Quantitative single-cell RNA-seq with unique molecular identifiers

Saiful Islam[1], Amit Zeisel[1], Simon Joost[2], Gioele La Manno[1], Pawel Zajac[1], Maria Kasper[2], Peter Lönnerberg[1] & Sten Linnarsson[1]

Single-cell RNA sequencing (RNA-seq) is a powerful tool to reveal cellular heterogeneity, discover new cell types and characterize tumor microevolution. However, losses in cDNA synthesis and bias in cDNA amplification lead to severe quantitative errors. We show that molecular labels—random sequences that label individual molecules—can nearly eliminate amplification noise, and that microfluidic sample preparation and optimized reagents produce a fivefold improvement in mRNA capture efficiency.

RNA-seq has become the method of choice for transcriptome analysis in tissues[1–3] and in single cells[4–7]. The two main challenges in single-cell RNA-seq are the efficiency of cDNA synthesis (which sets the limit of detection) and the amplification bias (which reduces quantitative accuracy). Published protocols have been reported to have limits of detection of between five and ten mRNA molecules[5–7], corresponding to a capture efficiency of around 10%, and all current methods use amplification, either by PCR or by *in vitro* transcription.

To correct for amplification bias, we[8] and others[9–11] have described how molecules can be directly counted through the use of unique molecular identifiers (UMIs). For single-cell RNA-seq, UMIs have been used as an internal validation control[12] but have not yet been explored as a direct, quantitative measure of gene expression. Molecule counting corrects for PCR-induced artifacts (**Supplementary Fig. 1**) and provides an absolute scale of measurement with a defined zero level. In contrast, standard RNA-seq uses relative measures such as reads per kilobase per million reads (RPKM), which mask differences in total mRNA content. For example, a gene may be 'upregulated' in terms of RPKM and have a decrease in absolute expression level if the total mRNA content also changes. Thus an absolute scale of measurement is crucial for interpreting transcriptional dynamics in single cells.

We applied molecule counting to mouse embryonic stem (ES) cells and used spike-in controls to monitor technical performance (similar results were obtained in independent experiments on ES cells, and hypothalamic and cortical primary neurons). During reverse transcription, each cDNA molecule was tagged with a 5-bp random sequence serving as UMI (**Fig. 1a** and **Supplementary Fig. 2**). We counted cDNA molecules by enumerating the total number of distinct UMIs aligned to each position (**Fig. 1b**). Mouse embryonic fibroblasts and damaged ES cells were removed on the basis of criteria established after sequencing (**Supplementary Note**).

UMIs will reflect molecule counts only if the number of distinct labels is substantially larger than the typical number of identical molecules. Approximately 10^5–10^6 mRNA molecules are present in a typical single mammalian cell, and up to 10,000 different genes may be expressed. However, many genes are expressed from multiple promoters (**Supplementary Fig. 3**) or have promoters with diffuse transcription start sites (**Supplementary Fig. 4**), so that the number of identical mRNA molecules is expected to be <100 for most genes. We therefore expected our 5-bp UMI, capable of distinguishing up to 1,024 molecules, to be sufficient. To confirm this, we determined the number of distinct UMIs that we could observe for each unique combination of sample barcode (i.e., cell) and genomic position (**Fig. 1c**). As expected, the vast majority of cases were represented by only a small number of UMIs (corresponding to a small number of cDNA molecules), and we did not find a single case of a fully saturated position. To obtain more accurate estimates of molecule counts at high expression levels, we corrected for the collision probability of UMIs (see Online Methods).

To ensure that successfully generated cDNA molecules are sequenced, it is crucial to sequence to a sufficient depth after amplification. We typically observed each UMI multiple times (**Fig. 1b**). Across all genes, the average number of reads per molecule was nine, with a distribution consistent with oversampling of most molecules (**Supplementary Fig. 5**).

To further demonstrate that UMIs labeled individual cDNA molecules, we examined genes containing a heterozygous single nucleotide polymorphism (SNP). If UMIs worked as intended, each UMI would be derived from a single molecule and therefore from a single allele. As a consequence, all reads derived from this UMI would carry the same allele. Indeed, the observed allele distribution confirmed that UMIs had correctly labeled single cDNA molecules derived from single alleles (**Fig. 1d**). In total, we found 47 informative SNPs, of which all showed the expected monoallelic pattern across UMIs.

