




Emergent statistical laws in single-cell transcriptomic dataSilvia Lazzardi * and Filippo Valle **Department of Physics, University of Turin and INFN, via P. Giuria 1, 10125 Turin, Italy*Andrea Mazzolini *Laboratoire de Physique de l'École Normale Supérieure (PSL University), CNRS,
Sorbonne Université and Université de Paris, 75005 Paris, France*Antonio Scialdone *Institute of Epigenetics and Stem Cells, Helmholtz Zentrum München, Feodor-Lynen-Straße 21, 81377 München, Germany
and Institute of Functional Epigenetics and Institute of Computational Biology, Helmholtz Zentrum München,
Ingolstädter Landstraße 1, 85764 Neuherberg, Germany*Michele Caselle  and Matteo Osella [†]*Department of Physics, University of Turin and INFN, via P. Giuria 1, 10125 Turin, Italy*

(Received 21 October 2022; accepted 24 March 2023; published 27 April 2023)

Large-scale data on single-cell gene expression have the potential to unravel the specific transcriptional programs of different cell types. The structure of these expression datasets suggests a similarity with several other complex systems that can be analogously described through the statistics of their basic building blocks. Transcriptomes of single cells are collections of messenger RNA abundances transcribed from a common set of genes just as books are different collections of words from a shared vocabulary, genomes of different species are specific compositions of genes belonging to evolutionary families, and ecological niches can be described by their species abundances. Following this analogy, we identify several emergent statistical laws in single-cell transcriptomic data closely similar to regularities found in linguistics, ecology, or genomics. A simple mathematical framework can be used to analyze the relations between different laws and the possible mechanisms behind their ubiquity. Importantly, treatable statistical models can be useful tools in transcriptomics to disentangle the actual biological variability from general statistical effects present in most component systems and from the consequences of the sampling process inherent to the experimental technique.

DOI: [10.1103/PhysRevE.107.044403](https://doi.org/10.1103/PhysRevE.107.044403)**I. INTRODUCTION**

Almost every cell of an organism has the same gene content, but how these genes are expressed ultimately defines the cellular phenotype. If the gene repertoire is the genomic vocabulary, the transcription program represents how the different words are actually used by different cells to determine the cell identity [1]. Single-cell RNA sequencing (scRNAseq) technologies have recently given access to these cell-specific transcription programs [2], and large-scale expression atlases have been compiled collecting thousands of single-cell expression profiles for all the major organs of different species [3–5].

The analogy between word statistics in a collection of texts and gene expression profiles in a large population of cells suggests that the transcriptome can be looked at as a complex component system [6]. Several complex systems of different nature and origin, from linguistics to biology, have an analogous modular structure with identifiable basic common building blocks that are used with different statistics.

This statistics should contain information about the generative processes and the architectural constraints of the system. Books are composed of words, genomes of different species are collections of genes of different evolutionary families or associated with different biological functions, and ecological niches are compositions of species with different abundances. Analogously, the transcriptional profiles of single cells are the sums of specific amounts of RNAs transcribed from a repertoire of common genes, and scRNAseq provides a picture of these profiles. The advantage of looking at single-cell transcriptomics as a complex component system is that a modeling framework and a set of analysis tools, based on statistical physics, have been developed for these systems. In fact, common statistical regularities have been characterized quantitatively in the different component systems described [7–11], and simple models have been proposed to explain their emergence. The first question we will address is if analogous statistical laws can be identified in large-scale transcriptomic data. While only a few general regularities have been already recognized in transcriptomic data [12–14], this work presents a detailed and systematic exploration across different datasets and experimental techniques. A quantitative systematic description is important for the development

*These authors contributed equally to this work.

[†]matteo.osella@unito.it

and test of simple statistical models that can capture the connections between seemingly independent emerging laws. These data-driven models can in turn be used as “null models,” for example to disentangle genuine biological variability from technical or statistical effects in the context of transcriptomics. In fact, the observed variability in single-cell expression experiments typically has several possible sources, such as technical noise due to the experimental techniques, the intrinsic stochasticity in gene expression, and the biological variability actually setting the cell identity in terms of cell type and cell state (such as the cell cycle stage) [1,15]. In particular, sampling noise associated with RNA capture and sequencing is inherent in RNA-sequencing techniques and can be a dominant source of noise, especially in single-cell transcriptomics where the starting RNA material is limited. This work focuses on cell atlases, as illustrative examples of large-scale scRNAseq datasets, with this complex systems perspective. We will identify emergent statistical laws in these datasets and assess their universality by comparison with properties of other component systems. Using a general and simple mathematical framework, we will show that several statistical properties of scRNAseq datasets can actually be explained as a result of the combination of heterogeneity in average expression levels and a sampling process. For example, we will show how this simple description can largely explain the empirical data sparsity in scRNAseq datasets, whose origin is still a debated topic in the field [16,17], without invoking convoluted or *ad hoc* assumptions.

On the other hand, some data features are not fully captured by this basic model, suggesting where to focus in order to extrapolate actual properties of biological variability. In fact, models based on empirical statistical laws, such as the one presented here or its potential future advancements, can be used as null models, for example to select the genes whose expression pattern is significantly divergent from the model expectation because of technical or biological reasons. We will show few illustrative examples of this model-driven gene-selection procedure. Analogously, the extent of the general discrepancy of a dataset properties from the model predictions should be related to intrinsic characteristics of the dataset, such as the degree of diversification of cell types. We will show that this indeed seems to be the case.

Data-driven statistical models are also instrumental to generate simple but realistic simulated datasets, where the key parameters are fully under control, in order to benchmark analysis methods or computational pipelines. The development of realistic benchmark datasets of controlled complexity is a crucial problem in the field [18], where a plethora of computational methods have been proposed, but it is not straightforward to quantitatively assess their relative performances.

Finally, we will discuss how the addition of transcriptomic data to the increasing large set of systems displaying seemingly universal statistical properties is a relevant case study in the context of model generation. In complex systems theory, different general models and principles behind these ubiquitous laws have been proposed, and new empirical examples such as transcriptomic data can be useful for model testing and selection. Therefore, this systematic exploration of statistical laws in single-cell transcriptomics could help to bridge the

gap between mathematicians and physicists building general quantitative descriptions of complex systems, and computational biologists that could use the same descriptions to extract useful biological information from large-scale datasets.

II. MATERIALS AND METHODS

A. Data sources

The Mouse Cell Atlas (MCA) was selected as the main illustrative dataset. In the MCA more than $\sim 4 \times 10^5$ single cells were profiled using scRNAseq from all major organs [3,19] using the microwell-sequencing technique, a high-throughput and low-cost scRNAseq platform. An advantage of this dataset is the use of unique molecular identifiers (UMIs) [20]. This technique allows the identification of the absolute number of unique RNA molecules detected by sequencing, thus eliminating the amplification noise. In the context of single-cell gene expression assays, this method provides a reliable estimate of the number of mRNAs detected for coding genes and an estimate of the transcriptome size sampled. Our analysis therefore mainly focuses on absolute molecule counts, rather than relying on normalization techniques, which are still a research area in sequencing data analysis.

We also analyzed the compendium of Tabula Muris (TM) for comparison. This atlas comprises an analogous number of cells from 20 organs and tissues [4,21] that were processed with the Smart-seq2 protocol [22], which produces a full-length transcriptome profiling but does not use UMIs.

A dataset of bulk RNA-sequencing of healthy human tissues from the Genotype-Tissue Expression (GTEx) Project [23] was used to test the results on population-averaged transcriptomic data.

Finally, we analyzed two additional single-cell datasets, relative to a human embryonic kidney (HEK) cell line and to mouse fibroblasts, profiled with the recently introduced Smart-seq3 protocol [24].

B. The data structure for component systems

A transcriptomic dataset, and more generally a component system, can be described by a matrix $\{n_i^c\}$ where each entry represents the counts relative to transcript (i.e. the component) $i \in \{1, \dots, N\}$ in cell (i.e., the realization) $c \in \{1, \dots, R\}$. N is the total number of different transcripts that could be present (the number of genes as a first approximation), which is essentially the vocabulary of our system. R is the number of cells analyzed. Each column of the data matrix is a vector $\{n_i^c\} = \{n_1^c, \dots, n_N^c\}$ that fully describes the expression profile of a single cell c . The size of the transcriptome of a cell captured in the experiment is defined as $M^c = \sum_{i=1}^N n_i^c$. While in other component systems, such as texts of natural language, this parameter is simply the size of the realization (e.g., the book size), in our context M^c represents the measured transcriptome size. Therefore, it does not necessarily correspond to the total number of transcripts in the cell because of the sampling process involved in RNA capture.

As described in the previous section, this work mainly focuses on two scRNAseq atlases: the MCA and the TM compendium. In these datasets, the total numbers of genes N (i.e., the genes with at least a single detected transcript) are

respectively around 38×10^3 and 23×10^3 , while the numbers of cells R are 34×10^3 and 41×10^3 . The distribution of transcriptome sizes M^c is quite broad and dataset dependent. In the MCA, the average number of UMIs per cell is $\simeq 1200$ and the distribution is reported in Fig. S1 of the Supplemental Material [25].

C. An analytical framework to model gene expression data

This section develops the mathematical framework that will be used to describe the datasets. Using this framework, simple null models can be built to characterize the expected statistical behaviors given the model assumptions. The same mathematical description was applied in the context of metagenomic data [11], while an analogous approach was previously introduced for scRNAseq data [26] and is the basis of a recent Bayesian procedure for data normalization [27].

The key underlying assumption is that the observed mRNA counts n_i^c are the combined result of the inherent biological variability between cells and of the sampling process due to RNA capture and sequencing [11,27]. Therefore, the probability of observing a specific expression profile $\{n^c\}$ in a cell c , from which M^c transcripts have been sequenced, is given by

$$P(\{n^c\}|M^c, \{f^c\}) = \frac{M^c!}{\prod_{i=1}^N n_i^c!} \prod_{i=1}^N (f_i^c)^{n_i^c}, \quad (1)$$

where f_i^c represents the true frequency of the mRNA i in cell c . The cell-to-cell variation in gene expression can be described by an unknown probability distribution $\rho(\{f\})$ setting the mRNA frequencies in the different cells. Thus, the probability of an expression profile in a cell with M observed transcripts becomes

$$P(\{n^c\}|M, \rho(\{f\})) = \int_0^\infty [df] \rho(\{f\}) \frac{M!}{\prod_i n_i^c!} \prod_i (f_i)^{n_i^c}. \quad (2)$$

Focusing on a single gene i (by marginalizing the above expression), we have the probability of observing n counts as

$$\begin{aligned} P_i(n|M) &= \int_0^\infty df \rho_i(f) \binom{M}{n} f^n (1-f)^{M-n} \\ &\simeq \int_0^\infty df \rho_i(f) \frac{e^{-fM} (fM)^n}{n!}. \end{aligned} \quad (3)$$

The Poisson approximation is valid when the number of mRNAs is large, $M \gg 1$, and even highly expressed genes occupy a small fraction of the total, which is typically the case in the datasets we want to analyze. In the presence of a dataset in which a gene transcript covers a large fraction of the total transcriptome for a non-negligible percentage of cells, the simplifying Poisson approximation does not hold. In those cases, the full model has to be considered and spurious negative correlations between genes can arise because of finite size effects. The distribution ρ_i captures the variability in expression of gene i due to both the different cell identities present in the cell population and the contribution from stochastic gene expression. On the other hand, the sampling variability is explicitly captured in the model by the binomial distribution. The average frequencies f_i can be directly estimated by the

empirical ones in the dataset [11]:

$$f_i \simeq \frac{1}{R} \sum_{c=1}^R \frac{n_i^c}{M^c}. \quad (4)$$

The ambitious goal would be to infer the distributions ρ_i , and to distinguish the different contributions to the biological expression variability. Here, instead, we first consider a simple limiting case in which the actual biological variability is negligible with respect to the sampling noise. In this case, the distributions ρ_i are extremely peaked with respect to the sampling noise and can be approximated with delta functions, i.e., $\rho_i(f) \simeq \delta(f - f_i)$. In this simplified scenario, the probability of observing an expression profile $\{n_i\}$ is given by the expression

$$P(\{n_i\}|M) = \frac{M!}{\prod_i n_i!} \prod_i (f_i)^{n_i}. \quad (5)$$

This model was previously analyzed to understand the origin of statistical regularities in different component systems [28].

