Computational and Structural Biotechnology Journal 18 (2020) 1830–1837

ELSEVIER

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Short review

Application of information theoretical approaches to assess diversity and similarity in single-cell transcriptomics

Michal T. Seweryn [a], Maciej Pietrzak [b,*], Qin Ma [b,*]

[a] Center for Medical Genomics, Jagiellonian University, Cracow, Poland
[b] Department of Biomedical Informatics, The Ohio State University, Columbus OH, United States

ARTICLE INFO

Article history:
Received 11 December 2019
Received in revised form 24 April 2020
Accepted 6 May 2020
Available online 21 May 2020

Keywords:
Single cell sequencing
Diversity and similarity assessment
Renyi's entropy
Renyi's divergence

ABSTRACT

Single-cell transcriptomics offers a powerful way to reveal the heterogeneity of individual cells. To date, many information theoretical approaches have been proposed to assess diversity and similarity, and characterize the latent heterogeneity in transcriptome data. Diversity implies gene expression variations and can facilitate the identification of signature genes; while, similarity unravels co-expression patterns for cell type clustering. In this review, we summarized 16 measures of information theory used for evaluating diversity and similarity in single-cell transcriptomic data, provide references and shed light on selected theoretical properties when there is a need to select proper measurements in general cases. We further provide an R package assembling discussed approaches to improve the researchers own single-cell transcriptome study. At last, we prospected further applications of diversity and similarity measures in support of depicting heterogeneity in single-cell multi-omics data.

1. Introduction

The single-cell transcriptomic data is able to elucidate the heterogeneous and state-dependent gene expression patterns. This pattern likely reflects both the unique functions of individual cells as well as the current and future trajectory of specialization/differ entiation [37]. With the rapid development of single-cell sequencing techniques, the gene expression profiles of hundreds to thousands single cells can be measured simultaneously [19]. The high-throughput single-cell transcriptomic experiments provide the opportunity to study the set of RNA transcripts relevant to current state of each cell. More than 50,000 transcriptomic profiles of cancer tissue samples have been collected since 2002. The gene expression values measured on a tissue sample is the ensemble expression of all its comprising cells, which are usually highly heterogeneous. Under such circumstances, the ensemble expression may not represent the expression pattern of any individual cells, thus the cell–cell difference within a tissue must be treated with a great caution. The scRNA-Seq technology emerged to unmask individual cellular properties, and scRNA-Seq data are ideal objects to build reference maps of single cell behaviors. What is more, with the use of certain machine learning techniques (such as clustering) one may extract relevant subsets of cells together with gene expression profiles which are used to define these subsets. Such approach allow to uncover the differences in the diversity of gene expression profiles of different cells, as well as the similarities between RNA expression profiles of cells in distinct clusters (functional subsets). This is particularly important when studying gene expression on a single cell level in highly heterogeneous tissues and/or in non-mendelian disorders [24]. For example, the highly heterogeneous expression patterns of cancer cells define the intra-tumor sub-populations with diverse drug-resistance levels and in turn result in different curative effects causing relapses after specific treatments [34]. Accurate identification of cell types based on the gene expression profiles is a necessary but pivotal step in analyzing single-cell transcriptome data. However, the high data sparsity and complexity leave a challenge towards the accurate measurement of the heterogeneity and similarity among cells and cell types [32,44].

Two concepts are commonly used to depict the heterogeneity in single-cell transcriptome data analyses, i.e. *diversity* and *similarity*. The *diversity* describes the variations of gene expression profiles among cells, and lays a foundation for feature extraction and selection in support of novel biological insights derivation [25]. Meanwhile, the analysis of *similarity* allows to group cells with relatively similar features, and enables the identification of cell types/states which also refers to the cell clustering [33].

* Corresponding authors.
E-mail addresses: michal.seweryn@wmii.uni.lodz.pl (M.T. Seweryn), pietrzak.20@osu.edu (M. Pietrzak), Qin.Ma@osumc.edu (Q. Ma).

https://doi.org/10.1016/j.csbj.2020.05.005
2001-0370/

Many algorithms and tools utilize the concepts of diversity and similarity in single-cell transcriptome data analysis. Juan et al. developed a novel biclustering method to separate regulatory signals and extract gene features by identifying the diversity among local-low-rank gene modules [45,43]. Kim *et al.* applied five common similarity measurements, including Euclidean, Manhattan distance, maximum distances, Pearson's correlation, and Spearman's correlation coefficients to measure the *diversity* and *similarity* for cell type prediction from single-cell transcriptomic data [22]. Results showed that the choice of similarity metric affects clustering performance, thereby leading to significant differences in cell-type identification. Moreover, the concept of entropy (which is associated with the uncertainty of a complex system) has been extensively used to evaluate the diversity of expression profiles among cells and lead to the identification of distinct cell states [40,16]. According to that, Guo et al. and Liu et al. used the single-cell entropy concept and proposed SLICE and scEGMM [16], respectively, quantifying the differentiation state of a given cell in an unbiased way, where the direction of the transition was correctly estimated form a cross-sectional data without sequential information. Moreover, Suo *et al.* evaluated the 'activity entropy' of co-regulated gene modules identified from single-cell transcriptomic data using the Jensen-Shannon Divergence, and unraveled the heterogeneous regulatory network [38]. Our attention is mostly focused around diversity and similarity indices that originate from information theory. We do not discuss the neither the multidimensional distance measures [39,5], the high-dimensional [14,10] or directional [4] dependency concepts which are applied in the analysis of single-cell transcriptomes. In this work, we review the methods for assessment of the diversity and similarity of transcription profiles in single cell systems. At the end we provide a R package that will allow the readers to test presented measures on their data.

Throughout the article, we consider the contingency table model, in which data (preferably gene counts) are arranged into a two-way $(m \times n)$ table $[c_{ij}]$, with columns representing n different cells and rows representing m genes that are potentially expressed in any of these cells. Often, one may want to refer to a given cell as being of 'a particular type', therefore under the term 'type of a cell' we understand a vector (profile) of relative gene expression (for some well defied subset of all genes - usually refered to as markers). In statistical terms, we consider n independent multinomial distributions $\boldsymbol{p}_1 = (c_{11}/\sum c_{i,1}, \ldots, c_{m,1}/\sum c_{i,1}), \ldots$, $\boldsymbol{p}_n = (c_{1,n}/\sum c_{i,n}, \ldots, c_{m,n}/\sum c_{in})$. We denote by $\boldsymbol{u}_m$ the vector of uniform probabilities on the set $\{1, 2, \ldots, m\}$and by $\Delta_{m-1}$the probability simplex in $\mathbb{R}^m_{\geqslant 0}$.