To improve cDNA synthesis efficiency, we implemented our protocol on a commercially available microfluidic platform (Fluidigm C1 AutoPrep) and carefully optimized the conditions for this device (Online Methods). To directly measure the efficiency of reverse transcription, we introduced a known number

[1]Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden. [2]Department of Biosciences and Nutrition, Karolinska Institutet, Stockholm, Sweden. Correspondence should be addressed to S.L. (sten.linnarsson@ki.se).

RECEIVED 27 SEPTEMBER; ACCEPTED 25 NOVEMBER; PUBLISHED ONLINE 22 DECEMBER 2013; DOI:10.1038/NMETH.2772

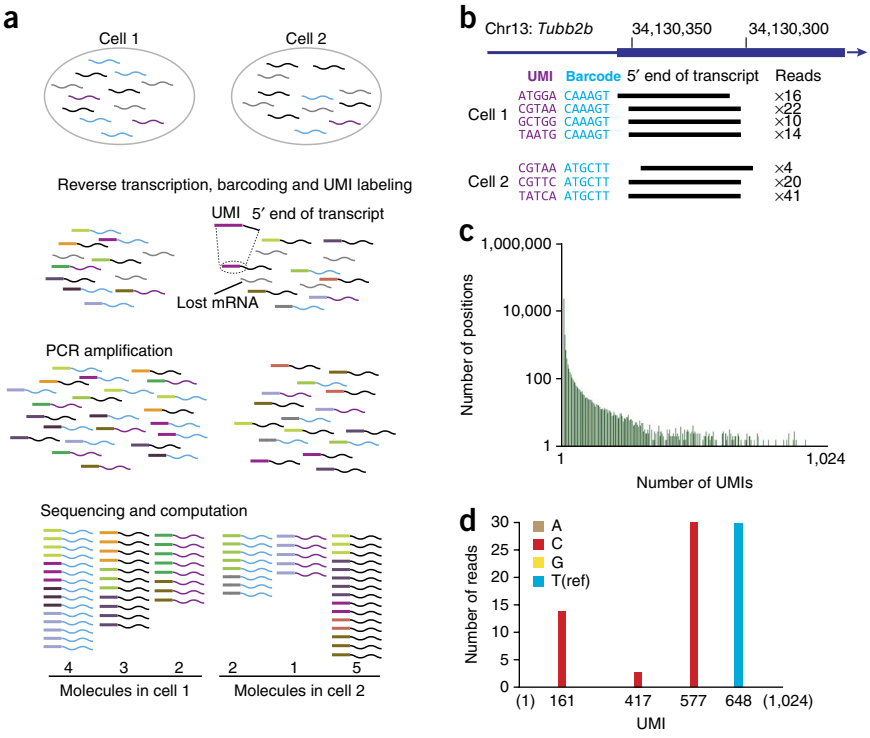

Figure 1 | Molecule counting using UMIs. (**a**) Overview of tagging single mRNA molecules with UMIs. Two cells are shown (top) containing mRNAs from different genes represented by distinct colors. UMIs are represented by colored boxes (middle and bottom); untagged mRNA molecules (gray, middle) were not reverse transcribed (bottom). (**b**) UMI alignment and mRNA molecule counting on a hypothetical example of reads aligned to *Tubb2b*. (**c**) Number of genomic positions that were assigned the given numbers of distinct UMIs. (**d**) UMIs observed in single ES cells for the *Exosc1* gene, which contained a SNP (rs13483630). Stacked bars indicate the number of reads carrying each possible UMI sequence, colored according to the allele observed in the read. Four distinct UMIs were observed, all monoallelic. Horizontal axis indicates the numbers of each UMI counted. Ref, reference allele.

of External RNA Controls Consortium (ERCC) control RNA molecules to each well. Counting the resulting number of detected cDNA molecules, we found an efficiency of 48 ± 5% s.d. (**Fig. 2** and **Supplementary Fig. 6**), a fivefold improvement over our previously published protocol[5]. We attribute the improvement both to the use of an integrated microfluidic device (reducing losses and minimizing background reactions) and to more optimized reagents, in particular the template-switching oligo design. Interestingly, using a separate set of optimizations, the recently published Smart-seq2 method achieved similarly improved capture effiency[13], suggesting that even higher efficiencies could be achieved by combining the protocols.

