Quantitative Transcription Factor Analysis of Undifferentiated Single Human Embryonic Stem Cells

Anders Ståhlberg,[1,2*] Martin Bengtsson,[3] Martin Hemberg,[4,5] and Henrik Semb[6*]

BACKGROUND: Human embryonic stem cells (hESCs) require expression of transcription factor genes *POU5F1* (POU class 5 homeobox 1), *NANOG* (Nanog homeobox), and *SOX2* [SRY (sex determining region Y)-box 2] to maintain their capacity for self-renewal and pluripotency. Because of the heterogeneous nature of cell populations, it is desirable to study the gene regulation in single cells. Large and potentially important fluctuations in a few cells cannot be detected at the population scale with microarrays or sequencing technologies. We used single-cell gene expression profiling to study cell heterogeneity in hESCs.

METHODS: We collected 47 single hESCs from cell line SA121 manually by glass capillaries and 57 single hESCs from cell line HUES3 by flow cytometry. Single hESCs were lysed and reverse-transcribed. Reverse-transcription quantitative real-time PCR was then used to measure the expression *POU5F1*, *NANOG*, *SOX2*, and the inhibitor of DNA binding genes *ID1*, *ID2*, and *ID3*. A quantitative noise model was used to remove measurement noise when pairwise correlations were estimated.

RESULTS: The numbers of transcripts per cell varied >100-fold between cells and showed lognormal features. *POU5F1* expression positively correlated with *ID1* and *ID3* expression ($P < 0.05$) but not with *NANOG* or *SOX2* expression. When we accounted for measurement noise, *SOX2* expression was also correlated with *ID1*, *ID2*, and *NANOG* expression ($P < 0.05$).

CONCLUSIONS: We demonstrate an accurate method for transcription profiling of individual hESCs. Cell-to-cell variability is large and is at least partly nonrandom because we observed correlations between core transcription factors. High fluctuations in gene expression may explain why individual cells in a seemingly undifferentiated cell population have different susceptibilities for inductive cues.

Embryonic stem cells (ESCs)[7] are pluripotent cells derived from the inner cell mass of the blastocyst. These cells are self-renewing, with the unique capacity to generate any cell type in the body. This capability is the basis for considering the human ESCs (hESCs) as an unlimited source of cells for replacement therapies and for the treatment of a wide range of diseases, such as diabetes mellitus and Alzheimer and Parkinson diseases *(1)*.

Undifferentiated ESCs require expression of transcription factor genes *NANOG*[8] (Nanog homeobox), *POU5F1* (POU class 5 homeobox 1; alias, *OCT4*), and *SOX2* [SRY (sex determining region Y)-box 2] to maintain their unique characteristics. Genetic approaches to increase or decrease the amounts of these transcription factors have caused rapid cell differentiation, indicating a key role for the NANOG, POU5F1, and SOX2 factors in maintaining ESCs in an undifferentiated state *(2–7)*. Furthermore, recent genome-wide promoter-binding studies have shown a high degree of complexity, in that NANOG, POU5F1, and SOX2 not only form an autoregulatory network in undifferentiated hESCs to promote self-renewal but also appear to block differentiation by repressing many of the essential cell fate regulators *(2)*. Gene expression

[1] Lundberg Laboratory for Cancer, Department of Pathology, Sahlgrenska Academy at University of Gothenburg, Gothenburg, Sweden; [2] TATAA Biocenter, Gothenburg, Sweden; [3] Department of Clinical Sciences, Lund University, Clinical Research Centre, UMAS, Malmö, Sweden; [4] Department of Bioengineering, Imperial College London, London, UK; [5] Department of Ophthalmology and Program in Neurobiology, Children's Hospital Boston, Harvard Medical School, Boston, MA; [6] Stem Cell Center, Lund University, Lund, Sweden.

* Address correspondence to: A.S. at Lundberg Laboratory for Cancer, Department of Pathology, Sahlgrenska Academy at University of Gothenburg, Gula Straket 8, SE- 413 45 Gothenburg, Sweden. Fax +46-31828733; e-mail anders.stahlberg@tataa.com. H.S. at Stem Cell Center, Lund University, BMC B10, SE-221 84 Lund, Sweden. Fax +46-462223600; e-mail henrik.semb@med.lu.se.

Received June 9, 2009; accepted September 9, 2009.
Previously published online at DOI: 10.1373/clinchem.2009.131433

[7] Nonstandard abbreviations: ESC, embryonic stem cell; hESC, human ESC; RT-qPCR, reverse-transcription quantitative real-time PCR.

[8] Genes: *POU5F1*, POU class 5 homeobox 1; *NANOG*, Nanog homeobox; *SOX2*, SRY (sex determining region Y)-box 2; *ID1*, inhibitor of DNA binding 1, dominant negative helix-loop-helix protein; *ID2*, inhibitor of DNA binding 2, dominant negative helix-loop-helix protein; *ID3*, inhibitor of DNA binding 3, dominant negative helix-loop-helix protein; *Mus musculus* gene *Id1*, inhibitor of DNA binding 1; *M. musculus* gene *Id2*, inhibitor of DNA binding 2; *M. musculus* gene *Id3*, inhibitor of DNA binding 3; *POLR2B*, polymerase (RNA) II (DNA directed) polypeptide B, 140kDa; *RPLP0*, ribosomal protein, large, P0; *SUZ12*, suppressor of zeste 12 homolog (*Drosophila*).

analysis of *NANOG*, *POU5F1*, and *SOX2* in cell populations of undifferentiated hESCs have led to the conclusion that these factors are coregulated in these cells *(2–5)*.

