Power Analysis of Single Cell RNA-Sequencing Experiments

Authors

Valentine Svensson*[1,2], Kedar Nath Natarajan*[1,2], Lam-Ha Ly[2], Ricardo J Miragaia[2,5], Charlotte Labalette[2,3,4], Iain C Macaulay[2], Ana Cvejic[2,3,4] & Sarah A Teichmann[1,2]

Affiliations

[1] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

[2] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

[3] Wellcome Trust – Medical Research Council Cambridge Stem Cell Institute, Cambridge, UK

[4] Department of Haematology, University of Cambridge, Cambridge, UK

[5] Centre of Biological Engineering, University of Minho, Campus de Gualtar, Braga, Portugal

* These authors contributed equally

Correspondence should be addressed to VS (vale@ebi.ac.uk) or SAT (st9@sanger.ac.uk).

Abstract

Single-cell RNA sequencing (scRNA-seq) has become an established and powerful method to investigate transcriptomic cell-to-cell variation, revealing new cell types, and providing insights into developmental processes and transcriptional stochasticity. A key question is how the variety of available protocols compare in terms of their ability to detect and accurately quantify gene expression. Here, we assess protocol sensitivity and accuracy on many published data sets based on spike-in standards and uniform data processing, which includes the development of a flexible Unique Molecular Identifier counting tool (https://github.com/vals/umis). We compare 15 protocols computationally and 4 protocols experimentally on batch-matched cell populations, in addition to investigating the impact of spike-in molecule degradation. Our analysis provides an integrated framework for comparing scRNA-seq protocols.

Introduction

The recent explosion in the development of protocols for sequencing the RNA of individual cells[1,2] has generated different approaches to capture cells, amplify cDNA, minimise biases, and utilise liquid handling platforms. Due to the tiny amount of starting material, considerable amplification is an integral step of all of these protocols. Consequently, it is important to assess the sensitivity and accuracy of the protocols in terms of numbers of RNA molecules detected. Previous studies have compared the performance of a limited number of protocols experimentally[3,4]. In this study, we assessed the performance of a large number of published scRNA-seq protocols based on their ability to quantify the expression of spike-in RNA of known concentration.

We define the sensitivity of a method as the minimum number of input RNA molecules required for a spike-in to be confidently detected (also known as the lower molecular detection limit, for a given sequencing depth), and accuracy as how close estimated relative abundances are to the known abundances of input molecules. High sensitivity permits the detection of very lowly expressed genes, while high accuracy implies that detected variations in expression reflect true biological differences in mRNA abundance across cells, rather than technical factors.

The ERCC (External RNA Controls Consortium)[5] spike-in standards consist of a mixture of 92 RNA species of varying length and GC content, present at 22 abundance levels spaced one fold-change apart from each other (Supplementary Fig. 1). Such spike-ins have been used to assess the reproducibility of standard RNA-seq protocols[6] and to judge the performance of differential expression tests on RNA-seq data[7]. In the context of scRNA-seq, ERCC spike-ins were first used with the CEL-seq protocol[8]. Here, we exploit spike-ins as a unified framework to compare the technical sensitivity and accuracy of different scRNA-seq protocol across various platforms, independent of the biological cell type investigated (Fig. 1).

Our analysis is subject to limitations (explored in depth in the Discussion). We rely on accurate reporting of spike-in volumes and dilutions by the original authors, which we have reconfirmed in a few cases by personal communication. In addition, spike-in molecules may not truly reflect endogenous mRNA capture efficiency in scRNA-seq owing to deviation from natural mRNA sequence features such as shorter polyA tails and the absence of mRNA binding proteins. Nevertheless, our approach allows us to compare the large number of protocols and platforms with published spike-in data, most of which are replicated across at least two different cell types and different laboratories (Supplementary Table 1). This reduces potential bias due to a specific

cell type or study.

Results

Our analysis spans 15 distinct experimental protocols encompassing 28 single-cell studies, including 17 studies that measured expression using full-length transcript coverage and 11 that used UMIs for digital quantification (Supplementary Table 1 and Online Methods). We also carried out 3 different scRNA-seq protocols on the Fluidigm C1 platform using batch-matched mouse embryonic stem cells (mESCs) with both ERCC and Spike-in RNA Variant (SIRV) spike-ins (Online Methods). SMARTer and Smart-Seq2 were performed in duplicate and STRT-seq was performed once. We also generated a high throughput droplet-based 10X Genomics Chromium dataset on ERCC spike-ins and human brain total RNA. In total, our analysis covers 18,123 publicly available samples comprising $30x10^9$ sequencing reads.

Using reported spike-in dilutions and volumes (Supplementary Table 1), we could calculate the absolute number of spike-in RNA molecules at different abundance levels across individual cell samples, thus permitting all data sets to be compared on the same scale.

scRNA-seq quantification accuracy

To assess the quantification accuracy of different protocols, we computed the Pearson product-moment correlation coefficient (□) between log transformed values of estimated ERCC RNA expression and input concentration for each individual cell or sample (Fig. 2A).