1.1. Diversity measures

The term 'diversity' is one of the key concepts in many fields of modern biological sciences (e.g. ecology, genetics). By diversity, one typically understands the abundance of elements of a given population. There are two main concepts associated with diversity and its measurement – that is richness and evenness.It is worth noting that here, we use evenness (in short) for the proportional abundance of species, and not for the value of the diversity index relative to its maximum. Richness translates into the number of genes that are expressed in a cell of a given type, whereas the second one corresponds to the relative abundances of gene expression profiles in a cell of a given type. Formally, consider a set of $m < \infty$ genes (RNAs) and a population $\boldsymbol{c} = (c_1, \ldots, c_m) \in \mathbb{N}^m_{\geqslant 0}$. We define diversity as.

Definition 1. For a given cell, i.e. a population $\boldsymbol{c} = (c_1, \ldots, c_m)$ of m genes, its richness is defined as the (often unknown) number $m_0 := |\{c_i : c_i \neq 0\}|$ its evenness is defined as $\boldsymbol{p} = (p_1, \ldots, p_m)$, $p_i := \frac{c_i}{\sum c_k}$ and its *diversity* or *fingerprint* is the vector $\mathcal{F}_{\boldsymbol{c}} = (v_1, \ldots, v_{\max_i c_i})$ where $v_k = |\{i : c_i = k\}|$. Any nonnegative, real function with values $D(\mathcal{F}_{\boldsymbol{c}}) \in \mathbb{R}_{\geqslant 0}$ is called a *measure of diversity* or an *index of diversity*.

For convenience we define the function D on the set of all nonnegative infinite sequences of natural numbers. It is also common to impose the following set of conditions on the diversity index. We will present these conditions as axioms.

Axiom 1. We shall say that a given diversity index D is:

- continuous if the multivariate function is continuous in each of its coordinate variables
- symmetric if D is invariant to any permutation of its variables
- maximal on uniform if (for a set number of genes m) D is maximized by the vector $\boldsymbol{u}_m$

Aside form these most common properties defined above, there are several which are of additional interest. One of these is the monotonicity of the diversity measure. Due to the multivariate character of the diversity measurement, there are several ways in which monotonicity may be defined. Below we present two such approaches.

Remark 1. Let $\mathbb{1}_{\mathbb{m}} = (\mathbb{1}, \ldots, \mathbb{1}) \in \mathbb{N}^{\mathbb{m}}_{\geqslant \mathbb{0}}$ and denote by $\mathcal{F}_{\mathbb{1}_{\mathbb{m}}}$ a vector with $v_1 = m$ and $v_i = 0$ for $i > 1$. We say that diversity index D is monotone on uniform profiles if $D(\mathcal{F}_{\mathbb{1}_{\mathbb{m}}})$ is nondecreasing in m.

Let $m > 0$, $\boldsymbol{p} = (\boldsymbol{p_1}, \ldots, \boldsymbol{p_m})$ be a normalized gene expression vector, $0 < i \leqslant m$ be a set gene. We say that the diversity index D is nondecreasing with respect to the transfer of mass (total amount of probability) from gene i (to a 'new' gene) if for any $0 < \epsilon < p_i$ $D(\boldsymbol{p}) \leqslant \boldsymbol{D}(\boldsymbol{q})$, where $\boldsymbol{q} = (p_1, \ldots, p_{i-1}, p_i - \epsilon, p_{i+1}, \ldots, p_m, \epsilon)$

As stated above, the first definition requires only that the diversity index is monotone with respect to the number of genes, given that all genes are equally abundant in a set of cells. At the same time, in the second condition one demands that the transfer of abundance (gene expression) from one existing gene to a new one always increases diversity. The detailed discussion of the mutual relation between the above conditions goes far beyond the scope of this short note.

As far as the transfer of abundance is concerned, there is one more property which is often imposed on the diversity indices. In this condition, one demands that the diversity increases as the probability mass is transferred from a gene with higher expression to a gene with lower expression.

Remark 2. Let $m > 0$, $\boldsymbol{p} = (p_1, \ldots, p_m)$ be a normalized gene expression vector and $0 < i, j \leqslant m$ be two genes, such that $p_i < p_j$. We say that the diversity index D is nondecreasing with respect to the transfer of mass from gene i to gene j if for any $0 < \epsilon < p_j - p_i$ we have $D(\boldsymbol{p}) \leqslant \boldsymbol{D}(\boldsymbol{q})$, where $\boldsymbol{q} = (p_1, \ldots, p_{i-1}, p_i + \epsilon, p_{i+1}, \ldots, p_{j-1}, p_j - \epsilon, p_{j+1}, \ldots, p_m)$.

This condition is different in its nature form the monotonicity with respect to transfer of mass to a new gene and is related to the well-known mathematical property of order-preserving functions (i.e. Schur-cocave functions).

1.2. Diversity Measures – examples

To date, there is a variety indices defined that are used in ecological, genetical and molecular studies. One of the most obvious ones is the richness - i.e. the number of genes expressed in a given cell. As simple as it seems to be, it still remains one of the most difficult to estimate in practical studies when the theoretical richness of the studied system is much larger than the sample size that is feasible to achieve. Recently, successful attempts have been made to estimate diversity measures and richness via species accumulation curves [8]. Yet, these general and elegant results do not provide 'optimal' (most suitable under any sampling model) methods for estimating diversity and/or richness. Additionally, it is worth noting that in gene-level studies the population abundances might be heavy-tailed and therefore the standard asymptotic theory for estimators of diversity might not apply (see [2,30]). In molecular studies, the difficulties in learning the missing mass may likely be best highlighted by the analysis of richness of T- and B-cell antigen receptor populations [3,13]. At the same time, the exact number of genes that are transcribed in the genome is not known - esp. with the variety of different classes of non-coding RNAs, some of which are highly unstable like eRNAs. We will discuss briefly the aspect of estimation of the missing mass in the subsequent section.

Below, we recall some of the best known and widely applied diversity indices with a short description of their properties. Contrary to the richness estimator, the Berger-Parker index takes only into account the relative expression level of the most abundant gene in a given cell, it is defined as:

$$D_{BP}(\boldsymbol{p}) := \frac{\boldsymbol{1}}{\max_{\boldsymbol{i}} \boldsymbol{p_i}}.$$

Although in many applications it is sufficient to use this index – and sometimes desired as it is robust to rare events – it has a rather pathological property being completely 'blind' to transfer of mass between non-abundant genes. The Simpson diversity index is defined as the probability that when taking a simple random sample form a given population of genes, twice we select the same entity, i.e. it holds that

$$D_{SI}(\boldsymbol{p}) := \sum_{\boldsymbol{i}} \boldsymbol{p_i^2}.$$

Even though the definition of this index is quite intuitive, the interpretation may be difficult – as high values of D_{SI} reflect low diversity. For this reason two other modifications of this index are more commonly used:

$$D_{GS}(\boldsymbol{p}) := \boldsymbol{1} - \boldsymbol{D_{SI}}(\boldsymbol{p})$$

and

$$D_{ISI}(\boldsymbol{p}) := \frac{\boldsymbol{1}}{\boldsymbol{D_{SI}}(\boldsymbol{p})},$$

where the first one is often referred to sa the Gini-Simpson index and the latter as the Inverse Simpson Index. Yet, these measures turn to be insufficiently sensitive to rare 'species'.

