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 that 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 clustering. 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. Our information-theoretic framework provides a formal way to assess the quality of data obtained from sc-Seq experiments and has 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 an appropriate null model – for example, as provided by a technical control or a proposed clustering – 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 $1/N$. From a Bayesian perspective this corresponds to taking the most non-committal uniform prior. 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 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 refer to this as the information unexplained for gene g, denoted $H_U(g)$ (see **Fig. 1** for a schematic). Formally,

$$H_U(g) = \log_2(N) - H_g, \tag{1}$$

where H_g is the entropy of the expression of g in the population (see **Methods** for full details). Intuitively, if the cell population is pure with respect to the expression of g then the assumption of homogeneity is correct and the information unexplained is zero. Conversely, the theoretical maximum for $H_U(g)$ 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 $H_U(g)$; (2) $H_U(g)$ 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, $H_U(g)$ measures the information obtained from the experiment concerning the expression of g.

In general, $H_U(g)$ 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 unexplained by the hypothesis of homogeneity 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.

Quality Control and Feature Selection

The quality of a sc-Seq data set may be assessed by calculating the information unexplained by an assumption of homogeneity for each gene profiled. Technical controls that do not contain cell mixtures are expected to broadly conform to this hypothesis;

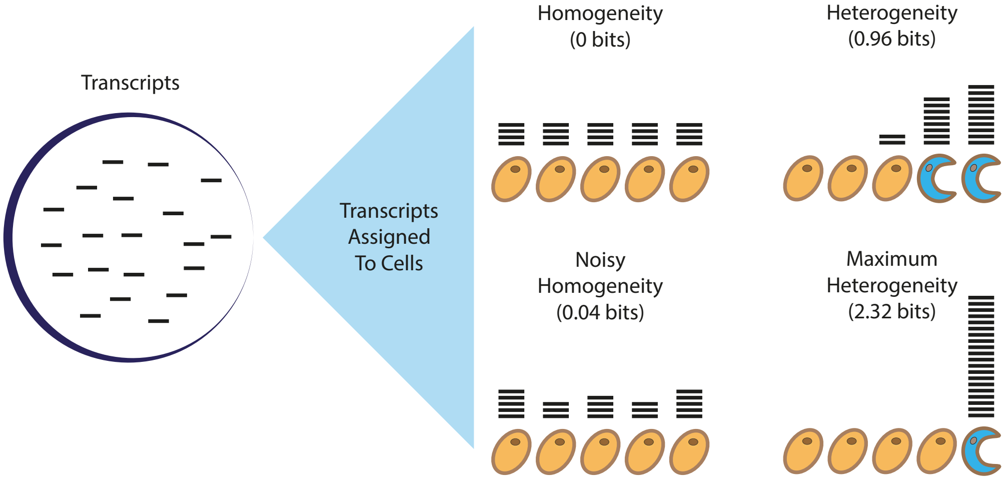
Transcripts
Transcripts Assigned To Cells
Homogeneity (0 bits)
Heterogeneity (0.96 bits)
Noisy Homogeneity (0.04 bits)
Maximum Heterogeneity (2.32 bits)

Figure 1. An information-theoretic view of sc-Seq data. Transcripts, or more generally counts, are assigned to cells after sequencing. If the population is pure, then the information unexplained by the hypothesis of homogeneity is zero (top left). In practice, the assignment process is stochastic, so will naturally result in a small non-zero information unexplained (bottom left). If the population is heterogeneous then transcripts are expressed preferentially in a subset of cells and the information unexplained 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).

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 five cancerous cell lines (3,918 cells) (Tian et al. 2019); and the *Tabula Muris*, an atlas of 12 mouse organs (55,656 cells) (Tabula Muris Consortium 2018). For each data set, we calculated both the observed and the expected value of H_U for each gene. The expected values were found by simulating the homogeneous null model, with adjustment for sequence depth (see **Methods**). The factors captured by this null simulation (namely sparsity, stochasticity and sequence depth variation) are unlikely to be the sole sources of variation in gene expression. Therefore, we only considered a gene g to be heterogeneously expressed if $H_U(g)$ exceeded that expected from the null simulation by 0.5 bits.

