shows the GC range per bin. (**e**) Mean fraction of reads aligning to the 3′-most 20% of the genes, the 5′-most 20% and the middle 60% for single-cell data ($n \geq 4$). (**f**) Principal-component analyses of single-cell gene expression data. Cells are colored according to preamplification enzyme and protocol variant (**Supplementary Table 4**).

for all comparisons) (**Fig. 1b**). The SMARTer buffer has a final $MgCl_2$ concentration of 6 mM, but we found that higher yield is obtained at higher concentration (9–12 mM). Finally, the average length of the preamplified cDNA increased by 370 nt when we administered dNTPs before the RNA denaturation rather than in the reverse transcription master mix ($P = 7.8 \times 10^{-9}$, $n = 23$, Student's *t*-test; **Fig. 1c**), presumably through mechanisms that stabilize the hybridization of RNA to the oligo-dT primer, consistent with earlier observations[13].

We generated single-cell cDNA libraries from 262 individual human or mouse cells (159 HEK293T, 34 DG-75, 30 C2C12 and 39 MEF cells) of different sizes (<50–120 µm) and RNA contents (<10–16 pg) (**Supplementary Table 3**). We demonstrate higher cDNA yields both with the use of the LNA-containing TSO (threefold increase, $P = 1.3 \times 10^{-3}$, $n \geq 6$, Student's *t*-test; **Fig. 1d**) and with betaine together with high Mg^{2+} concentrations (fourfold increase, $P = 3.7 \times 10^{-6}$, $n \geq 6$, Student's *t*-test; **Fig. 1e**).

The sensitivity and accuracy of single-cell methods are limited by the efficiencies of each sample-processing step. The SMARTer protocol uses bead purification to remove unincorporated adaptors[14] from the first-strand cDNA reaction before the preamplification with Advantage 2 Polymerase (Adv2). However, bead purification in small volumes poses a significant recovery challenge for liquid-handling automation. Interestingly, we noted that KAPA HiFi Hot Start (KAPA) DNA Polymerase efficiently amplified first-strand cDNA directly after reverse transcription, with no need for prior bead purification. Importantly, libraries preamplified without bead purification had no reduction in yield, but their average cDNA length was 450 nt greater ($P = 2.6 \times 10^{-12}$, $n \geq 13$, Student's *t*-test; **Fig. 1f**), demonstrating that KAPA preamplification improves cDNA generation and offers a viable approach for Smart-seq automation.

To assess the impact of the improved Smart-seq2 protocol on single-cell transcriptome profiling[7], we sequenced single HEK293T cell libraries generated using both SMARTer ($n = 4$) and variations of Smart-seq2 ($n = 35$) (**Supplementary Table 4**). Reads were aligned with STAR[15] and expression levels quantified as reads per kilobase gene model and million mapped reads (RPKM), as previously described[16]. We observed a substantial increase in our ability to detect gene expression (**Fig. 2a**) and lower technical variation for low- and medium-abundance transcripts (**Fig. 2b** and **Supplementary Fig. 2**). The improved sensitivity of the optimized protocol led to the detection of 2,372 more genes in each cell on average ($P = 0.016$, $n \geq 4$, Student's *t*-test; **Fig. 2c**). All these improvements were independently validated using an alternative RNA-seq alignment and analysis strategy (**Supplementary Fig. 3**). Moreover, we observed both better sensitivity and lower variability in single-cell transcriptome data generated with Smart-seq2 than for data available for Quartz-seq[5] (**Supplementary Fig. 4**). Although the sequenced libraries had mapping characteristics similar to those of SMARTer libraries, we noted a 7% increase in unmapped reads (**Supplementary Fig. 5**).

Several preamplification enzymes have lower GC bias than the Adv2 that is used with SMARTer[17], indicating that single-cell

profiling could also improve with cDNA preamplifications using KAPA. Indeed, preamplification using KAPA instead of Adv2 in Smart-seq2 allowed the detection of more genes at higher GC levels (**Fig. 2d** and **Supplementary Fig. 6**) and provided improved sensitivity and accuracy (**Supplementary Fig. 7**). Moreover, Smart-seq2 reads had more even coverage of both the 5′ and 3′ ends of the transcripts as they approached the expected fractions ($P = 2.7 \times 10^{-5}$, $P = 1.6 \times 10^{-3}$ for 5′ and 3′ ends, respectively, $n \geq 4$, Student's t-test; **Fig. 2e** and **Supplementary Figs. 8** and **9**). Importantly, global gene expression profiles from cells preamplified with KAPA and Adv2 separated on the first principal component (**Fig. 2f**), demonstrating that preamplification bias had a significant impact on the estimation of absolute expression levels. We also noted regions with aberrantly large numbers of aligned reads appearing systematically in Smart-seq irrespectively of preamplification enzyme. This necessitated filtering (**Supplementary Fig. 10**). Together, the data show that preamplification using KAPA improved GC tolerance and read coverage across transcripts, but they also suggest that comparing data generated using different amplifications procedures could be complicated.

