



OPEN

Transcriptome-wide Variability in Single Embryonic Development Cells

Vincent Piras^{1,2}, Masaru Tomita^{1,2} & Kumar Selvarajoo^{1,2}¹Institute for Advanced Biosciences, Keio University, 14-1 Baba-cho, 997-0035, Tsuruoka, Japan, ²Systems Biology Program, Graduate School of Media and Governance, Keio University, 5322 Endo, 252-0882, Fujisawa, Japan.SUBJECT AREAS:
COMPUTER MODELLING
CELLULAR NOISE
STOCHASTIC MODELLING
GENE EXPRESSIONReceived
22 July 2014Accepted
4 November 2014Published
20 November 2014Correspondence and
requests for materials
should be addressed to
K.S. (kumar@tck.keio.
ac.jp)

Molecular heterogeneity of individual molecules within single cells has been recently shown to be crucial for cell fate diversifications. However, on a global scale, the effect of molecular variability for embryonic developmental stages is largely underexplored. Here, to understand the origins of transcriptome-wide variability of oocytes to blastocysts in human and mouse, we examined RNA-Seq datasets. Evaluating Pearson correlation, Shannon entropy and noise patterns (η^2 vs. μ), our investigations reveal a phase transition from low to saturating levels of diversity and variability of transcriptome-wide expressions through the development stages. To probe the observed behaviour further, we utilised a stochastic transcriptional model to simulate the global gene expressions pattern for each development stage. From the model, we concur that transcriptome-wide regulation initially begins from 2-cell stage, and becomes strikingly variable from 8-cell stage due to amplification and quantal transcriptional activity.

Numerous studies on single cells have shown that individual molecules (genes, proteins or metabolites), within an iso-genic and -phenotypic cell population, can display highly variable expression levels. For example, immunofluorescence flow cytometry showed that Sca-1 expressions in multipotent murine hematopoietic cells follow a Gaussian-like distribution¹, and the monitoring of green fluorescent proteins in *Escherichia coli* displayed fluctuations in their expression levels over time².

Such heterogeneous or noisy characteristics have shown to play pivotal roles for the survival of species to diverse environmental conditions or for cell fate decisions³⁻⁵. Notably, it was demonstrated that regulating stochastic noise in the levels of *comK* in *Bacillus subtilis* was necessary to control cell fate decision under nutrient-deficient conditions⁶. For *Caenorhabditis elegans*, the intestinal cell fate process from early embryonic lineage was shown to be regulated by the variability in *end-1* expression, providing the basis for incomplete penetrance⁷. These studies have identified crucial single molecules that regulate heterogeneity or variability of single cells within a population. However, little work has been performed to investigate global responses, comprising the entire spectrum of molecular species, within single cells. In particular, the extent of transcriptome-wide expressions noise in the early mammalian development has yet to be determined.

In this paper, to understand global gene expression structure and noise patterns of single cells during mammalian developmental stages, we investigated transcriptome-wide RNA-Seq expressions of several cells during human⁸ and mouse⁹ embryogenesis. A total of 7 human and 10 murine cell origins, from oocytes to blastocysts, were analysed using high-dimensional statistical techniques, such as correlation metrics¹⁰⁻¹⁶, Shannon entropy¹⁷⁻¹⁹ and noise analyses²⁰.

Results

Single cell transcriptome expressions scatter increases along development stages. To observe gene expression variability between 2 single cells at each developmental stage, we plotted pair-wise distributions of single cell transcriptomes (Figure 1). For human, we noticed global expressions scatter is tightly constrained up to 2-cell stage, after which the scatter widened, especially for lowly expressed genes. For mouse, the scatter widened noticeably from middle of 2-cell stage. These data suggest that transcriptome-wide expression distributions become more variable along the developmental stages.

To better understand the variability and the effects of technical and biological noises, we performed transcriptome-wide correlation (similarity) analysis¹⁰⁻¹⁴ by comparing the expressions of two cells from the same cell origin. Although large expressions scatters are observed, especially for late developmental stages (Figure 1), the global averaged Pearson correlation coefficients between single cells of the same stage, as expected²⁰, are generally high (Figure 2A, dotted lines). However, the correlation coefficients between cells of distinct origins, are signifi-

