Supplementary information

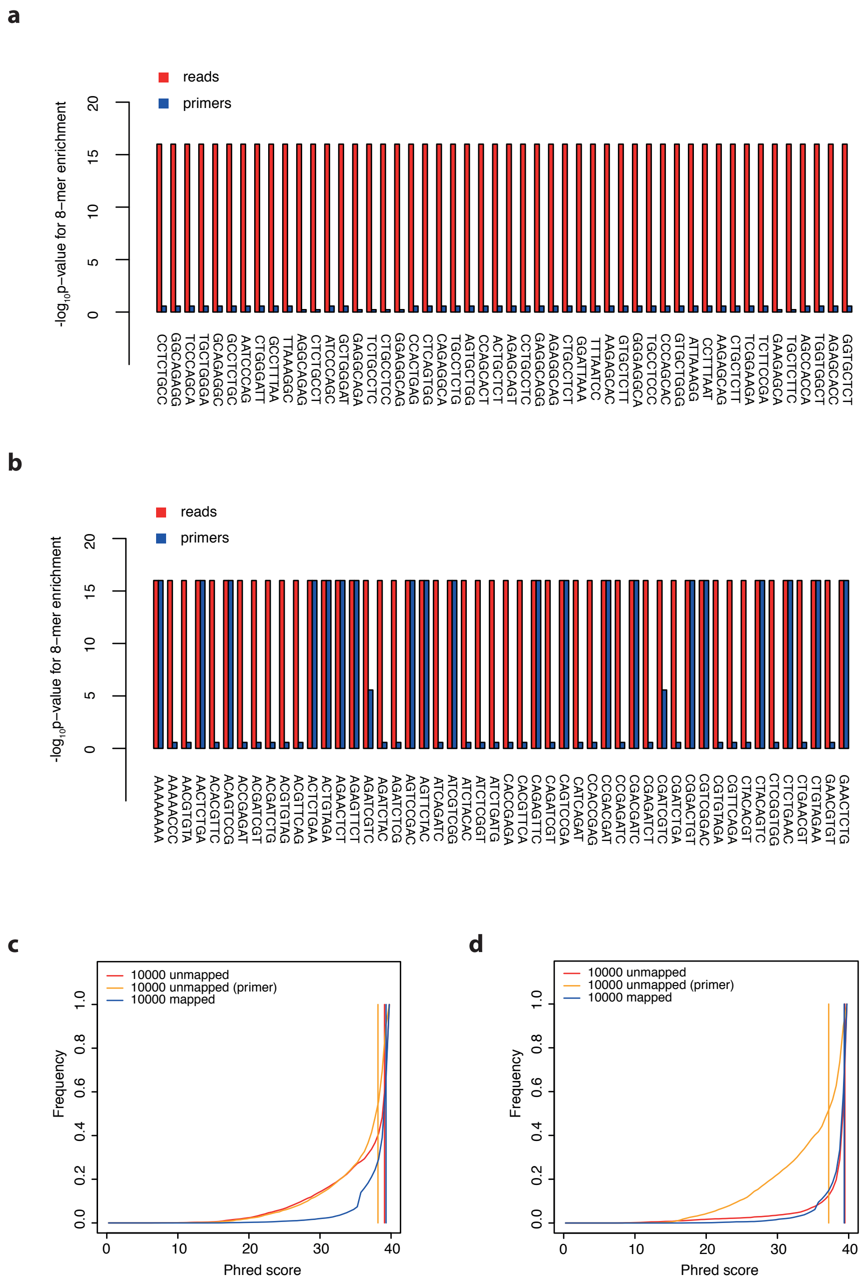
for

Validation of noise models for single-cell transcriptomics

Dominic Grün, Lennart Kester & Alexander van Oudenaarden

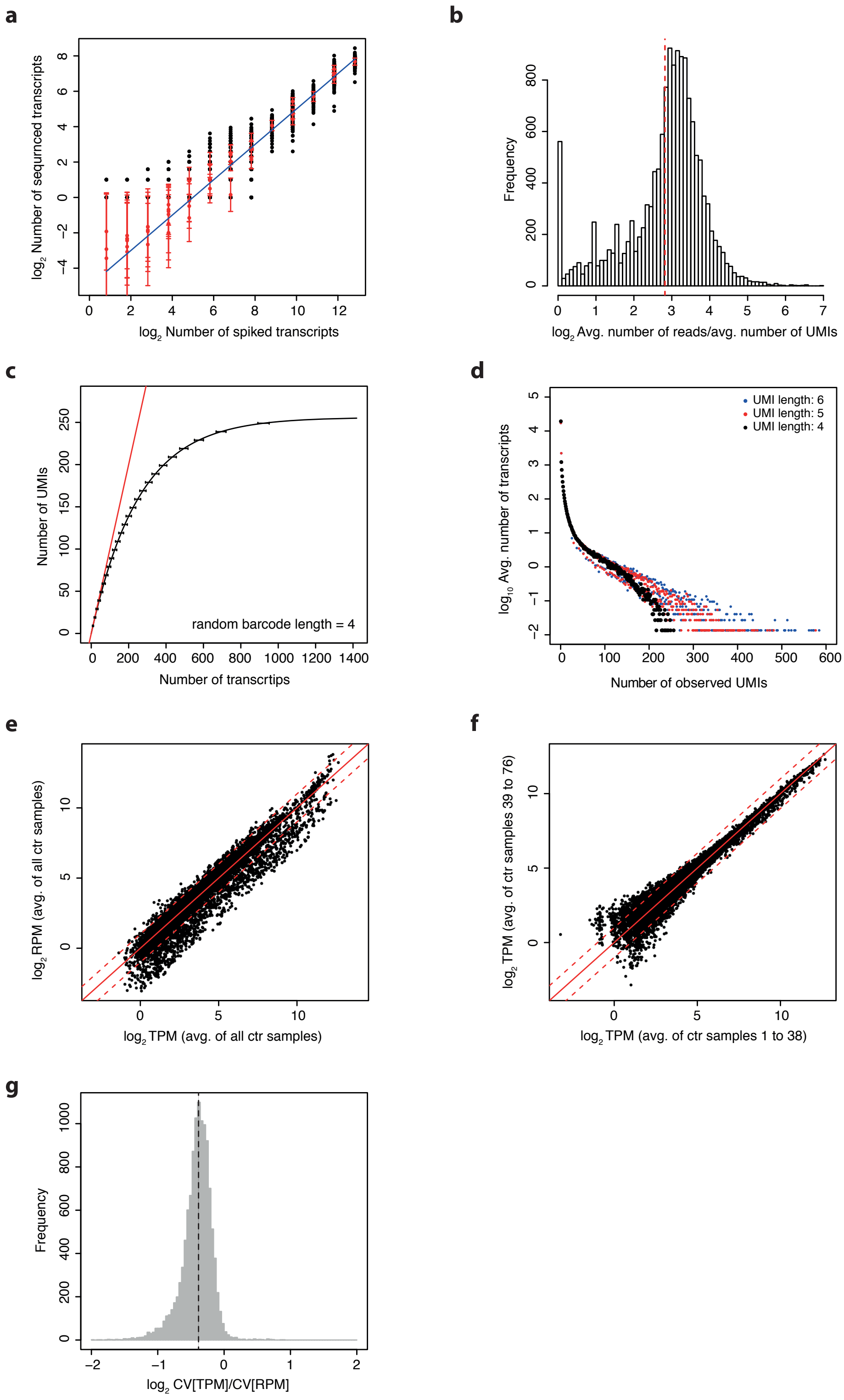
Nature Methods: doi:10.1038/nmeth.2930

Supplementary Figure 1



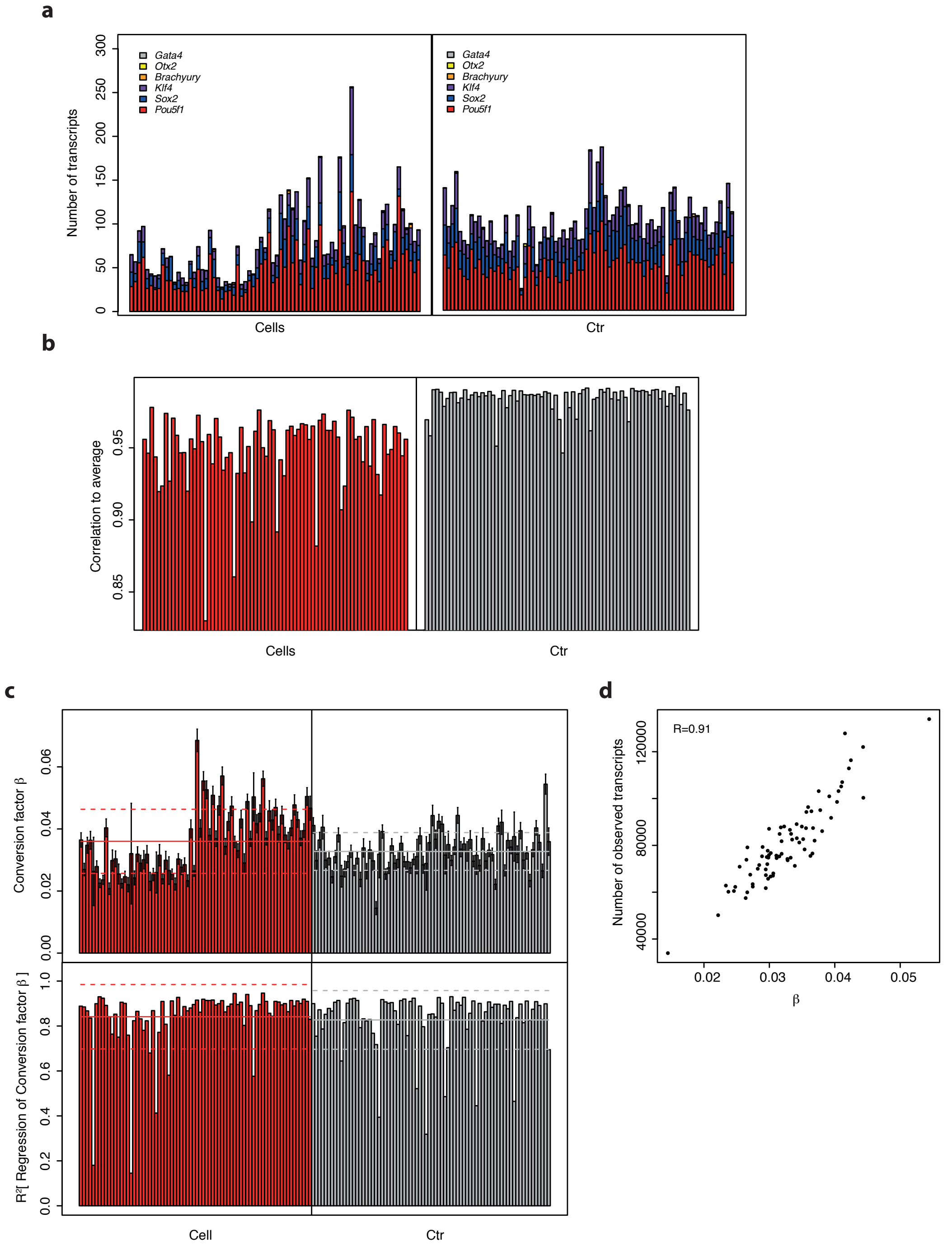
Analysis of primary sequencing data. (**a**) For the top fifty most highly enriched motifs in 10,000 randomly drawn mapped reads the hypergeometric enrichment p-value is shown (red). For comparison the enrichment p-value of the same motifs in the primer sequences is shown. (**b**) Same as (a), but for 10,000 randomly drawn non-mapped reads. A comparison of (a) and (b) reveals a strong overlap of motifs enriched in non-mapped reads and primer sequences. In (a) and (b) data for cells are shown. Results for controls are highly similar. (**c**) Cumulative distribution of average per reads Phred quality score for mapped reads, non-mapped reads and non-mapped reads containing primer derived 8-mers in cell samples. (d) Same as (c), but for control samples.

