EMERGENT STATISTICAL LAWS IN SINGLE-CELL TRANSCRIPTOMIC DATA

A PREPRINT

Silvia Lazzardi[1,†] Filippo Valle[1,†], Andrea Mazzolini[2], Antonio Scialdone[3,4,5], Michele Caselle[1] and Matteo Osella[1,*]
† **Equal contribution**

1 Physics Department, University of Turin and INFN, via P. Giuria 1, 10125 Turin, Italy
2 Laboratoire de physique de l'École normale supérieure (PSL University), CNRS, Sorbonne Université, and Université de Paris, 75005 Paris, France
3 Institute of Epigenetics and Stem Cells, Helmholtz Zentrum München; Feodor-Lynen-Strasse 21, 81377 München, Germany
4 Institute of Functional Epigenetics, Helmholtz Zentrum München; Ingolstädter Landstraße 1, 85764 Neuherberg, Germany
5 Institute of Computational Biology, Helmholtz Zentrum München; Ingolstädter Landstraße 1, 85764 Neuherberg, Germany
*mosella@to.infn.it

Abstract

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.

1 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, 4, 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 by words, genomes of different species are collections of genes of different evolutionary families or associated to different biological functions, ecological niches are compositions of species with different abundances. Analogously, the transcriptional profiles of single cells are the sum 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 modelling 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, 8, 9, 10, 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.

Identifying empirical laws and understanding their origin is a potentially useful step in developing statistical models that can disentangle genuine biological variability from technical or statistical effects. 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]. 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. However, some data features are not fully captured by this description, suggesting where to focus in order to extrapolate actual properties of biological variability.

Building step-by-step simple models of an increasing level of realism can be a useful approach in dissecting the motivations behind empirical patterns. For example, we will show how this model-driven approach can contribute to the understanding of data sparsity in scRNAseq datasets, which is a debated topic in the field [12, 13]. On the other hand, 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 trancriptomic data, can be useful for model testing and selection.

2 Materials and Methods

2.1 Data sources

The Mouse Cell Atlas (MCA) was selected as the main illustrative dasteset. In the MCA more than $\sim 4 * 10^5$ single cells where profiled using scRNAseq from all major organs [3]. An advantage of this dataset is the use of Unique Molecular Identifiers (UMI) [14]. This technique provides a number of unique mRNA molecules captured for each mRNA, without the complication of amplification noise, and thus a reliable estimate of the transcriptome size sampled.

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] that were processed with the Smart-seq2 protocol [15], which produces a full-length transcriptome profiling but does not use UMI.

A dataset of bulk RNA-sequencing of healthy human tissues from the Genotype-Tissue Expression (GTEx) Project [16] was used to test the results on population-averaged transcriptomic data.

Finally, we analyzed two additional single-cell datasets, relative to a HEK cell line and to mouse fibroblasts, profiled with the recently introduced Smart-seq3 protocol [17].

2.2 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) $i \in \{1, \ldots, N\}$ in cell (i.e., the realization) $c \in \{1, \ldots, R\}$. 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 $\{n_\cdot^c\} = \{n_1^c, \ldots, n_N^c\}$ 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 number 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 number 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 UMI per cell is $\sim 1\,200$ and the distribution is reported in Fig. S1.

2.3 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 has been previously introduced for scRNAseq data [18] and is at the basis of a recent Bayesian procedure for data normalization [19].

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, 19]. Therefore, the probability of observing a specific expression profile $\{n_\cdot^c\}$ in a cell c, from which M^c transcripts have been sequenced, is given by

$$P(\{n_\cdot^c\}|M^c) = \frac{M^c!}{\prod_{i=1}^N n_i^c!} \prod_{i=1}^N (f_i^c)^{n_i^c}, \tag{1}$$

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 $\rho(\{f_\cdot\})$ setting the mRNA frequencies in the different cells. Thus, the probability of an expression profile in a cell with M observed transcripts becomes

$$P(\{n_\cdot^c\}|M) = \int_0^\infty [df]\rho(\{f_\cdot\}) \frac{M!}{\prod_i n_i^c!} \prod_i (f_i)^{n_i^c} \tag{2}$$

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

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

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. 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]:

$$f_i \simeq \frac{1}{R} \sum_{c=1}^R \frac{n_i^c}{M^c} \tag{4}$$

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 $\{n_\cdot\}$ is given by the expression

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

This model was previously analyzed to understand the origin of statistical regularities in different component systems [20].

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.

3 Results

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

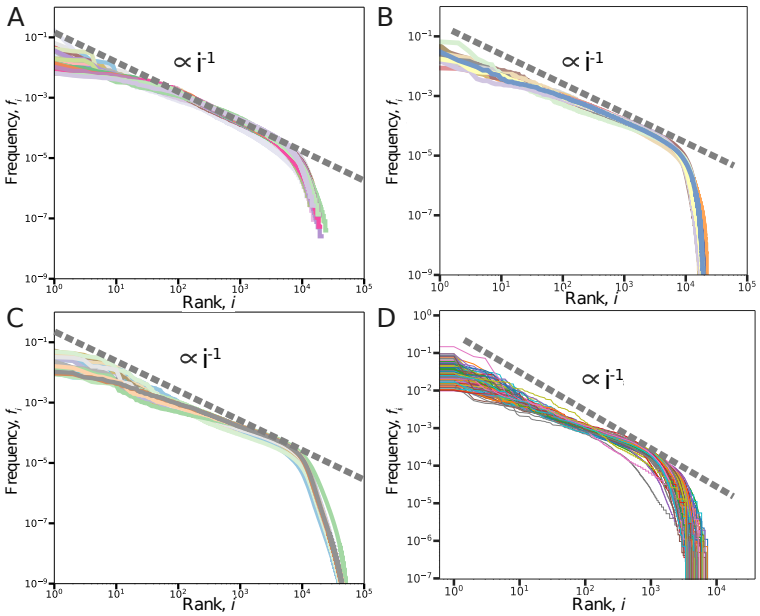

Figure 1: **A robust Zipf-like law for gene expression levels**. The average relative expression levels f_i are estimated as described by Equation 4 and reported as a function of their rank. The distributions reported correspond to averages over single cells belonging to different mouse organs from the Mouse Cell Atlas **(A)**, from 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. **(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). The dashed lines are just a reference power-law scaling with exponent -1.

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 [21]. For component systems in particular, the frequency of components is often well described by a power law known as Zipf's law [22, 21, 9, 6]. 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. Fig 1A 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 1B) is obtained by looking at the same organs in an alternative mouse expression atlas, i.e., Tabula Muris [4], in which different sequencing methods were adopted. Even if the relative gene expression levels measured in the two atlases are correlated, the variability is substantial (Fig. S4). 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 non-coding genes in the MCA, we still find the same distribution (Fig. S5).
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 [16] (Fig 1C). 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 [23, 24].

A natural question that can now be asked thanks to single-cell transcriptomics is if this emerging behaviour is a consequence of the averaging process or a property of the gene expression program of single cells. In other words, if gene expression in single cells displays actually a variety of cell-specific distributions that sum up in a Zipf-like law when averaged over a large cell population. Fig 1D 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 behaviour 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 empirical law essentially sets the only free parameters f_i of our null model described by Equation 5. Since this model only describes a sampling process given the empirical average frequencies of components, we can test what properties of the system can be explained merely by sampling effects and what features are instead potentially due to biological variability.