Next, we asked how the use of UMIs affected the overall quantification of gene expression in single cells and controls. For comparison, we analyzed the same data sets both using UMIs (counting molecules) and using reads in the

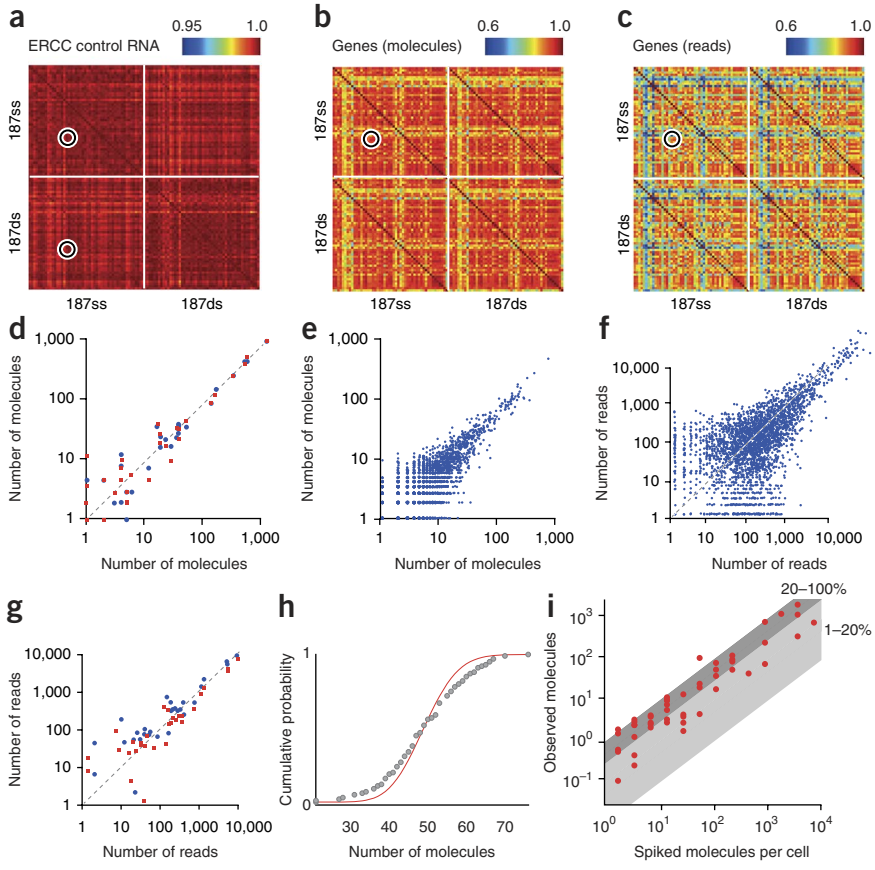

Figure 2 | Reproducibility of molecule counting. (**a**,**b**) Pairwise correlation coefficients calculated for ERCC spike-in control RNA (**a**) or endogenous genes (**b**), using molecule counts ($n = 41$) and prepared with (187ds) and without (187ss) nine additional cycles of library PCR. (**c**) Pairwise correlation coefficients as in **b** but counting reads instead of molecules ($n = 41$). (**d**) Scatterplot showing the pairwise comparison of two wells indicated in **a**. Red squares and blue dots show comparisons within and between libraries, respectively. (**e**) Scatterplot showing the two cells indicated in **b** based on molecule counts. (**f**) Scatterplot as in **e** but using reads instead of molecules. (**g**) Scatterplot as in **d** but using reads instead of molecules. (**h**) Distribution of molecule counts for a single ERCC spike-in transcript (gray dots) compared with the cumulative density function of the Poisson distribution (red line). (**i**) mRNA capture efficiency shown as observed molecule counts versus number of spiked-in molecules for ERCC control RNA transcripts. The shaded bands indicate efficiencies above (dark gray) and below (light gray) 20%. Each red dot represents the average of a single ERCC RNA across 96 wells. Similar results were obtained in one replicate experiment.

