Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in
Systems Biology

# Entropy as a measure of variability and stemness in single-cell transcriptomics

Olivier Gandrillon[1,2], Mathilde Gaillard[1,3], Thibault Espinasse[2,3], Nicolas B. Garnier[4], Charles Dussiau[5,6], Olivier Kosmider[5,6] and Pierre Sujobert[7]

**Abstract**

In the present article, we will try to clarify what entropy is, how it is computed, and we will show that the same quantity can be used to measure very different things. We will thus try to clarify for the biologist what is captured by entropy depending on its various usage.

**Addresses**

[1]Univ Lyon, ENS de Lyon, Univ Claude Bernard, CNRS UMR 5239, INSERM U1210, Laboratory of Biology and Modelling of the Cell, 46 Allée d'Italie Site Jacques Monod, Lyon, F-69007, France
[2]Inria Team Dracula, Inria Center Grenoble Rhône-Alpes, France
[3]Univ Lyon, Université Lyon 1, CNRS UMR5208, Institut Camille Jordan, 43 Blvd du 11 Novembre 1918, Villeurbanne-Cedex, F-69622, France
[4]Univ Lyon, Ens de Lyon, Univ Claude Bernard, CNRS UMR 5672, Laboratoire de Physique, Lyon, F-69342, France
[5]Université de Paris, Institut Cochin, CNRS UMR8104, INSERM U1016, Département Développement, Reproduction, Cancer, Paris, France
[6]Assistance Publique-Hôpitaux de Paris, Centre - Université de Paris, Service d'Hématologie Biologique, Hôpital Cochin, Paris, France
[7]Service d'Hématologie Biologique, Hospices Civils de Lyon, 165, Chemin du Grand Revoyet, Pierre-Bénite Cedex, 69495, France

Corresponding author: Gandrillon, Olivier (olivier.gandrillon@ens-lyon.fr)

**Current Opinion in Systems Biology** 2021, **27:**100348

This reviews comes from a themed issue on **Theoretical approaches to analyze single-cell data (2021)**

Edited by **Olivier Gandrillon** and **Michael Stumpf**

For complete overview of the section, please refer the article collection - Theoretical approaches to analyze single-cell data (2021)

Available online 23 May 2021

**https://doi.org/10.1016/j.coisb.2021.05.009**

2452-3100/© 2021 Elsevier Ltd. All rights reserved.

**Keywords**

Entropy, Single-cell transcriptomics, Stemness, Differentiation.

Guest Editor: Olivier Gandrillion had no involvement in the peer-review of this article and has no access to information regarding its peer-review. Full responsibility for the editorial process for this article was delegated to Michael Stumpf.

## Why use entropy?

Biology can be described as evolving between typological and population thinking, as proposed by Ernst Mayr [15]. In other words, biology might either focus on the type (i.e. the invariant or Plato's *eidos*), or focus on the population (in evolutionary biology), or the variation away from the type. Charles Darwin is of course the main contributor of this new way of thinking, where variation comes first. Statistics have evidently taken the main stage when it came to quantifying the extent of variability, through the concept of variance. More recently, concepts from Statistical Physics, like entropy, have been tentatively used for quantifying variability in biological systems. Initially forged for measuring the disorder or uncertainty in a physical system, then reused within the frame of information theory, entropy is now being increasingly used as a measure of variability and used as a proxy for stemness[1] in single-cell transcriptomics data.

There has been an occasional use of entropy for population-based measurements (see e.g. Refs. [2,25,49]), but the recent explosion of single-cell—based measurements has really led to a strong increase in the use of entropy as a relevant metric for quantifying variability at the single-cell level.

One of the first articles to advocate for the use of entropy as a measure of variability was Ref. [46]. The authors suggested that "This connection suggests a broad principle: at equilibrium, cell populations that are subject to strict regulatory constraints should exhibit well-defined and low entropy expression patterns, whereas those that are subject to weaker regulatory constraints should exhibit more diverse, higher entropy expression patterns. Viewing variability in this light indicates that PSC (pluripotent stem cells) populations may be more diverse than differentiated populations because they are subject to weaker regulatory constraints." The authors went on to propose an information-theoretic interpretation of stem cell dynamics that views cellular

[1] Stemness can be defined as the ability for a cell to self-renew and to give rise to multiple types of differentiated progeny. It has been proposed that stem cells should be considered as 'a state rather than an entity' [93] and that: 'The stem cell state is open minded' meaning that it will be having access to several differentiation programs [50].

multipotency as an instance of maximum entropy statistical inference [62].

Since this first proposal, there has been a vast amount of literature that made use of entropy for the analysis of single-cell transcriptomics data. Those types of data have to be seen as of *distributions* of values [47], because of the unavoidable stochastic nature of gene expression [39]. This massive use led to a situation where the use of the term 'entropy' can be misleading, because it measures different things.

We therefore propose in the present paper a simple typology that should help nonspecialists readers to better grasp the relevance of that term, and differentiate its various occurrences.

## What is entropy?

Entropy was first introduced by Rudolf Clausius to express the degradation of energy in thermodynamics at the time of the first industrial revolution [24]. Ludwig Boltzmann later revealed that this thermodynamical entropy could indeed be expressed as a measure of uncertainty, or *mixed-up-ness* of a physical system, that is, a measure of the unpredictability of the microscopic degrees of freedom that are unknown to an observer recording macroscopic variables [5]. In the 1940s, this measure of disorder was generalized by Claude E. Shannon to form the basis of a new scientific field: 'Information theory' [17,72]. Within this perspective, any signal, measurement or random variable can be assigned an entropy which measures its unpredictability, or in more celebrated terms its information contents.

Entropy is a functional of the probability density function (PDF). This simply means that given a signal $X$, with a PDF $p(x)$, which is nothing but the normalized histogram of $X$, the entropy can be computed using a formula that only involves $p$:

$$H(X) = -\int_S p(x)\log(p(x))dx.$$

In this expression, we have assumed that $X$ takes continuous values, for example, real numbers, and we have noted $S$ the support of the PDF, that is, the set of all possible values that $X$ can take. In that case, $p$ is said to be continuous and the entropy is often referred to as the *differential entropy*. Another important class of signals correspond to those which take discrete values, for example, integer values. In that case, $p$ is said to be discrete and the entropy is expressed by the famous Shannon formula

$$H(X) = -\sum_{m\in S} p(m)\log(p(m)).$$

In practice, and in particular, when dealing with single-cell transcriptomics, it is usually more comfortable to use discrete distributions. This is natural not only if data is composed of integers (for example, a number of molecules), but it is also the form that appears when constructing the histogram of $X$ to estimate p, whether it is discrete or continuous. The histogram is composed of successive bins indexed by an integer $m \in [1; M]$, where $M$ is the number of bins in the histogram (see Section 2 below). Each bin represents an interval of possible values for $X$, and $p(m)$ is related to the probability for the variable $X$ to have its value in the bin $m$.

Beware that the Shannon entropy computed after this binning procedure does not estimate consistently the differential entropy. Therefore, one has to be careful when comparing entropies of distributions using this binning step.

Here, we are facing two difficulties: first, theoretical conditions to ensure convergence of entropy computed after a binning procedure to the differential entropy are not straightforward [58,75]. Second, the true distribution $p$ is unknown, and needs to be estimated from the data.