Conventional bulk RNA-sequencing is more accurate than scRNA-seq protocols. Remarkably, the accuracy of scRNA-seq protocols is still high, and rarely do individual samples have a Pearson correlation lower than 0.6. The lower accuracy and variable Pearson correlations for individual cells within some protocols (GnT-Seq, CEL-Seq, MARS-Seq) may indicate variable success rates for these protocols.

scRNA-seq sensitivity

To investigate the technical sensitivity achieved for each sample and quantify inter-sample variability for each protocol, we devised a logistic regression model with detection of expression as the dependant variable. Our measure of sensitivity is the spike-in input level at which the probability of detection reaches 50% (Fig. 1b). Measuring the sensitivity of each sample individually avoids biases due to uneven batch sizes. It also avoids using detected spike-ins ratios at each abundance level, which would give poor resolution, because no more than seven

spike-ins share one abundance level.

scRNA-seq protocols are more sensitive than bulk RNA-seq and can detect very low numbers of input molecules (Fig. 2B). The sensitivity of scRNA-seq protocols varied over four orders of magnitude, and several protocols have the potential to detect as little as single digit input spike-in molecules (SMARTer (C1), CEL-Seq2 (C1), STRT-Seq, and inDrop). We observe high within-protocol variability in sensitivity, which can be attributed to sequencing depth; we quantify this in a section below to rank the protocols.

UMI efficiency in tag-counting protocols

The majority of scRNA-seq protocols utilise an UMI-tag counting strategy, in which a single unique random identifier sequence is added to each reverse transcribed mRNA molecule in order to achieve digital transcript quantification. This strategy has largely been applied to protocols that sequence short 5' or 3' RNA sequence tags and thus create cDNA libraries with extremely low complexity, which may lead to strong amplification biases. The UMI on each tag should allow one to remove these biases, as it is added prior to amplification[9]. The question then remains as to how *efficient* the entire scRNA-seq process is.

If $\square$ is the *UMI (counting) efficiency*, the underlying assumption is that the number of UMIs of a gene $\square = \square \cdot \square$, where $0 < \square < 1$ (Supplementary Fig. 2A) and $\square$ is the number of RNA molecules of a gene. We fitted this model for every UMI-tag sample and compared the results across protocols (Fig. 2C). The results recapitulate the logistic regression based measure for sensitivity, as samples with high efficiency have a low molecule detection limit (with the exception of MARS-Seq data, Supplementary Fig. 2B).

However, this measure might not be as appropriate as it appears. If we extend the model to $\square = \square \cdot \square^{\square}$, the best fit should give values of the *molecular exponent* $\square$ close to 1, if the underlying UMI counting assumption is correct. Instead, we find that the best fit is systematically lower than 1, with a mode of around 0.8 (Supplementary Fig. 2C). This implies a saturation of UMI counts as a function of input molecules. This can be explained partially (but not fully) by differences in UMI length between the different protocols (Supplementary Fig. 2D). For example, UMIs with length of 4 base pairs can only count up to 256 unique molecules, and have on average a molecular exponent of 0.6. However, even in protocols with UMIs of 10 base pairs (which can count over a million unique molecules), the molecular exponent is 0.8 per sample on average, and rarely reaches 1.

In conclusion, while UMIs should provide a way of removing amplification biases, the assumed absolute quantification does not seem to hold perfectly.

Endogenous transcripts are more efficiently captured than ERCC spike-ins

It is unclear to what extent sensitivity and accuracy calculations based on exogenous spike-ins apply to endogenous mRNA. On the one hand, ERCC spike-ins have shorter poly(A) tails than typical mRNA from mammalian cells[10], making them harder to capture by poly(T) priming. On the other hand, endogenous mRNA may have intricate secondary structure and can be bound to proteins, potentially reducing the efficiency of reverse transcription.

To investigate the relationship between endogenous and spike-in measurements, we analysed single molecule fluorescent in-situ hybridization (smFISH) data and CEL-seq data from the same mESC line and culture conditions[11] (molecule counts from Dominic Grün, personal communication). Based on data for 9 endogenous genes, CEL-Seq UMI counts correspond to 5-10% of smFISH counts, whereas average UMI counts for ERCC transcripts correspond to only 0.5-1% of input molecule counts (Supplementary Fig. 2E).

Although the number of transcripts is not large, this data suggests that endogenous RNA is much more efficiently captured and amplified than ERCC spike-in molecules, and that our sensitivity measures are likely to be underestimates. The accuracy metric is based on relative abundances, and is not affected by this. This difference in efficiency is important to consider if absolute molecule counts are to be inferred based on ERCC spike-ins.

Sensitivity is more dependent on sequencing depth than accuracy

The results of the per-sample accuracy and sensitivity analysis shows a large amount of within-protocol heterogeneity (Fig. 2A-B). Seeking to explain performance by technical factors, we find a relation with sequencing depth per sample, which researchers can control to fit their budgets and needs. We used a linear model that considers a global effect of sequencing depth, including diminishing returns (Online Methods). The model includes an individual corrected performance parameter for each protocol, which allows protocols to be ranked while accounting for the substantial technical factor of sequencing depth.

We find that accuracy does not strongly depend on sequencing depth (Fig. 3A). The best performing protocols in terms of accuracy are SUPeR-Seq (ρ=0.95), a total-RNA protocol for single cells, and CEL-Seq2 (ρ=0.94), which uses In Vitro Transcription (IVT) rather than PCR to amplify cDNA.

Since the model considers diminishing returns on the sequencing depth, we can identify from the model parameters that accuracy becomes saturated at as few as 250,000 reads, illustrating that it is not strongly dependent on sequencing depth. This also suggests that the expression levels of detected RNAs are generally accurate and quantitatively meaningful in scRNA-seq data.

By contrast, we find that technical sensitivity is critically dependent on sequencing depth, and sensitivity comparisons that do not account for differences in depth would be misleading (Fig. 3B). The sensitivity parameter of the model accounts for sequencing depth to allow for fair comparison, and we used this to rank protocols. The three protocols implemented in a C1 microfluidics system (CEL-Seq2 (C1), STRT-Seq (C1), and SMARTer (C1); number of molecules at a million reads, $\#\square = 2,3$ and 4 respectively) were the top performing protocols in terms of molecule detection. The matched microwell plate implementation of CEL-Seq2 has poorer sensitivity than the C1 implementation ($\#\square = 13$).

