BMC Bioinformatics

RESEARCH ARTICLE Open Access

An information-theoretic approach to single cell sequencing analysis

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

*Correspondence:
R.Sanchez-Garcia@soton.ac.uk;
bdm@soton.ac.uk

[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

Abstract

Background: Single-cell sequencing (sc-Seq) experiments are producing increasingly large data sets. However, large data sets do not necessarily contain large amounts of information.

Results: Here, we formally quantify the information obtained from a sc-Seq experiment and show that it corresponds to an intuitive notion of gene expression heterogeneity. We demonstrate a natural relation between our notion of heterogeneity and that of cell type, decomposing heterogeneity into that component attributable to differential expression between cell types (inter-cluster heterogeneity) and that remaining (intra-cluster heterogeneity). We test our definition of heterogeneity as the objective function of a clustering algorithm, and show that it is a useful descriptor for gene expression patterns associated with different cell types.

Conclusions: Thus, our definition of gene heterogeneity leads to a biologically meaningful notion of cell type, as groups of cells that are statistically equivalent with respect to their patterns of gene expression. Our measure of heterogeneity, and its decomposition into inter- and intra-cluster, is non-parametric, intrinsic, unbiased, and requires no additional assumptions about expression patterns. Based on this theory, we develop an efficient method for the automatic unsupervised clustering of cells from sc-Seq data, and provide an R package implementation.

Background

Advances in single-cell sequencing (sc-Seq) technologies have enabled us to profile thousands of cells in a single experiment [43]. In combination with advances in unsupervised analysis methods, particularly specialised clustering algorithms and dimensionality reduction techniques, these technologies have allowed us to dissect cellular identities in unprecedented detail and discover novel, functionally important, cell types [48]. 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, data is not the same as information; and large, complex, data sets do not necessarily convey 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 [17, 19]. 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 that quantifies the amount of information contained in a sc-Seq data set, and leads to a natural definition of gene expression heterogeneity. Our measure of gene expression heterogeneity decomposes into that which is explained by a given grouping of cells—a proposed clustering into cell types, for example—and that which remains unexplained.

Our framework differs from other approaches to heterogeneity decomposition (e.g. [14, 27]) by formally quantifying the information, in terms of gene expression heterogeneity, gained from a sc-Seq experiment concerning the mix of cell types present. The resulting measure of heterogeneity, and its decomposition into inter- and intra-cluster heterogeneity, is non-parametric, intrinsic, unbiased, and requires no *a priori* assumptions about gene expression patterns.

Our approach is mathematically precise, biologically intuitive and computationally simple to implement, enabling a practitioner to quickly assess the information content of a sc-Seq data set, in terms of gene expression heterogeneity, identify highly informative genes, and determine the extent to which observed patterns of gene expression variability are explained by the presence of a mixture of cell types in a cellular population. Furthermore, we provide an efficient unsupervised clustering algorithm of cells from sc-Seq data based on heterogeneity and an implementation as an R package.

Results

High-throughput single-cell analysis methods, such as single-cell sequencing, typically view cells as the objects of study and seek to compare cell identities with each other [4, 12, 14, 20, 24, 25, 35, 47]. However, this cell-centric view is less well suited to quantifying gene 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. For instance, variance, the standard measure of heterogeneity in the cell-centric view, may not be informative in multi-modal distributions [39].

In the context of sc-Seq, heterogeneity results from differential expression between cell types. In the simplest case, where one cell type expresses a gene highly and another lowly, such differential expression introduces bimodality in the cell-centric expression distribution. In a bimodal distribution, the population mean is no longer characteristic of either of the two subpopulations, making variance, as a measure of the spread about the mean, also misleading (see Smith and MacArthur [39]).

To meet these challenges, we introduce a novel gene-centric probabilistic view that seeks to more formally specify what is meant by gene expression homogeneity and heterogeneity. We show that this gene-centric view is better suited to quantifying gene expression heterogeneity in the context of a multimodal population, and formalises the biological intuition of expression homogeneity and heterogeneity.

Quantifying gene expression 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 gene g to the N cells profiled. Intuitively, 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 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 [22]. In the sc-Seq context, the KLD of an experimentally observed gene expression distribution from the homogeneous null model, described above, measures the amount of information that is lost by assuming that the gene is homogeneously expressed in the sequenced cell population. We will denote this quantity $I(g)$ and refer to it as the *heterogeneity* of g (see Fig. 1 for a schematic). Formally,

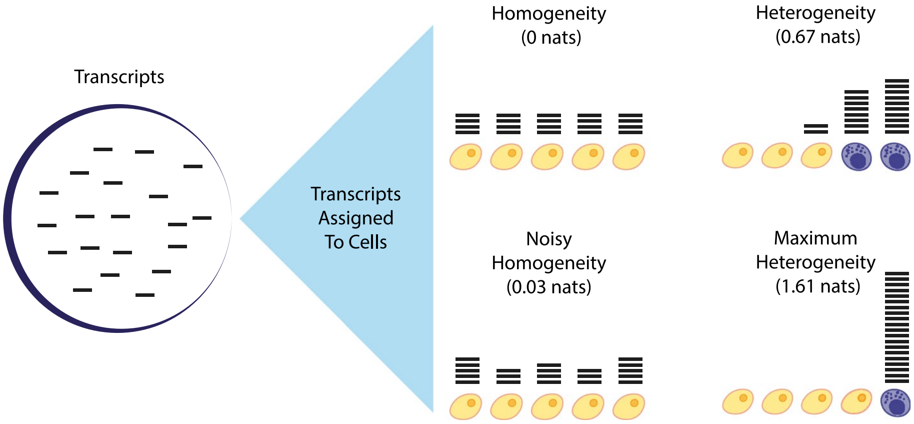

Transcripts
Transcripts Assigned To Cells
Homogeneity (0 nats)
Heterogeneity (0.67 nats)
Noisy Homogeneity (0.03 nats)
Maximum Heterogeneity (1.61 nats)

