Validation of noise models for single-cell transcriptomics

Dominic Grün[1–3], Lennart Kester[1–3] & Alexander van Oudenaarden[1,2]

Single-cell transcriptomics has recently emerged as a powerful technology to explore gene expression heterogeneity among single cells. Here we identify two major sources of technical variability: sampling noise and global cell-to-cell variation in sequencing efficiency. We propose noise models to correct for this, which we validate using single-molecule FISH. We demonstrate that gene expression variability in mouse embryonic stem cells depends on the culture condition.

© 2014 Nature America, Inc. All rights reserved.

The impact of stochastic gene expression on phenotypic variation has been subject to intense research during the past several years[1,2]. The availability of diverse single-cell sequencing methods[3–8] now permits the analysis of single-cell transcriptomes with high sensitivity[9]. However, owing to low amounts of input material, single-cell sequencing still suffers from substantial levels of technical noise. Experimental and computational strategies have been proposed to alleviate the impact of technical noise. For example, unique molecular identifiers (UMIs)[10–12] were recently used in single-cell sequencing[13–15] to reduce PCR amplification bias. In a previous study, a mathematical model was introduced for assessing statistical significance of observed cell-to-cell variability[16]. Although this model reveals whether expression noise significantly exceeds technical noise, it does not permit quantification of biological gene expression noise.

Here we present a method to accurately quantify technical and biological variability in the absolute numbers of mRNA molecules in single-cell sequencing experiments. To measure gene expression noise in mouse embryonic stem cells (mESCs), we hand-picked individual cells cultured in two-inhibitor (2i) medium[17], deposited them into single test tubes and spiked in a mixture of 92 synthetic RNAs covering a wide range of expression levels[18] (**Fig. 1a**). We then performed single-cell sequencing using the CEL-Seq (cell expression by linear amplification and sequencing) method[3] with small modifications (Online Methods). To quantify technical noise, we eliminated biological variability of cellular mRNA abundance by pooling thousands of cells and then splitting them into single-cell equivalents (~20 pg) of RNA (Online Methods). These pool-and-split control samples (hereafter referred to as controls) were sequenced in the same way as single cells (**Fig. 1a**). In total, we sequenced the transcriptome of 74 cells and 76 controls (**Supplementary Table 1**, **Supplementary Note 1** and **Supplementary Fig. 1**).

To count individual transcripts we integrated an UMI barcode consisting of four random nucleotides into the primer used for reverse transcription. A similar UMI method was recently described for another single-cell sequencing protocol[14]. For each transcript species, the number of observed UMIs was converted into transcript counts (Online Methods). The spike-in RNA of known composition and concentration (Online Methods) allowed us to convert the number of sequenced transcripts to actual abundances in cells and controls. By a linear regression of sequenced spike-in transcripts on the number of added spike-in transcripts, we inferred an average conversion factor β of ~0.034 (**Supplementary Fig. 2a**). We sequenced 102 million reads for 74 cells, and each transcript was sequenced seven times on average (**Supplementary Fig. 2b** and **Supplementary Table 1**), suggesting that, in theory, up to 600 cells could be sequenced on a single lane (assuming 120 million reads per lane). However, we would like to caution that combining a large number of samples could lead to underrepresentation of particular samples, and it is therefore advisable to stay below this theoretical maximum. We note that in our data, each individual cell was oversequenced at least fourfold.

The UMI allows reliable transcript counting up to ~500 copies, which, at a sensitivity of 3.4%, corresponds to ~15,000 transcripts per cell (**Supplementary Fig. 2c**). Consistent with this, we observed only a handful of transcripts with >200 unique barcodes (**Supplementary Fig. 2d**), and for 99% of the transcript species, <50 unique barcodes were sequenced. Therefore, an UMI of length 4 is sufficient.

We investigated the impact of UMI versus read-based expression quantification on technical noise by comparing average expression across controls in reads per million (RPM) and normalized mRNA abundance in transcripts per million (TPM). Substantial differences (greater than twofold) were observed across the entire dynamic range (**Supplementary Fig. 2e**,**f**), a result suggesting that the CEL-Seq method introduces substantial amplification bias. Technical noise, assessed by the coefficient of variation (CV), was reduced for almost all genes (**Fig. 1b** and **Supplementary Fig. 2g**) by about 50% on average, and for many genes by more than twofold, when expression was measured with UMIs instead of reads.

Expression of the pluripotency markers *Pou5f1*, *Sox2* and *Klf4* was high, whereas differentiation markers showed almost no expression in all samples, suggesting that we sequenced healthy nondifferentiated cells (**Supplementary Fig. 3a**,**b**).

To understand the origin of technical noise, we first computed the CV for all genes across control samples. For low-expression

[1]Hubrecht Institute-KNAW (Royal Netherlands Academy of Arts and Sciences), Utrecht, The Netherlands. [2]University Medical Center Utrecht, Cancer Genomics Netherlands, Utrecht, The Netherlands. [3]These authors contributed equally to this work. Correspondence should be addressed to A.v.O. (a.vanoudenaarden@hubrecht.eu).

