Power analysis of single-cell RNA-sequencing experiments

Valentine Svensson[1,2,6], Kedar Nath Natarajan[1,2,6], Lam-Ha Ly[2], Ricardo J Miragaia[2,3], Charlotte Labalette[2,4,5], Iain C Macaulay[2], Ana Cvejic[2,4,5] & Sarah A Teichmann[1,2]

Single-cell RNA sequencing (scRNA-seq) has become an established and powerful method to investigate transcriptomic cell-to-cell variation, thereby 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 assessed the protocol sensitivity and accuracy of many published data sets, on the basis of spike-in standards and uniform data processing. For our workflow, we developed a flexible tool for counting the number of unique molecular identifiers (https://github.com/vals/umis/). We compared 15 protocols computationally and 4 protocols experimentally for batch-matched cell populations, in addition to investigating the effects of spike-in molecular degradation. Our analysis provides an integrated framework for comparing scRNA-seq protocols.

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, minimize biases, and use liquid-handling platforms. Owing to the tiny amount of starting material, considerable amplification is an integral step in all of these protocols. Consequently, it is important to assess the sensitivity and accuracy of the protocols in terms of the number of RNA molecules detected. Previous studies have experimentally compared the performance of a limited number of protocols[3,4]. In this study, we assessed the performance of a large number of published scRNA-seq protocols on the basis of their ability to quantify the expression of spike-in RNAs of known concentration.

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

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

Our analysis was subject to limitations (described in depth in the Discussion). We relied on accurate reporting of spike-in volumes and dilutions by the original authors, which we reconfirmed by personal communication in several cases. 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 poly(A) tails and the absence of mRNA-binding proteins. Nevertheless, our approach allows for comparison across the large number of protocols and platforms with published spike-in data, most of which have been replicated across at least two different cell types and different laboratories (**Supplementary Table 1**). This methodology decreases potential bias due to a specific cell type or study.

RESULTS

Our analysis spanned 15 distinct experimental protocols encompassing 28 single-cell studies, including 17 studies that measured expression with full-length transcript coverage and 11 that used unique molecular identifiers (UMIs) for digital quantification (**Supplementary Table 1** and Online Methods). We also carried out three different scRNA-seq protocols on the Fluidigm C1 platform by using batch-matched mouse embryonic stem cells

[1]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge, UK. [2]Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK. [3]Centre of Biological Engineering, University of Minho, Braga, Portugal. [4]Wellcome Trust–Medical Research Council Cambridge Stem Cell Institute, Cambridge, UK. [5]Department of Haematology, University of Cambridge, Cambridge, UK. [6]These authors contributed equally to this work. Correspondence should be addressed to V.S. (vale@ebi.ac.uk) or S.A.T. (st9@sanger.ac.uk).