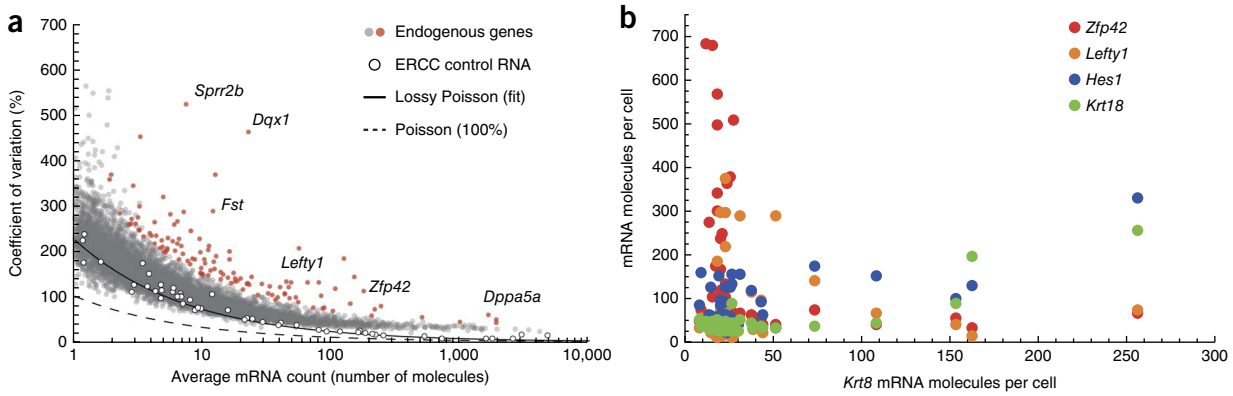

Figure 3 | Transcriptional noise in ES cells. (**a**) The coefficient of variation as a function of the mean number of mRNA molecules detected, for genes expressed in ES cells (n = 41 single ES cells; gray and red circles) and for internal ERCC controls spiked into every well (n = 41 replicates; white circles). The solid black line shows a lossy Poisson curve fit to the ERCC technical controls. The dashed line indicates the pure Poisson sampling noise. Genes with significant (α = 0.05; see Online Methods) excess noise are shown as red dots, and examples are indicated by name. (**b**) Coexpression of noisy genes. The expression of four genes, across 41 ES cells, is plotted against the expression of *Krt8*. Two branches were observed, interpreted as 'epiblast-like' (high *Krt8*, *Krt18* and *Hes1*) and 'pluripotent-like' (high *Zfp42* and *Lefty1*). Similar results were obtained in one replicate experiment.

conventional manner. We determined the pairwise correlation coefficients (**Fig. 2a–c**) and visualized typical examples as scatterplots (**Fig. 2d**,**e**,**g**,**h**). As expected, the quantitative precision was improved, especially at low molecule counts. The technical reproducibility was excellent, as demonstrated by correlation coefficients >0.95 for ERCC spike-in control RNA (**Fig. 2a**).

An important consequence of counting molecules is that the data can be displayed on an absolute and biologically meaningful scale, with a defined zero. The scales for read-count scatterplots (**Fig. 2g**,**h**) are arbitrary, as the total number of reads differs between wells (and could be increased arbitrarily through additional sequencing runs). Normalizing to reads per million (RPM) amounts only to scaling by a constant factor, and affects neither correlation coefficients nor scatterplots. (Normalizing to RPKM would distort the results, as we sequenced only the 5′ end of each mRNA, and thus read number was not proportional to gene length; **Supplementary Fig. 7**). In contrast, scales for molecule-counting scatterplots (**Fig. 2d**,**e**) are absolute and would not change appreciably if the number of reads were increased. We observed a smooth distribution of the number of counted molecules, consistent with accurate counting and approaching the theoretically optimal Poisson distribution (**Fig. 2h** and **Supplementary Fig. 8**).

Examining noise as a function of mRNA abundance, we found that the ERCC spike-in controls closely tracked the expected Poisson distribution (**Fig. 3a**). However, at low levels of expression, the capture efficiency limited our power to detect noise. This can be seen in the difference between the ERCC spike-in controls and the curve for a fully lossless and perfectly accurate measurement (Poisson). Nevertheless, this result demonstrates that little noise above that caused by inefficient reverse transcription was introduced during sample preparation.