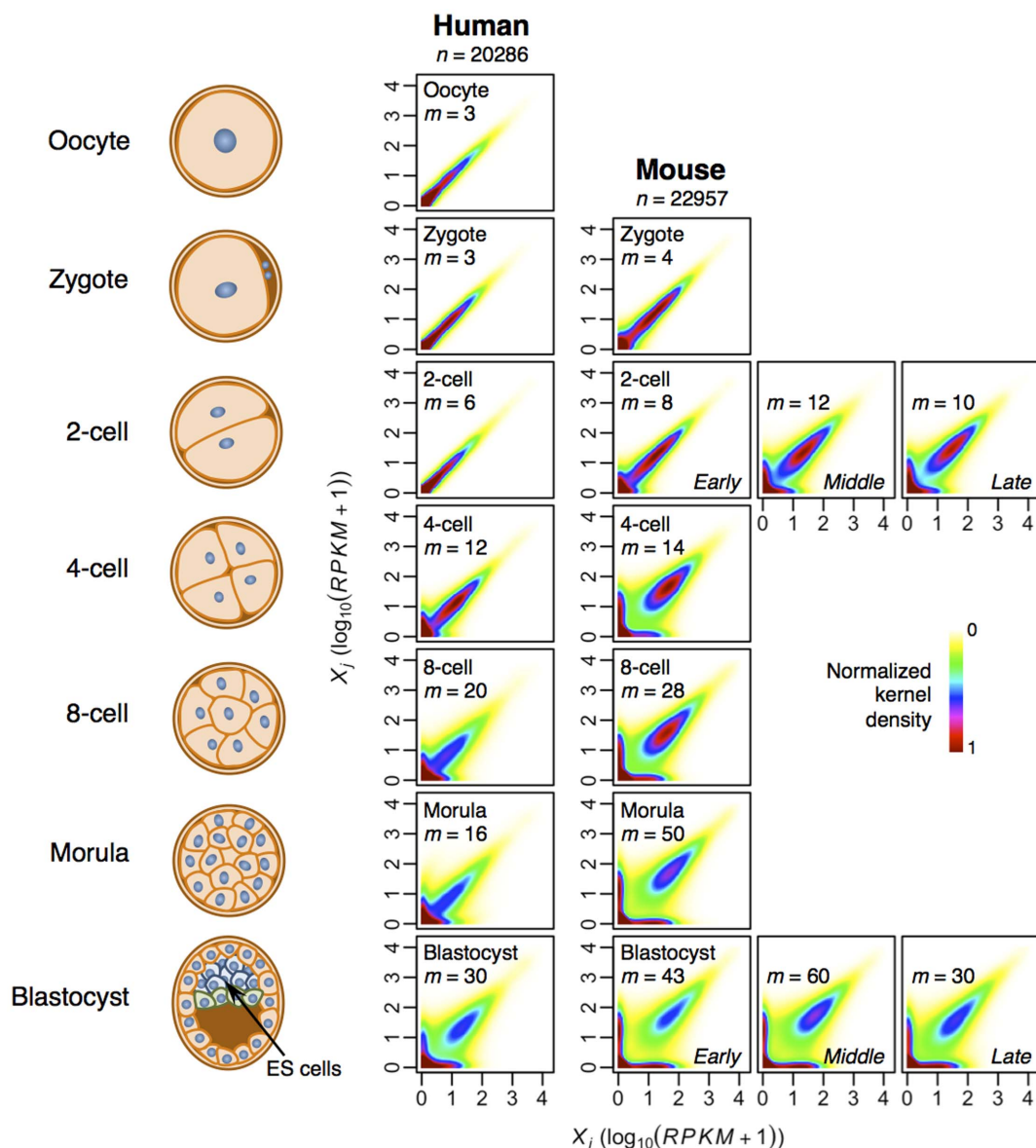


Figure 1 | Gene expression structures for developmental stages. Gene expression distributions, estimated from kernel density estimation (*kde2d* R function) of all genes expressions (RPKM) between all possible pairs of single cells in human and mouse, from oocytes to blastocysts. m is the total number of single cells and n is the total number of genes.

cantly lower (Figure 2A, solid lines and Supplementary Fig. S1 online). We further probed for non-linear relationships between the transcriptomes of cells of the same stage or different stages, and computed Spearman ranking metrics (non-linear monotonic relationships), distance correlation¹⁵ (strict statistical dependence) and maximum information coefficient¹⁶ (linear and non-linear associations) (see Materials and Methods). Remarkably, all metrics revealed similar trends compared with Pearson correlations (Supplementary Fig. S2 online). These results indicate that the global transcriptional program of developmental cells clearly deviates along the stages in time, with faster rate of deviation occurring for mouse when compared with human (Figure 2A, solid lines).

Next, we assessed the diversity of single cell transcriptomes using Shannon entropy, which measures the disorder of a high-dimensional system, where higher values indicate increasing disconnection between variables and zero value indicates order^{17–19} (Materials and Methods). For both human and mouse, Shannon entropies remained low in early stages but gradually increased from 2-cell (human) or 4-cell (mouse) stage, to reach high values for morula and blastocyst

(Figure 2B). This result, therefore, shows the disconnection or diversity of transcriptome-wide expressions increases during mammalian development.

Transcriptome-wide noise increases during developmental stages.

To further understand the effects of increasing entropy and diversity in single cell transcriptomes during embryogenesis, we quantified single cells' expressions scatter by computing transcriptome-wide average noise (renamed as total noise), η_{tot}^2 , i.e. summing the squared coefficient of variation²¹, defined as the variance (σ^2) of expression divided by the square mean expression (μ^2), for all genes (i) between all possible pairs of single cells (Materials and Methods). We observed that η_{tot}^2 is low during initial embryonic cell differentiation, but increases at later stages with significant increase from 2- to 4-cell stage onwards (Figure 3A). We also compared total noise for embryonic stem, normal somatic and cancer cells, and found similar values as obtained for later stage developmental cells (Supplementary Fig. S3 online). These data indicate that total noise stabilises at ~ 0.7 and may not increase further.

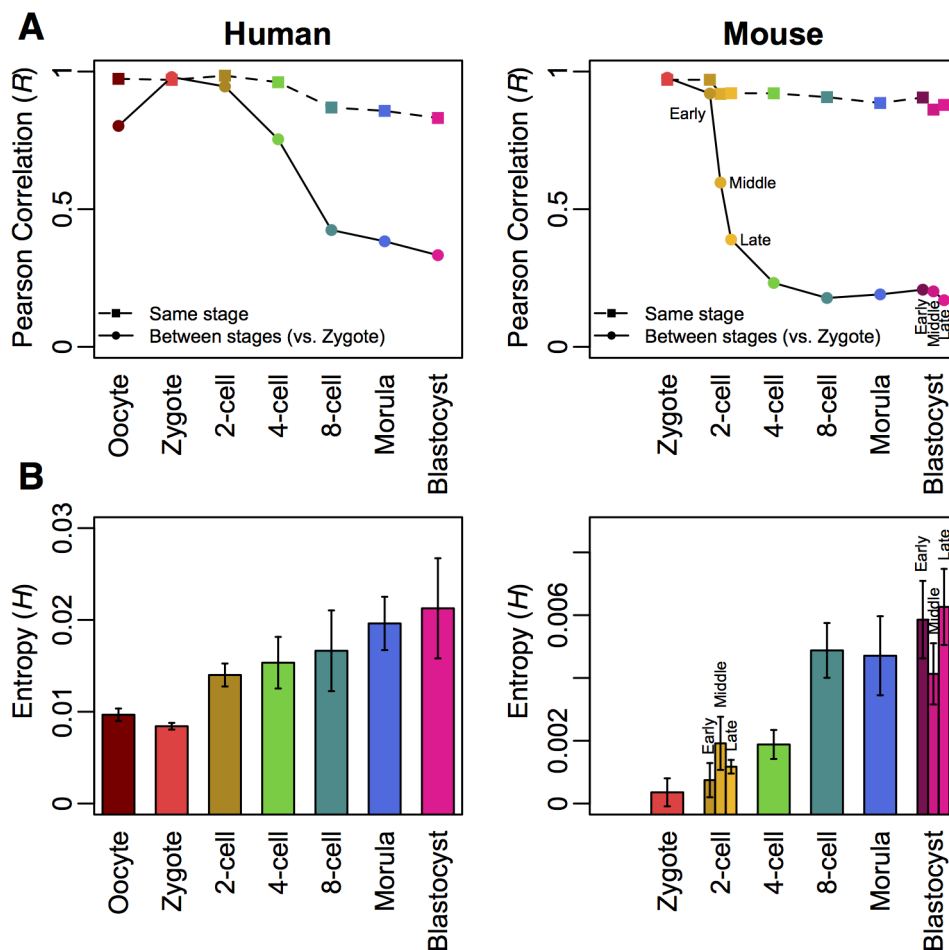


Figure 2 | Phase transition in single cell transcriptomes. (A) Pearson correlation, R , between transcriptomes of cells of the same development stage (dotted lines) and or between transcriptomes of zygote and other stages (solid lines) for human and mouse. (B) Shannon entropy (H) of single cell transcriptomes (average for m cells, error bars indicate 1 s.d.).

So far, we have analysed the entire transcriptome without setting any expression cut-off. It is known that lowly expressed genes in single cells are dominated by stochastic and/or technical noises, which reduce their between cells correlation values, while highly expressed genes show more deterministic expressions²². Thus, we delineated total noise to investigate average noise for every group of 500 genes, between pairs of single cells, as a function of mean expressions (μ) starting from the lowest expressions.