3.2 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 behaviour can be observed by looking at protein domain frequencies in genomes of different species [6]. A double scaling was also observed in natural language [25], where it was tentatively explained by a model with two different classes of words: common words (high rank) composing a core vocabulary and the rest of more specific words in a vast vocabulary. Analogously, two different groups of genes can be distinguished in bulk transcriptomic data: a core of highly expressed genes with active promoters and a second group of lowly expressed and putatively non-functional transcripts [26]. 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), 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). Fig 2 shows that this exponential tail is also enriched in non-coding 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 behaviour with exponent close to -1 (Fig 2). In fact, analysing the fitting error from a single power law (details of the procedure are reported in Fig. S7), three different regimes can be identified . They are approximately captured by two power laws with different exponents before the exponential tail. 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. A similar law with three regimes was observed in a quantitative transcriptomic study of fission yeast [27]. The gene expression levels in single cells (Fig 1D) also seem to generally display three classes, but the higher level of fluctuations does not allow a refined analysis.

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 to the basic protein translation process (e.g, "structural constituent of ribosome" or "translation") with Benjamini-corrected P-values lower than 10^{-20} (GO enrichement analyses were performed using DAVID [28]). Moreover, 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.

Therefore, the first regime appears to be composed by highly expressed genes related to basic functions that are common across different organs. This first core is followed by actively expressed genes that are more tissue specific and whose expression follows the classic Zipf's law with exponent -1.

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

A complex biological system such as an organ is composed by 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

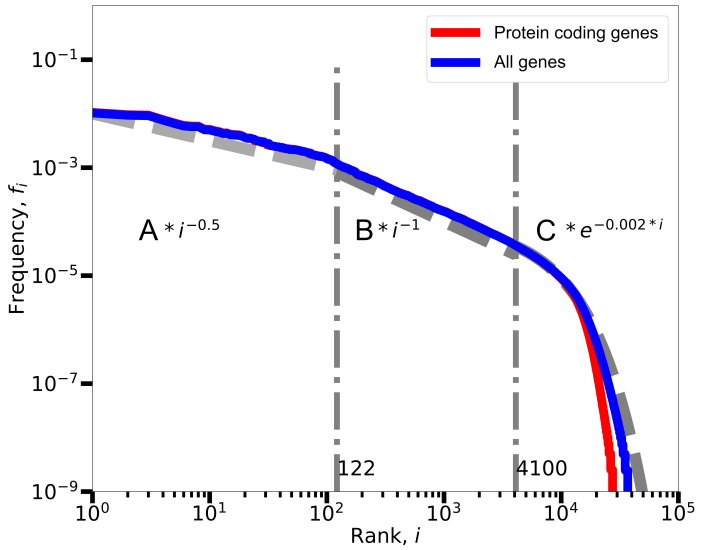

Figure 2: **Multiple regimes in the rank-plot of the average expression levels.** By considering the fitting error using a power law function in a window with variable width (Fig. S7), we were able to identify the part of the distribution well explained by a single power law, and consequently the two other regimes for highly expressed and lowly expressed genes. Excluding or including the non-coding genes from the analysis only influences the exponential tail of the distribution.

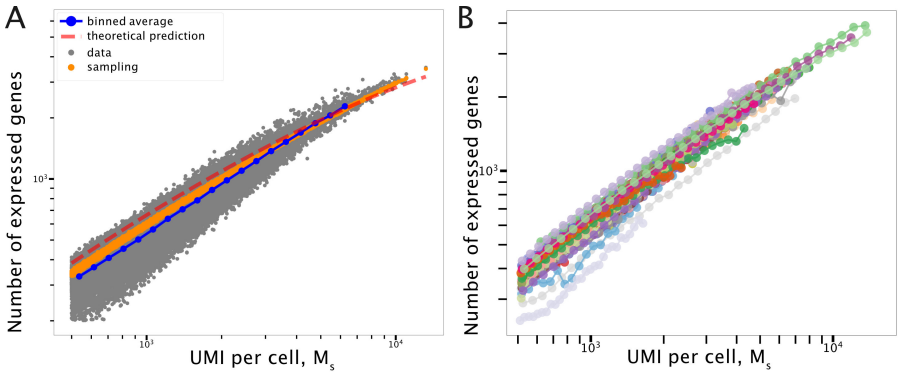

Figure 3: **The number of detected different transcripts follows Heaps' law (A)** The number of mRNAs with at least one detected transcript $h(M)$ is reported as a function of the transcriptome size as measured by the number of UMI, i.e., $M^c = \sum_i n_i^c$. Each point in the scatter plot thus corresponds 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 (red dashed line). **(B)** The same sublinear average scaling is approximately conserved in all organs reported in the MCA.

basic 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' law [29, 9, 30]. 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. Fig 3A shows the number of different mRNAs as a function of the total number of UMI as an estimate of the total number of detected mRNAs. This analysis cannot be naturally applied in absence of UMI, 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 the model presented in Equation 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 [30] 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. 3A) 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 [30] that the expression for $h(M)$ simplifies to

$$\langle h(M) \rangle = N - \sum_{i=1}^{i^*} \left(1 - A i^{-\gamma_1}\right)^M - \sum_{i=i^*+1}^{i^{**}} \left(1 - B i^{-\gamma_2}\right)^M - \sum_{i=i^{**}+1}^{N} \left(1 - C e^{-k*i}\right)^M. \tag{8}$$

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

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

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 3A.
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 3B.

3.4 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 UMI 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 $\sigma(h)^2$ as a function of the average value $\langle h \rangle$. Fig 4A 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

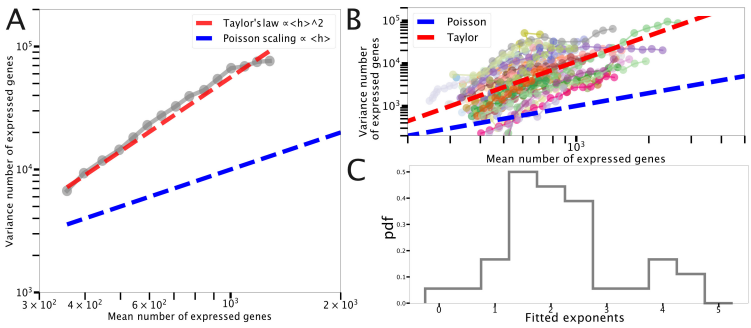

Figure 4: **Fluctuation scaling in the number of detected transcripts follows Taylor's law. (A)** The variance in the number of measured expressed genes is reported as a function of its average value for all cells in the MCA. Data are compared to a quadratic scaling (red dashed line) and to the Poisson 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 $C\langle h\rangle^k$.

empirical variance displays a quadratic scaling with the average vocabulary size, i.e., $\sigma^2(h) \sim \langle h\rangle^2$ that is clearly distinguishable from the sampling prediction. Focusing on single cells belonging to the same organ the phenomenology is quite diversified, also due to the reduced cell numbers (Fig 4B). However, fitting these organ-specific fluctuation curves with the function $C\langle h\rangle^k$ still leads to a distribution of exponents peaked on 2 (Fig 4C). This suggests that the 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, 31, 32]. Therefore, a general explanation, which goes beyond the specific properties of expression profiles, could be at the origin of this scaling.

3.5 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^2 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 criterion is often used to reduce the number of features (i.e., genes) to consider in further analysis [33, 34]. 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.