RECEIVED 9 SEPTEMBER 2016; ACCEPTED 7 FEBRUARY 2017; PUBLISHED ONLINE 6 MARCH 2017; DOI:10.1038/NMETH.4220

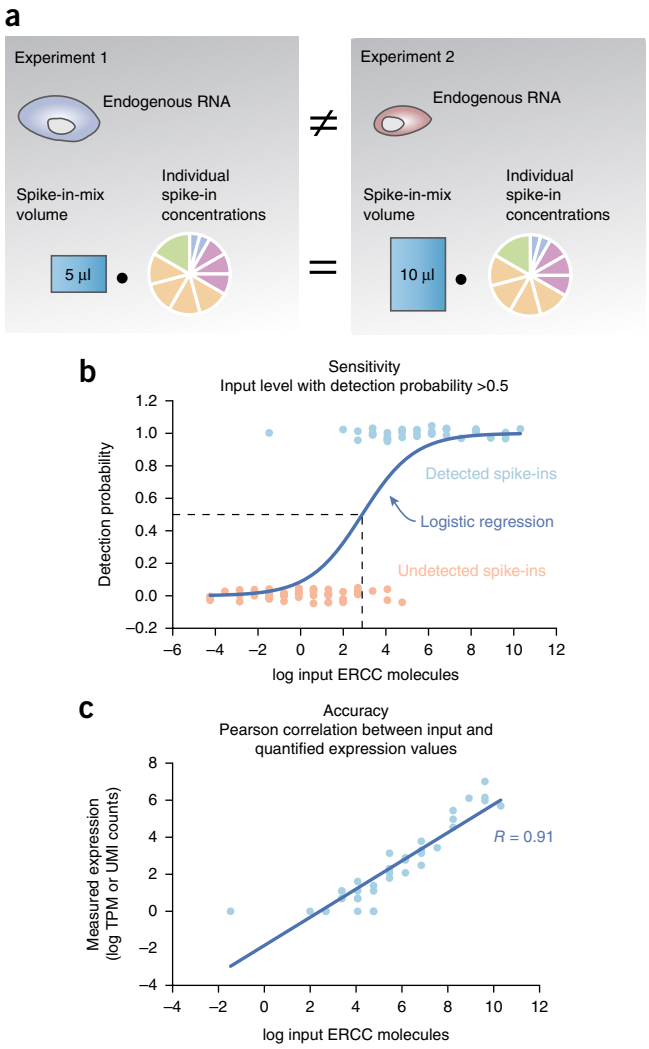

Figure 1 | Strategy for scRNA-seq protocol comparison. (**a**) Endogenous mRNA levels vary by cell type and condition and cannot be used to compare protocols applied to different cell types. By contrast, protocols can be compared, regardless of cell type, by measuring the same spike-in RNA standards added at known concentrations to all experiments. (**b**,**c**) We define two global technical performance metrics for spike-ins: sensitivity, the number of input spike-in molecules at the point at which the probability of detection reaches 50% (**b**), and accuracy, the Pearson product-moment correlation (*R*) between estimated expression levels and actual input RNA-molecule concentration (ground truth) (**c**). TPM, transcripts per million.

(mESCs) with both ERCC and Spike-in RNA Variant (SIRV) spike-ins (Online Methods). SMARTer and Smart-seq2 were performed in duplicate, and single-cell tagged reverse transcription (STRT)-seq was performed once. We also generated a high-throughput droplet-based 10× Genomics Chromium data set on ERCC spike-ins and human brain total RNA. In total, our analysis covered 18,123 publicly available samples comprising 30×10^9 sequencing reads.

Using reported spike-in dilutions and volumes (**Supplementary Table 1**), we calculated 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 (*R*) 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 individual samples rarely have a Pearson correlation less than 0.6. The lower accuracy and variable Pearson correlations for individual cells within some protocols (genome and transcriptome sequencing (G&T-seq), CEL-seq, and massively parallel single-cell RNA-seq (MARS-seq)) may indicate variable success rates for these protocols.

scRNA-seq sensitivity

To investigate the technical sensitivity achieved for each sample and to quantify the intersample variability for each protocol, we devised a logistic regression model with detection of expression as the dependent variable. Our measure of sensitivity was the spike-in input level at which the probability of detection reached 50% (**Fig. 1b**). Measuring the sensitivity of each sample individually avoided biases due to uneven batch sizes. This approach also avoided the need to use detected spike-in ratios at each abundance level, which would have resulted in poor resolution, because no more than seven spike-ins share one abundance level.

scRNA-seq protocols are more sensitive than bulk-RNA sequencing 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 (SMARTer (C1), CEL-seq2 (C1), STRT-seq, and inDrop) have the potential to detect as little as single-digit input spike-in molecules. We observed high within-protocol variability in sensitivity, which may have been attributable to sequencing depth; as described below, we quantified this variability to rank the protocols.

UMI efficiency in tag-counting protocols

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

If *E* is the UMI (counting) efficiency, the underlying assumption is that the number of UMIs of a gene (*U*) is equal to $E \times M$, where $0 < E < 1$ (**Supplementary Fig. 2a**), and *M* 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 recapitulated the logistic-regression-based measure for sensitivity, because samples with high efficiency had a low molecular-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 $U = E \times M^c$, the best fit should yield