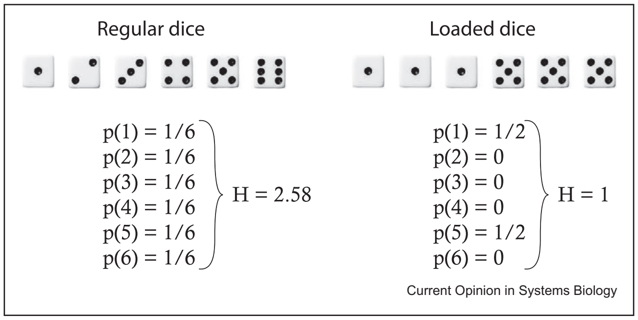
To fix the ideas, instead of Shanon entropy, it is possible [85] to compute $-\sum_{m\in S}\widehat{p}(m)\log\left(\frac{\widehat{p}(m)}{w(m)}\right)$, where $w(m)$ corresponds to the length of the $m$th bin and $\widehat{p}(m)$ correspond to an estimation of $p(m)$ (for instance, the proportion of the sample that falls in bin $m$).

At that stage, it is worth mentioning that natural estimation of entropy exhibits a tendency to underestimation. Figure 4 in the Glossary illustrates this property. However, this bias vanishes as the number of observations increases [3]. Of course, many more advanced methods have been proposed to improve estimation. We have chosen not to present these methods here, but encourage the curious reader to dig deeper into estimation methods (To cite but a few: 'Miller—Madow estimate' [51], Chao and Shen construction [13], Hausser and Strimmer 'James—Stein Shrinkage estimator' [31], 'Best Upper Bounds' Paninsky estimate [57], or, for differential entropy, kernel estimation [38] or k-nearest-neighbor estimation [42]). All these estimators may also be used for Information Theoretical Measures for instance to build Information-based Gene Regulatory Networks [11].

Other authors (e.g. Ref. [36]) proposed to use a normalized entropy, for instance, by rescaling Shannon entropy with its maximum value $\log(M)$: $-\frac{1}{\log(M)}\sum_{m\in S}\widehat{p}(m)\log(\widehat{p}(m))$ to get a quantity between 0 and 1.

The entropy as defined by Shannon is a measure of the uncertainty or information contained in a signal. If the probability is uniform (think about a regular dice with 6 perfectly equi-probable values, $p(m) = p = 1/6$ for all $m \in [1..6]$), then the uncertainty is maximal, and so is the entropy. On the contrary, if one bin or one of the possible outcomes has much larger probability (think about a loaded dice), then the uncertainty is smaller, and so is the entropy (see Figure 1).

**Figure 1**



Current Opinion in Systems Biology

In the case of a regular dice, the entropy is equal to – (1/6 × $\log_2(1/6)$) × 6, that is 2.58. In the case of the loaded dice, the entropy is equal to – (1/2 × $\log_2(1/2)$) × 2, that is 1.

The extreme case where only a single value has a nonvanishing probability — which is then equal to 1 — corresponds to a Dirac distribution (see Glossary below), which has an entropy equal to 0, according to the Shannon formula.

In addition, the differential entropy evolves as the logarithm of the variance of the distribution: if one considers the reduced variable $y = (x - \mu)/\sigma$, where $x$ denotes the expected value of $X$ and $\sigma$ its standard deviation, then

$$H(X) = H(Y) + \log \sigma,$$

which indicates that the entropy of a data set $X$ increases with its standard deviation.[2]

Entropy is therefore a measure of the uncertainty, understood as the possible surprise of an outcome, and hence a measure of the information contained in the probability distribution. One part of the entropy is due to the standard deviation of the signal, whereas another part is related to the shape of its distribution. As such, it offers a fully relevant measure of variability [10].

## How to compute entropy

In this review, we will focus on the use of entropy for analyzing single-cell omics data. Those are information regarding the molecular content of one cell at a time [90]. We will more specifically focus on single-cell transcriptomics data, where a number of mRNA molecules for different genes is determined at the single-cell level.

There are a number of open issues regarding the proper handling of such single-cell mRNA expression values. It ranges from the definition of a proper statistical model to account for the specific nature of those data, including a high proportion of null values [14,66], up to normalization issues [16]. We refer the interested reader to a recent review [44].

Single-cell mRNA expression has been shown to be well fitted using a Gamma distribution ([1]; see Glossary below). It has been shown that a negative binomial distribution, the discrete version of the Gamma law can be derived under certain conditions (i.e. in a bursty regime) from a mathematical analysis of a two-state model for gene expression [71]. If one assumes that this is a correct model for the data, and that the data are good enough for a proper estimation of the parameters of the Gamma distribution, then one can derive the entropy from the analytical expression of the Gamma distribution.[3]

The alternative path consists of nonparametric estimations of the entropy, as follows:

A single-cell omics experiment will generate values, for example, Ct values in the case of an RTqPCR experiment or the number of mRNAs detected into a number of cells in the case of an RNA-seq experiment. From such continuous or discrete data, one must first estimate the probability distribution the values were drawn from to calculate the entropy. Although this might seem from a biologist's point of view like a trivial task, especially in the discrete case of RNA-seq data, a closer examination shows that this is a really difficult question (see e.g Refs. [32,43,57]). It can be rephrased as how good is the estimate the entropy $H(X)$ given by:

$$\widehat{H}(X) = -\sum_{m \in M} \widehat{p}_m \log(\widehat{p}_m)$$

where $\widehat{p}_m$ is the (normalized) frequency at which one can find cells or genes values in the given bin indexed by $m$. One should note here that this value is computed for one cell, or one gene (see below), in other terms, it is a univariate distribution. No multivariate version of entropy has been proposed for the single-cell field, although it might

[2] The entropy of $X$ can be understood as the sum of two terms: the entropy of $Y$ (the normalized version of $X$, which is independent of $\sigma$), which relates to the shape of the PDF, and log $\sigma$, which describes the width of the PDF.

[3] $H = \alpha - ln(\beta) + ln\Gamma(\alpha) + (1 - \alpha)\psi(\alpha)$, with $\alpha = \mu^2/\sigma$ and $\beta = \mu/\sigma$, $\Gamma$ being the Gamma function and $\psi$ being the digamma function.

help to better characterize data sets of intrinsically multivariate nature.

$\widehat{p}_m$ is not the real probability but an *estimate* of that probability based on a limited data set, and on the choice of the bins. The question then arises as to what is the best binning procedure, and how it impacts the resulting calculation.

As an example, say we have measured the diameter of five cells:

| Cell | Diameter ($\mu$m) |
|---|---|
| C1 | 4.5 |
| C2 | 3.8 |
| C3 | 5.3 |
| C4 | 4.7 |
| C5 | 4.2 |

One can now decide to regroup those cells by the following size class (bin):

| Bin | Count | PDF |
|---|---|---|
| [3.5; 4.0] | 1 | 1/5 |
| [4.0; 4.5] | 1 | 1/5 |
| [4.5; 5.0] | 2 | 2/5 |
| [5.0; 5.5] | 1 | 1/5 |

The data is then partitioned in discrete boxes (bins) and in each of these, one can calculate $\widehat{p}_m$, and thus deduce the estimate $\widehat{H} = 7/5\log 5 - 4/5\log 2 = 1.699$. But one can also decide to use a more compact view of the original data set. This can simply be achieved by increasing the size of the bins to get:

| Bin | Count | PDF |
|---|---|---|
| [3.5; 4.5] | 2 | 2/5 |
| [4.5; 5.5] | 3 | 3/5 |

which leads to an estimate $\widehat{H} = 4/5\log(2/5) + 9/5\log(3/5) = 1.653$, that is smaller than the former one.