Based on the model, we find that sensitivity saturates at about 4.5 million reads per sample. The increase in read depth from 1 million reads to 4.5 million reads per sample results in marginally increased sensitivity; less than a fold change. However the increase from 100,000 reads to 1 million reads per sample results in increased sensitivity of an order of magnitude. Thus, we recommend considering 1 million reads per sample as a good target for saturated gene detection.

It should be noted that not all studies need to saturate detection, especially in cases where the genes of interest are highly expressed. It is equally important to note that sequencing depth is a technical feature, and the number of genes detected depend on the depth. Therefore, sequencing depth must be taken into account when performing and computationally analyzing scRNA-seq, even for compositional expression units such as TPM (Transcripts Per Million).

Degradation of spike-ins does not explain performance variation between experiments

Our performance analysis inherently assumes the gold standard annotation of the spike-ins to be correct. However, due to the labile nature of RNA, it can get degraded during the course of normal reagent handling. To quantify the impact of degradation, we subjected spike-in molecules (both ERCCs and SIRVs) to repeated freeze-thaw cycles (Online Methods). Additionally, as a measure of complete/full degradation, we left the spike-ins either at room

temperature or at 37°C overnight. The freeze-thaw cycles emulate normal handling and upon comparing these to our in-house protocols, we observed an overall small effect on accuracy, which was similar between protocols (Fig. 4A).

Spike-in degradation directly impinges on the effective spike-in dilution in a sample, which is a central factor for calculating the technical sensitivity. We observed that normal handling accounts for molecule limit differences within an order of magnitude, even when spike-ins are subjected to as many as six freeze-thaw cycles. The sensitivity metric for samples subjected to conditions as extreme as overnight degradation (room temperature or 37°C) had two orders of magnitude difference compared to other samples, similar to the difference between protocols (Fig. 4A).

SIRV spike-ins recapitulate accuracy results based on ERCC spike-ins

All the studies mentioned above are based on the ERCC spike-ins, which have bacterial sequence composition. To ensure that our conclusions are generally applicable, we also analysed the SIRV spike-in mix, consisting of 69 artificial transcripts that mimic the splicing patterns of 7 human genes and allow RNA isoform assessment. The SIRV mix E2 contains these isoforms across four abundance levels. As SIRVs only span four abundance levels, they are not compatible with sensitivity analysis, so we focused on accuracy. To compare accuracy using ERCC and SIRV standards, we performed two matched scRNA-seq comparisons (Smart-seq2, SMARTer and STRT-seq on C1 system) using mESCs with both spike-ins (Fig. 4b).

We observe that accuracy is systematically lower when using SIRVs. This is expected, since the ambiguous read assignment to the isoforms introduces a noise element. Overall, we observed a similar pattern of relative accuracy based on SIRVs and ERCCs between our SMARTer and Smart-Seq2 experiments. The STRT-Seq samples had very poor accuracy, as expected since the 5’ transcript tags alone cannot distinguish between different mRNA isoforms.

This experiment provides quantitative evidence that mRNA splice form variation can be inferred at the single cell level using the appropriate protocol. Comparing the protocols, accuracy calculated based on SIRVs recapitulates accuracy based on ERCCs, indicating that spike-in batch variability does not in general explain differences between protocols.

Endogenous mRNA amount does not affect performance metrics based on spike-ins

cDNA is generated from both endogenous mRNA and spike-in RNA during library preparation;

thus, spike-ins are less likely to be sampled if the amount of mRNA is higher. To verify that discrepancy in endogenous mRNA levels (due to e.g. cell type differences) does not affect performance metrics, we investigated published data where information on empty (spike-in RNA alone) and non-empty (mRNA and spike-ins present) samples was provided for the same batch of cells. We compared accuracy and sensitivity between empty and non-empty samples from three studies and found equivalent results, confirming that endogenous mRNA content does not affect performance metrics (Fig. 4C). We quantified the equivalence using 95% Confidence Interval (CI) based equivalence analysis[12] (Online Methods). We found that the empty median CI is 100% contained within the non-empty median CI for accuracy, and 84% contained for sensitivity.

Impact of freeze-thaw cycles on spike-in abundance

To quantify RNA degradation rates in our freeze-thaw experiment, we added single mESCs to individual wells and performed the Smart-seq2 protocol. We compared the spike-in content to the endogenous mRNA content within each well, and related this to the number of freeze-thaw cycles.

We made a predictive Bayesian model of mRNA degradation (see Online Methods) with a degradation rate parameter δ. Sampling from the posterior distribution of δ when applying the model to ERCC spike-ins, we found a degradation rate of 19±0.7% per freeze-thaw cycle (95% CI, Supplementary Fig. 3, see Fig. 4d for posterior predictions). We also applied the mRNA degradation model to SIRVs, and found a similar degradation rate of 18.5±0.1%. However, the SIRV measurements were more noisy, likely due to mapping uncertainty (see Discussion). Overall, our data approximates a 20% degradation rate of spike-ins in each freeze-thaw cycle during normal sample handling.

While we did not observe a large variation in molecular detection limit or accuracy due to normal handling, the relative abundance of spike-ins in a sample is strongly affected by freeze-thaw cycles. This means that the inference of total mRNA in cells based on spike-ins might prove problematic. As we also found that the degradation rate was conserved between ERCC and SIRV spike-ins, the approximately 20% degradation rate per freeze-thaw cycle may hold for RNA in general.