Probably the best known examples of diversity indices are rooted in the information theory, see e.g. Tóthmérész [41], Ricotta [36], Keylock [21] – in particular, the Shannon entropy function

$$H_1(\boldsymbol{p}) = -\sum \boldsymbol{p_i} \log(\boldsymbol{p_i}).$$

By the Jensen inequality we have that $H_1(\boldsymbol{p}) \leqslant \log(\boldsymbol{m}) = \boldsymbol{H_1}(\boldsymbol{u_m}) = \boldsymbol{H_1}(\mathcal{F}_{\mathbb{1}_{\mathbb{m}}})$ and that H_1 is monotone. The H_1 has numerous appealing properties – aside form the ones discussed of note is the additivity of entropy on product distributions and the chain rule for entropy, which states that the joint entropy of gene expression profiles in two cells equals the sum of the marginal entropy in one cell and the conditional entropy of gene expression in the second cell given the expression profile in the first cell. This particular property is not shared by a family of entropies defined by A. Renyi:

Example 1 (*Renyi Entropy*). The Renyi entropy of order $\alpha \in [0, \infty]$ is given by

$$H_\alpha(\boldsymbol{p}) = \frac{\boldsymbol{1}}{\boldsymbol{1} - \alpha} \log \left(\sum \boldsymbol{p_i^\alpha} \right) \tag{1}$$

for $\boldsymbol{p} \in \Delta_{\boldsymbol{m-1}}$, with the limiting cases of interest $H_0(\boldsymbol{p}) = \log(\boldsymbol{m})$, $H_1(\boldsymbol{p}) = -\sum \boldsymbol{p_i} \log \boldsymbol{p_i}$and $H_\infty(\boldsymbol{p}) = -\log(\max_{\boldsymbol{i}} \boldsymbol{p_i})$.The case $\alpha = 2$ in the above example is sometimes known as the *Rao quadratic entropy* [31] with the obvious relation to the Simpson index. It is worth noting that the Renyi entropy of order $\alpha < 1$ puts more weight on genes with low expression levels and the Renyi entropy of order $\alpha > 1$ puts more weight on the highly expressed ones [35]. There exist another generalization of the Shannon entropy, which is associated with non-extensive statistical mechanics and is called Tsallis family of entropies [42]. One of the most important property of these diversity indices is that the Tsallis entropy of order $\alpha \neq 1$ is non-additive on product distributions. The formal definition goes as follows.

Example 2 (*Tsallis Entropy*). The Tsallis entropy of order $\alpha \in [0, \infty]$ is given by

$$T_\alpha(\boldsymbol{p}) = \frac{\boldsymbol{1}}{\alpha - \boldsymbol{1}} \left(\boldsymbol{1} - \sum \boldsymbol{p_i^\alpha} \right) \tag{2}$$

for $\boldsymbol{p} \in \Delta_{m-1}$. For these two parametric families of entropy functions, it was suggested that the analysis of diversity should be performed as a function of the parameter α. Such approach has been applied in various settings and is referred to as the analysis of diversity profiles (see Tóthmérész [41]).All diversity indices defined above have values on completely different scales, therefore it is challenging to compare the diversities of cells measured with the use of different indices. This may be done via the concept of the *effective number of species* (ENS) Jost [20]. Let $\boldsymbol{c}$ be an arbitrary population and D be a diversity measure monotone with respect to the uniform distribution. For any y of the form $y = m + \alpha$ $(0 \leqslant \alpha \leqslant 1)$ define $D(\mathcal{F}_{\mathbb{1}_{\mathbb{m}+\alpha}}) := (1 - \alpha)D(\mathcal{F}_m) + \alpha D(\mathcal{F}_{m+1})$. The ENS based on D is the smallest solution $y = y_0$ of the equation

$$D(\mathcal{F}_{\mathbb{1}_{\mathbb{y}}}) = D(\mathcal{F}_c).$$

Note that the ENS is typically less than the species number m and may be noninteger. The ENS for the class of Renyi entropies are called Hill numbers – i.e. we have.

Example 3. Set $D = H_\alpha$ and a population $\boldsymbol{p} \in \mathbb{N}_{\geqslant \boldsymbol{0}}$. The number $m_\alpha = (\sum p_i^\alpha)^{\frac{1}{1-\alpha}}$ for integer m is the ENS for this diversity index. Note that for non-integer values of m we define ENS by the linear interpolation. Note that the majority of diversity indices considered do share a number of important properties, but at the same time are highly non-linear functions, and thus one encounters certain challenges when considering for instance the analysis of bias for the so-called plug-in estimators. At the same time, since for the single-cell experiments we are in the small n large p regimen, it is unclear how does the undersampling bias affects the naive estimators. We shall deal with these issues in the following section.

1.3. Diversity Measures – estimators

In the present section, due to economy of space, we only consider the problem of estimating the richness and the generalized family of Renyi diversity indices. One of the best known estimators of population total is the Horvitz-Thompson estimator, which is defined as follows (see [27]):

Example 4. Let $\boldsymbol{c}$ be a population of genes (i.e. a cell of a given type) and assume that a sample gene expression profile S is generated based on the scRNA-seq experiment with sequencing depth N. Then the Horvitz-Thompson estimator of the population total is given by:

$$M_{HT}(\boldsymbol{c}, S) := \sum_{i=1}^{M} \delta_i(S)\pi_i(S), \tag{3}$$

where $\delta_i(S)$ is the indicator of the event that the i-th gene was observed at least once in our sample profile, and $\pi_i(S) := 1 - (1 - p_i)^N$ is the inclusion probability of a gene i in a profile of size N. In practical applications the π_i is unknown and is substituted with its nonparametric estimator.