RECEIVED 16 OCTOBER 2013; ACCEPTED 25 MARCH 2014; PUBLISHED ONLINE 20 APRIL 2014; DOI:10.1038/NMETH.2930

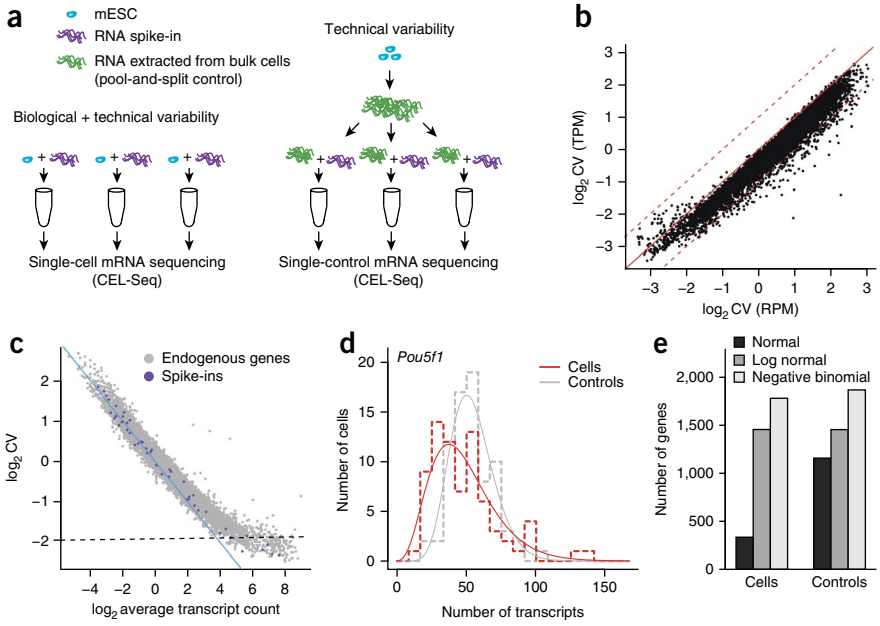

Figure 1 | Analysis of gene expression noise with single-cell mRNA sequencing. (**a**) Hand-picked single mESCs were spiked with foreign RNA (left) and sequenced with a modified CEL-Seq[3] protocol (Online Methods). To measure technical noise, RNA aliquots from bulk cells were treated likewise (right). (**b**) CV computed on the basis of transcripts per million (TPM) versus reads per million (RPM). The diagonal (red solid line) and twofold change intervals (red dashed lines) are indicated. (**c**) CV across control samples as a function of average expression. Blue line indicates CV for a hypothetical Poissonian distribution; dashed line represents CV computed from the s.d. of β, i.e., for global tube-to-tube variability. (**d**) Count distribution of *Pou5f1* transcripts across cells and controls (dashed lines) fitted by negative binomials (solid lines). (**e**) Different functions were fitted to the count distribution in cells and controls. The goodness of fit was assessed by a χ^2 test. The bar plot shows the number of genes for which a given distribution was not rejected (χ^2 test $P > 0.01$).

genes, the dependence of the CV on the expression level was consistent with Poissonian sampling noise (**Fig. 1c**). Here, the s.d. of transcript number scales with the square root of the mean. With increasing expression, however, the CV starts to exceed sampling noise and ultimately approaches a constant value independent of the expression level (**Fig. 1c**). We reasoned that a constant CV arises from a noise component that implies a linear dependence of the s.d. of the transcript number on the mean. This type of noise could originate from global tube-to-tube variability in sequencing efficiency, which was quantified by the conversion factor β. The distribution of β across individual samples indicates an approximately twofold range of variability (**Supplementary Fig. 3c**). The high correlation ($R = 0.91$) between β and the total number of sequenced RNAs per sample (**Supplementary Fig. 3d**) confirmed that β reliably quantifies sequencing efficiency. From the s.d. of β, we calculated the CV component explained by