There is an abundant literature as to what should be the proper way of defining a relevant binning size. It has been shown that extreme bin size (either too small or too large) introduces a bias in the estimation [60]. Sophisticated methods have been proposed for a proper bin size estimate, ranging from the use of Doane's rule [19,59], an extension of Sturges' formula, to the use of the Bayesian Blocks algorithm, a method designed to find an optimal binning for a set of values without enforcing uniform bin width [67,78]. The very existence of so many methods shows how the proper binning still is an open and difficult question. For more advanced methods for data-driven binnings, we can refer, for instance, to Ref. [7] or [45] in the context of density estimation or to Ref. [75] for some consistency results for entropy estimation.

Such a binning issue is relevant for scRTqPCR data where the initial values are expressed as a function of the Ct which is a continuous value. It is also relevant in the case of discrete values like the counting of molecules in scRNAseq experiments: one can choose a number of bins that is equal to the highest number of molecules detected in any cell of the data set (and hence a bin size equal to 1), or increase the bin size up to having only two bins: one for the zero expression level and the other one for any nonzero expression level [89].

## When to study entropy?

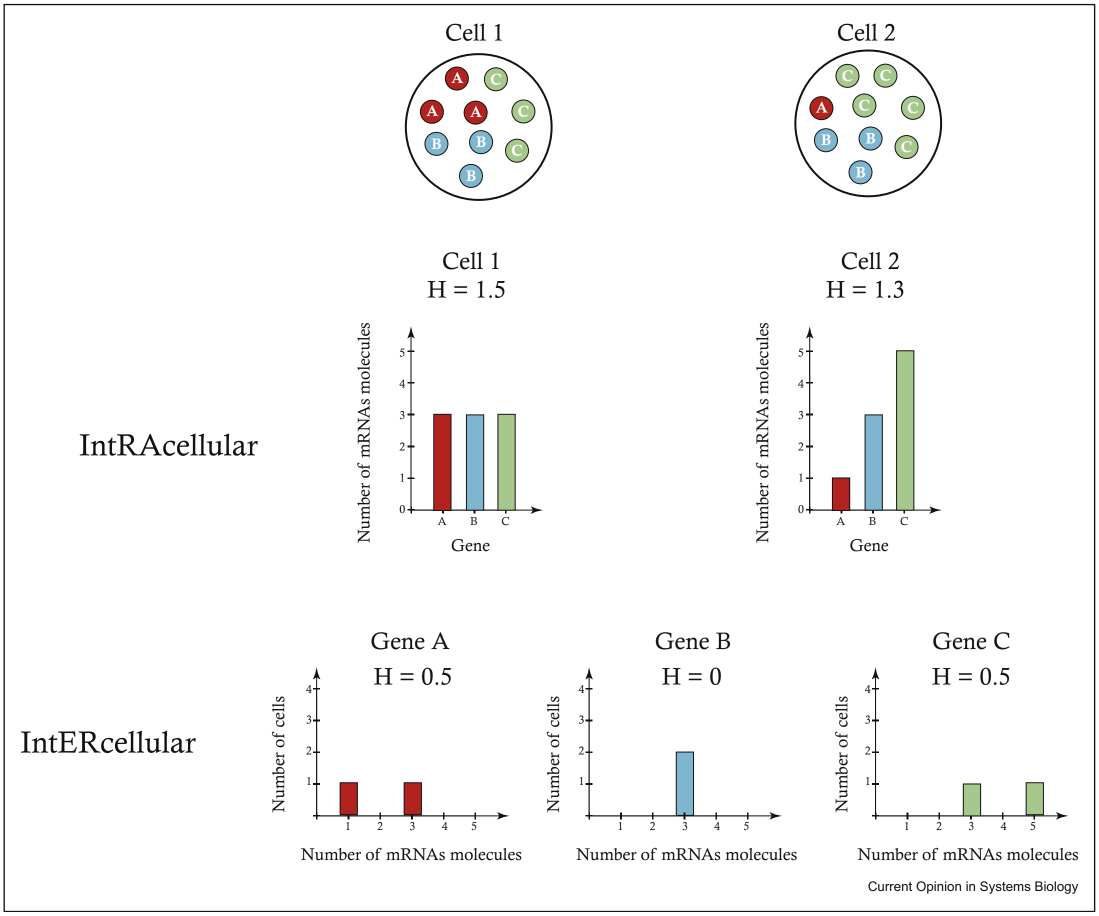
We would first like to point out that one can calculate two sorts of entropies: an IntRAcellular Entropy and an IntERcellular Entropy (Figure 2). The IntRAcellular Entropy is endowing a CELL with an entropy value, whereas the IntERcellular Entropy is endowing a GENE with an entropy value. In other words, in the IntRAcellular Entropy, one tries to capture the heterogeneity of the transcriptional state of a given cell. In the IntERcellular Entropy, one tries to capture the heterogeneity of the transcriptional state of a given gene across a population of cells.

Based on a thorough literature search, we can then refine this proposal and classify entropy for single-cell omics data analysis into four different categories: one can estimate an entropy per cell or an entropy per gene, and one can use external information (like Gene Ontology of a protein–protein interaction network) to compute this entropy or obtain it directly from a given distribution. All the available literature can be classified in such a way (Table 1 below).

### IntRAcellular entropy requiring an external information

In the SLICE algorithm [30], Gene Ontology (GO) clusters define the external information. Genes are first attributed to those GO clusters, and then the algorithm computes the entropy of the clusters distributions. A high entropy is linked with a flat distribution where all GO clusters are more or less equally represented (i.e. contains a similar number of genes), and there are many potential functions harbored by the cells, whereas a

**Figure 2**



Current Opinion in Systems Biology

A schematic description of the two different entropies. On the first line are displayed two cells with their mRNA content for three genes. The second line shows how one can estimate an IntRAcellular Entropy. Here the entropy for Cell 1 is equal to $-(1/3 \times \log_2(1/3)) \times 3$, that is 1.5. The entropy for Cell 2 is equal to $-((1/9 \times \log_2(1/9)) + (3/9 \times \log_2(3/9)) + (5/9 \times \log_2(5/9)))$, that is 1.3. The mean IntRAcellular entropy is therefore of 1.4. The third line shows how one can estimate an IntERcellular Entropy. Here the entropy for Gene B is null, and the entropy for both Gene A and C is equal to $-(1/2 \times \log_2(1/2))$, that is 0.5. The mean IntERcellular Entropy is therefore of 0.33.

reduced entropy is linked with a distribution that concentrates on a smaller amount of clusters/functions, being the hallmark of a differentiated cell.

For the single-cell entropy (SCENT) algorithm, the external information comes from a protein–protein interaction (PPI) network [80]. scRNA-Seq values are superimposed on top of this PPI network. Once more, immature stem cells are defined by more promiscuous signaling pathways, whereas more mature cells are characterized by a narrower distribution. This was an extension of an entropy measure that was initially defined over bulk tissue data [6].