Although microarray studies have been useful for understanding gene regulation and expression in different cell types, substantial variability within a population of cells has also been demonstrated. These deviations from the population mean are often referred to as noise, and the noise level can be appreciable for transcription *(8–10)*. One source of such noise is the production of mRNA in bursts *(10–12)*. An important consequence of the noise is that cell populations will exhibit some degree of variability even if the cells are genetically identical and have been exposed to the same environment *(13)*. Collectively, the results from measurements of gene expression in single cells indicate a strong stochastic element that causes highly variable expression, which in turn means that data obtained from cell population measurements cannot be extrapolated to individual cells.

To advance our understanding of the transcriptional regulation of *NANOG*, *POU5F1*, and *SOX2* in undifferentiated hESCs, we have performed quantitative expression studies of these genes in individual cells. The aim of this study was to determine to what extent the observed heterogeneity of 2 genes that are presumed to be correlated is a consequence of noise in the measurements and to what extent it is the result of underlying cell-to-cell variations. In addition, we have included in our study inhibitor of DNA binding genes *ID1*, *ID2*, and *ID3* (inhibitors of DNA binding 1, 2, and 3, dominant negative helix-loop-helix protein), the expression of which is known to inhibit neuroectoderm differentiation in undifferentiated mouse ESCs *(14)*, although the importance of the *ID* genes in hESCs has yet to be confirmed. Furthermore, NANOG, POU5F1, and SOX2 have been experimentally verified to bind at the promoter regions of *ID1* and *ID2 (2)*. In situ analyses of *ID* gene expression during development have demonstrated widespread expression of *ID1*, *ID2*, and *ID3* throughout the developing organism, from early gestation through birth and with considerable overlap in the expression patterns of *ID1* and *ID3 (15)*. Mice (*Mus musculus*) lacking *Id1*, *Id2*, or *Id3* [inhibitors of DNA binding 1, 2, and 3] are all viable, whereas mice without both *Id1* and *Id3* die at day E13.5 *(16)*. We applied reverse-transcription quantitative real-time PCR (RT-qPCR) to the analysis of mRNA in single cells. This method is characterized by a wide dynamic range, high reproducibility, and a sensitivity sufficient to detect single molecules *(17, 18)*.

We show that individual hESCs can be collected by either glass capillaries or flow cytometry. Transcripts of *ID1*, *ID2*, *ID3*, *NANOG*, *POU5F1*, and *SOX2* can be accurately measured with RT-qPCR. We also provide a mathematical model for determining correlations between genes, which can compensate for experimental noise, thereby revealing additional statistically significant correlations. Our results reveal that the numbers of transcripts in individual hESCs are highly variable and that the degree of *POU5F1* expression does not correlate with that of *SOX2* or *NANOG*, whereas the expression of members of the *ID* gene family correlates with the expression of *POU5F1* and *SOX2*. Gene expression profiling of single hESCs allows us to study cell heterogeneity and to understand why individual hESCs differentiate differently.

Materials and Methods

hESC CULTURES

Cell lines SA121 *(19)* (Cellartis) and HUES3 *(1)* were used for in vitro experiments with hESCs in accordance with Swedish ethics guidelines. Undifferentiated SA121 cells were maintained on mitotically inactivated mouse embryonic fibroblasts. Half of the medium was changed every 2 days, and SA121 cells were passaged manually every 4–7 days onto fresh mouse embryonic fibroblasts, as previously described *(19)*. HUES3 was enzymatically passaged and cultivated as previously described *(1)*.

COLLECTION OF SINGLE CELLS AND CELL LYSIS

hESCs were rinsed with PBS (200 mg/L KCl, 200 mg/L KH_2PO_4, 8 g/L NaCl, 2.16 g/L Na_2HPO_4-$7H_2O$; Invitrogen) and then dissociated into a suspension of single cells with TrypLE Select (Invitrogen) or with Trypsin, 0.05% (1X) with EDTA 4Na (Invitrogen) for 3 min at 37 °C. TrypLE Select and trypsin were inactivated with hESC medium *(1)*, and the cells were replated onto petri dishes. Individual cells were collected with heat-treated glass pipettes (Hilgenberg) mounted on a micromanipulator over an inverted microscope. Pipettes were emptied into a 200-μL plastic tube with 2 μL lysis solution containing 5 mL/L IGEPAL CA-630, 50 mmol/L Tris-HCl pH 8.0, 140 mmol/L NaCl, and 1.5 mmol/L $MgCl_2$ (all Sigma-Aldrich). Tubes were then immediately heated to 80 °C for 5 min and stored at −80 °C until reverse transcription. A more detailed description of this procedure has been reported *(20, 21)*. To collect single cells by flow cytometry, we used either the FACSDiva or FACSVantage instrument (both BD Biosciences). The flow cytometry instrument was manually calibrated to deposit single cells in the center of each collection tube. 7-Aminoactinomycin D (Sigma-Aldrich) was added as a viability marker in the sorting procedure. hESCs were kept in PBS containing 25 mL/L fetal bovine serum (BIOCHROM) before cell sorting.

Total RNA was purified with the GenElute Mammalian Total RNA Purification Kit (Sigma–Aldrich).