As expected, we noticed noise is relatively high for the lowly expressed portion of the transcriptome ($\eta^2(\mu) \sim 2$ for $\mu < 0.1$ for all cell types) (Figure 3B). This portion of transcriptome is usually discarded due to low signal-to-noise ratios. Considering noise patterns above this threshold ($\mu \geq 0.1$), we observed all patterns followed the relationship, $\eta^2(\mu) = \alpha/\mu + \beta$, where α and β are proportionality constant and asymptotic value, respectively (Supplementary Fig. S4 online). That is, noise scales with the inverse of mean values reaching asymptotic values at higher expressions for all cell types and species, including other embryonic stem cells, somatic cells and human cancer cells²³ (Supplementary Fig. S3 online). Notably, this relationship was also previously observed for other high-throughput datasets^{21,24}.

Fitting the noise pattern of each cell type and species, we observed α and β increase during the human development process; oocyte, zygote and 2-cells stages show similar patterns, 4-cell stage shows transitional values, while 8-cell, morula and blastocyst stages reach higher values. Similar trends were also found for mouse, nevertheless, the transition between low and high values occurred earlier at middle of 2-cell stage (Figure 3C). These results show that transcrip-

ptome-wide noise increased along the developing stages and stabilises from 8-cell stage. Note that other grouping sizes, e.g. 50, 100, 1000 genes (Supplementary Fig. S5 online), did not affect the overall increase in transcriptome-wide noise along the developing stages.

Deciphering the origins of noise patterns using a stochastic transcriptional model. To understand the transcriptional mechanisms governing increasing noise patterns along the development process, we utilised a single cell transcriptional model based on Gillespie stochastic algorithm. We simulated transcriptome data constituting 20,000 gene ‘units’, where each gene dynamics is governed by 3 ordinary differential equations with 5 kinetic parameters^{25,26}; transcription rate (s), degradation rate (δ), promoter activation (k_{on}) and deactivation (k_{off}) rate constants, where both continuous gene promoter activation ($k_{off} = 0$) and quantal (bursty) dynamics ($k_{off} > 0$) can be simulated (Figure 4A). The transcriptional amplification process²⁷, i.e. number of transcripts produced per activation event, is controlled by ϕ .

Using the model, we simulated gene expressions for various conditions by controlling the 5 kinetic parameters. Firstly, simulations were performed by choosing the transcription (s) and degradation (δ) rate parameters for each gene from statistical distributions found in experimental data^{28,29} (Supplementary Fig. S6 online). The three other parameters were kept at default constant values ($\phi = 1$, $k_{off} = 0$ and $k_{on} = 0.5$). In this theoretical setting, generated noise is entirely intrinsic (and Poisson), and decreased with mean expressions with α constant and $\beta = 0$ for the whole range of gene expressions (Figure 4B, panel 1). It is conceivable that none of the development stages follow this

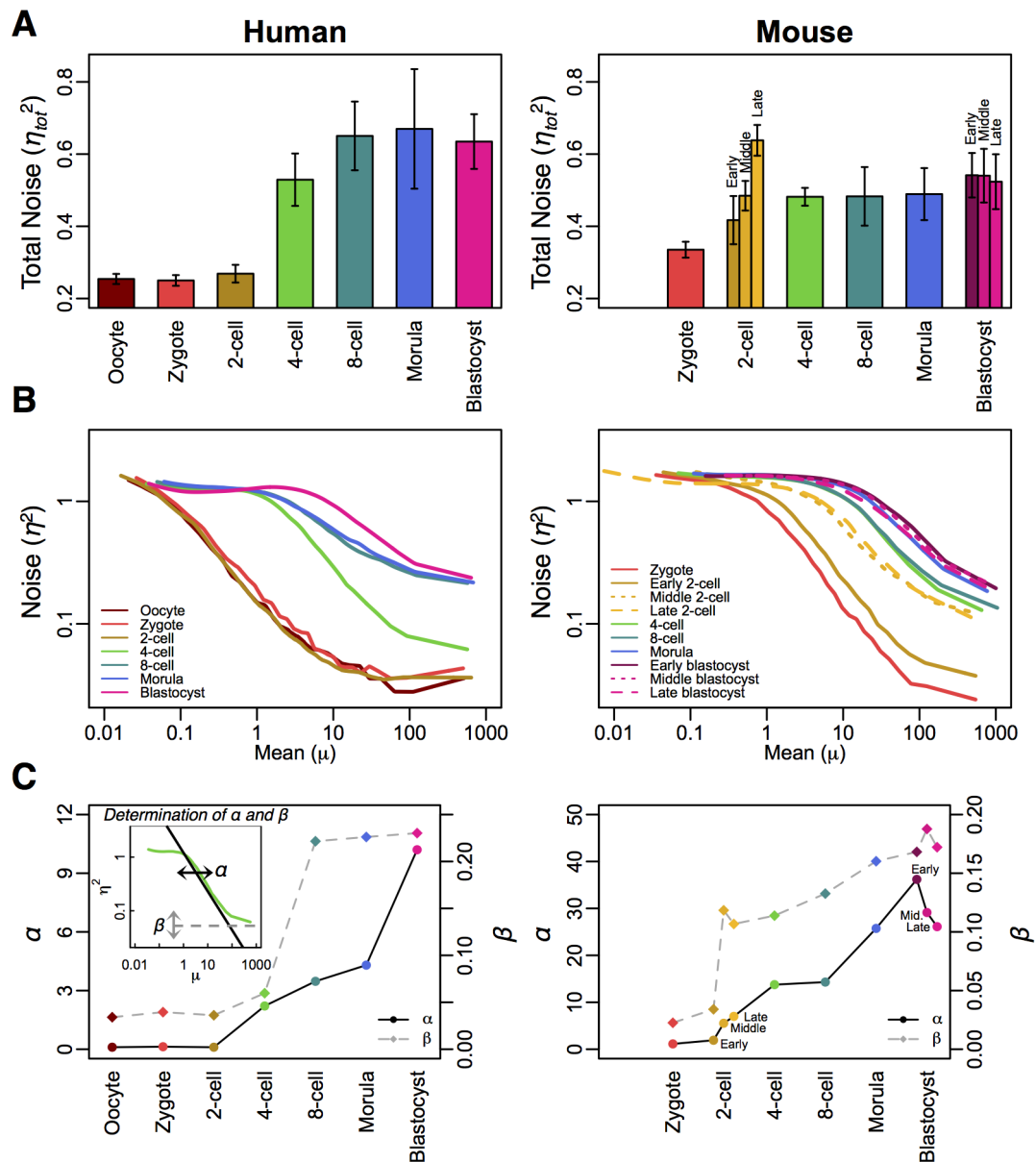


Figure 3 | Transcriptome-wide noise patterns. (A) Total noise (η_{tot}^2) of single cells for each development stage (average for m cells, error bars indicate 1 s.d.). (B) Noise (η^2) vs. mean (μ) expression patterns for each development stage. (C) Plots of α and β against cell stage. The insert illustrates how α and β are determined.

idealised condition, which does not contain noise due to extrinsic or other non-Poisson factors, such as technical noise^{30,31}.

To consider expression-independent noise in our simulations, such as technical and environmental variability, we included additive and multiplicative random white noise^{32,33} to our simulated expressions. As a result, we observed an increase of β , while α remained unchanged (Figure 4B, panel 2). This is because the white noises do not interfere with the transcriptional regulation.