For the single-cell transcriptome profiling, the issue of undersampling bias may arises naturally due to huge diversity of gene abundances per cell as well as the limitations on data collection (cost of sequencing at very high depth). This problem of estimating the missing mass is not new and has been previously studied for example in the context of TCR sequencing. The core concept in analyzing the missing mass is the *sample coverage*. We describe it in detail and first recall the following

Definition 2 (*Coverage*). Let $\boldsymbol{X} = (X_1, \dots, X_m)$ denote a multinomial random variable $Mult(n, \boldsymbol{p})$ and set $I_i = 1$ if $X_i > 0$ and $I_i = 0$ otherwise. The $\boldsymbol{X}$-based *sample coverage* is given by

$$C = \sum p_i \delta_i(\boldsymbol{X}).$$

The sample coverage is a random variable and may be interpreted as the posterior probability of discovering a new multinomial class in the next sample given the information in the current sample. For that reason it is used in molecular and ecological studies to estimate the probability of discovering a new "species" in a population [29]. The sample coverage is not available without knowing the normalized abundance vectors a priori, yet, the following empirical estimate, known as the Good-Turing coverage estimator, [15] proves to be an excellent approximation. The Good-Turng *empirical sample coverage* is given by

$$\widehat{C} = 1 - \frac{f_1}{n}, \tag{4}$$

where the symbol f_i denotes the number of genes (species) in $\boldsymbol{X}$ observed exactly i times, so that $\sum f_i = n$. The issue of estimation of the sample coverage and the missing mass remains an active area of research. The properties (consistency and normality) of the Good-Turing estimator were originally studied by [11,12]. More recently, a necessary and sufficient condition for the asymptotic normality of $\widehat{C}$ was given by Zhang et al. [46]. At the same time, sufficient and necessary conditions for the possibility of learning the missing mass via any empirical estimator have been recently given [30]. These conditions are associated with the tails of the distribution to be estimated and based on the regularly varying functions. There are several ideas on how to apply the coverage adjustment and Horvitz-Thompson corrections to reduce the under-sampling bias in entropy estimation (see [9]). In particular, the estimators constructed via adjusting the exponent of the Renyi and/or Tsallis entropies [35] are seen to put more weight on the less frequent species and hence, intuitively at least, should be reduce the under-sampling bias. Moreover, as long as the sample coverage converges to unity, the adjusted estimates are consistent under mild regularity conditions. It is worth noting that the alternative approach which uses species accumulation curves has been proposed by Chao and Jost [8]. This method provides stable results and allows to reduce the bias of both richness and diversity profile estimators. We indicate the need for comprehensive (both theoretical and simulation-based) comparison of this approach with other (more standard methods) in regimens when the number of genes is much higher than the sample size. Yet such comparative studies go beyond the scope of the current short note.

Definition 3. Under the above notation, define the following coverage and Horvitz-Thompson adjusted Renyi/Tsallis diversity index by

$$H_\alpha^{(n)}(\boldsymbol{p}) = \frac{1}{1-\alpha} \log \left(\sum \frac{p_i^\alpha}{1 - (1 - p_i)^n} \right)$$

and

$$T_\alpha^{(n)}(\boldsymbol{p}) = \frac{1}{\alpha - 1} \left(1 - \sum \frac{p_i^\alpha}{1 - (1 - p_i)^n} \right).$$

Moreover define the coverage adjusted estimators of Renyi/Tsallis entropies of order α by

$$H_{\alpha C}(\boldsymbol{p}) = \frac{1}{1 - \alpha C} \log \left(\sum p_i^{\alpha C} \right) \quad \text{and}$$

$$T_{\alpha C}(\boldsymbol{p}) = \frac{1}{\alpha C - 1} \left(1 - \sum p_i^{\alpha C} \right).$$

The Horvitz-Thompson estimator is probably the best known example of the estimator of population total in survey sampling. Yet, there seems to be a better alternative with superior performance under mild conditions, namely the Hajek estimator (see [17]). It has a general form of a weighted average, where the weights are the inverse inclusion probabilities. To best our knowledge, such adjustments, based on the Hajek estimator, to general entropy functions have not yet been made. Therefore we propose the following example.

Example 5. Under the above notation, define the following coverage and Hajek adjusted Renyi/Tsallis diversity index by

$$H_\alpha^{(hn)}(\boldsymbol{p}) = \frac{1}{1-\alpha} \log \left(\frac{1}{\sum (1 - (1 - p_i)^n)} \sum \frac{p_i^\alpha}{1 - (1 - p_i)^n} \right)$$

and

$$T_\alpha^{(hn)}(\boldsymbol{p}) = \frac{1}{\alpha - 1} \left(1 - \frac{1}{\sum (1 - (1 - p_i)^n)} \sum \frac{p_i^\alpha}{1 - (1 - p_i)^n} \right).$$

It is of note, that under additional regularity conditions one may define a diversity index which accounts for the unseen probability mass explicitly and is directly related to the coverage-adjusted Tsallis entropy index. Below, we present a heuristic reasoning how such estimators may be constructed.

Remark 3. Let us consider a triangular array of probabilities (p_{ni}) such that $\log n \sum_i p_{ni}(1 - p_{ni})^{n-1} \to 0$ and denote by W_n a random variable such that $P\left(W_n = \log \frac{1}{p_i}\right) = p_i$. Assume that an i.i.d. sample $\mathcal{S}_n$ of size k_n is drawn from the distribution $\left(p_{n1}, \dots, p_{nm(n)}\right)$. By the Taylor expansion (heuristically) we have

$$\frac{\sum_{i\in\mathcal{S}_n}\boldsymbol{p}_i^{\widehat{EC}}-1}{1-E\widehat{C}}=\frac{\sum_{i\in\mathcal{S}_n}\boldsymbol{p}_i\left(\boldsymbol{p}_i^{\widehat{EC}-1}-1\right)}{1-E\widehat{C}}$$
$$=\sum_{i\in\mathcal{S}_n}\boldsymbol{p}_i\log\frac{1}{\boldsymbol{p}_i}+\frac{1-E\widehat{C}}{2}\sum_{i\in\mathcal{S}_n}\boldsymbol{p}_i\log^2\frac{1}{\boldsymbol{p}_i}+R,$$

where $R=o\left(\frac{1-E\widehat{C}}{2}\sum_{i\in\mathcal{S}_n}\boldsymbol{p}_i\log^2\frac{1}{\boldsymbol{p}_i}\right)$. Therefore we have

$$\frac{\sum_{i\in\mathcal{S}_n}\boldsymbol{p}_i^{\widehat{EC}}-1}{1-E\widehat{C}}-\sum_i p_i\log\frac{1}{p_i}=\sum_{i\in\mathcal{S}_n}\boldsymbol{p}_i\log\frac{1}{\boldsymbol{p}_i}-\sum i\in\mathcal{S}_n p_i\log\frac{1}{p_i}$$
$$+\frac{1-E\widehat{C}}{2}\sum_{i\in\mathcal{S}_n}\boldsymbol{p}_i\log^2\frac{1}{\boldsymbol{p}_i}-\sum_{i\notin\mathcal{S}_n}p_i\log\frac{1}{p_i}+R.$$