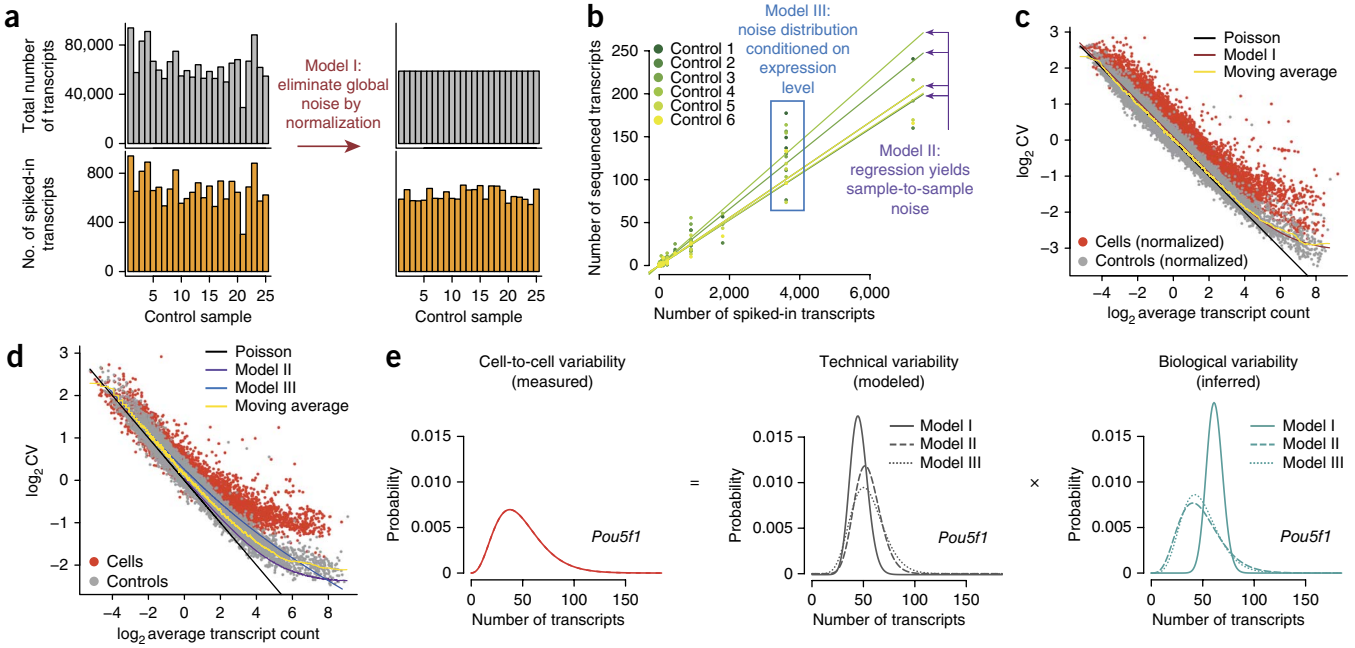

Figure 2 | Modeling of technical variability and inference of biological noise. (**a**) Schematic of transcript normalization for model I. (**b**) Linear regression of the transcript counts on the predicted spike-in molecule numbers. Different slopes reflect varying efficiencies between samples. (**c**,**d**) CV as a function of average expression in cells and controls with (**c**) and without (**d**) count normalization. Predictions are indicated for model I in **c** and models II and III in **d**. The moving average and the CV for Poissonian noise are also shown. (**e**) Schematic of biological noise inference. Cell-to-cell noise (red) was fitted by a negative binomial, and a deconvolution of technical noise (gray) yields biological variability (turquoise). Example distributions are shown for *Pou5f1*. Noise distributions are similar for models II and III but narrower for model I owing to the elimination of global variability by normalization.

© 2014 Nature America, Inc. All rights reserved.