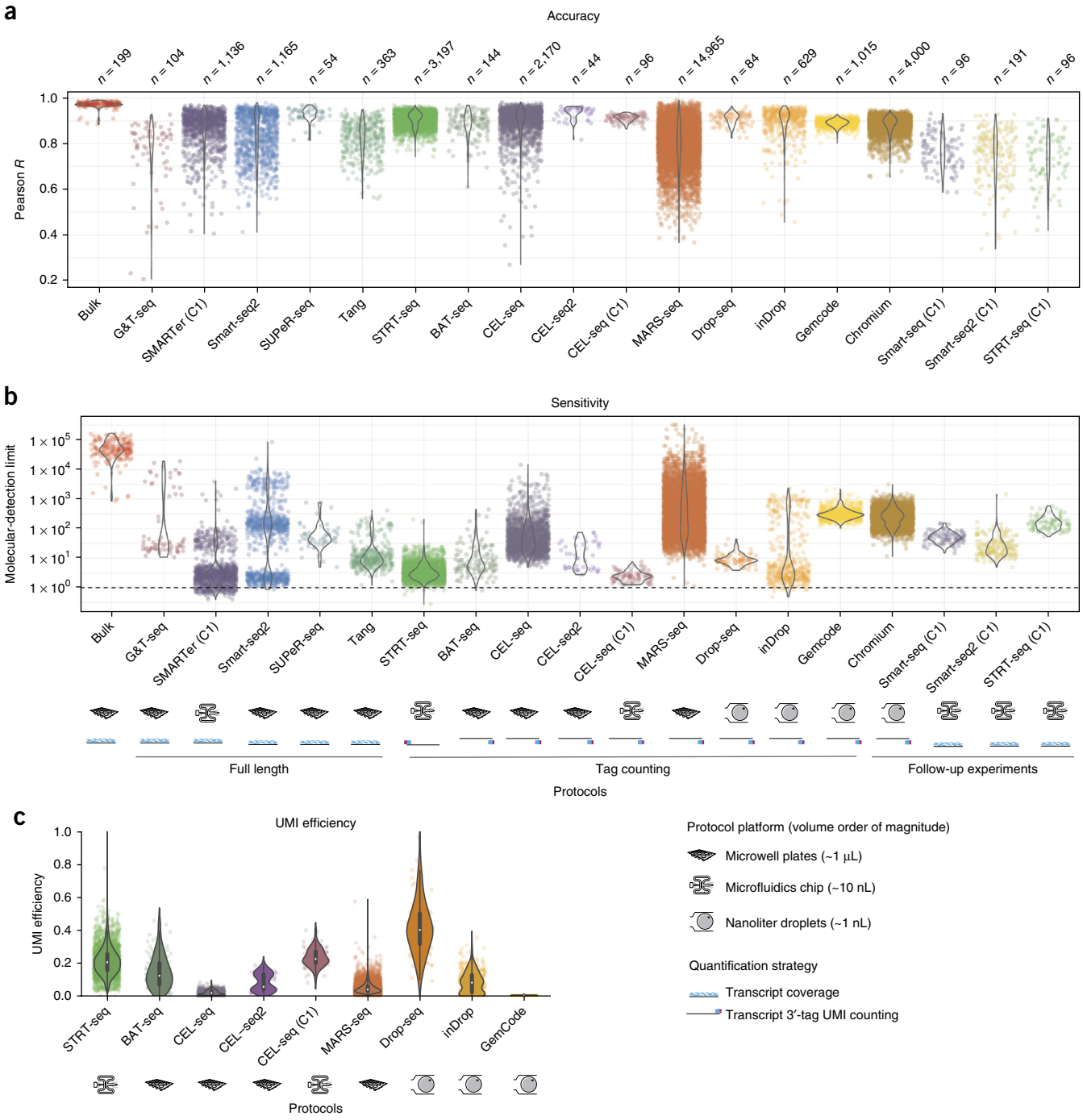

Figure 2 | Performance metrics for scRNA-seq protocols. (**a**) Accuracy. Distributions of Pearson correlations (*R*) for all samples, stratified by protocol (without accounting for sequencing depth). BAT-seq, barcoded 3′-specific sequencing. (**b**) Sensitivity. Distributions of molecular-detection 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 samples, based on UMI-tag counting, stratified by protocol. Boxes, quartiles; whiskers, full range of values; white dots, median.

values of the molecular exponent *c* close to 1, if the underlying UMI counting assumption is correct. Instead, we found that the best fit was systematically lower than 1, with a mode of approximately 0.8 (**Supplementary Fig. 2c**). This finding suggested a saturation of UMI counts as a function of input molecules and may be partially (but not fully) explained by differences in UMI length among the different protocols (**Supplementary Fig. 2d**). For example, UMIs with a length of 4 bp are able to count up to only 256 unique molecules and had a molecular exponent of 0.6, on average. However, even in protocols with UMIs of 10 bp (which are able to count over 1 million unique molecules), the molecular exponent was 0.8 per sample, on average, and rarely reached 1. In conclusion, whereas UMIs should provide a way of removing amplification biases, the assumed absolute quantification does not appear to hold true perfectly.

Endogenous transcripts are more efficiently captured than ERCC spike-ins

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

To investigate the relationship between endogenous and spike-in measurements, we analyzed single-molecule fluorescence *in situ* hybridization (smFISH) data and CEL-seq data from the same mESC line and culture conditions[11] (molecule counts from D. Grün, Max Planck Institute of Immunobiology and Epigenetics, personal communication). On the basis of data for nine endogenous genes, CEL-seq UMI counts corresponded to 5–10% of smFISH counts, whereas the average UMI counts for ERCC transcripts corresponded to only 0.5–1% of input-molecule counts (**Supplementary Fig. 2e**).

Although the number of transcripts was not large, these data suggested 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 was based on relative abundance and was not affected by underestimated capture. This difference in efficiency is important to consider if absolute molecule counts are to be inferred on the basis of ERCC spike-ins.

Sensitivity is more dependent than accuracy on sequencing depth

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

We found that accuracy does not strongly depend on sequencing depth (**Fig. 3a**). The best-performing protocols in terms of accuracy were single-cell universal poly(A)-independent RNA-seq (SUPeR-seq) ($R = 0.95$), a total-RNA protocol for single cells, and CEL-seq2 ($R = 0.94$), which uses *in vitro* transcription rather than PCR to amplify cDNA.

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

By contrast, we found 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 parameter 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 one million reads ($\#M$) of 2, 3, and 4, respectively) were the top-performing protocols in terms of molecular detection. The matched microwell-plate implementation of CEL-seq2 had poorer sensitivity than the C1 implementation ($\#M = 13$).