We will compare the predictions of this simple model with empirical expression data. The idea is to understand what can actually be explained from the natural heterogeneity in average expression levels and from pure statistical effects due to the sampling process. One advantage of this model is that it is analytically treatable and provides mathematical predictions that can be directly tested against data. The situation of a dominant sampling noise can also be easily simulated. The transcript frequencies f_i can be estimated from data with Eq. (4) and an ensemble of synthetic cells can be generated by randomly drawing M^c transcripts, with M^c values matching the empirical ones. The resulting surrogate datasets reproduce precisely the average expression levels and the sampling depth of the empirical dataset.

III. RESULTS

A. Robust emergence of Zipf's law for the gene expression levels at different scales

One of the hallmarks of complex systems, from real-world networks to natural language, is a high level of heterogeneity, which is often epitomized by the emergence of power-law distributions [29]. For component systems in particular, the frequency of components is often well described by a power law known as Zipf's law [6,9,29,30]. In natural language, this law describes the distribution of word frequencies in a corpus of texts, typically reported as a rank plot. In the context of transcriptomics, this would translate in a law for the distribution of gene expression levels in a large-scale dataset. Figure 1(a) reports the rank plot of the relative expression levels f_i calculated by averaging across cells belonging to the same organ (different curves correspond to different organs) in the Mouse Cell Atlas (MCA) [3]. The distribution is largely compatible with a power-law decay with an exponent close to -1 , as in the classic Zipf's law, followed by an exponential tail. The shape of the distribution does not depend on the specific dataset or on the experimental technique used. An essentially identical plot [Fig. 1(b)] is obtained by looking at the same organs in an alternative mouse expression atlas, i.e., Tabula Muris [4], in which different sequencing methods

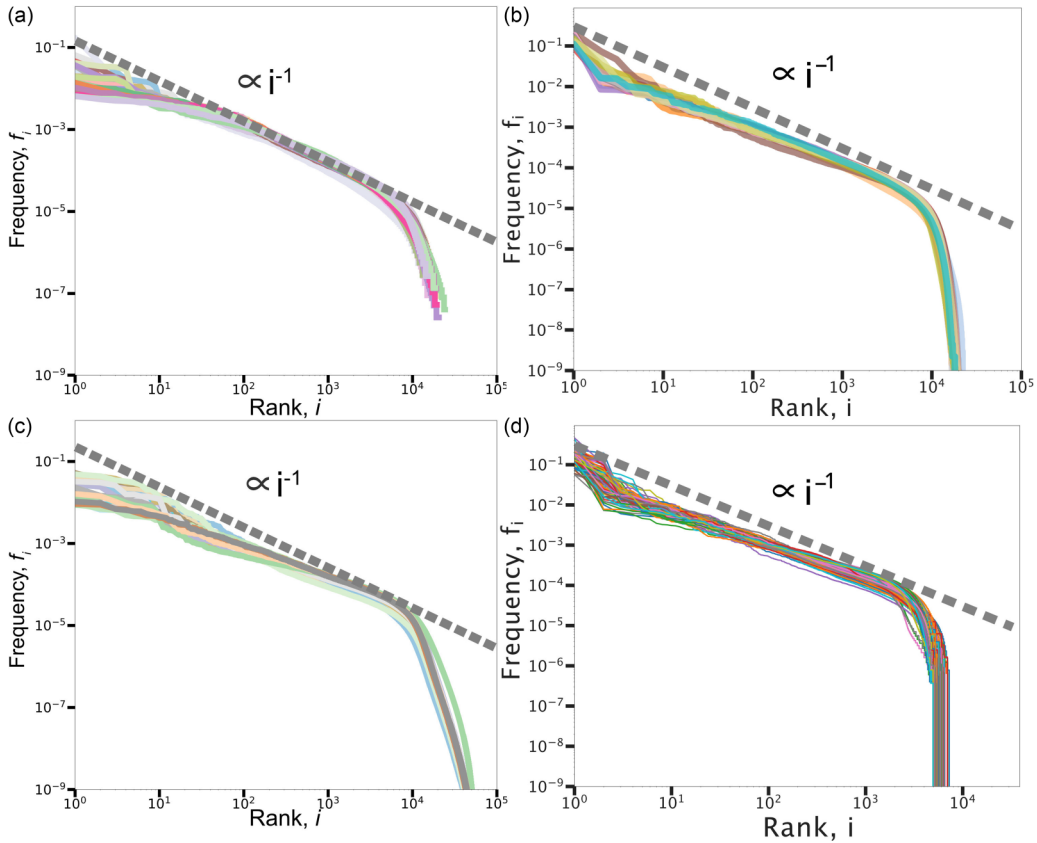


FIG. 1. A robust Zipf-like law for gene expression levels. The average relative expression levels f_i are estimated as described by Eq. (4) and reported as a function of their rank. The distributions reported correspond to averages over single cells belonging to different mouse organs from the Mouse Cell Atlas (a), from the Tabula Muris database (b), and to bulk RNA sequencing data from samples of healthy human organs in the GTEx database (c). Each curve corresponds to a single organ or tissue and the corresponding color code is reported in Fig. S2 [25]. (d) The relative gene expression levels evaluated in single cells (without averaging) follow an analogous Zipf-like trend. We report the distribution relative to 100 cells from the heart sample in Tabula Muris. Similar results can be obtained from other organs or from the Mouse Cell Atlas (Fig. S3 [25]). The dashed lines are just a reference power-law scaling with exponent -1 .

were adopted. Even if the relative gene expression levels measured in the two atlases are correlated, the variability is substantial (Fig. S4 [25]). Besides biological variability, the two atlases adopt different protocols, therefore it is not so surprising that the measured expression levels are not perfectly conserved. However, Zipf's law is robustly emerging. Also limiting the analysis to noncoding genes in the MCA, we still find the same distribution (Fig. S5 [25]). This statistical property seems indeed very general and not limited to scRNAseq data or to the specific species in analysis. For example, the same law emerges considering bulk RNA sequencing measurements across healthy tissues in human from the GTEx database [23] [Fig. 1(c)]. This result corroborates previous observations based on microarray and SAGE (serial analysis of gene expression) datasets that reported a power-law distribution of gene expression levels across different species and experimental conditions [31,32], as well as previous observations based on RNA sequencing data [13,14].

A natural question that can now be asked thanks to single-cell transcriptomics is if this emerging behavior is a consequence of the averaging process or a property of the gene expression program of single cells. Figure 1(d) shows an illustrative example of the gene expression distributions in single

cells. Besides some variability, the distributions recapitulate the population ones reported in the other panels. Therefore, the Zipf-like behavior is an inherent property of single-cell expression profiles.

In conclusion, Zipf's law appears to be a robustly emerging statistical property of gene expression data from bulk to single-cell experiments. This law sets the only free parameters f_i of our null sampling model [Eq. (5)] that can be used to test what properties of the system can be explained merely by sampling effects.

B. A Zipf's law with multiple regimes

At a coarse grained view, the rank plot of gene expression levels can be described as a power law followed by an exponential tail. The presence of a double scaling in the component frequency distribution again is a general feature of several component systems. A similar behavior can be observed by looking at protein domain frequencies in genomes of different species [6], where an exponential tail can be identified after the power-law scaling. A double scaling was also observed in natural language [33], where it was tentatively explained by a model with two different classes of words: common words

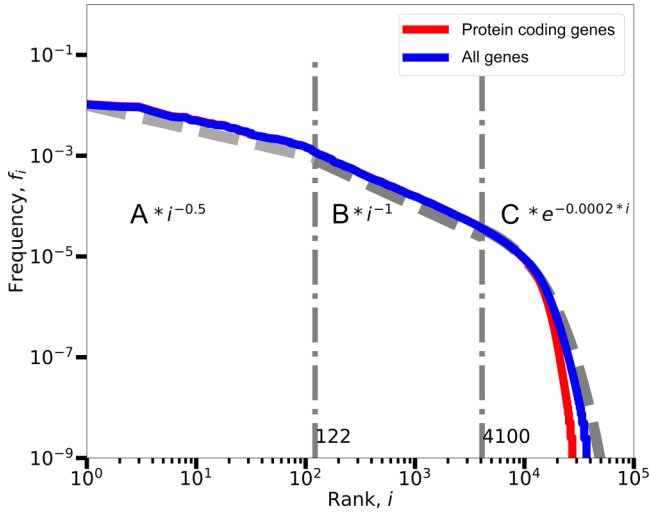


FIG. 2. Multiple regimes in the rank plot of the average expression levels. By considering the fitting error using a power law function in a window with variable width (Fig. S7 [25]), we were able to identify the part of the distribution well explained by a single power law, and consequently the two other regimes for highly expressed and lowly expressed genes. Excluding or including the noncoding genes from the analysis only influences the exponential tail of the distribution. The figure refers to the MCA dataset as an illustrative example.

(high rank) composing a core vocabulary and the rest of more specific words in a vast vocabulary. This double scaling was not characterized in previous analysis of the Zipf's law based on RNA sequencing data [13,14]. However, two different groups of genes can be analogously distinguished in bulk transcriptomic data: a core of highly expressed genes with active promoters and a second group of lowly expressed and putatively nonfunctional transcripts [12]. This distinction was originally based on an observed bimodality in the histogram of expression levels in several bulk experiments. The same trend is present also in scRNAseq data from mouse organs (Fig. S6 [25]), and it is reflected in the drastic change of regime in the Zipf's law, where the exponential tail contains the lowly expressed genes (Fig. 1). Figure 2 shows that this exponential tail is also enriched in noncoding genes, that are indeed generally lowly expressed.

However, a more detailed and quantitative analysis indicates that also the top highly expressed genes deviate from the general power-law behavior with exponent close to -1 (Fig. 2). In fact, it is possible to identify three different regimes that are approximately captured by two power laws with different exponents before the exponential tail. In order to quantitatively support these observations, we selected a window of ranks (e.g., identified by the dashed-dotted lines in Fig. 2) over which we performed a power-law fit to the frequencies. For different positions of the window boundaries, the coefficient of determination $R^2 = 1 - \frac{\sum_i |g(i) - f_i|^2}{\sum_i (f - f_i)^2}$ can be evaluated to select the range of ranks where the distribution is best explained by a power-law function. In the expression above, $g(i) = B \times i^{-\gamma}$ is the power-law function of rank i obtained by fitting, while f_i are the empirical frequencies with average \bar{f} . Once the boundaries and the best power-

law fit of the central part of the distribution are defined by this procedure, the first regime (low ranks and high frequencies) can be fitted with an independent power-law function $g(i) = A \times i^{-\gamma_1}$, and the third regime (high ranks and low frequencies) with an exponential function $g(i) = C \times e^{-\gamma_3 i}$. Figure S7 [25] presents a more detailed illustration of this fitting procedure.