REVERSE TRANSCRIPTION

Reaction tubes containing single lysed cells in lysis solution (described above) supplemented with 0.5 mmol/L of each of the 4 deoxynucleotides (Sigma-Aldrich), 2.5 μmol/L oligo(dT) (Invitrogen), and 2.5 μmol/L random hexamers (Invitrogen) were heated to 65 °C for 5 min and then chilled on ice. We then added 50 mmol/L Tris-HCl (pH 8.3), 75 mmol/L KCl, 3 mmol/L $MgCl_2$, 5 mmol/L dithiothreitol, 20 U RNaseOUT, and 100 U SuperScript III (all Invitrogen) to a final volume of 10 μL. Samples were incubated at 50 °C for 90 min and then inactivated enzymatically at 70 °C for 15 min, as previously described *(22)*.

QUANTITATIVE REAL-TIME PCR

Two instruments were used for real-time PCR measurements: the ABI PRISM 7900HT Sequence Detection System (Applied Biosystems) and the LightCycler 2.0 Real-Time PCR System (version 4.6; Roche Diagnostics). Reactions of 10 or 20 μL contained 10 mmol/L Tris (pH 8.3), 50 mmol/L KCl, 3 mmol/L $MgCl_2$, 0.3 mmol/L of each deoxynucleotide, 1 U JumpStart *Taq* polymerase (all Sigma-Aldrich), 0.5× SYBR Green I (from a 10 000× concentrate; Invitrogen), and 400 nmol/L of each primer (MWG-Biotech). BSA (0.1 g/L; Fermentas) was added to LightCycler reactions, and 1× Reference Dye for Quantitative PCR (Sigma-Aldrich) was used as a passive reference dye in ABI PRISM 7900HT reactions. Real-time PCRs started with 3 min of preincubation at 95 °C, followed by 50 amplification cycles. The following temperature profile was used (with LightCycler/ABI PRISM 7900HT incubation times): denaturation at 95 °C for 0/20 s, annealing at 60 °C for 10/20 s, and elongation at 73 °C for 15/20 s. All primers except those for *SOX2* were designed to span an intron to avoid amplification of genomic DNA. *SOX2* lacks introns, and therefore its amplification could be biased by genomic amplification. BLAST searches, however, revealed no pseudogenes for *SOX2*. The genomic background of 2 DNA copies can therefore be disregarded, compared with the degree of *SOX2* expression (approximately 250 mRNA copies). The primer sequences are shown in Table 1 in the Data Supplement that accompanies the online version of this article at http://www.clinchem.org/content/vol55/issue12. All assays of single cells were optimized not to generate primer dimers before cycle 40, to have a PCR efficiency of at least 80%, and to amplify all known splice forms documented by the National Center for Biotechnology Information (NCBI). Calibration curves with purified PCR products (QIAquick PCR Purification Kit; Qiagen) were used to establish the linearity of the assays. PCR efficiencies for *ID1*, *ID2*, *ID3*, *NANOG*, *POU5F1*, *POLR2B* [polymerase (RNA) II (DNA directed) polypeptide B, 140kDa], *RPLP0* (ribosomal protein, large, P0), *SOX2*, and SUZ12 [suppressor of zeste 12 homolog (*Drosophila*)] were 93%, 93%, 86%, 96%, 88%, 86%, 92%, 95%, and 88%, respectively ($r^2 > 0.99$ for calibration curve slopes). Formation of the correct PCR products was confirmed by electrophoresis on 20 g/L agarose gels for all assays and by melting-curve analysis of all samples. Purified PCR products were quantified with the NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies/Thermo Scientific) to generate calibration curves for all assays and thereby allow absolute quantification. RT-qPCR data analysis was performed as previously described *(17)*. Additional information according to MIQE (Minimum Information for Publication of Quantitative Real-Time PCR Experiments) guidelines *(23)* is shown in Table 1 in the online Data Supplement. Cell population data were normalized against those for *RPLP0*. Potential reference genes were evaluated with NormFinder and a reference panel of 10 genes (Human Endogenous Control Gene Panel; TATAA Biocenter). For calculations of RT-qPCR CIs for low transcript numbers, 16-plicate RT-qPCR was carried out with diluted, purified total RNA from hESCs at 6 different RNA concentrations (concentrations were prepared by making 4-fold serial dilutions). These data were also used in our mathematical model for correlation.

SINGLE-CELL MEASUREMENTS

Approximately 20 target molecules per PCR are needed for accurate RT-qPCR quantification *(20)*, and this requirement was satisfied for most cells and genes used in this study. By keeping the dilution between the reverse-transcription and real-time PCR steps at a minimum, we obtained the highest reproducibility *(20)*. All qPCR results are expressed as the number of cDNA molecules. To calculate the number of mRNA molecules requires determining the reverse-transcription efficiency. We have previously shown how the reverse-transcription efficiency can be determined from serial dilutions of known cDNA calibrators *(20)*, and this efficiency is usually <100% *(22)*. Reverse-transcription efficiencies are shown in Table 2 in the online Data Supplement. Because cDNA is single-stranded, 1 cycle was subtracted from the measured value when we calibrated with calibration curves based on double-stranded PCR products *(24)*. All transcript numbers are related to a single cell and not to reference genes. The use of constantly expressed reference genes for sample comparison, which is appropriate at the cell-population level, is not valid at the level of single cells because of the occurrence of transcriptional bursts

(11, 12). The experimental variation in each assay is shown in Figs. 1 and 2 in the online Data Supplement.