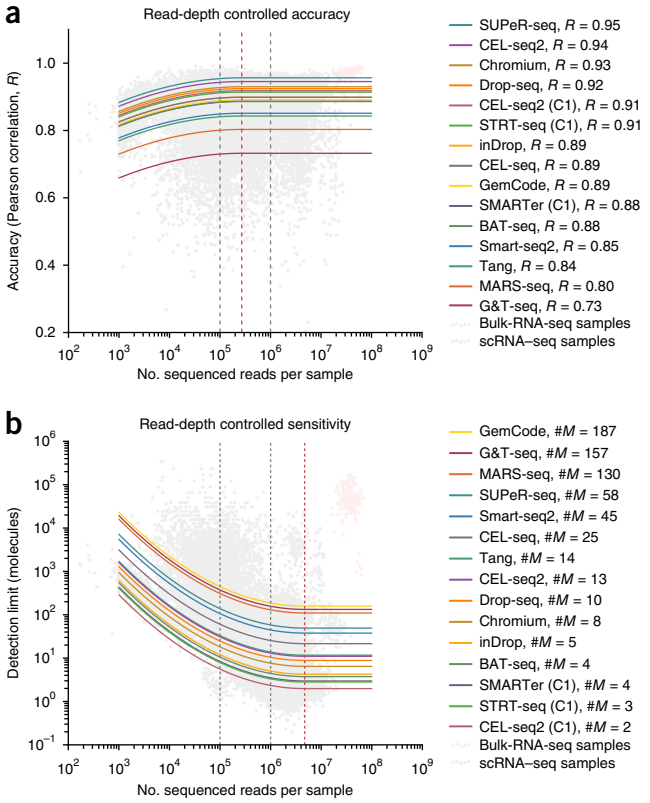

Figure 3 | Performance metrics after accounting for sequencing depth. (**a**,**b**) Models of accuracy and sensitivity with a global dependency on sequencing depth, considering diminishing returns, with a distinct corrected performance parameter for each protocol. Each model has 26 parameters and is fitted to $n = 20{,}717$ samples. Bulk data (pink triangles) are displayed only 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 on the basis of predicted correlation (R) 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, whereas moving from 100,000 reads to 1 million reads corresponds to an order-of-magnitude gain in sensitivity (dashed black lines). Protocols are ordered by performance on the basis of predicted detection limit ($\#M$, number of molecules at 1 million reads).

On the basis of the model, we found that the sensitivity saturates at approximately 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, of less than a onefold 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.

Notably, not all studies need to saturate detection, especially in cases in which the genes of interest are highly expressed. It is