Supplementary Figure 2



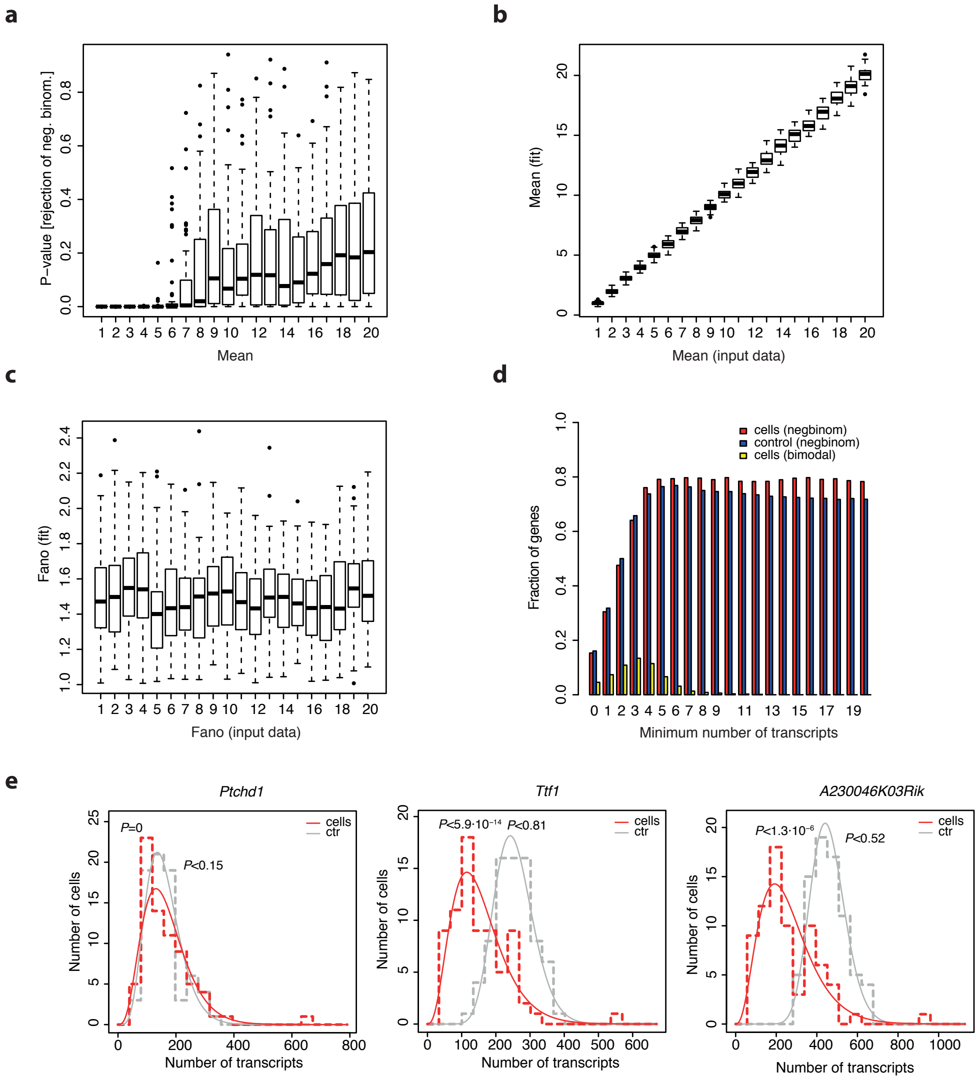
Unique molecular identifiers (UMIs) allow quantification of absolute transcript number. (**a**) Number of sequenced spike-in RNAs in individual samples as a function of the estimated number of molecules based on spike-in concentration. The intercept represents the conversion factor β and corresponds to the fraction of RNA recovered by sequencing (3.3% for controls and 3.6% for cells). The red dots and error bars indicate the mean and standard deviation, respectively, of each spike-in across samples. (**b**) Histogram of $\log_2$-ratio between number of reads and number of UMIs per gene. Each UMI was sequenced on average seven times (see Supplementary Table 1). Transcripts were thus over-sequenced sevenfold. (**c**) Predicted number of transcripts as a function of the number of observed UMIs (Methods). Genes, sequenced with more than 500 transcripts can still be reliably quantified using UMIs of length four. Since in our hands CEL-seq has a sensitivity of < 10%, genes expressed at several thousand copies can still be quantifies with UMIs of length 4. Error bars for number of UMIs at a given expression level were derived based on random counting statistics. A direct translation of UMIs into transcript counts (red line) only works at low expression. (**d**) Average number of transcripts across cells with a given number of observed UMIs. Less than 10% of the cells contain few transcripts with >200 sequenced UMIs. A 4 bp random barcode as UMI is thus sufficient. Simulated distributions of the number of UMIs for the derived transcript abundance are displayed for a UMI length of 5 and 6 bases. The distributions become distinguishable at ~150 UMIs per gene. Although a saturation of the 4bp UMI distribution is apparent, transcript numbers can be predicted with confidence even at 200 UMIs per gene (see error bars in (c)). (**e**) Comparison of mean expression estimates across all cells based on transcript counts (TPM), versus read counts (RPM). Differences greater than twofold can be observed across the entire dynamic range, suggesting that many genes are affected by PCR amplification bias. (**f**) Cells were split into two groups and average expression in TPM was compared between these two groups. In contrast to (e) strong differences only occur at low expression due to sampling noise. (**g**) Histogram of $\log_2$-fold change in the coefficient of variation (CV) computed based on reads per million (RPM) versus transcripts per million (TPM).

Supplementary Figure 3



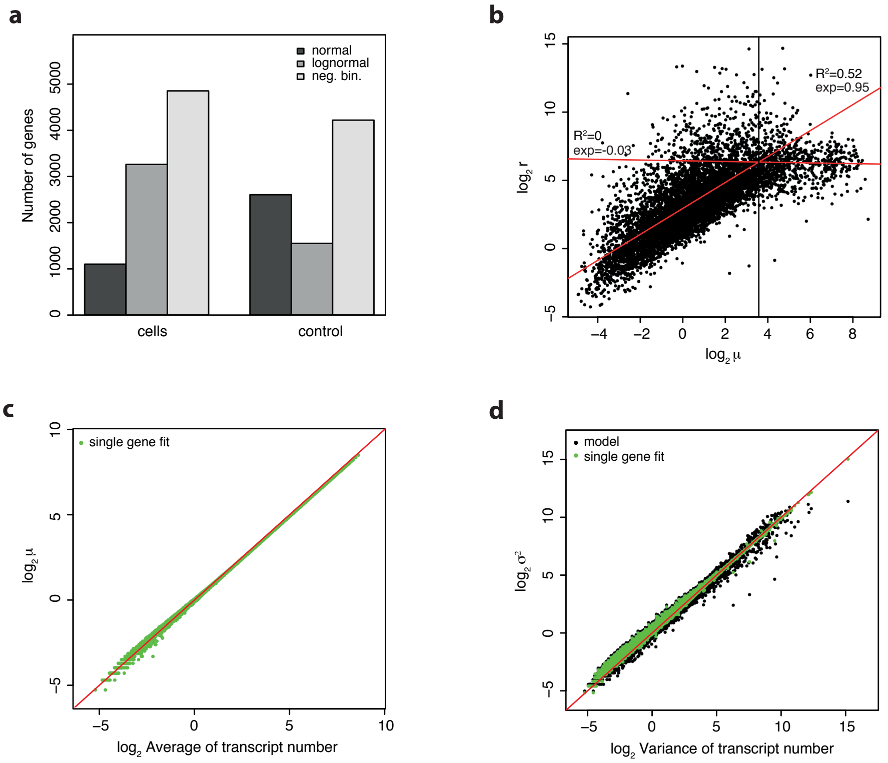
Gene expression in single mESCs is highly correlated. (**a**) Number of sequenced transcripts for pluripotency markers *Pou5f1*, *Sox2*, and *Klf4* as well as differentiation markers *Brachyury* (mesoderm), *Otx2* (ectoderm), and *Gata4* (endoderm) in cells and control samples. In contrast to robustly expressed plutipotency markers, differentiation markers are only very lowly expressed in all samples. Sequencing for cells and controls was performed in two batches each, due to a limited number of sample barcodes. The second batch of cells yielded on average more transcripts than the first. (**b**) Correlation between the number of sequenced transcripts in single cells or controls and the average across all cell or control samples, respectively. Correlation is high for all cells but on average lower and more variable than for control samples. (**c**) Spike-in RNA (1 µl of ERCC92 group 1 spike-in mix[1] diluted to 1:2,500,000 was added to each sample) was used to estimate the number of transcripts per cell. A regression of sequenced spike-in transcripts on the predicted number of spike-in molecules yields the conversion factor β which reflects the sequenced fraction of RNA with an average of 0.033 in controls and 0.036 in cells. The conversion factor is shown in the upper panel. Error bars correspond to the standard error of the regression parameters. R^2 between sequenced and predicted number of spike-ins is shown in the lower panel. Mean and standard deviation are indicated as solid and broken black lines, respectively. (**d**) Scatter plot of the total number of sequenced transcripts in control samples versus the conversion factor β. The correlation (R = 0.91) suggests that variability of β reflects variable efficiencies for different tubes.