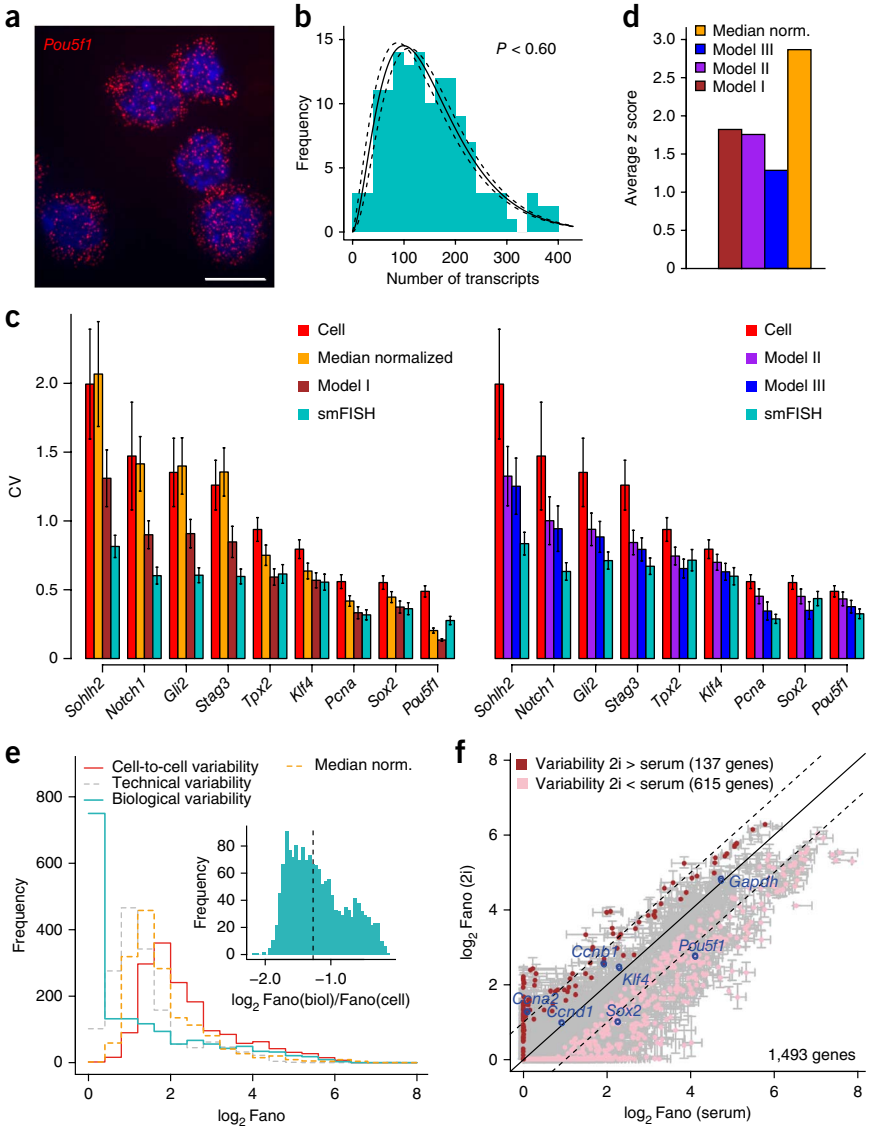

Figure 3 | Validation of predicted biological variability by smFISH. (**a**) *Pou5f1* transcripts labeled with Cy5 using smFISH in mESCs (maximal *z* projections). Single molecules appear as diffraction-limited spots. Nuclei were stained with DAPI. Scale bar, 10 µm. (**b**) Count distribution for smFISH on *Pou5f1* (>100 cells) and a negative binomial fit (black line) with uncertainty interval (dashed lines). The *P* value for rejecting a negative binomial was computed by a χ^2 test. (**c**) CV measured in cells, as inferred after deconvolution of technical noise and measured with smFISH. In the model I comparison, the CV after normalization of transcript counts without deconvolution of sampling noise is also shown. Error bars are derived from estimated standard errors of the numerical fits. (**d**) *z* score of deviations between models and smFISH-based CVs averaged across genes. The *z* score after normalization (Median norm.) of transcript counts without deconvolution of sampling noise is also shown. (**e**) Distribution of Fano factors as measured in cells and controls and as inferred for biological variability using model III. The distribution after normalization of transcript counts without deconvolution of sampling noise is also shown (Median norm.). The inset shows a histogram of $\log_2$ fold changes between Fano factors before and after deconvolution of technical noise. (**f**) Scatter plot of Fano factors in the serum versus 2i conditions. Genes that have different Fano factors within their error bars between the two conditions are colored. Error bars are based on standard errors of fitting parameters.

global tube-to-tube variability as $\Delta\beta/\beta$ (Online Methods), which coincided with the observed constant CV level for highly expressed transcripts (Online Methods and **Fig. 1c**).

We concluded that the dominant source of technical noise depends on expression level: for low-expression genes, it is sampling noise, and for high-expression genes, it is global tube-to-tube variability in sequencing efficiency.

In order to model the observed technical noise components, we investigated the distribution of transcript counts (**Fig. 1d**). We performed maximum-likelihood fits of various distributions and found that a negative binomial explained the distribution for the largest fraction of genes (**Fig. 1e** and **Supplementary Fig. 4**).

In a first model (model I), we eliminated tube-to-tube variability by normalizing counts in each sample to the cross-sample median (**Fig. 2a**) and subsequently inferred parameters of the expression distribution for the controls in order to model technical noise (Online Methods and **Supplementary Fig. 5**).

Models II and III are based on the raw-transcript count distributions. For these models, global tube-to-tube variability of sequencing efficiency was derived from the statistics of the sequenced spike-ins, which were used to calculate a model-specific β (Online Methods, **Fig. 2b** and **Supplementary Figs. 6** and **7**). A brief description of the models is provided in **Supplementary Note 2**.

Across the entire dynamic range of transcript expression, the model-derived CVs yielded a good approximation of the average gene-specific CV (**Fig. 2c**,**d**).

We quantitatively inferred true biological gene expression noise by deconvolving out technical variability from the transcript distributions measured in cells (**Fig. 2e**), which were also fit by negative binomials (**Fig. 1d**,**e** and **Supplementary Fig. 5a**). We assumed endogenous mRNA abundance to follow a negative binomial, which is supported by a physical model of bursting expression[19]. Inferred parameters of the biological distribution were well defined for robustly expressed genes (**Supplementary Fig. 8**), and we found a clear reduction of inferred biological versus measured cell-to-cell noise (**Fig. 2e** and **Supplementary Fig. 8**).

To validate our biological noise predictions, we selected four stem cell markers (*Pou5f1*, *Sox2*, *Klf4* and *Pcna*), a moderately expressed gene (*Tpx2*) and four genes with low expression (*Sohlh2*, *Notch1*, *Gli2* and *Stag3*). These genes cover most of the dynamic range of transcript expression (**Supplementary Fig. 9a**). We performed single-molecule FISH (smFISH) experiments on these genes in independently cultured cell populations, i.e., single mRNAs were labeled with a fluorescent dye and counted by

© 2014 Nature America, Inc. All rights reserved.

microscopic imaging[20] in >100 cells per gene (Online Methods and **Fig. 3a**,**b**). Transcript quantification by smFISH is highly sensitive and accurate[20]. We therefore compared sequencing-derived biological CVs directly to the CV computed for the smFISH data. For comparison with model I, transcript abundance obtained by smFISH was normalized to the cell area. The CVs predicted by all three models were overall in good agreement with the smFISH-derived CVs (**Fig. 3c**). At low expression, our predictions overestimated the biological noise measured by smFISH. However, for *Gli2* and *Stag3*, which were on average expressed at 12 transcripts per cell as determined from the smFISH data, the biological noise predicted by models II and III already overlapped with the smFISH measurement.

We note that model I, owing to the normalization, predicts concentration noise, whereas models II and III estimate noise of the actual transcript number. Notably, the normalization alone, without deconvolving sampling noise, overestimates biological noise (**Fig. 3c**).

A model comparison based on z scores for the deviation of smFISH- and sequencing-derived noise estimates indicates that model III performs best (**Fig. 3d**). We speculate that model III outperforms model II owing to individual fits of the conversion factor at distinct expression levels.

We used our smFISH data to analyze sensitivity of CEL-Seq. Mean expression correlated strongly between both methods (**Supplementary Fig. 9b**), but sensitivity of smFISH was eightfold higher. Assuming 100% sensitivity of smFISH yields an estimated CEL-Seq sensitivity of 12.5% (**Supplementary Fig. 9c**). In contrast, the observed spike-in counts suggest an efficiency of only 3.4%. This is presumably an underestimate, and a possible explanation could be RNA degradation due to the age of the spike-in batch or other technical reasons. More speculatively, cellular RNA could for unknown reasons be more protected during cell lysis than spike-in RNA. Because the actual sensitivity of smFISH has been shown to be >80% (ref. 20), we estimate that the true sensitivity of our CEL-Seq experiments is on the order of 10%, which corresponds to a mean of 500,000 transcripts per mESC.

