Measuring the Information Obtained from a Single-Cell Sequencing Experiment

Michael J. Casey[1,2,3], **Rubén J. Sánchez-García**[1,2,3], **and Ben D. MacArthur**[1,2,3,4,*]

[1]Mathematical Sciences, University of Southampton, Southampton, UK
[2]Institute for Life Sciences, University of Southampton, Southampton, UK
[3]The Alan Turing Institute, London, UK
[4]Centre for Human Development, Stem Cells and Regeneration, Faculty of Medicine, University of Southampton, Southampton, UK
[*]Correspondence to Ben MacArthur (bdm@soton.ac.uk)

ABSTRACT

Single-cell sequencing (sc-Seq) experiments are producing increasingly large data sets. However, large data sets do not necessarily contain large amounts of information. Here, we introduce a formal framework for assessing the amount of information obtained from a sc-Seq experiment, which can be used throughout the sc-Seq analysis pipeline, including for quality control, feature selection and cluster evaluation. We illustrate this framework with some simple examples, including using it to quantify the amount of information in a single-cell sequencing data set that is explained by a proposed clustering, and thereby to determine cluster quality. Our information-theoretic framework provides a formal way to assess the quality of data obtained from sc-Seq experiments and the effectiveness of analyses performed, with wide implications for our understanding of variability in gene expression patterns within heterogeneous cell populations.

Introduction

Advances in single-cell sequencing (sc-Seq) technologies have enabled us to profile thousands of cells in a single experiment (Svensson et al. 2018). In combination with advances in unsupervised analysis methods, particularly specialised clustering algorithms and dimensionality reduction methods, these technologies have allowed us to dissect cellular identities in unprecedented detail and discover novel, functionally important, cell types (Trapnell 2015). The goal of most sc-Seq studies (except those focused on methodology development) is to extract biological information, often concerning the mix of cell types present in the tissue sample, from the data obtained. Yet, information is not the same as data; and large, complex, data sets do not necessarily impart useful or usable information. Notably, current single-cell profiling technologies typically produce noisy data for numerous technical reasons, including low capture rate, sparsity due to shallow sequencing and batch effects (Kharchenko et al. 2014, Hicks et al. 2018). Consequently, the relationship between biological information and sc-Seq data is complex and incompletely understood. There is, therefore, a need for formal, quantitative, methods to evaluate this relationship.

To address this challenge we propose an information-theoretic framework, based on a formal yet simple definition of expression heterogeneity, that can be used to precisely measure the amount of information contained in a sc-Seq data set. Our method decomposes the information obtained from an sc-Seq experiment into that which is explained by a grouping of cells – for example, as provided by time-point labels or a proposed clustering of cells as cell types – and that which remains unexplained, and so allows assessment of the extent to which known or inferred mechanisms explain observed expression patterns, or fail to do so. The method is simple to compute and can be used to quickly assess the quality of a sc-Seq data set, compare data quality across platforms or experiments, identify features of importance and contrast competing clustering

protocols.

Results

Single-cell analysis methods typically view cells, identified with vectors of expression, as the objects of study and seek to compare cell identities with each other (Kiselev et al. 2019). However, this cell-centric view is not well suited to quantifying expression heterogeneity, which is concerned with patterns of variation that arise from the mixing of cell types within a population and may vary from gene to gene (Smith & MacArthur 2017). To address this issue, we will take an alternative gene-centric probabilistic view that seeks to more formally specify what is meant by expression homogeneity and heterogeneity.

Consider the expression pattern of a single gene g of interest in a population of N distinct cells. Assume that in total M transcripts of g are identified in the cell population (i.e. across all N cells profiled). Note that M represents the observed transcript count, which may differ from the true count due to technical artefacts. Now consider the stochastic process of assigning the M identified transcripts of g to the N cells profiled. The population is homogeneous with respect to expression of g if all the cells are statistically the same with respect to its expression. Mathematically, this means that the M transcripts of g will be assigned to the N cells independently and equiprobably – i.e. each transcript will be assigned to each cell with probability $1/N$. Conversely, if the population is heterogeneous with respect to expression of g (that is, it consists of a mix of cell types, each expressing the gene differently) then transcripts will not be assigned uniformly, but rather will be assigned preferentially to distinct subsets of cells. Heterogeneity in experimentally observed patterns of expression (and therefore important biology) can, therefore, be assessed in terms of deviation from this hypothetical homogeneous null model.