**Table 1**

**The four entry table for entropy use in single-cell omics. In the case of [41,91], the entropy calculation is based upon a *Reference Cell* (see below), which can be either an external input or based on an intrinsic calculation from the data set. There is no known example of an IntERcellular Entropy requiring external information, but it could be easily designed, for example, using a given gene expression distribution as a reference.**

| | IntRAcellular Entropy | IntERcellular Entropy |
|---|---|---|
| Requires external information | **SLICE** [30], **scEntropy** [41], **MCE** [74], **SCENT** [80], **scEntropy** [91] | |
| Without external information | **dpath** [27], **StemId** [28], **single-cell entropy** [41], **Shannon entropy** [59], **Palantir** [70], **CEE** [86], **scEntropy** [91], **scRCMF** [92], | **Shannon entropy** [21,61,78,89] |

**Table 2**

**Monotonous decrease in cell-based entropy. Cells 1 and 2 are stem cells, 3 and 4 are progenitor cells, and 5 and 6 are mature cells characterized by the elevated expression of Gene 1. All the distribution gets progressively concentrated on one gene.**

| | Gene 1 | Gene 2 | Gene 3 | Gene 4 | Entropy |
|---|---|---|---|---|---|
| Cell 1 | 1 | 2 | 4 | 3 | 2 (High) |
| Cell 2 | 1 | 2 | 3 | 4 | 2 (High) |
| Cell 3 | 5 | 7 | 0 | 0 | 1.5 (Medium) |
| Cell 4 | 0 | 0 | 8 | 5 | 1.5 (Medium) |
| Cell 5 | 15 | 0 | 0 | 0 | 0.8 (Low) |
| Cell 6 | 18 | 0 | 0 | 0 | 0.8 (Low) |

A similar philosophy was introduced in Ref. [74] with the Markov Chain Entropy (MCE) where the authors compute a Markov transition matrix (see Glossary below) through a PPI. They demonstrate that the MCE behavior is driven by a correlation between mRNA expression and network connectivity, and conclude that it outperforms three other single-cell potency models.

### IntRAcellular entropy without an external information

The StemId algorithm [28] computes a cell-based entropy $H = -\sum_{i,j} p_{i,j} \log(p_{i,j})$, where $p_{i,j} = n_{i,j}/N$ and $n_{i,j}$ equals the number of transcripts of gene i in cell j. N equals the total number of transcripts in each cell, which has been normalized. From this, the authors compute an entropy score for each cluster, as defined by k-medoids clustering, to identify stem cell clusters [28]. The authors assume that a more promiscuous transcriptome, reflected in a higher entropy, would be expected for stem cells, when a more confined transcriptome would be expected for a mature cell type.

The Palantir algorithm [70] computes yet another entropy value, that also requires the construction of a Markov transition matrix (see Glossary) with branches. In contrast with MCE, this transition is built from the data, and not from an external information. One can then compute a vector of branch probabilities to reach each of the possible terminal states, and compute the entropy of such a vector. It allows to define a differentiation potential in the sense that immature cells have a higher probability to reach a large number of terminal states. This indicator, as the previous ones, decreases monotonously as differentiation proceeds.

The dpath algorithm computes a metagene entropy [27]. It is based on the construction of metagenes (vectors of genes) using a weighted Poisson non-negative matrix factorization (wp-NMF) method, and on the computation of an entropy on the metagene coefficients V, a probabilistic simplex that indicated the relative weight of each metagene in each cell. The metagene entropy of cell m is then defined as $-\sum_{k=1}^{K} V_{km}\log(V_{km})$. This metagene entropy serves as a measure of how many distinct programs (parts) are active (expressed) in a cell and was significantly higher in progenitor cells compared with more differentiated cells. It therefore allowed the ranking of cells based on their differentiation potential.

**Table 3**

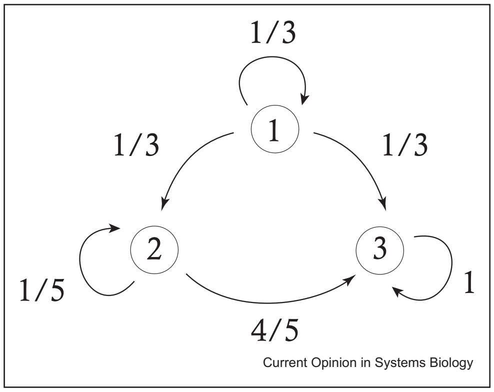
**Nonmonotonous behavior of gene-based entropy. Cells 1 and 2 are stem cells, 3 and 4 are progenitor cells, and 5 and 6 are mature cells characterized by the expression of Gene 1. The intermediary phase is characterized by a state of uncertainty in which entropy is peaking [26,54,61].**

| | Gene 1 | Gene 2 | Gene 3 | Gene 4 | Mean entropy |
|---|---|---|---|---|---|
| Cell 1 | 1 | 2 | 4 | 3 | |
| Cell 2 | 1 | 2 | 3 | 4 | |
| Ent | 0 | 0 | 1 | 1 | 0.5 (Low) |
| Cell 3 | 5 | 7 | 0 | 0 | |
| Cell 4 | 0 | 0 | 8 | 5 | |
| Ent | 1 | 1 | 1 | 1 | 1 (High) |
| Cell 5 | 15 | 0 | 0 | 0 | |
| Cell 6 | 18 | 0 | 0 | 0 | |
| Ent | 1 | 0 | 0 | 0 | 0.25 (Low) |

Up to that point, all observations concluded toward a monotonous decrease of IntRAcellular Entropy as a function of the differentiation process.

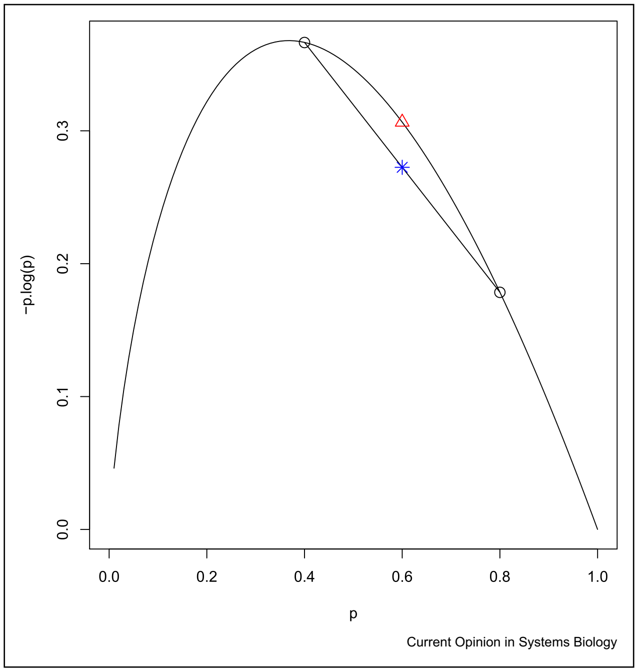
The only exception concern the study of the very early development in human and mouse [41,59,92]. In the first two studies, the authors observe a steady increase from the oocyte up to the blastocyst stage [41,59]. They propose that such an increase in heterogeneity would have been necessary for cell fate diversifications [59]. Note that the second exception uses the single-cell entropy (scEntropy), where $p_i(y)$ is the distribution

**Figure 3**



Transition graph between three states. In this example, each arrow corresponds to a probability of transition within one time unit, between two states.

**Figure 4**



Current Opinion in Systems Biology

Illustration of the bias in entropy estimation. The function $g$: $p \mapsto -p\log(p)$ is mapped here. For any random variable $\widehat{p}$, we have, $\mathbb{E}[g(\widehat{p})] \le g(\mathbb{E}[\widehat{p}])$ thanks to the concavity of $g$. For instance, if $\widehat{p}$ takes values 0.4, 0.8 (black dots) with probability 1/2, then $\mathbb{E}[\widehat{p}] = p_0 = 0.6$ and $\mathbb{E}[g(\widehat{p})]$ is given by the $y$-coordinate of the blue dot, whereas $g(p_0)$ is given by the $y$-coordinate of the red triangle. Hence, $\mathbb{E}[-\widehat{p}\log(\widehat{p})] \le -p_0\log(p_0)$. By summing over all bins, this illustrates why $\mathbb{E}[\widehat{H}(X)] = -\sum_i \mathbb{E}[\widehat{p}_i\log(\widehat{p}_i)] \le -\sum_i p_i\log(p_i) = H(p)$.

density of the components $y_{i,j}$ in $y_i = x_i - r$, r being the gene expression vector r of a *Reference Cell* [41]. The third study, using the scRCMF algorithm, which is a very similar NMF approach as Ref. [27], describes a transient increase of entropy during what the author describes as 'transition states' [92]. The authors do not discuss the discrepancy between their results and the previously published ones.