Considering all the cells in the MCA, highly expressed genes (around 100 genes) follow a power law with exponent close to -0.5 , while the central part of the distribution is well described by an exponent close to -1 as in the classic Zipf's law. Interestingly, a very similar law with three regimes was observed in a quantitative transcriptomic study of fission yeast [34]. The same behavior can be observed by considering the different tissues in the MCA separately (Fig. S8 [25]). Also the gene expression levels in single cells [Fig. 1(d)] seem to generally display three classes, but the higher level of fluctuations does not allow a refined analysis.

A Zipf's law with three regimes emerges also across different datasets, as can be qualitatively observed from Fig. 1. The precise values of the boundaries and of the fitted exponents are dataset dependent as reported in the Supplemental Material [25], Tables S1 and S2. However, the general trend appear to be conserved: few highly expressed genes with a flatter expression distribution are followed by a central region of expression levels well described by a power law with exponent close to -1 . Finally, the distribution shows an exponential tail for lowly expressed genes.

The highly expressed genes in the first regime belong to specific functional classes. For example, the most enriched gene ontology (GO) categories for the genes with rank lower than 100 in the MCA are associated with the basic protein translation processes (e.g. "structural constituent of ribosome" or "translation") with Benjamini-corrected P values lower than 10^{-20} . GO enrichment analysis was performed using DAVID repository [35] and cross-checked using Metascape [36]. The lists of the most-enriched GO terms are reported in the Supplemental Material (Tables S3 and S4 [25]), where also the links to the full gene lists are presented. The genes in this first regime are quite common across organs, e.g., around 70% of them is in the top 100 highly expressed genes in at least half of the organs. In particular, 35 genes appear in the first regime of the 70% of the organs (Table S5 [25]). These genes present an enrichment for GO terms such as "ribosome" and "ribosome subunit" with P values lower than 10^{-20} . Therefore, the first regime is composed of highly expressed genes related to basic functions. This first core is followed by actively expressed genes that are more tissue specific and whose expression approximately follows the classic Zipf's law with exponent -1 .

C. The average number of detected transcripts follows Heaps's law as predicted by a sampling process

A complex biological system such as an organ is composed of multiple cell types with transcription programs differentiated according to their functional role. Even the repertoire of genes that have to be transcribed is expected to vary from cell to cell as a function, for example, of the level of specialization of the cellular phenotype. Therefore, a basic

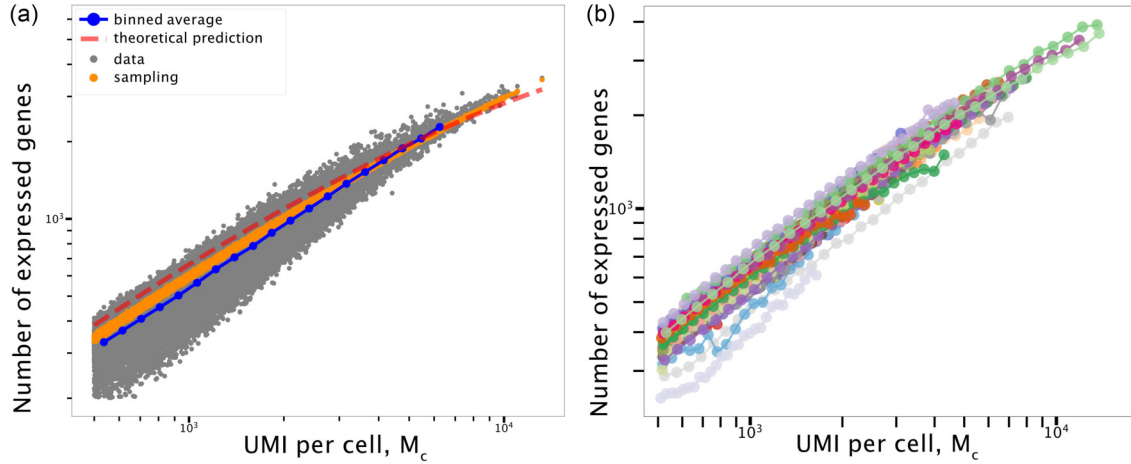


FIG. 3. The number of detected different transcripts follows Heaps’s law. (a) The number of mRNAs with at least one detected transcript $h(M)$ is reported as a function of the transcriptome size as measured by the number of UMIs, i.e., $M^c = \sum_i n_i^c$. Each point in the scatter plot thus corresponds to a single cell, for the illustrative example of cells in the bone marrow from the MCA. The average empirical sublinear scaling (blue dots) is compared to the results of a stochastic sampling process using detailed simulations (orange dots) and analytical predictions from Eq. (8) (red dashed line). (b) The same sublinear average scaling is approximately conserved in all organs reported in the MCA.

observable difference between single-cell expression profiles could be the total number of genes that are actually transcribed. Resuming the analogy with texts of natural language, different texts typically use a different vocabulary (i.e., total number of different words), and the size of the vocabulary can depend on several factors such as the author style or the topic complexity. However, the average vocabulary of texts empirically displays a specific and well conserved sublinear scaling with the text size, known as Heaps’s law [9,37,38]. Again, an analogous law relates the number of different genes or protein domains to the genome size in prokaryotes [10]. Transcriptomic data present the additional complication that the number of detected transcripts also depends on the sampling process due to RNA capture. This naturally introduces a dependence on the sampling efficiency which is proportional to M^c , i.e., the total number of captured transcripts from a cell, c . Figure 3(a) shows the number of different mRNAs as a function of the total number of UMIs as an estimate of the total number of detected mRNAs. This analysis cannot be naturally applied in the absence of UMIs, since a reliable measure of the sample size is needed.

The sublinear power-law scaling is very similar to the one found in other component systems [9,10]. This empirical trend can be compared with predictions from the model presented in Eq. (5). The model assumption is that the probability of observing a specific mRNA i in the sampling process is only determined by its empirical average frequency f_i . It is easy to show [38] that according to this model the probability of not observing a mRNA given the total number of transcripts sampled M is well approximated by

$$P_i(0|M) \simeq e^{-f_i M}. \tag{6}$$

From this expression, we can calculate the expected number of detected different transcripts h as

$$\langle h(M) \rangle = N - \sum_{i=1}^N P_i(0|M) \simeq N - \sum_{i=1}^N e^{-f_i M}, \tag{7}$$

where N is the total number of possible mRNAs, given by the number of genes considered in the experiment, which is around 30×10^3 . The formula above reproduces well the results of direct simulations of the sampling process (see the Methods section for details) reported as orange dots, and also captures quite accurately the empirical average scaling. Therefore, the observed repertoire of expressed genes in these scRNAseq experiments is on average mostly determined by the sampling process. This trend has to be carefully taken into account in order to reliably estimate the biological variability in transcript repertoires.

A quantitative difference between the empirical average number of expressed genes [blue line in Fig. 3(a)] and the expectation from sampling (orange line) can be observed. In fact, the sampling model slightly overestimates the empirical trend. In other words, cells typically express a lower number of genes to a higher expression level than expected. This small discrepancy is linked to the statistics of zero values that will be discussed in detail in the following sections.

As illustrated in Fig. 2, two power-law regimes followed by an exponential decay can be sketched for the expression levels. The model can be simplified by exploiting this observation. Instead of considering all the f_i values as free parameters that have to be inferred from data, we can assume the double power-law scaling, with exponents γ_1 and γ_2 estimated by fitting, and the exponential tail for low frequency components. In this case, it can be shown [38] that the expression for $h(M)$ simplifies to

$$\begin{aligned} \langle h(M) \rangle = N - \sum_{i=1}^{i^*} (1 - Ai^{-\gamma_1})^M - \sum_{i=i^*+1}^{i^{**}} (1 - Bi^{-\gamma_2})^M \\ - \sum_{i=i^{**}+1}^N (1 - Ce^{-ki})^M. \end{aligned} \tag{8}$$

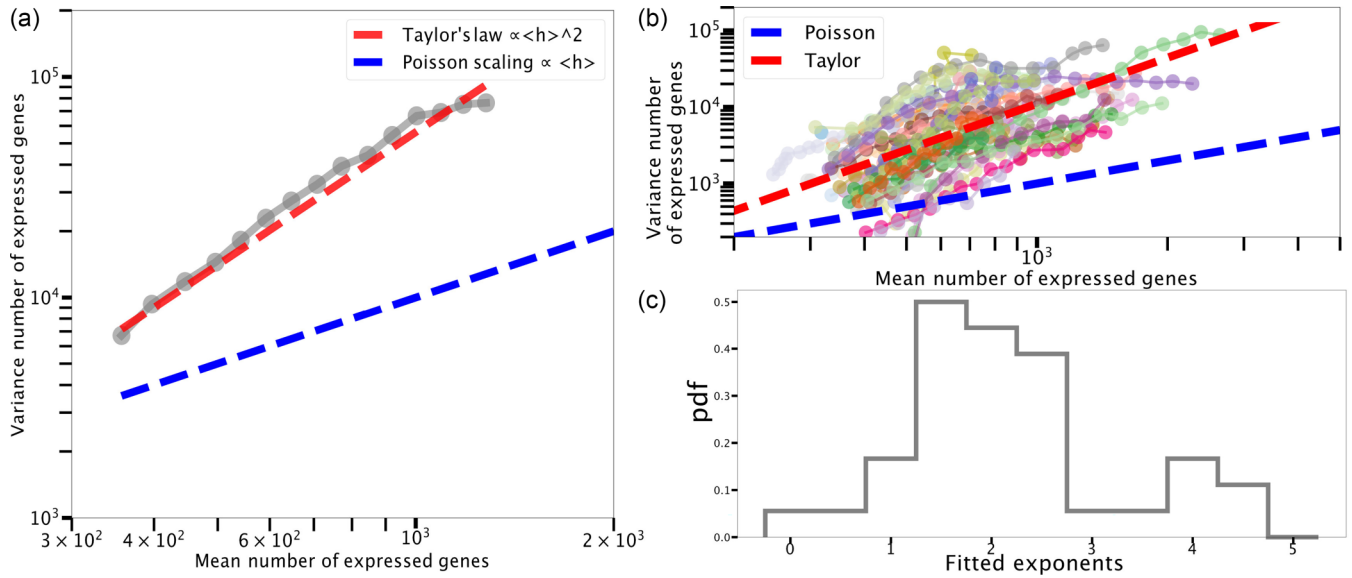


FIG. 4. Fluctuation scaling in the number of detected transcripts follows Taylor's law. (a) The variance in the number of measured expressed genes is reported as a function of its average value for all cells in the MCA. Data are compared to a quadratic scaling (red dashed line) and to the Poisson scaling predicted by a sampling process (blue dashed line). (b) The fluctuation scaling is conserved by considering separately different organs and tissues. (c) Probability density function of the exponents k obtained by fitting the curves in panel (b) with $C\langle h \rangle^k$.