To determine the extent of technical variability in the single-cell transcriptome profiling with Smart-seq2, we generated sequencing libraries from dilution series of HEK293T cells (100, 50 and 10 cells) and total RNA (1 ng, 100 pg, 10 pg). Technical losses and variations were small when analyzing ten cells or more, but considerable variability exists at the single-cell level (**Supplementary Fig. 11a–d**), as previously observed[7]. It is informative to contrast the technical variability measured in the dilution experiment with the biological variability present in cells of the same or different cell type origin. To this end, we sequenced additional single-cell transcriptomes from DG-75 ($n = 7$), C2C12 ($n = 6$) and MEF ($n = 7$) cells. For low-abundance transcripts, the observed variability between cells was mainly of a technical nature, whereas in medium- and high-abundance transcripts, variability between cells was mainly biological (**Supplementary Fig. 11e**).

Another attractive feature of Smart-seq2 is the cost-effective generation of single-cell RNA-seq libraries (**Supplementary Table 5**) using off-the-shelf reagents. Currently, Smart-seq2 is limited to poly(A)$^+$ RNAs and does not retain strand or molecule information, although it is compatible with partial-molecule counting[18]. The modifications we suggest will also improve other single-cell methods that rely on template switching, including those carried out on microfluidic chips (such as Fluidigm C1 chips) or inside emulsion droplets.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Raw sequence reads and expression level tables are available at the Gene Expression Omnibus (GSE49321).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank D. Topcic and N. Volakakis for providing cells, D. Ramsköld for contributing code and the members of the Sandberg laboratory for constructive comments on the work. This work was supported by European Research Council Starting Grant 243066 (R.S.), Swedish Foundation for Strategic Research FFL4 (R.S.) and Swedish Research Council grants 2008-4562 (R.S.) and 2010-6844 (Å.K.B.).

AUTHOR CONTRIBUTIONS

S.P. developed the protocol, picked cells, generated cDNA and sequencing libraries, and wrote the manuscript; Å.K.B. performed computational analyses, prepared figures and wrote the manuscript; O.R.F. conceived and designed LNA-based oligos; S.S. picked cells; G.W. contributed to protocol development. R.S. designed the study and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Maruyama, K. & Sugano, S. *Gene* **138**, 171–174 (1994).
2. Carninci, P. & Hayashizaki, Y. *Methods Enzymol.* **303**, 19–44 (1999).
3. Das, M., Harvey, I., Chu, L.L., Sinha, M. & Pelletier, J. *Physiol. Genomics* **6**, 57–80 (2001).
4. Tang, F. *et al. Nat. Methods* **6**, 377–382 (2009).
5. Sasagawa, Y. *et al. Genome Biol.* **14**, R31 (2013).
6. Zhu, Y.Y., Machleder, E.M., Chenchik, A., Li, R. & Siebert, P.D. *Biotechniques* **30**, 892–897 (2001).
7. Ramsköld, D. *et al. Nat. Biotechnol.* **30**, 777–782 (2012).
8. Islam, S. *et al. Genome Res.* **7**, 1160–1167 (2011).
9. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. *Cell Rep.* **2**, 666–673 (2012).
10. Plessy, C. *et al. Nat. Methods* **7**, 528–534 (2010).
11. Petersen, M. & Wengel, J. *Trends Biotechnol.* **21**, 74–81 (2003).
12. Rees, W.A., Yager, T.D., Korte, J. & Von Hippel, P.H. *Biochemistry* **32**, 137–144 (1993).
13. Huang, L., Lee, J., Sitaraman, K. & Gallego, A. *Life Technologies Focus* **22:1**, 6–7 (2000).
14. SMARTer. *Ultra Low RNA Kit for Illumina Sequencing*, Protocol PT5163-1, version PR0X3693 (Clontech Laboratories, 2011).
15. Dobin, A. *et al. Bioinformatics* **29**, 15–21 (2013).
16. Ramsköld, D., Wang, E.T., Burge, C.B. & Sandberg, R. *PLoS Comput. Biol.* **5**, e1000598 (2009).
17. Quail, M.A. *et al. Nat. Methods* **9**, 10–11 (2011).
18. Shalek, A.K. *et al. Nature* **498**, 236–240 (2013).