Discussion

A previous study showed[13] that ERCC read alignment varies widely between libraries and

platforms, with some spike-ins having reproducibly poor behavior. This raises the question of whether spike-ins are suitable for the calibration of absolute expression values. The ERCC spike-ins have short poly-A tails ranging from 20 to 26 bases long (the majority are 24 bases) in comparison to eukaryotic mRNAs of 250 base long poly-A tails[10]. This suggests that poly-T priming of ERCC spike-ins might be less efficient than for endogenous mRNA. Furthermore, ERCC spike-ins are not capped at the 5' end, which may lead to reduced template switching efficiency (used in several protocols) as compared to endogenous mRNAs[14]. Lastly, unlike endogenous mRNAs, spike-in RNA are not naturally bound by mRNA-binding proteins or have secondary structures.

Our comparison of smFISH values, a gold standard for absolute mRNA quantification, with spike-in values, suggests that endogenous RNA is detected more efficiently than spike-ins by about one order of magnitude. Therefore, it is important to highlight that the "spike-in molecule detection limit" may underestimate the detection limit for endogenous RNA, and should only be used as a relative sensitivity measure to rank protocols. The global ranking of protocol sensitivity remains relevant, and accuracy is unaffected by these issues, as all ERCC spike-ins within a sample are equally affected.

A perfect comparison would implement each protocol in multiple laboratories using a single stock of reagents and mRNA dilution ladders as standards. Having multiple scientists carry out each protocol would allow the effects of skill to be excluded. A control ladder of mRNA would eliminate issues arising from differences between synthetic spike-ins and mRNA. While the majority of the protocols we have investigated here have been reproduced by at least two distinct laboratories (Supplementary Table 1), we cannot completely rule out the impact of technical proficiency on protocol performance.

We showed that handling and batch variation in ERCC dilutions leads to smaller variations in performance than we see between protocols (Fig. 4A). Nevertheless, in certain published experiments, spike-ins may have been greatly degraded with an impact on our performance metrics. In addition to these caveats, it is important to note that our assessment was performed on currently available data, and does not necessarily reflect the full potential or suitability of a given protocol.

The scRNA-seq protocols that we analyzed provide tremendously powerful and high resolution techniques for unbiased genome-wide dissection of cell populations and their transcriptional regulation. We show that while these protocols vary widely in their detection sensitivity, with

lower limits between 1 and 1,000 molecules per cell, their accuracy in quantification of gene expression is generally high. Sensitivity depends on sequencing depth, but this is less critical for accuracy. However, both sensitivity and accuracy are closely dependent on the scRNA-seq protocol used to generate the data. Protocols with high sensitivity are more suitable for analyzing lowly expressed genes or for additional insights into more subtle gene expression differences affecting individual cell states, but may be less suitable for other scenarios.

Our comparison also suggests that miniaturized scRNA-seq reaction volumes increase sensitivity and provide a good return on investment when sequencing around a million reads per sample. Future improvements of protocols and decreases in the price of sequencing will further boost our ability to answer new questions in biology using single cell transcriptomics.

Accession codes

Primary accessions

ArrayExpress

- **E-MTAB-5480**
- **E-MTAB-5481**
- **E-MTAB-5482**
- **E-MTAB-5483**
- **E-MTAB-5484**
- **E-MTAB-5485**
- **E-MTAB-5486**

Referenced accessions

Gene Expression Omnibus

- **GSE53334**[15]
- **GSE65785**[16]
- **GSE67833**[17]
- **GSE53386**[18]
- **GSE71318**[19]
- **GSE46980**[9]
- **GSE60361**[20]
- **GSE60768**[21]
- **GSE54695**[11]
- **GSE78779**[22]

- **GSE54006**[23]
- **GSE72857**[24]
- **GSE63473**[25]
- **GSE65525**[26]

European Genome-phenome Archive

- **EGAS00001001204**[27]

European Nucleotide Archive

- **ERP010108**[27]
- **ERP005640**[28]
- **ERP006670**[29]
- **ERP010952**[30]
- **ERP013160**[27]

Sequence Read Archiva

- **SRP030617**[3]
- **SRP041736**[31]
- **SRP033209**[32]
- **SRP055153**[33]
- **SRP045422**[34]
- **SRP047290**[35]
- **SRP025171**[36]
- **SRP050499**[37]
- **SRP073767**[38]

ArrayExpress

- **E-MTAB-3346**[39]
- **E-MTAB-3624**[39]

Data Availability

All data used in this study have been deposited at ArrayExpress, and summary tables are provided as supplementary files.

Acknowledgements

We are grateful to O Stegle and J K Kim for helpful discussions and comments on the manuscript. We thank M Lynch for support with the C1 experiments, X Chen for spike-in discussions and M Quail for help with 10x Chromium experiments. We extend our gratitude to S

Linnarsson and A Zeisel for invaluable support in implementing STRT-Seq in our lab and help with sequencing the STRT-library. We also thank D Grun for sharing smFISH molecule counts. Finally we thank R Kirchner for many improvements to the umis tool. The study was supported by Cancer Research UK grant number C45041/A14953 to A Cvejic and C Labalette, European Research Council project 677501 - ZF_Blood to A Cvejic and a core support grant from the Wellcome Trust and MRC to the Wellcome Trust – Medical Research Council Cambridge Stem Cell Institute. The ERC grant ThSWITCH to S A Teichmann (grant no. 260507) and a Lister Institute Research Prize to S A Teichmann. K N Natarajan was supported by the Wellcome Trust Strategic Award “Single cell genomics of mouse gastrulation”.

Author Contributions

VS and SAT conceived the study. VS and LL annotated and processed all data. VS conceived and implemented the umis tool. VS conceived and performed the performance modeling of the data. VS, RJM, and KNN. designed the in-house experiments. KNN optimised and implemented the protocols. The degradation experiments were designed by VS, ICM, RJM, and KNN, who performed the experiments. IM and CL performed zebrafish experiments under the supervision of AC. VS and LL designed the degradation model, and LL implemented the model. VS, KNN, and SAT wrote the manuscript.