Overall, the elimination of technical noise yields biological noise estimates substantially lower than noncorrected cell-to-cell noise (**Fig. 3e**). We note that the inferred biological Fano factor has to be divided by β to obtain the actual biological Fano factor to account for the limited sensitivity (**Supplementary Figs. 9d** and **10**).

We compared our noise predictions to a recently published method[16], which identifies genes with substantial biological noise without inferring the actual noise level. We observed that genes with biological noise clearly exceeding sampling noise according to our analysis, including the genes validated by smFISH, were not identified by this method (**Supplementary Fig. 11**).

Additionally, we demonstrated the validity of our approach for another sequencing technique based on PCR amplification of starting material[14] (**Supplementary Note 3** and **Supplementary Fig. 12**).

Finally, we applied our method to investigate gene expression noise in mESCs cultured in the 2i condition[17] in comparison to the traditional serum culture. In 2i medium, mESCs acquire a state of naive pluripotency with reduced heterogeneity in morphology and expression of pluripotency markers[17]. To test whether reduced gene expression variability in 2i medium affects a larger number of genes, we sequenced 44 cells and 56 serum culture controls (**Supplementary Fig. 13a**) and validated our CV predictions by smFISH (**Supplementary Fig. 13b**).

After discarding genes with fewer than five transcripts per cell on average, we found that Fano factors are on average 1.4-fold higher for serum culture (**Supplementary Fig. 14a**). Out of 1,493 genes above our expression cutoff in both culture conditions, 615 genes were more variable, whereas only 137 genes were less variable in the serum condition (**Fig. 3f**).

Notably, genes correlating with the cell cycle (*Ccna2*, *Ccnb1* and *Ccnd1*), or housekeeping genes, such as *Gapdh*, were not more variable in general (**Fig. 3f**). Increased variability in the serum versus 2i condition was also confirmed by a CV-based comparison (**Supplementary Fig. 14b**) and supported by smFISH for *Pou5f1*, *Sox2* and *Pcna* (**Supplementary Figs. 14c** and **15**). An extended analysis of the differential expression variability in the two conditions is presented in **Supplementary Note 4** and **Supplementary Table 2**.

METHODS

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

Accession codes. Gene Expression Omnibus: RNA-seq data are deposited under accession number GSE54695.

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

ACKNOWLEDGMENTS

This work was supported by a European Research Council Advanced grant (ERC-AdG 294325-GeneNoiseControl) and a Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) Vici award.

AUTHOR CONTRIBUTIONS

D.G., L.K. and A.v.O. conceived the methods. D.G. developed the noise models, performed all computations and wrote the manuscript. L.K. performed all experiments and corrected the manuscript. A.v.O. guided experiments, data analysis and writing of the manuscript, and corrected the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Munsky, B., Neuert, G. & van Oudenaarden, A. *Science* **336**, 183–187 (2012).
2. Eldar, A. & Elowitz, M.B. *Nature* **467**, 167–173 (2010).
3. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. *Cell Rep.* **2**, 666–673 (2012).
4. Sasagawa, Y. *et al. Genome Biol.* **14**, R31 (2013).
5. Tang, F. *et al. Nat. Methods* **6**, 377–382 (2009).
6. Ramsköld, D. *et al. Nat. Biotechnol.* **30**, 777–782 (2012).
7. Islam, S. *et al. Genome Res.* **21**, 1160–1167 (2011).
8. Picelli, S. *et al. Nat. Methods* **10**, 1096–1098 (2013).
9. Shapiro, E., Biezuner, T. & Linnarsson, S. *Nat. Rev. Genet.* **14**, 618–630 (2013).
10. Kivioja, T. *et al. Nat. Methods* **9**, 72–74 (2012).
11. Shiroguchi, K., Jia, T.Z., Sims, P.A. & Xie, X.S. *Proc. Natl. Acad. Sci. USA* **109**, 1347–1352 (2012).
12. Hug, H. & Schuler, R. *J. Theor. Biol.* **221**, 615–624 (2003).
13. Shalek, A.K. *et al. Nature* **498**, 236–240 (2013).
14. Islam, S. *et al. Nat. Methods* **11**, 163–166 (2014).
15. Jaitin, D.A. *et al. Science* **343**, 776–779 (2014).
16. Brennecke, P. *et al. Nat. Methods* **10**, 1093–1095 (2013).
17. Ying, Q.-L. *et al. Nature* **453**, 519–523 (2008).
18. The External RNA Controls Consortium. *Nat. Methods* **2**, 731–734 (2005).
19. Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y. & Tyagi, S. *PLoS Biol.* **4**, e309 (2006).
20. Raj, A., van den Bogaard, P., Rifkin, S.A., van Oudenaarden, A. & Tyagi, S. *Nat. Methods* **5**, 877–879 (2008).

© 2014 Nature America, Inc. All rights reserved.