It is natural to treat the term $\sum_{i\in\mathcal{S}_n}\boldsymbol{p}_i\log\frac{1}{\boldsymbol{p}_i}$ as a good estimator of the 'observed part' of the Shannon entropy – which is $\sum_{i\in\mathcal{S}_n}p_i\log\frac{1}{p_i}$. By the above equation we note that in the coverage adjusted Tsallis entropy an additional term – which is $\frac{1-E\widehat{C}}{2}\sum_{i\in\mathcal{S}_n}\boldsymbol{p}_i\log^2\frac{1}{\boldsymbol{p}_i}$ – is added to correct for the 'unobserved part' of the Shannon entropy $\sum_{i\notin\mathcal{S}_n}p_i\log\frac{1}{p_i}$. Therefore the question arises on the goodness of such adjustment for 'unseen part' of the entropy. Let us note that $1-E\widehat{C}$ is the estimator of the unseen probability mass (in the Good-Turing sense). Therefore the correction term in the adjusted entropy estimator is seen to be equal to the probability of the unseen part of the distrubution times the estimator of the second moment of the random variable W_n on the observed part of the ditribution. Now, the question is whether the approximation

$$\frac{\sum_{i\notin\mathcal{S}_n}p_i\log\frac{1}{p_i}}{1-E\widehat{C}}\approx\sum_{i\in\mathcal{S}_n}\boldsymbol{p}_i\log^2\frac{1}{\boldsymbol{p}_i}.$$

that is associated with the coverage adjusted Tsallis entropy may be corrected for certain types of distributions, and in turn be a better estimator of the Shannon entropy may be proposed. Let us look at the l.h.s. of the above equation as the expectation of the random variable W_n conditionally on the unobserved probability mass. With this interpretation in mind we note the relation to the mean excess function of a random variable – here, we are assuming that the unobserved part of the distribution consists mostly of rare events. The mean excess function is widely used in actuarial mathematics in the context of power-law distributions. Thus we assume that the random variable W_n has a distributon such that the sequence of probabilities $(p_i)_{k\in\{1,\dots,m(n)\}}$ is a regularly varying (in Karamata's sense) sequence. Under such assumption we note that if $(p_i)_{k\in\{1,\dots,m(n)\}}$ is regularly varying with index smaller then -1 then, by the Karamata theorem, for sufficiently lagre $\delta>0$, we have

$$\frac{\sum_{i:\log\frac{1}{p_i}>\delta}p_i\log\frac{1}{p_i}}{P\left(W_n^{(1)}>\log\frac{1}{\delta}\right)}\approx\delta.$$

It is now clear, that in such case one may improve upon the unseen part of the entropy by setting

$$\widehat{H}^{PL}=\sum_{i\in\mathcal{S}_n}\boldsymbol{p}_i\log\frac{1}{\boldsymbol{p}_i}+\left(1-\widehat{C}\right)\log n.$$

Moreover, in the case when $(p_i)_{k\in\{1,\dots,m(n)\}}$ is regularly varying with index -1, which means that $p_i:=\frac{l_n(k)}{k}$ where $l_n(k)$ is some slowly varying function, and such that $\sup_n\sum_i p_i\log\frac{1}{p_i}<\infty$ from more detailed analysis based upon Karamata's theorem one may conclude that the slower the decay of the function l_n, the better the performance of the adjusted Tsallis entropy estimator is.

The analysis of diversity allows for comparison of certain summary statistics of the expression profile, however, in order to compare expression profiles of two or more cells (and in turn classify the cells into groups/clusters) one takes a different approach, based on the idea of similarity measures. The theory of similarity measures is related to the concept of pseudometric spaces in mathematics, yet to date (and best our knowledge) there is no unifying definition of a measure of similarity.

1.4. Similarity measures

The comparisons of expression profiles of cells in highthroughtput single-cell transcriptomic experiments are fundamental concepts that allow for the analysis of heterogeneity of cell types in a given population as well as detection of 'novel' cell types via an unsupervised approach [22]. In this section, we consider n cells $\boldsymbol{c}_1,\boldsymbol{c}_2,\dots\boldsymbol{c}_n$, each with at most m genes being expressed, so that $\boldsymbol{c}_i\in\mathbb{N}^m_{\geqslant 0}$ for $i=1\dots,n$. We aim to compare the supports $supp(\boldsymbol{c}_i)$ of $\boldsymbol{c}_i$ – i.e. quantify the $S_n=\cap_{k=1}^n supp(\boldsymbol{c}_k)$. We start with pairwise comparisons. The two most widely used overlap indices are the Jaccard index and the Sorensen index.

Example 6 (*Jaccard and Sorenson indices*). Let $(\boldsymbol{c}_1,\boldsymbol{c}_2)\in\mathbb{N}^m_{\geqslant 0}\times\mathbb{N}^m_{\geqslant 0}$ be a pair of cells the Jaccard (J) and the Sorensen index (L) of similarity are defined as follows

$$J(\boldsymbol{c}_1,\boldsymbol{c}_2)=\frac{\sum\min(c_{i1},c_{i2})}{\sum(c_{i1}+c_{i2})-\sum\min(c_{i1},c_{i2})}$$

$$L(\boldsymbol{c}_1,\boldsymbol{c}_2)=\frac{2\sum\min(c_{i1},c_{i2})}{\sum(c_{i1}+c_{i2})}.$$

Both the J and L indices, together with various modifications, are widely used not only in ecological studies but also in molecular immunology and transcriptomics (see, e.g., [7,18,23,26]). At the same time, the geometric interpretation leads to a new family o similarity measures, which are defined via the angle (or an appropriate angular measure) between two population vectors in $\mathbb{R}^m_{\geqslant 0}$. The interpretation is obvious since the greater the angle, the more dissimilar two populations tend to be. The most popular of such measures is the Morisita-Horn index [28], which gives the cosine of an angle between a pair of standardized population vectors.

Example 7 (*Morisita-Horn index and Bhattacharyya's coefficient*). Let $(\boldsymbol{c}_1,\boldsymbol{c}_2)\in\mathbb{N}^m_{\geqslant 0}\times\mathbb{N}^m_{\geqslant 0}$ be a pair of population vectors. The Morisita-Horn index (MH) is defined as

$$MH(\boldsymbol{c}_1,\boldsymbol{c}_2)=\frac{2\sum_k\frac{c_{k1}}{\sum c_{i1}}\frac{c_{k2}}{\sum c_{i2}}}{\sum_k\left(\frac{c_{k1}}{\sum c_{i1}}\right)^2+\sum_k\left(\frac{c_{k2}}{\sum c_{i2}}\right)^2}$$

or in terms of the inner products of the normalized populations $\boldsymbol{p_1},\boldsymbol{p_2}$,

$$MH(\boldsymbol{p_1},\boldsymbol{p_2})=\frac{\boldsymbol{2p_1p_2}}{\boldsymbol{p_1^2}+\boldsymbol{p_2^2}}.$$

We have that $0\leqslant MH(\boldsymbol{p}_1,\boldsymbol{p}_2)\leqslant 1$ and it attainis its minimum/maximum when $\boldsymbol{c}_1\perp\boldsymbol{c}_2$ and $\boldsymbol{c}_1=\boldsymbol{c}_2$, respectively. As expected from the low-dimensional intuition this measure tends to be overly sensitive to the highly abundant genes of $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$. It is therefore desirable to often use a more suitable index, known as the Bhattacharyya (BC) coefficient and defined as the cosine of an angle between the vectors $\sqrt{\boldsymbol{p_1}}=\left(\sqrt{\boldsymbol{p_{11}}},\dots,\sqrt{\boldsymbol{p_{m1}}}\right)$ and $\sqrt{\boldsymbol{p_2}}=\left(\sqrt{\boldsymbol{p_{12}}},\dots,\sqrt{\boldsymbol{p_{m2}}}\right)$, i.e.

$$BC(\boldsymbol{p_1},\boldsymbol{p_2}) = \sum_k \sqrt{\frac{\boldsymbol{c_{k1}}}{\sum \boldsymbol{c_{i1}}}}\sqrt{\frac{\boldsymbol{c_{k2}}}{\sum \boldsymbol{c_{i2}}}} = \sum (\boldsymbol{p_{i1}}\boldsymbol{p_{i2}})^{\boldsymbol{1/2}}.$$

The above considerations were generalized by the authors of [35] to a slightly more general form of a geometric overlap index, parametrized by two nonnegative numbers, which allows it to put more weight on rare (resp. abundant) genes. The new index may be viewed as an extension of the indices presented in the previous example.

