SCIENTIFIC REPORTS
nature research

OPEN

Persistent features of intermittent transcription

Michael Wilkinson[1,2*], Spyros Darmanis[1], Angela Oliveira Pisco[1] & Greg Huber[1]

Single-cell RNA sequencing is a powerful tool for exploring gene expression heterogeneity, but the results may be obscured by technical noise inherent in the experimental procedure. Here we introduce a novel parametrisation of sc-RNA data, giving estimates of the probability of activation of a gene and its peak transcription rate, which are agnostic about the mechanism underlying the fluctuations in the counts. Applying this approach to single cell mRNA counts across different tissues of adult mice, we find that peak transcription levels are approximately constant across different tissue types, in contrast to the gene expression probabilities which are, for many genes, markedly different. Many genes are only observed in a small fraction of cells. An investigation of correlation between genes activities shows that this is primarily due to temporal intermittency of transcription, rather than some genes being expressed in specialised cell types. Both the probability of activation and the peak transcription rate have a very wide ranges of values, with a probability density function well approximated by a power law. Taken together, our results indicate that the peak rate of transcription is a persistent property of a gene, and that differences in gene expression are modulated by temporal intermittency of the transcription.

Using nucleic acid polymerase technologies, it has become possible to quantify mRNA transcription processes with ever greater sensitivity[1]. Recently, it has become possible to obtain quantitatively reliable counts of individual gene transcripts from single cells[2]. This technology has the potential to reveal new insights into the mechanisms and organization of transcription processes. This paper reports on a statistical analysis of the *Tabula Muris* dataset of mRNA transcriptions from individual mouse cells, derived from a range of distinct tissue types[3].

The number of 'reads' of mRNA from individual cells is highly variable, with many cells yielding a zero count for a particular gene, while a few cells from the same tissue type might yield hundreds or even thousands of reads. This variation is often ascribed to 'dropouts', viewed as a technical consequence of the stochastic nature of the DNA polymerization reaction. However, in the case of the *Tabula Muris* dataset, we find that the variability of the counts is markedly different between different genes, and for many genes the counts are much more variable than those of exogenous RNA sequences which are 'spiked in' with known concentrations. The variability cannot, therefore, be explained solely as a technical artifact, and we should therefore consider other, biological interpretations.

In this paper we use single-cell mRNA counts to estimate the probability p that each gene is being expressed in a given cell. These probabilities vary greatly, and we find that a significant fraction of genes are expressed with very low probability. There are at least two possible explanations for the wide variability of the gene *expression probability* p, illustrated schematically in Fig. 1:

- **C**ase A: the cell population could be inhomogeneous, with cells from a given tissue differentiating into many different types, which express the same set of genes continuously. A small value of p is a consequence of a gene being expressed in a rare, specialized cell type. In the schematic of Fig. 1(a) genes with labels i and j are expressed by different types of rare cells.
- **C**ase B: It could be that the activity of a cell within a given tissue-type population is time-dependent. In this case a small value of p is interpreted as an indication that the gene is turned off for most of the time. In the schematic illustration Fig. 1(b), a cell which is actively expressing a gene has a count equal to a *peak activity* α, but for most of the time the gene is not being transcribed.

It is desirable to distinguish clearly between these possibilities. This can be done by considering the coefficients of correlation for expression of different genes in the same cell. The evidence (discussed in the section on *gene activity correlations* below) strongly favors case B above as a model for explaining the occurrence of genes which are expressed with a low probability.

[1]Chan Zuckerberg Biohub, 499 Illinois Street, San Francisco, CA, 94158, USA. [2]School of Mathematics and Statistics, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK. *email: michael.wilkinson@czbiohub.org

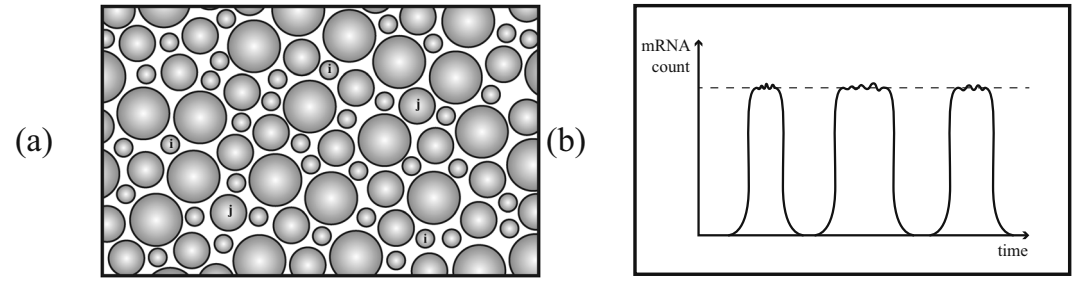

Figure 1. Schematic contrasting two explanations for genes being observed with low probability in single cell mRNA counts. (**a**) Genes *i* and *j* may be transcribed continuously in rare, specialized cell types. (**b**) Genes may be intermittently active in all cells.

Earlier studies of the time dependence of the expression of single genes have shown evidence that transcription of mRNA occurs in 'bursts'. This comes from direct observation of the time-dependence of mRNA transcription[4,5]. It has been remarked that indirect evidence for bursting also comes from the observation that the variance of single-cell mRNA reads is typically much larger than the mean[5,6].

'Bursty' transcription is regarded as being a consequence of a quite specific type of stochastic process, involving binding and un-binding of transcription factors from the DNA[7–10]. More recently, by fitting the statistics of single-cell mRNA reads to stochastic models of bursty transcription[11,12], burst sizes and burst frequencies have been ascribed to individual genes[13]. The work of Larsson *et al.*[13] emphasizes correlations between the kinetic parameters of the bursting model and aspects of the structure of the gene (such as its size) or of its promoters (such as TATA elements).