Fig 5A shows the CV^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.

Note that this 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 [35, 36]. This description of gene expression 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 [37], and there could be an additional basal technical noise besides sampling fluctuations [18]. All these factors can introduce non-Poisson fluctuations in the data that are not due to functional biological 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. Consider a random variable x with average $\langle x\rangle = \mu$ and variance $\sigma_x^2 = \mu$. We can now define a new variable $y = kx$, where k is a constant describing a supposedly constant amplification factor. The mean

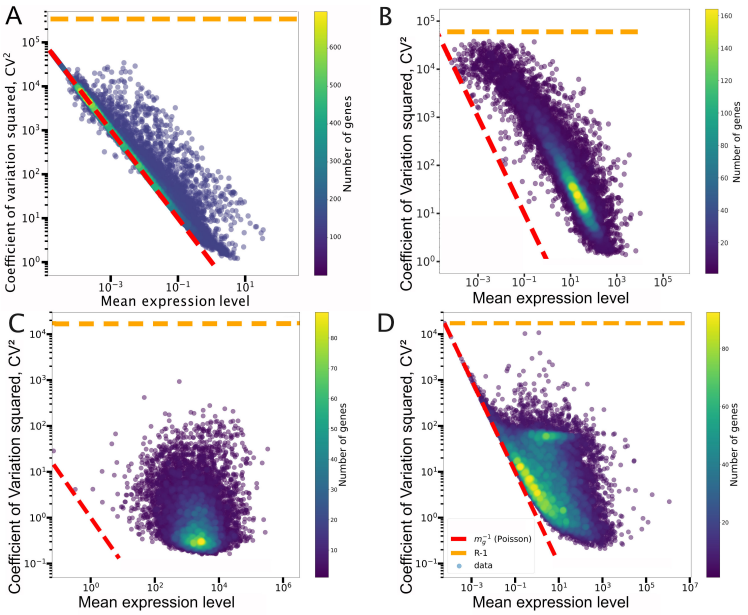

Figure 5: **The Coefficient of Variation CV^2 as a function of the average expression.** The red dashed lines report the Poisson scaling,i.e., $\sigma_n^2 = \langle n \rangle$. The horizontal orange dashed lines correspond to the maximum possible value of $CV^2 = R - 1$, which is achieved if a gene is 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 non-coding 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.

and variance of y are simply given by $\langle y \rangle = k\mu$ and $\sigma_y^2 = k^2\mu$. Therefore, the CV^2 of the scaled variable y is still $CV_y^2 = \frac{1}{\mu}$ but can be written as a function of $\langle y \rangle$ as $\log(CV_y^2) = -\log(\mu) = -\log\left(\frac{\langle y \rangle}{k}\right) = -\log\left(\langle y \rangle\right) + \log(k)$.

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 5C). The empirical variability is far from the sampling limit and the CV seems typically weakly dependent on the average expression level. However, focusing on non-coding genes (Fig 5D), 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 RNA-seq data.

This analysis shows that the sampling process sets a lower bound to 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. This makes the observation that several statistical empirical trends can be captured just from Zipf's law and a sampling process even more surprising.

3.6 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 non-zero count). Fig 6A reports the occurrence

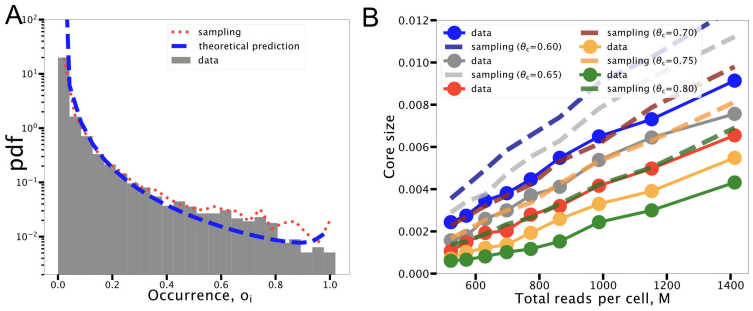

Figure 6: **The occurrence distribution of gene expression. (A)** The probability of observing a mRNA in a fraction o_i of cells is reported for the illustrative example of cells in the bone marrow as profiled in the MCA. **(B)** The empirical fraction of genes expressed in at least θ cells (different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

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

$$o_i = 1 - \frac{1}{R}\sum_{c=1}^{R} P_i(0|M^c) \simeq 1 - \frac{1}{R}\sum_{c=1}^{R} e^{-f_i M^c}. \tag{10}$$

From this expression, the probability density of the occurrences can be extracted. It is reported as a dotted red line in Fig 6A 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 UMI ($M \simeq 1\,500$ transcripts in this case). The resulting expression is

$$p(o) = \frac{(1-o)^{\frac{1}{M-1}}}{\gamma M N \alpha^{1/\gamma}[1-(1-o)^{\frac{1}{M}}]^{1/\gamma+1}}. \tag{11}$$

Despite the crude approximations, this analytical prediction (blue dashed line in Fig 6A) 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 6A): ubiquitously expressed genes, or core genes, seems 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

$$c(M) = \frac{M^{\frac{1}{\gamma}}}{\alpha^{\frac{1}{\gamma}} N}(-log(1-\theta))^{-\frac{1}{\gamma}}, \tag{12}$$

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 6B). 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. 6B). The empirical scaling can be compared with direct simulations of the sampling process or with the equivalent numerical integration of Eq. 10 (dashed lines in Fig. 6B). While the linear trend described by Eq. 12 is observed in both data and simulations, the empirical curves 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.

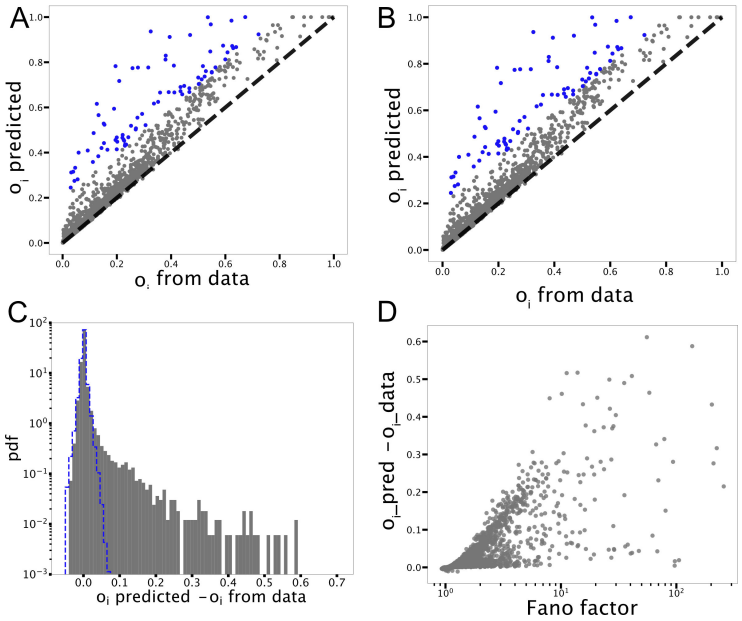

Figure 7: **Explaining the empirical occurrences from average expression and sampling**. The number of cells in which a transcript is present $o_{i\text{data}}$ is reported with respect to the occurrence predicted by sampling $o_{i\text{predicted}}$ for **(A)** Bladder and **(B)** Muscle. **(C)** The distribution of $o_{real} - o_{predicted}$ 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 $o_{real} - o_{predicted}$ in Muscle as an illustrative example.