The Kullback-Leibler divergence (KLD, also known as the relative entropy) is a measure of the information encoded in the difference between an observed and null distribution (Kullback & Leibler 1951). In the sc-Seq context, the KLD of an experimentally observed gene expression distribution from the homogeneous null model measures the extent to which the gene is heterogeneously expressed, or, more precisely, the amount of information that is lost by assuming that the transcripts of g are uniformly distributed in the sequenced cell population. We will denote this quantity $I(g)$ (see **Fig. 1** for a schematic). Formally,

$$I(g) = \log_2(N) - H(g), \tag{1}$$

where $H(g)$ is the entropy of the expression of g in the population (see **Methods** for full details). Since sc-Seq data is noisy, rather than working with the experimentally observed proportions of transcripts assigned to each cell, we will instead adopt a Bayesian approach to estimate the cellular frequencies (using the James-Stein estimator, see **Methods** and **Fig. 1**; Hausser & Strimmer (2009)) and subsequent calculations.

Intuitively, if the cell population is unstructured with respect to the expression of g then the assumption of homogeneity is correct and $I(g) = 0$. Conversely, the theoretical maximum for $I(g)$ is $\log_2(N)$, which is achieved when $H(g) = 0$ and all transcripts of the gene are assigned to the same cell (see **Fig. 1**). Note that: (1) we do not need to know, or model, the particular expression distribution of g in the population, and so no *a priori* assumptions about expression patterns are required to calculate $I(g)$; (2) $I(g)$ is agnostic concerning missing readings so long as they are distributed uniformly at random; (3) since it quantifies the deviation from the homogeneous null model, $I(g)$ measures the information obtained from the experiment concerning the expression of g.

In general, $I(g)$ is associated with cellular diversity: the more distinct cell sub-populations present in a sample, and the more those sub-populations differ from one another with respect to their expression of g, the more information there will be. Thus, it represents a parsimonious measure of expression heterogeneity that makes minimal assumptions concerning expression patterns and imposes minimal technical requirements on data collection methodology or quality. This measure can be used as the basis for numerous aspects of the sc-Seq pipeline, including quality control, feature selection and cluster evaluation.

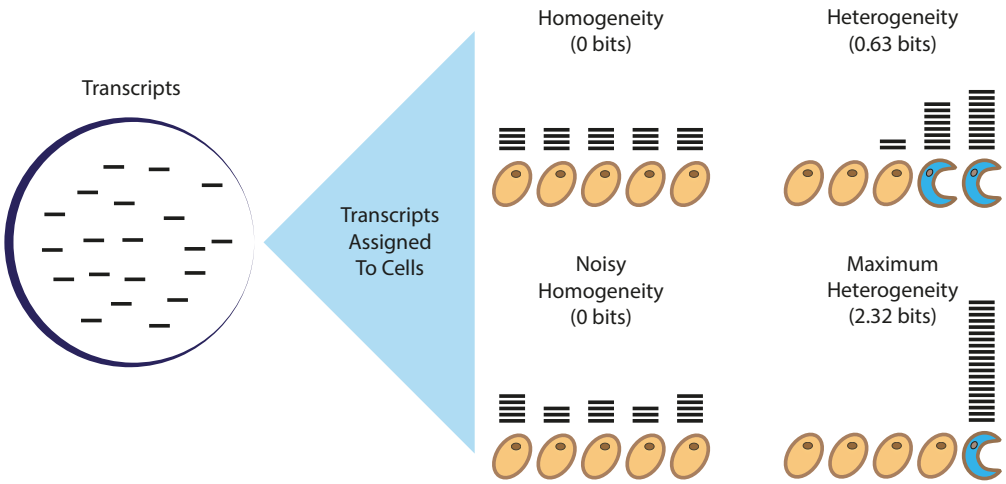

Figure 1. An information-theoretic view of sc-Seq data. Transcripts, or more generally counts, are assigned to cells after sequencing. If the population has no structure, then the information is zero (top left). In practice, as the assignment process is stochastic, and so there will always be some deviation from homogeneity, but the effect on information obtained is buffered by using a shrinkage estimator to estimate frequencies (bottom left). If the population is heterogeneous then transcripts are expressed preferentially in a subset of cells and the information is large (top right), reaching a maximum (at $\log_2(N)$, where N is the number of cells) when only one cell expresses the gene (bottom right).

Quality Control and Feature Selection

The quality of a sc-Seq data set may be assessed by calculating $I(g)$ for each gene profiled. Technical controls that do not contain cell mixtures are expected to broadly conform to the hypothesis of homogeneity; while biological samples consisting of cell mixtures are expected to deviate, with the extent of deviation relating to the complexity of the sample. To illustrate this we analyzed three single-cell RNA-sequencing data sets: *Svensson*, a technical control of RNA spike-ins (4,000 cell equivalents) (Svensson et al. 2017); *Tian*, a mixture of three cancerous cell lines (902 cells) (Tian et al. 2019); and the *Tabula Muris*, an atlas of 12 mouse organs (55,656 cells) (Tabula Muris Consortium 2018).