Although none of the development stages fitted this pattern, we observed the expression-independent noise values at higher expression levels of early stages could be achieved by setting a certain threshold of random white noise, such as $\beta = 0.03$ for human and 0.02 for mouse (Supplementary Fig. S4 online). We, therefore, set these values as the level of extrinsic noise for all the subsequent simulations, assuming it remains unchanged across each cell type for a particular species.

Next, we increased the number of transcripts produced per activation event, which resulted in amplification of the scatter in expressions

(Supplementary Fig. S7 online), with corresponding increase in noise structure (Figure 4B, panel 3). Since ϕ controls transcriptional amplification, it is conceivable that increasing ϕ will proportionally increase α . However, β was not affected.

For human, the simulated noise structures matched experimental patterns of oocyte, zygote, and 2-cell stages for low ϕ and 4-cell stages for high ϕ (Figure 4C). For later stages, although higher values of ϕ improved the simulation results, as noted earlier, β could not be increased to fit the experimental patterns (Figure 4B, panel 4). For mouse, the trend of increasing over developmental stages is also observed (Figure 3B), however, the values are generally higher for early stages (Figure 4C and Supplementary Fig. S8 online). Like human data, the mouse simulations also did not match for the values of β (Supplementary Fig. S8 online).

To improve the simulation results, that is to specifically increase β values, we next explored the parameters governing bursty transcriptional dynamics: k_{on} and k_{off} . Previous experiments in mouse ES cells³⁴ suggest $0.1 \leq k_{on} \leq 2.5 \text{ h}^{-1}$ (median, 0.5 h^{-1}) and $3 \leq k_{off}$

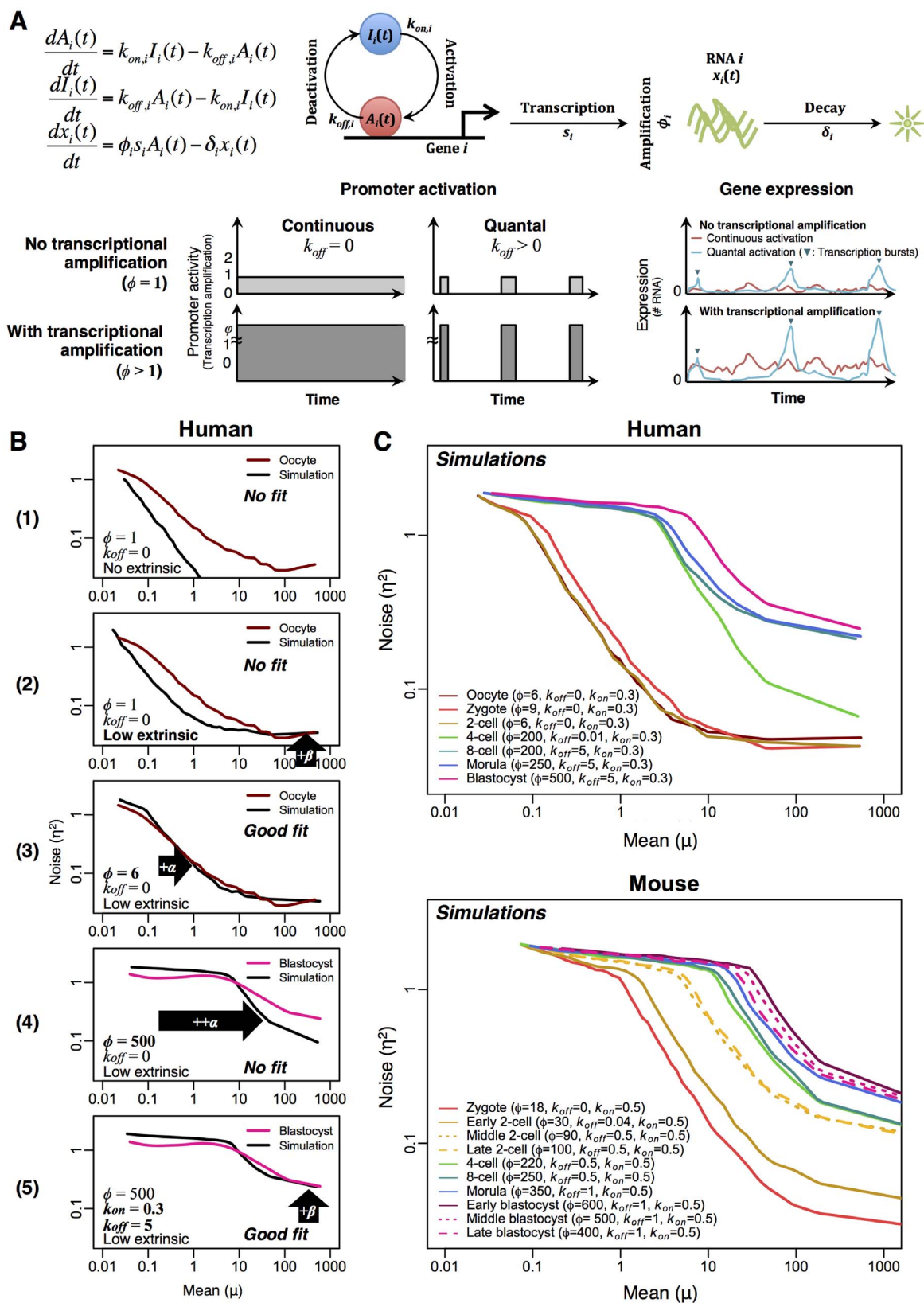


Figure 4 | Simulations of transcriptome-wide expressions. (A) Top: single cell transcriptional model. Bottom: promoter activity (left and middle panels) and gene expressions (right panels) for continuous and quantal simulations, without and with transcriptional amplification. (B) Simulations (panel 1, black curves) using continuous activation, no amplification and no extrinsic noise. Addition of extrinsic noise only (panel 2), and with low transcriptional amplification (panel 3) allowed better fit of simulations to early development stages patterns (e.g. oocyte, brown curve). Higher amplification with extrinsic noise (panel 4). Higher amplification and quantal activation with extrinsic noise (panel 5). Panel 5 shows a good fit with late stage developmental cells. (C) Simulated noise vs. mean expression patterns for all developmental stages in human (top) and mouse (bottom). Note that $k_{on} = 0.5$ was used for mouse simulations and $k_{on} = 0.3$ for human.



$\leq 200 \text{ h}^{-1}$ (median, 14 h^{-1}). To obtain β values that fit the human experimental patterns of later developmental stages, we initially set $k_{on} = 0.5$ and increased k_{off} to 5. However, the value for β obtained from simulations was still too low, and further increase of k_{off} did not produce major increase in β . We, therefore, reduced the value of k_{on} such that $k_{on} = 0.3$ to obtain a good fit (Figure 4B, panel 5). For mouse, $k_{on} = 0.5$ and $k_{off} > 0.5$ were required to fit the experimental β values. Again, further increases of k_{off} did not significantly increase β (Supplementary Fig. S8 online).