3.7 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 [13, 12, 38]. 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. 10). This prediction can be directly compared with the empirical occurrence as in Fig 7A and B. The first observation is that the density of points is mostly located on the diagonal, where the number of zero 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 7C (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 [39, 40], are generally not needed. This observation is in line with recent analysis of scRNAseq data based on UMI [13, 19, 40].

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 7B 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 over-simplification 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 (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., $\sigma^2(n)/\langle n \rangle$) 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 7D). 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 [41]. This excess of zero values can derive from technical reasons (often called "drop-outs"), 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 $o_{predicted} - o_{data}$ (depicted in color in Fig 7A and B) in different organs and performed a GO enrichment analysis. Some categories are over-represented (reported in Table S3 in the Supporting Information). The presence of enriched GO categories already indicates that if a drop-out phenomenon is present, it is not random across genes. However, enriched categories such as "Ribosomal proteins" or "Blood microparticle" most likely indicate technical reasons for the presence of their zero values, for example linked to 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 the Zip's law (Table S2). Therefore, the few empirical zero counts cannot be explained from sampling given the high average expression levels. In this example, the sampling model is 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. 7C). For example, the small fraction of genes that are detected in more cells than expected given their average expression (negative values in Fig. 7C) can be explained by sampling fluctuations.

3.8 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 [17]. The Smart-seq3 protocol combines high sensitivity with the use of UMI, 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. Fig 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. S8. 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. 8A). The trend predicted by the sampling process for Heaps' law (Fig. 8B) 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 [42, 43].

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. 8D). Thanks to the protocol 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 (Fig. 8E and F), in perfect analogy with our results for the MCA. Typical occurrence fluctuations only due to sampling (blue line in Fig. 8D) 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).

In conclusion, the presented statistical laws seem 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.

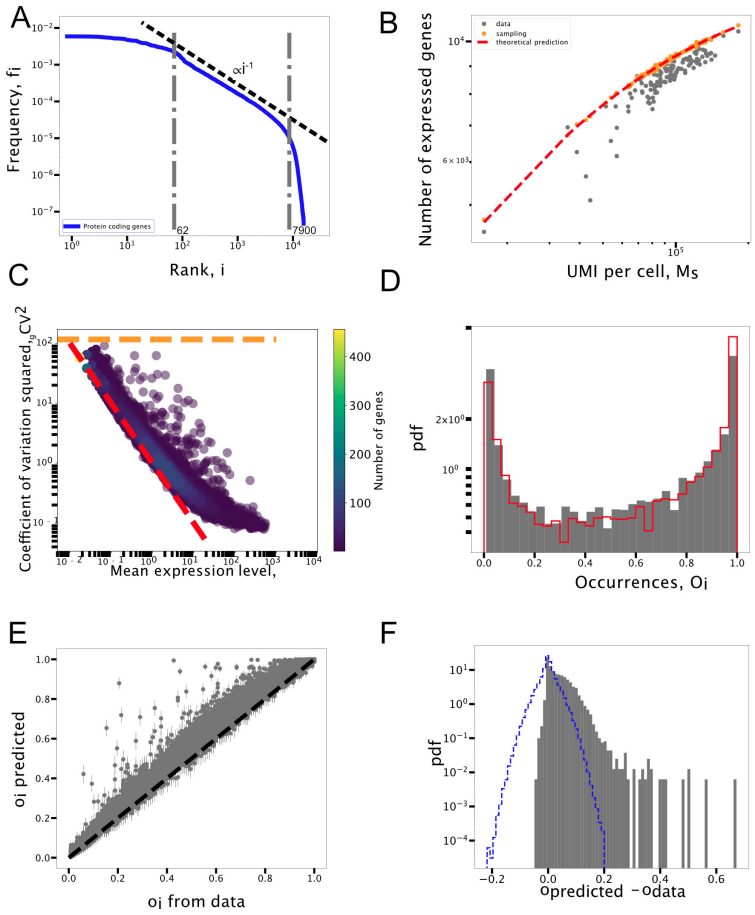

Figure 8: **Emergent statistical laws from a HEK cell line profiled with Smart-seq3**. (**A**) Zipf's law with the three scaling regimes. (**B**) Heaps' law: number of detected transcripts as a function of the total number of UMI per cell. (**C**) Expression variability CV^2 as a function of the average expression level for all detected transcripts. (**D**) The empirical transcript occurrence distribution compared with the model expectation (red continous line). (**E**) The relation between the empirical occurrences o_{data} and the expected values from the sampling model o_{pred}. Error bars represent the variability (one sigma) between different realizations of the sampling process. (**F**) Distribution of $o_{pred} - o_{data}$. As in Fig. 7 C, the blue dashed line identifies the differences compatible with sampling fluctuations.

4 Discussion

The identification of statistical laws is a key step in designing effective descriptions of complex systems [44]. Leveraging on 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 lead to simple models of cell physiology that can explain several large-scale gene expression patterns using just few key parameters such as the growth rate [45]. 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 in mouse. Strikingly, analogous laws are ubiquitously found in different complex component systems from linguistics to ecology [9, 11, 46, 6].

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 modelling 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 [21, 47]. It has been previously reported for gene expression at the population level [23], but we showed that it is an intrinsic property of single cells robustly emerging in different datasets.

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' law and the U-shaped statistics of shared components [6, 30]. 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 analysis 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' 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 [31, 48, 11], 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 [13, 12, 40, 38]. 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 [40]. Despite this general trend, the model is a tool to identify specific deviations. A fraction of genes are indeed expressed in less cells than expected (i.e., they have an excess of zero counts). While we discussed few examples of genes with these atypical zeros that seems related to technical errors, we do not exclude the possibility that gene regulation has a role in other cases. 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.

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, 49, 25]. 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' law. The recently proposed sample-space-reducing process can be also be ascribed to this class of stochastic growth models [50, 30]. 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-occurence in a realization [51, 8, 20]. 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' laws [20]. 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 governs the hidden structure behind the data [52, 53]. 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 behaviour of different genetic circuits [54, 55]. 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.

Conflict of Interest Statement

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.

Author Contributions

MO designed of the research. SL and FV performed the analyses and built the figures. MO, SL, FV wrote the article. MO, SL, FV, AM, AS and MC read and edited the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was partially supported by the "Departments of Excellence 2018 - 2022" Grant awarded by the Italian Ministry of Education, University and Research (MIUR) (L.232/2016).

Acknowledgments

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

Data Availability Statement

The notebooks and the code to reproduce this study can be found in the *The single cell transcriptome as a component system* GitHub's repository at `https://github.com/SilviaLazzardi/Emergent_Laws_in_scRNA-seq_Data`.

References

[1] Allon Wagner, Aviv Regev, and Nir Yosef. Revealing the vectors of cellular identity with single-cell genomics. *Nature Biotechnology*, 34(11):1145–1160, 2016.

[2] Ehud Shapiro, Tamir Biezuner, and Sten Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, 14(9):618–630, 2013.

[3] Xiaoping Han, Renying Wang, Yincong Zhou, Lijiang Fei, Huiyu Sun, Shujing Lai, Assieh Saadatpour, Ziming Zhou, Haide Chen, Fang Ye, et al. Mapping the mouse cell atlas by microwell-seq. *Cell*, 172(5):1091–1107, 2018.