Example 8 (*PG index*). For any pair $(\boldsymbol{p_1},\boldsymbol{p_2}) \in \Delta_{\boldsymbol{m-1}} \times \Delta_{\boldsymbol{m-1}}$ and $\alpha,\beta \in (0,\infty)$ the power-geometric (or PG) index of overlap is given by

$$PG_{\alpha,\beta}(\boldsymbol{p_1},\boldsymbol{p_2}) = \frac{\sum \boldsymbol{p}_{\boldsymbol{i1}}^{\alpha}\boldsymbol{p}_{\boldsymbol{i2}}^{\beta}}{\sum \boldsymbol{p}_{\boldsymbol{i1}}^{\boldsymbol{2}\alpha} + \sum \boldsymbol{p}_{\boldsymbol{i2}}^{\boldsymbol{2}\beta}}$$

For $m = 2$ the PG index is seen as the cosine of an angle between the vectors $\boldsymbol{p}^{\alpha} := \left(\boldsymbol{p}_{\boldsymbol{11}}^{\alpha},\ldots,\boldsymbol{p}_{\boldsymbol{m,1}}^{\alpha}\right)$ and $\boldsymbol{p}^{\beta} := \left(\boldsymbol{p}_{\boldsymbol{12}}^{\beta}\ldots\boldsymbol{p}_{\boldsymbol{m,2}}^{\beta}\right)$. If $\alpha < 1, \beta < 1$, the *PG* index is less affected by the amount of intersection among the most abundant species than the Morisita-Horn index. Similarly to the Renyi entropy, the PG overlap measure can therefore be seen as putting more weight on rare or abundant species, depending on the value of the parameters α, β. In analogy with a diversity profile, we refer to the function $\alpha \rightarrow PG_{\alpha,\alpha}$ as a *similarity profile*. Note that for $\alpha = \beta = 1$ this is simply the Moristita-Horn index.

Perhaps the two most widely used families of similarity measures for clustering problems are the indices associated with: (1) certain distance metrics in high dimensional Euclidean spaces, and (2) mesures of correlation. Below, we recall the selected standard definitions for completeness:

Example 9. Let $(\boldsymbol{c}_1,\boldsymbol{c}_2) \in \mathbb{N}_{\geqslant 0}^m \times \mathbb{N}_{\geqslant 0}^m$ be a pair of population vectors. For a set $m \geqslant 1$ we define the Minkowski distance (MK) as

$$MK(\boldsymbol{c}_1,\boldsymbol{c}_2) = \left(\sum_i |\ c_{i,1} - c_{i,2}|^m\right)^{1/m}.$$

It is known that the main pitfall of the Minkowski distance is that it is dominated by the highly expressed genes (or the distance between the most abundant genes). This led to a modified definition of the weighted Minkowski distance, yet this consideration goes beyond the scope of this note. Closely related to this measure, yet different in its concept as it is dependent on the 'geometry of the data', is the Mahalanobis distance, which is defined as the weighted Euclidean distance (Minkowski with $m = 2$), where the weights are empirically defined by the covariance matrix Σ between data points – i.e. we have

$$MA(\boldsymbol{c}_1,\boldsymbol{c}_2) := \sqrt{(c_1 - c_2)^T\Sigma^{-1}(c_1 - c_2)}.$$

One of the major issues with MA is that for the calculation of this distance, one needs to invert the covariance matrix, which is not a trivial task when strong linear relationships are present.

At the same time, the two most popular similarity measures based on the measures o statistical dependency (correlation) are worth mentioning – the Pearson and Spearmann correlations are defined as:

$$PC(\boldsymbol{c}_1,\boldsymbol{c}_2) := \left|\left(\sum \frac{(c_{i,1} - \mu_1)(c_{i,2} - \mu_2)}{\sqrt{\sum (c_{i,1} - \mu_1)^2}\sqrt{\sum (c_{i,2} - \mu_2)^2}}\right)\right|$$

and

$$SC(\boldsymbol{c}_1,\boldsymbol{c}_2) := \frac{6\sum d_i^2}{(n^2 - 1)n},$$

where $d_i := rank(c_{i,1}) - rank(c_{i,2})$ and μ_1 and μ_2 are population averages of $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ respectively. The last formula is valid only if there are no ties in the sample.

The two measures of similarity based on correlation coefficents are examples of the use of measures of statistical dependence in defining similarity between data objects. Yet, both of these measures perform well almost exclusively in cases when linear dependence is considered. Another way of measuring similarity is by adapting certain concepts from information theory and signal processing – it relies on the so-called (Renyi) divergence measures. Such measures have been successfully utilized in the context of independence testing in various settings. For instance, the problem of testing for independence in contingency tables is a long standing problem, dating back to the works of Karl Pearson, and still remains unsolved for sparse tables. Below, we recall the definition of Renyi divergence which has proven useful in this setting (see [1]).

Example 10 (*Renyi Divergence*). For a pair of normalized populations $(\boldsymbol{p_1},\boldsymbol{p_2}) \in \Delta_{\boldsymbol{m-1}} \times \Delta_{\boldsymbol{m-1}}$, their *Renyi divergence* of order $\alpha \in [0,\infty]$ is given by

$$F_{\alpha}(\boldsymbol{p_1},\boldsymbol{p_2}) = \frac{\boldsymbol{1}}{\alpha - \boldsymbol{1}}\log\left(\sum \frac{\boldsymbol{p}_{\boldsymbol{i1}}^{\alpha}}{\boldsymbol{p}_{\boldsymbol{i2}}^{\alpha-\boldsymbol{1}}}\right).$$

Note that in the limiting cases we have $F_1(\boldsymbol{p_1},\boldsymbol{p_2}) = -\sum \boldsymbol{p_{i1}}\log\left(\frac{\boldsymbol{p_{i1}}}{\boldsymbol{p_{i2}}}\right)$, which is the Kullblack-Leibler divergence, and $F_{\infty}(\boldsymbol{p_1},\boldsymbol{p_2}) = -\log\left(\max_{\boldsymbol{i}}\frac{\boldsymbol{p_{i1}}}{\boldsymbol{p_{i2}}}\right)$, as well as $F_{\frac{1}{2}}(\boldsymbol{p_1},\boldsymbol{p_2}) = -\boldsymbol{2}\log \boldsymbol{BC}(\boldsymbol{p_1},\boldsymbol{p_2})$.