Overall, from these simulations, we concur that distinct transcriptional mechanisms govern transcriptome-wide expressions during embryonic development. Notably, we showed that α is governed solely by transcriptional amplification whereas β is controlled by both extrinsic noise and quantal transcriptional activation.

Discussion

Recent studies on single cells have shown that individual molecules (genes, proteins or metabolites), within a homogenous cell population, can display highly variable expression values^{1,2,6,7,35,36}. This variability has been linked to biological noise or the stochastic nature of molecular network regulations^{37–39}. However, there is a general lack in the investigation of global regulatory mechanisms at an omics-wide scale for single cell behaviour. Studying global properties has been instrumental in interpreting collective mechanisms of living organisms, for example, the innate immune response to invading pathogens¹³ or the attractor states of cell differentiation process⁴⁰. Here, to understand the global noise patterns of single developmental cells, we investigated RNA-seq transcriptome-wide expressions of oocytes to blastocysts in human and mouse.

Firstly, by studying the distribution of gene expressions between single cells, we observed the expression scatter increased from 2-cell to 4-cell stage onwards in both human and mouse (Figure 1). Next, we examined the Pearson correlation and Shannon entropy for each developmental stage. Again, we observed that expressions become more variable from the 2-cell stage (Figure 2). Subsequently, the global noise character of single cells was investigated by quantifying the squared coefficient of expression variations over mean expression values. Here, we observed clear transition of noise patterns occurring between 2-cell and 8-cell stage (Figure 3).

To understand the noise patterns, we developed a stochastic transcriptional model and estimated the parameter values to match each developmental cell pattern (Figure 4). From the model, we concur that the early developmental stages are mainly dominated by low transcriptional activity. For these stages, the number of transcripts produced per activation event, ϕ , is low. The lower overall transcription in oocytes and early zygote is consistent with i) transcriptional silencing and ii) stochastic degradation of maternal RNA that has been observed from oocytes to 4-cell stage in humans^{41,42}. Transcriptional silencing is likely due to chromatin condensation state that prevents transcriptional machinery from reaching gene promoters^{41,43,44}.

To track the gene expression profiles of common maternal^{45,46} and zygotic genes⁴⁵, we plotted their relative expressions (Supplementary Fig. S9 online), and found 2 and 3 major temporal clusters, respectively, for 137 maternal and 116 zygotic highly expressed genes (Figure 5 and Table S1 online). Notably, the maternal genes (e.g. *Cdh3*, *Dppa5*, *Mos*, *Npm2*, *Zp1*, *Zp2*) showed dominant decay profiles of RNA expressions, indicating lack of transcription process (Figure 5A). However, the zygotic genes and genes expressed in embryonic stem cells^{47,48} (e.g. *Klf4*, *Lin28a*, *Myc*, *Nanog*, *Pou5f1*, *Sox2*) showed transcriptional process significantly increasing after the 4-cell stage (Figure 5B). The high transcriptions can be due to instructive signaling pathways, or multiple rounds of transcription reinitiation by RNA polymerase^{26,49,50}. The observation of high transcriptome-wide noise for the middle stage developmental cells indicates the generation of heterogeneity in gene expressions between

individual cells. Such heterogeneity has been shown to be necessary for cell fate diversifications^{1,6}.

For the later stage developmental cells, on top of high transcriptional process, the cells possess quantal activation of most transcription factors, or are subject to more extrinsic variability such as phenotypic diversity among individual cells. These factors increase the general expression scatter and noise levels. However, investigating expression-independent random noise in our single cell transcriptional model simulations suggest that the levels of extrinsic and/or technical noise in our RNA-Seq data for all cells are relatively low ($\eta_{tot}^2 \sim 0.25$). That is, the relatively high levels of noise for later stages stem from quantal activation rather than technical biases, or in certain cases, such as blastocyst cells, may result from phenotypic variability, as blastocysts consist of different cellular subtypes. Conversely, since phenotypic variability among more homogenous 8-cell stage is similar to blastocyst (Figure 1), we believe that quantal promoter activation is crucial for the increase of noise scatter along developmental stages. Notably, such quantal promoter activation has been noted to occur for single cell organisms such as *E. coli*³, and has been shown to be important for the cell fate decision of *B. subtilis*⁶.

Overall, our investigations on the transcriptome-wide expressions of the early mammalian developmental stages reveal increasing variability and noise patterns across the mammalian development process. These data suggest different stages of the cell differentiation process can be better understood by investigating the transcriptome-wide noise patterns. In conclusion, our systemic approach provides novel insights into the transcriptome-wide expression and noise patterns for development cells, and the underlying nature of the transcriptional mechanisms.

Methods

Single cell datasets. Single cell RNA-Seq datasets were downloaded from Gene Expression Omnibus (GEO) database from previously published data for 7 human⁸ (GSE36552) and 10 mouse⁹ (GSE45719) developmental stages. All datasets were obtained through Illumina sequencing systems. Each dataset contains the RPKM values (Reads Per Kilobase Mapped) for $n \sim 20,000$ genomic features, which is proportional to the number of transcripts of coding and non-coding genes (and splicing variants).

Correlation analyses. To quantify transcriptome variability we utilised correlation metrics, which are widely employed to compare global relationships between high-throughput datasets^{10–14}.

The Pearson correlation between two transcriptomes, X and Y , containing n gene expressions, is obtained by $R(X, Y) = \frac{\sum_{i=1}^n (x_i - \mu_X)(y_i - \mu_Y)}{(\sigma_X \sigma_Y)}$, where x_i and y_i are the i^{th} observation in the vectors X and Y respectively, μ_X and μ_Y are the average values of each transcriptome, and σ_X and σ_Y are the corresponding standard deviations. The Spearman correlation coefficient between transcriptomes X and Y is defined by $\rho(X, Y) = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^n (r_{x,i} - r_{y,i})^2$, where $r_{x,i}$ and $r_{y,i}$ are the ranks of the i^{th} observations x_i and y_i in vectors X and Y respectively. Both correlations were computed using the *cor* function of R *stats* package (<http://www.r-project.org/>).

Pearson and Spearman correlations respectively measure linear and non-linear monotonic relationships between two vectors, where $R = 1$ (respectively $\rho = 1$) if the two vectors are identical, and $R = 0$ (respectively $\rho = 0$) if there is no linear or monotonic relationships between the vectors. However, both metrics do not detect other non-linear relationships, therefore null correlation values do not imply statistical independence. To obtain a more stringent measure of statistical dependence between transcriptomes X and Y , we used Distance Correlation¹⁵ (*dCor*), where $dCor(X, Y) = 0$ if and only if the two vectors are statistically independent. Maximum Information Coefficient¹⁶ (*MIC*) can also be used to detect other types of non-linear associations between the transcriptomes, by calculating mutual information of the vectors using an automated non-parametric approach for binning. The computation of Distance Correlation values was performed using the *dcor* function of the R *energy* package, and the computation of Maximum Information Coefficient with the *mine* function of the R *minerva* package, with default parameters.