The Cellular Entropy Estimator [86] is defined as $-\sum_{j=1}^{K} P_{i,j}\log(P_{i,j})$, where $(P_{i,j})$ is the probability for the cell i to be attributed to the cluster j by the SoptSC algorithm [87]. As such it is more a measure of the ability of a cell to transition to a new state than a measure of stemness per se.

Up to this point, most of the studies have shown that the very early development shows a monotonous increase in cellular entropy, whereas differentiation from adult stem cells shows a monotonous decrease in entropy. The only two exceptions are the use of scEntropy, which was shown to increase during the differentiation of iPSCs into cardiomyocytes [91] and the use of scRCMF during mammalian preimplantation development [92]. In the first case, it could be the choice of the *Reference Cell* that influences such a directionality of entropy changes. In the second case, it would be interesting to see the result of the application of dpath [27], a very similar algorithm, to an early mammalian development data set.

## IntERcellular entropy without an external information

In any case, the IntRAcellular entropy shows mostly a monotonous behavior. This is in sharp contrast with the study by Ref. [61]. In this work, we proposed an IntERcellular version of Shanon entropy, computed from the raw data. For each gene, one can compute its entropy from its distribution in a population of cells. By measuring single-cell gene expression during the differentiation of chicken erythroid progenitors, we demonstrated the existence of a surge in variability in gene expression in the midst of this erythroid differentiation process, with a final differentiated state reaching a lower entropy than the initial one. This was the first example of a differentiation process where IntERcellular Entropy could evolve in a nonmonotonous way.

Very similar results have been obtained in a different setting [78]: the authors have been measuring single-cell expression by RTqPCR during the differentiation of mouse embryonic stem cells along the neuronal lineage. Using a very similar version of Shanon entropy as in Ref. [61], they also observed an increase in variability at the beginning of the differentiation process.

Such a nonmonotonous behavior of entropy was further corroborated by Ref. [89]. To solve the binning issue (see upper), the authors argue that only two obviously separated levels can be easily distinguished: the zero expression level and the greater-than-zero expression level. This led them to propose a binary Shannon entropy, calculated on two bins. Although this might seem like a very different measure than Ref. [61] or [78], they also demonstrate the existence of a surge in entropy in long-term hematopoietic stem cell differentiation as well as in EML cell line erythroid commitment.

More recently, we computed a gene-based Shannon entropy on Single-cell RNA-seq data from normal and pathological human bone marrow. We demonstrated the surge in entropy in the main hematopoietic differentiation pathways on normal and myelodysplastic syndromes samples. We also showed that entropy is increased in hematopoietic stem cell of myelodysplastic syndromes as compared to age-matched control, suggesting a role for gene expression variability in the pathophysiology of this disease [21].

One key point in computing an IntERcellular Entropy lies within the proper definition of the group of cells on which the gene entropy is computed. In Refs. [61,78], the cells have been harvested at different time points offering a natural cell grouping scheme. In Ref. [89], the authors group the cells by predefined cell types, based on known gene expression patterns. In Ref. [21], we propose a more data-driven approach: the cells are first ordered by their pseudo-time as assessed by Slingshot [76], and the entropy is calculated on sliding windows across such a pseudo-time. Various reordering scheme do exist [8,9,68], but they have not yet been systematically assessed for their ability to reorder cells in an entropy-relevant manner.

### Why IntRAcellular and IntERcellular entropy differ?

At that stage, it is remarkable that almost all techniques aiming at measuring an IntRAcellular entropy do show a monotonous decrease in entropy as the differentiation proceeds, whereas all IntERcellular metrics point toward a nonmonotonous behavior, that displays a maximum before it decreases again.

To propose an explanation for those observations, we would like to make a reasoning on a toy data set, represented in Tables 2 and 3. We display in those matrices six cells, ranging from stem to mature, and four genes. The stem cells are characterized by a low and promiscuous expression of all the genes, thereby preserving their potential for different lineages choices. As differentiation proceeds, some genes get repressed and some get activated. This tends to concentrate the gene expression pattern on a smaller number of genes, thereby reducing the entropy. Ultimately, the mature cell type is characterized by the expression of a small number of genes (here one), and a low entropy.

To account for the surge in IntERcellular entropy, one should now focus on the intermediate 'progenitor' population. In this case, one proposes that cells are passing through a stage of so-called 'hesitant' behavior [54]. During this stage, each cell explores (at its own pace and independently of cell division) many different possibilities before reaching a stable combination of genes to be expressed. This peak of uncertainty can be captured by entropy measurements (see Table 3), but also through the reconstruction of transcriptional uncertainty landscapes [26], or through the mathematical definition of a potential energy of a population of cells [83].

## Discussion

It is no wonder why the entropy concept has become so popular in the era of single-cell omics. Indeed, even though the stochastic nature of gene expression has been anticipated for quite some times [35,37], single-cell studies have established its unavoidable nature [23,39,79]. This led to the critical need for tools that can help to make sense of distribution-based evidence [47] away from the classical mean-based vision [39].

Although all the examples in this review are from the single-cell transcriptomics field, one should note that most, if not all, of the known molecular techniques are being adapted to single-cell studies [90]. The resultant high-dimensional single-cell data generated require new theoretical approaches and analytical algorithms for effective visualization and interpretation. Statistical physics tools and especially information-theoretic ones, like entropy, therefore became largely used. There are of course more uses of entropy in biology, both within (see e.g. Ref. [40]) and outside of the single-cell omics field (see e.g. Refs. [16,34]), but that would be beyond the scope of this review to analyze them all.

Beyond the need for a consensual binning procedure (see upper), one currently missing tool is the absence of a statistical test for comparing the entropy of two distributions and to decide whether or not their entropy is significantly different. We are currently working on this issue.

Other indicators of variability do exist, from the variance, up to the coefficient of variation (CV; $\sigma/\mu$), the Fano factor ($\sigma^2/\mu$), and the normalized variance ($\sigma^2/\mu^2$). Some authors have proposed a measure of 'transcriptional noise' based on pairwise cell—cell distanced calculated as $d = \sqrt{(1-\rho)/2}$, with $\rho$ being the Spearman's rank correlation coefficient [52].

One has to state here that entropy stands out from those measures, as it can be shown to measure the variability of the elements within a given distribution, and that its expression is not arbitrary, as it is the only linear indicator for such a concept [10].

All measures of molecular variability are influenced by a variety of potentially confounding factors, including a gene's mean expression [22]. Some authors proposed to condition CV values on mean by computing the residuals of a nonparametric loess regression of CV-mean [82] to eliminate such a mean dependency.

Such bias advocates for care to be taken when drawing inferences about the role of biological variability using such indicators. Nevertheless, our review shows that the use of entropy was instrumental in highlighting the connection between stemness and uncertainty [62] or between differentiation and an increase in molecular variability [20,29,53], leading to the concept of noise-induced differentiation [12,33,65,84].

Altogether, those results strongly support the view that cell differentiation is a probabilistic process [37,56]. This is rooted in the fact that cells are neither machines [55], nor simple information processing devices. Cells (like all living systems) are rooted within a physicochemical world to which they belong. Their specific complexity nevertheless sometimes led to the idea that they should be treated differently than classical physicochemical systems [69]. But their nature of dynamical systems is exemplified here by the relevance of the use of a statistical physics concept, like entropy.