The stochastic model used in these studies has four distinct transition rates, and in principle, measurement of the distribution of counts can be used to determine three ratios of these transition rates. The measured count distributions are, however, subject to large uncertainties because of the instability of the polymerisation reactions and other technical factors, and the count distributions of a given gene in a given tissue can show marked differences between different experimental protocols. Determining the parameters is also made more difficult because, in some regions of parameter space, fitting the three parameters of the model is an ill-conditioned problem. A further complication is that the activity of different genes may be correlated, implying that a model in which transcription of different genes is determined by independent stochastic processes may be too simple. For these reasons it is desirable to adopt a more direct approach to characterizing gene activity, which does not depend upon specific mechanisms underlying the fluctuations, and which is agnostic about the form of the distribution of counts. Our approach estimates two parameters, namely the probability p that a gene is active, and the typical number of transcripts which are present when the gene is active, α. It is complementary to the approach of fitting parameters to the stochastic model, in that it is simple to implement and does not assume any specific mechanism for the fluctuating gene activity. We explain the connection between our parameters and those of the stochastic model in an appendix.

Because 'bursting' transcription has an association with a particular stochastic model which may not be the only viable explanation of the phenomena, we shall refer to *intermittent* transcription in this text. Further information, such as investigation of correlations between gene expression, are required to distinguish between intermittent transcription and gene expression by specialist cells (that is, between cases A and B above). The analysis of the *Tabula Muris* dataset presented here supports the view that intermittent transcription is an ubiquitous phenomenon, extending to genes which are only expressed at very low levels. The fact that mRNA counts of the intermittent genes are highly variable implies that genes are being turned on and off very slowly, on a timescale much longer than the lifetime of mRNA (which appears to be at least one hour in most mammalian cells[14]).

Because the *Tabula Muris* data set contains information from a range of different tissue types, we are able to assess the effects of cellular differentiation on gene expression. We find evidence that the peak rate of transcription, α, is a *persistent* attribute of a gene, which takes similar values in all of the tissues that were surveyed. The probability of transcription is found to be much more variable. These observations are consistent with gene expression being controlled by turning genes on and off, rather than via continuous regulation of their rate of transcription.

Because we have data for a very large number of genes (approximately 2×10^4 genes are included in the database), we can describe the statistics of gene attributes by probability distributions, for example we could introduce a probability density function (PDF) $P(\alpha)$ for the gene activity, α, such that the fraction of genes which have activity in a small range between α and $\alpha + \delta\alpha$ is $P(\alpha)\delta\alpha$. Both the peak transcription level α and the probability of expression p vary over a large range. We present evidence that the probability density of α is well approximated by a power law. The probability density of the gene expression probability varies markedly between different tissues. In the concluding section we speculate about whether intermittent transcription is necessarily a stochastic phenomenon, as is frequently supposed[8–10], and why it might confer advantages in organizing activity within a cell.

A Two-Parameter Characterization of Genes

The Tabula Muris data. The *Tabula Muris* dataset[15] combines single-cell mRNA count data from two different technologies: microfluidic droplet-based 3′-end counting, which provides a survey of thousands of cells per organ at relatively low coverage, and fluorescence-activated cell sorting (FACS)-based full length transcript

analysis, which provides higher sensitivity and coverage[3]. The 'droplet' data uses unique molecular identifier (UMI) sequences to label individual mRNA molecules, providing a precise assay of the number of mRNA molecules from a given gene that have been amplified. The 'FACS' data set records the number of reads after amplification, without counting the number of individual molecules which have been captured. We processed both data set finding that, while not quantitatively equivalent, they are sufficiently consistent to justify our qualitative conclusions.

The FACS-based dataset lists the counts of mRNA molecules obtained from individual cells, for 23,433 protein-coding genes (as annotated in the dataset), obtained from samples of 17 different tissues: the number of sampled cells ranges from 866 (kidney) to 6,007 (heart). In addition to the endogenous mRNA, the samples from each cell were 'spiked' with known concentrations of exogenous mRNA, with 96 different combinations of sequences and concentrations. The dataset also contains the number of reads for each of these 'spike-in' sequences. The droplet data was obtained from samples of 12 different tissue types: there were a total of 55,638 cells, ranging from 624 heart cells to 11,258 trachea cells.

In the case of the FACS data, we only included genes where more than 500 total counts were recorded, neglecting counts of less than 10 in individual cells. In the case of the droplet data, we applied a quality threshold of at least 1000 total reads and at least 500 genes detected.

Gene parameters. The data lists the number of reads M_{ijk} of mRNA for a gene with index i, in a cell with index j, from a tissue with index k. The polymerization process is inherently unstable and the experimental parameters may not be exactly the same for all cells. It was found that the total count for a cell, $\sum_i M_{ijk}$, varies over a large range (approximately two decades). For this reason, we considered 'normalizing' the counts by dividing M_{ijk} by the total number of counts for each cell (which would be appropriate if our studies were aimed, for example, at characterizing differences between different tissue types). However, evidence from the 'spike-in' sequences salted into the FACS data (discussed section on *characterization of spiked-in sequences* below) indicates that most of the variability of the total count for individual cells is real, rather than a technical artifact. For this reason, we did not normalize the counts.