Next, we examined endogenous transcriptional noise in ES cells. We would expect noise across these measurements to include the sampling noise introduced by reverse transcription as well as biological noise due to stochastic or bursty transcription and to oscillatory and regulated gene expression. We were surprised to find that most genes were expressed at noise levels approaching the Poisson limit (**Fig. 3a**). At all levels of expression, the noise measured as coefficient of variation (s.d. divided by the mean) closely tracked the predicted Poisson limit, but with a slight excess above the technical noise (compare ERCC controls with endogenous genes).

Although most genes showed low levels of noise, using a conservative threshold (see Online Methods) we found a set of 118 significantly noisy genes in ES cells (P = 0.05; **Fig. 3a** and **Supplementary Table 1**). Interestingly, several of these noisy genes have been previously demonstrated to show heterogeneous (*Zfp42*, also known as *Rex*)[14] or oscillatory (*Hes1*)[15] expression in ES cells. *Nanog* was also heterogeneously expressed ($P < 0.05$), as expected[16], but was excluded by our stringent noise criterion. Furthermore, pathway analysis identified an over-representation of genes involved in the transforming growth factor–β signaling pathway (*Id2*, *Id3*, *Fst*, *Lefty1* and *Lefty2*), which has recently been shown to regulate ES cell heterogeneity and self-renewal[17]. These results directly validate our approach and suggest that the noisy genes we discovered were not simply affected by stochastic transcription but reflect more complex biological heterogeneity.

We hypothesized that noisy expression partially reflected a kind of 'resonant state' between pluripotency and early differentiation. In agreement with this idea, the genes encoding keratins 8, 18 and 19 (Krt8, Krt18 and Krt19)—which are known to be expressed at low levels in mouse ES cells but highly induced upon embryoid body formation and in epiblast stem cells[18]—were among the noisy genes and were coexpressed in a small subset of cells (**Fig. 3b**). Using *Krt8* to separate these cellular substates, we found that cells ranged between two extreme states characterized by high expression of *Hes1*, *Krt8* and *Krt18* (epiblast-like state) and *Zfp42* and *Lefty1* (pluripotent-like state), respectively. Thus the observed variations in these genes reflect a stochastic distribution of ES cells in regulatory substates rather than intrinsic fluctuations in transcription. A more detailed analysis of these findings will be published elsewhere.

The finding that the majority of genes expressed in ES cells had low levels of noise contrasts with some previous findings of large and widespread intrinsic transcriptional noise[19,20]. It is plausible that the discrepancy in findings can be partly explained by methodological differences. It is difficult to estimate the contribution of technical noise in imaging-based methods, and previous authors have generally assumed that all observed noise was of a biological

origin. On the other hand, we lost some mRNA molecules during cDNA synthesis, which would mask some biological noise, especially at low expression levels. Nevertheless, the notion that gene expression in general is extremely bursty seems to be incompatible with the levels of noise that we observed.

In conclusion, we have shown that quantitatively accurate single-cell RNA-seq can uncover oscillatory and heterogeneous gene expression within a single cell type. We anticipate that accurate molecule counting will be an important approach for future single-cell transcriptome analyses.

METHODS

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

Accession codes. Gene Expression Omnibus: supplementary data are available under accession code GSE46980.

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

ACKNOWLEDGMENTS

Illumina sequencing experiments were performed by A. Johnsson and A. Juréus (Karolinska Institutet). ES cells were generously provided by the Karolinska Center for Transgene Technologies. Recombinant Tn5 transposase was a generous gift from R. Sandberg (Karolinska Institutet). We are grateful for the advice and support of B. Jones and B. Fowler (Fluidigm). This work was supported by grants from the Swedish Research Council (STARGET) and the European Research Council (261063) to S.L. and from the Swedish Cancer Society, Swedish Research Council and Ragnar Söderberg Foundation to M.K.