ONLINE METHODS

Experiments using total RNA. RNA experiments were performed using the Control Total RNA supplied with the SMARTer Ultra Low RNA Kit for Illumina Sequencing (Clontech), extracted from mouse brain. One microliter of a 1 ng/µl solution was used in the reverse transcription for each total RNA experiment.

Single-cell cDNA isolation. Single HEK293T (human), DG-75 (human), C2C12 (mouse) and MEF (mouse) cells were manually picked under the microscope after resuspension in PBS and TrypLE Express (Gibco) in order to keep them floating. To make sure that only single cells were collected, the solution was visually inspected under the microscope and was discarded if multiple cells were observed. Volume of liquid was kept as low as possible, usually below 0.5 µl and preferably below 0.3 µl. Cells were then transferred to a 0.2 ml thin-wall PCR tube containing 2 µl of a mild hypotonic lysis buffer composed of 0.2% Triton X-100 (Sigma) and 2 U/µl of RNase inhibitor (Clontech). Cells already picked were kept on ice throughout the process or stored at −80 °C if not used immediately.

SMARTer, Smart-seq2 and variants of the protocol. We generated RNA-seq libraries from total RNA and individual cells using different protocols. First, SMARTer cDNA libraries were generated from total RNA and single cells using the Smart-seq protocol[7] following manufacturer's instructions (see SMARTer Ultra Low RNA Kit for Illumina Sequencing manual). After PCR preamplification, 5 ng of cDNA were used for the tagmentation reaction and processed exactly in the same way as described below. Libraries were also generated from total RNA and individual cells with an improved protocol (called Smart-seq2) outlined in detail below. Additionally, we generated sequencing libraries from cells using small variations of the Smart-seq2 protocol, with variations in TSO amounts, TSO sequence or PCR enzyme as detailed in **Supplementary Table 4**. Sequencing results for the variants are reported in **Supplementary Figures 2**, **5**–**8** and **10**. Libraries generated with all protocols were handled identically unless otherwise stated.

Reverse transcription. Total RNA or single-cell lysates (see above) were mixed with 1 µl of anchored oligo-dT primer (10 µM, 5′-AAGCAGTGGTATCAACGCAGAGTACT$_{30}$VN-3′, where "N" is any base and "V" is either "A", "C" or "G"; Biomers.net) and 1 µl of dNTP mix (10 mM, Fermentas), denatured at 72 °C for 3 min and immediately placed on ice afterwards. Seven microliters of the first-strand reaction mix, containing 0.50 µl SuperScript II reverse transcriptase (200 U/µl, Invitrogen), 0.25 µl RNAse inhibitor (40 U/µl, Clontech), 2 µl Superscript II First-Strand Buffer (5×, Invitrogen), 0.25 µl DTT (100 mM, Invitrogen), 2 µl betaine (5 M, Sigma), 0.9 µl $MgCl_2$ (100 mM, Sigma), 1 µl TSO (10 µM, the complete list of the oligos can be found in **Supplementary Table 1**) and 0.1 µl nuclease-free water (Gibco), were added to each sample. Reverse transcription reaction was carried out by incubating at 42 °C for 90 min, followed by 10 cycles of (50 °C for 2 min, 42 °C for 2 min). Finally, the reverse transcriptase was inactivated by incubation at 70 °C for 15 min.