IMMUNOSTAINING

hESCs were fixed in 40 g/L paraformaldehyde (Merck) for 20 min, permeabilized for 30 min in PBS (200 mg/L KCl, 200 mg/L KH_2PO_4, 8 g/L NaCl, 2.16 g/L Na_2HPO_4-$7H_2O$; Invitrogen) containing 5 mL/L Triton X-100 (Sigma–Aldrich), and blocked in PBS with 50 g/L skim milk powder (Merck) for 30 min. The cells were incubated overnight at 4 °C with a primary antibody against POU5F1 (Santa Cruz Biotechnology) and then incubated with a Cy3-conjugated secondary antibody (Jackson Immunoresearch Laboratories) at room temperature for 1 h. Cell nuclei were counterstained with 4′,6-diamidino-2-phenylindole (Invitrogen).

MATHEMATICAL MODEL FOR CORRELATION

For a given gene, we assume that the observed number of transcripts, X, is the sum of 2 components: the true biological amount of mRNA, X_B, and the noise stemming from the measurement, X_T. We assume that the noise has $E[X_T] = 0$ ($E[X_T]$ means "expectation value of X_T") and that it is uncorrelated to the biological component or to the measurement noise of every other gene. From these assumptions, it follows that the covariance for a pair of genes X and Y is:

$$\mathrm{Cov}(X_B + X_T, Y_B + Y_T) = \mathrm{Cov}(X_B, Y_B). \quad (1)$$

Furthermore, we assume that the variance of a gene, σ^2, can be written as the sum of 2 components, $\sigma^2 = \sigma_B^2 + \sigma_T^2$, which correspond to the inherent biological fluctuations and to the measurements, respectively. Because the measurement noise has $E[X_T] = 0$, it follows that the normalized variance can be written as:

$$\eta^2 = \eta_T^2 + \eta_B^2 = \frac{\sigma_T^2 + \sigma_B^2}{\mu^2}, \quad (2)$$

where η_T^2 is the noise strength of measurement ($\eta^2 = \sigma^2/\mu^2$), η_B^2 is the biological variability, and $E[X_B] = \mu$. η_T^2 and η_B^2 can be computed with the strategy described in *(20)*. Defining $\beta = \eta_T^2/\eta_B^2$, we may write the variance as:

$$\sigma = (1 + \beta)\eta_B^2\mu^2. \quad (3)$$

Starting from the definition of the correlation coefficient, ρ, and using Eqs. 1 and 3, we obtain:

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{\mathrm{Cov}(X,Y)}{\sigma_{X,B}\sigma_{Y,B}\sqrt{(1+\beta_X)(1+\beta_Y)}}$$
$$= \frac{\rho_B}{\sqrt{(1+\beta_X)(1+\beta_Y)}}, \quad (4)$$

where ρ_B is the true biological correlation coefficient. Using Eq. 4, we can estimate the reduction in correlation (γ) due to the measurement noise as:

$$\gamma = \frac{1}{\sqrt{(1+\beta_X)(1+\beta_Y)}}. \quad (5)$$

Because the denominator in Eq. 5 will always be >1, it follows that $0 \le \gamma \le 1$. We may think of γ as a multiplicative factor, inversely proportional to the magnitude of the measurement noise, that describes the loss of correlations due to the measurement noise. With the values in Table 2 in the online Data Supplement, we can calculate the γ factor for each pair of genes, and, via the relation $\gamma_B = \rho/\gamma$, we obtain the true biological correlation. Using a t-test for the statistic:

$$t = \rho\frac{\sqrt{N-2}}{\sqrt{1-\rho^2}}, \quad (6)$$

with $N - 2$ degrees of freedom, where N is the number of data points, we may calculate the significance of ρ_B *(25)*.

Results and Discussion

We manually collected 47 undifferentiated hESCs (cell line SA121) with glass capillaries and performed RT-qPCR. Fig. 1 shows that transcripts for at least 1 gene were detected in 45 cells (96%). The 95% CIs for the RT-qPCR measurements of *POU5F1*, *NANOG*, *SOX2*, *ID1*, *ID2*, and *ID3* are shown in Fig. 2. Because the number of transcripts in a single cell is finite, only a limited number of genes may be accurately measured for each cell. We have previously shown that the inherent imprecision of the RT-qPCR measurements is significantly increased when there are <20 copies of mRNA molecules present at the start. In this study, we have >20 transcripts for most cells and genes, a result that implies that the biological variability is substantially greater than the variation introduced by the measurements *(20)*. The RT-qPCRs were run sequentially for each single cell, which meant that roughly one sixth (approximately 15%) of the cell's content was analyzed each time. Thus, the theoretical lower limit at which our experimental setup becomes less reproducible is approximately 20 × 6 = 120 transcripts when 6 genes are analyzed. The most frequently detected genes are expressed in quantities greater than this number (Table 1).

The collection step of our protocol requires that hESCs be dissociated into single cells. To evaluate the effect of this treatment, we analyzed total mRNA from a small cell population before enzymatic dissociation, after enzymatic dissociation, and after enzymatic dissociation followed by a 3-h incubation in cell medium.