The count data was processed to produce two statistics for each gene, denoted by α and p. The α variable is a measure of the peak level of transcription of a gene, and the p variable characterizes the probability that a given cell will be expressing that gene. If N_k is the number of cells for tissue type k, for a given gene with index i, we calculate the mean μ_{ik} and variance σ_{ik}^2 of the counts M_{ijk}, defined by

$$\mu_{ik} = \frac{1}{N_k}\sum\nolimits_j M_{ijk}\ , \quad \sigma_{ik}^2 = \frac{1}{N_k}\sum\nolimits_j M_{ijk}^2 - \mu_{ik}^2. \tag{1}$$

From the means and variances we can construct the quantities

$$p_{ik} \equiv \frac{\mu_{ik}^2}{\sigma_{ik}^2}. \tag{2}$$

Let us consider what value of p_{ik} would be expected according to a model where the transcription process is intermittent, in the sense that it is either 'on', with a small probability p, or else 'off', and where the 'on' state results in a count equal to α. According to this model, the mean and variance would be $\mu = \alpha p$ and $\sigma^2 = p(1-p)\alpha^2$, so that if $p \ll 1$ then $p \approx \mu^2/\sigma^2$, in agreement with Eq. (2). This indicates that if the transcription occurs intermittently, with the probability of being 'on' being $p \ll 1$, then p_{ik} is a measure of the probability of a cell expressing gene i in tissue k at a significant rate. Note that Eq. (2) was motivated by a simple model, under the assumption that p is small. We remark that the reciprocal quantity σ^2/μ^2 has previously been used as an estimator for 'noise' in studies of protein transcription[16].

If the number of counts when a gene is active is α, then the mean count is equal to $\mu = \alpha p$. Given an estimate of p, Eq. (2), we can then estimate α by writing $\alpha = \mu/p$. This indicates that the quantity

$$\alpha_{ik} \equiv \frac{\sigma_{ik}^2}{\mu_{ik}} \tag{3}$$

is a measure of the peak level of transcription of gene i in tissue type k. Thus every gene i in tissue k can be characterized by two parameters: a probability that it is active, p_{ik}, and a peak activity level α_{ik}, across different tissue types, labeled by k. If, when transcription of a gene is 'turned on', mRNA is produced at a rate R_p and degraded at a rate R_d, then the number of UMIs in the droplet data is expected to be $\alpha \sim R_\mathrm{p}/R_\mathrm{d}$. We cannot distinguish directly whether the variability in α is primarily due to variations in the rate of production or the rate of degradation.

Both quantities, p_{ik} and α_{ik}, have a broad range of values spanning two and three decades respectively. For this reason, it is useful to use logarithmic variables

$$I_{ik} = -\ln\left(p_{ik}\right)\ , \quad A_{ik} = \ln\left(\alpha_{ik}\right) \tag{4}$$

so that the statistics are not dominated by properties of the largest values. The negative sign in the definition of I ensures that this variable takes positive values. The A_{ik} will be referred to as the *activity* of gene i in tissue k, and I_{ik} will be termed its *intermittency*.

We emphasise that our parameters p and α are not intended to be an accurate description of the behaviour of a particular gene, because of they do not attempt to deal with the technical uncertainties in the single-cell mRNA counts. They are, rather, a means to make comparisons of the expression of different genes within the same data set.

Characterization of spiked-in sequences. There were 96 exogenous sequences (denoted by labels ERCC − 000xx, where xx is a two-digit number) 'spiked' into the FACS samples with known concentrations. These were used to give an indication of the reproducibility of the experimental procedure. Because the PCR process is unstable, it amplifies fluctuations, such as stochastic variation of counts or small differences in the experimental parameters between reaction cells. For this reason, it may be advantageous to 'normalize' the counts, replacing M_{ijk} by

$$M'_{ijk} = \frac{M_{ijk}}{\sum_i M_{ijk}} \tag{5}$$

(that is, the normalized values are the *fractions* of the total count for a given cell represented by this gene, rather than the raw count). If the variation of the total count for a cell were primarily due to technical limitations of the experiment, it would be preferable to use these normalised counts.

An alternative scenario is that the total number of mRNA molecules in a cell may have large variations, which are not due to measurement errors. In this case the measurements would be distorted by applying the normalization. We used the spike-in sequences to provide a test of whether it is appropriate to normalize the data. Fluctuations of the logarithm of the counts of spike-in sequences give an indication of the fractional errors. We compared the variance of $\ln M$ (un-normalized) to the variance of $\ln M'$ (normalized counts) for the exogenous sequences, and found that the variance of the latter was considerably larger (approximately five time higher than the variance of the un-normalized counts). This was mainly due to cells with very low total counts, which cause the genes which are present to be greatly exaggerated if normalization is carried out. Even after eliminating cells in the lower quartile of the total count, the variance of $\ln (M')$ was still substantially greater than that of $\ln (M)$. For this reason, our statistics used the raw (un-normalized) count data.

Because a given exogenous sequence is spiked into all cells at the same concentration, it should, ideally, have zero intermittency. In practice, because of the sources of technical variability mentioned above, the spike-in sequences have non-zero and apparently random values of the intermittency I_{ik}. We find few of the spike-ins which are present at higher concentrations (above 200 amol/μl) give values of the intermittency parameter greater than ln (5) (corresponding to a gene being active with probability less than $p = 0.2$). Accordingly, we regard all genes with $I >$ ln (5) as being intermittent.

While other groups have proposed quite complex schemes for normalizing single-cell mRNA counts[17,18], the use of normalization for the purposes of the present study did not appear to be advantageous.

Statistical Observations and Interpretations

Gene activity correlations. In order to distinguish between two possible models for explaining small values of p, cases **A** and **B** discussed in the Introduction and illustrated in Fig. 1, we used the *Tubula Muris* dataset to examine correlation coefficients of the gene activities: these are

$$C_{ii',k} = \frac{\langle M_{ijk} M_{i'jk} \rangle - \mu_{ik}\mu_{i'k}}{\sigma_{ik}\sigma_{i'k}} \tag{6}$$

where the angle brackets denote an average over the set of cells in tissue type k. Figure 2 shows the probability density function (PDF) of the correlation coefficients for heart and liver tissue, using the droplet data (other tissues give similar results). The distributions of both the positive and the negative correlation coefficients are displayed. Most of the correlation coefficients are extremely small and it appears implausible that these small correlation coefficients have any statistical significance. Including all gene pairs, positive and negative coefficients occurred in roughly equal numbers, but Fig. 2 shows that among the larger correlation coefficients, there is a much smaller proportion of negative coefficients.