Substantially more information was obtained from the Tian and Tabula Muris data sets than from Svensson, with 99% of genes in the Svensson technical control having less than 0.5 bits of information (**Fig. 2a-c**). Selecting highly informative genes on the basis of thresholding (here chosen to be 1 bit), we found an order of magnitude more genes were heterogeneously expressed in the Tabula Muris data compared to Tian (5272 and 462 respectively). This difference in highly informative genes matches with expectation; the Tabula Muris is a complex data set containing numerous cell types from different organs, while the Tian data is a controlled mixture of three cell types.

These results indicate that: (1) $I(g)$ is a simple measure of quality for technical controls, which (2) can be used to select potentially informative features for further study.

Cluster Evaluation

Let S be a discrete clustering of cells (i.e. an assignment of cells to a finite set of C non-intersecting sub-populations) and assume that each sub-population is homogeneous with respect to the expression of a gene of interest, g. The null model used to assess deviation from homogeneity may be easily modified to account for the presence of multiple homogeneous cell sub-populations encoded in S.

In the **Methods** we show that $I(g)$ can be decomposed into two parts: one part that quantifies the extent to which the sub-populations defined by S are homogeneous, which we term the *micro-heterogeneity* denoted $h_S(g)$; and one part that quantifies how variably the gene is expressed, in aggregate, between sub-populations, which we term the *macro-heterogeneity*