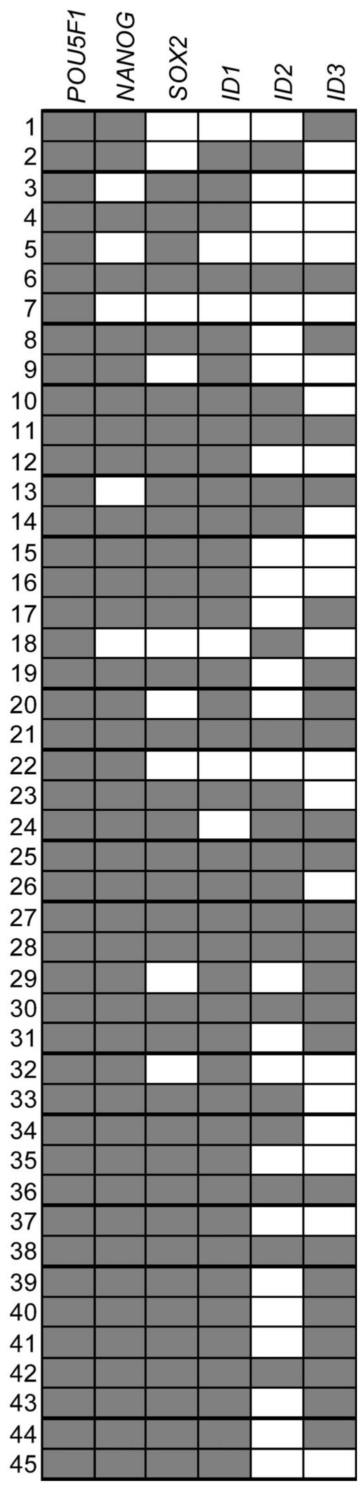

Fig. 1. Detection of transcripts in 45 individual hESCs.

Filled boxes indicate the presence of *POU5F1*, *NANOG*, *SOX2*, *ID1*, *ID2*, or *ID3* mRNA.

The incubation time was chosen to represent the time elapsed between the first and the last manually collected cells. Only small variations in gene expression (<1.5-fold) could be detected between the 3 dissociation procedures (see Fig. 3 in the online Data Supplement). Moreover, we observed no correlation between transcript counts and the time of cell collection, indicating that cell handling did not impose any additional differentiation. Thus, we conclude that the applied cell-collection method had no major impact on hESC differentiation.

To further verify our method, we used flow cytometry to collect single hESCs from another cell line (HUES3, n = 57; see Tables 3 and 4 in the online Data Supplement). Instead of the *ID* genes (which were slightly upregulated when maintained in PBS; see Fig. 3 in the online Data Supplement), we analyzed the *SUZ12* and *POLR2B* genes in the second cell line. SUZ12 is a subunit of polycomb repressive complex 2, and POLR2B is a subunit of RNA polymerase 2. SUZ12 was recently shown to block expression of differentiation genes in undifferentiated hESCs *(26)*. Data from this experiment are in accordance with those obtained for the manually picked hESCs, but with marginally lower mean expression values for *POU5F1* and *NANOG*.

Fig. 3 shows the distribution of *POU5F1* mRNA copy numbers in the undifferentiated hESC population. *POU5F1* expression ranges from approximately 190 to approximately 24 000 mRNA copies/cell (>100-fold variation), with a median expression of approximately 7400 transcripts. Immunofluorescence analysis revealed high variation in the amount of POU5F1 protein as well (see Fig. 4 in the online Data Supplement) *(27)*. Median expression values for the remaining transcription factor genes (Table 1) indicate that *NANOG* and *ID1*, like *POU5F1*, are abundantly expressed (approximately 2800 and 1200 transcript copies/cell, respectively), whereas *SOX2*, *ID2*, and *ID3* are expressed at a much lower level (approximately 250, 53, and 81 transcript copies/cell, respectively). The expression of *ID2* and *ID3* is relatively low, and therefore the expression of these genes is more prone than the other genes to measurement errors due to the decreased reproducibility of the PCR *(20)*.

Lognormal distributions are common in biological processes, and they typically arise when the underlying variables affect the outcome in a multiplicative manner *(28)*. Transcript numbers in individual mammalian cells have been reported to often appear lognormally distributed *(11, 21, 29)*. Indeed, for all 6 genes we can reject the hypothesis that their transcript numbers are normally distributed. After a logarithmic transformation, we find that the normality hypothesis for *NANOG*, *ID1*, *ID2*, and *ID3* cannot be rejected by