Entropy analysis. To assess the diversity of single cell transcriptomes, we used Shannon entropy^{17–19}. Shannon entropy measures the disorder of a high-dimensional system, where higher values indicate increasing disorder. Entropy of each single cell transcriptome, X , is defined as $H(X) = -\sum_{i=1}^n p(x_i) \log_2 p(x_i)$, where $p(x_i)$ represents the probability of gene expression value $x = x_i$. Entropy values were obtained through binning approach and the number of bins, $b = 26$, was determined

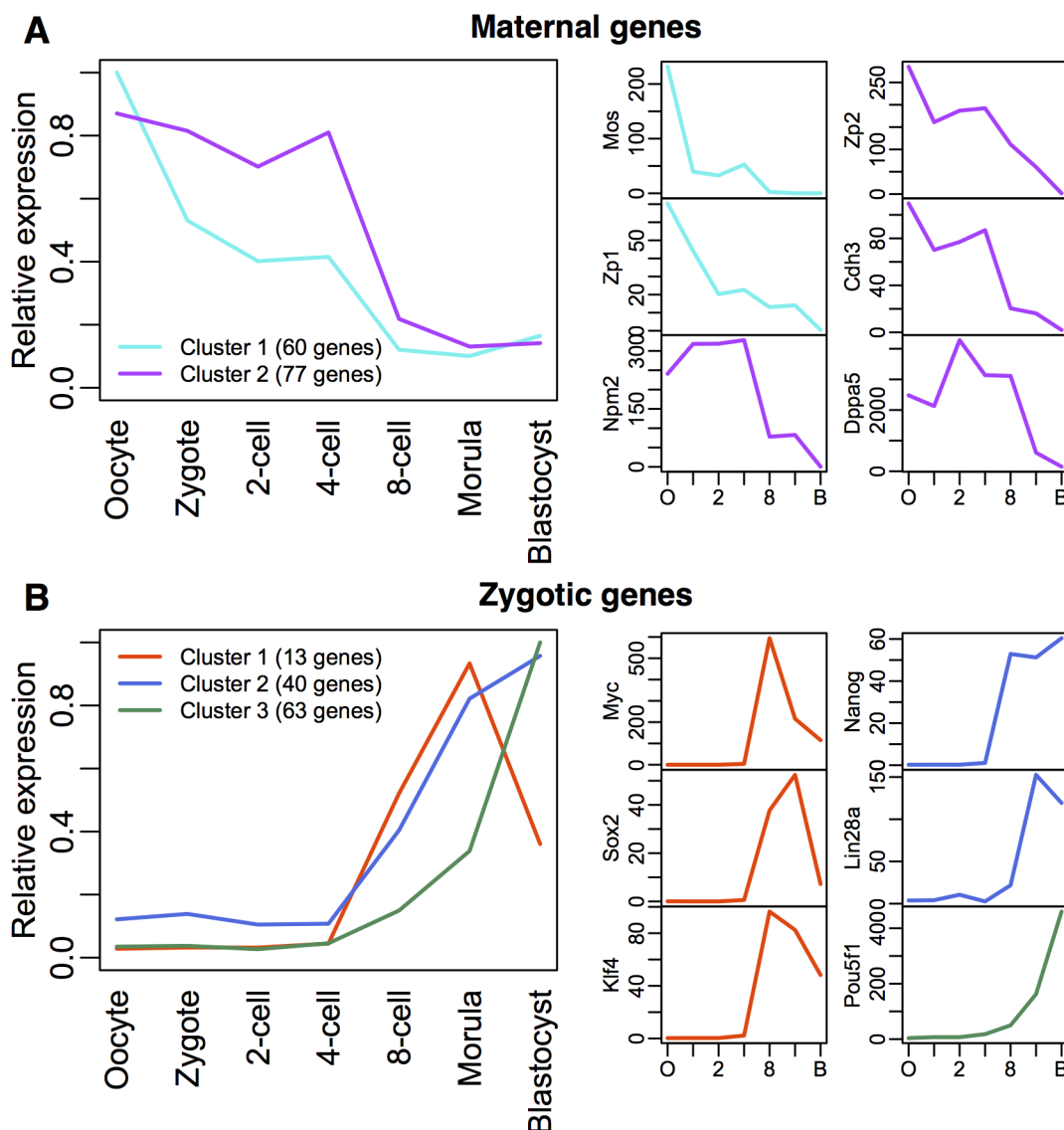


Figure 5 | Gene expression profiles of maternal and zygotic genes during development. Average gene expression profiles of (A) maternal (2 clusters) and (B) zygotic (3 clusters) genes during human embryo development. Clusters were obtained using k-means clustering. From the initial lists of genes obtained from Xue et al.⁴⁵ (Table S1), we retained maternal genes with expression values reaching peak before 4-cell stages and zygotic genes with expression values reaching peak from 8-cell stage onwards. Right panels show the expression values in all development stages of individual maternal genes (*Cdh3*, *Dppa5*, *Mos*, *Npm2*, *Zp1* and *Zp2*; obtained from Kocabas et al.⁴⁶) expressed in oocytes, and zygotic genes involved in later embryonic stages and stem cells (*Myc*, *Klf4*, *Pou5f1* (*Oct4*), *Sox2*, *Lin28a* and *Nanog*; obtained from Takahashi et al.⁴⁷ and Yu et al.⁴⁸).

from the data using Doane's rule⁵¹, such as $b(X) = 1 + \log_2 n + \log_2(1 + |g_X|/\sigma_g)$, where g_X is the skewness of the expression distribution of each sample, and $\sigma_g = \sqrt{6(n-2)/(n+1)(n+3)}$. The computation of entropy values was performed using the maximum likelihood implementation (*entropy.empirical*) of the R *entropy* package.

Transcriptome-wide average noise. To quantify between single cells' expressions scatter, we computed Transcriptome-wide average noise for each cell type, defined as $\eta_{tot}^2 = \frac{1}{n} \sum_{i=1}^n \eta_i^2$, where n is the number of genes and η_i^2 is the pairwise noise of the i^{th} gene (variability between any two cells), defined as $\eta_i^2 = \frac{2}{m(m-1)} \sum_{j=1}^{m-1} \sum_{k=j+1}^m \eta_{ijk}^2$, where m is the number of cells and η_{ijk}^2 is the expression noise of the i^{th} gene, defined by the variance divided by the squared mean expression¹⁸ in the pair of cells (j,k), such as $\eta_{ijk}^2 = \sigma_{ijk}^2 / \mu_{ijk}^2$, where $\mu_{ijk} = (x_{ij} + x_{ik})/2$ is the average value of the i^{th} gene in the pair of single cells (j,k), and $\sigma_{ijk}^2 = (x_{ij} - x_{ik})^2 / 2$, is the corresponding variance.