Competing financial interests

The authors declare no competing financial interests.

References

1. Macaulay, I. C. & Voet, T. Single cell genomics: advances and future perspectives. *PLoS Genet.* **10,** e1004126 (2014).
2. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* **16,** 133–145 (2015).
3. Wu, A. R. *et al.* Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods* **11,** 41–46 (2014).
4. Ziegenhain, C. *et al.* Comparative analysis of single-cell RNA sequencing methods. *bioRxiv* 035758 (2016). doi:10.1101/035758

5. External RNA Controls Consortium. Proposed methods for testing and selecting the ERCC external RNA controls. *BMC Genomics* **6,** 150 (2005).
6. Jiang, L. *et al.* Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* **21,** 1543–1551 (2011).
7. Munro, S. A. *et al.* Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures. *Nat. Commun.* **5,** 5125 (2014).
8. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* **2,** 666–673 (2012).
9. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11,** 163–166 (2014).
10. Viphakone, N., Voisinet-Hakil, F. & Minvielle-Sebastia, L. Molecular dissection of mRNA poly(A) tail length control in yeast. *Nucleic Acids Res.* **36,** 2418–2433 (2008).
11. Grün, D., Kester, L. & van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nat. Methods* **11,** 637–640 (2014).
12. Walker, E. & Nowacki, A. S. Understanding equivalence and noninferiority testing. *J. Gen. Intern. Med.* **26,** 192–196 (2011).
13. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* **32,** 903–914 (2014).
14. Kapteyn, J., He, R., McDowell, E. T. & Gang, D. R. Incorporation of non-natural nucleotides into template-switching oligonucleotides reduces background and improves cDNA synthesis from very small RNA samples. *BMC Genomics* **11,** 413 (2010).
15. Ferreira, T. *et al.* Silencing of odorant receptor genes by G protein βγ signaling ensures the expression of one odorant receptor per olfactory sensory neuron. *Neuron* **81,** 847–859 (2014).

16. Owens, N. D. L. *et al.* Measuring Absolute RNA Copy Numbers at High Temporal Resolution Reveals Transcriptome Kinetics in Development. *Cell Rep.* **14,** 632–647 (2016).
17. Llorens-Bobadilla, E. *et al.* Single-Cell Transcriptomics Reveals a Population of Dormant Neural Stem Cells that Become Activated upon Brain Injury. *Cell Stem Cell* **17,** 329–340 (2015).
18. Fan, X. *et al.* Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. *Genome Biol.* **16,** 148 (2015).
19. Dang, Y. *et al.* Tracing the expression of circular RNAs in human pre-implantation embryos. *Genome Biol.* **17,** 130 (2016).
20. Zeisel, A. *et al.* Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347,** 1138–1142 (2015).
21. Velten, L. *et al.* Single-cell polyadenylation site mapping reveals 3' isoform choice variability. *Mol. Syst. Biol.* **11,** 812 (2015).
22. Hashimshony, T. *et al.* CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17,** 77 (2016).
23. Jaitin, D. A. *et al.* Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343,** 776–779 (2014).
24. Paul, F. *et al.* Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell* **163,** 1663–1677 (2015).
25. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161,** 1202–1214 (2015).
26. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161,** 1187–1201 (2015).
27. Macaulay, I. C. *et al.* G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12,** 519–522 (2015).
28. Mahata, B. *et al.* Single-cell RNA sequencing reveals T helper cells synthesizing steroids

de novo to contribute to immune homeostasis. *Cell Rep.* **7,** 1130–1142 (2014).

29. Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* (2015). doi:10.1038/nbt.3102
30. Scialdone, A. *et al.* Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* **85,** 54–61 (2015).
31. Pollen, A. A. *et al.* Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* **32,** 1053–1058 (2014).
32. Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509,** 371–375 (2014).
33. Padovan-Merhar, O. *et al.* Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. *Mol. Cell* **58,** 339–352 (2015).
34. Sansom, S. N. *et al.* Population and single-cell genomics reveal the Aire dependency, relief from Polycomb silencing, and distribution of self-antigen expression in thymic epithelia. *Genome Res.* **24,** 1918–1931 (2014).
35. Wilson, N. K. *et al.* Combined Single-Cell Functional and Gene Expression Analysis Resolves Heterogeneity within Stem Cell Populations. *Cell Stem Cell* **16,** 712–724 (2015).
36. Streets, A. M. *et al.* Microfluidic single-cell whole-transcriptome sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 7048–7053 (2014).
37. Guo, F. *et al.* The Transcriptome and DNA Methylome Landscapes of Human Primordial Germ Cells. *Cell* **161,** 1437–1452 (2015).
38. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *bioRxiv* 065912 (2016). doi:10.1101/065912
39. Brennecke, P. *et al.* Single-cell transcriptome analysis reveals coordinated ectopic gene-

expression patterns in medullary thymic epithelial cells. *Nat. Immunol.* **16,** 933–941 (2015).

Fig. legends

Fig. 1

Strategy for scRNA-seq protocol comparison.

(a) Endogenous mRNA levels vary by cell type and condition and cannot be used to compare protocols applied to different cell types. By contrast, protocols can be compared, regardless of cell type, by measuring the same spike-in RNA standards added at known concentrations to all experiments. (**b,c**) We define two global technical performance metrics based on spike-ins: **(b)** Sensitivity, the number of input spike-in molecules at the point where the probability of detection reaches 50%, and **(c)** Accuracy, the Pearson product-moment correlation (□) between estimated expression levels and actual input RNA molecule concentration (ground truth).

Fig. 2

Performance metrics for scRNA-seq protocols.