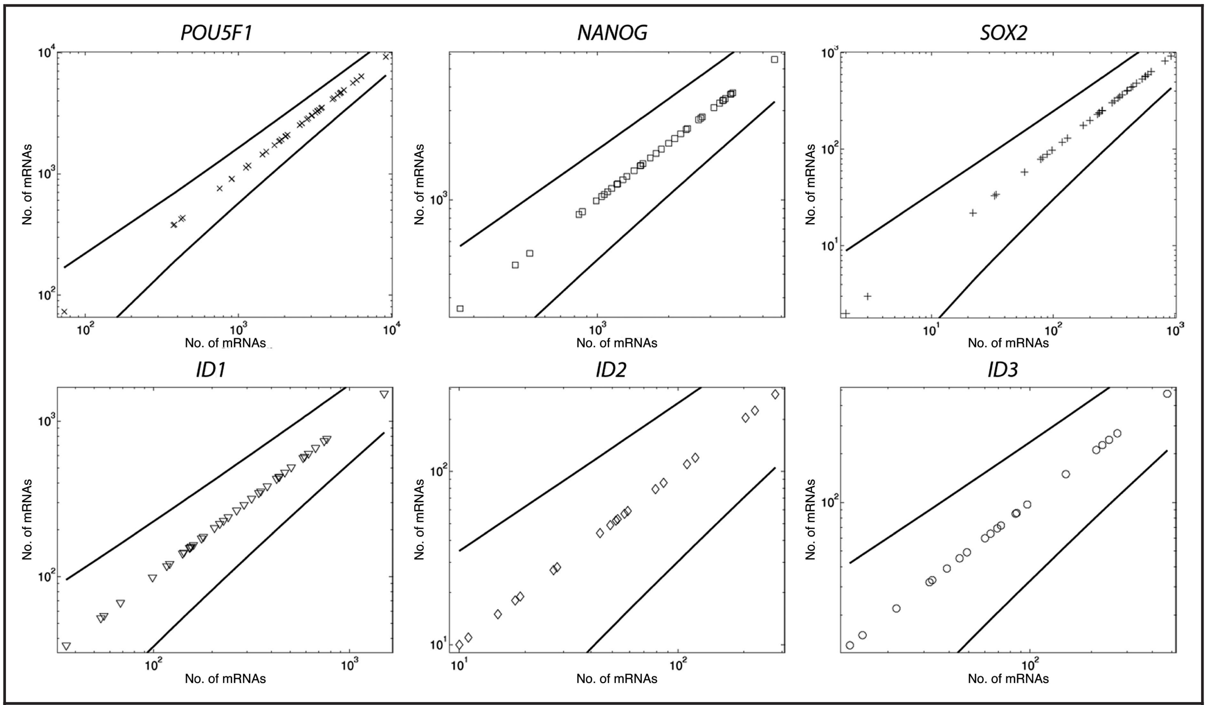

Fig. 2. CIs for RT-qPCR measurements.

Indicated are 95% CIs for different amounts of mRNA for *POU5F1*, *NANOG*, *SOX2*, *ID1*, *ID2*, and *ID3*. Data points represent the gene expression of each individual hESC. The CIs are based on a lognormal distribution with parameters estimated from the experimental data.

the Shapiro–Wilk normality test at a 95% confidence level. In lognormally distributed populations, a cell's characteristic degree of expression is best represented by the geometric mean, which corresponds to the median for this distribution (Table 1). Use of the arithmetic mean (which is an appropriate summary statistic for the normal distribution) would overestimate the number of transcripts in a typical cell. A consequence of the skewed distribution in transcript numbers is that a majority of the transcripts for a particular gene originate from a minority of the cells in the population. For example, the 5 hESCs with the highest numbers of *ID1* transcripts contributed to 34% of all transcripts for this gene, whereas the 5 lowest-expressing cells contributed

Table 1. Statistical parameters describing gene expression in single hESCs.

Gene	No. of cells[a]	Median[b]	Geometric mean[c]	Log_{10} geometric mean (SD)	Skewness[d]
POU5F1	45	7400	5500	3.74 (0.40)	−1.34
NANOG	40	2800	2700	3.43 (0.28)	−0.71
SOX2	36	250	180	2.25 (0.60)	−1.58
ID1	39	1200	1200	3.07 (0.36)	−0.15
ID2	20	53	50	1.70 (0.42)	0.05
ID3	24	81	89	1.95 (0.43)	0.13

[a] Number of cells expressing the tested gene ($N_{total} = 47$).

[b] Median value represents the number of transcripts for the experimental median cell.

[c] *NANOG*, *ID1*, *ID2*, and *ID3* were lognormally distributed; *POU5F1* and *SOX2* were not ($P < 0.05$, Shapiro–Wilk normality test). Data represent the number of transcripts.

[d] A negative skewness value indicates that the lognormal distribution is skewed toward lower expression values.

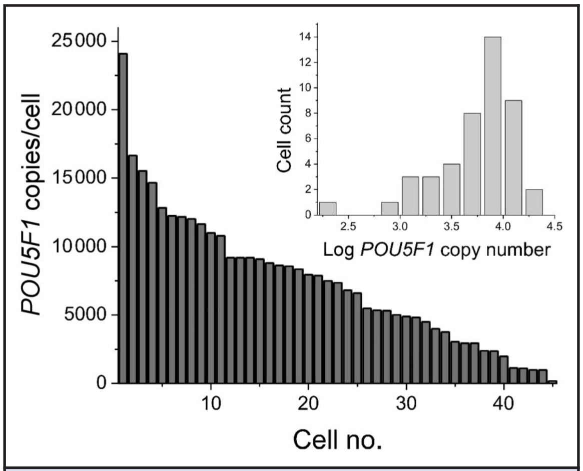

Fig. 3. ***POU5F1*** **expression in single hESCs.**

Shown are 45 *POU5F1*-expressing cells ranked by the number of transcript copies, from high to low. Inset presents the corresponding histogram of transcript copy number on a $\log_{10}$ scale.

Table 2. Spearman correlation coefficients of expression values in single hESCs.[a]

	POU5F1	*NANOG*	*SOX2*	*ID1*	*ID2*	*ID3*
POU5F1	1	0.17	0.08	**0.50**	0.04	**0.60**
		(0.22)	(0.14)	(0.63)	(0.06)	(0.79)
NANOG		1	0.24	0.05	0.01	0.19
			(**0.44**)	(0.07)	(0.01)	(0.28)
SOX2			1	0.37	0.31	0.04
				(0.64)	(0.68)	(0.06)
ID1				1	−0.12	0.21
					(−0.19)	(0.29)
ID2					1	−0.04
						(−0.02)
ID3						1

[a] Spearman correlation coefficients compensated for technical variation in RT-qPCR are shown within parentheses. Underscored values indicate ≥99% significance; boldface values indicate ≥95% significance (Holm–Bonferroni correction).

to only 2%. This result means that data from pooled cell populations are biased toward cells with high expression, a phenomenon that may lead to incorrect interpretations.