AUTHOR CONTRIBUTIONS

S.L. conceived the study. S.I., A.Z. and P.Z. developed the protocols and performed the RNA-seq experiments. S.J. and M.K. prepared cells. A.Z., G.L.M. and S.I. developed the tagmentation protocol. P.L. wrote the bioinformatics pipeline. P.L., S.L. and A.Z. analyzed data. S.L. and S.I. wrote the manuscript with support from all authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. *Nat. Methods* **5**, 621–628 (2008).
2. Cloonan, N. *et al. Nat. Methods* **5**, 613–619 (2008).
3. Wang, Y. *et al. Nat. Genet.* **40**, 1478–1483 (2008).
4. Tang, F. *et al. Nat. Methods* **6**, 377–382 (2009).
5. Islam, S. *et al. Genome Res.* **21**, 1160–1167 (2011).
6. Ramsköld, D. *et al. Nat. Biotechnol.* **30**, 777–782 (2012).
7. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. *Cell Rep.* **2**, 666–673 (2012).
8. Kivioja, T. *et al. Nat. Methods* **9**, 72–74 (2012).
9. Shiroguchi, K., Jia, T.Z., Sims, P.A. & Xie, X.S. *Proc. Natl. Acad. Sci. USA* **109**, 1347–1352 (2012).
10. Fu, G.K., Hu, J., Wang, P.H. & Fodor, S.P. *Proc. Natl. Acad. Sci. USA* **108**, 9026–9031 (2011).
11. Casbon, J.A., Osborne, R.J., Brenner, S. & Lichtenstein, C.P. *Nucleic Acids Res.* **39**, e81 (2011).
12. Shalek, A.K. *et al. Nature* **498**, 236–240 (2013).
13. Picelli, S. *et al. Nat. Methods* **10**, 1096–1098 (2013).
14. Carter, M.G. *et al. Gene Exp. Patterns* **8**, 181–198 (2008).
15. Kobayashi, T. *et al. Genes Dev.* **23**, 1870–1875 (2009).
16. Chambers, I. *et al. Nature* **450**, 1230–1234 (2007).
17. Galvin-Burgess, K.E., Travis, E.D., Pierson, K.E. & Vivian, J.L. *Stem Cells* **31**, 48–58 (2013).
18. Maurer, J. *et al. PLoS ONE* **3**, e3451 (2008).
19. Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y. & Tyagi, S. *PLoS Biol.* **4**, e309 (2006).
20. Dar, R.D. *et al. Proc. Natl. Acad. Sci. USA* **109**, 17454–17459 (2012).

ONLINE METHODS

Cell culture. R1 mouse ES cells (not authenticated by short tandem repeat profiling but routinely used for successful generation of chimeras and tested for mycoplasma contamination) were grown in ES1 medium on irradiated mouse embryonic fibroblast feeder cells and harvested at confluency. After trypsinization, the feeder cells were given time to settle to obtain a pure R1 ES cell suspension. To reduce contamination with dead and dying cells, dead cells were subsequently stained with Red Fixable Dead Cell Stain (Life Technologies) and dead and early apoptotic cells were depleted with annexin V–conjugated microbeads (Miltenyi Biotec). The cells were resuspended at a concentration of 400 cells per microliter in ES1 medium with 10% DMSO, divided into aliquots and frozen at −80 °C.

Cell capture and imaging. A 30-µL aliquot of ~12,000 cells was thawed, and 20 µL C1 Suspension Reagent was added (all 'C1' reagents were from Fluidigm, Inc.). Five microliters of this mix were loaded according to the manufacturer's protocol on a C1 Single-Cell AutoPrep IFC microfluidic chip designed for 10- to 17-µm cells, and the chip was then processed on a Fluidigm C1 instrument using the 'mRNA Seq: Cell Load (1772x/1773x)' script. This captured one cell in each of up to 96 capture chambers and took ~30 min. The plate was then transferred to an automated microscope (Nikon TE2000E), and an image was acquired from each site using µManager (http://micro-manager.org/), which took <15 min.

Lysis, reverse transcription and PCR. The plate was returned to the lab and 20 µL lysis buffer (0.15% Triton X-100, 1 U/µL TaKaRa RNase inhibitor, 4 µM reverse transcription primer C1-P1-T31, 5% C1 Loading Reagent and 1:50,000 Life Technologies ERCC Spike-In Mix 1), reverse transcription mix (1× SuperScript II First-Strand Buffer supplemented with 3 mM $MgCl_2$, 1.5 mM dNTP, 4 mM DTT, 3.3% C1 Loading Reagent, 1.8 µM template-switching oligo C1-P1-RNA-TSO, 1.5 U/µL TaKaRa RNase inhibitor and 18 U/µL Life Technologies Superscript II reverse transcriptase) and PCR mix (1.1× Clontech Advantage2 PCR buffer, 440 µM dNTP, 530 nM PCR primer C1-P1-PCR-2 (**Supplementary Table 2**), 5% C1 Loading Reagent and 2× Advantage2 Polymerase Mix) were added to the designated wells according to the manufacturer's instructions, but using the indicated mixes in place of the corresponding commercial reagents. The plate was then returned to the Fluidigm C1 and the 'mRNA Seq: RT + Amp (1772x/1773x)' script was executed, which took ~8.5 h and included lysis, reverse transcription and 21 cycles of PCR. When the run finished, the amplified cDNA was harvested in a total of 13 µL C1 Harvesting Reagent and quantified on an Agilent BioAnalyzer. The typical yield was 1 ng per microliter (**Supplementary Fig. 9**).