The factors A , B , C are defined by imposing normalization and continuity conditions between the three regimes:

$$\begin{aligned} A(i^*)^{-\gamma_1} &= B(i^*)^{-\gamma_2}, \\ B(i^{**})^{-\gamma_2} &= C e^{-k(i^{**})}, \\ A \sum_{i=1}^{i^*} i^{-\gamma_1} + B \sum_{i=i^*}^{i^{**}} i^{-\gamma_2} + \sum_{i=i^{**}}^N C e^{-ki} &= 1. \end{aligned} \quad (9)$$

i^* is the rank at which the change of power-law exponent is estimated, while i^{**} is the rank at which the exponential regime starts. This is the theoretical prediction reported as a dashed red line in Fig. 3(a).

If the sampling process is the dominant factor setting the repertoire of observed transcripts, the trend should not depend crucially on the biology of the system in analysis. Indeed, the sublinear scaling is well conserved across different organs as reported in Fig. 3(b).

D. Variability in the repertoire of expressed genes follows Taylor's law and reveals deviations from a sampling process

As discussed in the previous section, the scaling of the average number of detected genes can be well explained as a result of the sampling process. However, there is substantial variability in the empirical data, i.e., cells with the same total number of UMIs can have expression repertoires of largely different sizes. The question is if this variability can be again explained as sampling fluctuations. The model provides a precise prediction for the variance σ_h^2 as a function of the average value $\langle h \rangle$. Fig. 4(a) compares the model prediction of a Poisson scaling (blue dashed line) with the empirical scaling (grey dots) evaluated over all the cells in the MCA dataset in order to have large statistics. The empirical variance displays a power-law scaling with the average vocabulary size that is not compatible with a Poisson scaling. Fitting the

empirical scaling with the function $C\langle h \rangle^k$ leads to an exponent $k = 1.64 \pm 0.18$. This value is significantly different from the Poisson scaling expected from sampling ($Z = 3.5$) and more compatible with the quadratic scaling ($R^2 = 0.94$ and $Z < 2$) that has been observed for several other complex systems [11,39,40].

Focusing on single cells belonging to the same organ the phenomenology is quite diversified, also due to the reduced cell numbers [Fig. 4(b)]. With the caveat that the reduced statistics makes the fitting procedure less robust, we can still fit the organ-specific fluctuation curves with the function $C\langle h \rangle^k$ finding a distribution of exponents peaked on 2 [Fig. 4(c)]. Although the distribution is quite large, this suggests that an approximately quadratic scaling is an inherent property of the transcriptome diversification and it is not only due to differences between organs. Interestingly, we have found an emergent statistical law that cannot be explained by the sampling process inherent to RNA sequencing, and that can thus contain information on biological variability. However, this quadratic fluctuation scaling is yet again a common feature of several complex component systems, from linguistics to ecology, known as Taylor's law [11,39,40]. Therefore, a general explanation, which goes beyond the specific properties of expression profiles, could be at the origin of this scaling.

E. Poisson noise sets the lower bound and the scaling of gene expression cell-to-cell variability

A commonly analyzed property of cell-to-cell variability in single-cell expression studies is the coefficient of variation ($CV = \sigma_n / \langle n \rangle$) of gene expression levels across cells. In particular, the CV_i^2 of each gene i is often reported as a function of the mean gene expression level in order to identify highly variable genes at a given average value. In fact, this

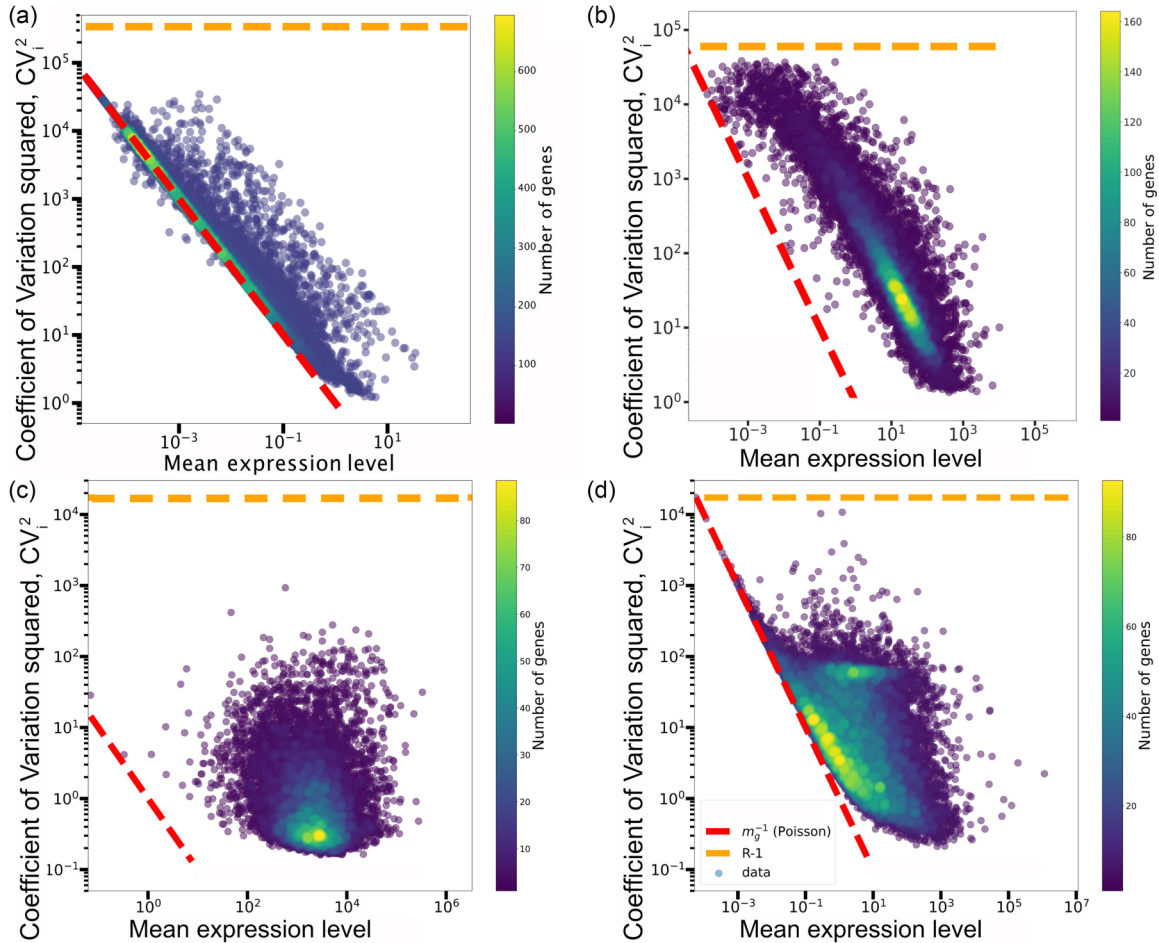


FIG. 5. The coefficient of variation CV_i^2 as a function of the average expression. The red dashed lines report the Poisson scaling, i.e., $\sigma_n^2 = \langle n \rangle$. The horizontal orange dashed lines correspond to the maximum possible value of $CV^2 = R - 1$, which is achieved if a gene is expressed in only one cell. The different panels correspond to data from (a) Mouse Cell Atlas, (b) Tabula Muris, (c) GTEX limited to protein coding genes, (d) GTEX limited to noncoding genes. As explained in the Methods section, for the MCA database we considered the UMI counts, while for Tabula Muris and GTEX the raw read counts are reported.

criterion is often used to reduce the number of features (i.e., genes) to consider in further analysis [41,42]. We analyzed this fluctuation scaling in the two scRNAseq atlases, and in a bulk RNAseq large-scale experiment (GTEX; see the Methods section) for comparison.

Figure 5(a) shows the CV_i^2 scaling for cells in the MCA. The red dashed line is the analytical prediction (confirmed by simulations) of the expected scaling for a sampling process, which is basically a Poisson scaling. The measured values follow essentially the same scaling, but the observed gene expression fluctuations are larger than the Poisson prediction, which essentially sets the lower bound of measurable variability.

The analogous plot for the Tabula Muris dataset displays the same scaling but with a clear shift. This can be simply explained by the amplification process used before sequencing. In fact, a Poisson random variable multiplied by a constant has a translated CV^2 .

In general, the random variable representing the observed mRNA counts x for the different genes has average value $\langle x \rangle = \mu$ (set by the true average expression level) and a variance at least with a Poisson scaling $\sigma_x^2 = c \mu$ introduced

by the sampling process. The fluctuations can be larger depending on the true expression distribution, hence the unknown factor c that could have an additional dependence on μ . We can now define the new variable $y = kx$, where k is a constant describing a supposedly constant amplification factor. The mean and variance of y are simply given by $\langle y \rangle = k\mu$ and $\sigma_y^2 = k^2 c \mu$. Therefore, the CV^2 can be written as a function of $\langle y \rangle$ as

$$\begin{aligned} \log_{10}(CV_y^2) &= -\log_{10}\left(\frac{k^2 c \mu}{k^2 \mu^2}\right) \\ &= -\log_{10}\langle y \rangle + \log_{10}(kc) \\ &\simeq -\log_{10}\langle y \rangle + \log_{10}(k), \end{aligned} \quad (10)$$

where the last equality derives from the observation [Fig. 5(a)] that when there is no amplification, i.e., $k = 1$, most of the genes display a true Poisson scaling, thus $c = 1$. This naturally implies that amplification leads to a simple translation of the CV^2 scaling, precisely as in Fig. 5(b).

Figures 5(c) and 5(d) are instead obtained using data from GTEX (see the Methods section) and each point represents a tissue in a bulk RNA-sequencing experiment. In bulk RNA

sequencing, the sampling is performed on a large sample of cells, and we are essentially averaging over N cells the random variable representing the gene expression value. Indeed, the sampling process is performed on the sum of mRNAs from all N cells, and finally the counts are typically normalized with a factor proportional to the number of cells. Therefore, in bulk RNAseq experiments, the normalized number of detected transcripts of a gene should have mean $\langle x \rangle \simeq \mu$, as simply set by the average expression in single cells, and variance $\sigma_x^2 \simeq \frac{\mu}{N}$, which is the variance of the average of N independent random variables with a Poisson scaling. The amplification again introduces a constant factor k , and the CV_y of the new amplified observable y (with average $\langle y \rangle = k\mu$) is

$$\begin{aligned} \log_{10}(\text{CV}_y^2) &= \log_{10}\left(\frac{k^2\mu}{N} \frac{1}{k^2\mu^2}\right) \\ &= -\log_{10}\langle y \rangle + \log_{10}\left(\frac{k}{N}\right). \end{aligned} \quad (11)$$

Equation (11) shows that the amplification factor k is suppressed by the term N (i.e., the number of cells sequenced) in the CV^2 expression for bulk RNAseq experiments. This explains why in Fig. 5(d) the data still show the Poisson trend without the translation observed for single cells [Fig. 5(b)].

The CV^2 has a natural upper bound that is reached when a gene is expressed in only one cell. If the only not-zero count is n , the average expression is n/R (where R is the number of cells), and the CV^2 is $R - 1$, which does not depend on n . This upper bound is reported as a dashed orange line in Fig. 5. Note that the highest variability reported in the Tabula Muris atlas closely approaches the bound.