Based on this threshold, we did not identify any genes that deviated from homogeneity in the Svensson data set (**Fig. 2a**), as expected from this technical control. However, in both the Tian (**Fig. 2b**) and the Tabula Muris (**Fig. 2c**) data sets, we identified numerous heterogeneously expressed genes. Notably, there were strikingly more heterogeneously expressed genes in the Tabula Muris data than the Tian data (4,430 and 973 respectively), reflecting the fact that the Tabula Muris is a complex data set containing numerous cell types from different organs, while the Tian data is a controlled mixture of 5 cell types. In total 2.6%, 10.3% and 11.6% of the information obtained from the experiment was unaccounted for by the homogeneous null model for the Svensson, Tian and Tabula Muris experiments respectively (see **Methods** for details). These results indicate that: (1) information unexplained is a simple measure of quality for technical controls, which (2) can be used to select potentially informative features for further study; and (3) a remarkable amount of the information obtained from even complex sc-Seq experiments can be explained by an assumption of homogeneity.

Cluster Quality

Let S be a discrete clustering of cells (i.e. an assignment of cells to a finite set of 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 information unexplained under the assumption of homogeneity may be easily modified to account for the presence of multiple

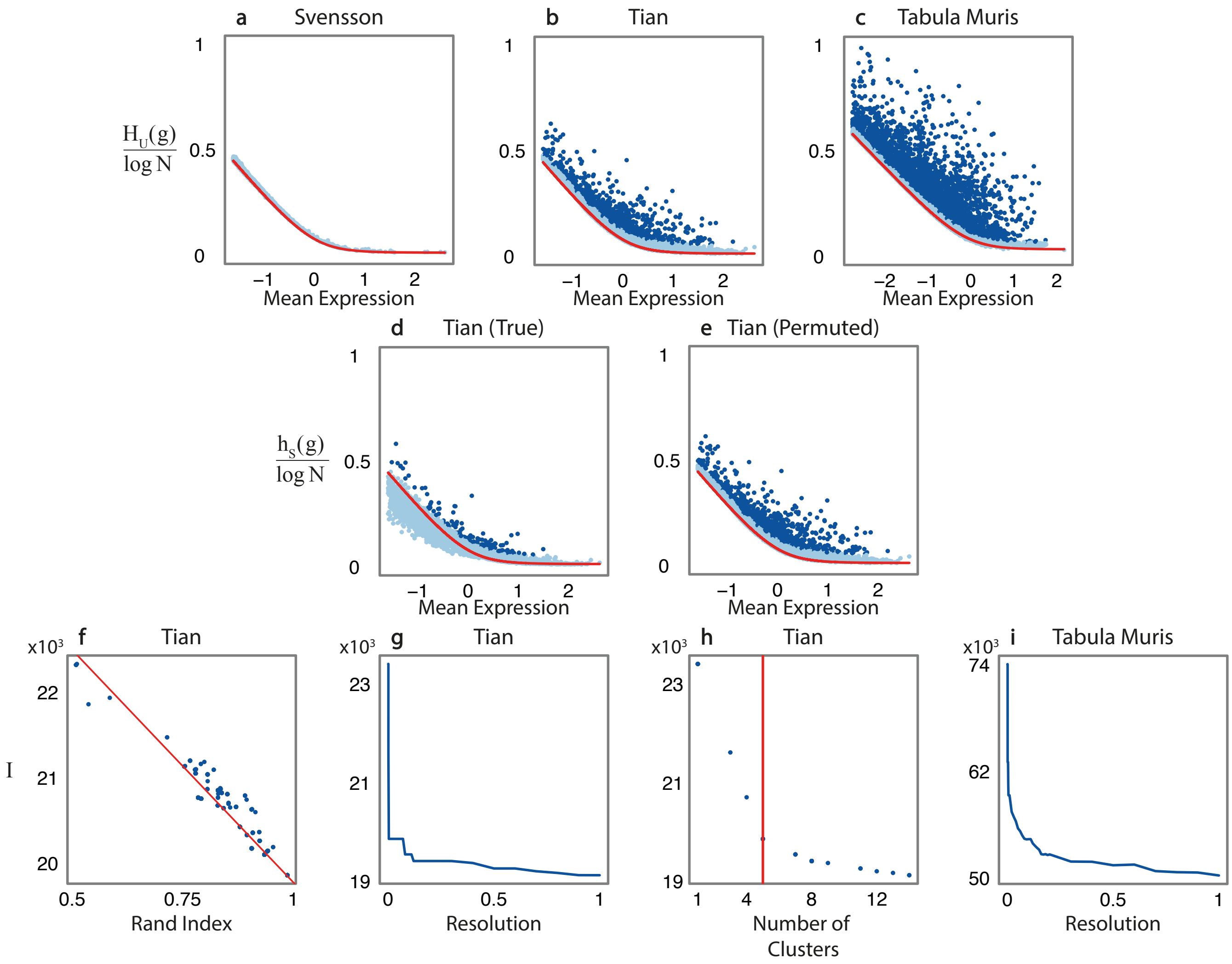

Figure 2. Information unexplained as a measure of quality. a-c Information unexplained $H_U(g)$, for each gene in the Svensson (a), Tian (b) and Tabula Muris (c) data sets, normalised by the theoretical maximum (log number of cells), against $\log_{10}$ mean expression. Each gene is a point. The red-line is the expected information unexplained from the depth-adjusted null model. Genes with more than 0.5 bits information unexplained are highlighted in dark blue. **d-e** Micro-heterogeneity, h_s, for the Tian data with clustering provided by **d** the genotyped cell annotation and **e** that expected from randomly permuted annotations (2,000 permutations). **f** Total information unexplained, I, against Rand index for k-means clustering (5 centres, 1,000 repeats). Each point is a single realization of k-means clustering. The red-line shows linear regression. **g-h** Elbow plots of total information unexplained I, against cluster resolution (g) and cluster number (h) for the Tian data. The red line indicates the true cluster number (5). **i** Elbow plot of total information unexplained I against cluster resolution for the Tabula Muris data. In panels (g-i) the clustering was performed using the Louvain method (see **Methods**).

homogeneous cell sub-populations encoded in S.

In the **Methods** we show that $H_U(g)$ can be decomposed into two parts: one part related to the extent to which the proposed clustering is compatible with an assumption of local homogeneity in each sub-population of cells, which we term the *micro-heterogeneity* denoted $h_S(g)$; and one part related to how variably the gene is expressed, on average, between sub-populations, which we term the *macro-heterogeneity* denoted $H_S(g)$.

In particular, we show that

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

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**).