[4] Tabula Muris Consortium and others. Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature*, 562(7727):367, 2018.

[5] Xiaoping Han, Ziming Zhou, Lijiang Fei, Huiyu Sun, Renying Wang, Yao Chen, Hyde Chen, Jingjing Wang, Huanna Tang, Wenhao Ge, Yincong Zhou, Fang Ye, Mengmeng Jiang, Junqing Wu, Yanyu Xiao, Xiaoning Jia, Tingyue Zhang, Xiaojie Ma, qi Zhang, and Guoji Guo. Construction of a human cell landscape at single-cell level. *Nature*, 581, 05 2020.

[6] Andrea Mazzolini, Marco Gherardi, Michele Caselle, Marco Cosentino Lagomarsino, and Matteo Osella. Statistics of Shared Components in Complex Component Systems. *Physical Review X*, 8:021023, Apr 2018.

[7] Erik van Nimwegen. Scaling laws in the functional content of genomes. *Power Laws, Scale-Free Networks and Genome Biology*, pages 236–253, 2006.

[8] Tin Yau Pang and Sergei Maslov. Universal distribution of component frequencies in biological and technological systems. *Proceedings of the National Academy of Sciences*, 110(15):6235–6239, 2013.

[9] E. G. Altmann and M. Gerlach. Statistical Laws in Linguistics. In Mirko Degli Esposti, Eduardo G. Altmann, and François Pachet, editors, *Creativity and Universality in Language*, pages 7–26. Springer International Publishing, Cham, 2016.

[10] M. Cosentino Lagomarsino, A. L. Sellerio, P. D. Heijning, and B. Bassetti. Universal features in the genome-level evolution of protein domains. *Genome Biology*, 10(1):R12, 2009.

[11] Jacopo Grilli. Macroecological laws describe variation and diversity in microbial communities. *Nature Communications*, 11:4743, 2020.

[12] Justin D. Silverman, Kimberly Roche, Sayan Mukherjee, and Lawrence A. David. Naught all zeros in sequence count data are the same. *Computational and Structural Biotechnology Journal*, 18:2789–2798, 2020.

[13] Valentine Svensson. Droplet scrna-seq is not zero-inflated. *Nature Biotechnology*, 38(2):147—150, February 2020.

[14] Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell rna-seq with unique molecular identifiers. *Nature Methods*, 11(2):163, 2014.

[15] Simone Picelli, Omid R Faridani, Åsa K Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. Full-length rna-seq from single cells using smart-seq2. *Nature Protocols*, 9(1):171–181, 2014.

[16] GTEx Consortium and others. The genotype-tissue expression (gtex) pilot analysis: multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[17] Michael Hagemann-Jensen, Christoph Ziegenhain, Ping Chen, Daniel Ramsköld, Gert-Jan Hendriks, Anton JM Larsson, Omid R Faridani, and Rickard Sandberg. Single-cell rna counting at allele and isoform resolution using smart-seq3. *Nature Biotechnology*, 38(6):708–714, 2020.

[18] Allon M Klein, Linas Mazutis, Ilke Akartuna, Naren Tallapragada, Adrian Veres, Victor Li, Leonid Peshkin, David A Weitz, and Marc W Kirschner. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161(5):1187–1201, 2015.

[19] Jérémie Breda, Mihaela Zavolan, and Erik van Nimwegen. Bayesian inference of gene expression states from single-cell rna-seq data. *Nature Biotechnology*, pages 1–9, 2021.

[20] Andrea Mazzolini, Jacopo Grilli, Eleonora De Lazzari, Matteo Osella, Marco Cosentino Lagomarsino, and Marco Gherardi. Zipf and Heaps laws from dependency structures in component systems. *Physical Review E*, 98:012315, Jul 2018.

[21] Mark EJ Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46(5):323–351, 2005.

[22] George Kingsley Zipf. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books, 2016.

[23] Chikara Furusawa and Kunihiko Kaneko. Zipf's law in gene expression. *Physical Review Letters*, 90(8):088102, 2003.

[24] Hiroki R. Ueda, Satoko Hayashi, Shinichi Matsuyama, Tetsuya Yomo, Seiichi Hashimoto, Steve A. Kay, John B. Hogenesch, and Masamitsu Iino. Universality and flexibility in gene expression from bacteria to human. *Proceedings of the National Academy of Sciences*, 101(11):3765–3769, 2004.

[25] M. Gerlach and E. G. Altmann. Stochastic model for the vocabulary growth in natural languages. *Physical Review X*, 3(2):021006, 2013.

[26] Daniel Hebenstreit, Miaoqing Fang, Muxin Gu, Varodom Charoensawan, Alexander van Oudenaarden, and Sarah A Teichmann. Rna sequencing reveals two major classes of gene expression levels in metazoan cells. *Molecular Systems Biology*, 7(1):497, 2011.

[27] Samuel Marguerat, Alexander Schmidt, Sandra Codlin, Wei Chen, Ruedi Aebersold, and Jürg Bähler. Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. *Cell*, 151(3):671–683, 2012.

[28] Brad T Sherman, Richard A Lempicki, et al. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature Protocols*, 4(1):44, 2009.

[29] Harold Stanley Heaps. *Information retrieval, computational and theoretical aspects*. Academic Press, 1978.

[30] Andrea Mazzolini, Alberto Colliva, Michele Caselle, and Matteo Osella. Heaps' law, statistics of shared components, and temporal patterns from a sample-space-reducing process. *Physical Review E*, 98(5), nov 2018.

[31] Zoltán Eisler, Imre Bartos, and János Kertész. Fluctuation scaling in complex systems: Taylor's law and beyond. *Advances in Physics*, 57(1):89–142, 2008.

[32] M. Gerlach and E. G. Altmann. Scaling laws and fluctuations in the statistics of word frequencies. *New Journal of Physics*, 16(11):113010, 2014.

[33] Philip Brennecke, Simon Anders, Jong Kyoung Kim, Aleksandra A Kołodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, Vladimir Benes, Sarah A Teichmann, John C Marioni, et al. Accounting for technical noise in single-cell rna-seq experiments. *Nature Methods*, 10(11):1093, 2013.

[34] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888 – 1902.e21, 2019.

[35] A. Raj and A. van Oudenaarden. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2):216–226, 2008.

[36] Alma Dal Co, Marco Cosentino Lagomarsino, Michele Caselle, and Matteo Osella. Stochastic timing in gene expression for simple regulatory strategies. *Nucleic Acids Research*, 45(3):1069–1078, 12 2016.

[37] David M Suter, Nacho Molina, David Gatfield, Kim Schneider, Ueli Schibler, and Felix Naef. Mammalian genes are transcribed with widely different bursting kinetics. *Science*, 332(6028):472–474, 2011.

[38] Tae Hyun Kim, Xiang Zhou, and Mengjie Chen. Demystifying "drop-outs" in single-cell umi data. *Genome Biology*, 21(1):1–19, 2020.

[39] Peter V Kharchenko, Lev Silberstein, and David T Scadden. Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11(7):740–742, 2014.

[40] Abhishek K Sarkar and Matthew Stephens. Separating measurement and expression models clarifies confusion in single cell rna-seq analysis. *Nature Genetics*, pages 1–8, 2021.