In bulk RNA sequencing data, the effects of sampling and stochasticity in gene expression should be averaged out by extracting RNA from a large number of cells. In fact, the CV profile is radically different if calculated on protein-coding genes in the GTEx dataset [Fig. 5(c)]. The empirical variability is far from the sampling limit and the CV seems typically weakly dependent on the average expression level. However, focusing on noncoding genes [Fig. 5(d)], which are typically lowly expressed, we observe the Poisson trend emerging again. This indicates that, as expected, sampling has to be carefully taken into account when the variability of lowly expressed genes is analyzed even in the context of bulk RNAseq data.

This analysis shows that the sampling process sets a lower bound on the measured variability in gene expression, and data from single-cell RNA sequencing are generally close to this bound. Clearly, the observed expression variability is not captured by our simple null model which focuses on sampling and thus can only produce Poisson distributions.

Note that a Poisson scaling could also be explained by a simple model of stochastic gene expression in which transcription and degradation are modeled as reactions with constant rates [43,44]. This description also leads to Poisson distributions for mRNAs and thus to the same scaling of the CV. However, transcription can be more complex than a birth-death process, for example it can be characterized by bursts of expression [45]. Models accounting for bursty production naturally lead to overdispersed expression distributions such the negative binomial distribution (or its continuous analog

gamma distribution), or even to more complex bimodal distributions [45–47]. Moreover, the presence of extrinsic noise, i.e., fluctuations in global cellular factors [48], can induce a constant CV with respect to the average expression [49], and there could be an additional basal technical noise besides sampling fluctuations in scRNAseq data [26]. Indeed, empirical CV^2 values typically present a double scaling, with a Poisson-like dependence for lowly expressed genes (where intrinsic noise and/or sampling effects are relevant) and a constant “floor” noise at higher level of expression both for proteins and for mRNAs [27,49]. This constant floor noise is also evident in bulk transcriptomic data (Fig. 5), as well as for highly expressed genes in scRNAseq datasets, although the double scaling is more evident for single-cell datasets obtained with the Smart-seq3 technology that will be introduced in a following section. A constant CV is precisely equivalent to a fluctuation scaling of the type $\sigma_n^2 \propto \langle n \rangle^2$. While this functional form is again reminiscent of Taylor’s law [39], gene expression fluctuations are not equivalent to the fluctuations in the number of expressed genes, i.e., the fluctuations around Heaps’s law that we previously described (Fig. 4). In other words, a quadratic scaling of expression fluctuations, for example due to stochastic gene expression, does not necessarily induce a quadratic scaling of the fluctuations around Heaps’s law. A simple model experiment can be used to prove this point. We assume that genes have gamma-distributed expression levels, with mean values following a Zipf-like law. We also assume that the CV of the expression levels is constant, thus it follows a Taylor’s law for expression fluctuations, by appropriately fixing the variance of each gamma distribution. These gamma-distributed expression levels are then randomly sampled to obtain a CV^2 that displays a double scaling as in many empirical observations (Fig. S9A [25]). However, the fluctuations in the number of expressed genes, i.e., the analog of Fig. 4, can still show an approximately Poisson scaling (Fig. S9B [25]). Therefore, more realistic models of stochastic gene expression can in principle explain the empirical levels and the scaling properties of expression fluctuations, but they do not necessarily reproduce the Heaps’s law fluctuations.

F. The statistics of transcript sharing

While the repertoire of expressed genes can be highly cell specific, it is natural to expect a certain degree of overlap between the genes that have to be expressed in different cells. This overlap should depend on the specific gene functions and on the similarity of the cell types in analysis. For example, we intuitively expect a core set of genes, linked to basic cellular functions, to be expressed in essentially every cell. In order to quantify the statistics of the overlaps between the expression profiles of different cells, we analyzed the occurrence distribution. The occurrence o_i of a transcript is defined as the fraction of cells in which it is detected (i.e., it has a nonzero count). Figure 6(a) reports the occurrence distribution for cells belonging to a single tissue (the bone marrow in the example). Surprisingly, most of the genes appear to be expressed in very few cells and the number of genes expressed in all cells seems negligible. However, a quantitative comparison with the null model suggests that this is mainly an effect of the sampling process. In fact, given the empirical average expression levels

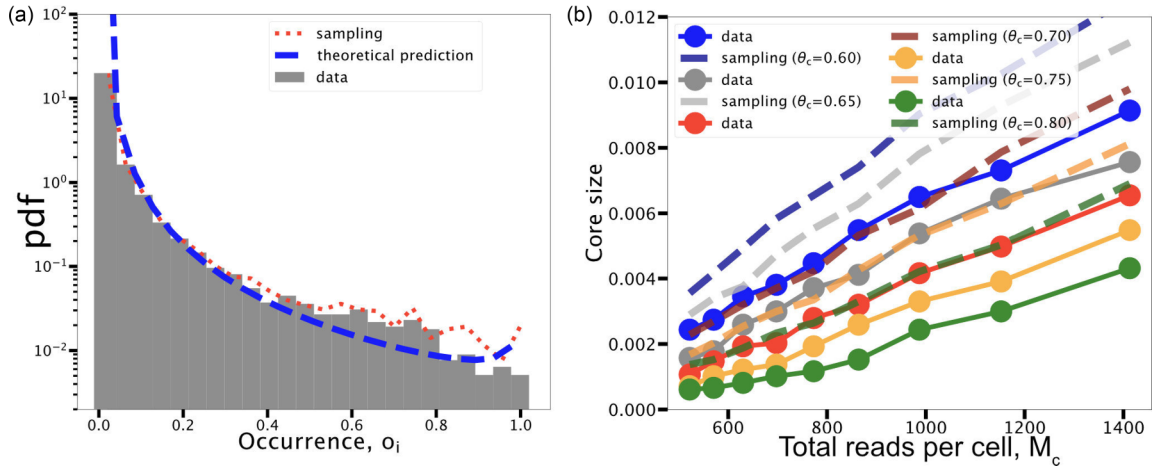


FIG. 6. The occurrence distribution of gene expression. (a) The probability of observing a mRNA in a fraction o_i of cells is reported for the illustrative example of cells in the bone marrow as profiled in the MCA. (b) The empirical fraction of genes expressed in at least θ cells (different θ values correspond to different curves) is compared with the corresponding predictions of the sampling process.

f_i , the sampling model [Eq. (5)] gives the expected occurrence for each gene i as

$$o_i = 1 - \frac{1}{R} \sum_{c=1}^R P_i(0|M^c) \simeq 1 - \frac{1}{R} \sum_{c=1}^R e^{-f_i M^c}. \quad (12)$$

From this expression, the probability density of the occurrences can be extracted. It is reported as a dotted red line in Fig. 6(a) and provides a good approximation of the empirical distribution. An equivalent result can be obtained from direct simulations of the sampling process.

As previously shown [6], the occurrence distribution takes a particularly simple functional form if we approximate the distribution of relative expression levels with a single power law ($f_i \sim i^{-\gamma}$, with $\gamma \simeq -0.8$), and we assume that all cells have the same average number of total UMIs ($M \simeq 1500$ transcripts in this case). The resulting expression is

$$p(o) = \frac{(1-o)^{\frac{1}{M-1}}}{\gamma M N \alpha^{1/\gamma} [1 - (1-o)^{\frac{1}{M}}]^{1/\gamma+1}}. \quad (13)$$

Despite the crude approximations, this analytical prediction [blue dashed line in Fig. 6(a)] can still reproduce reasonably well the data, and can thus be used in general for an easy first prediction of the effect of sampling noise on mRNA occurrences. While a simple Poisson sampling process can largely explain the shape of the occurrence distribution, there are quantitative differences. In particular, there is a clear difference between the two distributions for high occurrence levels [Fig. 6(a)]: ubiquitously expressed genes, or core genes, seem under-represented in the data. A more detailed comparison can be done by explicitly looking at the core size and how it scales with the total number of sequenced transcripts M . The core size c can be defined as the fraction of genes expressed in at least a fraction θ of the cells in the population, i.e., genes with $o_i > \theta$. Considering again the approximation of a power-law distribution of average expression levels, the sampling process predicts a specific scaling for the core size with the sample size M [6]. The core size is indeed described

by the expression

$$c(M) = \frac{M^{\frac{1}{\gamma}}}{\alpha^{\frac{1}{\gamma} N}} [-\log_{10}(1-\theta)]^{-\frac{1}{\gamma}}, \quad (14)$$

where $\alpha = \sum_i i^{-\gamma}$ is a normalization and $\gamma \simeq 0.8$ is estimated from data. Thus, the scaling is expected to be approximately linear if γ is close to 1. This qualitative prediction is confirmed by empirical data [Fig. 6(b)]. In this plot, core sizes, defined by different values of θ , are measured over cells with a different number of detected transcripts M [dots in Fig. 6(b)]. The empirical scaling can be compared with direct simulations of the sampling process or with the equivalent numerical integration of Eq. (12) [dashed lines in Fig. 6(b)]. The linear trend described by Eq. (14) is observed in both data and simulations. However, the empirical curves have slightly smaller slopes and they systematically show smaller core sizes. In other words, given the average gene expression levels, there is a smaller than expected number of genes that can be detected in a large fraction of the cell population. The origin of this discrepancy is closely related to the statistics of zero values in scRNAseq datasets that will be addressed in more detail in the next section.

G. Predicting presence from transcript abundance and the statistics of zero values

The sparsity of scRNAseq data and the possible origins of the detected zero values have been, and still are, an active field of research and debate [16,17,50]. Using our simple null model, we can identify what level of data sparsity is expected from sampling only. Moreover, we can isolate genes whose zero statistics is unexpected, and thus possibly linked to biological variability or technical noise not included in our description. As discussed in the previous section, the sampling model provides a prediction for the occurrence o_i , i.e., the number of cells in which the count is not zero, for each gene [Eq. (12)]. This prediction can be directly compared with the empirical occurrence as in Figs. 7(a) and 7(b). The first observation is that the density of points is mostly located on the diagonal, where the number of zero