1.4.1. Information Index

Based on this definition is a similarity measure defined by the authors of [35]. It is a version of the generalized mutual information statistic in two-way tables, and may be therefore viewed an information-theoretical extension of the standard Pearson χ^2-statistic (note that under appropriate conditions, that is 'near' independence of column and rows, the χ^2-statistic is a first order Taylor expansion-based approximation to the Mutual Information). Let $\boldsymbol{P} = [p_{ij}] := \left[\frac{c_{ij}}{\sum_{kl}c_{kl}}\right]$ be a normalized matrix with columns $\boldsymbol{p_1},\boldsymbol{p_2},\ldots,\boldsymbol{p_n}$. Denote also $p_{i\circ} = \sum_j p_{ij}$, $p_{\circ j} = \sum_i p_{ij}$ and the row and column marginals as $\boldsymbol{P}_{\circ} = (p_{\circ 1},\ldots,p_{\circ n}) \in \Delta_{n-1}$, $\boldsymbol{P}^{\circ} = (p_{1\circ},\ldots,p_{m\circ}) \in \Delta_{m-1}$, as well as $\boldsymbol{Q} = \boldsymbol{P}_{\circ} \otimes \boldsymbol{P}^{\circ} := [p_{i\circ}p_{\circ j}]$. The main idea behind the $I - index$ is to measure the 'strength' of the dependence between marginals of the contingency table, instead of e.g. quantifying the pairwise similarity of its columns-specific frequencies as the independence in this table means that the column vectors are proportional. The index is scaled, so as to take its values in the unit interval. The $I - index$ is defined as follows.

Definition 4. For any real $m \times n$ matrix $\boldsymbol{C}$ of nonnegative entries, the *I*-index of order $\alpha \in (0,2)$ is defined as

$$I_{\alpha}(\boldsymbol{C}) = 1 - F_{\alpha}(\boldsymbol{P},\boldsymbol{Q})/H_{2-\alpha}(\boldsymbol{P}_{\circ}).$$

Remark 4. Note that in the case of $\alpha = 1$ we have

$$Q_1(\boldsymbol{C}) = \frac{H_1(\boldsymbol{P}_{\circ}) + H_1(\boldsymbol{P}^{\circ}) - H_1(\boldsymbol{P})}{H_1(\boldsymbol{P}_{\circ})}$$

which is the *mutual information index* scaled by the Shannon entropy of the column-marginal $\boldsymbol{P}_{\circ}$.

It follows from the definition that in the case when $\alpha > 1$ the *I*-index puts more weight on the entries of $\boldsymbol{P}$ with positive dependence (i.e. when $p_{ij} \geqslant p_{i\circ}p_{\circ j}$) and in the case when $\alpha < 1$, it puts more weight on the entries with negative dependence (i.e. when

$p_{ij} \leqslant p_{i\circ}p_{\circ j}$). This feature makes it potentially useful for analyzing the dependence structure of a contingency table (see, for example, [1]).In our setting, the positive dependence between the distributions $\boldsymbol{P}$ and $\boldsymbol{Q} = \boldsymbol{P}_\circ \otimes \boldsymbol{P}^\circ$ intuitively means that if a gene in one population is highly expressed, it also tends to be highly expressed in the remaining populations, with the reverse statement being true for the negative dependence.

2. Divo package

Aforementioned tools to asses the similarity and diversity of expression profiles in single cell systems are assembled in the R package 'divo'. The package is available from CRAN repositoryhttps://CRAN.R-project.org/package=divo. For readers' convenience, we prepared an example code and dataset with brief instruction on GitHub:https://github.com/MPiet11/Seweryn2020_MiniReview.git. Using this code, the readers will be able to test the functions described in this mini review on their own data.

3. Summary

The recent development of single-cell transcriptomics offers a significant opportunity for understanding the heterogeneous signatures in individual cells, and diversity and similarity measures have been applied to evaluate the relationships in gene expressions and cell types, respectively. We overviewed seven diversity and nine similarity measures and showcased their applications in single-cell transcriptome data analyses. Additionally, we assembled these measures into an R package, named divo, to facilitate wide applications for the community.

Despite much progress in analyzing the raw single-cell transcriptome data, diversity and similarity measures can also help evaluate advanced downstream analysis results, such as gene co-expression networks [6], leading to a more comprehensive understanding of differences and commonalities among cells. Given the generality of assessment measures in information theory, they have the potential to be applied to other single-cell omics data, e.g., genomics, epigenetics, proteomics, and metabolomics. Furthermore, the fast-developed single-cell multi-omics data provides a holistic view to better characterize cell heterogeneity by considering the synergistic effects among different omic layers. Such integrative data is a trend in the future single-cell study and leaving a challenge for the optimization of diversity and similarity measures.