From a more biological perspective and especially in the field of cancer, further single-cell study-based entropy analysis of splice variants [48] may help us to better understand the consequences of splicing alteration in the disease establishment and progression as previously suggested in bulk analysis [63]. Moreover, future single-cell analysis could consider some poorly explored players of the cells, such as miRNAs [88] or lncRNAs [18], which could vary significantly during normal and pathological processes of differentiation. Lastly, all these considerations have to be explored again in approaches including the consequences of spatial cell—cell interactions [64,73,77], for example, between hematological stem cells and their niche [4].

In conclusion, we hope the reader is now convinced of the versatility and power of entropy to quantify the extent of variability contained in a distribution, and that next time, he/she will investigate how to unravel the importance of variability during a biological process, he/she will not hesitate to make the best use of entropy as a relevant measure of variability.

## Glossary

### Dirac

The **Dirac** delta is a mathematical object $\delta$ that is characterized by the following relation, for any $a$, $b$:

$$\int_a^b \delta(x)\mathrm{d}x = \begin{cases} 1 \text{ if } 0 \in (a,b) \\ 0 \text{ otherwise} \end{cases}$$

It can be loosely thought of as the limit of function on the real line which are picking at the origin, and taking the value zero everywhere else, and which is also constrained to satisfy the identity

$$\int_{-\infty}^{\infty} \delta(x)\mathrm{d}x = 1.$$

Intuitively, it corresponds to the distribution of a constant random variable $X = 0$:

$$\mathbb{P}(X \in (a,b)) = \int_a^b \delta(x)\mathrm{d}x = \begin{cases} 1 \text{ if } 0 \in (a,b) \\ 0 \text{ otherwise} \end{cases}$$

Such random variable cannot be described properly by a PDF.

### Gamma distribution

The **Gamma** distribution is a two-parameter family of continuous probability distributions for positive real values. It is highly versatile and it occurs frequently in models used in engineering, business, or biology for which the variables are always positive and the results are skewed (unbalanced), and therefore not captured by the more 'classical' normal distribution.

### Markov transition matrix

A **Markov transition matrix** describes how a random system evolves as a function of time. Formally, a transition matrix $P$ is a real nonnegative square matrix with each row summing to 1. It number of rows is equal to the number of possible states for the system, and the $i$th row is the probability distribution of the state at time $t + 1$, *knowing* it is in state $i$ at time $t$. Hence, $P_{ij}$ corresponds to the probability for the system to be in state $j$ at time $t + 1$ *knowing* it is in state $i$ at time $t$, see Figure 3 for a trivial example.

In this case, the corresponding transition matrix is given by

$$P = \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 0 & 1/5 & 4/5 \\ 0 & 0 & 1 \end{pmatrix}$$

One can then compute an entropy per row, and one can see in the example that this entropy will be decreasing from type 1 ($H = 1.58$) to type 3 ($H = 0$).

### Biased estimation

The maximum likelihood estimation $\widehat{H}(X)$ of Shannon entropy is always **negatively biased** [3]:

$$\mathbb{E}_p[\widehat{H}(X)] \leq H(p),$$

which means that it is, *in average*, underestimated. Figure 4 tries to give an insight for this fact.

## Note added in proof

While this manuscript was being revised, we became aware of a review including similar views, and more [81] which will be of interest for the readers of the present review.

## Conflict of interest statement

Nothing declared.

## Acknowledgements

This work has been supported by grants from the Rhône—Alpes Complex Systems Institute (IXXI) and Institut Convergence PLASCAN (ANR-17-CONV-0002). We also thank the BioSyL Federation and the LabEx Ecofect (ANR-11-LABX-0048) of the University of Lyon for inspiring scientific events. We warmly thank Gaël Yvert and Franck Picard for critical reading of the manuscript.

## References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
** of outstanding interest

1. * Albayrak C, Jordi CA, Zechner C, Lin J, Bichsel CA, Khammash M, Tay S: **Digital quantification of proteins and mRNA in single mammalian cells**. *Mol Cell* 2016, **61**:914–924.

A technical tour-de-force: counting mRNA and proteins from the same gene at the single cell level.

2. Anavy L, Levin M, Khair S, Nakanishi N, Fernandez-Valverde SL, Degnan BM, Yanai I: **BLIND ordering of large-scale transcriptomic developmental timecourses**. *Development* 2014, **141**:1161–1166.

3. Antos A, Kontoyiannis I: **Convergence properties of functional estimates for discrete distributions**. *Random Struct Algorithm* 2001, **19**:163–193.

4. Baccin C, Al-Sabah J, Velten L, Helbling PM, Grunschlager F, Hernandez-Malmierca P, Nombela-Arrieta C, Steinmetz LM, Trumpp A, Haas S: **Combined single-cell and spatial transcriptomics reveal the molecular, cellular and spatial bone marrow niche organization**. *Nat Cell Biol* 2020, **22**:38–48.

5. Balian R: *From microphysics to macrophysics*. Springer-Verlag; 1991.

6. Banerji CR, Miranda-Saavedra D, Severini S, Widschwendter M, Enver T, Zhou JX, Teschendorff AE: **Cellular network entropy as the energy potential in Waddington's differentiation landscape**. *Sci Rep* 2013, **3**.

7. Barron AR, Gyorfi L, van der Meulen EC: **Distribution estimation consistent in total variation and in two types of information divergence**. *IEEE Trans Inf Theor* 1992, **38**:1437–1454.

8. Bergen V, Lange M, Peidli S, Wolf FA, Theis FJ: **Generalizing RNA velocity to transient cell states through dynamical modeling**. *bioRxiv* 2019, 820936.

9. Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, Hill AJ, Zhang F, Mundlos S, Christiansen L, Steemers FJ, Trapnell C, Shendure J: **The single-cell transcriptional landscape of mammalian organogenesis**. *Nature* 2019, **566**:496–502.

10. Carcassi G, Aidala CA, Barbour J: *Variability as a better characterization of shannon entropy*. 2019.

11. Chan TE, Stumpf MPH, Babtie AC: **Gene regulatory network inference from single-cell data using multivariate information measures**. *Cell systems* 2017, **5**:251–267.

12. * Chang HH, Hemberg M, Barahona M, Ingber DE, Huang S: **Transcriptome-wide noise controls lineage choice in mammalian progenitor cells**. *Nature* 2008, **453**:544–547.

Demonstrates a functional connection between transcriptome fluctuations andcell decision making

13. Chao A, Shen T-J: **Nonparametric estimation of shannon's index of diversity when there are unseen species in sample**. *Environ Ecol Stat* 2003, **10**:429–443.

14. Choi K, Chen Y, Skelly DA, Churchill GA: **Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics**. *Genome Biol* 2020, **21**, 183.

15. Chung C: **On the origin of the typological/population distinction in ernst mayr's changing views of species, 1942-1959**. *Stud Hist Philos Sci C Stud Hist Philos Biol Biomed Sci* 2003, **34**: 277–296.

16. Cole MB, Risso D, Wagner A, DeTomaso D, Ngai J, Purdom E, Dudoit S, Yosef N: **Performance assessment and selection of normalization procedures for single-cell RNA-seq**. *Cell Syst* 2019, **8**:315–328 e8.

17. Cover T, Thomas J: *Elements of information theory*. 2nd ed. Wiley; 2006.

18. Deng Y, Luo H, Yang Z, Liu L: **Lncas2cancer: a comprehensive database for alternative splicing of lncrnas across human cancers**. *Brief Bioinform* 2020, **22**.