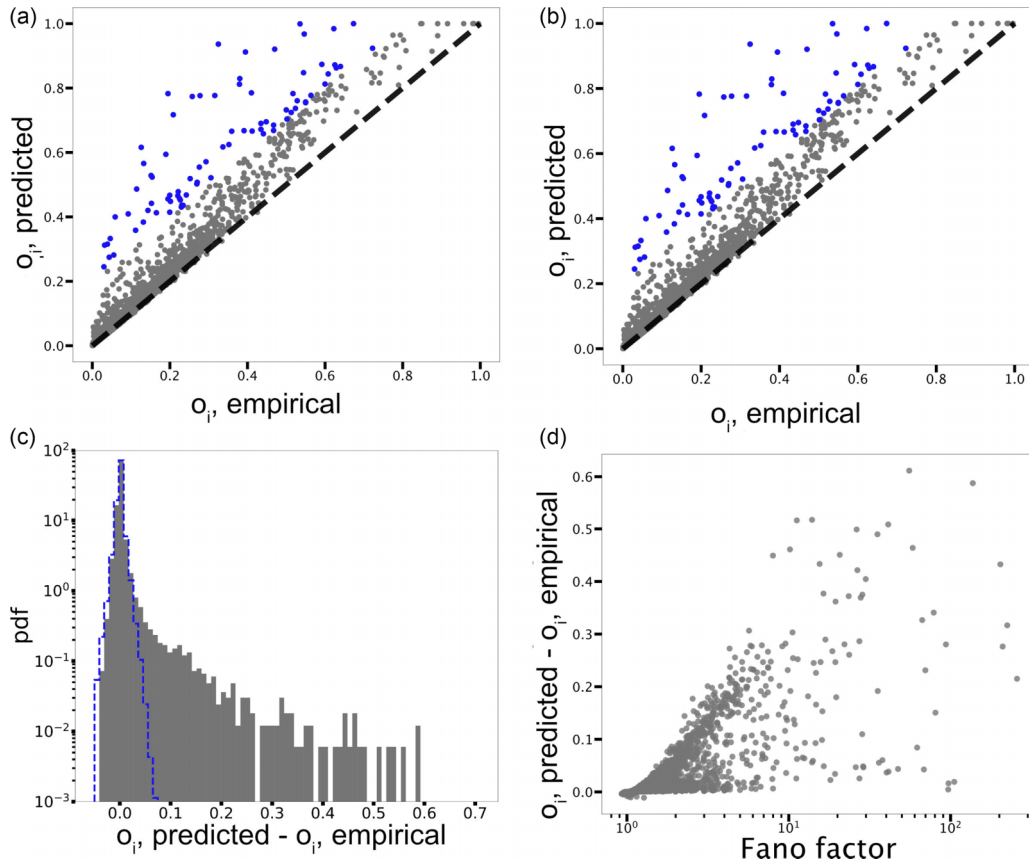


FIG. 7. Explaining the empirical occurrences from average expression and sampling. The number of cells in which a transcript is present, $o_{i,\text{empirical}}$, is reported with respect to the occurrence predicted by sampling $o_{i,\text{predicted}}$ for (a) bladder and (b) muscle. (c) The distribution of $o_{i,\text{predicted}} - o_{i,\text{empirical}}$ is extremely peaked on zero, but with a right tail indicating a higher level of sparsity in empirical data. The blue dashed line represents the distribution of the typical occurrence deviations between sampling realizations, thus providing an estimate of the deviations that are compatible with sampling fluctuations. (d) Scatter plot of the relation between the Fano factor ($\sigma_n^2/\langle n \rangle$) of expression levels and $o_{i,\text{predicted}} - o_{i,\text{empirical}}$ in muscle as an illustrative example.

values predicted coincides with the sampling expectation. In fact, the probability density of the differences between the predicted and the empirical occurrences is extremely peaked on zero as reported in Fig. 7(c) (note that the y axis is in logarithmic scale). Therefore, the zero values are precisely those expected from sampling for most genes, and this result suggests that complex zero-inflated models, which are often introduced to capture the data sparsity [51,52], are generally not needed. This observation is in line with recent analysis of scRNAseq data based on UMIs [17,27,52]. However, the fraction of data points that deviates from the diagonal are mostly above it, showing that indeed genes whose zero count statistics is not well described by sampling have an excess of zero values. This leads to a general level of data sparsity that is slightly underestimated by the model. For example, 97% of entries in the Muscle dataset reported in Fig. 7(b) are zero values, while the sampling process predicts 96% null entries on average. Such a minor discrepancy should not be surprising since our model is an intended oversimplification of the system. For example, we are not considering the inherent variability due to stochasticity in gene expression. We are approximating the true gene expression distributions as delta functions (see Methods section), which become Poisson

distributions only through the sampling process. The underestimation of expression variability is explicitly depicted in Fig. 5, where the variance of empirical expression values is often larger than Poisson. The Fano factor or index of dispersion (i.e., $\sigma_n^2/\langle n \rangle$) can be used to quantitatively measure how large is the deviation. In our case, it measures how far is a gene expression variability from the sampling prediction. As intuitively expected, the Fano factor is correlated with the difference between the predicted and observed occurrences for each gene [Fig. 7(d)]. However, the scatter plot does not show a clear and simple relation, thus suggesting a complex interplay between expression variability and zero count statistics. Given its inherent simplifications, the model provides a simple and quantitative way to select the genes whose zero value counts are “atypical” and thus that are potentially interesting for further analysis [53]. This excess of zero values can derive from technical reasons (often called “dropouts”), from biological variability due to cell-type heterogeneity or eventually from particularly noisy promoters. Subsequent analysis of the selected genes could select the most likely contributions. While we leave this step for future work and specific applications, we propose an illustrative example. We selected the genes with the top values of $o_{i,\text{predicted}} - o_{i,\text{empirical}}$ [de-

picted in color in Figs. 7(a) and 7(b)] in different organs and performed a GO enrichment analysis. Some categories are over-represented. The presence of enriched GO categories already indicates that if a dropout phenomenon is present, it is not random across genes. All the significantly enriched categories are reported in Table S6 of the Supplemental Material [25] together with links to the full gene lists. Highly enriched categories could indicate biological signals as well as technical reasons not captured by a simple sampling. For example, the presence of general categories such as “ribosomal proteins,” “extracellular exosome,” or “blood microparticle” could be due to the insertion of few zero values at the initial filtering procedure. Ribosomal proteins are expected to be expressed in essentially every cell type at a relatively high level. This is generally the case in the data, since those genes are in the high-rank region of Zipf’s law (Table S2 [25]). Therefore, the few empirical zero counts cannot be explained from sampling given the high average expression levels. In this example, the sampling model could be a potentially useful check of the technical procedures. Note that sampling is a stochastic process, thus the predicted number of zeros for a given gene has a confidence interval determined by sampling fluctuations. These fluctuations can be evaluated using an ensemble of sampling realizations and measuring the distribution of gene occurrences across the ensemble and the typical differences between sampling predictions that can be compared with the empirical ones [blue dashed line in Fig. 7(c)]. For example, the small fraction of genes that are detected in more cells than expected given their average expression [negative values in Fig. 7(c)] can be explained by sampling fluctuations. As a further test of the significance of analyzing the deviations from our data-driven null model, we consider a few datasets composed of cells from different organs in the MCA, and we focus again on genes whose zero value statistics significantly differ from expectation. The rationale is that genes whose expression is, for example, tissue specific, will typically have expression distributions far from the null model expectation when they are evaluated over cells belonging to different organs. This would also reflect in their atypical zero value statistics. In fact, the genes selected based on their anomalous occurrences (as the coloured dots in Fig. 7) across cells from “brain” and “ovary” are significantly enriched in the GO term “myelin sheath” (P value 6.8×10^{-16}) which is clearly a tissue specific function that characterize only a fraction of the cells in analysis. The same GO term appears when cells from “brain” and “muscle” are considered (P value of 8.4×10^{-11}). Considering cells from “muscle” and “blood” instead leads to the selection of genes significantly associated with “haptoglobin” and “hemoglobin” (P values $< 10^{-6}$). The GO term “spermatogenesis” is associated with genes with atypical occurrences in a dataset joining cells from “ovary” and “testis.” The full lists of the selected genes in these illustrative examples are reported in the Supplemental Material [25] (Tables S7–S10). Besides gene selection, a dataset of overall deviation from the prediction of the corresponding null model should be related to the inhomogeneity of the cellular transcription programs it includes, thus ultimately to the “complexity” of the cell population analyzed. As a preliminary test of this hypothesis, we focus again on the number of detected genes in datasets composed of cells be-

longing to a different number of organs. The number of organs included can be used as a rough measure of the dataset inhomogeneity. We observe a clear correlation between this number and the statistical deviation from the sampling model (Fig. S10 [25]). Although we are only focusing on the discrepancy between model and data in the gene occurrences (thus essentially on the Heaps’ law), this result suggests that the null model is a useful tool to measure and quantify intrinsic properties of the dataset.

H. Checking the robustness of the statistical laws

As a further test of the robust emergence of the described scaling laws, we analyzed two additional datasets of cells profiled with the recently introduced protocol Smart-seq3 [24]. The Smart-seq3 protocol combines high sensitivity with the use of UMIs, providing reliable molecule counts and data matrices typically with a lower degree of sparsity. Despite the differences in the protocol and in the cell types considered, the same phenomenological laws reported for the large-scale Mouse Cell Atlas are clearly observable. Also in this case, the general trends can be framed in our analytical framework and partially explained by a sampling process. Figure 8 shows some of these laws for the example of a HEK cell line, while the analogous results for mouse fibroblasts are reported in Fig. S11 [25]. The rank plot of the average expression levels is again a Zipf-like law characterized by three clearly distinguishable regimes with a central power-law scaling with exponent close to -1 [Fig. 8(a)]. Interestingly, the UMI-based datasets present a fitted exponent with values very close to the classic -1 (Table S2 [25]). The trend predicted by the sampling process for Heaps’ law [Fig. 8(b)] is compatible with the empirical number of detected transcripts, with a slight sampling overestimation that is linked to the zero-value statistics. The cell-to-cell variability in expression levels shows a Poisson scaling for lowly expressed genes. However, given the higher sensitivity of the Smart-seq3 protocol, the regime of approximately constant CV, which was observed for bulk data (Fig. 5), is now detectable for highly expressed genes, where sampling effects are less dominant. Interestingly, this double scaling of the CV^2 is analogous to the one reported for protein fluctuations in large-scale single-cell experiments based on fluorescence [49,54].

In the cell line considered for Fig. 8, the empirical fraction of zero values is 45%, while the sampling expectation is 41%. Therefore, the zero statistics is again largely explained by sampling effects. Indeed, the occurrence distribution is largely recapitulated by the sampling model [Fig. 8(d)]. Thanks to the protocol’s higher sensitivity, leading to lower data sparsity and larger realization sizes M , occurrences clearly display the typical U-shaped distribution that robustly emerges in several complex component systems [6]. The observable deviations from the model only derive from the small fraction of transcripts that present more zero values than expected [Figs. 8(e) and 8(f)], in perfect analogy with our results for the MCA. Typical occurrence fluctuations only due to sampling [blue line in Fig. 8(d)] are expected to be larger for this dataset with respect to the MCA because of the lower number of cells profiled (i.e., 117 cells for the HEK cell line). For this reason, we considered the fibroblasts dataset which contains 369 cells