Since the micro- and macro-heterogeneity are related by Eq. (1), either can be used as a simple measure of cluster quality. Since, informally, the micro-heterogeneity $h_S(g)$ assesses the amount of information in the expression pattern of g that is unexplained by the proposed clustering (the better S explains the experimentally observed expression patterns the smaller $h_S(g)$ will be) we will use here as a measure of cluster quality.

To illustrate this idea, we considered the Tian data further. In this data, cell type annotations are known *a priori* via genotyping. We compared the information unexplained by this ground truth clustering (**Fig. 2d**) to that unexplained by random permutations of the ground truth (**Fig. 2e**). As expected, the ground truth clustering explained more information than all random permutations for 99.5% of genes (2,000 permutations, $p < 0.05$, FDR corrected).

These results indicate that by simple information-theoretic reasoning we are able to quantify the amount of information explained by a given clustering on the expression of a single gene. However, a key strength of single-cell methods is that they allow the simultaneous profiling of thousands of genes. Very similar reasoning may be used to calculate the information explained by 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 unexplained by a given clustering, by examining the sum $I = \sum_g h_S(g)$. From a Bayesian perspective this corresponds to taking the most non-committal multivariate uniform prior. The total information unexplained I is a simple, easily computed, measure for cluster quality that favours grouping of cells into homogeneous sub-populations and is minimized (at zero) if and only if the proposed clustering accounts for all the heterogeneity contained in the sc-Seq data set.

To demonstrate this concept, we again made use of the ground truth annotation provided in the Tian data set. Normalising I by total information obtained ($\sum_g H_U(g)$), the ground truth annotation left 85.1% information unexplained compared to 99.78% ($\pm 0.01\%$) left unexplained by random permutations. To investigate further we conducted k-means clustering (5 centres, 1,000 repeats) and calculated the total information unexplained for each realization. We found a strong negative correlation between similarity to the ground truth annotation (determined by Rand index) and information unexplained (-0.97, Pearson's correlation coefficient) (**Fig. 2f**) (Hastie et al. 2009, Rand 1971). Collectively, these results indicate that information unexplained is simple way to assess annotation quality for individual genes. Moreover, they indicate that a substantial amount of information unexplained is due to associations between genes, which are not included in the multivariate homogeneous null model. The extent of these associations may also be quantified using the information unexplained.

In general, the total information unexplained decreases with increasing cluster number (Shannon 1948) and will tend to zero as the number of clusters tends to N, the number of cells profiled. There is, therefore, a trade-off between the number of clusters included in a proposed clustering and the total amount of information unexplained. Similar issues arise with other measures of cluster quality, and a variety of different methods exist for identifying the appropriate number of clusters in a data set (Tibshirani et al. 2001, Rousseeuw 1987).