An important question relates to the functional significance of the high variation in mRNA amounts. *POU5F1*, *NANOG*, and *SOX2* encode transcription factors that regulate how the cell differentiates, and their expression is required to maintain the pluripotent state. Mathematical modeling suggests a switch-like behavior of these genes, meaning that the genes have a high tolerance for fluctuations in mRNA concentrations before differentiation is initiated *(30)*. Some of the measured cells have very low numbers of *POU5F1* and *SOX2* transcripts. Close inspection of the individual cells reveals that all 5 cells not expressing *NANOG* have low *POU5F1* expression (see Fig. 5 in the online Data Supplement). We hypothesize that this pattern is an early sign of differentiation, and we plan to investigate this concept further in future studies. Furthermore, some cells with low or no *POU5F1* and *NANOG* expression have considerable expression of *SOX2*, which may be an early indication of differentiation toward an ectodermal cell fate. *SOX2*, in contrast to *POU5F1* and *NANOG*, is not down-regulated in early differentiation toward ectoderm [*(31)* and unpublished data]. When undifferentiated hESCs are exposed to inductive cues, only a fraction of the cells will respond, and then at different times. The discovery that large variations in the expression of genes encoding different transcription factors involved in self-renewal exist in individual cells could explain why individual cells in a seemingly undifferentiated cell population have different susceptibilities for differentiation. We also point out, however, that our hESC cultures are regularly validated by POU5F1 and NANOG staining and that nonstained cells in principle are not observed. We therefore speculate that our observed signs of early differentiation at the mRNA level indicate a reversible state of the cells.

We also confirmed previous results that *POU5F1*, *SOX2*, and *NANOG* expressions correlate at the population level (data not shown). Surprisingly, this finding was not the case at the single-cell level (Table 2; Fig. 5 in the online Data Supplement). Correlation coefficients ranged from 0.08 to 0.24, indicating no or only very weak correlation between these genes. The lack of correlation was confirmed with another cell line (HUES3; see Table 4 in the online Data Supplement). *POU5F1*, *SOX2*, and *NANOG* were recently suggested to all share the same feedback regulatory mechanism *(2, 4)*; however, these findings are inconsistent with our data from single cells, which suggest a more complex transcriptional regulation. Our data allow us only to detect simultaneous correlations, so we cannot exclude a coupling between 2 genes with a more complicated temporal pattern or at the protein level. Interestingly, further analysis showed that *POU5F1* expression was positively correlated with *ID1* and *ID3* expression (Table 2). This result is consistent with detailed promoter analyses, which have shown that *ID1* and *ID3* have similar regulatory elements *(32)*. We observed a correlation between *ID1*, *ID3*, and *POU5F1*, suggesting that the transcription factors encoded by these genes may be coregulated in undifferentiated hESCs. The lack of a correlation between *ID1* and *ID3* may be due to the exclusion of cells with no *ID1* or *ID3* transcripts. Correlation data are therefore somewhat biased toward highly expressing cells, and we cannot exclude a correlation between *ID1* and *ID3*.

The high noise levels of individual genes in combination with the measurement uncertainties could potentially obscure real correlations. We previously developed a mathematical framework for quantifying the measurement noise for RT-qPCR *(20)*. In brief, by measuring the experimental noise for different mRNA and cDNA concentrations, we may model the measurement noise as a function of the molecules for each gene. The mathematical noise model allows us to determine the contribution of measurement noise to the total observed noise, which includes the biological variability. We have extended this methodology (for details see Materials and Methods) to allow us to estimate the decrease in correlation due to the measurement noise. The values in parentheses in Table 2 show the estimated biological correlations once the measurement noise has been removed. Interestingly, 3 more pairs of genes (*SOX2* and *NANOG*, *SOX2* and *ID1*, and *SOX2* and *ID2*) now become significantly correlated. These additional correlations further support the hypothesis that *ID* genes prevent differentiation and are functionally involved in self-renewal. We conclude that measurement noise in one gene is enough to decrease the observed correlation (see Fig. 6 in the online Data Supplement). Assuming that β (the fraction of noise from the measurement) is independent of the number of cells analyzed, we estimate that approximately 600 cells are needed to statistically verify a correlation of 0.08. For a correlation of 0.22, 80 cells are needed ($P < 0.05$).

In summary, we have demonstrated that RT-qPCR is an accurate and sensitive method to measure the expression levels of multiple genes in single hESCs. We show that the transcription of *POU5F1*, *SOX2*, and *NANOG* is highly variable among undifferentiated hESCs. The lack of strong correlations suggests that these genes, although important regulators of self-renewal, are independently regulated during self-renewal. Instead, we found that *POU5F1* and *SOX2* are correlated with the *ID* genes during hESC self-renewal. Our results support the notion that seemingly heterogeneous cell cultures, as shown by staining, can have important differences at the mRNA level as well. This finding could help explain why culture cells respond differently to differential cues.

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.*

Authors' Disclosures of Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the Disclosures of Potential Conflict of Interest form. Potential conflicts of interest:*

Employment or Leadership: None declared.
Consultant or Advisory Role: None declared.
Stock Ownership: None declared.
Honoraria: None declared.
Research Funding: The Lund Stem Cell Center, the Juvenile Diabetes Research Foundation, the Swedish Research Council, and the JDRF Center for Beta Cell Therapy in Europe. A. Ståhlberg, the Swedish Research Council and the Socialstyrelsens Foundation; M. Hemberg, the Engineering and Physical Sciences Research Council UK and the Children's Hospital Ophthalmology Foundation.
Expert Testimony: None declared.