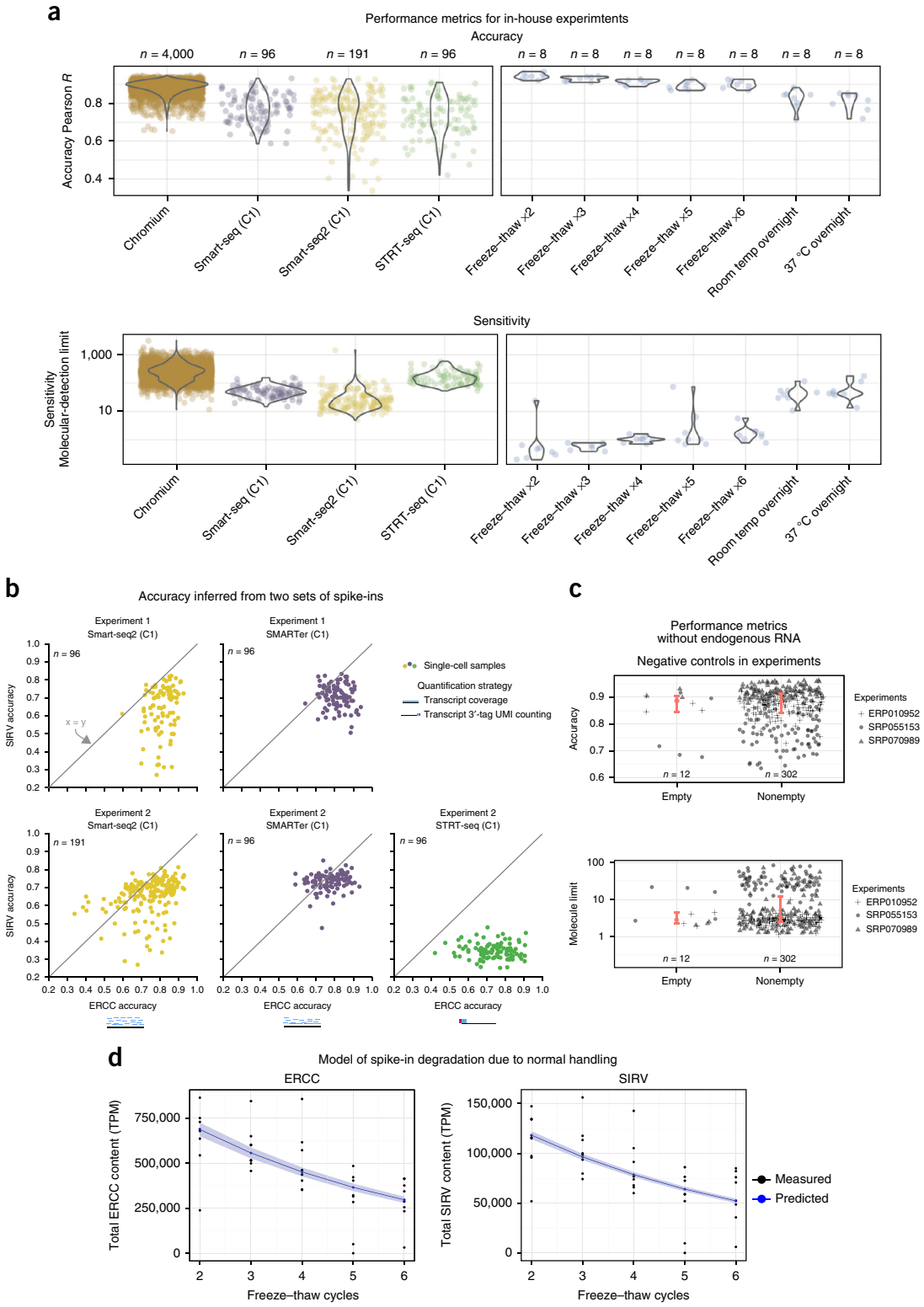

Figure 4 | Effects of various factors on performance metrics. (**a**) Batch effects and RNA degradation. Performance distributions for three protocols, implemented as a single batch, on the Fluidigm C1 (left) and 10× Chromium (far left; different batch) platforms. Performance distributions of spike-ins measured after freeze–thaw cycles, with normal (two or three cycles) to critical degradation (six cycles, left overnight at room temperature). Temp, 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, under multiple protocols. (**c**) Endogenous mRNA amount does not affect performance metrics. Comparison of performance metrics between empty (lacking endogenous mRNA) and nonempty samples from three published data sets shows similar performance and no bias due to the 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.

equally important to note that sequencing depth is a technical feature, and the number of genes detected depends on the depth. Therefore, sequencing depth must be taken into account when performing and computationally analyzing scRNA-seq data, even for compositional expression units such as transcripts per million.

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

Our performance analysis inherently assumed the gold-standard annotation of the spike-ins to be correct. However, owing to its labile nature, RNA can be degraded during the course of normal reagent handling. To quantify the effects of such degradation, we subjected spike-in molecules (both ERCCs and SIRVs) to repeated freeze–thaw cycles (Online Methods). Additionally, as a measure of complete or full degradation, we left the spike-ins either at room temperature or at 37 °C overnight. The freeze–thaw cycles emulated normal handling, and by comparing samples at different degradation levels, we observed a small overall effect on accuracy and sensitivity, which was similar to the variation within a protocol (**Fig. 4a**).

Spike-in degradation directly impinges on the effective spike-in dilution in a sample and is a central factor for calculating the technical sensitivity. We observed that normal handling accounted for molecule-limit differences within an order of magnitude, even when spike-ins were 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), compared with other samples, had a difference of two orders of magnitude, which was similar to the difference between protocols (**Fig. 4a**).

SIRV spike-ins recapitulate accuracy results with ERCC spike-ins

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

We observed that the accuracy was systematically lower when SIRVs were used. This result was expected, because the ambiguous read assignment to the isoforms introduced a noise element. Overall, when using SIRVs and ERCCs, we observed a similar pattern of relative accuracy between our SMARTer and Smart-seq2 experiments. The STRT-seq samples had very poor accuracy, as was expected, because the 5′ transcript tags alone cannot distinguish among different mRNA isoforms.

This experiment provided quantitative evidence that mRNA splice-form variation can be inferred at the single-cell level when the appropriate protocol is used. Comparing the protocols, we found that the accuracy calculated when SIRVs were used recapitulated the accuracy when ERCCs were used, thus indicating that spike-in batch variability does not generally explain differences among protocols.

Endogenous mRNA amount does not affect performance metrics with 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 high. To verify that discrepancies in endogenous mRNA levels (due to, for example, cell-type differences) do not affect performance metrics, we investigated published data in which information on empty (spike-in RNA alone) and nonempty (mRNA and spike-ins present) samples have been reported for the same batch of cells. We compared accuracy and sensitivity between empty and nonempty samples from three studies and found equivalent results, thus confirming that endogenous mRNA content does not affect performance metrics (**Fig. 4c**). We quantified the equivalence through 95% confidence interval (CI)-based equivalence analysis[12] (Online Methods). We found that the empty median CI was 100% contained within the nonempty median CI for accuracy and was 84% contained for sensitivity.

Effects 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 the results to the number of freeze–thaw cycles.

We made a predictive Bayesian model of mRNA degradation (Online Methods) with a degradation-rate parameter p. Sampling from the posterior distribution of p when applying the model to ERCC spike-ins, we found a degradation rate of 19 ± 0.7% per freeze–thaw cycle (mean ± 95% CI, **Supplementary Fig. 3**; posterior predictions in **Fig. 4d**). 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, probably because of mapping uncertainty (described in Discussion). Overall, our data approximated a 20% degradation rate of spike-ins in each freeze–thaw cycle during normal sample handling.