(A) Accuracy. Distributions of Pearson correlations (□) for all samples stratified by protocol (without accounting for sequencing depth). **(B) Sensitivity.** Distributions of molecule detection limits for all samples stratified by protocol (without accounting for sequencing depth). n, number of samples. The implementation platforms and quantification strategies are indicated below the protocols. **(C) UMI Efficiency.** Distributions of UMI counting efficiencies in UMI-tag counting based samples, stratified by protocol.

Fig. 3

Performance metrics after accounting for sequencing depth.

(**a,b**) Models of accuracy and sensitivity with a global dependency on sequencing depth considering diminishing returns, with a distinct corrected performance parameter for each protocol. Each model has 26 parameters, and is fitted to n=20,717 samples. Bulk data (pink triangles) are only displayed for context. Solid curves show the predicted dependence on sequencing depth. **(A)** Accuracy is only marginally dependent on sequencing depth. Saturation occurs at 270,000 reads per cell in the model (dashed red line). Protocol names are ordered by performance based on predicted correlation □ at 1 million reads. **(B)** Sensitivity is critically dependent on sequencing depth. Saturation occurs at 4.6 million reads per cell (dashed red

line). The gain from 1 to 4 million reads per sample is marginal, while we note that moving from 100,000 reads to 1 millions reads corresponds to an order of magnitude gain in sensitivity (dashed black lines). Protocols are ordered by performance based on predicted detection limit (#, number of molecules at 1 million reads).

Fig. 4

Impact of various factors on performance metrics.

(A) Batch effects and RNA degradation. Performance distributions for three protocols implemented as a single batch, on the Fluidigm C1 (left) and 10x Chromium (far-left; different batch) platforms. Performance distributions of spike-ins measured after freeze-thaw cycles, normal (2-3 cycles) to critical degradation (6 cycles, left overnight at room temperature). **(B) Accuracy estimates across both ERCC and SIRV spike-ins are similar.** Accuracy (Pearson correlation) of both ERCC and SIRV spike-ins inferred across two replicates using multiple protocols. **(C) Endogenous mRNA amount does not affect performance metrics**. Comparison of performance metrics between empty (lacking endogenous mRNA) and non-empty samples from 3 published datasets shows similar performance and no bias due to presence of endogenous mRNA. Red dot, median. Red bar, 95% CI of median, estimated with bootstraps. **(D) Model of relative spike-in abundance degradation during normal handling**. Posterior predictions from Bayesian exponential decay model, for both ERCCs and SIRVs (decay parameter 19% and 18.5%, respectively). Confidence bands correspond to 95% CI from posterior parameter distribution.

Online Methods

Mouse embryonic stem cell culture

Wildtype E14 mouse ES cells (kindly provided by Pentao Liu, Wellcome Trust Sanger Institute) were cultured on gelatin coated dishes using Knockout DMEM (#10829; Gibco), 15% Fetal Calf Serum (FB-1001/500; batch tested from Labtech), 1x Penicillin-Streptomycin-Glutamine (#10378-016; Gibco), 1x MEM NEAA (11140-035; Gibco), 2-mercaptoethanol (31350-010; Gibco) and 1000U Leukemia Inhibitory Factor (LIF; #ESG1107). Mycoplasma-free tested mESC were passaged every 2-3 days.

SMARTer, Smart-Seq2 and STRT-Seq on C1

E14 mESCs were trypsinized to obtain single cell suspension and passed through 30µm filter

(CellTrics; #04-0042-2316). Cells were processed using the C1 Single Cell Auto Prep System (Fluidigm; #100-7000 and #100-6209) following the manufacturers protocol (#100-5950 B1). Briefly, we perform SMARTer, Smart-seq2 and STRT-Seq each across three small C1 Open App IFCs (5-10µm; #100- 5759). The specific sample preparation steps for the three protocols (SMARTer[3,28,29,31,32], Smart-seq2[40] and STRT-Seq[9,11,20,23]) were downloaded from Fluidigm Script Hub. Dissociated single cells were loaded and captured on C1 Open App IFCs, followed by manual inspection to demarcate empty well, doublets or debris containing wells. Two different spike-in RNA control sets were used for batch-matched comparison of different protocols, 92 ERCC spike-ins (#4456740; Lot# 1411014; Ambion) and 69 SIRV spike-ins (#SKU025.03; E2 Spike-in RNA Variant Control Mixes; Lexogen), were mixed (0.5µl 1:500 diluted ERCCs + 0.6µl 1:500 diluted SIRVs) and added to respective lysis buffer master mixes for SMARTer (20µl), Smart-Seq2 (27µl) and STRT-seq (20µl). 9µl of the respective lysis master mix is added to each Open App C1 IFCs. The subsequent steps (cell lysis, cDNA synthesis by reverse transcription and PCR reaction) are performed as described on Fluidigm Script Hub.

SMARTer and Smart-Seq2 on C1

E14 mESCs were trypsinized to obtain single cell suspension and passed through 30µm filter (CellTrics; #04-0042-2316). Single cell suspension was processed using SMARTer and Smart-seq2 in parallel across two C1 Single Cell Auto Prep System (Fluidigm; #100-7000 and #100-6209) following the manufacturer's protocol (#100-5950 B1). Smart-seq2 protocol was downloaded from Fluidigm Script Hub. The cells were loaded, captured on C1 Open App IFCs, followed by manually inspection. Both ERCC and SIRV spike-ins were mixed (0.5µl 1:500 diluted ERCCs + 0.6µl 1:500 diluted SIRVs) and added to respective Lysis buffer master mixes for SMARTer (20µl) and Smart-Seq2 (27µl). The subsequent steps (cell lysis, cDNA synthesis by reverse transcription and PCR reaction) are performed as described on Fluidigm Script Hub.