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, or preparation or approval of manuscript.

Acknowledgments: We thank Ann-Katrin Häger for technical assistance.

References

1. Cowan CA, Klimanskaya I, McMahon J, Atienza J, Witmyer J, Zucker JP, et al. Derivation of embryonic stem-cell lines from human blastocysts. N Engl J Med 2004;350:1353–6.
2. Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, et al. Core transcriptional regulatory circuitry in human embryonic stem cells. Cell 2005;122:947–56.
3. Chambers I, Colby D, Robertson M, Nichols J, Lee S, Tweedie S, Smith A. Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. Cell 2003; 113:643–55.
4. Loh YH, Wu Q, Chew JL, Vega VB, Zhang W, Chen X, et al. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nat Genet 2006;38:431–40.
5. Mitsui K, Tokuzawa Y, Itoh H, Segawa K, Murakami M, Takahashi K, et al. The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. Cell 2003; 113:631–42.
6. Niwa H, Miyazaki J, Smith A. Quantitative expression of Oct-3/4 defines differentiation, dedifferentiation or self-renewal of ES cells. Nat Genet 2000;24:372–6.
7. Assou S, Carrour TL, Tondeur S, Ström S, Gabelle A, Marty S, et al. A metaanalysis of human embryonic stem cell transcriptome integrated into a Web-based expression atlas. Stem Cells 2007; 25:961–73.
8. Blake WJ, Kaern M, Cantor CR, Collins JJ. Noise in eucaryotic gene expression. Nature 2003;422: 633–7.
9. Raser JM, O'Shea EK. Noise in gene expression: origins, consequences, and control. Science 2005; 309:2010–3.
10. Raj A, van Oudenaarden A. Nature, nurture, or chance: stochastic gene expression and its consequences. Cell 2008;135:216–26.
11. Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S. Stochastic mRNA synthesis in mammalian cells. PLoS Biol 2006;4:e309.
12. Ross IL, Browne CM, Hume DA. Transcription of individual genes in eukaryotic cells occurs randomly and infrequently. Immunol Cell Biol 1994; 72:177–85.
13. Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic gene expression in a single cell. Science 2002;297:1183–6.
14. Ying QL, Nichols J, Chambers I, Smith A. BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3. Cell 2003;115:281–92.

15. Ruzinova MB, Benezra R. Id proteins in development, cell cycle and cancer. Trends Cell Biol 2003;13:410–8.
16. Lyden D, Young AZ, Zagzag D, Yan W, Gerald W, O'Reilly R, et al. Id1 and Id3 are required for neurogenesis, angiogenesis and vascularization of tumour xenografts. Nature 1999;401:670–7.
17. Nolan T, Hands RE, Bustin SA. Quantification of mRNA using real-time RT-PCR. Nat Protoc 2006;1:1559–82.
18. Peixoto A, Monteiro M, Rocha B, Veiga-Fernandes H. Quantification of multiple gene expression in individual cells. Genome Res 2004;14:1938–47.
19. Heins N, Englund MC, Sjöblom C, Dahl U, Tonning A, Bergh C, et al. Derivation, characterization, and differentiation of human embryonic stem cells. Stem Cells 2004;22:367–76.
20. Bengtsson M, Hemberg M, Rorsman P, Ståhlberg A. Quantification of mRNA in single cells and modelling of RT-qPCR induced noise. BMC Mol Biol 2008;9:63.
21. Bengtsson M, Ståhlberg A, Rorsman P, Kubista M. Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. Genome Res 2005;15:1388–92.
22. Ståhlberg A, Kubista M, Pfaffl MW. Comparison of reverse transcriptases in gene expression analysis. Clin Chem 2004;50:1678–80.
23. Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, et al. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clin Chem 2009;55:611–22.
24. Ståhlberg A, Håkansson J, Xian X, Semb H, Kubista M. Properties of the reverse transcription reaction in mRNA quantification. Clin Chem 2004;50:509–15.
25. Glenberg AM, Andrzejewski M. Learning from data: an introduction to statistical reasoning. 3rd ed. New York: Lawrence Erlbaum Associates; 2007.
26. Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS, Kumar RM, et al. Control of developmental regulators by polycomb in human embryonic stem cells. Cell 2006;125:301–13.
27. Ware CB, Nelson AM, Blau CA. A comparison of NIH-approved human ESC lines. Stem Cells 2007;24:2677–84.
28. Limpert E, Stahel WA, Abbt M. Log-normal distributions across the sciences: keys and clues. Bioscience 2001;51:341–52.
29. Warren L, Bryder D, Weissman IL, Quake SR. Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR. Proc Natl Acad Sci U S A 2006;103:17807–12.
30. Chickarmane V, Troein C, Nuber UA, Sauro HM, Peterson C. Transcriptional dynamics of the embryonic stem cell switch. PLoS Comput Biol 2006;2:e123.
31. Rex M, Orme A, Uwanogho D, Tointon K, Wigmore PM, Sharpe PL, Scotting PJ. Dynamic expression of chicken Sox2 and Sox3 genes in ectoderm induced to form neural tissue. Dev Dyn 1997;209:323–32.
32. Lopez-Rovira T, Chalaux E, Massague J, Rosa JL, Ventura F. Direct binding of Smad1 and Smad4 to two distinct motifs mediates bone morphogenetic protein-specific transcriptional activation of Id1 gene. J Biol Chem 2002;277:3176–85.