Transcriptome-wide noise patterns. To elucidate transcriptome-wide noise patterns, we sorted the transcriptome into groups of $w = 500$ genes from low to high expression values for each pair of cells (j,k). We formed $G = n/w$ groups, and obtained

the average gene expression of each group for each pair of cells, $\mu_{jk}(g) = \frac{1}{w} \sum_{i \in G} \mu_{ijk}$, and the average gene expression noise of all genes contained in the group,

$\eta_{jk}^2(g) = \frac{1}{w} \sum_{i \in G} \eta_{ijk}^2$, for each pair of cells. We finally obtained characteristic whole transcriptome noise patterns $\mu(g) = \frac{2}{m(m-1)} \sum_{j=1}^{m-1} \sum_{k=j+1}^m \mu_{jk}(g)$ and

$\eta^2(g) = \frac{2}{m(m-1)} \sum_{j=1}^{m-1} \sum_{k=j+1}^m \eta_{jk}^2(g)$ by averaging the patterns of all pairs of cells.

For simplicity, we used $\mu = \mu(g)$ and $\eta^2 = \eta^2(g)$ in the main and following texts. As a result, the reported curve for each stage is the average pattern of all single cells pairs. We fitted noise as function of mean expressions, $\eta^2 = f(\mu)$,

$$\eta^2 = \begin{cases} 2, & \mu < \alpha/(2-\beta) \\ \alpha/\mu + \beta, & \mu \geq \alpha/(2-\beta) \end{cases} \quad (1)$$

using nonlinear least squares piecewise curve fitting with the *nls* R function, to obtain the values of α and β for each pair of cells in all development stages.

Transcriptome simulations. Each transcriptome model consists of a set of 20,000 genes, and each gene's expression is obtained from the stochastic simulation⁵² of a



telegraph process²⁵, such as the system of rate equations that governs gene expression x_i of the i^{th} gene is

$$\begin{cases} \frac{dA_i(t)}{dt} = k_{on,i}I_i(t) - k_{off,i}A_i(t) \\ \frac{dI_i(t)}{dt} = k_{off,i}A_i(t) - k_{on,i}I_i(t) \\ \frac{dx_i(t)}{dt} = \varphi_i s_i A_i(t) - \delta_i x_i(t) \end{cases} \quad (2)$$

where the promoter activation of each gene is defined by a 2-state model, active (A_i) and inactive (I_i) states, and the transition rates between states are defined by two parameters, $k_{on,i}$ and $k_{off,i}$. s_i is the transcription rate when the promoter is active, φ_i is the transcription amplification factor and δ_i is the degradation rate constant. The distribution of RNA degradation rate constants, δ_i , was obtained from transcriptome-wide RNA half-lives data in differentiating mouse ES cells²⁸ and human B cells²⁹, and fitted to a lognormal distribution with mean (log scale) and standard deviation parameters, $\mu = -2.24$ and $\sigma = 0.61$ for mouse and $\mu = -1.58$ and $\sigma = 0.73$ for human (Supplementary Fig. S6 online).

RNA transcription rates, s_i , were estimated from the same data^{28,29}, such as $s_i = x_i \delta_i / (k_{on,i} + k_{off,i}) / (\varphi_i k_{on,i})$, where x_i is the gene expression value (read counts) of the i^{th} gene, taken from a Zipf's law distribution with exponent $r = 0.8$ to fit the expression range in our data (Supplementary Fig. S6 online, inserts). As a result, we fitted the estimated values using a lognormal distribution for $x_i \delta_i$, with parameters $\mu = 0.17$ and $\sigma = 2.67$ for mouse and $\mu = 0.73$ and $\sigma = 2.53$ for human (Supplementary Fig. S6 online). Since the model simulates integer read counts expressions, to obtain corresponding RPKM values, we multiplied our simulated values by a normalization constant, Γ , defined as the average ratio between RPKM values of all genes and corresponding number of reads ($\Gamma = 0.03$ for the human dataset, and 0.12 for mouse).

To generate transcriptome-wide expressions, we assigned different values for degradation rate constant, δ_i , and transcription rate, s_i , for each gene, and set the values of φ_i , $k_{on,i}$ and $k_{off,i}$ identical for all genes. To test variable values of φ_i , $k_{on,i}$ and $k_{off,i}$ for each gene, we generated transcriptome-wide expressions using a Poisson distribution for φ_i with parameter $\lambda = \varphi_i$. Our result showed no noticeable difference in the patterns between fixed value or Poisson distributed φ_i .

Similarly we simulated transcriptomes with promoter activation kinetics that vary between genes. We estimated the distributions of $k_{on,i}$ and $k_{off,i}$ from previously observed experimental distributions of promoter 'on' and 'off' time intervals³⁴ ($\tau_{on,i}$ and $\tau_{off,i}$). From the data, we observed the distributions of $\tau_{on,i}$ and $\tau_{off,i}$ could be approximated by an exponential distribution with parameter $\lambda = 10$, and a lognormal distribution with parameters $\mu = 0.69$ and $\sigma = 1$ respectively. Since $k_{on,i} = 1/\tau_{off,i}$ and $k_{off,i} = 1/\tau_{on,i}$ ³⁵, we found median $k_{on,i} = 0.5 \text{ h}^{-1}$ ($0.1 \sim 2.5 \text{ h}^{-1}$ range) and median $k_{off,i} = 14 \text{ h}^{-1}$ ($3 \sim 200 \text{ h}^{-1}$). We then compared the simulations using fixed $k_{on,i} = 0.5$ and $k_{off,i} = 14$ or variable $k_{on,i}$ and $k_{off,i}$ and found no significant change in the patterns.

To account for non-intrinsic variations, we introduced different levels of additive and multiplicative white (Gaussian) noise to the simulated data^{32,33}. Additive noise is achieved by adding a different random value to each gene in each cell, and multiplicative noise, by multiplying all gene expressions in the same cell by a random number such as, $\log x_{ij} = \omega_j (\log x_{ij} + \varepsilon_{ij})$, where x_{ij} is the expression of i^{th} gene in the j^{th} cell including non-intrinsic noise, ε_{ij} represents additive noise for the i^{th} gene in the j^{th} cell, and ω_j is the multiplicative noise that affects the j^{th} cell. ε_{ij} is chosen from a normal distribution ($\varepsilon_{ij} \in N(0, \varepsilon \log x_{ij})$) with mean and standard deviation parameters, $\mu = 0$ and $\sigma = \varepsilon \log x_{ij}$, where ε represents the level of additive noise. ω_j is log-normal distributed, such as $\omega_j \in N(0, \log \omega)$ with parameters $\mu = 0$ and $\sigma = \omega$ respectively, and ω is the level of multiplicative noise. We used $\varepsilon = 0.08$ and $\omega = 0.1$ to simulate non-intrinsic noise of human dataset (equivalent to $\beta \sim 0.03$), and $\varepsilon = 0.06$ and $\omega = 0.1$ for mouse ($\beta \sim 0.02$).