CRediT authorship contribution statement

Michal T. Seweryn: Conceptualization, Methodology, Writing - review & editing, Software. **Maciej Pietrzak:** Conceptualization, Methodology, Writing - review & editing, Software. **Qin Ma:** Conceptualization, Methodology, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by an R01 award 1R01GM131399-01 from the National Institute of General Medical Sciences of the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] Agresti A, Kateri M. Categorical Data Analysis. Springer; 2011.
[2] F. Ayed, M. Battiston, F. Camerlenghi, S. Favaro. On consistent estimation of the missing mass, 2018..
[3] Briney B, Inderbitzin A, Joyce C, Burton DR. Commonality despite exceptional diversity in the baseline human antibody repertoire. Nature 2019;566 (7744):393.
[4] Budden DM, Crampin EJ. Information theoretic approaches for inference of biological networks from continuous-valued data. BMC Syst. Biol. 2016;10 (1):89.
[5] Cai S, Georgakilas GK, Johnson JL, Vahedi G. A cosine similarity-based method to infer variability of chromatin accessibility at the single-cell level. Front. Genetics 2018;9.
[6] Carlson MRJ, Zhang B, Fang Z, Mischel PS, Horvath S, Nelson SF. 03). Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. BMC Genomics 2006;7:40.
[7] Chao A, Chazdon RL, Colwell RK, Shen T-J. A new statistical approach for assessing similarity of species composition with incidence and abundance data. Ecol. Lett. 2005;8(2):148–59.
[8] Chao A, Jost L. Estimating diversity and entropy profiles via discovery rates of new species. Methods Ecol. Evol. 2015;6(8):873–82.
[9] Chao A, Shen T-J. Nonparametric estimation of shannon's index of diversity when there are unseen species in sample. Environ. Ecolog. Stat. 2003;10 (4):429–43.
[10] Clark GW, Ackerman SH, Tillier ER, Gatti DL. Multidimensional mutual information methods for the analysis of covariation in multiple sequence alignments. BMC Bioinformatics 2014;15(1):157.
[11] Esty WW. The efficiency of good's nonparametric coverage estimator. Ann. Stat. 1986:1257–60.
[12] Esty WW et al. A normal limit law for a nonparametric estimator of the coverage of a random sample. Ann. Stat. 1983;11(3):905–12.
[13] Franckaert D, Liston A. Expression diversity adds richness to t cell populations. Immunity 2016;45(5):960–2.
[14] S. Ghazanfar, Y. Lin, X. Su, D.M. Lin, E. Patrick, Z.G. Han, J.C. Marioni, J.Y.H. Yang. Investigating higher order interactions in single cell data with schot. bioRxiv, 2019..
[15] Good IJ. The population frequencies of species and the estimation of population parameters. Biometrika 1953;40(3–4):237–64.
[16] Guo M, Bao EL, Wagner M, Whitsett JA, Xu Y. Slice: determining cell differentiation and lineage based on single cell entropy. Nucleic Acids Res. 2016;45(7):e54.
[17] J. Hájek, Comment on a paper by d. basu. Foundations of statistical inference 236, 1971..
[18] Hsieh C-S, Zheng Y, Liang Y, Fontenot JD, Rudensky AY. An intersection between the self-reactive regulatory and nonregulatory t cell receptor repertoires. Nature Immunol. 2006;7(4):401.
[19] Hwang B, Lee JH, Bang D. 08). Single-cell rna sequencing technologies and bioinformatics pipelines. Exp. Mol. Med. 2018;50(8):96.
[20] Jost L. Entropy and diversity. Oikos 2006;113(2):363–75.
[21] Keylock C. Simpson diversity and the shannon–wiener index as special cases of a generalized entropy. Oikos 2005;109(1):203–7.
[22] T. Kim, I.R. Chen, Y. Lin, A.Y.-Y. Wang, J.Y.H. Yang, P. Yang. Impact of similarity metrics on single-cell rna-seq data clustering. Briefings in bioinformatics, 2018..
[23] Komatsu N, Mariotti-Ferrandiz ME, Wang Y, Malissen B, Waldmann H, Hori S. Heterogeneity of natural foxp3+ t cells: a committed regulatory t-cell lineage and an uncommitted minor population retaining plasticity. Proc. Nat. Acad. Sci. 2009;106(6):1903–8.
[24] Kulkarni A, Anderson AG, Merullo DP, Konopka G. Beyond bulk: a review of single cell transcriptomics methodologies and applications. Curr. Opin. Biotechnol. 2019;58:129–36.
[25] Q. Liu, C.A. Herring, Q. Sheng, J. Ping, A.J. Simmons, B. Chen, A. Banerjee, W. Li, G. Gu, R.J. Coffey, Y. Shyr, K.S. Lau. Quantitative assessment of cell population diversity in single-cell landscapes. PLOS Biology 16(10), (2018) e2006687..
[26] Luecken MD, Theis FJ. Current best practices in single-cell rna-seq analysis: a tutorial. Mol. Syst. Biol. 2019;15(6).
[27] Magnussen S. A horvitz–thompson-type estimator of species richness. Environmetrics 2011;22(7):901–10.
[28] Magurran AE. Measuring Biological Diversity. John Wiley & Sons; 2013.
[29] Mao CX, Lindsay BG. A poisson model for the coverage problem with a genomic application. Biometrika 2002;89(3):669–82.
[30] Mossel E, Ohannessian M. On the impossibility of learning the missing mass. Entropy 2019;21(1):28.
[31] Nayak TK. An analysis of diversity using rao's quadratic entropy. Sankhya: Indian J. Stat. , Series B 1986:315–30.
[32] Ntranos V, Kamath GM, Zhang JM, Pachter L, Tse DN. Fast and accurate single-cell rna-seq analysis by clustering of transcript-compatibility counts. Genome Biol. 2016;17(1):112.
[33] R. Qi, A. Ma, Q. Ma, Q. Zou. Clustering and classification methods for single-cell rna-sequencing data. Briefings in Bioinformatics, 2019..
[34] Rambow F, Rogiers A, Marin-Bejar O, Aibar S, Femel J, Dewaele M, Karras P, Brown D, Chang YH, Debiec-Rychter M, Adriaens C, Radaelli E, Wolter P, Bechter O, Dummer R, Levesque M, Piris A, Frederick DT, Boland G, Flaherty KT, van den Oord J, Voet T, Aerts S, Lund AW, Marine J-C. 2019/12/01). Toward

minimal residual disease-directed therapy in melanoma. Cell 2018;174(4):843–855.e19.

[35] Rempala GA, Seweryn M. Methods for diversity and overlap analysis in t-cell receptor populations. J. Math. Biol. 2013;67(6–7):1339–68.

[36] Ricotta C. Through the jungle of biological diversity. Acta Biotheoretica 2005;53(1):29–38.

[37] Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat. Rev. Genet. 2015;16(3):133–45.

[38] Suo S, Zhu Q, Saadatpour A, Fei L, Guo G, Yuan G-C. Revealing the critical regulators of cell identity in the mouse cell atlas. Cell Rep. 2018;25(6):1436–1445.e3.

[39] Tang H, Zeng T, Chen L. High-order correlation integration for single-cell or bulk rna-seq data analysis. Front. Genetics 2019;10:371.

[40] Teschendorff AE, Enver T. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. Nature Commun. 2017;8(1):15599.

[41] Tóthmérész B. Comparison of different methods for diversity ordering. J. Vegetation Sci. 1995;6(2):283–90.

[42] Tsallis C. Introduction to Nonextensive Statistical Mechanics: Approaching a Complex World. Springer Science & Business Media; 2009.

[43] Wan C, Chang W, Zhang Y, Shah F, Lu X, Zang Y, Zhang A, Cao S, Fishel ML, Ma Q, Zhang C. Ltmg: a novel statistical modeling of transcriptional expression states in single-cell rna-seq data. Nucleic Acids Res. 2019;47(18):e111.

[44] Woo J, Winterhoff BJ, Starr TK, Aliferis C, Wang J. 08). De novo prediction of cell-type complexity in single-cell rna-seq and tumor microenvironments. Life Science Alliance 2019;2(4):e201900443.

[45] Xie J, Ma A, Zhang Y, Liu B, Cao S, Wang C, Xu J, Zhang C, Ma Q. Qubic2: a novel and robust biclustering algorithm for analyses and interpretation of large-scale rna-seq data. Bioinformatics 2019.

[46] Zhang C-H, Zhang Z, et al. Asymptotic normality of a nonparametric estimator of sample coverage. Ann. Stat. 2009;37(5A):2582–95.