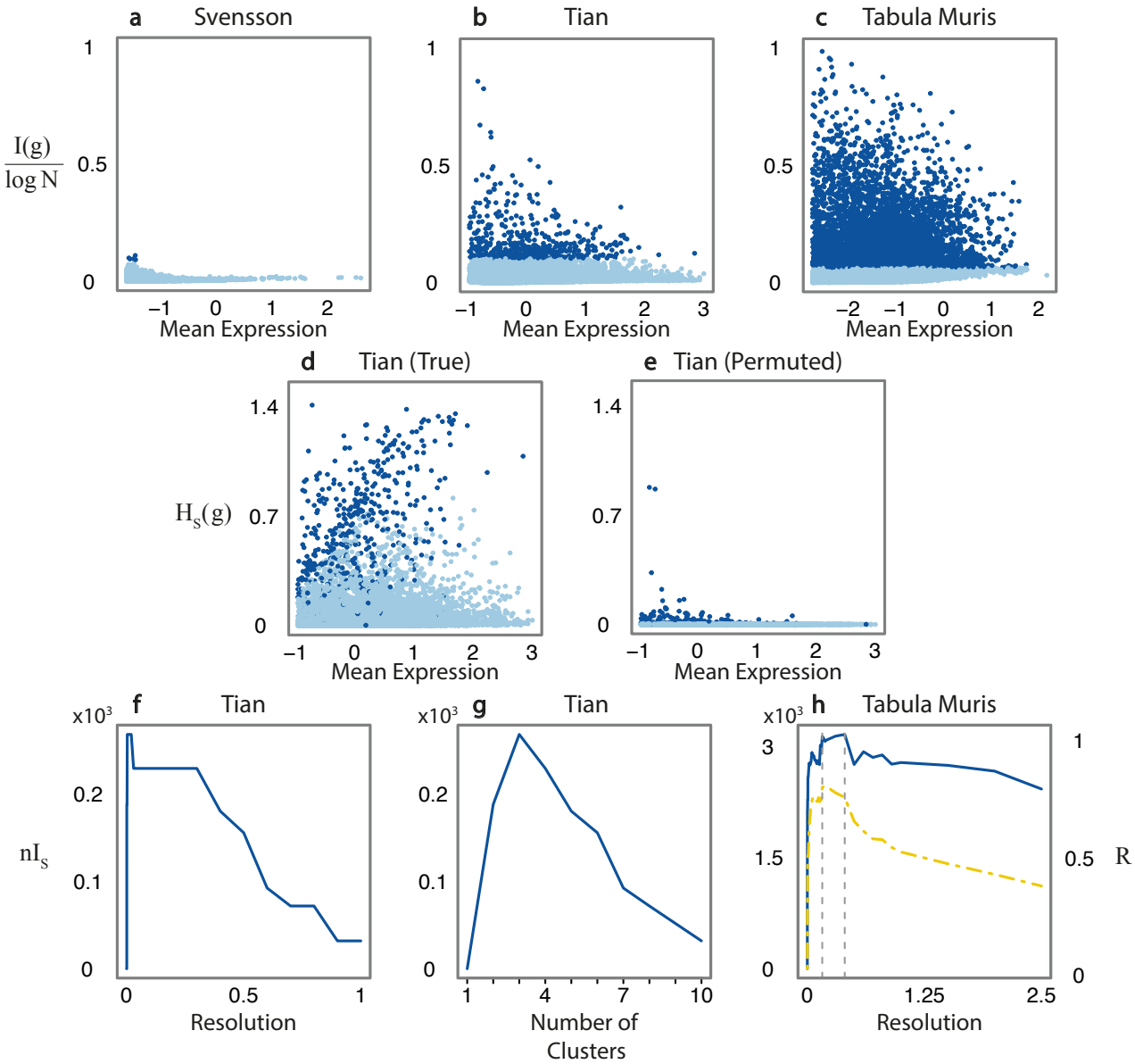

Figure 2. Information as a measure of quality. a-c $I(g)$ for each gene in the (a) Svensson, (b) Tian and (c) Tabula Muris data sets, normalised by the theoretical maximum (log number of cells), against $\log_{10}$ mean expression. Each gene is a point. Genes with $I(g) > 1$ bit are highlighted in dark blue. **d-e** Macro-heterogeneity, H_s, for the Tian data with clustering provided by (d) the genotyped cell annotation and (e) a random permutation of the genotyped annotation. Gene colours inherited from (b). **f-h** Net information explained, nI_S, against (f) cluster resolution and (g) cluster number for the Tian data. **h** nI_S against cluster resolution for the Tabula Muris data. In panel **f** the dashed gold line is the adjusted Rand Index (R) for each clustering with respect to the 'true' annotation provided in (Tabula Muris Consortium 2018). The Seurat (v3) clustering pipeline was used to generate clusterings in panels **f-h** (see **Methods**).

denoted $H_S(g)$.

In particular, we show that

$$I(g) = H_S(g) + h_S(g), \tag{2}$$

for any proposed clustering S. Thus, the information obtained from an experiment concerning the expression of a gene g can be explicitly related to both local and global patterns of variation. Full mathematical details of this decomposition are provided in the Methods. We have also produced an R package for its calculation (see **Methods**).

For a proposed cluster to meaningfully represent a cell type it must be associated with differential expression of at least some subset of genes. Thus, $H_S(g)$ is a simple measure of the biological information in S with respect to expression of g, and can therefore be used as basis for cluster evaluation. While we will anchor our analysis on $H_S(g)$, we note that the micro-heterogeneity $h_S(g)$ is an alternative, quantitatively equivalent measure for cluster evaluation since they are related by Eq. 2.

To illustrate this idea, we considered the Tian data further. In this data, cell type annotations are known *a priori* via genotyping. We can therefore confirm macro-heterogeneity as the measure of biological information by comparing the genotype derived cluster structure (**Fig. 2d**) to random permutations of this ground-truth (**Fig. 2e**). As expected, the ground-truth clustering explained significantly more information than random permutations for 91% of genes (1,000 permutations, $p < 0.05$, one-sided, FDR corrected).

These results indicate that by simple information-theoretic reasoning we are able to quantify the amount of information in the expression pattern of a single gene that is explained by a given clustering. However, a key strength of single-cell methods is that they allow the simultaneous profiling of thousands of genes. This reasoning may be extended to an entire single-cell profiling experiment by assuming that each gene is an independent source of information and making use of the fact that information from independent sources is additive (Shannon 1948). Thus, we can determine the total information explained by a given clustering S, by examining the sum $I_S = \sum_g H_S(g)$. The total information explained by S is a simple, easily computed, measure for cluster evaluation that favours grouping of cells into homogeneous sub-populations and is maximised (at $\sum_g I(g)$) if and only if the proposed clustering divides the population into homogeneous sub-populations and thereby accounts for all the heterogeneity contained in the sc-Seq data set. In this case S is the maximum entropy partition, and may be considered as the most parsimonious way of assigning cell identities to C classes.

To demonstrate this concept, we again made use of the ground-truth annotation provided in the Tian data set. Differential expression between the ground truth sub-populations, as quantified by the macro-heterogeneity, accounted for 736 bits of information; while the presence of heterogeneity within sub-populations of cells, as quantified by the micro-heterogeneity, accounted for 2180 bits of information. In contrast, the macro-heterogeneity of randomized clusters only accounted for 14.6 ± 1.7 bits (mean $\pm$ standard deviation) of information.

All else being equal, the total information explained increases with increasing cluster number (Shannon 1948) and will tend to $\sum_g I(g)$ as the number of clusters tends to N, the number of cells profiled. We can account for this increase by quantifying the amount of information required to encode a given cluster structure, $H(S)$ and penalizing clusterings accordingly. Taking $H(S)$ as the cost, in bits, of a given clustering, we may therefore denote the net information explained by $nI_S = I_S - \lambda H(S)$. The scaling factor, λ, is introduced because the information obtained from each gene may be degenerate with respect to cluster structure, i.e. multiple genes may be differentially expressed in the same cluster(s). A natural choice is $\lambda = \sum_g I(g)/\log N$, for which $nI_S = 0$ when the number of clusters is zero or N. The net information explained is then a global measure for cluster evaluation, independent of cluster number, which is maximized when clustering is optimized.