- Chang, H. H., Hemberg, M., Barahona, M., Ingber, D. E. & Huang, S. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature* **453**, 544–547 (2008).
- Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183–1186 (2002).
- Eldar, A. & Elowitz, M. B. Functional roles for noise in genetic circuits. *Nature* **467**, 167–173 (2010).
- Selvarajoo, K. Understanding multimodal biological decisions from single cell and population dynamics. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **4**, 385–399 (2012).
- Selvarajoo, K. Uncertainty and certainty in cellular dynamics. *Front. Genet.* **4**, 68 (2013).
- Maamar, H., Raj, A. & Dubnau, D. Noise in gene expression determines cell fate in *Bacillus subtilis*. *Science* **317**, 526–529 (2007).
- Raj, A., Rifkin, S. A., Andersen, E. & van Oudenaarden, A. Variability in gene expression underlies incomplete penetrance. *Nature* **463**, 913–918 (2010).
- Yan, L. *et al.* Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.* **20**, 1131–1139 (2013).
- Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**, 193–196 (2014).
- Lo, A. W. & MacKinlay, A. C. Stock market prices do not follow random walks: evidence from a simple specification test. *Rev. Financ. Stud.* **1**, 41–66 (1988).

- Stewart, T. R. A decomposition of the correlation coefficient and its use in analyzing forecasting skill. *Weather Forecast.* **5**, 661–666 (1990).
- Amati, L. *et al.* Measuring the cosmological parameters with the Ep, i-Eiso correlation of Gamma-Ray Bursts. *Mon. Not. R. Astron. Soc.* **391**, 577–584 (2008).
- Tsuchiya, M. *et al.* Emergent genome-wide control in wildtype and genetically mutated lipopolysaccharides-stimulated macrophages. *PLoS One* **4**, e4905 (2009).
- Tsuchiya, M., Piras, V., Giuliani, A., Tomita, M. & Selvarajoo, K. Collective dynamics of specific gene ensembles crucial for neutrophil differentiation: the existence of genome vehicles revealed. *PLoS One* **5**, e12116 (2010).
- Székely, G. J., Rizzo, M. L. & Bakirov, N. K. Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35**, 2769–2794 (2007).
- Reshef, D. N. *et al.* Detecting novel associations in large data sets. *Science* **334**, 1518–1524 (2011).
- Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423, 623–656 (1948).
- Schneider, T. D., Stormo, G. D., Gold, L. & Ehrenfeucht, A. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* **188**, 415–431 (1986).
- Weiss, O., Jiménez-Montaño, M. A. & Herzel, H. Information content of protein sequences. *J. Theor. Biol.* **206**, 379–386 (2000).
- Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
- Bar-Even, A. *et al.* Noise in protein expression scales with natural protein abundance. *Nat. Genet.* **38**, 636–643 (2006).
- Piras, V., Tomita, M. & Selvarajoo, K. Is central dogma a global property of cellular information flow? *Front. Physiol.* **3**, 439 (2012).
- Ramsköld, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30**, 777–782 (2012).
- Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10**, 1093–1085 (2013).
- Peccoud, J. & Ycart, B. Markovian modeling of gene-product synthesis. *Theor. Popul. Biol.* **48**, 222–234 (1995).
- Thattai, M. & van Oudenaarden, A. Intrinsic noise in gene regulatory networks. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 8614–8619 (2001).
- Lin, C. Y. *et al.* Transcriptional amplification in tumor cells with elevated c-Myc. *Cell* **151**, 56–67 (2012).
- Sharova, L. V. *et al.* Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells. *DNA Res.* **16**, 45–58 (2009).
- Friedel, C. C., Dölken, L., Ruzsics, Z., Kosziniowski, U. H. & Zimmer, R. Conserved principles of mammalian transcriptional regulation revealed by RNA half-life. *Nucleic Acids Res.* **37**, e115 (2009).
- Swain, P. S., Elowitz, M. B. & Siggia, E. D. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 12795–12800 (2002).
- Hilfinger, A. & Paulsson, J. Separating intrinsic from extrinsic fluctuations in dynamic biological systems. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 12167–12172 (2011).
- Nussbaum, M. Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Stat.* **24**, 2399–2430 (1996).
- Furusawa, C. & Kaneko, K. Epigenetic feedback regulation accelerates adaptation and evolution. *PLoS One* **8**, e61251 (2013).
- Suter, D. M. *et al.* Mammalian genes are transcribed with widely different bursting kinetics. *Science* **332**, 472–474 (2011).
- Taniguchi, Y. *et al.* Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**, 533–538 (2010).
- Ibáñez, A. J. *et al.* Mass spectrometry-based metabolomics of single yeast cells. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 8790–8794 (2013).
- Pedraza, J. M. & van Oudenaarden, A. Noise propagation in gene networks. *Science* **307**, 1965–1969 (2005).
- Eldar, A. *et al.* Partial penetrance facilitates developmental evolution in bacteria. *Nature* **460**, 510–514 (2009).
- Hornung, G. *et al.* Noise-mean relationship in mutated promoters. *Genome Res.* **22**, 2409–2417 (2012).
- Huang, S., Eichler, G., Bar-Yam, Y. & Ingber, D. E. Cell fates as high-dimensional attractor states of a complex gene regulatory network. *Phys. Rev. Lett.* **94**, 128701 (2005).
- Braude, P., Bolton, V. & Moore, S. Human gene expression first occurs between the four- and eight-cell stages of preimplantation development. *Nature* **332**, 459–461 (1988).
- Tadros, W. & Lipshitz, H. D. The maternal-to-zygotic transition: a play in two acts. *Development* **136**, 3033–3042 (2009).
- Debey, P. *et al.* Competent mouse oocytes isolated from antral follicles exhibit different chromatin organization and follow different maturation dynamics. *Mol. Reprod. Dev.* **36**, 59–74 (1993).
- De La Fuente, R. Chromatin modifications in the germinal vesicle (GV) of mammalian oocytes. *Dev. Biol.* **292**, 1–12 (2006).
- Xue, Z. *et al.* Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* **500**, 593–597 (2013).
- Kocabas, A. M. *et al.* The transcriptome of human oocytes. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 14027–14032 (2006).
- Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human Fibroblasts by defined factors. *Cell* **131**, 861–872 (2007).



48. Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318**, 1917–1920 (2007).
49. Dieci, G. & Sentenac, A. Detours and shortcuts to transcription reinitiation. *Trends Biochem. Sci.* **28**, 202–209 (2003).
50. Chubb, J. R., Trcek, T., Shenoy, S. M. & Singer, R. H. Transcriptional pulsing of a developmental gene. *Curr. Biol.* **16**, 1018–1025 (2006).
51. Doane, D. P. Aesthetic frequency classification. *Am. Stat.* **30**, 181–183 (1976).
52. Gibson, M. A. & Bruck, J. Efficient exact stochastic simulation of chemical systems with many species and many channels. *J. Phys. Chem. A.* **104**, 1876–1889 (2000).
53. Dar, R. D. *et al.* Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 17454–17459 (2012).

Acknowledgments

This work was supported by the Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research J13108 (K.S.), and Tsuruoka City, Yamagata Prefecture.

Author contributions

V.P. and K.S. analysed and interpreted the data, developed the gene expression model, and performed the simulations. V.P., M.T. and K.S. wrote and reviewed the manuscript.

Additional information

Supplementary information accompanies this paper at <http://www.nature.com/scientificreports>

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Piras, V., Tomita, M. & Selvarajoo, K. Transcriptome-wide Variability in Single Embryonic Development Cells. *Sci. Rep.* **4**, 7137; DOI:10.1038/srep07137 (2014).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder in order to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>