Although 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 was strongly affected by freeze–thaw cycles. Hence, the inference of total mRNA in cells when spike-ins are used might prove problematic. Because 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 true for RNA in general.

DISCUSSION

A previous study has shown[13] that ERCC read alignment varies widely across libraries and platforms, and some spike-ins have reproducibly poor behavior, thus raising 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, which have 250-base-long poly(A) tails[10]. Hence, poly(T) priming of ERCC spike-ins might be less efficient than that for endogenous mRNA. Furthermore, ERCC spike-ins are not capped at the 5′ end, thus possibly leading to decreased template-switching efficiency (used in several protocols) as compared with that for endogenous mRNAs[14]. Finally, unlike endogenous mRNAs, spike-in RNAs are not naturally bound by mRNA-binding proteins, nor do they have secondary structures.

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

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

We showed that handling and batch variation in ERCC dilutions led to smaller variations in performance than those observed among protocols (**Fig. 4a**). Nevertheless, in certain published experiments, spike-ins may have been greatly degraded and consequently may have affected 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, whereas 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 sequencing depth 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 weakly expressed genes, or for gaining additional insights into 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 approximately 1 million reads per sample are sequenced. Future improvements in protocols and decreases in the price of sequencing should further boost the ability to answer new questions in biology by using single-cell transcriptomics.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

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

ACKNOWLEDGMENTS

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 discussions on spike-ins, and M. Quail for help with 10× Chromium experiments. We extend our gratitude to S. Linnarsson and A. Zeisel for invaluable support in implementing STRT-seq in our laboratory and for help with sequencing the STRT library. We also thank D. Grün for sharing smFISH molecule counts. Finally we thank R. Kirchner for many improvements to the umis tool. This study was supported by Cancer Research UK grant C45041/A14953 to A.C. and C.L.; European Research Council project 677501–ZF_Blood to A.C.; a core support grant from the Wellcome Trust and MRC to the Wellcome Trust–Medical Research Council Cambridge Stem Cell Institute; ERC grant ThSWITCH to S.A.T. (grant 260507); and a Lister Institute Research Prize to S.A.T. K.N.N. was supported by the Wellcome Trust Strategic Award 'Single cell genomics of mouse gastrulation'. We thank P. Liu (Wellcome Trust Sanger Institute) for providing cells.

AUTHOR CONTRIBUTIONS

V.S. and S.A.T. conceived the study. V.S. and L.-H.L. annotated and processed all data. V.S. conceived and implemented the umis tool. V.S. conceived and performed the performance modeling of the data. V.S., R.J.M., and K.N.N. designed the in-house experiments. K.N.N. optimized and implemented the protocols. The degradation experiments were designed by V.S., I.C.M., R.J.M., and K.N.N., who performed the experiments. I.C.M. and C.L. performed zebrafish Smart-seq2 experiments under the supervision of A.C. V.S. and L.H.L. designed the degradation model, and L.H.L. implemented the model. V.S., K.N.N., and S.A.T. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

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. Preprint at http://biorxiv.org/content/early/2016/06/29/035758/ (2016).
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).

ONLINE METHODS

Mouse embryonic-stem-cell culture. Wild-type E14 mouse ES cells (kindly provided by P. Liu, Wellcome Trust Sanger Institute) were cultured on gelatin-coated dishes with Knockout DMEM (10829; Gibco), 15% fetal calf serum (FB-1001/500; batch tested from Labtech), 1× penicillin–streptomycin–glutamine (10378-016; Gibco), 1× MEM NEAA (11140-035; Gibco), 2-mercaptoethanol (31350-010; Gibco), and 1,000 U leukemia inhibitory factor (LIF; ESG1107). mESCs tested free of mycoplasma contamination were passaged every 2 or 3 d.

SMARTer, Smart-seq2 and STRT-seq on C1. E14 mESCs were trypsinized to obtain a single-cell suspension and were passed through a 30-µm filter (CellTrics; 04-0042-2316). Cells were processed with a C1 Single Cell Auto Prep System (Fluidigm; 100-7000 and 100-6209), according to the manufacturer's protocol (100-5950 B1). Briefly, we performed SMARTer, Smart-seq2, and STRT-seq each across three small C1 Open App IFCs (5–10 µm; 100-5759). The specific sample-preparation steps for the three protocols (SMARTer[3,15–18], Smart-seq2[19], and STRT-aeq[9,11,20,21]) were downloaded from the Fluidigm Script Hub. Dissociated single cells were loaded and captured on C1 Open App IFCs, and this was followed by manual inspection to demarcate empty wells, 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 (0.5 µl 1:500-diluted ERCCs + 0.6 µl 1:500-diluted SIRVs) and added to respective lysis buffer master mixes for SMARTer (20 µl), Smart-seq2 (27 µl), and STRT-seq (20 µl). 9 µl of the respective lysis master mix was added to each Open App C1 IFC. The subsequent steps (cell lysis, cDNA synthesis by reverse transcription, and PCR reaction) were performed as described in the Fluidigm Script Hub.