We now argue that these data distinguish between cellular differentiation (case **A**) and intermittent transcription (case **B**). Let us consider the consequences of assuming that small expression probabilities arise solely as a consequence of cellular differentiation, and assume that two different low-expression genes, labelled by index numbers i and i', are only expressed in different rare cell types, C_1 and C_2 respectively. This implies that if gene i is being expressed, so that $M_{ijk} > 0$, then the cell is of type C_1, so that gene i' is not expressed, implying that $M_{i'jk} = 0$. Similarly, if $M_{i'jk} > 0$ we are dealing with a cell of type C_2, and $M_{ijk} = 0$. If rarely expressed genes result from cellular differentiation, then pairs of rare genes would not usually be expressed in the same cell. If the genes with index i and i' are only expressed in different cell types, then $\langle M_{ijk} M_{i'jk} \rangle$ would be equal to zero, because the counts M_{ijk} and $M_{i'jk}$ would never be non-zero in the same cell for different genes. This implies that the correlation coefficient, given in Eq. (6) would be negative, and equal to $-\mu_{ik}\mu_{i'k}/\sigma_{ik}\sigma_{i'k}$. According to this cellular differentiation model, many values of

$$K_{ii',k} = \frac{C_{ii',k}\sigma_{ik}\sigma_{i'k}}{\mu_{ik}\mu_{i'k}} \tag{7}$$

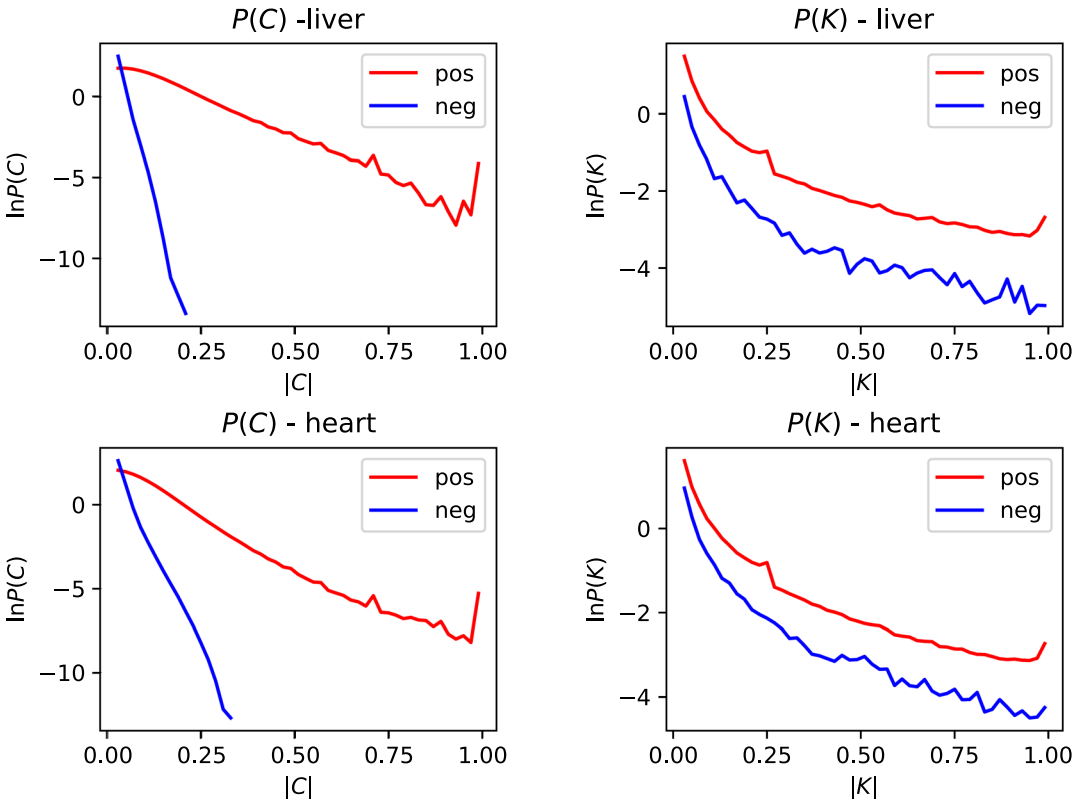

Figure 2. Distribution of the magnitude of gene-activity correlation coefficients, C (Eq. (6)), and of the transformed correlation coefficients, K (Eq. (7)), for both heart and liver cells. In each case we show the distributions for both positive and negative correlations. In cases where there is a significant correlation of gene activity, positive correlations are far more common than negative ones. If two genes were only active in different cells, their K coefficient would be expected to equal -1. More generally, negative correlations are a signature of genes being expressed in different cells. The preponderance of positive correlation coefficients over negative ones strongly favors the intermittent transcription model.

would be equal to -1 whenever genes i and i' are not expressed in the same type of cell. We infer that if cellular differentiation is the primary cause of observing genes with a low probability, then those coefficients $K_{ii',k}$ which correspond to genes which are never expressed in the same cell would be close to -1. In Fig. 2 we display the PDF of the positive and negative values of $K_{ii',k}$ for heart and liver droplet data. There are very few values of $K_{ii',k}$ close to -1 indicating that very few genes are being expressed exclusively in different cell types.

More generally, if genes i and i' have a strong tendency to to be expressed in different cell types, without being strictly mutually exclusive, this will lead to the correlation coefficients $C_{ii',k}$ being negative. However Fig. 2 indicates that there are more positive correlation coefficients than negative ones. This strongly favors the intermittent transcription model.