Supplementary Figure 4



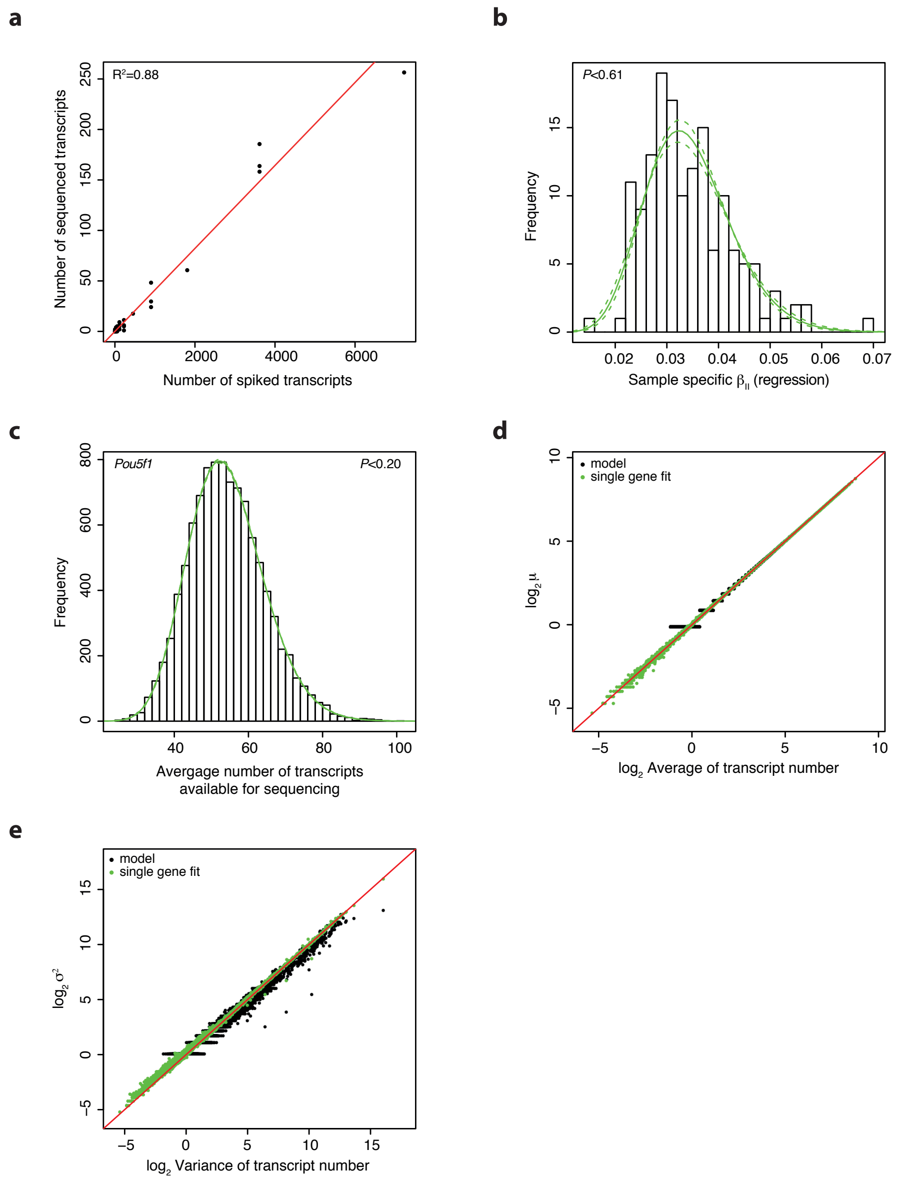
Data fitting confidence depends on the mean and is similar in cells and controls. We conducted simulations to investigate the dependence of the p-value obtained by a negative binomial fit to a set of random numbers sampled from a negative binomial distribution. To reflect the properties of our data we chose a size of the data set equal to the number of cells in 2i conditions. To perform the simulation we generated random numbers from distributions with varying mean and a size parameter that scales with the mean. To these random numbers we then fitted a negative binomial. At each mean value we sampled 50 times. (**a**) Distribution of p-values for rejecting a negative binomial as a function of the mean. For mean values of 5 or lower the fits are almost always rejected. (**b**) The mean obtained by the fit as a function of the mean of the input distribution. (**c**) The Fano factor obtained from the fit as a function of the mean of the input distribution. Even at low mean values the fitting procedure yields overall good estimates of the true parameters. In (a-c) the size parameter equals two times the mean, yielding a fixed dispersion of 1.5, independent of the mean. (**d**) Histogram of the fraction of genes for which the negative binomial fit is not rejected ($P > 0.01$) in cells (red) and control samples (blue) as a function of minimum mean expression. The fraction of genes with confident fits saturates at a minimum average of five transcripts per cell. About 80% of the genes follow a negative binomial. Importantly, this number is similar in cells and control samples, suggesting that the reason for rejecting a negative binomial is not in general increased cell-to-cell variability, due to, for instance, bimodality. We screened the remaining cases manually and found that in all cases a single or a few tail events caused the fit to fail. We checked in more detail for bimodal distributions and identifed genes, which display bimodal distributions in the cells but not in the controls using Hartigan’s test for bimodality ($P < 0.01$). The fraction of bimodal genes (yellow) is always low and converges against zero for highly expressed genes. (**e**) Three examples of highly expressed genes for which the negative binomial fit is rejected in cells but not in the control samples. These examples are representative for the observation that the most frequent reason for a fit rejection is the presence of a single or few tail events. Similar cases are also found among the control samples. The *P*-value for a χ^2-test for rejection of a negative binomial fit is given for the distribution in cells and controls.

Supplementary Figure 5



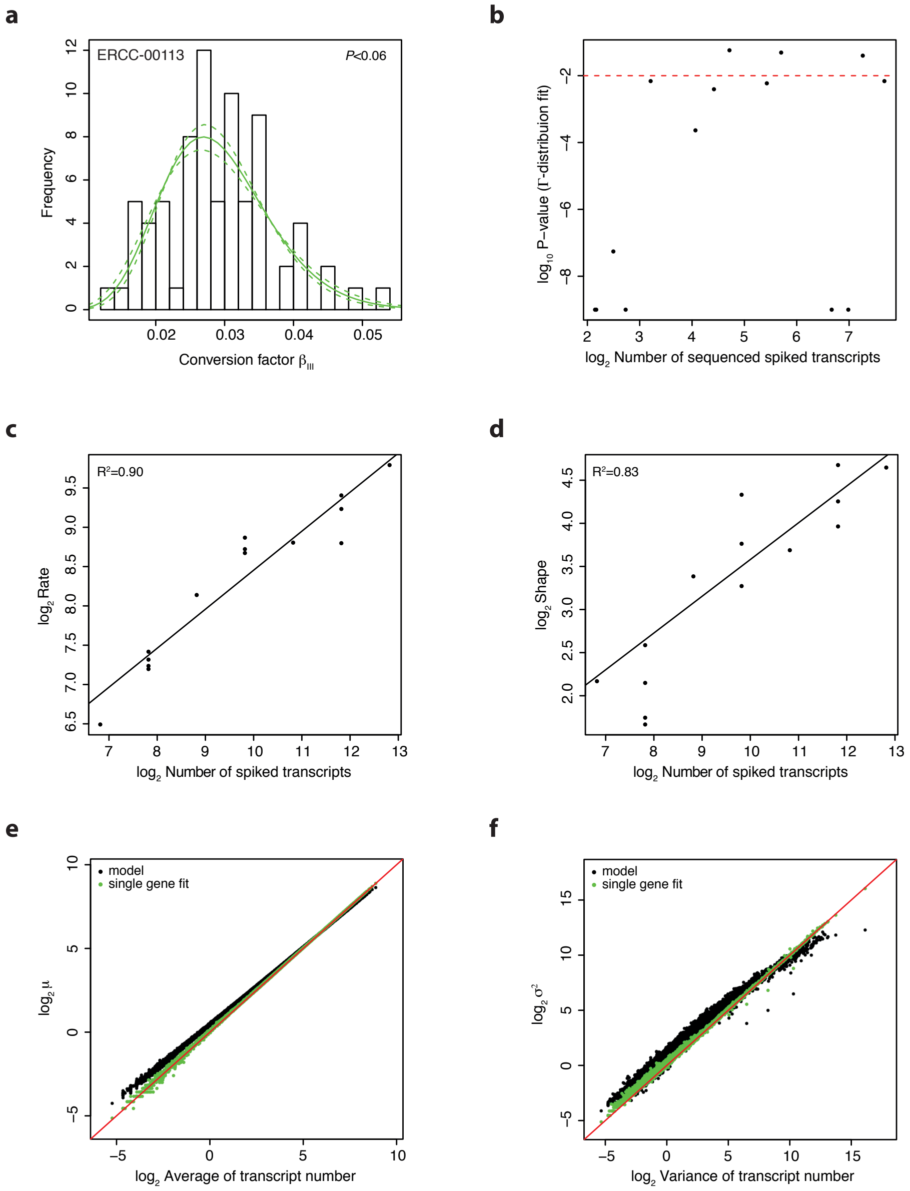
Modelling technical variability: Model I. (**a**) The transcript count in control samples was normalized to the cross-sample median. Different functions (normal distribution, lognormal distribution, and negative binomial distribution) were fitted to the normalized count distribution in cells and controls. The goodness of fit was assessed by a χ^2-test. The barplot shows the number of genes for which a given distribution was not rejected (χ^2-test $P > 0.01$). For 1,567 out of 1,992 genes (79%) expressed at >5 transcripts a negative binomial fit was not rejected ($P > 0.01$). (**b**) For negative binomials fitted to the normalized count distribution across control samples, the dispersion parameter r is plotted against the mean μ. The parameter dependence is fitted by a piecewise linear function in logarithmic space (red lines). The piecewise linear dependence of the dispersion parameter on the mean was used as parameter input for the technical noise distribution. The exponent for the parameter dependence and R^2 of the fit are given in the plot. (**c**) The average expression μ inferred from negative binomial fits is plotted against the mean directly computed from normalized counts across control samples. The red line marks the diagonal. (**d**) The expression variance σ^2 inferred from negative binomial fits using the dispersion parameter obtained from single gene fits (black) or as computed from the piecewise linear dependence on μ (green) is plotted against the variances directly computed from normalized counts across control samples. The red line marks the diagonal.