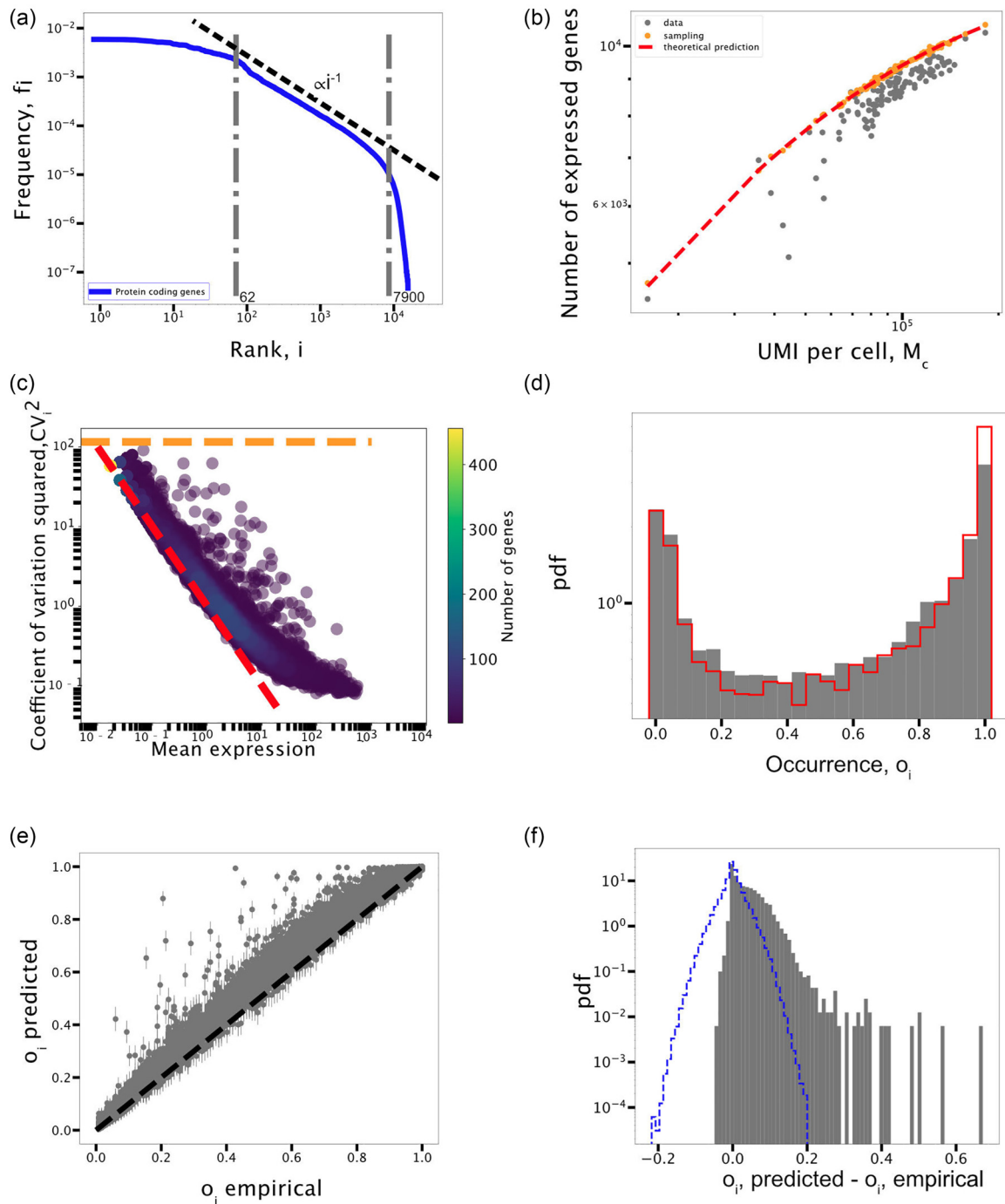


FIG. 8. Emergent statistical laws from a HEK cell line profiled with Smart-seq3. (a) Zipf's law with the three scaling regimes. (b) Heaps's law: number of detected transcripts as a function of the total number of UMIs per cell. (c) Expression variability CV^2 as a function of the average expression level for all detected transcripts. (d) The empirical transcript occurrence distribution compared with the model expectation (red continuous line). (e) The relation between the empirical occurrences $o_{i,\text{empirical}}$ and the expected values from the sampling model $o_{i,\text{predicted}}$. Error bars represent the variability (one sigma) between different realizations of the sampling process. (f) Distribution of $o_{i,\text{predicted}} - o_{i,\text{empirical}}$. As in Fig. 7(c), the blue dashed line identifies the differences compatible with sampling fluctuations.

to select genes with more zeros than expected and perform GO enrichment analysis. The results show cell-cycle and cell-division related terms as enriched (Table S11 [25]). Indeed, in a nonsynchronized proliferative cell line such as the one in analysis, we should expect nontrivial expression distributions for genes related to the cell-cycle progression [24].

As an additional robustness test we also considered single cells profiled with the alternative 10x genomics protocol in the Tabula Muris database. Figure S12 [25] reports the summary of the statistical laws obtained for bone marrow cells as an illustrative example. Basically, there are no significant differences in the emerging laws (by comparing for example with

Fig. 8), suggesting that they do not crucially depend on the specific sequencing protocol (as long as UMIs are present to remove amplification factors).

In conclusion, the presented statistical laws seem to be a robust emergent property of single-cell RNAseq data. The proposed mathematical framework provides an explanation for most of the general trends, and thus can be a useful simple null model to identify significant deviations of biological or technical origin.

IV. DISCUSSION

The identification of statistical laws is a key step in designing effective descriptions of complex systems [55]. Leveraging large-scale regularities, phenomenological models can be built, in the spirit of statistical physics, to capture relevant system properties without focusing on a detailed description of the high number of degrees of freedom. For example, the presence of quantitative empirical laws in cell composition of fast-growing bacteria has led to simple models of cell physiology that can explain several large-scale gene expression patterns using just a few key parameters such as the growth rate [56]. Analogously, the emergence of different cell identities and their organization in tissues and organs is driven at the molecular level by the complex orchestration of the expression of large sets of genes. However, simple coarse-grained descriptions can be hopefully extracted without resorting to all the molecular details. As a first step in this direction, we identified several statistical laws emerging in single-cell transcriptomic profiles using large-scale expression atlases of mouse tissues. Strikingly, analogous laws are ubiquitously found in different complex component systems from linguistics to ecology [6,9,11,57].

An additional complication of scRNAseq data is the presence of a sampling process inherent to the experimental technique. Therefore, the observed expression statistics is due to a combination of natural cell-to-cell variability and stochastic sampling. We focused on modeling the sampling process given a basic system property, which is the specific average heterogeneity of gene expression levels described by the classic Zipf's law. This law is apparently a hallmark of several component systems [29,58]. While it has been previously reported for gene expression values [13,14,31], we showed that it is an intrinsic property of single cells robustly emerging in different datasets, and that different regimes can be identified apparently related to the gene functions.

The proposed simple model essentially neglects biological expression fluctuations and tests what can be explained from sampling only. In this framework, there is a natural predicted connection between Zipf's law and other statistical regularities such as Heaps's law and the U-shaped statistics of shared components [6,38]. We first showed that indeed these additional empirical laws emerge in transcriptomic data, and second that they can be well explained as consequences of stochastic sampling. This result suggests that downstream analyses typically performed on these datasets, such as clustering to identify cell types or expression fold-change analysis, have to carefully take into account sampling and the statistical regularities it generates.

However, we identified some clear deviations from sampling predictions. Specifically, the empirical variability in the cell expression repertoires, captured by the fluctuation scaling of the Heaps's law, cannot be reproduced by the model. This result could conceal a biological motivation linked to the differentiation of expression programs in different cell types. However, the very same scaling is a recurrent feature of several complex systems, often called Taylor's law [11,39,40], suggesting a more general mechanism behind its emergence.

This fluctuation scaling is closely linked to the statistics of zero values, which is a central theme in scRNAseq data [16,17,50,52]. In this regard, we first showed that the vast majority of zero counts in the data can be simply explained as a sampling effect. Therefore, there is not a clear indication that complex (and parameter rich) models, such as zero-inflated models, are needed to capture the technical noise [52]. Despite this general trend, the model is a tool to identify specific deviations. A possible application of a data-driven null model that captures general statistical properties of scRNAseq data is to focus on the empirical deviations from its expectation. For example, in several datasets a fraction of genes are expressed in less cells than expected from sampling (i.e., they have an excess of zero counts). We discussed few examples in which the atypical expression distribution of these genes can be explained by their associated biological functions, suggesting the potential use of our null model for gene selection. Interestingly, a similar approach to identify informative genes by comparing their expression variability to a simple random null model was recently proposed [59], precisely leveraging the analogy between linguistics and transcriptomic data.

The sampling model provides essentially a lower bound for the number of zero values of a transcript given its average expression level. This should be expected since the expression variability in the model only derives from sampling, and thus does not match the typical CV values observed.

The subsequent step, which we leave for future work, would be to include progressively more realistic models of the stochastic process of gene expression in order to leave out from the description only the cell-to-cell variability coming from the diversity of gene expression programs in the cell population. The price of increasing the complexity of null models is that more parameters have to be introduced and inferred from data. More realistic models of gene expression simply correspond to the selection of an appropriate distribution ρ in Eq. (3). As previously discussed, a natural choice could be the gamma distribution, since it is a good description for bursty stochastic gene expression [46]. If $\rho(f)$ is a gamma distribution and the sampling is still a Poisson process, it is easy to show that the expected distribution for the observed counts should be a negative binomial distribution [52]. The negative binomial is indeed the standard overdispersed distribution often used to fit RNA-sequencing data [17]. While we leave to future work the quantitative analysis of the different emerging statistical laws using this model, we can anticipate that a gamma-based model can generally better reproduce the statistical properties of empirical data, including the zero-value statistics already quite well captured by sampling alone. This is not surprising given the larger number of degrees of

freedom with respect to the sampling model analyzed here. Therefore, we should expect a better fit of the U-shaped occurrence distributions and a more precise estimate of the core size.

Statistical laws have also been observed and studied using complex systems approaches at the basic level of DNA sequences [60–62]. The possibility of a link between general statistical properties at the nucleotide level and the emerging laws for expression patterns here described is captivating, but still to be explored.

Finally, this work adds single-cell transcriptomics to the list of complex component systems displaying statistical laws that are seemingly universal. However, the specificity of transcriptomic data can provide useful indications and constraints to the research of general models and principles behind these laws. Many of the models proposed for the emergence of Zipf's law in component systems are based on a stochastic growth process. Some examples are classic models based on the Yule-Simon process, on the Chinese restaurant process, on Polya urns, or on the preferential attachment principle [10,33,63]. Basically, these generative mechanisms assume a reuse or duplication of existing components proportional to their current frequencies, and a parallel innovation process that adds new components from a vocabulary. These simple ingredients (with some general prescriptions) are sufficient to reproduce Zipf's law and the average sublinear scaling of Heaps's law. The recently proposed sample-space-reducing process can be also ascribed to this class of stochastic growth models [38,64]. The description can be appropriate for texts that are generated by the writing process through the progressive addition of words, or for the evolutionary processes that shaped genome composition by duplicating, removing or discovering/transferring new genes. However, the composition of a cell transcriptome is not naturally described by this type of processes, since single transcripts are not progressively added in the cell.

Few alternative compelling mechanisms have been proposed that do not rely on a growth process and could thus apply to the case of transcriptomic data. A possibility is that components have specific networks of dependencies and that these functional relations determine their co-occurrence in a realization [8,28,65]. In the transcriptomics case, this would translate in an underlying unobserved network of gene-gene dependencies for example due to correlated functions. Models

based on this network assumption can generate Zipf's and Heaps's laws [28]. Even more generally, power-law distributions can naturally arise if the observed variables (i.e., the expression levels) are affected by fluctuating latent variables that govern the hidden structure behind the data [66,67]. Gene expression is controlled by several latent factors that defines the state of the cell and are not directly observed in transcriptomic datasets. These latent factors can be highly variable and thus can naturally generate Zipf's law under certain quite general conditions. One simple example of a hidden variable is the physiological state of the cell, for example described by the growth rate, which is known to strongly influence the gene expression program and the behavior of different genetic circuits [68,69]. Analogously, the cell-cycle stage, the cell type, or the slowly varying concentration of key enzymes can in principle represent latent variables that have a specific variability in our system and affect gene expression.