PCR preamplification. In the original Smart-seq protocol, purification with Ampure XP beads is performed after first-strand cDNA synthesis. PCR is then carried out directly on the cDNA immobilized on the beads, after adding 2 µl Advantage 2 Polymerase Mix (50×, Clontech), 5 µl Advantage 2 PCR Buffer (10×, Clontech), 2 µl dNTP mix (10 mM, Clontech), 2 µl IS PCR primer (12 µM, Clontech) and 39 µl nuclease-free water to a final reaction volume of 50 µl. In our experiments we did not purify the cDNA after reverse transcription but just added the same PCR master mix, taking into account that the volume after first-strand cDNA synthesis is 10 µl and adjusting the amount of water accordingly. The reaction was incubated at 95 °C for 1 min, then cycled 15 times for total RNA experiments (18 for single cells) between (95 °C 15 s, 65 °C 30 s, 68 °C 6 min), with a final extension at 72 °C for 10 min. A second modification was the replacement of Advantage 2 Polymerase mix with KAPA HiFi HotStart ReadyMix (KAPA Biosystems). Purification after first-strand cDNA synthesis was omitted also in this case. The PCR master mix had the following composition: 25 µl KAPA HiFi HotStart ReadyMix (2×, KAPA Biosystems), 1 µl ISPCR primers (10 µM, 5′-AAGCAGTGGTATCAACGCAGAGT-3′, Biomers.net) and 14 µl nuclease-free water (Gibco). The program used was as follows: 98 °C 3 min, then 15 cycles (18 for cells) of (98 °C 15 s, 67 °C 20 s, 72 °C 6 min), with a final extension at 72 °C for 5 min. Regardless of the PCR protocol used, PCR was purified using a 1:1 ratio of AMPure XP beads (Beckman Coulter), with the final elution performed in 15 µl of EB solution (Qiagen). Library size distribution was checked on a High-Sensitivity DNA chip (Agilent Bioanalyzer) after a 1:5 dilution (for total RNA experiments) or undiluted (in single-cell experiments). The expected average size should be around 1.5–2.0 kb (depending on cell type), and the fraction of fragments below 300 bp should be negligible. To evaluate the performance of the different modifications introduced in the protocol, we relied on the amount of cDNA comprised in the interval 300–9,000 bp in the Agilent Bioanalyzer plot.

Tagmentation reaction and final PCR amplification. Five nanograms of cDNA were then used for the tagmentation reaction carried out with Nextera DNA Sample Preparation kit (Illumina), with the addition of 25 µl of 2× Tagment DNA Buffer and 5 µl of Tagment DNA Enzyme, in a final volume of 50 µl. The tagmentation reaction was incubated at 55 °C for 5 min and then purified with DNA Clean & Concentrator-5 kit (Zymo Research), with a final elution in 20 µl Resuspension Buffer (RSB) from the Nextera kit. The whole volume was then used for limited-cycle enrichment PCR, along with 15 µl of Nextera PCR Primer Mix (NPM), 5 µl of Index 1 primers (N7xx), 5 µl of Index 2 primers (N5xx) and 5 µl of PCR Primer Cocktail (PPC). A second amplification round was performed as follows: 72 °C 3 min, 98 °C 30 s, then 5 cycles of (98 °C 10 s, 63 °C 30 s, 72 °C 3 min). Purification was done with a 1:1 ratio of AMPure XP beads and samples were loaded on a High-Sensitivity DNA chip to check the quality of the library, while quantification was done with Qubit High-Sensitivity DNA kit (Invitrogen). Libraries were diluted to a final concentration of 2 nM and pooled, and 10 pmol were sequenced on Illumina HiSeq 2000.

Serial dilution experiments. To evaluate the technical sensitivity and variability of our protocol, we performed serial dilution experiments both with total RNA extracted from HEK293T cells and with different amounts of HEK293T cells. For the RNA experiment we collected a pellet corresponding to 10^6 cells and

split it into two parts. On the first part we performed extraction of total RNA using the RNeasy Mini kit (Qiagen) and assessed its quality and quantity on an Agilent RNA Nano 6000 chip (Agilent Technologies). The RNA was then diluted down to obtain final concentrations of 1 ng, 100 pg and 10 pg/µl, and 1 µl from each of them was used in three technical replicates, except for the 10-pg experiment, where only two replicates are available.

The second part of the cell pellet was resuspended in PBS to a final concentration of 100,000 cells/ml (100 cells/µl). Part of the suspension was diluted down to 50 and 10 cells/µl. One microliter of each dilution was then added to a 0.2-ml tube containing 2 µl of lysis buffer and RNAse inhibitor and processed as described above for the total RNA. Three technical replicates are available for these experiments as well.

Statistical analyses of cDNA yield and length. Performances of the different protocols were evaluated with regard to cDNA yield and average cDNA length according to the Bioanalyzer in the range of 300–9,000 bp. For mouse brain total RNA samples, each variable was evaluated in a pairwise manner selecting a set of experiments where all other variables are identical. Within that set of experiments, the significance for a change in yield or length, between the two variables, was evaluated using Student's *t*-tests and Wilcoxon rank-sum tests (**Supplementary Table 1**, sheet B). In the HEK293T cell experiments, each optimized experimental setting was compared to each other setting, as well as to the SMARTer protocol, using Student's *t*-test and Wilcoxon rank-sum test (**Supplementary Table 3**, sheet B). To remove the impacts of potential cell aggregates in the single-cell picking procedures, we first analyzed all replicates per conditions for outliers in cDNA yield. We estimated the Huber robust mean and s.d. (Huber's proposal 2, implemented in python package statsmodels) and transformed each observation into s.d. from the mean ([observation-mean]/stdev). Observations were flagged as outliers if they were more than 2.5 s.d. from the mean, and those samples were not included in downstream analyses. All analysis and items in **Figure 1** were produced using R.