Tagmentation and isolation of 5′ fragments. Amplified cDNA was simultaneously fragmented and barcoded by 'tagmentation', i.e., using Tn5 DNA transposase to transfer adaptors to the target DNA. Ninety-six different 10× transposome stocks (6.25 µM barcoded adaptor C1-TN5-*x*, 40% glycerol, 6.25 µM Tn5 transposase, where *x* denotes a well-specific barcode) were prepared, each with a different barcode sequence (**Supplementary Table 2**). Six microliters of harvested cDNA was mixed with 5 µL tagmentation buffer (50 mM TAPS-NaOH, pH 8.5, 25 mM $MgCl_2$ and 50% DMF), 11.5 µL nuclease-free water and 2.5 µL 10× transposome stock. The mix was incubated for 5 min at 55 °C then cooled on ice. Dynabeads MyOne Streptavidin C1 beads (100 µL) were washed in 2× BWT (10 mM Tris-HCl, pH 7.5, 1 mM EDTA, 2 M NaCl, 0.02% Tween-20) then resuspended in 2 mL 2× BWT. Twenty microliters of beads were added to each well and incubated at room temperature for 5 min. All fractions were pooled, the beads were immobilized and the supernatant removed (thus removing all internal fragments and retaining only the 5′- and 3′-most fragments). The beads were then resuspended in 100 µL TNT (20 mM Tris, pH 7.5, 50 mM NaCl, 0.02% Tween), washed in 100 µL Qiagen Qiaquick PB, then washed twice in 100 µL TNT. The beads were then resuspended in 100 µL restriction mix (1× NEB NEBuffer 4, 0.4 U/µL PvuI-HF enzyme), designed to cleave 3′ fragments carrying the PvuI recognition site. The mix was incubated for 1 h at 37 °C, then washed three times in TNT. Finally, the single-stranded library was eluted by resuspension of the beads in 100 µL 100 mM NaOH, incubation for 5 min, removal of the beads and addition of 100 µL 100 mM HCl and 50 µL neutralization buffer (200 mM Tris, pH 7.5, 0.05% Tween-20). At this point, the typical yield was 1–5 nM single-stranded library. Optionally (but not recommended), the library was further amplified for nine cycles as previously described[21]. This additional amplification was used only for the 187ds sample (**Fig. 2**).

Illumina high-throughput sequencing. The library was quantified by quantitative PCRusing KAPA Library Quant (Kapa Biosystems) and then sequenced on an Illumina HiSeq 2000 instrument using C1-P1-PCR-2 as the read 1 primer, and C1-TN5-U as the index read primer. Reads of 50 bp were generated along with 8-bp index reads corresponding to the cell-specific barcode. Reagent costs were reduced more than four-fold compared to commercial kits (**Supplementary Fig. 10**).

Data analysis. Every read that was considered valid by the Illumina HiSeq control software was processed and filtered as follows: (i) any 3′ bases with a quality score of B were removed; (ii) the well-identifying barcode was extracted from the 5′ end of the read; (iii) if the read ended in a poly(A) sequence leaving <25 transcript-derived bases, the read was discarded; (iv) the UMI was extracted, and if any of the UMI bases had a Phred score <17, the read was discarded; (v) a maximum of nine template-switching generated Gs were removed from the 5′ end of the transcript-derived sequence; (vi) if the remaining sequence consisted of fewer than six non-A bases or a dinucleotide repeat with fewer than six other bases at either end, the read was discarded.