19. Doane DP: **Aesthetic frequency classifications**. *Am Statistician* 1976, **30**:181–183.

20. Domingues AF, Kulkarni R, Giotopoulos G, Gupta S, Vinnenberg L, Arede L, Foerner E, Khalili M, Adao RR, Johns A, Tan S, Zeka K, Huntly BJ, Prabakaran S, Pina C: **Loss of kat2a enhances transcriptional noise and depletes acute myeloid leukemia stem-like cells**. *Elife* 2020, **9**.

21. Dussiau C, Boussaroque A, Gaillard M, Bravetti C, Zaroili L, Knosp C, Friedrich C, Asquier P, Willems L, Quint L, Bouscary D, Fontenay M, Espinasse T, Plesa A, Sujobert P, Gandrillon O, Kosmider O: **Hematopoietic differentiation is characterized by a transient peak of cell-to-cell gene expression variability in normal and pathological conditions**. *bioRxiv* 2021.

22. Eling N, Morgan MD, Marioni JC: **Challenges in measuring and understanding biological noise**. *Nat Rev Genet* 2019, **20**: 536–548.

23. * Elowitz MB, Levine AJ, Siggia ED, Swain PS: **Stochastic gene expression in a single cell**. *Science* 2002, **297**:1183–1186.

Laid the foundation for the study of gene expression variability at the single cell level using a two reporter system in E. Coli

24. Fermi E: *Thermodynamics*. Dover; 1956.

25. Fuhrman S, Cunningham MJ, Wen X, Zweiger G, Seilhamer JJ, Somogyi R: **The application of Shannon entropy in the identification of putative drug targets**. *Biosystems* 2000, **55**:5–14.

26. Gao NP, Gandrillon O, Páldi A, Herbach U, Gunawan R: **Universality of cell differentiation trajectories revealed by a reconstruction of transcriptional uncertainty landscapes from single-cell transcriptomic data**. *bioRxiv* 2020. 04.23.056069, 2020.

27. Gong W, Rasmussen TL, Singh BN, Koyano-Nakagawa N, Pan W, Garry DJ: **Dpath software reveals hierarchical haemato-endothelial lineages of Etv2 progenitors based on single-cell transcriptome analysis**. *Nat Commun* 2017, **8**:14362.

28. * Grun D, Muraro MJ, Boisset JC, Wiebrands K, Lyubimova A, Dharmadhikari G, van den Born M, van Es J, Jansen E, Clevers H, de Koning EJ, van Oudenaarden A: **De novo prediction of stem cell identity using single-cell transcriptome data**. *Cell Stem Cell* 2016, **19**:266–277.

The first use of IntRAcellular entropy to characterize stemness at the single cell level.

29. Guillemin A, Duchesne R, Crauste F, Gonin-Giraud S, Gandrillon O: **Drugs modulating stochastic gene expression affect the erythroid differentiation process**. *PloS One* 2019, **14**, e0225166.

30. Guo M, Bao EL, Wagner M, Whitsett JA, Xu Y: **SLICE: determining cell differentiation and lineage based on single cell entropy**. *Nucleic Acids Res* 2017, **45**:e54.

31. Hausser J, Strimmer K: **Entropy inference and the james-stein estimator, with application to nonlinear gene association networks**. *J Mach Learn Res* 2009, **10**.

32. Hausser J, Strimmer K: **Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks**. *J Mach Learn Res* 2009, **10**:1469–1484.

33. Hoffmann M, Chang HH, Huang S, Ingber DE, Loeffler M, Galle J: **Noise-driven stem cell and progenitor population dynamics**. *PloS One* 2008, **3**, e2922.

34. Jenkinson G, Pujadas E, Goutsias J, Feinberg AP: **Potential energy landscapes identify the information-theoretic nature of the epigenome**. *Nat Genet* 2017, **49**:719–729.

35. Ko MS: **A stochastic model for gene induction**. *J Theor Biol* 1991, **153**:181–194.

36. Kumar U, Kumar V, Kapur JN: **Normalized measures of entropy**. *Int J Gen Syst* 1986, **12**:55–69.

37. Kupiec JJ: **A probabilistic theory for cell differentiation, embryonic mortality and DNA C-value paradox**. *Speculations Sci Technol* 1983, **6**:471–478.

38. Lake DE: **Nonparametric entropy estimation using kernel densities**. *Methods Enzymol* 2009, **467**:531–546.

39. Levsky JM, Singer RH: **Gene expression and the myth of the average cell**. *Trends Cell Biol* 2003, **13**:4–6.
**

The first demonstration of cell-to-cell variability in human cells.

40. Liu B, Li C, Li Z, Wang D, Ren X, Zhang Z: **An entropy-based metric for assessing the purity of single cell populations**. *Nat Commun* 2020, **11**:3155.

41. Liu J, Song Y, Lei J: **Single-cell entropy to quantify the cellular order parameter from single-cell RNA-seq data**. *Biophys Rev Lett* 2020, **15**:35–49.

42. Lombardi D, Pant S: **Nonparametric k-nearest-neighbor entropy estimator**. *Phys Rev* 2016, **93**, 013310.

43. Lord WM, Sun J, Bollt EM: **Geometric k-nearest neighbor estimation of entropy and mutual information**. *Chaos* 2018, **28**, 033114.

44. Luecken MD, Theis FJ: **Current best practices in single-cell RNA-seq analysis: a tutorial**. *Mol Syst Biol* 2019, **15**, e8746.

45. Lugosi G, Nobel A: **Consistency of data-driven histogram methods for density estimation and classification**. *Ann Stat* 1996, **24**:687–706.

46. MacArthur BD, Lemischka IR: **Statistical mechanics of pluripotency**. *Cell* 2013, **154**:484–489.

47. Mar JC: **The rise of the distributions: why non-normality is important for understanding the transcriptome and beyond**. *Biophys Rev* 2019:89–94.

48. Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM, Wold BJ: **From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing**. *Genome Res* 2014, **24**:496–510.

49. Martinez O, Reyes-Valdes MH: **Defining diversity, specialization, and gene specificity in transcriptomes through information theory**. *Proc Natl Acad Sci U S A* 2008, **105**:9709–9714.

50. Mikkers H, Frisen J: **Deconstructing stemness**. *EMBO J* 2005, **24**:2715–2719.

51. Miller G: **Note on the bias of information estimates**. *Inf Theory Psychol Probl Methods* 1955, **2**:95–100.

52. Mohammed H, Hernando-Herraez I, Savino A, Scialdone A, Macaulay I, Mulas C, Chandra T, Voet T, Dean W, Nichols J, Marioni JC, Reik W: **Single-cell landscape of transcriptional heterogeneity and cell fate decisions during mouse early gastrulation**. *Cell Rep* 2017, **20**:1215–1228.

53. Moris N, Edri S, Seyres D, Kulkarni R, Domingues AF, Balayo T, Frontini M, Pina C: **Histone acetyltransferase KAT2A stabilizes pluripotency with control of transcriptional heterogeneity**. *Stem Cell* 2018, **36**:1828–1838.

54. Moussy A, Cosette J, Parmentier R, da Silva C, Corre G, Richard A, Gandrillon O, Stockholm D, Paldi A: **Integrated time-lapse and single-cell transcription studies highlight the variable and dynamic nature of human hematopoietic cell fate commitment**. *PLoS Biol* 2017, **15**, e2001867.