[41] Tallulah S Andrews and Martin Hemberg. M3drop: dropout-based feature selection for scrnaseq. *Bioinformatics*, 35(16):2865–2867, 2019.

[42] Yuichi Taniguchi, Paul J Choi, Gene-Wei Li, Huiyi Chen, Mohan Babu, Jeremy Hearn, Andrew Emili, and X Sunney Xie. Quantifying e. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991):533–538, 2010.

[43] John RS Newman, Sina Ghaemmaghami, Jan Ihmels, David K Breslow, Matthew Noble, Joseph L DeRisi, and Jonathan S Weissman. Single-cell proteomic analysis of s. cerevisiae reveals the architecture of biological noise. *Nature*, 441(7095):840–846, 2006.

[44] Martin Gerlach and Eduardo G Altmann. Testing statistical laws in complex systems. *Physical Review Letters*, 122(16):168301, 2019.

[45] Matthew Scott, Carl W Gunderson, Eduard M Mateescu, Zhongge Zhang, and Terence Hwa. Interdependence of cell growth and gene expression: origins and consequences. *Science*, 330(6007):1099–1102, 2010.

[46] Eugene V Koonin. Are there laws of genome evolution? *PLoS Computational Biology*, 7(8):e1002173, 2011.

[47] S. K. Baek, S. Bernhardsson, and P. Minnhagen. Zipf's law unzipped. *New Journal of Physics*, 13(4):043004, 2011.

[48] Martin Gerlach and Eduardo G Altmann. Scaling laws and fluctuations in the statistics of word frequencies. *New Journal of Physics*, 16(11):113010, 2014.

[49] Francesca Tria, Vittorio Loreto, Vito Domenico Pietro Servedio, and Steven H Strogatz. The dynamics of correlated novelties. *Scientific Reports*, 4(1):1–8, 2014.

[50] B Corominas-Murtra, R Hanel, and S Thurner. Understanding scaling through history-dependent processes with collapsing sample space. *Proceedings of the National Academy of Sciences USA*, 112(17):5348–5353, 2015.

[51] Iacopo Iacopini, Sta ša Milojević, and Vito Latora. Network dynamics of innovation processes. *Physical Review Letters*, 120:048301, Jan 2018.

[52] David J. Schwab, Ilya Nemenman, and Pankaj Mehta. Zipf's law and criticality in multivariate data without fine-tuning. *Physical Review Letters*, 113:068102, Aug 2014.

[53] Laurence Aitchison, Corradi N, and Peter E. Latham. Zipf's law arises naturally when there are underlying, unobserved variables. *PLoS Computational Biology*, page e1005110, 2016.

[54] Stefan Klumpp, Zhongge Zhang, and Terence Hwa. Growth rate-dependent global effects on gene expression in bacteria. *Cell*, 139:1366–1375, 2009.

[55] Matteo Osella and Marco Cosentino Lagomarsino. Growth-rate-dependent dynamics of a bacterial genetic oscillator. *Physical Review E*, 87:012726, Jan 2013.

Supporting information

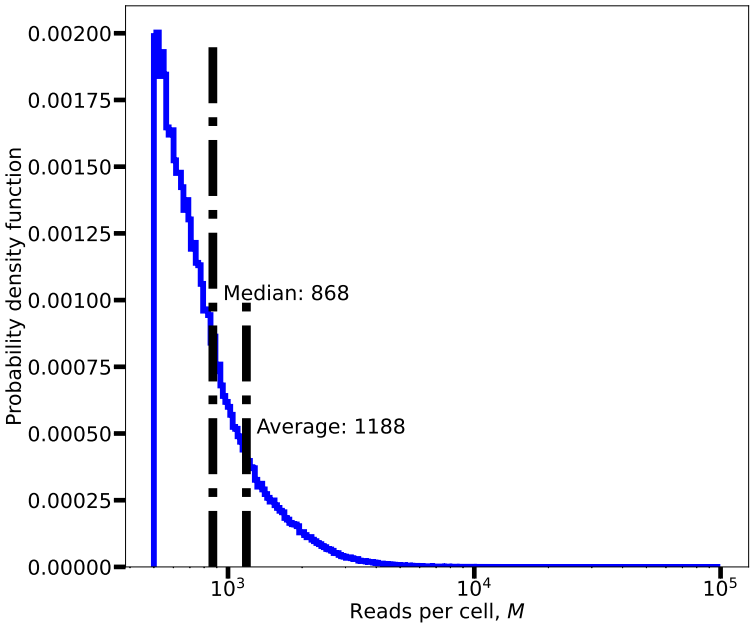

Figure S1: **Histogram of the trascriptome sizes in the Mouse Cell Atlas.** The trascriptome size M_j represent the total number of UMI detected from cell j. Mean and median values are reported for reference.

Figure S2: **Color legend for the different organs and tissues.** All the plots with multiple organs or tissues in the main text refers to this legend.

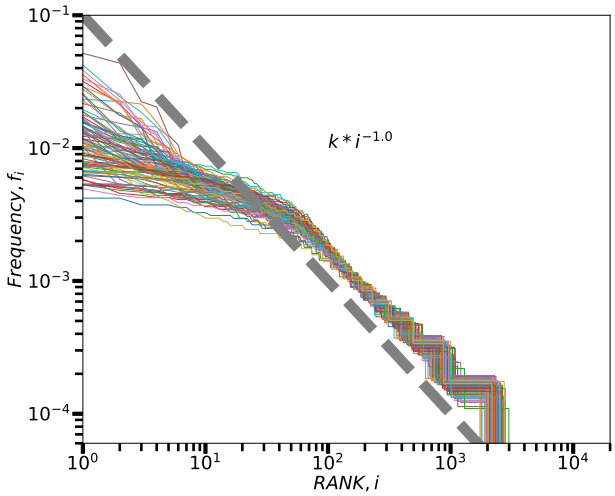

Figure S3: **Single-cell Zifp's law from the Mouse Cell Atlas.** Rank plot of the expression levels of the 100 most sequenced cells from the bone marrow in the Mouse Cell Atlas.

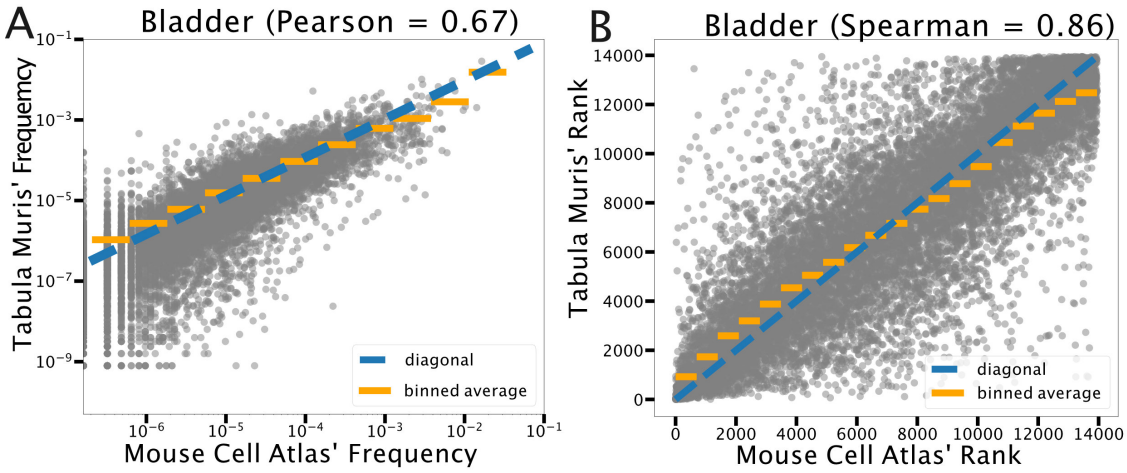

Figure S4: **Conservation of expression levels and corresponding ranking across the two different cell atlases. (A)** Scatter plot of the relative abundances of gene transcripts as reported in the Mouse Cell Atlas with respect to the Tabula Muris compendium for one illustrative organ (equivalent results are obtained looking at the other organs). **(B)** The correlation between the ranking of the relative abundances in the two atlases. Both measures indicate a clear correlation but the variability is substantial.