To illustrate, we used nI_S to determine the optimum number of clusters in the Tian data. Based on the Seurat v3 clustering pipeline, in which cluster number is indirectly determined by the resolution hyperparameter, we generated a range of clusterings and found the corresponding values of nI_S (Hafemeister & Satija 2019). The resulting maxima in nI_S with respect to resolution (**Fig. 2f**) corresponded to three cellular identities (**Fig. 2g**). Furthermore, the clustering associated with this peak was a strong match for the true genotyped cellular identities (0.99 adjusted Rand index, a supervised measure of clustering similarity), indicating that nI_S is a powerful measure for hyperparameter selection.

When we conducted a similar analysis for the Tabula Muris data set, we found that no Seurat produced clustering met or

exceeded the net information explained of the original author-generated annotation, consisting of 56 cell types, which we will take as the 'true' labelling. Instead, we found a plateau in net information explained at Seurat resolution corresponding to 28-36 clusters (**Fig. 2h**). While this suggests the presence of fewer cell types than previously suggested, the resulting clusterings showed a strong similarity to the 'true' labels, as measured by the adjusted Rand Index (**Fig. 2h**). Moreover net information explained and adjusted Rand Index were strongly positively correlated (Pearson's correlation coefficient 0.86; see **Fig. 2h**). Collectively, these results suggests that observed the plateau in net information explained is a robust indicator that a range of different cell type annotations are compatible with the data, and warn against over-reliance of one clustering technology (here Seurat). Under some circumstances it is also likely that certain cellular identities are determined by continuous processes (such as developmental or maturation processes) that are not well resolved by discrete annotations.

Methods

Data Collection

Count matrices for each data set were downloaded from their respective repositories (see **Availability of Data and Materials**). For the Tabula Muris data set, the count matrices of the various tissues were concatenated using the `Matrix` package in R.

Data Pre-processing

For each data set those genes with less than 100 total transcripts in total (i.e. across all cells) were excluded from further analysis.

Mathematical Details

Suppose we conduct an experiment and measure the expression of a gene g in a population of N cells. Let m_i be the number of transcripts associated with cell i in the population and let $\sum_i m_i = M$. Let $p_g(i) = m_i/M$ be the fraction of transcripts of gene g expressed by cell i, for each $1 \leq i \leq N$. Thus, $\sum_{i=1}^{N} p_g(i) = 1$. In the argument that follows we will focus on expression patterns of g in generality and so we will drop the g subscript from our notation from now on.

Let $X = X_g$ be the discrete random variable on the set $\{1, 2, \ldots N\}$ with probabilities

$$x_i = \lambda' \frac{1}{N} + (1 - \lambda') p_g(i), \tag{3}$$

where λ' is the shrinkage intensity given by

$$\lambda' = \frac{1 - \sum_{i=1}^{N} (p_g(i))^2}{(M-1)(\sum_{i=1}^{N} (\frac{1}{N} - p_g(i))^2)}, \tag{4}$$

which is the James-Stein estimator, with the uniform distribution as the shrinkage target (Hausser & Strimmer 2009). Shrinkage approaches deal with undersampling and so are able to correct for the substantial sparsity observed in sc-Seq data. The shrinkage probabilities given in Eq. (3) are a compromise between the observed probabilities $p_g(i)$, which are unbiased but with a high variance, and the target uniform probabilities, which are biased but with low variance (Hausser & Strimmer 2009, Chan et al. 2017).

The Shannon entropy of X is, by definition,

$$H(X) = -\sum_{i=1}^{N} x_i \log x_i. \tag{5}$$

By convention, we assume that $0 \cdot \log 0 = 0$ and take logarithms to the base 2, so the entropy is measured in bits.

The Shannon entropy is a measure of the information, or uncertainty, in the outcomes of X. It has a minimum value of zero, when $x_i = 1$ for some i, that is g is expressed in only one cell in the population (i.e. $m_i = M$ for some i and $m_j = 0$ for all $j \neq i$)

and its maximum value is $\log(N)$, when $x_i = 1/N$ for all i, that is, when the gene is uniformly expressed in the cell population. The entropy may therefore be considered as a measure of the *homogeneity* of expression of the gene g in the cell population profiled. By contrast, the quantity $\log(N) - H(X)$ also ranges between zero and $\log(N)$, yet is minimized when the gene is homogeneously expressed and so is a simple measure of expression *heterogeneity*, which we will denote $I(g)$. We can rewrite this as

$$\log(N) - H(X) = \sum_{i=1}^{N} x_i \log(N) + \sum_{i=1}^{N} x_i \log(x_i) = \sum_{i=1}^{N} x_i \log(N x_i). \tag{6}$$