The code and the Jupyter notebooks needed to reproduce the analyses and the figures described in this study can be found in a GitHub repository [70]. The datasets analyzed during the current study are available from independent previously published studies

(i) in the Mouse Cell Atlas repository [19]. Data are also available on Gene Expression Omnibus (GEO) through the accession number [GSE108097](#);

(ii) in the Tabula Muris repository [21];

(iii) in the Genotype Tissue Expression (GTEx) project [23] repository;

(iv) Smart-Seq3 data have been deposited by their authors under ArrayExpress [E-MTAB-8735](#) at the European Bioinformatics Institute.

ACKNOWLEDGMENTS

We would like to thank Jacopo Grilli, Matteo Cereda and Sarah Perrone for useful discussions.

M.O. designed the research. S.L. and F.V. performed the analyses and generated the figures. M.O., S.L., and F.V. wrote the article. M.O., S.L., F.V., A.M., A.S., and M.C. read and edited the manuscript. All authors contributed to the article and approved the submitted version.

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

-
- [1] A. Wagner, A. Regev, and N. Yosef, *Nat. Biotechnol.* **34**, 1145 (2016).
 [2] E. Shapiro, T. Biezuner, and S. Linnarsson, *Nat. Rev. Genet.* **14**, 618 (2013).
 [3] X. Han, R. Wang, Y. Zhou, L. Fei, H. Sun, S. Lai, A. Saadatpour, Z. Zhou, H. Chen, F. Ye *et al.*, *Cell* **172**, 1091 (2018).
 [4] Tabula Muris Consortium *et al.*, *Nature (London)* **562**, 367 (2018).
 [5] X. Han, Z. Zhou, L. Fei, H. Sun, R. Wang, Y. Chen, H. Chen, J. Wang, H. Tang, W. Ge, Y. Zhou, F. Ye, M. Jiang, J. Wu, Y.

- Xiao, X. Jia, T. Zhang, X. Ma, q. Zhang, and G. Guo, *Nature (London)* **581**, 303 (2020).
 [6] A. Mazzolini, M. Gherardi, M. Caselle, M. Cosentino Lagomarsino, and M. Osella, *Phys. Rev. X* **8**, 021023 (2018).
 [7] E. van Nimwegen, in *Power Laws, Scale-Free Networks and Genome Biology* (Springer, New York, 2006), p. 236.
 [8] T. Y. Pang and S. Maslov, *Proc. Natl. Acad. Sci. USA* **110**, 6235 (2013).
 [9] E. G. Altmann and M. Gerlach, in *Creativity and Universality in Language*, edited by M. Degli Esposti, E. G. Altmann, and F. Pachet (Springer, Cham, 2016), pp. 7–26.

- [10] M. Cosentino Lagomarsino, A. L. Sellerio, P. D. Heijning, and B. Bassetti, *Genome Biol.* **10**, R12 (2009).
- [11] J. Grilli, *Nat. Commun.* **11**, 4743 (2020).
- [12] D. Hebenstreit, M. Fang, M. Gu, V. Charoensawan, A. van Oudenaarden, and S. A. Teichmann, *Mol. Syst. Biol.* **7**, 497 (2011).
- [13] M. Borella, G. Martello, D. Risso, and C. Romualdi, *Bioinformatics* **38**, 164 (2021).
- [14] B. Wang, *PLoS One* **15**, e0230594 (2020).
- [15] O. Stegle, S. A. Teichmann, and J. C. Marioni, *Nat. Rev. Genet.* **16**, 133 (2015).
- [16] J. D. Silverman, K. Roche, S. Mukherjee, and L. A. David, *Comput. Struct. Biotechnol. J.* **18**, 2789 (2020).
- [17] V. Svensson, *Nat. Biotechnol.* **38**, 147 (2020).
- [18] S. Mangul, L. S. Martin, B. L. Hill, A. K.-M. Lam, M. G. Distler, A. Zelikovsky, E. Eskin, and J. Flint, *Nat. Commun.* **10**, 1393 (2019).
- [19] G. Guo, https://figshare.com/articles/dataset/HCL_DGE_Data/7235471 accessed January 2020.
- [20] S. Islam, A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. Lönnerberg, and S. Linnarsson, *Nat. Methods* **11**, 163 (2014).
- [21] The Tabula Muris Consortium, Single-cell RNA-seq data from Smart-seq2 sequencing of FACS sorted cells, https://figshare.com/projects/Tabula_Muris_Transcriptomic_characterization_of_20_organ_and_tissues_from_Mus_musculus_at_single_cell_resolution/27733 accessed January 2020 for FACS gene counts and metadata and September 2022 for microfluidic droplets (GSE109774).
- [22] S. Picelli, O. R. Faridani, Å. K. Björklund, G. Winberg, S. Sagasser, and R. Sandberg, *Nat. Protoc.* **9**, 171 (2014).
- [23] GTEx Consortium *et al.*, *Science* **348**, 648 (2015); <https://www.gtexportal.org/home/datasets> accessed January 2020.
- [24] M. Hagemann-Jensen, C. Ziegenhain, P. Chen, D. Ramsköld, G.-J. Hendriks, A. J. Larsson, O. R. Faridani, and R. Sandberg, *Nat. Biotechnol.* **38**, 708 (2020).
- [25] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevE.107.044403> for all the additional Figures and Tables mentioned in this paper.
- [26] A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner, *Cell* **161**, 1187 (2015).
- [27] J. Breda, M. Zavolan, and E. van Nimwegen, *Nat. Biotechnol.* **39**, 1008 (2021).
- [28] A. Mazzolini, J. Grilli, E. De Lazzari, M. Osella, M. C. Lagomarsino, and M. Gherardi, *Phys. Rev. E* **98**, 012315 (2018).
- [29] M. E. Newman, *Contemp. Phys.* **46**, 323 (2005).
- [30] G. K. Zipf, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology* (Addison-Wesley, Cambridge, MA, 1949).
- [31] C. Furusawa and K. Kaneko, *Phys. Rev. Lett.* **90**, 088102 (2003).
- [32] H. R. Ueda, S. Hayashi, S. Matsuyama, T. Yomo, S. Hashimoto, S. A. Kay, J. B. Hogenesch, and M. Iino, *Proc. Natl. Acad. Sci. USA* **101**, 3765 (2004).
- [33] M. Gerlach and E. G. Altmann, *Phys. Rev. X* **3**, 021006 (2013).
- [34] S. Marguerat, A. Schmidt, S. Codlin, W. Chen, R. Aebersold, and J. Bähler, *Cell* **151**, 671 (2012).
- [35] B. T. Sherman, R. A. Lempicki *et al.*, *Nat. Protoc.* **4**, 44 (2009).
- [36] Y. Zhou, B. Zhou, L. Pache, M. Chang, A. H. Khodabakhshi, O. Tanaseichuk, C. Benner, and S. K. Chanda, *Nat. Commun.* **10**, 1523 (2019).
- [37] H. S. Heaps, *Information Retrieval, Computational and Theoretical Aspects* (Academic, New York, 1978).
- [38] A. Mazzolini, A. Colliva, M. Caselle, and M. Osella, *Phys. Rev. E* **98**, 052139 (2018).
- [39] Z. Eisler, I. Bartos, and J. Kertész, *Adv. Phys.* **57**, 89 (2008).
- [40] M. Gerlach and E. G. Altmann, *New J. Phys.* **16**, 113010 (2014).
- [41] P. Brennecke, S. Anders, J. K. Kim, A. A. Kołodziejczyk, X. Zhang, V. Proserpio, B. Baying, V. Benes, S. A. Teichmann, J. C. Marioni *et al.*, *Nat. Methods* **10**, 1093 (2013).
- [42] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck III, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija, *Cell* **177**, 1888 (2019).
- [43] A. Raj and A. van Oudenaarden, *Cell* **135**, 216 (2008).
- [44] A. D. Co, M. C. Lagomarsino, M. Caselle, and M. Osella, *Nucleic Acids Res.* **45**, 1069 (2017).
- [45] D. M. Suter, N. Molina, D. Gatfield, K. Schneider, U. Schibler, and F. Naef, *Science* **332**, 472 (2011).
- [46] N. Friedman, L. Cai, and X. S. Xie, *Phys. Rev. Lett.* **97**, 168302 (2006).
- [47] V. Shahrezaei and P. S. Swain, *Proc. Natl. Acad. Sci. USA* **105**, 17256 (2008).
- [48] P. S. Swain, M. B. Elowitz, and E. D. Siggia, *Proc. Natl. Acad. Sci. USA* **99**, 12795 (2002).
- [49] Y. Taniguchi, P. J. Choi, G.-W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. S. Xie, *Science* **329**, 533 (2010).
- [50] T. H. Kim, X. Zhou, and M. Chen, *Genome Biol.* **21**, 196 (2020).
- [51] P. V. Kharchenko, L. Silberstein, and D. T. Scadden, *Nat. Methods* **11**, 740 (2014).
- [52] A. K. Sarkar and M. Stephens, *Nat. Genet.* **53**, 770 (2021).
- [53] T. S. Andrews and M. Hemberg, *Bioinformatics* **35**, 2865 (2019).
- [54] J. R. Newman, S. Ghaemmaghami, J. Ihmels, D. K. Breslow, M. Noble, J. L. DeRisi, and J. S. Weissman, *Nature (London)* **441**, 840 (2006).
- [55] M. Gerlach and E. G. Altmann, *Phys. Rev. Lett.* **122**, 168301 (2019).
- [56] M. Scott, C. W. Gunderson, E. M. Mateescu, Z. Zhang, and T. Hwa, *Science* **330**, 1099 (2010).
- [57] E. V. Koonin, *PLoS Comput. Biol.* **7**, e1002173 (2011).
- [58] S. K. Baek, S. Bernhardsson, and P. Minnhagen, *New J. Phys.* **13**, 043004 (2011).
- [59] M. Gerlach, H. Shi, and L. A. N. Amaral, *Nat. Mach. Intell.* **1**, 606 (2019).
- [60] G. Cristadoro, M. Degli Esposti, and E. G. Altmann, *Sci. Rep.* **8**, 15817 (2018).
- [61] R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C. K. Peng, M. Simons, and H. E. Stanley, *Phys. Rev. Lett.* **73**, 3169 (1994).
- [62] M. Lynch and J. S. Conery, *Science* **302**, 1401 (2003).
- [63] F. Tria, V. Loreto, V. D. P. Servedio, and S. H. Strogatz, *Sci. Rep.* **4**, 5890 (2014).
- [64] B. Corominas-Murtra, R. Hanel, and S. Thurner, *Proc. Natl. Acad. Sci. USA* **112**, 5348 (2015).

- [65] I. Iacopini, S. c. v. Milojević, and V. Latora, *Phys. Rev. Lett.* **120**, 048301 (2018).
- [66] D. J. Schwab, I. Nemenman, and P. Mehta, *Phys. Rev. Lett.* **113**, 068102 (2014).
- [67] L. Aitchison, C. N., and P. Latham, *PLoS Comput. Biol.* **12**, e1005110 (2016).
- [68] S. Klumpp, Z. Zhang, and T. Hwa, *Cell* **139**, 1366 (2009).
- [69] M. Osella and M. C. Lagomarsino, *Phys. Rev. E* **87**, 012726 (2013).
- [70] S. Lazzardi and F. Valle, Emergent Statistical Laws in Single-Cell Transcriptomic Data (2022), <https://zenodo.org/record/6302674>.