To illustrate, we used the elbow heuristic with total information unexplained I as the quality metric, to determine the optimum number of clusters in the Tian data, using the Seurat clustering pipeline (Hafemeister & Satija 2019). We observed an evident elbow in the Seurat resolution hyperparameter (which indirectly determines cluster number; **Fig. 2g**), corresponding to five cellular identities (**Fig. 2h**). The resulting clustering is a strong association (0.99 Rand index) to the ground truth genotype annotation, indicating that total information unexplained can be used to successfully optimize clustering. A similar analysis for the Tabula Muris data also indicated a strong elbow, suggesting that information unexplained can also be used to identify clusters in complex data sets (**Fig. 2i**).

In conclusion, we have presented a simple framework, based on an information-theoretic interpretation of cell type, that can be used throughout the sc-Seq analysis pipeline. For large cell atlas data-sets, such as the human cell landscape, multiple levels of clustering are often required to identify subtle differences between cell types (Han et al. 2020). In our framework, the information unexplained is a simple measure of the extent to which the expression pattern of a given gene is explained by a proposed clustering. Substantial information unexplained can, therefore, be used to identify clusters of cells and/or genes that would benefit from further investigation. If an observed gene expression pattern in a population of cells remains unexplained by more refined clustering, then it may be that it cannot be well described by a uniform mixture model. This implies that the observed expression patterns are not consistent with the presence of a mixture of homogeneous cell sub-populations. Such patterns are particularly interesting since they indicate that more complex expression dynamics, such as those associated with temporal oscillations, may be present and worth further investigation.

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 $x_i = p_g(i)$. The Shannon entropy of X is, by definition,

$$H(X) = -\sum_{i=1}^{N} x_i \log x_i. \tag{3}$$

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 $H_U(g)$. We can rewrite this as

$$\log(N) - H(X) = \sum_{i=1}^{N} x_i \log(N) + \sum_{i=1}^{N} x_i \log(x_i) = \sum_{i=1}^{N} x_i \log(N x_i). \tag{4}$$

The Kullback-Leibler divergence of a discrete probability distribution $p_1, \ldots, p_N$ from a discrete probability distribution

$q_1,\dots,q_N$ is, by definition,

$$D_{\mathrm{KL}}(P||Q) = \sum_{i=1}^{N} p_i \log\left(\frac{p_i}{q_i}\right), \tag{5}$$

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,

$$H_U(g) = \log(N) - H(X) = D_{\mathrm{KL}}(X||U), \tag{6}$$

where U denotes to the uniform distribution on the set $\{1,2,\dots N\}$. 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.

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 $H_U(g)$, 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 $S_1,\dots,S_C$ of sizes $N_1,\dots,N_C$. 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, that is,

$$y_k = \sum_{i\in S_k} x_i. \tag{7}$$

This gives another discrete random variable Y with probability distribution $y_1,\dots,y_C$, on the set $\{1,2,\dots C\}$. For each $k = 1,\dots,C$, 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 $H_U(g)$ in terms of Y and Z_k, as follows:

$$H_U(g) = \log(N) - \sum_{i=1}^{N} x_i \log\left(\frac{1}{x_i}\right), \tag{8}$$

$$= \log(N) - \sum_{k=1}^{C} \sum_{i\in S_k} x_i \log\left(\frac{1}{x_i}\right), \tag{9}$$

$$= \log(N) - \sum_{k=1}^{C} y_k \sum_{i\in S_k} \frac{x_i}{y_k}\left(\log\left(\frac{1}{x_i/y_k}\right) + \log\left(\frac{1}{y_k}\right)\right), \tag{10}$$

$$= \log(N) - \sum_{k=1}^{C} y_k \underbrace{\sum_{i\in S_k} \frac{x_i}{y_k} \log\left(\frac{1}{x_i/y_k}\right)}_{H(Z_k)} - \sum_{k=1}^{C} y_k \sum_{i\in S_k} \frac{x_i}{y_k} \log\left(\frac{1}{y_k}\right), \tag{11}$$

$$= \log(N) - \sum_{k=1}^{C} y_k H(Z_k) - \underbrace{\sum_{k=1}^{C} \log\left(\frac{1}{y_k}\right) \overbrace{\sum_{i\in S_k} x_i}^{y_k}}_{H(Y)}, \tag{12}$$

$$= \log(N) - \sum_{k=1}^{C} y_k H(Z_k) - H(Y), \tag{13}$$

$$= \underbrace{\log(N) - H(Y) - \sum_{k=1}^{C} y_k \log(N_k)}_{A} + \underbrace{\sum_{k=1}^{C} y_k \log(N_k) - \sum_{k=1}^{M} y_k H(Z_k)}_{B}. \tag{14}$$