Supplementary Figure 6



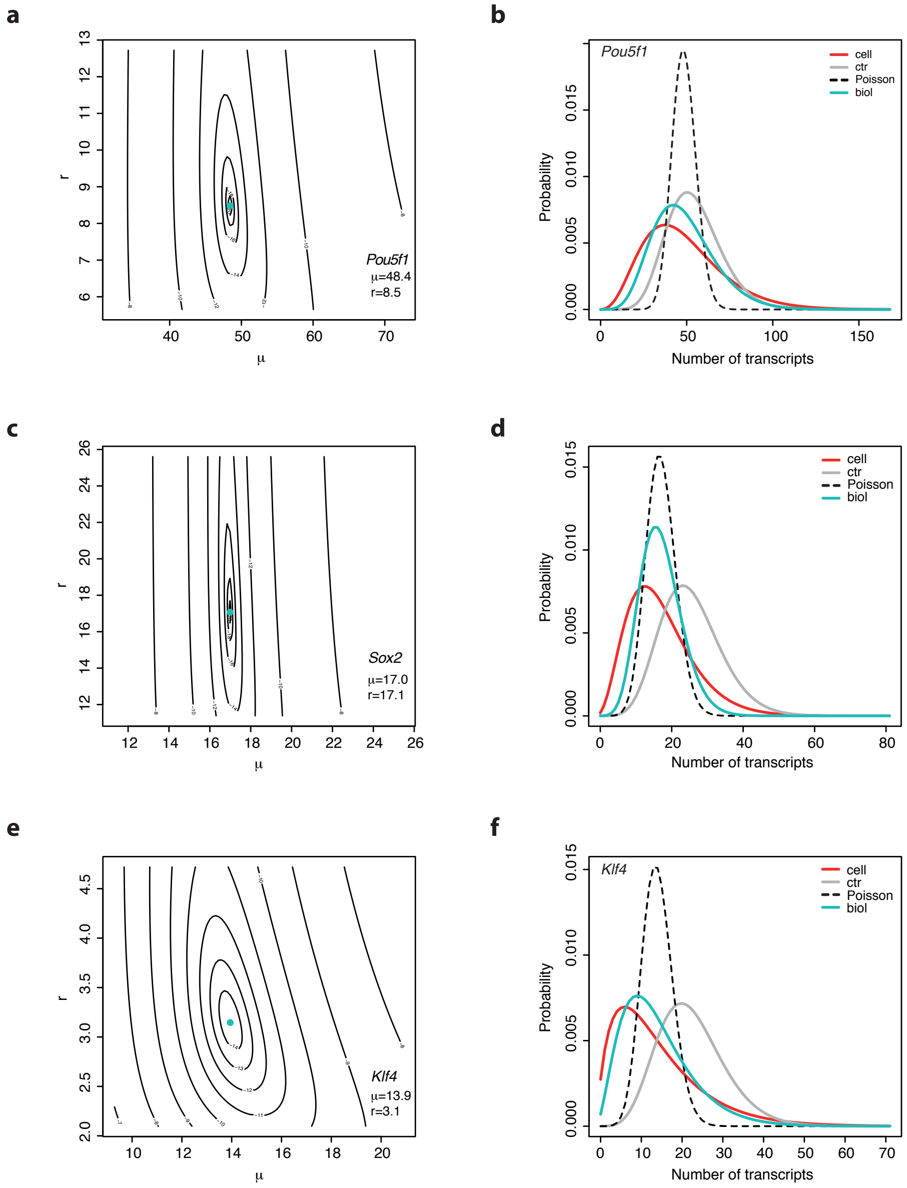
Modelling technical variability: Model II. (**a**) A regression of the number of sequenced transcripts on the predicted number of spike-in molecules was computed for each control sample. The plot shows the data and the regression line (red) for a single control sample. The slope corresponds to the conversion factor β_{II}. (**b**) Histogram of β_{II} across all cells and control samples. The count distribution was fitted by a Γ-distribution (solid green line), which was not rejected by a χ^2-test ($P < 0.61$). The uncertainty of the fit is indicated as broken green line. (**c**) Total technical variability across control samples follows a product of a Poisson distribution, capturing the sampling noise, and the Γ-distribution reflecting variability in global efficiency. The product distribution cannot be computed analytically, but was simulated for a wide range of parameters and again found to correspond to a Γ-distribution. The histogram shows the product distribution for the mean expression of *Pou5f1* and the Γ-distribution of conversion factors as predicted for control samples. The Γ-distribution was not rejected by a χ^2-test ($P < 0.20$). (**d**) The expression average μ inferred from negative binomial fits using the mean parameter obtained from single gene fits (black) or as computed from the model (green) is plotted against the mean directly computed from counts across control samples. The red line marks the diagonal. (**e**) The expression variance σ^2 inferred from negative binomial fits using the dispersion parameter obtained from single gene fits (black) or as computed from the model (green) is plotted against the variances directly computed from counts across control samples. The red line marks the diagonal.

Supplementary Figure 7



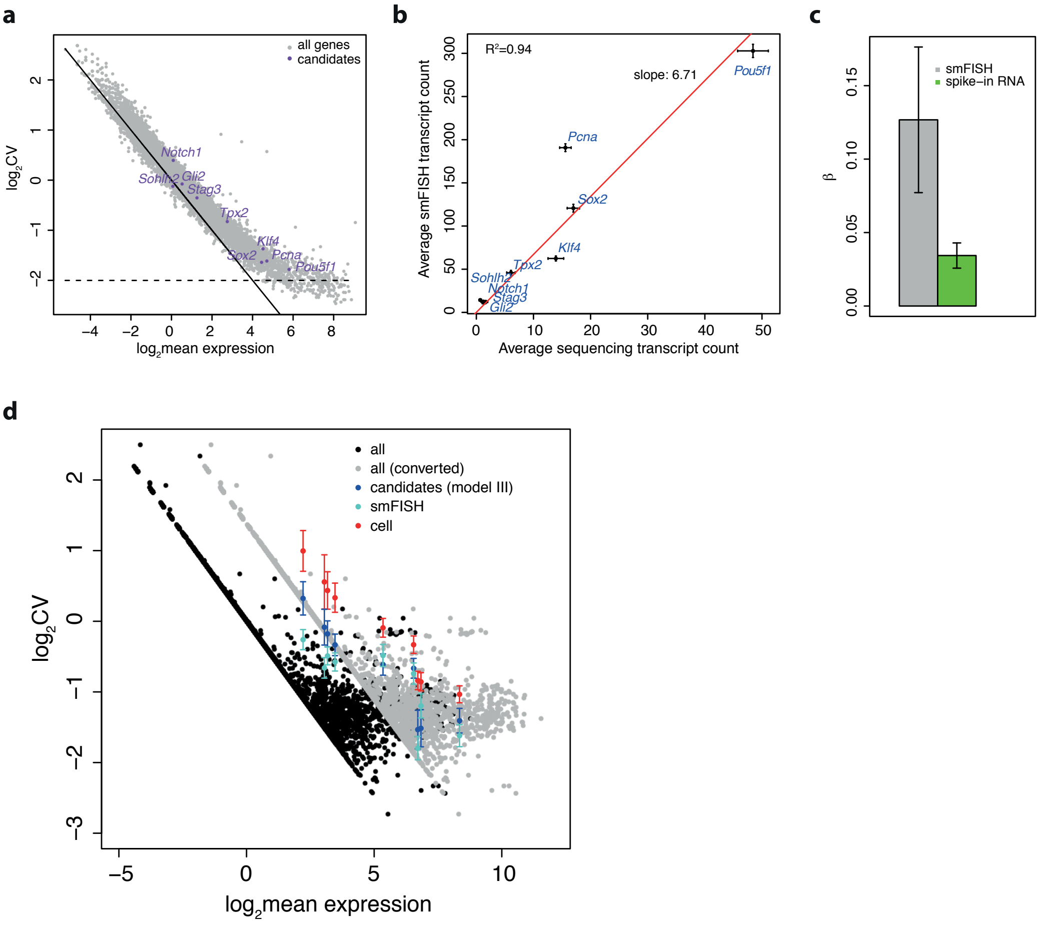
Modelling technical variability: Model III. (**a**) The conversion factor β_{III} was computed for a given spike-in species in all samples as the ratio of sequenced and added number of spike-in transcripts, and a Γ-distribution was fitted to this distribution. The Γ-distribution was not rejected by a χ^2-test ($P < 0.06$). The plot displays the count histogram of β_{III} and the maximum likelihood fit. The standard error of the fit is indicated as broken green line. (**b**) A χ^2-test was performed for spike-in count distributions at all available expression levels. The $\log_{10}$ P-value is plotted against the expression level. For data points below the broken red line, the Γ-distribution fit was rejected. However, the performance of a χ^2-test strongly depends on the number of data points and increasing the number of samples would most likely lead to better fits. (**c**) Linear regression of the rate parameter of the Γ-distribution fits on the number of spike-in RNAs in logarithmic space ($R^2 = 0.90$). (**d**) Linear regression of the shape parameter of the Γ-distribution fits on the number of spike-in RNAs in logarithmic space ($R^2 = 0.83$). (**e**) The expression average μ inferred from negative binomial fits using the mean parameter obtained from single gene fits (green) or as computed from the model (black) is plotted against the mean directly computed from counts across control samples. The red line marks the diagonal. (**f**) The expression variance σ^2 inferred from negative binomial fits using the dispersion parameter obtained from single gene fits (green) or as computed from the model (black) is plotted against the variances directly computed from counts across control samples. The red line marks the diagonal.