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

Spike-in degradation experiment using Smart-seq2 on plates. We used a 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 to produce a spike-in master mix (1:200 final dilution; termed '×2 freeze–thaw'). The spike-in master mix was divided among three tubes: one incubated overnight at 37 °C (condition 1), one incubated overnight at room temperature (condition 2), and one incubated 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 2–5 min, and an aliquot was collected and refrozen in dry ice. We repeated this freeze–thaw cycle an additional five times (conditions 3–7). All the spike-in mixes (conditions 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 X-100, 1:20 RNase inhibitor, 10 mM oligo $d(T)_{30}$ VN, and 10 mM dNTPs) into each well. 1 µl of spike-in mix per condition (conditions 1–7) was added to each well columnwise, such that each column represented a single condition with eight replicate wells. E14 mESCs were filtered through a 30-µm filter and FACS-sorted (BD Influx; BD Biosciences) into a 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 before the Smart-seq2 protocol was performed as previously described[19].

Library preparation and sequencing. Representative cDNA from single cells across three C1 runs and Smart-seq2 (on plates) was assessed with High Sensitivity DNA chips for the Agilent Bioanalyzer (5067-4626 and 5067-4627; Agilent Technologies). Single-cell cDNA from SMARTer[3,15–18] and Smart-seq2 C1 IFCs and Smart-seq2 (on plates) was tagmented and pooled to generate libraries by using an Illumina Nextera XT DNA sample-preparation kit (Illumina; FC-131-1096) with 96 dual-barcoded indices (Illumina; FC-131-1002). The library cleanup and sample pooling was performed with AMPure XP beads (Agencourt Biosciences; A63880). All protocols were as described in the Fluidigm protocol (100-5950), Fluidigm Script Hub, and Smart-seq2 protocol[19]. The STRT-seq libraries were generated and sequenced at the Karolinska Institutet as previously described[9,20]. The single-cell libraries from SMARTer and Smart-seq2 C1 IFCs and Smart-seq2 (on plates) were sequenced across 1 lane of a HiSeq V4 (Illumina) by using 75-bp/125-bp paired-end sequencing.

10× Genomics Chromium experiment. A Single Cell Gel Bead kit (120217), Single cell chip kit (120219) and Single cell library kit (120218) were used along with a 10× GemCode Single Cell Instrument, per the manufacturer's specifications and manuals (document CG00011; revision B). Equal volumes of control brain RNA (3 µl; FirstChoice Human Brain Total RNA; AM7962) and ERCC spikes (3 µl 1:4 dilution; 4456653) were mixed to produce a '2× control RNA + ERCC' master mix. We further diluted this mixture to '1× control RNA + ERCC' with PCR-grade water. We generated two single-cell master-mix preparations with 3 µl of 2× control RNA + ERCC and 1× control RNA + ERCC instead of single-cell suspension (adjusted with 34.4 µl nuclease-free water). The remaining protocol was performed according to the manufacturer's manual (document CG00011; revision B). Each 10× library was sequenced across a HiSeq2500 (2× lanes; rapid run), per Wellcome Trust Sanger Institute sequencing guidelines.

Data sources. Raw read data from published studies were downloaded from either ENA or SRA, as listed in **Supplementary Table 1**. These included Gene Expression Omnibus accession codes GSE53334 (ref. 22), GSE65785 (ref. 23), GSE67833 (ref. 24), GSE53386 (ref. 25), GSE71318 (ref. 26), GSE46980 (ref. 9),

doi:10.1038/nmeth.4220

GSE60361 (ref. 20), GSE60768 (ref. 27), GSE54695 (ref. 11), GSE78779 (ref. 28), GSE54006 (ref. 21), GSE72857 (ref. 29), GSE63473 (ref. 30), and GSE65525 (ref. 31); European Genome-phenome Archive accession code EGAS00001001204 (ref. 32); European Nucleotide Archive accession codes ERP010108 (ref. 32), ERP005640 (ref. 15), ERP006670 (ref. 16), ERP010952 (ref. 33), and ERP013160 (ref. 32); Sequence Read Archive accession codes SRP030617 (ref. 3), SRP041736 (ref. 17), SRP033209 (ref. 18), SRP055153 (ref. 34), SRP045422 (ref. 35), SRP047290 (ref. 36), SRP025171 (ref. 37), SRP050499 (ref. 38), and SRP073767 (ref. 39); and ArrayExpress accession codes E-MTAB-3346 (ref. 40) and E-MTAB-3624 (ref. 40).

Information regarding the concentration and volume of the 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 nonstandard names annotating the ERCC spike-ins, and through personal communication with the authors, we received a table for converting these to the names provided by Life Technologies. Additionally we were informed by the authors that the final spike-in dilution noted as 1:50,000 in Islam *et al.*[9] had actually been 1:20,000.