Expression A may be rewritten as:

$$\begin{aligned} A &= \log(N) - H(Y) - \sum_{k=1}^{C} y_k \log(N_k), && (15)\\ &= \sum_{k=1}^{C} y_k \log(N) - \sum_{k=1}^{C} y_k \log\left(\frac{1}{y_k}\right) - \sum_{k=1}^{C} y_k \log(N_k), && (16)\\ &= \sum_{k=1}^{C} y_k \log\left(\frac{y_k}{N_k/N}\right), && (17)\\ &= D_{\mathrm{KL}}(Y||U_{\mathrm{group}}). && (18) \end{aligned}$$

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 unexplained by 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:

$$\begin{aligned} B &= \sum_{k=1}^{C} y_k \log(N_k) - \sum_{k=1}^{C} y_k H(Z_k), && (19)\\ &= \sum_{k=1}^{C} y_k \left(\log(N_k) - H(Z_k)\right), && (20)\\ &= \sum_{k=1}^{C} y_k \sum_{i\in S_k} \frac{x_i}{y_k} \log\left(N_k \frac{x_i}{y_k}\right), && (21)\\ &= \sum_{k=1}^{C} y_k \sum_{i\in S_k} \frac{x_i}{y_k} \log\left(\frac{x_i/y_k}{1/N_k}\right), && (22)\\ &= \sum_{k=1}^{C} y_k D_{\mathrm{KL}}(Z_k||U_k). && (23) \end{aligned}$$

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 $H_U(g)$ 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:

$$H_U(g) = H_S(g) + h_S(g). \qquad (24)$$

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 $H_U(g)$. 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. Again, this corresponds to the least informative Bayesian prior. Because the information from independent sources is additive (Cover & Thomas 2012), the total information unexplained in sc-Seq data set by a given clustering is therefore given by the sum:

$$I = \sum_g h_S(g). \tag{25}$$

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 **Availability of Data and Materials**).

Null Model

Typically in sc-Seq data there is substantial variation in cellular count-depths, i.e. the total number of transcripts expressed in each cell. This variation is thought to be largely technical, not biological (Hafemeister & Satija 2019). We accounted for this variation as follows. Let l_i be the total number of transcripts associated with cell i (i.e. across all genes) and denote $\sum_i l_i = L$. Now let $\rho_i = l_i/L$ be the proportion of transcripts observed in total that are expressed in cell i, for each $1 \leq i \leq N$. This determines a discrete random variable Γ with probability distribution $\rho_1, \rho_2 ..., \rho_N$ on the set $\{1, 2, \ldots N\}$, which determines the likelihood that a given transcript will be assigned to a given cell, taking into account sequencing depth.

To realize the null homogeneous model computationally we repeatedly assigned (300 times) the M transcripts experimentally associated with each gene g to the N cells profiled independently using the probability measure Γ.

Genotype Annotation

Genotyped cell annotations for the Tian data are available in the repository metadata. Random clustering was generated by randomly permuting the genotyped annotation 2,000 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. All clustering was carried out on the normalised data. The inbuilt R function `kmeans` was used with default parameters. `Seurat` clustering (Louvain community detection) was carried out with default parameters, with the exception of the resolution, as described in (Stuart et al. 2019). The Rand index was calculated using the `rand.index` function in the R package `fossil`. Pearson's correlation coefficient was found using the inbuilt R function `cor`.

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_5cl.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 for the calculation of all information theoretic quantities is available as an R package at `https://github.com/mcaseySoton/infohet`

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

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.

Han, X., Zhou, Z., Fei, L., Sun, H., Wang, R., Chen, Y., Chen, H., Wang, J., Tang, H., Ge, W. et al. (2020), 'Construction of a human cell landscape at single-cell level', *Nature* **581**(7808), 303–309.

Hastie, T., Tibshirani, R. & Friedman, J. (2009), *The elements of statistical learning: data mining, inference, and prediction*, Springer Science & Business Media.

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.

Rand, W. M. (1971), 'Objective criteria for the evaluation of clustering methods', *Journal of the American Statistical association* **66**(336), 846–850.

Rousseeuw, P. J. (1987), 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis', *Journal of computational and applied mathematics* **20**, 53–65.

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.

Tibshirani, R., Walther, G. & Hastie, T. (2001), ‘Estimating the number of clusters in a data set via the gap statistic’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**(2), 411–423.

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