Spike-in degradation experiment using Smart-Seq2 on plates

We used new tube of Spike-ins, ERCC (#4456740; Lot# 1412014; Ambion) and SIRV (E2 mix; #SKU025.03; Lot#216651530; Lexogen) for this experiment. Briefly, 1:100 dilutions of ERCCs and SIRVs were mixed together resulting in spike-in master mix (1:200 final dilution; termed ‘x2 Freeze-thaw’). The spike-in master mix was split between three tubes; one left overnight at 37°C (Condition 1), one left overnight at room temperature (Condition 2) and third kept overnight at -80°C. The following day the third tube (from -80°C) was subjected to multiple freeze-thaw cycle wherein the tube was thawed at room temperature for 2-5minutes, an aliquot was taken

and re-freezed in dry ice. We repeated this freeze-thaw cycle an additional 5 times (Condition 3 to Condition 7). All the spike-in mixes (Condition 1-7) were subsequently diluted to a final 1:1000,000 dilution. A 96-well plate for Smart-seq2 was prepared by dispensing 2 µl Smart-Seq2 lysis buffer (0.2%Triton, 1:20 RNAse inhibitor, 10mM Oligo-dT_{30}VN, 10mM dNTPs) across each well. 1µl of spike-in mix per condition (Condition 1-7) was added to each well column-wise such that each column represented a single condition with 8 replicate wells. E14 mESCs were filtered through a 30µm filter and FACS sorted (BD Influx; BD Biosciences) into 96-well plate. The first three wells (row-wise) across the 96-well plate received matched bulk 500, 50 and 5 cells, and all other wells received a single cell. The 96-well plate was immediately spun and frozen on dry-ice prior to Smart-seq2 protocol as previously described[40].

Library preparation and Sequencing

Representative cDNA from single cells across three C1 runs and Smart-Seq2 (on plates) were assessed using High Sensitivity DNA chips for Bioanalyzer (5067-4626 and 5067-4627; Agilent Technologies). Single cell cDNA from SMARTer[3,28,29,31,32] and Smart-Seq2 C1 IFCs and Smart-seq2 (on plates) was tagmented and pooled to make libraries using Illumina Nextera XT DNA sample preparation kit (Illumina; FC-131-1096) with 96 dual barcoded indices (Illumina; FC-131-1002). The library clean-up and sample pooling was performed using AMPure XP beads (Agencourt Biosciences; A63880). All protocols are described in the Fluidigm protocol (100-5950), Fluidigm Script Hub and Smart-seq2 protocol[40]. The STRT-Seq libraries were made and sequenced at Karolinska Institutet as previously described[9,20]. The Single cell libraries from SMARTer and Smart-Seq2 C1 IFCs and Smart-seq2 (on plates) was sequenced across 1 lane of HiSeq V4 (Illumina) using 75bp/125bp paired-end sequencing.

10x Genomics Chromium experiment

The Single Cell Gel Bead kit (#120217), Single cell chip kit (#120219) and Single cell library kit (#120218) were used along with 10x GemCode Single Cell Instrument as per manufacturer specifications and manuals (Document # CG00011; Revision B). Equal volumes of control brain RNA (3µl; FirstChoice Human Brain Total RNA; #AM7962) and ERCC spikes (3µl 1:4 dilution; #4456653) were mixed to form a ‘2x Control RNA+ERCC’ master mix. We further diluted this to ‘1x Control RNA+ERCC’ with PCR grade water. We made two single cell master mix preparation using 3µl of ‘2x Control RNA+ERCC’ and ‘1x Control RNA+ERCC’ respectively instead of single cell suspension (adjusted with 34.4 µl Nuclease-Free water). The remaining protocol was followed as per manufacturer's manual (Document # CG00011; Revision B). Each

10x library was sequenced across HiSeq2500 (2x lanes; Rapid Run) as per Wellcome Trust Sanger Institute sequencing guidelines.

Data Sources

Raw read data from published studies was downloaded from either ENA or SRA, as listed with accession numbers in Supplementary Table 1. Information regarding concentration and volume of ERCC mix in each sample was gathered from the original publications (also indicated in Supplementary Table 1) or through direct communication with authors in ambiguous cases.

The expression table for mESC-STRT had non-standard names annotating the ERCC spike-ins, and through personal communication with the authors we were given a table for converting these to the names as provided by Life Technologies. Additionally we were informed by the authors that the final spike-in dilution noted as 1:50000 in Islam et al[9] had actually been 1:20000.

The concentrations of the ERCC solution in the Dendritic-MARS table was ambiguous as there were two different values in the GEO table and in the text of the paper. Communication with the authors clarified that these referred to different volumes. The volume and dilution described in the GEO table was used. Thirty samples were excluded as they were annotated as not having had ERCC spike-ins added to them.

For the K562-SMART data it was unclear which data sets had used spike-ins, and personal communication with the authors provided the names of the two batches which had spike-ins added.

A table with notes on individual data sets is provided (Supplementary Table 1).

RNA-Seq data processing

For coverage based data, relative abundances were quantified using Salmon[41] 0.6.0, with library type parameter -l IU and the optional flag --biasCorrect. The Salmon transcriptome indices were built by adding ERCC sequences to cDNA sequences from Ensembl. For samples with mouse background, this was Ensembl 83 cDNA annotation of GRCm38.p4. For samples with human background, this was cDNA annotation from Ensembl 78 of GRCh38, and for samples with zebrafish background, the Ensembl 77 annotation of Zv9. Finally, for samples with frog background, this was Ensembl 84 annotation of JGI4.2.