ONLINE METHODS

Cell culture. Serum medium consists of DMEM (Gibco) supplemented with 15% FCS (Gibco), 2 mM GlutaMAX (Gibco), 0.1 mM MEM non-essential amino acids, 0.1 mM β-mercaptoethanol (Sigma), 1% penicillin/streptomycin (Gibco) and 1,000 U LIF/ml (ESGRO). J1 embryonic stem cells were obtained from the Koch institute for Integrative Cancer Research at MIT and were tested mycoplasma free. J1 cells have not recently been authenticated. J1 cells cultured in serum medium were maintained on a monolayer of MEFs. 2i medium consists of 50% DMEM/F12 and 50% N2B27 supplemented with 2 mM GlutaMAX, 0.1 mM MEM non-essential amino acids, 0.1 mM β-mercaptoethanol, 1% pen/strep, 1,000 U LIF/ml, 1 μM PD0325901 (Stemgent) and 3 μM CHIR99021 (Stemgent). J1 embryonic stem cells cultured in 2i medium were maintained on culture dishes coated with gelatin. J1 cells cultured in serum medium were passaged every 2–3 d by dissociation with trypsin (Gibco). J1 cells cultured in 2i medium were passaged every 2–3 d by dissociation with Accutase (Gibco).

Cell picking. Cells were dissociated into a single-cell suspension and picked under a stereomicroscope using a 30-μm glass capillary and mouth pipette. Picked cells were deposited in the lid of a 0.5-ml LoBind Eppendorf tube and snap frozen in liquid nitrogen. For the pool-and-split controls, approximately 1 million cells were lysed, the amount of RNA was quantified on a bioanalyzer (Agilent) using the Eukaryote Total RNA Pico kit. 20-pg aliquots of total RNA were used for each pool-and-split control.

CEL-Seq library preparation. Single cells were processed using the previously described CEL-Seq technique[3], with a few alterations. A 4-bp random barcode as unique molecular identifier (UMI) was added to the primer in between the cell-specific barcode and the poly(T) stretch. Instead of the 70 °C lysis step, cells were lysed by adding 0.05% IGEPAL CA 630 (Sigma) to the first-strand synthesis mix. Libraries were sequenced on an Illumina Hi-Seq 2500 using 50-bp paired-end sequencing.

Single-molecule FISH. Probe libraries were designed and fluorescently labeled as previously described[20]. All probe libraries consist of 48 oligonucleotides of 20-bp length (see **Supplementary Table 3** for probe sequences) complementary to the coding sequence of the genes. Cells were hybridized overnight with probes at 30° C, as previously described[20]. 4′,6-Diamidino-2-phenylindole (DAPI) was added during washes. Images were acquired on a PerkinElmer spinning disc confocal microscope with a 100× oil-immersion objective (numerical aperture, 1.4) using PerkinElmer Volocity software. Images were recorded as stacks with a *z* spacing of 0.3 μm. Diffraction-limited dots corresponding to single mRNA molecules were automatically detected using custom Matlab software based on previously described algorithms[20]. Briefly, the images were first filtered using a three-dimensional Laplacian-of-Gaussian filter, followed by selection of the intensity threshold at which the number of connected components was least sensitive to the threshold.

Quantification of transcript abundance. Paired-end reads obtained by CEL-Seq were aligned to the transcriptome using BWA[21] with default parameters. The transcriptome contained all RefSeq gene models based on the mouse genome release mm10 downloaded from the UCSC genome browser[22] and contained 31,109 isoforms derived from 23,480 gene loci. The right mate of each read pair was mapped to the ensemble of all RefSeq transcripts and to the set of 92 ERCC spike-ins[18] in the sense direction. Reads mapping to multiple loci were distributed uniformly. For each cell barcode, we counted the number of UMIs for every transcript and aggregated this number across all transcripts derived from the same gene locus. On the basis of binomial statistics (see below), we converted the number of observed UMIs into transcript counts. We discarded all genes that were not expressed with at least a single transcript in at least two cells and two control samples. For mESCs in 2i conditions, we retained 11,555 genes above this expression threshold; and for mESCs cultured in serum, 11,701 genes were expressed above this minimum level. Samples with fewer than ten sequenced *Pou5f1* transcripts were discarded because sequencing efficiency was low or the cells are potentially undergoing differentiation.

Conversion of UMI count to transcript number. For each gene i, $k_{o,i}$ denotes the number of observed UMIs and $k_{n,i}$ the number of non-observed UMIs. The total number K of UMIs is given by

$$K = k_{o,i} + k_{n,i} \tag{1}$$

The probability of not observing $k_{n,i}$ UMIs for a gene with m_i copies is given by

$$\left(1 - \frac{1}{K}\right)^{m_i} = \frac{k_{n,i}}{K} \tag{2}$$

which can be solved for the number of sequenced transcripts m_i

$$m_i = \frac{\ln\left(1 - \frac{k_{o,i}}{K}\right)}{\ln\left(1 - \frac{1}{K}\right)} \cong -K \ln\left(1 - \frac{k_{o,i}}{K}\right) \tag{3}$$

CV component from global tube-to-tube variability. The contribution of tube-to-tube variability to the CV can be computed from the statistics of sequencing efficiencies across tubes. These efficiencies are quantified by the conversion factor β, which was obtained for each sample by a regression of the number of sequenced spike-in transcripts on the number of spike-in molecules added to the sample. Given the s.d. of β ($\Delta\beta$) and the number of transcripts for gene i (n_i), the CV caused by tube-to-tube variability, or efficiency noise, can be computed with an s.d. σ_i derived from the number of sequenced transcripts and $\Delta\beta$

$$\mathrm{CV}_i^{\mathrm{eff}} = \frac{\sigma_i}{n_i} = \frac{\frac{\Delta\beta}{\beta} n_i}{n_i} = \frac{\Delta\beta}{\beta} \tag{4}$$