Variation between tissue types. For each gene, we can also define the average of the activity and of the intermittency across different tissues, denoting these by $\bar{A}_i$ and $\bar{I}_i$ respectively (we used a simple average, rather than one weighted by the numbers of cells in each tissue). It is interesting to consider the fluctuations of activity and intermittency between different tissue types, described by the following quantities:

$$\Delta A_{ik} = A_{ik} - \bar{A}_i, \quad \Delta I_{ik} = I_{ik} - \bar{I}_i. \tag{8}$$

Figure 3 shows 'heatmaps' (that is, a 2d histogram color coded for the occupancy of the bins) showing the density of ΔI, ΔA across all combinations of tissues and genes (these are plotted separately for the two experimental protocols). The most significant feature is that the values of the activity fluctuations ΔA show substantially smaller dispersion than the variation of the intermittency, ΔI. The variances of the data points generating Fig. 3 are $\mathrm{Var}(\Delta A) = 0.157$ and $\mathrm{Var}(\Delta I) = 1.28$ respectively for the droplet data, and $\mathrm{Var}(\Delta A) = 0.457$ and $\mathrm{Var}(\Delta I) = 1.02$ for the FACS data. Another interesting feature is that there are two distinct sub-sets of genes: the bright spot at the center of the heatmap shows that many genes show little variation in their intermittency between different tissues, whereas others show a marked variation in their probability of expression. While both data sets show clear evidence that the variation of I is greater than the variation of A, the heatmaps for the two data sets are quite different. The markedly lower value of the variance of ΔA_{ik} in the droplet data is principally due to there being a significant number of genes with just one count in one cell in this data set.

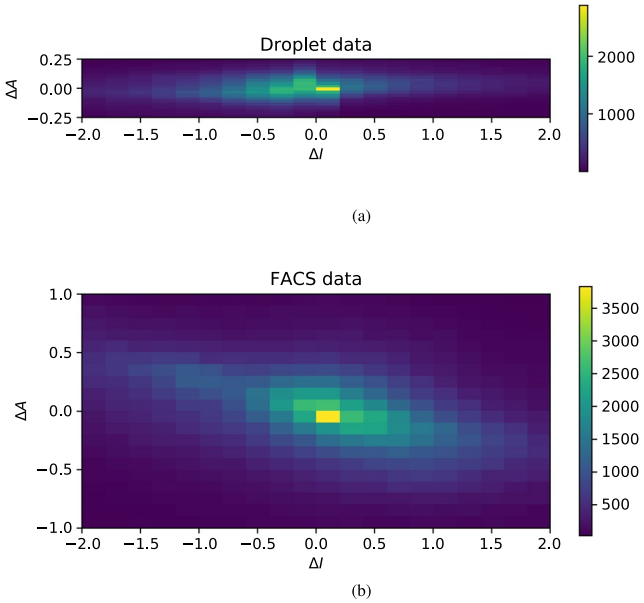

Figure 3. 'Heatmap' showing the dispersion of the gene activity and intermittency, ΔA and ΔI, across different tissues. The activity of a gene is almost constant for all tissue types. Some genes show a marked dispersion of their intermittency, while others have an almost constant probability of being represented by an mRNA transcript. The histogram uses 400 rectangular bins in a 20×20 lattice: (**a**) droplet data, (**b**) FACS data. These data indicate that the action of a gene in different tissues is primarily modulated by varying the probability p that the gene is expressed, while the peak activity α shows little variation across tissues.

The 'heatmap' in Fig. 3 suggests that the peak transcription levels α_i are intrinsic properties of the genes, unaffected by cellular differentiation or by the mechanisms regulating transcription. It might be expected that, at least in some genes, regulation of activity would involve the interaction of transcription factor proteins which could modulate the rate of gene expression by partially blocking the transcription process. If this control molecule were binding and detaching from the DNA on a timescale shorter than the lifetime of mRNA, the gene activity A would differ between tissues. Figure 3, however, shows that differences in activity of a gene between different tissues are (for most genes) small. The fact that many genes show evidence of intermittency implies that the timescales for turning transcription on and off are longer than the mRNA lifetime, which is typically taken to be at least one hour, and often considerably longer[14]. Our observations are, therefore, consistent with a picture of gene expression being regulated by switching transcription between 'on' and 'off' states on a timescale with is longer than the mRNA lifetime.

Distribution of gene parameters. We investigated the PDF of the gene activity A_{ik}, illustrated in Fig. 4, separately for each tissue, showing results for both experimental protocols. Because we have argued that the activity is approximately constant, we might expect that the distribution of the activity A will be very similar for different tissues, and the results for the droplet case confirm this. In the case of the FACS data we see curves which are similar, but shifted horizontally, indicating that the overall activity is different in different tissues. Figure 4 also shows the result of applying a 'normalization' to make the tissue mean equal to the overall mean, in order to account for the fact that cells in different tissues may have different levels for physiological activity, implying that their mean total mRNA count may differ. With this tissue-dependent normalization, the distributions of A from different tissue types are very similar.

In the case of the distribution of the gene activities A, there are marked differences between the two different experimental protocols. In particular, in the FACS data there is a 'tail' of the distribution corresponding to genes which have very low activities, which is absent from the distribution of activities obtained from the droplet data. The lowest activity genes in the droplet data correspond to events where a single UMI is recorded in a single cell.

We also investigated the PDF of the average of the activity over different genes, $\bar{A}_i$ (we shall use an overbar to denote an average over tissue types). The tail of the PDF of $\bar{A}$, illustrated in Fig. 5, can be approximated by an exponential function, $\exp(-\mu\bar{A})$, for some coefficient μ, when $\bar{A}$ is large. The corresponding PDF of the activity parameter $\alpha = \exp(A)$ is a power-law, of the form $P(\alpha) \sim \alpha^{-(\mu+1)}$ for large values of α. Figure 5 is consistent with $P(\alpha)$ having a power-law form for both data sets, but the exponents are different: for the droplet data we have $P(\alpha) \sim \alpha^{-2.35}$, whereas $P(\alpha) \sim \alpha^{-2.65}$ for the FACS data. The data in Fig. 5 indicates that, while most genes have similar values of the activity parameter, there are a few which have much larger peak activity. We speculate that this can be explained by models in which transcription is impeded by slowly transcribing base sequences. Rare, highly active genes are those which happen not to have the slowly-transcribing sequences.