The Kullback-Leibler divergence of a discrete probability distribution $p_1, \ldots, p_N$ from a discrete probability distribution $q_1, \ldots, q_N$ is, by definition,

$$D_{\mathrm{KL}}(P||Q) = \sum_{i=1}^{N} p_i \log\left(\frac{p_i}{q_i}\right), \tag{7}$$

with the provision that $q_i = 0$ implies $p_i = 0$, and the convention that $0 \cdot \log(\frac{0}{0}) = 0$. From this definition it is clear that our measure of heterogeneity is simply the Kullback-Leibler divergence of the observed expression distribution from the uniform null distribution. Thus,

$$I(g) = \log(N) - H(X) = D_{\mathrm{KL}}(X||U), \tag{8}$$

where U denotes to the uniform distribution on the set $\{1, 2, \ldots N\}$. Since the Kullback-Leibler divergence is the amount of information that is lost when Q is used to approximate P, our measure of heterogeneity is, therefore, the amount of information left unexplained by assuming the expression distribution of g is homogeneous. It is the amount of information obtained from measurement.

A crucial property of the Kullback-Leibler divergence is that it is additively decomposable with respect to arbitrary groupings (Shorrocks 1980). Informally, this means that if we have a clustering of the cells into disjoint groups then $I(g)$, can be reconstructed from within- and between-group heterogeneities.

Let S be a clustering that unambiguously assigns each cell in the sample into one of C non-intersecting sub-populations $S_1, \ldots, S_C$ of sizes $N_1, \ldots, N_C$. Note that $\sum_{k=1}^{C} N_k = N$, the total number of cells. Let y_k be the fraction of transcripts associated with cells in sub-population S_k, adjusted by shrinkage estimator. That is,

$$y_k = \sum_{i \in S_k} x_i. \tag{9}$$

This gives another discrete random variable Y with probability distribution $y_1, \ldots, y_C$, on the set $\{1, 2, \ldots C\}$. For each $k = 1, \ldots, C$, we can also assess the heterogeneity of the sub-population S_k by considering the random variable Z_k with probability distribution $z_i = x_i / y_k$ on the set $i \in S_k$.

We may rewrite $I(g)$ in terms of Y and Z_k, as follows:

$$
\begin{aligned}
I(g) &= \log(N) - \sum_{i=1}^{N} x_i \log\left(\frac{1}{x_i}\right), && (10)\\
&= \log(N) - \sum_{k=1}^{C} \sum_{i \in S_k} x_i \log\left(\frac{1}{x_i}\right), && (11)\\
&= \log(N) - \sum_{k=1}^{C} y_k \sum_{i \in S_k} \frac{x_i}{y_k} \left(\log\left(\frac{1}{x_i/y_k}\right) + \log\left(\frac{1}{y_k}\right)\right), && (12)\\
&= \log(N) - \sum_{k=1}^{C} y_k \underbrace{\sum_{i \in S_k} \frac{x_i}{y_k} \log\left(\frac{1}{x_i/y_k}\right)}_{H(Z_k)} - \sum_{k=1}^{C} y_k \sum_{i \in S_k} \frac{x_i}{y_k} \log\left(\frac{1}{y_k}\right), && (13)\\
&= \log(N) - \sum_{k=1}^{C} y_k H(Z_k) - \underbrace{\sum_{k=1}^{C} \log\left(\frac{1}{y_k}\right) \overbrace{\sum_{i \in S_k} x_i}^{y_k}}_{H(Y)}, && (14)\\
&= \log(N) - \sum_{k=1}^{C} y_k H(Z_k) - H(Y), && (15)\\
&= \underbrace{\log(N) - H(Y) - \sum_{k=1}^{C} y_k \log(N_k)}_{A} \underbrace{+ \sum_{k=1}^{C} y_k \log(N_k) - \sum_{k=1}^{M} y_k H(Z_k)}_{B}. && (16)
\end{aligned}
$$

Expression A may be rewritten as:

$$
\begin{aligned}
A &= \log(N) - H(Y) - \sum_{k=1}^{C} y_k \log(N_k), && (17)\\
&= \sum_{k=1}^{C} y_k \log(N) - \sum_{k=1}^{C} y_k \log\left(\frac{1}{y_k}\right) - \sum_{k=1}^{C} y_k \log(N_k), && (18)\\
&= \sum_{k=1}^{C} y_k \log\left(\frac{y_k}{N_k/N}\right), && (19)\\
&= D_{\mathrm{KL}}(Y || U_{\mathrm{group}}). && (20)
\end{aligned}
$$

This is the Kullback-Leibler divergence of Y from the uniform distribution U_{group} in which $p_k = N_k/N$ for $k = 1, \ldots, C$. Since y_k is the proportion of transcripts assigned to cluster S_k, this is the information explained by any deviation from the assumption that the clusters are homogeneous in their expression of g (i.e. they all express g at the same level). Since it is a measure of the extent to which the population deviates from a homogeneous macroscopic mixture we will term this contribution the *macro-heterogeneity* of g with respect to S, denoted H_S.