Supplementary Figure 8



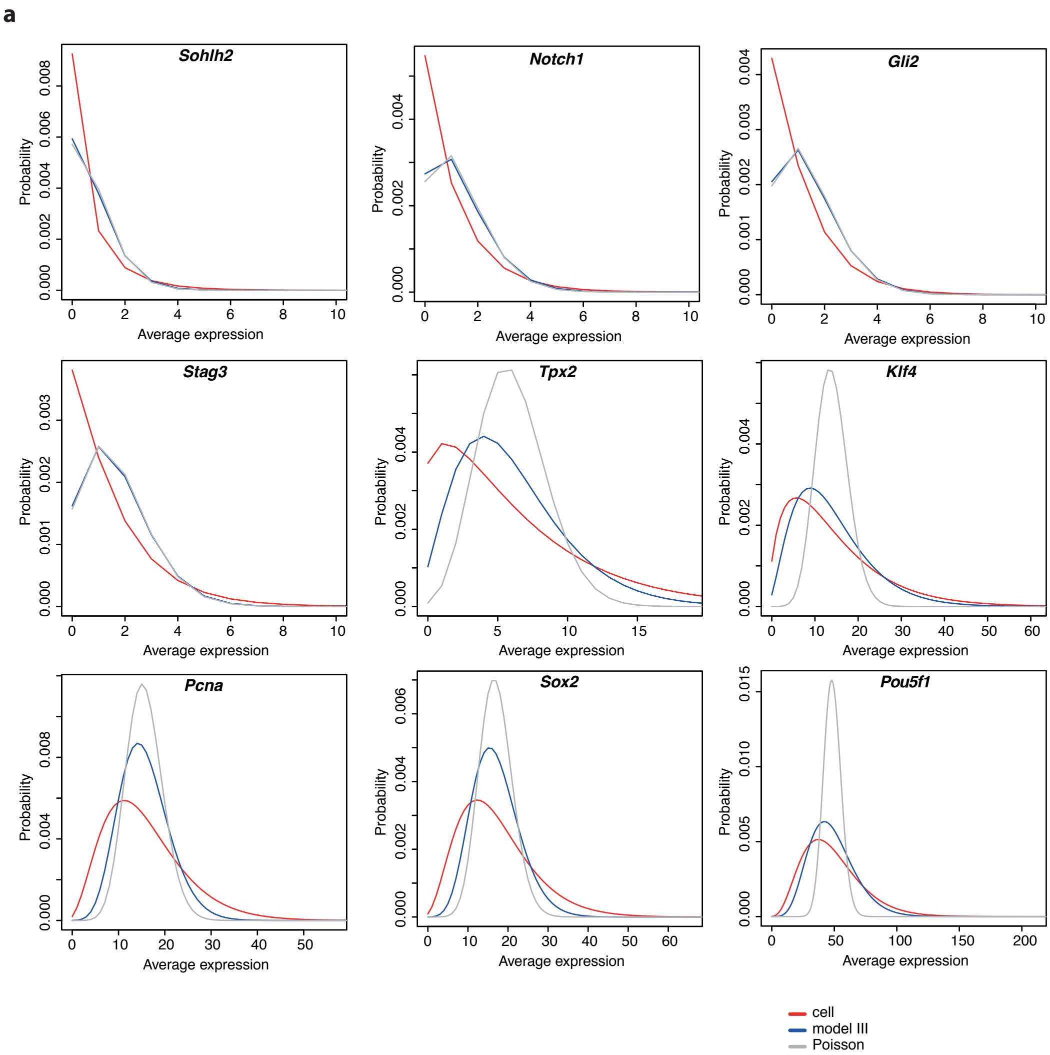
Inference of biological variability. A negative binomial describing the number of transcripts available for CEL-seq and the biological variability between single mESCs was obtained by deconvolving out technical variability from the count distribution measured for mESCs. Distribution parameters μ (mean) and r (dispersion) were inferred by a minimization procedure. (**a**) The contour plot shows the $\log_2$ expression subject to minimization (Methods, equation (21)) as a function of the inferred parameters for *Pou5f1*. The parameter combination for the unique minimum identified by the algorithm is highlighted in turquoise and parameter values are given in the lower right corner. (**b**) Count distribution for *Pou5f1* transcripts measured in cells (red), control samples (grey), and biological count distribution after deconvolution of technical variability (turquoise). A Poisson distribution is shown to visualize the degree of overdispersion. (**c**) Contour plot for *Sox2*. Details as in (a). (**d**) Count distributions for *Sox2*. Details as in (b). (**e**) Contour plot for *Klf4*. Details as in (a). (**f**) Count distributions for *Klf4*. Details as in (b).

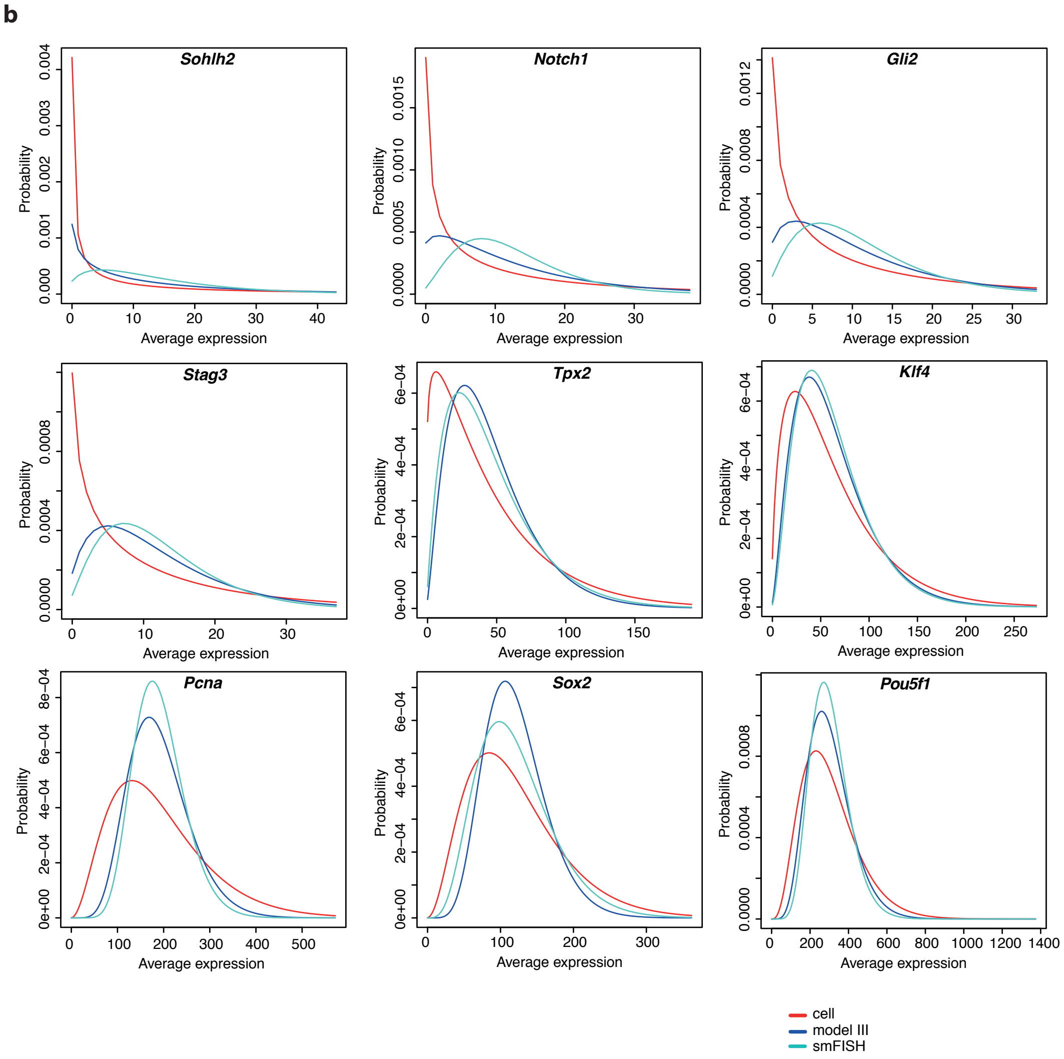
Supplementary Figure 9



Sensitivity of CEL-seq compared to smFISH. (**a**) The CV measured in cells as a function of average expression across all cells is shown. The candidates validated by smFISH (purple) cover most of the dynamic range. The Poissonian noise level (black solid line) and the global efficiency noise level are indicated (black broken line). (**b**) Comparison of transcript counts obtained by CEL-seq and smFISH. Error bars are derived from the standard error of the fit. A linear regression (red line) indicates that smFISH is almost seven times as sensitive as CEL-seq. (**c**) The conversion factor β as computed from spike-ins and as derived form smFISH (assuming 100%-sensitivity of smFISH). (**d**) The plot shows the predicted biological CV (model III) as a function of average expression in the pool of sequenced transcripts (black) and after conversion of average expression to the expected level in cells based on the smFISH data (grey). The black dots at low expression fall on a line corresponding to the Poissonian limit permitted by a negative binomial and become overdispersed after conversion. For the validated candidates the model prediction (blue), the CV measured by CEL-seq (red), and the CV measured by smFISH (turquoise) are shown (after conversion of average expression). Since our method cannot infer underdispersed biological count distributions, we will tend to overestimate noise for lowly expressed genes.