The PDF of the intermittency I is illustrated in Fig. 6 separately for each of the tissue types. For each tissue, the PDF of I shows a peak close to $I = 0$, corresponding to a sub-set of genes which are not intermittent (that is,

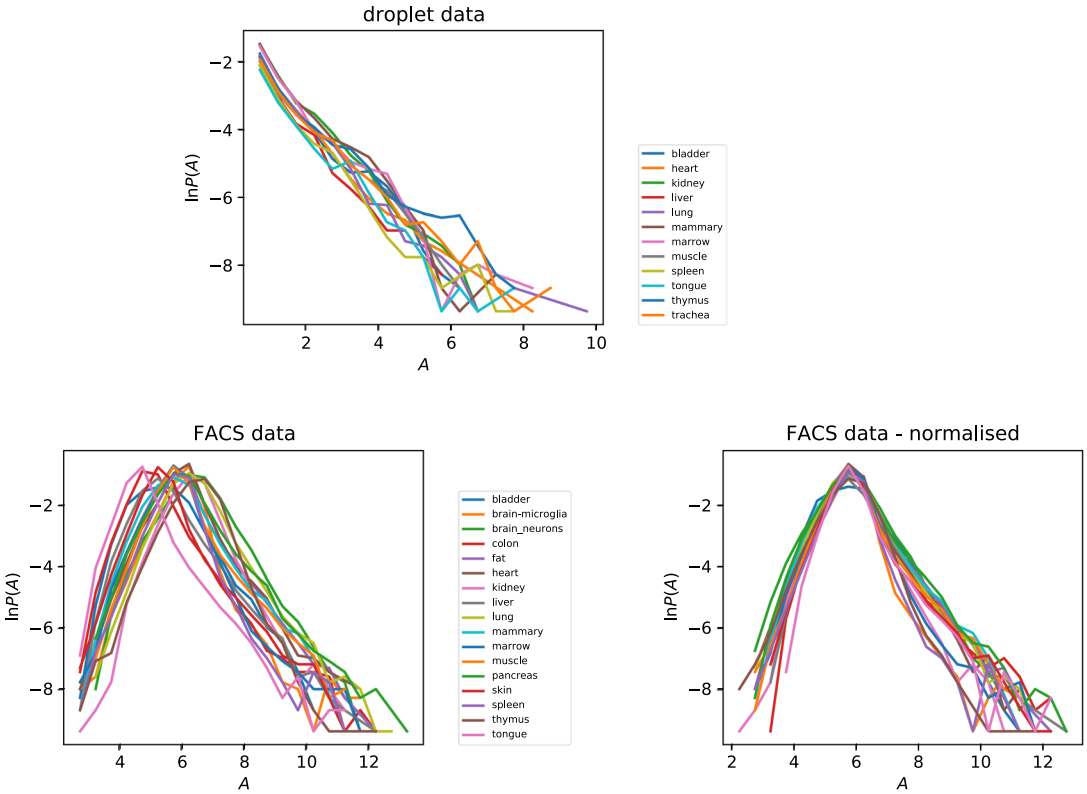

Figure 4. Distribution of gene activity parameter A_{ik} for different tissues, for both droplet and FACS data, showing that the activity parameter varies by orders of magnitude. The marked difference between the plots for the two datasets reflects the greater sensitivity of the FACS protocol. In the case of the FACS data we also exhibit the effect of normalizing to equalise the mean activity of different tissues.

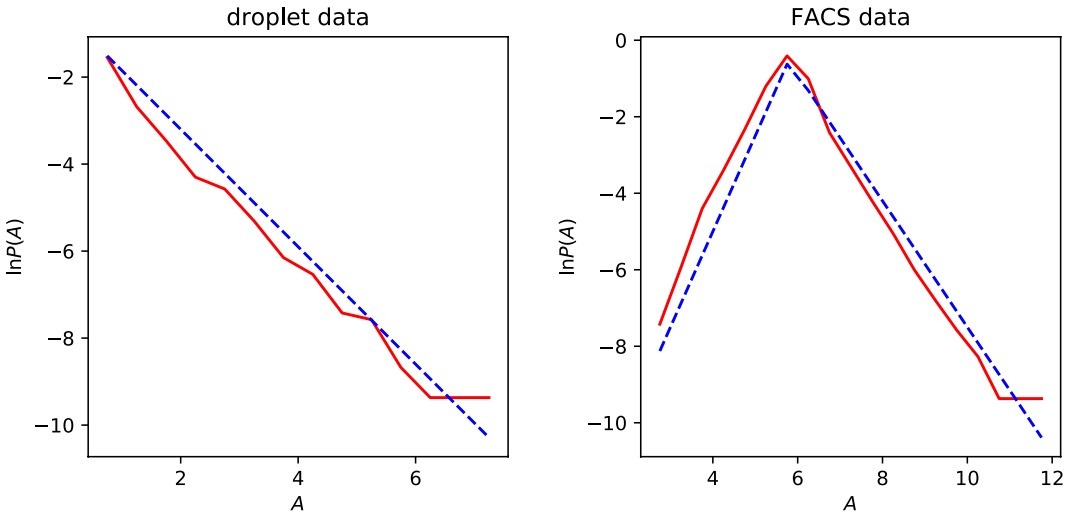

Figure 5. Distribution of tissue-averaged gene activity $\bar{A}$ for droplet data and FACS data. The dashed lines are a guide to the eye. The lines in the droplet data plot has slope -1.35, and those in the FACS plot have slopes $+2.5$ and -1.65. These data are consistent with the PDF of α having a power-law tail for large values of α.