Expression B may be rewritten as:

$$B = \sum_{k=1}^{C} y_k \log(N_k) - \sum_{k=1}^{C} y_k H(Z_k), \tag{21}$$

$$= \sum_{k=1}^{C} y_k \left(\log(N_k) - H(Z_k)\right), \tag{22}$$

$$= \sum_{k=1}^{C} y_k \sum_{i \in S_k} \frac{x_i}{y_k} \log\left(N_k \frac{x_i}{y_k}\right), \tag{23}$$

$$= \sum_{k=1}^{C} y_k \sum_{i \in S_k} \frac{x_i}{y_k} \log\left(\frac{x_i/y_k}{1/N_k}\right), \tag{24}$$

$$= \sum_{k=1}^{C} y_k D_{\mathrm{KL}}(Z_k || U_k). \tag{25}$$

This is the weighted sum of the Kullback-Leibler divergences of the empirical distributions of Z_k (i.e. the observed gene expression distribution in group S_k) from the uniform distribution U_k on S_k (in which $p_i = 1/N_k$ for each $i \in S_k$). It is the expected information unexplained by the assumption that the population consists of a mixture of homogeneous sub-populations according to the clustering S (where the expectation is taken with respect to the probability measure provided by Y). Since it is a measure of the expected extent to which the proposed sub-populations deviate from the homogeneous microscopic null model, we will term this contribution the *micro-heterogeneity* of g with respect to S, denoted $h_S(g)$.

Taken together these results show that $I(g)$ can be decomposed into two well-defined parts that encode properties of the global and local structure of the expression distribution of g respectively, as:

$$I(g) = H_S(g) + h_S(g). \tag{26}$$

The Kullback-Leibler divergence is always non-negative and hence so are $H_S(g)$ and $h_S(g)$ for any S. Thus, both quantities range from zero to $I(g)$. If S places one cell in each group (i.e. $C = N$) then $Z_k = U_k$ for all k and thus $H_S(g) = 0$. Conversely, if S places all cells in one group (i.e. $C = 1$) then $Y = U_S$ and thus $h_S(g) = 0$. In this case, the information unexplained by the assumption of homogeneity is equivalent to the information unexplained by the trivial clustering.

Although phrased for a single gene g these notions may be easily extended to the multivariate setting. In this case, the homogeneous null model is obtained by assuming that each gene is expressed homogeneously and independently. Because the information from independent sources is additive (Cover & Thomas 2012), the total information explained in sc-Seq data set by a given clustering is therefore given by the sum:

$$I_S = \sum_g H_S(g). \tag{27}$$

All else being equal, I_S will increase with an increasing number of clusters. We therefore require a regularisation term, preferably one that does not sacrifice interpretability. Keeping to information-theoretic quantities, entropy itself has found use in a range of regularisation problems (Schiebinger et al. 2019).

We take the entropy of the clustering itself, S, as our regularisation term. Let n_k be the fraction of cells in the subset S_k, i.e. $n_k = N_k/N$. The entropy of a proposed clustering is

$$H(S) = -\sum_{k=1}^{C} n_k \log n_k, \tag{28}$$

which will be zero if S places all cells in the same group (i.e. $C = 1$) and at a maximum, $\log(N)$, when all cells belong to a unique cluster (i.e. $C = N$). All else being equal, $H(S)$ will increase with the number of clusters, and so penalizes complex clusterings.

The regularised total information explained is:

$$nI_S = I_S - \lambda H(S), \tag{29}$$

where λ is the regularisation scaling factor. Given both terms are in units of bits, we can interpret the above equation as the net information explained, where $H(S)$ is the cost in information of a given clustering.

In this context, $\lambda > 1$ reflects degeneracy in the clustering. Choosing λ so that the net information explained is zero by definition when either $C = 1$ or $C = N$, we take:

$$\lambda = \frac{\sum_g I(g)}{\log(N)}. \tag{30}$$

As λ is proportional to $\sum_g I(g)$, the net information explained is sensitive to the total information obtained.

Computational Implementation

All the information-theoretic measures described above were calculated using the R package `infohet`, which we have developed and made freely available (see **Code Availability**).

Genotype Annotation

Genotyped cell annotations for the Tian data were taken from the repository metadata. The random clusterings were generated by randomly permuting the genotyped annotation 1,000 times. Comparison of true and shuffled annotations was performed, correcting for multiple testing by false discovery rate using the R function *p.adjust*.