After filtering, the reads were aligned to the genome using the Bowtie aligner[22], allowing for up to three mismatches and up to 24 alternative mappings for each read. The genome included an artificial chromosome, containing a concatamer of the ERCC control sequences. Any reads with no alignments were realigned against another artificial chromosome, containing all possible splice junctions arising from the exons defined by the known transcript variants. Reads mapping within these splice junctions were translated back to the corresponding actual genomic positions.

The UCSC transcript models[23] were used for the expression level calculation. If a locus had several transcript variants, the exons of these were merged to a combined model that represented

all expression from the locus. To account for incomplete cap site knowledge, the 5′ ends of all models were extended by 100 bases, but not beyond the 3′ end of any upstream nearby exon of another gene of the same orientation.

The annotation step was performed barcode by barcode. For every unique mapping (genomic position and strand) the number of reads in each UMI was counted. Any multiread that had one or more repeat mappings that was outside exons was assigned randomly as one of these repeats and contributed to the summarized read count of that repeat class. Else, if it had one or more mappings to exons, it was assigned at the exon where it was closest to the transcript model 5′ end, even if the sequence was repeat-like. If it had no exon mapping, it was assigned randomly at one of the mappings. After assigning reads, the number of molecules at each mapping position was determined by the number of distinct UMIs observed. To account for UMIs that stem from PCR-induced mutations or sequencing errors, any UMI that had fewer reads than 1/100 of the average of the nonzero UMIs was excluded. The raw UMI count was corrected for the UMI collision probability (important only at high UMI counts) as described[10]. The expression level of each transcript model was calculated as the total number of molecules assigned to all its possible mapping positions. MEFs and damaged cells were removed on the basis of quality control criteria (**Supplementary Note** and **Supplementary Figs. 11–13**).

Detecting noisy genes. We searched for noisy genes by selecting genes whose noise (measured as CV) was high compared to the pure Poisson distribution (**Fig. 3a**, dashed line), i.e., where

$$\sigma_{\text{Gene}} > 3.7\sigma_{\text{Poisson}} + 0.3$$

The rationale for this selection was twofold: first, that 4 s.d. was enough to exclude all spike-in controls, thus minimizing the false discovery rate; second, that there remained a residual 30% noise even at high molecule counts (for example, compare endogenous genes with the spike-in molecules at >1,000 molecules per cell in **Fig. 3a**). We found 167 such noisy genes in total.

To validate the findings, we performed the following statistical test. We noted that the technical noise distribution (**Fig. 3**) closely followed that of a Poisson, but its CV was inflated by a constant factor. This can be modeled as a loss factor f, such that only a fraction f of the molecules are actually observed, which inflates the CV from $1/\sqrt{\lambda}$ to $1/\sqrt{f\lambda}$ as f approaches zero. To determine the loss factor, we used maximum likelihood to fit the $\text{CV} = 1/\sqrt{f\lambda}$ function to the ERCC spike-in controls, as indicated in **Figure 3a**, and found $f = 0.20$. Note that the loss factor combines the effect of actual losses with all other sources of noise, such as differences in reaction conditions, sequencing depth and pipetting errors.

We then analyzed each gene in comparison with the ERCC controls, with the null hypothesis that the observed CV_{obs} of the gene was equal to that of a hypothetical ERCC control with the same mean expression. We obtained an expected distribution of CVs by repeatedly (500 times) calculating the CV of 41 (equal to the number of cells) random samples from the Poisson distribution with mean $f\lambda$ (where $f = 0.20$). We then verified by Pearson's chi-square goodness-of-fit test that the sampled CVs were normally distributed (omitting genes where $P < 0.05$ for this test). Finally, we fit a normal distribution to the sampled CVs and used its cumulative density function to calculate the P value of CV_{obs} along this distribution, applying a Bonferroni correction to control for multiple testing. We considered a gene noisy when its corrected significance was $\alpha < 0.05$.

On the basis of these criteria, we found a total of 118 noisy genes in ES cells (**Supplementary Table 1**). We performed gene set enrichment analysis using DAVID[24].

21. Islam, S. *et al. Nat. Protoc.* **7**, 813–828 (2012).
22. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. *Genome Biol.* **10**, R25 (2009).
23. Meyer, L.R. *et al. Nucleic Acids Res.* **41**, D64–D69 (2013).
24. Huang, D.W. *Nat. Protoc.* **4**, 44–57 (2009).

doi:10.1038/nmeth.2772