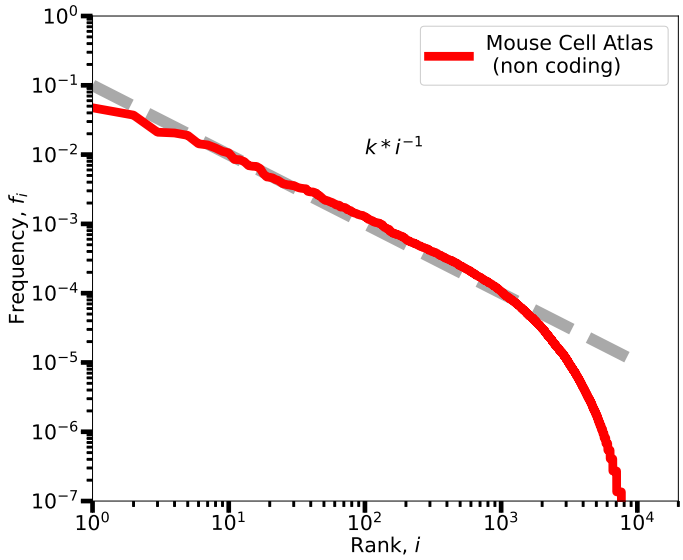

Figure S5: **Zipf's law for non-coding genes.** Expression frequencies f_i of non-coding genes from the Mouse Cell Atlas are reported in a rank plot. The power-law scaling followed by an exponential tail is compatible with the Zipf's law identified for coding genes.

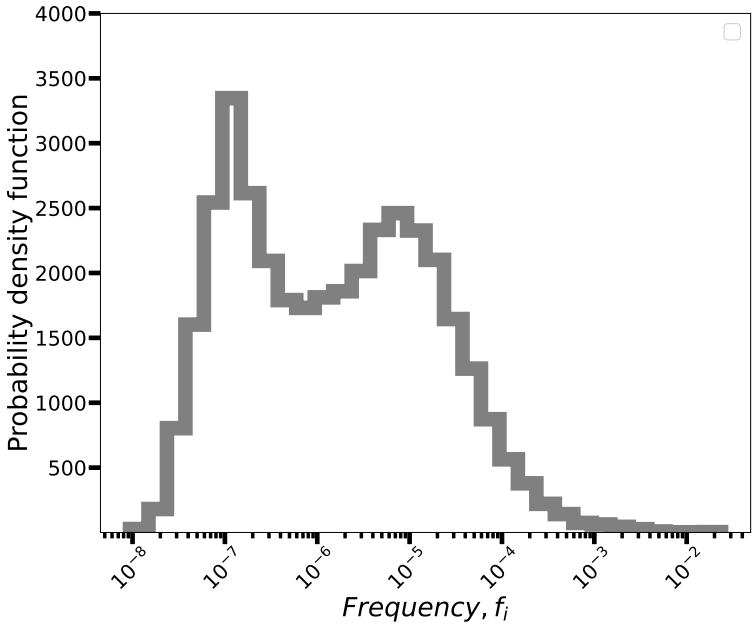

Figure S6: **Bimodality in the probability density function of expression levels**. We report the histogram of the estimated average transcript frequencies as $f_i = \frac{\sum_j \frac{n_{ij}}{M_j}}{R}$. The distribution has a bimodal shape with two maximum values at around 10^{-7} and 10^{-5}. These values correspond to the second and the third regimes of the Zipf's law we discussed.

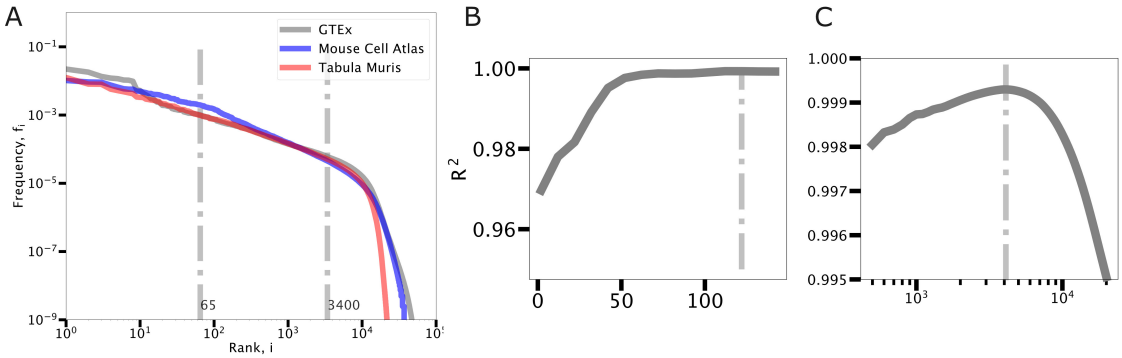

Figure S7: **Procedure for the identification of the three regimes in Zipf's law.** **(A)** The average rank plot for each dataset and the average bounds. Given window of frequencies, we performed a fit evaluation for each position of the window boundaries. The left bound was changed from 2 to 122. Given a left bound position, the right bound moved between 500 and 20000. For each boundary positions we estimated the coefficient of determination $R^2 = 1 - \frac{\sum_i (f(x_i)-y_i)^2}{\sum_i (\bar{y_i}-y_i)^2}$, as reported in panels **(B)** and **(C)** for the MCA example. f is the power-law function, x_i the rank, y_i the frequency and $\bar{y_i}$ the average of the y_i values. The boundaries providing the larger R^2 were selected to identify the regimes. With fixed boundaries, we finally fitted with a power law the first and second regimes (high frequencies) and with an exponential function $f(i) = a * e^{-b*i}$ the low frequency tail.

Dataset	Left Bound	Right bound
Mouse Cell Atlas	122	4100
Tabula Muris	12	1000
GTEx	62	5100
Average	65	3400
SmartSeq3 HEK	62	7900
SmartSeq3 Fibroblast	82	1100

Table S1: Boundaries separating the three regimes of the Zipf's law (by optimizing R^2) in each dataset analysed.