Three models for the deconvolution of technical noise. We developed three models for technical noise, which were inferred from transcript count distributions measured by CEL-Seq for spike-in RNAs or pool-and-split control samples. All models are based on a negative binomial distribution fitted to the count

© 2014 Nature America, Inc. All rights reserved.

histogram (**Fig. 1d,e** and **Supplementary Fig. 5a**). A negative binomial

$$\mathrm{NB}(n;\mu,r)=\left(\frac{r}{r+\mu}\right)^{r}\frac{\Gamma(r+n)}{n!\Gamma(r)}\left(\frac{\mu}{r+\mu}\right)^{n} \quad (5)$$

is governed by two parameters, the average μ and the dispersion parameter r. A negative binomial is frequently used to model over-dispersed count data. For instance, it has been used to describe gene expression variability across replicates of RNA-seq on bulk cells and to infer differential expression[23,24].

The dependence of the variance of a negative binomial on the mean is controlled by r

$$\sigma^2=\mu+\frac{1}{r}\mu^2 \quad (6)$$

and converges to a Poisson distribution for large values of r. At low values of r, the noise inferred from a negative binomial exceeds random sampling, i.e., Poissonian noise.

From a phenomenological perspective, two main sources of technical noise have to be taken into account. First, individual transcripts are sampled from a pool of available transcripts for CEL-Seq. This noise component obeys Poissonian statistics, and thus the CV is inversely proportional to the square root of the mean. Second, we observed variability in the total number of sequenced transcripts (**Supplementary Fig. 3c**), which we term efficiency noise. The s.d. explained by this noise component (**Supplementary Fig. 3c**) scales linearly with mean expression and therefore yields a constant CV (**Fig. 1c**). Whereas the sampling noise dominates at low expression, the efficiency noise is dominant for highly expressed genes. This crossover appears as a bend of the CV as a function of μ in logarithmic space (**Fig. 1c**). The dispersion parameter r will thus be a function of μ, and the goal of all three models is the derivation of this dependence in order to characterize technical noise for arbitrary expression levels.

Model I. In the first model the impact of efficiency noise is eliminated, to a certain degree, by normalizing the total transcript count in each control to the median transcript number across controls (**Fig. 2a**). Negative binomials are subsequently fitted to the normalized count distributions. The dispersion parameter of these fits is found to display a piecewise linear dependence on the mean in log space (**Supplementary Fig. 5b**).

$$\log_2(r)=a+b\log_2(\mu) \quad (7)$$

Separate linear fits were performed for $\log_2(\mu)<4$ and $\log_2(\mu)>4$, and the piecewise dependences inferred from these fits were used to define r. The slopes of 0.95 and −0.03, for low and high values of μ, respectively, correspond to the Poissonian and the residual efficiency noise regime (the residual noise not eliminated by the normalization). Compared to non-normalized expression, the Poissonian regime expands to much higher expression (compare **Fig. 2c** and **Fig. 2d**).

Model II. Model II infers efficiency noise from the variability of the conversion factor β_{II}, which was obtained for each sample by a linear regression of the number of sequenced spike-in transcripts on the number of added spike-in molecules predicted on the basis of the spike-in concentration (**Fig. 2b** and **Supplementary Figs. 3c** and **6a**). The distribution of β_{II} was fitted by a Γ distribution (**Supplementary Fig. 6b**). The number of transcripts of gene i available for sequencing, λ_i, can now be written as

$$\lambda_i=\beta_{\mathrm{II}}n_i \quad (8)$$

n_i is the number of transcripts in a control sample or of a given spike-in and thus follows a Poisson distribution. The distribution of λ_i is therefore given as a product of a Γ distribution and a Poisson distribution. We confirmed by simulation across a wide range of parameters that the product distribution again corresponds to a Γ distribution. Analytical integration of the product distribution was not possible. Therefore, the shape parameter a and the rate parameter b of the product distribution were simulated for transcript counts ranging from 0 to 100,000.

The number of sequenced transcripts then follows a Poisson distribution

$$P(m_i)=\frac{\lambda_i^{m_i}e^{-\lambda_i}}{m_i!} \quad (9)$$

with rate λ_i, which is a Γ-distributed random variable

$$\lambda_i\sim\Gamma(a_i,b_i) \quad (10)$$

A Poisson distribution with a Γ-distributed rate again yields a negative binomial

$$P(m_i)\sim\mathrm{NB}(\mu_i=a_i/b_i;r_i=a_i) \quad (11)$$

which was used to describe technical noise within model II.