Read alignments and gene-expression estimation. Single-cell libraries were sequenced with Nextera dual indexes (i7+i5) on Illumina HiSeq 2000, giving 43bp reads after demultiplexing. The reads were aligned to human (hg19) or mouse (mm10) genomes using STAR v2.2.0 (ref. 15) with default settings and filtered for uniquely mapping reads. Gene expression values were calculated as RPKM values for each transcript in Ensembl release 69 and RefSeq (February 2013) using rpkmforgenes[16]. Comparisons between protocols in **Figure 2a–c** were generated on depth-normalized libraries, using 17 million randomly selected reads per library to compute expression levels (RPKMs). In parallel, we aligned reads from all libraries with TopHat2 (ref. 19) using default settings (but with –segment-length 21, segment-mismatches 1) and using RefSeq gene and transcript annotations (February 2013). Gene expression values were calculated for RefSeq transcripts as FPKMs using Cufflinks 2.1.1 (ref. 20). As cufflinks reported inconsistent FPKM values for short genes, transcripts shorter than 500 nt were removed.

Single-cell RNA-seq sensitivity and variability. Analyses of gene detection in single HEK293T cells (**Fig. 2a** and **Supplementary Figs. 2a** and **7a**) were calculated over all possible pairs of technical replicates from each experimental setting. Genes were binned by expression level in the two samples, and was considered detected if it had an RPKM above 0.1 in both samples. The mean for all possible pairs of technical replicates within a group was used together with 90% confidence intervals computed using the adjusted Wald method. Analyses of variation (**Fig. 2b** and **Supplementary Figs. 2b** and **7b**) were also calculated on pairs of samples, binning genes by the mean of log expression, excluding genes below 0.1 RPKM in either sample. As gene expression levels across single cells are often log normally distributed[21], we calculated absolute difference in log10 expression values by multiplying mean variation in a bin with 0.886.

Analyses of read coverage and GC tolerance. Gene body coverage was calculated using RSeQC-2.3.4 (ref. 22) for the longest transcript of all protein coding genes (**Fig. 2e** and **Supplementary Figs. 8** and **9**), and normalizing the read count at each position by the number of isoforms covering that position. Gene detection at different GC content was calculated using longest transcript for all protein coding RefSeq genes that were binned by GC content into ten equal-sized bins, and the numbers of genes with no detection, or detection at different RPKM cutoffs were calculated (**Fig. 2d** and **Supplementary Fig. 6**).

Re-analyses of published single-cell data. Single-cell transcriptome data generated with Quartz-seq[5] and 3′-end poly(A) tailing[23] were downloaded from SRA (Quartz-seq: SRP017173; Tang *et al.*: GSE20187). Quartz-seq data was processed identically to SMARTer and Smart-seq2 samples using STAR and rpkm for genes (see above), whereas published alignments for SOLiD data[23] were used.

Read peak analyses. Some genes displayed unexplained peaks with high density of reads within the gene body. To identify these regions, we divided the gene bodies of each gene into 101 equally sized bins, and each gene with at least one bin with >5 s.d. read density over the mean read distribution within that gene was analyzed further. In this analysis we discarded genes with low expression (those with fewer than around 2,000–10,000, reads depending on the sequencing depth per cell). The number of such genes in each cell is represented in **Supplementary Figure 10a**. And the genes with peaks in the highest number of HEK239T cells are displayed as heatmaps in **Supplementary Figure 10b**, illustrating that the peaks are consistently found at the same position in all experiments.

19. Kim, D. *et al. Genome Biol.* **14**, R36 (2013).
20. Trapnell, C. *et al. Nat. Biotechnol.* **28**, 511–515 (2010).
21. Bengtsson, M., Ståhlberg, A., Rorsman, P. & Kubista, M. *Genome Res.* **15**, 1388–1392 (2005).
22. Wang, L., Wang, S. & Li, W. *Bioinformatics* **28**, 2184–2185 (2012).
23. Tang, F. *et al. Cell Stem Cell* **6**, 468–478 (2010).

doi:10.1038/nmeth.2639