expressed with high probability). The distributions of the intermittency I do differ between different tissues, in accord with the discussion in section sec: 3.1 above, but we can say that the PDF of ln (I) is very approximately constant over three decades, indicative of a very broad distribution of the intermittency. For the same tissue, the distributions of the intermittency are quite different for the two experimental protocols. This may reflect the greater sensitivity of the FACS experiments. Figure 6 indicates that a sub-set of genes (with $I \approx 0$) are transcribed continuously, while some others are transcribed with very low probability. The genes which are transcribed with

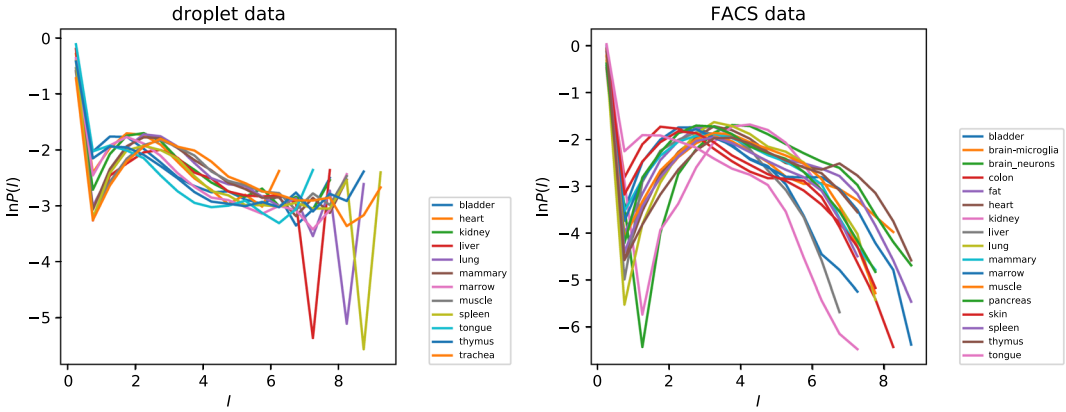

Figure 6. Distributions of gene intermittency parameter I by tissue type, for both droplet and FACS data.

very low probability may be concerned with control functions. The regulatory processes in a cell are probably organized in a hierarchical manner. The fact that the PDF of I is approximately uniform is interesting, and may be indicative of how the control of transcription is controlled.

It might be suspected that genes which have a regulatory function would be required to produce very little protein, and that they might therefore have a low activity parameter A, as well as having a high degree of intermittency, I, This suggests that there would be a negative correlation between A and I. However, the evidence from Fig. 3 indicates that the expression of genes is regulated via modulation of I rather than A, indicating that the latter quantity is an intrinsic attribute of the gene, and not susceptible to control by gene regulation networks. This argument would suggest that there may be no significant correlation between I and A. We also examined the correlation coefficient $C_{\bar{A}\bar{I}}$ between $\bar{A}$ and $\bar{I}$, finding $C_{\bar{A}\bar{I}} = -0.235$ for the droplet dataset and $C_{\bar{A}\bar{I}} = -0.129$ for the FACS dataset, showing that genes with a high intermittency have a slight tendency to have low activity. The low degree of correlation between A and I is consistent with the picture that the gene activity α is an intrinsic property of a gene.

Discussion

Our studies of gene correlations in single-cell mRNA counts for mouse cells distinguish between two possible reasons why some genes are rarely transcribed. They support the view that rarely observed genes are being transcribed intermittently in most cells within a tissue, as opposed to being expressed continuously in specialized sub-populations of cells.

We proposed a characterization of intermittent transcription by assigning two parameters to every gene, namely a peak transcription level α and a probability of transcription, p. We found that both variables have a very wide range of values, calling for a statistical analysis in terms of logarithmic variables: the activity $A = \ln \alpha$ and the intermittency $I = -\ln p$.

The wide range of different tissues types in the *Tabula Muris* datasets enable us to gather evidence about the effects of the differentiation of tissues on the transcription process. Single-cell mRNA sequence data provide support for the view that the gene transcription rate α in the 'on'-state is an intrinsic property of the genes, and that differentiation of tissues leads to variation in the probability p that transcription occurs. The timescale for turning genes on or off must be slow compared to the lifetime of mRNA molecules (otherwise the level of mRNA would remain nearly constant), but our data do not permit these timescales to be estimated reliably.

This picture of intermittent transcription is consistent with direct observations of 'bursting' transcription of genes in other systems. Our results are consistent with the hypothesis that intermittent transcription is ubiquitous in mammalian cells, extending to rarely expressed genes where it would be difficult to observe directly. Our approach is complementary to earlier studies[11–13], which assign parameters to genes using a stochastic model, based upon telegraph-noise processes[8–10]. This stochastic model does not address why bursting processes appear to play an important role in systems such as mammalian cells, where homeostasis is an important principle. Our results do suggest reasons why intermittent transcription should be ubiquitous. Intermittent transcription might be used by cells as a matter of necessity, because our results on the constancy of α across different tissue types indicate that cells only have 'on-off' control, rather than a continuously variable transcription rate. However, beyond being a matter of necessity, intermittent transcription may offer definite advantages over the cellular differentiation model. It makes sense to 'stage' operations so that proteins for specific purposes are produced only after their interaction partners are in place. If all the components of complex cellular systems were produced all the time then they would have difficulty finding the complementary sites with which they should bind. Limits on the rate of ribosomal translation imply that a cell cannot be efficiently performing all of its functions all of the time. These arguments about using intermittent transcription to organize the efficient operation of a cell would suggest that, rather than modeling bursting as a purely stochastic phenomenon, we should look for a model of bursting transcription which is determined by a complex dynamical system which can turn genes on and off in a time-ordered sequence.