Clustering

Each data set was normalised, with feature selection, using the `Seurat` pipeline, as described in (Hafemeister & Satija 2019) using default parameters. Clustering (Louvain community detection) was carried out with default parameters, with the exception of the resolution, as described in (Stuart et al. 2019). The Adjusted Rand index, based on the annotation provided in (Tabula Muris Consortium 2018), was calculated using the `adj.rand.index` function in the R package `fossil` (Vavrek 2011).

Declarations

Authors' contributions

Conceptualization, MJC, RSG, BDM; Software, MJC; Investigation, MJC, RJSG and BDM; Writing – Original Draft, MJC, RJSG and BDM; Writing – Review & Editing, MJC, RJSG and BDM; Visualization, MJC; Supervision, RJSG and BDM. All authors have seen and approved the manuscript.

Acknowledgements

We thank P. Stumpf, J. Egan and B. Kitching-Morley for helpful discussions.

Funding

RSG and BDM were partially supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1.

Availability of Data and Materials

The Svensson data was downloaded as file "svensson_chromium_control.h5a" from https://data.caltech.edu/records/1264

The Tian data was downloaded as file "sincell_with_class.RDat" from https://github.com/LuyiTian/sc_mixology

The Tabula Muris data was downloaded as set of files "Single-cell RNA-seq data from microfluidic emulsion (v2)" from https://tabula-muris.ds.czbiohub.org

Code Availability

Code for the calculation of all information theoretic quantities is available as an R package at https://github.com/mjcasy/infohet

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

Chan, T. E., Stumpf, M. P. & Babtie, A. C. (2017), 'Gene regulatory network inference from single-cell data using multivariate information measures', *Cell systems* **5**(3), 251–267.

Cover, T. M. & Thomas, J. A. (2012), *Elements of information theory*, John Wiley & Sons.

Hafemeister, C. & Satija, R. (2019), 'Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression', *Genome Biology* **20**(1), 1–15.

Hausser, J. & Strimmer, K. (2009), 'Entropy inference and the james-stein estimator, with application to nonlinear gene association networks.', *Journal of Machine Learning Research* **10**(7).

Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A. (2018), 'Missing data and technical variability in single-cell rna-sequencing experiments', *Biostatistics* **19**(4), 562–578.

Kharchenko, P. V., Silberstein, L. & Scadden, D. T. (2014), 'Bayesian approach to single-cell differential expression analysis', *Nature methods* **11**(7), 740–742.

Kiselev, V. Y., Andrews, T. S. & Hemberg, M. (2019), 'Challenges in unsupervised clustering of single-cell rna-seq data', *Nature Reviews Genetics* **20**(5), 273–282.

Kullback, S. & Leibler, R. A. (1951), 'On information and sufficiency', *The annals of mathematical statistics* **22**(1), 79–86.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P. et al. (2019), 'Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming', *Cell* **176**(4), 928–943.

Shannon, C. E. (1948), 'A mathematical theory of communication', *Bell system technical journal* **27**(3), 379–423.

Shorrocks, A. F. (1980), 'The class of additively decomposable inequality measures', *Econometrica: Journal of the Econometric Society* pp. 613–625.

Smith, R. C. & MacArthur, B. D. (2017), 'Information-theoretic approaches to understanding stem cell variability', *Current Stem Cell Reports* **3**(3), 225–231.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck III, W. M., Hao, Y., Stoeckius, M., Smibert, P. & Satija, R. (2019), 'Comprehensive integration of single-cell data', *Cell* **177**(7), 1888–1902.

Svensson, V., Natarajan, K. N., Ly, L.-H., Miragaia, R. J., Labalette, C., Macaulay, I. C., Cvejic, A. & Teichmann, S. A. (2017), 'Power analysis of single-cell rna-sequencing experiments', *Nature methods* **14**(4), 381.

Svensson, V., Vento-Tormo, R. & Teichmann, S. A. (2018), 'Exponential scaling of single-cell rna-seq in the past decade', *Nature protocols* **13**(4), 599–604.

Tabula Muris Consortium (2018), 'Single-cell transcriptomics of 20 mouse organs creates a tabula muris.', *Nature* **562**(7727), 367.

Tian, L., Dong, X., Freytag, S., Lê Cao, K.-A., Su, S., JalalAbadi, A., Amann-Zalcenstein, D., Weber, T. S., Seidi, A., Jabbari, J. S. et al. (2019), 'Benchmarking single cell rna-sequencing analysis pipelines using mixture control experiments', *Nature methods* **16**(6), 479–487.

Trapnell, C. (2015), 'Defining cell types and states with single-cell genomics', *Genome research* **25**(10), 1491–1498.

Vavrek, M. J. (2011), 'Fossil: palaeoecological and palaeogeographical analysis tools', *Palaeontologia Electronica* **14**(1), 16.