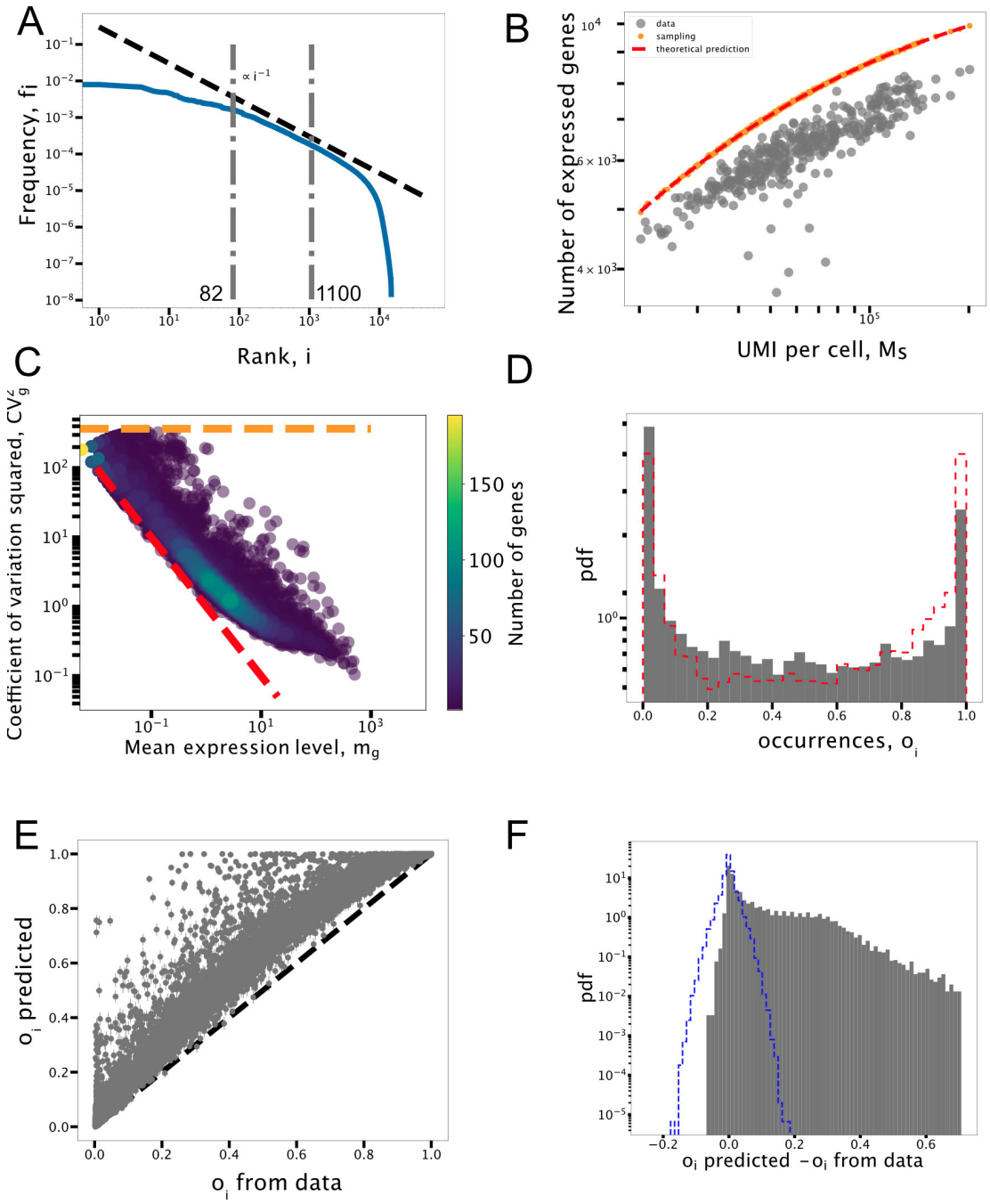

Figure S8: **Statistical laws from fibroblast cells profiled with Smart-seq3 protocl**. The Figure is the equivalent of Figure 8 in the main text. (**A**) Zipf's law. (**B**) Heaps' law. (**C**) CV^2 scaling with the average expression level. (**D**) Transcript occurrence distribution compared with the model expectatio. (**E**)The comparison of predicted occurrences o_{pred} versus empirical ones o_{data}. (**F**) Distribution of $o_{pred} - o_{data}$ and the corresponding distribution from sampling variability.

GO enrichment analysis

We reported in this section all the terms and the Benjamini corrected P-values of Gene Ontology (GO) enrichment analyses discussed in the main text. They were performed using DAVID.

Organ	Related GO	Benjamini
Bladder	Ribosomal protein	$1 * 10^{-100}$
	ribosome	$4.3 * 10^{-92}$
	Ribosome	$7.3 * 10^{-92}$
	Ribo nucleo protein	$1.1 * 10^{-88}$
Brain	myelin sheath	$1.9 * 10^{-29}$
	extracellular exosome	$4.6 * 10^{-15}$
	extracellular space	$4.3 * 10^{-7}$
	Acetylation	$2.8 * 10^{-6}$
Liver	Ribosomal protein	$1.3 * 10^{-31}$
	Ribo nucleo protein	$1.9 * 10^{-27}$
	structural constituent of ribosome	$1.9 * 10^{-27}$
	ribosome	$1.3 * 10^{-27}$
	substantia nigra development	$1.8 * 10^{-5}$
Muscle	Ribosomal protein	$7.2 * 10^{-68}$
	structural constituent of ribosome	$1.8 * 10^{-61}$
	Ribosome	$5.3 * 10^{-61}$
	ribosome	$1.9 * 10^{-60}$
	cytosolic small ribosomal subunit	$1.3 * 10^{-27}$
Pancreas	Ribosomal protein	$8.6 * 10^{-30}$
	ribosome	$7.2 * 10^{-26}$
	Ribosome	$1.3 * 10^{-25}$
	Ribonucleoprotein	$2.9 * 10^{-25}$
	Ribonucleoprotein	$4.4 * 10^{-59}$
	structural constituent of ribosome	$2.9 * 10^{-24}$

Table S2: **Most enriched GO categories for highly expressed genes.** Genes with high levels of expression belonging to the first regime of the Zipf's law in different organs were selected for the analysis.

Organ	Related GO	Benjamini
Bladder	extracellular exome	$8.4 * 10^{-17}$
	secreted	$1.3 * 10^{-7}$
	extracellular matrix	$2.3 * 10^{-7}$
	disulfide bond	$1.1 * 10^{-6}$
Brain	blood microparticle	$1.1 * 10^{-5}$
	haptoglobin-hemoglobin complex	$3.1 * 10^{-5}$
	extracellular exosome	$3.6 * 10^{-5}$
	haptoglobin binding	$5.1 * 10^{-5}$
Liver	blood microparticle	$3.3 * 10^{-10}$
	extracellular exosome	$1.8 * 10^{-9}$
	haptoglobin-hemoglobin complex	$1.1 * 10^{-4}$
	extracellular space	$1.2 * 10^{-4}$
Muscle	extracellular space	$2.7 * 10^{-10}$
	Ribosomal protein	$4.5 * 10^{-10}$
	Acetylation	$7.6 * 10^{-12}$
	Muscle protein	$4.5 * 10^{-10}$
Pancreas	extracellular space	$6.3 * 10^{-19}$
	Secreted	$2.0 * 10^{-17}$
	Disulfite bond	$2.9 * 10^{-17}$
	signal peptide	$2.2 * 10^{-16}$

Table S3: **Most enriched GO categories for genes whose predicted occurrence is significantly higher than the empirical one.** We selected genes with $o_{predicted} > o_{real} + 0.2$ (as highlighted in Figure 7 of the main text). The complete gene lists are available from `https://github.com/SilviaLazzardi/The_single_cell_transcriptome_as_a_component_system/tree/main/MouseCellAtlas/OrealOpred_lists`.