55. Nicholson DJ: **Is the cell really a machine?** *J Theor Biol* 2019, **477**:108–126.

56. Paldi A: **Stochastic or deterministic? That is the question. *Organisms***. *J Biol Sci* 2020, **4**:77–79.

57. Paninski L: **Estimation of entropy and mutual information**. *Neural Comput* 2003, **15**:1191–1253.

58. Piera FJ, Parada P: **On convergence properties of shannon entropy**. *Probl Inf Transm* 2009, **45**:75–94.

59. Piras V, Tomita M, Selvarajoo K: **Transcriptome-wide variability in single embryonic development cells**. *Sci Rep* 2014, **4**.

60. Purwani S, Nahar J, Twining C: **Analyzing bin-width effect on the computed entropy**. *AIP Conf Proc* 2017, **1868**, 040008.

61. Richard A, Boullu L, Herbach U, Bonnafoux A, Morin V, Vallin E, Guillemin A, Papili Gao N, Gunawan R, Cosette J, Arnaud O, Kupiec JJ, Espinasse T, Gonin-Giraud S, Gandrillon O: **Single-cell-based analysis highlights a surge in cell-to-cell molecular variability preceding irreversible commitment in a differentiation process**. *PLoS Biol* 2016, **14**, e1002585.
**

The first demonstration of a non-monotonous behavior of IntERcellular entropy during a differentiation sequence.

62. Ridden SJ, Chang HH, Zygalakis KC, MacArthur BD: **Entropy, ergodicity, and stem cell multipotency**. *Phys Rev Lett* 2015, **115**:208103.

63. Ritchie W, Granjeaud S, Puthier D, Gautheret D: **Entropy measures quantify global splicing disorders in cancer**. *PLoS Comput Biol* 2008, **4**, e1000011.

64. Rodriques SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR, Welch J, Chen LM, Chen F, Macosko EZ: **Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution**. *Science* 2019, **363**:1463–1467.

65. Safdari H, Kalirad A, Picioreanu C, Tusserkani R, Goliaei B, Sadeghi M: **Noise-driven cell differentiation and the emergence of spatiotemporal patterns**. *PloS One* 2020, **15**, e0232060.

66. Sarkar A, Stephens M: **Separating measurement and expression models clarifies confusion in single cell RNA-seq analysis**. *bioRxiv* 2020.
**

A very thoughtful and relevant analysis of the statistical nature of scRNAseq data

67. Scargle JD, Norris JP, Jackson B, Chiang J: **Studies in astronomical time series analysis. VI. Bayesian block representations**. *Astrophys J* feb 2013, **764**:167.

68. Schiebinger G, Shu J, Tabaka M, Cleary B, Subramanian V, Solomon A, Gould J, Liu S, Lin S, Berube P, Lee L, Chen J, Brumbaugh J, Rigollet P, Hochedlinger K, Jaenisch R, Regev A, Lander ES: **Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming**. *Cell* 2019, **176**:928–943 e22.

69. Schrödinger E: *What is life? The physical aspect of the living cell.* Cambridge University Press; 1944.

70. Setty M, Kiseliovas V, Levine J, Gayoso A, Mazutis L, Pe'er D: **Characterization of cell fate probabilities in single-cell data with Palantir**. *Nat Biotechnol* 2019, **37**:451–460.

71. Shahrezaei V, Swain PS: **Analytical distributions for stochastic gene expression**. *Proc Natl Acad Sci U S A* 2008, **105**:17256–17261.

72. Shannon CE: **A mathematical theory of communication**. *Bell Syst Tech J* 1948, **27**:379–423.

73. Shao X, Lu X, Liao J, Chen H, Fan X: **New avenues for systematically inferring cell-cell communication: through single-cell transcriptomics data**. *Protein Cell* 2020, **11**:866–880.

74. Shi J, Teschendorff AE, Chen W, Chen L, Li T: **Quantifying Waddington's epigenetic landscape: a comparison of single-cell potency measures**. *Brief Bioinform* 2018. 20180040.

75. Silva JF, Parada P: **On the convergence of shannon differential entropy, and its connections with density and entropy estimation**. *J Stat Plann Inference* 2012, **142**:1716–1732.

76. Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, Purdom E, Dudoit S: **Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics**. *BMC Genom* 2018, **19**:477.

77. Strell C, Hilscher MM, Laxman N, Svedlund J, Wu C, Yokota C, Nilsson M: **Placing RNA in context and space - methods for spatially resolved transcriptomics**. *FEBS J* 2019, **286**:1468–1481.

78. Stumpf PS, Smith RCG, Lenz M, Schuppert A, Müller F-J, Babtie A, Chan TE, Stumpf MPH, Please CP, Howison SD, Arai F, MacArthur BD: **Stem cell differentiation as a non-markov stochastic process**. *Cell Syst* 2017, **5**:268–282.
**

A statistical mechanics view of a differentiation process

79. Symmons O, Raj A: **What's luck got to do with it: single cells, multiple fates, and biological nondeterminism**. *Mol Cell* 2016, **62**:788–802.

80. Teschendorff AE, Enver T: **Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome**. *Nat Commun* 2017, **8**:15599.

81. Teschendorff Andrew E, Feinberg Andrew P: **Statistical mechanics meets single-cell biology**. *Nat Rev Genet* 2021.

82. Triqueneaux G, Burny C, Symmons O, Janczarski S, Gruffat H, Yvert G: **Cell-to-cell expression dispersion of B-cell surface proteins displays genetic variation among humans**. *Communications Biology* 2020, **3**:346.

83. Ventre E, Espinasse T, Brehier C-E, Calvez V, Lepoutre T, Gandrillon O: **Reduction of a stochastic model of gene expression: Lagrangian dynamics gives acces to basins of attraction as cell types and metastability**. *bioRxiv* 2020. 09.04.283176, 2020.

84. Villani M, Barbieri A, Serra R: **A dynamical model of genetic networks for cell differentiation**. *PloS One* 2011, **6**, e17703.

85. Wallis K: *A note on the calculation of entropy from histograms*. MPRA Paper; 2006.

86. Wang S, Drummond ML, Guerrero-Juarez CF, Tarapore E, MacLean AL, Stabell AR, Wu SC, Gutierrez G, That BT, Benavente CA, Nie Q, Atwood SX: **Single cell transcriptomics of human epidermis identifies basal stem cell transition states**. *Nat Commun* 2020, **11**:4239.

87. Wang S, Karikomi M, MacLean AL, Nie Q: **Cell lineage and communication network inference via optimization for single-cell transcriptomics**. *Nucleic Acids Res* 2019, **47**: e66.

88. Wang S, Tu J, Wang L, Lu Z: **Entropy-based model for miRNA isoform analysis**. *PloS One* 2015, **10**, e0118856.

89. Wiesner K, Teles J, Hartnor M, Peterson C: **Haematopoietic stem cells: entropic landscapes of differentiation**. *Interface Focus* 2018, **8**:20180040.

90. Xing QR, Cipta NO, Hamashima K, Liou YC, Koh CG, Loh YH: **Unraveling heterogeneity in transcriptome and its regulation through single-cell multi-omics technologies**. *Front Genet* 2020, **11**:662.

91. Ye Y, Yang Z, Zhu M, Lei J: **Using single-cell entropy to describe the dynamics of reprogramming and differentiation of induced pluripotent stem cells**. *Int J Modern Phys B* 2020, **34**, 2050288.

92. Zheng X, Jin S, Nie Q, Zou X: **scRCMF: identification of cell subpopulations and transition states from single-cell transcriptomes**. *IEEE Trans Biomed Eng* 2020, **67**: 1418–1428.

93. Zipori D: **The nature of stem cells: state rather than entity**. *Nat Rev Genet* 2004, **5**:873–878.