Model III. In model III, efficiency noise was also inferred from the conversion factor β. This time, however, β_{III} was calculated for each spike-in in each sample as the number of sequenced spike-in transcripts divided by the added number of spike-in transcripts (**Fig. 2b**), and a Γ distribution was fitted to the distribution of β_{III} across samples for each spike-in species (**Supplementary Fig. 7a,b**). The random variables λ_i drawn from these distributions multiplied by the mean expression level of the corresponding spike-in

$$\lambda_i=\beta_{\mathrm{III}}(\overline{n_i})\overline{n_i} \quad (12)$$

represent the number of transcripts in the pool available for sequencing. If the distribution of β_{III} is governed by the parameters a and b,

$$\beta_{\mathrm{III}}(\overline{n_i})\sim\Gamma(a(\overline{n_i}),b(\overline{n_i})) \quad (13)$$

then λ_i follows a Γ distribution with rescaled parameters,

$$\lambda_i\sim\Gamma(a(\overline{n_i}),b(\overline{n_i})/\overline{n_i}) \quad (14)$$

and the number of sequenced transcript again obeys a Poisson distribution with rate λ_i, which is identical to a negative binomial

$$P(m_i)\sim\mathrm{NB}(\mu=\overline{n_i}\,a(\overline{n_i})/b(\overline{n_i});r=a(\overline{n_i})) \quad (15)$$

© 2014 Nature America, Inc. All rights reserved.

doi:10.1038/nmeth.2930

We performed a linear regression of the shape and the rate parameter on the mean expression level in log space (**Supplementary Fig. 7c,d**),

$$\log_2(a(\overline{n_i})) = k_a + f_a\overline{n_i} \tag{16}$$

and

$$\log_2(b(\overline{n_i})) = k_b + f_b\overline{n_i} \tag{17}$$

which allowed us to derive an analytical formula for the dependence of the dispersion parameter of the negative binomial on the mean

$$r = 2^{k_a + f_a\frac{k_b - k_a}{1 + f_a - f_b}}\mu^{\frac{f_a}{1 + f_a - f_b}} \tag{18}$$

A negative binomial for technical noise was defined based on equation (18) for arbitrary expression levels.

Deconvolution of technical and biological noise. In general, the expression noise $P_{\text{cell}}(n)$ measured for an arbitrary gene among single cells by any method that suffers from technical noise can be written as a convolution of biological cell-to-cell expression noise $P_{\text{biol}}(n)$ and technical noise $Q_{\text{ctr}}(\Delta n)$

$$P_{\text{cell}}(n) = \sum_{m \geq 0} Q_{\text{ctr}}(n - m)P_{\text{biol}}(m) \tag{19}$$

With the negative binomial obtained from a fit to $P_{\text{cell}}(m)$ and the technical noise distributions inferred by the three models, the convolution can be expressed in terms of negative binomials for each gene i

$$\text{NB}(n; \mu_i^{\text{cell}}, r_i^{\text{cell}}) = \sum_{m \geq 0} \text{NB}(n; m, r(m))\text{NB}(m; \mu_i^{\text{biol}}, r_i^{\text{biol}}) \tag{20}$$

with $r(m)$ derived by the noise models. As explained in the main text, we also assume a negative binomial distribution for the biological noise, and its parameters μ_i^{biol} and r_i^{biol} can be inferred by minimizing the absolute difference of the two sides of equation (20)

$$\text{argmin}_{\mu_i^{\text{biol}}, r_i^{\text{biol}}} \left| \text{NB}(n; \mu_i^{\text{cell}}, r_i^{\text{cell}}) - \sum_{m \geq 0} \text{NB}(n; m, r(m))\text{NB}(m; \mu_i^{\text{biol}}, r_i^{\text{biol}}) \right| \tag{21}$$

For the numerical optimization we used a quasi-Newton method with box-constraints[25] implemented in the R "optim" function.

A physical model of bursting transcription. In a general physical model of gene expression, the promoter of a gene can be assumed to be in either an active state (A) or inactive state (I). Switching from the active to the inactive state and vice versa occurs at rate k_{off} and k_{on}, respectively. In the active state, a gene is transcribed at rate k_{T} and the transcript m decays at rate γ.

The model can be summarized by the following set of dynamic equations:

$$\begin{aligned} \text{A} &\xrightarrow{k_{\text{off}}} \text{I} \\ \text{I} &\xrightarrow{k_{\text{on}}} \text{A} \\ \text{A} &\xrightarrow{k_{\text{T}}} m \\ m &\xrightarrow{\gamma} \varnothing \end{aligned} \tag{22}$$

The model was previously described and solved analytically[19]. For $k_{\text{off}} \gg \gamma$, the solution is given by

$$\rho(m) = \left(1 + \frac{k_{\text{T}}}{k_{\text{off}}}\right)^{-\frac{k_{\text{on}}}{\gamma}} \frac{\Gamma\left(\frac{k_{\text{on}}}{\gamma} + m\right)}{\Gamma\left(\frac{k_{\text{on}}}{\gamma}\right)\Gamma(m+1)}\left(\frac{\frac{k_{\text{T}}}{k_{\text{off}}}}{1 + \frac{k_{\text{T}}}{k_{\text{off}}}}\right)^m \tag{23}$$

which corresponds to a negative binomial distribution with mean

$$\mu = \frac{k_{\text{T}}}{\gamma} \times \frac{k_{\text{on}}}{k_{\text{off}}} \tag{24}$$

and dispersion parameter

$$r = \frac{k_{\text{on}}}{\gamma} \tag{25}$$

21. Li, H. & Durbin, R. *Bioinformatics* **26**, 589–595 (2010).
22. Meyer, L.R. *et al. Nucleic Acids Res.* **41**, D64–D69 (2013).
23. Anders, S. & Huber, W. *Genome Biol.* **11**, R106 (2010).
24. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. *Bioinformatics* **26**, 139–140 (2010).
25. Byrd, R.H., Lu, P., Nocedal, J. & Zhu, C. *SIAM J. Sci. Comput.* **16**, 1190–1208 (1995).

© 2014 Nature America, Inc. All rights reserved.