Finally, we remark that studies of splicing of mRNA should be capable of yielding additional information about the timescales of the intermittent gene transcription, by an extension of the approach described by LaManno *et al.*[19].

Appendix

Here we discuss the relationship between the parameters of the stochastic model for gene activity[11–13], and our parameters p and α, defined by Eqs. (2) and (3).

The stochastic model considers the DNA of a gene to have two states: A (active) and I (inactive). Conversion between these states is a telegraph noise process, with rate constants k_{on} (for the process I $\rightarrow$ A) and k_{off} (for A $\rightarrow$ I). The A state is transcribed to make mRNA, with rate constant k_{tr}, which is subsequently destroyed at a rate k_{deg}. If M is the population of the mRNA the steady-state mean and variance of this count are given by Peccoud and Ycart[7]: in our notation

$$\langle M\rangle = \frac{k_{on}}{k_{on}+k_{off}}\frac{k_{tr}}{k_{deg}} = p_0 M_0, \quad \mathrm{Var}(M) = p_0 M_0 + p_0(1-p_0)M_0^2\frac{k_{deg}}{k_{on}+k_{off}+k_{deg}} \tag{9}$$

where

$$p_0 = \frac{k_{on}}{k_{on}+k_{off}}, \quad M_0 = \frac{k_{tr}}{k_{deg}} \tag{10}$$

Note that p_0 is the probability that the DNA is in the active (A) state, and M_0 is the mean count that would be obtained if the DNA were always in the A state.

Our own parameters p and α, defined by Eqs. (2) and (3), are intended to provide a simple estimate of the probability that a gene is active, and of its peak activity level, when $p \ll 1$. We should consider the circumstances under which p and α are approximately equal to the parameters p_0 and M_0 defined in (10). Because we exclude genes with very low counts, we may assume that $M_0 \gg 1$, and we have then

$$\begin{aligned} p &\sim \frac{p_0}{1-p_0}\frac{k_{deg}+k_{on}+k_{off}}{k_{deg}} \\ \alpha &\sim M_0(1-p_0)\frac{k_{deg}}{k_{deg}+k_{on}+k_{off}} \end{aligned} \tag{11}$$

When $p_0 \ll 1$ and the degradation rate k_{deg} satisfies $k_{deg} \gg k_{on}+k_{off}$, we find that our parameters p and α can indeed be identified with the parameters p_0 and M_0 of the stochastic model.

Received: 13 November 2019; Accepted: 10 January 2020;
Published online: 21 February 2020

References

1. Saiki, R. *et al.* Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**, 487–91 (1988).
2. Zhang, K. *et al.* Sequencing genomes from single cells by polymerase cloning. *Nature Biotechnology* **24**, 680–6 (2006).
3. The Tabula Muris Consortium, Quake, S. R., Wyss-Coray, T. & Darmanis, S. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–72 (2018).
4. Chubb, J. R., Trcek, T., Shenoy, S. M. & Singer, R. H. Transcriptional pulsing of a developmental gene. *Current Biology* **16**, 1018–25 (2006).
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. Chubb, J. R. & Liverpool, T. B. Bursts and pulses: insights from single cell studies into transcriptional mechanisms. *Current Opinion in Genetics and Development* **20**, 478–84 (2010).
7. Peccoud, J. & Ycart, B. Markovian modelling of gene product synthesis. *Theoretical Population Biology* **48**, 222–234 (1995).
8. Paulsson, J. Models of stochastic gene expression. *Physics of Life Reviews* **2**, 157–175 (2005).
9. Corrigan, A. M., Tunnacliffe, E., Cannon, D. & Chubb, J. R. A continuum model of transcriptional bursting. *ELIFE* **5**, e13051 (2016).
10. Chubb, J. R. Gene regulation: stable noise. *Current Biology* **26**, R60–82 (2016).
11. Kim, J. K. & Marioni, J. C. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biol.* **14**, R7 (2013).
12. Grün, D., Lennart, K. & van Oudenaarden, A. Validation of noise models for single-cell transcriptomics. *Nature Methods* **11**, 637–640 (2014).
13. Larsson, A. J. M. *et al.* Genomic encoding of transcriptional burst kinetics. *Nature* **565**, 251–4 (2018).
14. Chen, C.-Y. A., Ezzeddine, N. & Shyu, A.-B. Messenger rna half-life measurements in mammalian cells. *Methods Enzymol.* **448**, 335–357 (2008).
15. Tabula Muris Consortium. Data set archived at, https://tabula-muris.ds.czbiohub.org (2018).
16. Taniguchi, Y. *et al.* Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329**, 533–8 (2010).
17. Townes, F. W., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. Feature selection and dimension reduction for single cell RNA-seq based on a multinomial model. *Genome Biol.* **20**, 295, https://doi.org/10.1186/s13059-019-1861-6 (2019).
18. Vieth, B., Parekh, S., Ziegenhain, C., Enard, W. & Hellmann, I. A systematic evaluation of single cell RNA-seq analysis pipelines. *Nature Communications* **10**, 4667, https://doi.org/10.1038/s41467-019-12266-7 (2019).
19. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).

Acknowledgements
We thank Joshua Batson, Olga Botvinnik and Emma Lundberg for comments. M.W. thanks the Chan Zuckerberg Biohub for its hospitality. G.H. thanks the Kavli Institute for Theoretical Physics (supported by grant NSF-PHY-1748958) for its hospitality.

Author contributions
M.W. executed the study, with guidance and contributions from S.D., A.P. and G.H. All authors contributed to writing the manuscript, and reviewed the manuscript before submission.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to M.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.