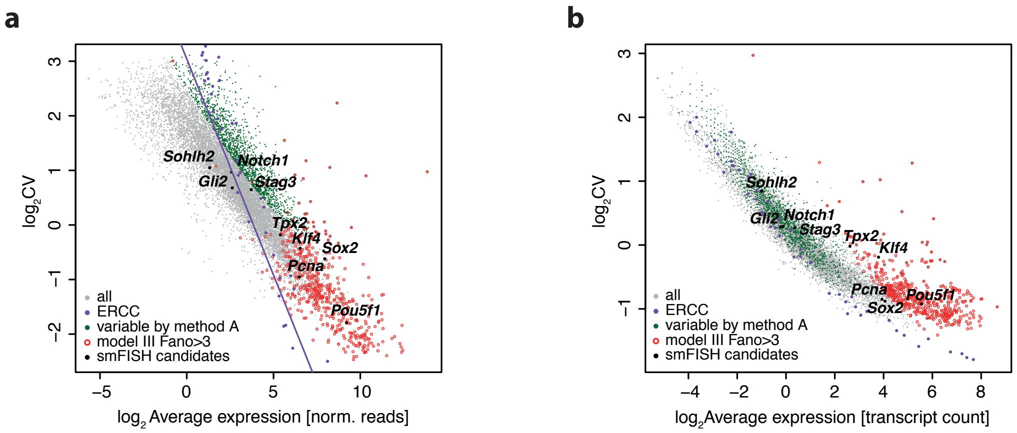
Supplementary Figure 10





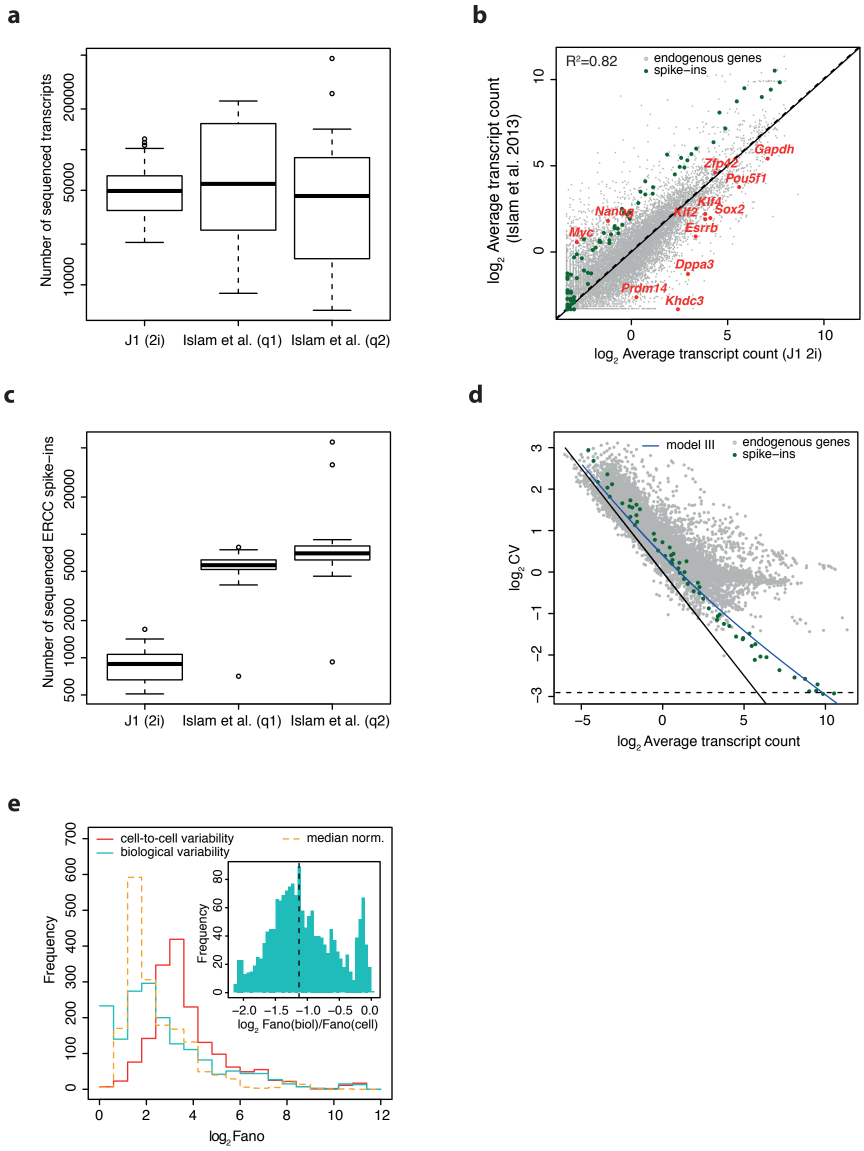
Comparison of count distribution derived from CEL-seq and smFISH. (**a**) For each gene the count distribution as measured in cells by CEL-seq (red), the inferred biological count distribution for model III (blue) and a Poisson distribution (grey) with the same mean are shown. (**b**) For each gene the smFISH derived count distribution (turquoise) is shown together with the count distribution as measured in cells by CEL-seq (red) and with the inferred biological count distribution for model III (blue) after converting the sequencing based distributions to the smFISH derived mean. The converted count distributions were computed with the sequencing derived CV and the smFISH derived expression level. A comparison of (a) and (b) reveals that an almost Poissonian biological count distribution becomes overdispersed after conversion.

Supplementary Figure 11



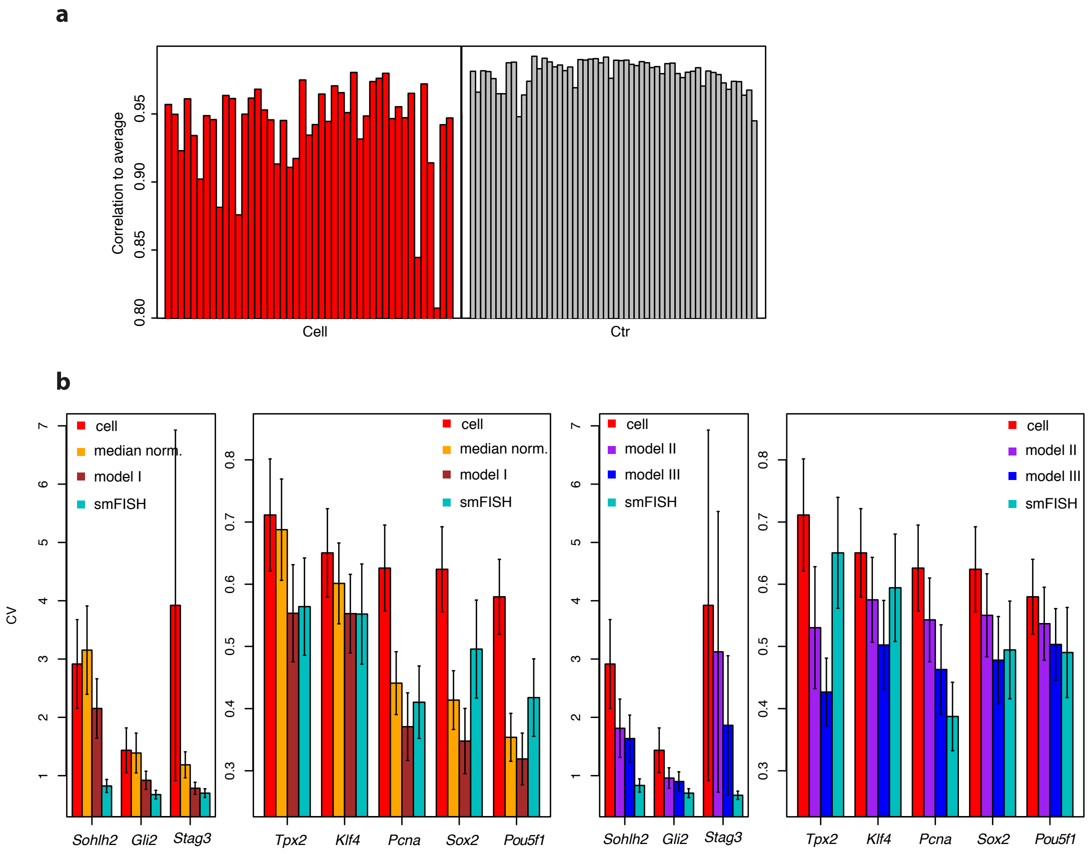
Comparison of noise predictions to a previously published method for inference of differential variability. To compare our biological noise predictions to a recently published method[2] (hereafter referred to as method A) that predicts genes with substantial levels of biological noise we first processed our sequencing data for the 2i condition by the published pipeline. Genes with substantial levels of biological noise according to this method were compared to the set of genes with a biological Fano factor > 3 as predicted by our model III. Method A was published for read based quantification instead of UMI based transcript counting. (**a**) The plot shows the CV as a function of average expression after normalizing our read data by method A. (**b**) The plot shows the CV as a function of the average transcript number per cell quantified based on UMIs. In (a) and (b) genes with substantial levels of biological noise are highlighted in green and genes with a Fano factor >3 are highlighted in red. Spike-ins (purple) and candidates validated by smFISH (black) for our method are also highlighted. In (a) a linear fit is shown for spike-ins to outline the level of technical noise. Surprisingly, method A identifies variable genes almost exclusively at low expression and the noise level of these genes still overlaps with the technical noise level indicated by the spike-ins. The more highly expressed genes that acquire a biological Fano factor > 3 based on model III show noise levels clearly exceeding technical noise as measured by spike-ins in both normalization schemes. We point out that the highly expressed stem cell markers (*Pou5f1*, *Sox2*, *Pcna*, *Klf4*) and the moderately expressed gene *Tpx2* for which we validate biological noise substantially higher than Poissonian sampling noise (**Supplementary Fig. 10**), were not identified as significantly variable genes by method A.