All coverage-based datasets were sequenced using Illumina paired-end sequencing, with read

lengths between 75 and 150 base pairs.

In order to process all UMI-based data in a coherent way, we developed a quantification strategy based on pseudo-mapping, and counting up evidence for (transcript, UMI) pairs.

The principle is to transfer information from a (UMI, tag) pair to a (transcript, UMI) pair based on which transcript the tag maps to. Since UMI-based methods only use 3' or 5' end tags of cDNA, which can be as short as 25bp, mapping of these tags are commonly ambiguous. Our strategy for this is to weight a (UMI, tag) pair by the number of transcripts the tag maps to. After (UMI, tag) pairs were mapped with either RapMap[42] or Kallisto[43] in pseudobam mode, only (transcript, UMI) pairs with a user specified minimum amount of evidence are counted (default 1). This can be either on the gene or transcript level. In the 10x Genomics Chromium data we detected 70,000 and 45,000 droplets with respect to the samples. For the sake of computational memory efficiency we uniformly sampled 2000 droplets out of all detected droplets to count the umi tags per droplet.

Code Availability

We implemented the UMI counting strategy in a publicly available command line tool which we call 'umis'. The tool as available at https://github.com/vals/umis/ as well as in the Python Package Index, and in Bioconda. Version 0.3.0, used for this paper, is submitted as Supplementary Software.

Analysis

An ERCC spike-in was considered detected when the estimated TPM of that ERCC was greater than zero. For UMI-based data, a spike-in is detected when at least one copy of an ERCC molecule is inferred.

The amount of input spike-in molecules for each spike, for each sample, in each experiment was calculated from the final concentration of ERCC spike-in mix in the sample.

Calculation of the accuracy of an individual sample was done by the Pearson correlation between input concentration of the spike-ins and the measured expression values. If less than 8 spike-ins were observed, the accuracy was set to infinity, as we consider this to be insufficient evidence to estimate the accuracy.

For the logistic regression model of each sample's detection limit, the probability of detecting a spike-in at a given input level is modeled by the logistic function:

$$\square(\square\square\square\square\square\square\square\square\square_{\square}) = \frac{1}{1+\square^{-(\square\cdot\square\square\square(\square_{\square})+\square)}} + \square.$$

We used the LogisticRegression class from the linear_model module of the machine learning package scikit-learn[44]. The fit was performed with the liblinear solver and the optional argument fit_intercept=True. The logistic regression analysis was limited to samples with at least eight spike-ins detected. The detection limit was chosen as the molecular abundance where the logistic regression model passes 50% detection probability:

$$\square\square\square\square\square\square\square\square\square\ \square\square\square\square\square = -\frac{\square}{\square}.$$

To investigate the UMI efficiency of UMI based protocols, we used a linear model where the only parameter was the efficiency:

$$\square\square\square_{\square} = \square\cdot\square_{\square} + \square.$$

As we mention in the text though, the data fits a model much better where there is a non-one exponent parameter on the number of input molecules:

$$\square\square\square_{\square} = \square\cdot\square_{\square}^{\ \square} + \square.$$

When we model the relation between read depth and performance metrics for individual protocols, we use a linear model with a quadratic term for read depth to capture diminishing returns on investment. The model considers the read depth effect to be global, and has a categorical performance parameter for each protocol:

$$\square\square\square\square\square\square\square_{\square} = \square^{2}\cdot\square\square\square_{10}(\square\square\square\square\square\square_{\square}) + \square\cdot\square\square\square_{10}(\square\square\square\square\square\square_{\square}) + \square\square\square\square\square\square\square\square\square\square\square\square_{\square\square\square\square\square\square\square\square\square} + \square.$$

Here the performance metric will plateau and saturate when

$$\square\square\square_{10}(\square\square\square\square\square\square) = -\frac{\square}{2\square}.$$

The linear models were fitted and analysed using the OLS regression function in the statsmodels Python package.

In the spike-in degradation model the degradation rate p and the cellular fraction F were inferred by a Bayesian approach using Stan[45] (R package rstan v 2.10.1). The model was specified as the following: p was sampled from a uniform distribution between 0 and 1, F_i for each spike-in i was drawn from a normal distribution with mean 0.5 and standard deviation 1. F_{ij} was estimated

by a normal distribution with mean F_i *$(1-p)^j$, where j was the j-th freeze-thaw cycle and standard deviation sigma sampled from a uniform distribution between 0 and 20. The model was run with 5000 iteration steps, 1000 warm up steps and 4 chains.

Confidence intervals with regard to accuracy and sensitivity for non-empty and empty wells were estimated by bootstrapping. Therefore, study SRP055153, ERP010952 and SRP070989 were pooled together separating non-empty and empty wells, respectively. For each group, sample sizes of 20 were randomly picked with replacement and median of the bootstrapped samples was determined. This process was repeated with a number of 1,000 iterations. Having sorted the bootstrapped estimates, we determined the median and the 2.5th and 97.5th percentiles of the distributions for non-empty and empty wells. All data needed for our analysis is provided as Supplementary Table 2.

References

40. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10,** 1096–1098 (2013).
41. Patro, R., Duggal, G. & Kingsford, C. Salmon: Accurate, Versatile and Ultrafast Quantification from RNA-seq Data using Lightweight-Alignment. *bioRxiv* 021592 (2015). doi:10.1101/021592
42. Srivastava, A., Sarkar, H., Gupta, N. & Patro, R. RapMap: a rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes. *Bioinformatics* **32,** i192–i200 (2016).
43. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34,** 525–527 (2016).
44. Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830 (2011).
45. Carpenter, B., Gelman, A., Hoffman, M., Lee, D. & Goodrich, B. Stan: A probabilistic programming language. *J. Stat. Softw.* (2016).