The concentrations of the ERCC solution in the dendritic-MARS table was ambiguous, because 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 were used. Thirty samples were excluded because 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.

Notes on individual data sets are provided in **Supplementary Table 1**.

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

All coverage-based data sets were sequenced with Illumina paired-end sequencing with read lengths between 75 and 150 bp.

To process all UMI-based data in a coherent manner, we developed a quantification strategy based on pseudomapping and counting evidence for transcript–UMI pairs.

The principle was to transfer information from a UMI–tag pair to a transcript–UMI pair according to which transcript the tag mapped to. Because UMI-based methods use only 3′- or 5′-end tags of cDNA, which may be as short as 25 bp, mapping of these tags is commonly ambiguous. Our strategy was to weight a UMI–tag pair according to the number of transcripts to which the tag mapped. 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 were counted (default 1) at either the gene or the transcript level. In the 10× 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 2,000 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 is 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 work, is provided as **Supplementary Software**.

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

The calculation of the accuracy of an individual sample was determined with the Pearson correlation between input concentration of the spike-ins and the measured expression values. If fewer than eight spike-ins were observed, the accuracy was set to infinity, because we considered this level 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 was modeled by the logistic function:

$$p(detected_i) = \frac{1}{1 + e^{-(a \times log(M_i) + b)}} + \epsilon$$

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 at which the logistic regression model passes 50% detection probability:

$$detection\ limit = -\frac{b}{a}$$

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

$$UMI_i = E \times M_i + \epsilon$$

However, as mentioned in the main text, the data fit the model much better when there is a non-one exponent parameter on the number of input molecules:

$$UMI_i = E \times M_i^c + \epsilon$$

When we modeled the relationship between the read depth and performance metrics for individual protocols, we used 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:

$$metric_i = a^2 \times \log_{10}(reads_i) + b \times \log_{10}(reads_i) + performance_{protocol} + \epsilon$$

Here, the performance metric plateaus and saturates when

$$\log_{10}(reads) = -\frac{b}{2a}$$

The linear models were fitted and analyzed with 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 with Stan[45] (R package rstan v 2.10.1). The model was specified as the following: the prior for p was the uniform distribution between 0 and 1, and F_i for each spike-in i had their priors defined as the normal distribution with a mean of 0.5 and an s.d. of 1. F_{ij} was modeled by a normal distribution with mean $F_i \times (1-p)^j$, where j is the j-th freeze–thaw cycle, and s.d. σ had the uniform distribution between 0 and 20 as a prior. The model posterior was sampled with 5,000 iteration steps, 1,000 warm-up steps and four chains.

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

Data availability. All data generated in this study have been deposited in the ArrayExpress database under accession codes E-MTAB-5480, E-MTAB-5481, E-MTAB-5482, E-MTAB-5483, E-MTAB-5484, E-MTAB-5485, and E-MTAB-5486. Summary tables are provided as supplementary files.

15. 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).
16. Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **33**, 155–160 (2015).
17. 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).
18. Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509**, 371–375 (2014).
19. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
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. 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).
22. 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).
23. 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).
24. 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).
25. Fan, X. *et al.* Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. *Genome Biol.* **16**, 148 (2015).
26. Dang, Y. *et al.* Tracing the expression of circular RNAs in human pre-implantation embryos. *Genome Biol.* **17**, 130 (2016).
27. Velten, L. *et al.* Single-cell polyadenylation site mapping reveals 3′ isoform choice variability. *Mol. Syst. Biol.* **11**, 812 (2015).
28. Hashimshony, T. *et al.* CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 77 (2016).
29. Paul, F. *et al.* Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* **163**, 1663–1677 (2015).
30. Macosko, E.Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
31. Klein, A.M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
32. Macaulay, I.C. *et al.* G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12**, 519–522 (2015).
33. Scialdone, A. *et al.* Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* **85**, 54–61 (2015).
34. 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).
35. 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).
36. 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).
37. Streets, A.M. *et al.* Microfluidic single-cell whole-transcriptome sequencing. *Proc. Natl. Acad. Sci. USA* **111**, 7048–7053 (2014).
38. Guo, F. *et al.* The transcriptome and DNA methylome landscapes of human primordial germ cells. *Cell* **161**, 1437–1452 (2015).
39. Zheng, G.X.Y. *et al.* Massively parallel digital transcriptional profiling of single cells. Preprint at http://biorxiv.org/content/early/2016/07/26/065912/ (2016).
40. 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).
41. Patro, R., Duggal, G., Love, M.I., Irizarry, M.A. & Kingsford, C. Salmon provides accurate, fast, and bias-aware transcript expression estimates using dual-phase inference. Preprint at http://biorxiv.org/content/early/2016/08/30/021592/ (2015).
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. *et al.* Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
45. Carpenter, B., Gelman, A., Hoffman, M., Lee, D. & Goodrich, B. Stan: A probabilistic programming language. *J. Stat. Softw.* **76**, 1–32 (2017).

doi:10.1038/nmeth.4220