Supplementary Figure 12



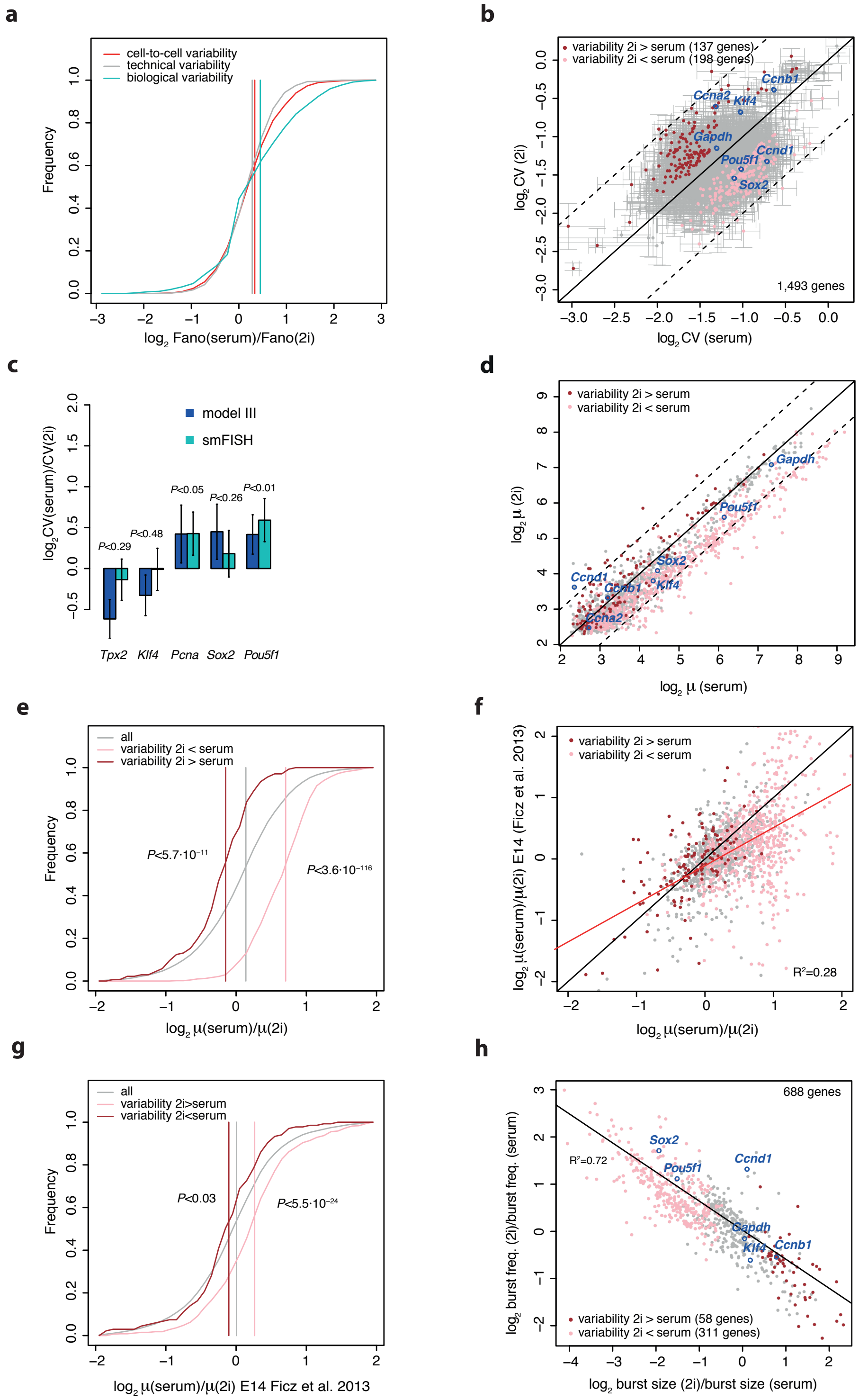
Model application to a PCR based single cell sequencing method[3]. (**a**) Comparison of the distribution of total transcript counts of endogenous genes across all sequenced cells for our J1 cells in 2i condition and the published R1 mESC data. For the latter we show the distribution after processing the sequencing data with a modified version of our pipeline and applying the same filtering steps as Islam et al.[3] (q1), and for the published raw molecule count data (q2). We found that average transcript number quantified with q1 and q2 are highly correlated (R^2=0.84) and yield similar transcript numbers. (**b**) Correlation between transcript counts in our J1 cells in 2i condition and the published R1 data quantified with our pipeline (q1). The diagonal (black solid line) and a regression for the offset (black broken line), representing the average fold change, are shown. A separate regression of the offset is shown for the ERCC spike-ins (green broken line). A number of pluripotency related genes is highlighted in red. (**c**) Same as in (a), but for the set of 92 ERCC spike-ins. Based on method q1, we measure two times more spike-in RNA while according to the published count data (q2) eight times more spike-in RNA was sequenced. The authors indicated that in total 28,000 ERCC transcripts were spiked into each sample, which is a factor of 1.12 more than in our method. (**d**) The CV as a function of the average transcript number for the published R1 data. Model III (blue line) provides a good estimate of the technical noise represented by the spike-in data (green). Poissonian noise (black solid line) and noise due to cell-to-cell variability of sequencing efficiency (black broken line) are indicated. The CV of endogenous genes measured in cells is also indicated (grey). The authors did not sequence pool-and-split controls. (**e**) Distribution of Fano factors as measured in cells, and as inferred for biological variability using model III for the published R1 data. The distribution after normalizing transcript numbers to the median without deconvolution of sampling noise is also shown. The inset shows a histogram of $\log_2$-fold changes between Fano factors before and after deconvolution of technical noise.

Supplementary Figure 13



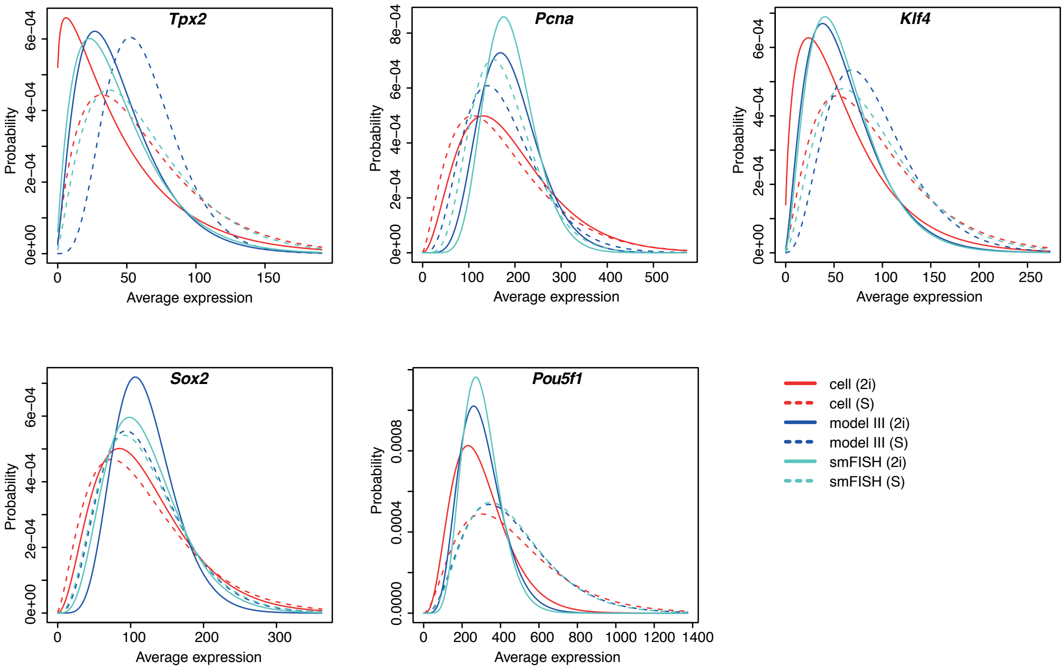
Validation of predicted biological variability by smFISH for mESCs cultured in serum. (**a**) Correlation between the number of sequenced transcripts in single mESCs or controls cultured in serum and the average across all cells or control samples, respectively. Correlation is high for all cells but on average lower and more variable than for control samples. (**b**) Validation of the inferred biological CV for the stem cell markers *Pou5f1*, *Sox2*, *Klf4* and *Pcna,* the moderately expressed gene *Tpx2*, and the lowly expressed genes *Sohlh2*, *Gli2*, and *Stag3*. To validate model I, smFISH derived transcript counts were normalized by cell area. Residual variability of total mRNA content remains and explains deviations from the model predictions. Shown is the CV measured in cells, the inferred biological CV based on the three models, and the CV measured with smFISH. In the model I comparison, the CV after normalizing to the median transcript number without deconvolution of sampling noise is shown. Error bars are derived from estimated standard errors of the negative binomial distribution parameters obtained by numerical fits. Lowly and more highly expressed genes are shown in separate plots to increase visual resolution. The ordering of the genes is the same as in **Figure 3c**. Data for Notch1 are not shown because on average only 0.6 transcripts per cell have been sequenced and the uncertainty of the CV prediction is large. In comparison to mESCs cultured in 2i, smFISH data for serum culture contain more technical noise due to background fluorescence in feeder cells. This explains that the smFISH derived CV is on average slightly increased in comparison to the sequencing derived value. The uncertainty (larger error bars) of the sequencing derived values is also increased compared to 2i conditions due to the smaller number of sequenced cells (74 mESCs in 2i versus 44 mESCs in serum).

Supplementary Figure 14



Biological noise depends on culture conditions of mESCs. (**a**) Cumulative distribution of $\log_2$-fold changes between Fano factors for cells grown in serum and 2i condition. Shown are the distributions for cells (red), controls (grey) and the inferred biological distribution (turquoise). The medians are indicated as vertical lines. (**b**) Scatter plot of the CV in serum versus 2i condition. Genes that have different CVs within their error bars between the two conditions are colored in pink (CV(serum) > CV(2i)) and brown (CV(serum) < CV(2i)). Stem cell markers (*Pou5f1*, *Sox2*) cell cycle related genes (*Ccna2*, *Ccnd1*, *Ccnb1*) and a housekeeping gene (*Gapdh*) are highlighted in blue. Error bars are based on standard errors of fitting parameters. (**c**) Comparison of CV $\log_2$-fold changes in serum versus 2i condition as predicted by model III and as measured with smFISH. Error bars are derived from estimated standard errors of the negative binomial distribution parameters obtained by numerical fits. Only more highly expressed genes, which were included in the noise comparison between the two conditions are shown. The model predicts correctly the sign of the variability change between both conditions and gives a good estimate of the difference predicted by smFISH. The p-value for the smFISH based CV fold change to be significantly different from zero is indicated for each gene. *Pou5f1* and *Pcna* show significantly enhanced CVs in serum versus 2i condition. (**d**) Scatter plot of average expression μ in serum versus 2i condition. Coloring as in **Figure 3f**. (**e**) Cumulative distribution of fold changes between serum and 2i conditions for all genes, and genes with differential variability. While genes with increased variability in serum versus 2i condition show increased expression in serum condition the opposite is true for genes with decreased variability. Wilcoxon's rank sum test p-value for a comparison to the distribution of all genes is indicated. (**f**) Comparison of observed average expression fold change between serum and 2i condition to the fold change observed by bulk sequencing of E14 ESCs in serum versus 2i condition extracted from the literature [4]. A linear regression (red line) reveals a good correlation ($R = 0.53$) between data obtained in different cell lines with different methods. (**g**) Same as in (e) but based on recently published bulk data in E14 mESCs. (**h**) Scatter plot of $\log_2$-fold changes between 2i and serum condition in burst frequency versus size. Coloring as in **Figure 3f**. Only data points for genes with a Fano-factor > 1.5 in both conditions are shown.

Supplementary Figure 15



Comparison of count distributions derived from CEL-seq and smFISH in serum versus 2i condition. For each gene the smFISH derived count distribution (turquoise) is shown together with the count distribution as measured in cells by CEL-seq (red) and with the inferred biological count distribution for model III (blue) after converting the sequencing based distributions to the smFISH derived mean. The converted count distributions were computed with the sequencing derived CV and the smFISH derived expression level. Distributions are displayed for 2i (solid lines) and serum (broken lines) condition. All tested candidates with confident noise prediction (> 5 transcripts per cell on average) were included.

Supplementary Table 1

a

2i	Lane 1			
	Control 1	Control 2	Cells 1	Cells 2
# of reads	41,751,147	23,764,721	31,287,097	20,117,132
# of mapped reads	21,636,338	13,001,381	11,002,275	8,407,015
% of mapped reads	51.8%	54.7%	35.2%	41.8%
2i	Lane 2			
	Control 1	Control 2	Cells 1	Cells 2
# of reads	42,211,075	24,027,498	31,273,378	20,307,476
# of mapped reads	21,711,713	13,020,604	10,889,938	8,418,155
% of mapped reads	51.4%	54.2%	34.8%	41.5%

b

serum	Lane 1			
	Control 1	Control 2	Cells 1	Cells 2
# of reads	34,439,319	31,821,460	33,219,585	30,956,532
# of mapped reads	13,100,787	10,003,769	9,962,452	8,735,205
% of mapped reads	38.0%	31.4%	30.0%	28.2%
serum	Lane 2			
	Control 1	Control 2	Cells 1	Cells 2
# of reads	34,143,410	32,160,773	33,616,668	31,252,645
# of mapped reads	13,604,363	10,736,439	10,679,513	9,408,096
% of mapped reads	39.8%	33.4%	31.8%	30.1%

c

2i	Number	Mapped reads per cell	Mapped UMIs per cell	Transcripts per cell
cells	74	239,866	39,742	44,170
controls	76	421,411	59,419	69,242

d

serum	Number	Mapped reads per cell	Mapped UMIs per cell	Transcripts per cell
cells	44	229,342	33,792	39,700
controls	56	284,557	55,310	64,520

Read statistics. (**a**) For cells grown in 2i culture condition and the corresponding controls, two libraries were sequenced. Material of each library was split and sequenced on two lanes. On each lane cells and controls from all libraries were barcoded with Illumina barcodes and sequenced together. The table contains the number of reads, and number and fraction of mapped reads. (**b**) Same as in (a), but for cells grown in serum culture. (**c**) For cells and controls grown in 2i culture condition, the number of cells and control samples is given and the average number of reads, UMIs and transcripts per cell is indicated. (**d**) Same as in (c), but for cells grown in serum culture.

Supplementary Table 2 (online)

GO terms enriched among genes with increased expression variability in serum versus 2i culture condition. A full list of GO terms enriched among genes with increased gene expression variability in serum versus 2i condition can be found in Supplementary Table 2 online. Enriched biological processes and enriched molecular functions are given as separate lists. Only significantly enriched GO-terms (P < 0.05) were included. The lists indicate the GO-term ID, the hypergeometric P-value, the odds ratio, the expected number of genes associated with each GO-term, the observed number of genes for each GO-term, the size of the GO-term (total number of genes associated) and a short description. For the inference of over-represented GO terms, the set of differentially variable genes was compared to the universe of all genes expressed in the two conditions. The GOstats package was used to compute GO enrichment in R.

Supplementary Table 3 (online)

Probe set composition of smFISH probes used. Each column represents a probe set for the gene specified in the column header. All probes were labeled on the 3' end with TMR, Alexa594 or Cy5. A full list of all probe sequences can be found in Supplementary Table 3 online.

Supplementary Note 1

Analysis of primary sequencing data

Sequencing of our cell and control libraries on an Illumina HiSeq2500 platform yielded ~120 million reads per lane. Of those, we could map 35%-42% and 51%-55% of the reads in our cell and control libraries, respectively, to our transcript models comprising RefSeq gene models and ERCC spike-in sequences (Supplementary Table 1 and Methods). To investigate whether the additional reads derived from introns, we mapped to pre-mRNA, which only led to a minor increase of 5% mapped reads for cells and controls. To test whether the remaining reads derived from primer sequence due to self- or cross-hybridization, we performed an 8mer enrichment analysis since direct mapping of the entire reads to primer sequences did not yield substantial mappings. We measured a strong overlap of 8mers enriched in primer sequences and reads that did not map to our gene models (**Supplementary Fig. 1a, b**). Out of the top 100 most highly enriched 8mers in primer sequences, at least one is contained in 3,605 out of 10,000 non-mapped reads, but only in 489 out of 10,000 mapped reads. These data suggest that a major fraction of non-mapped reads derive from primer sequences. For mapped and non-mapped reads, the sequencing quality was high in cells and controls. In all cases > 75% of all reads have and average Phred score of 30 or higher corresponding to 99.9% basecall accuracy (**Supplementary Fig. 1c, d**). We note that we could also map 29% of all reads for a recently published PCR based single cell sequencing method[3] and that our derived transcript counts correspond well to the published count data.

Supplementary Note 2

Three models for technical noise in single cell sequencing data

In the first model (model I) we eliminated most of the tube-to-tube variability by normalizing counts in each sample to the cross-sample median (**Fig. 2a**). Subsequently, we fitted negative binomials to the normalized distributions (**Supplementary Fig. 5a**). A negative binomial is controlled by two parameters, the mean and the dispersion parameter, and we observed a piecewise linear dependence of the dispersion parameter on the mean (**Supplementary Fig. 5b**). This dependence defines the technical noise distribution for all expression levels. Mean and variance obtained by model I were in excellent agreement with the corresponding quantities directly calculated from transcript counts (**Supplementary Fig. 5c, d**).

For model II and III global tube-to-tube variability of sequencing efficiency was derived from the statistics of the sequenced spike-ins. To calculate β, we inferred a Γ-distribution for the fraction of transcripts of a given gene available for sequencing. Molecules are randomly sampled from this pool. The number of sequenced transcripts thus corresponds to a Poisson distribution with a Γ-distributed rate, which equals a negative binomial and therefore can be fitted to the observed distribution.

In model II, β_{II} was determined for each sample from a regression of the number of sequenced spike-in transcripts on the number of spike-in molecules added to each sample (**Fig. 2b** and **Supplementary Fig. 6a**). The statistics of β_{II} were fitted by a Γ-distribution (**Supplementary Fig. 6b**) and, after superimposing Poissonian sampling noise, parameters for the negative binomial that describes technical noise within model II could be derived (**Supplementary Fig. 6c**). The first two moments of this distribution were in very good agreement with moments directly computed from transcript counts (**Supplementary Fig. 6d, e**).

In model III, the tube-to-tube variability was conditioned on the expression level (**Fig. 2b**). For each spike-in species, β_{III} was computed in every sample and a Γ-distribution was fitted to the distribution of β_{III} across all samples (**Supplementary Fig. 7a, b**). The dependence of the Γ-distribution parameters on mean expression could be explained by a simple linear model (**Supplementary Fig. 7c, d**). Using these dependencies, we inferred parameters of the negative binomial that describe the technical noise. The first two moments of this distribution corresponded well to the count derived moments (**Supplementary Fig. 7e, f**). A detailed description of all three models is given in Online Methods. Importantly, all three models do not require pool-and-split controls. Measuring spike-in RNA of known concentration across the entire dynamic range is sufficient to fit all models.

Supplementary Note 3

Comparison to a PCR based single cell sequencing method

We tested the performance of our approach on data generated by a different sequencing technique, based on PCR amplification of starting material. We applied model III to recently published single cell sequencing data for R1 mESCs generated by this method with integrated UMIs of length five[3]. Overall, the average number of transcripts was very similar to our J1 cells in 2i condition (**Supplementary Fig. 15a**), and correlation of average expression was high (R^2 = 0.82) (**Supplementary Fig. 15b**). This was surprising, since Islam et al. reported an efficiency of 48%. Even at high expression we did not see a saturation effect in the comparison of transcript numbers, suggesting that the UMI of length 4 was sufficient. After extending the annotated spike-in sequences with 5’ leader sequence that is likely derived from in vitro transcription of these sequences and was identified from sequencing reads (personal communication), we could reproduce the published spike-in expression which was six-fold higher than for our J1 cells sequenced in 2i condition (**Supplementary Fig. 15c**). This could either mean that the mRNA content of J1 and R1 cells is six-fold different, or that the spike-in RNA is less degraded in comparison to cellular RNA for the published data. Regardless of these differences, we found that our model reproduces the technical noise in these data very well, suggesting that the dominating sources were again sampling noise and global tube-to-tube variability, although the latter was strongly reduced compared to our data (**Supplementary Fig. 15d**). This was not surprising, since amplification was carried out on a microfluidic platform.

Supplementary Note 4

Exploring the origin of differential gene expression noise in 2i versus serum

To investigate the origin of differential variability in 2i condition versus serum culture condirton we first compared mean expression in both conditions and found that genes with increased variability were on average also more highly expressed in serum (**Supplementary Fig. 14d, e**). We could confirm the observed expression differences between the two conditions in recently published bulk RNA-seq data for a different ES cell line[4] (**Supplementary Fig. 14f, g**). Intuitively, high expression should not affect the Fano factor and even lead to lower variability quantified by the CV. However, this is only true for continuous transcription at a fixed rate, but not for bursting transcription. We therefore mapped our inferred parameters on a two state model for bursting transcription[5] and computed quantities that scale with burst size and frequency (Online Methods). The change in these quantities when switching from serum to 2i condition indicates that increased variability in serum culture is a consequence of lower burst frequency, but larger burst size (**Supplementary Fig. 14h**).

We finally explored the function of the more variable genes and, using Gene Ontology (GO) enrichment analysis. We compared GO terms of the more variable genes to a background of genes with similar expression (more than five transcripts) and discovered an over-representation of diverse functional categories related to RNA processing and protein production (**Supplementary Table 2**). For instance, translation initiation factors were more than threefold enriched (hypergeometric test $P < 2 \times 10^{-3}$) and structural constituents of the ribosome were more than twofold enriched among variable genes (hypergeometric test $P < 3 \times 10^{-3}$). Expression of these genes was at the same time 1.6-fold higher in serum condition. This observation provides evidence that anabolic activity is not only increased, but also more variably when culturing mESCs in serum versus 2i medium. Notably, higher expression of genes associated with anabolic activity in serum is consistent with a previous finding that anabolic activity is enhanced by Gsk3, which is inhibited in 2i medium[6].

Taken together, we discovered a global increase of gene expression noise due to less frequent, but larger expression bursts when culturing mESCs in serum versus 2i medium, and identified enrichment of genes involved in mRNA processing and protein synthesis among the more variable genes.

References

1. Baker, S. C. *et al.* The External RNA Controls Consortium: a progress report. *Nat. Methods* **2,** 731–4 (2005).

2. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10,** 1093–5 (2013).

3. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11,** 163–6 (2014).

4. Ficz, G. *et al.* FGF signaling inhibition in ESCs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency. *Cell Stem Cell* **13,** 351–9 (2013).

5. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.* **4,** e309 (2006).

6. Suzuki, T. *et al.* Inhibition of AMPK catabolic action by GSK3. *Mol. Cell* **50,** 407–19 (2013).