Fig. 1 An information-theoretic view of sc-Seq data. Transcripts, or more generally counts, of a given gene (shown here as horizontal bars) are assigned to cells after sequencing. If the cell population is homogeneous with respect to the expression of g, then the heterogeneity $I(g)$ will be zero (top left population, $I(g) = 0$). In practice, the transcript assignment process is stochastic, and so there will always be some deviation from this ideal (bottom left population, $I(g)$ small). (Note that the technical effects of this stochasticity on the information obtained may be reduced by using a shrinkage estimator to determine the distribution of transcripts (see "Methods" Section)). If the population is heterogeneous, then transcripts may be preferentially expressed in a subset of cells and the information obtained from the experiment, as measured by $I(g)$ will be larger (top right population, $I(g)$ large), reaching a maximum at $\log(N)$, where N is the number of cells sequenced, when only one cell expresses the gene (bottom right population, $I(g) = \ln(5) \approx 1.61$largest). Note that the population heterogeneity $I(g)$ is independent of any decomposition of the cell populations into subpopulations (shown here as yellow and purple cells, for illustration). However, given any grouping of the cells into subpopulations, $I(g)$ can be formally decomposed as the sum of the heterogeneity explained by within and in-between subpopulations (see "Results" Section and Fig. 3). This decomposition, but not the overall value of $I(g)$, does depend on the chosen assignment of cells to subpopulations

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

where x_i^g is the fraction of transcripts of gene g expressed by cell i, for each $1 \leq i \leq N$, and $H(g)$ is the entropy of the expression of g in the population (see "Methods" Section for full details). For technical reasons, explained in the "Methods" Section, we will use natural logarithms in all calculations; $I(g)$ therefore has units of nats.

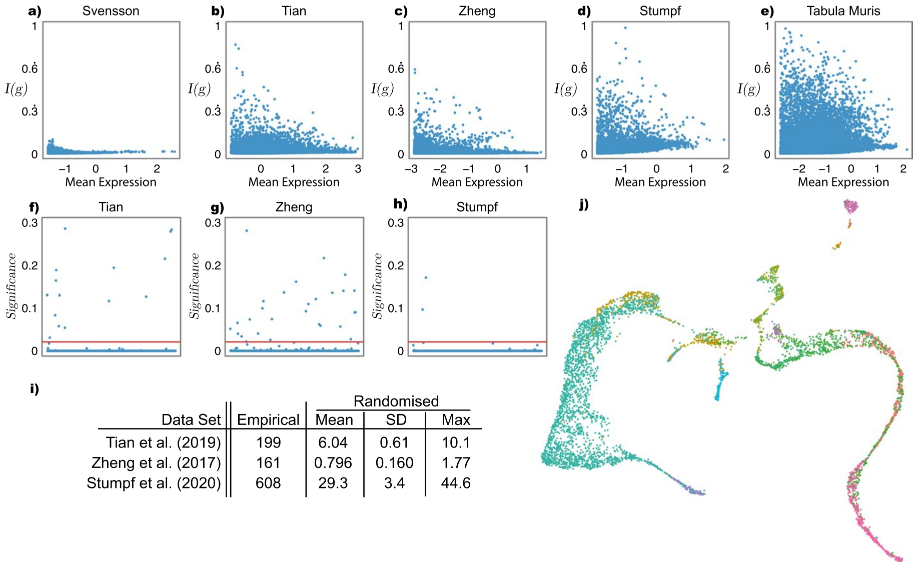
Intuitively, if the cell population is unstructured with respect to the expression of g (i.e. if the cells are approximately interchangeable with respect to the expression of g) then the assumption of homogeneity is correct and $I(g) \approx 0$. Conversely, the theoretical maximum for $I(g)$ is $\log(N)$, 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, so no *a priori* assumptions about expression patterns are required to calculate $I(g)$; (2) $I(g)$ is agnostic concerning missing readings or counts 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 practice, $I(g)$ 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 larger $I(g)$ will be. Thus, $I(g)$ is a parsimonious measure of expression heterogeneity that makes minimal assumptions concerning expression patterns and, therefore, imposes minimal technical requirements on data collection methodology or quality. As such, it can be used as the basis for numerous aspects of the sc-Seq pipeline, including feature selection and unsupervised clustering.

To validate these uses, we considered a series of single-cell RNA-sequencing benchmarking data sets: *Svensson*, a technical control [42]; *Tian*, a mixture of three cancerous cell lines [46]; *Zheng*, a set of FACS separated peripheral blood mononuclear cells; *Stumpf*, a sampling of cells from mouse bone marrow [41]; and the *Tabula Muris*, a mouse cell atlas with cells from twelve organs [44]. Fig. 2a–e shows plots of $I(g)$ for each gene profiled in these experiments. These plots confirm the intuition that the more distinct types present in a sample, and the more those cell types differ from one another with respect to their gene expression patterns, the greater the population heterogeneity will be, as measured by $I(g)$.

These results indicate that heterogeneity (as defined here) can be used for feature selection. For example, heterogeneously expressed genes can be identified by 1) ordering them in descending ordering by $I(g)$, and selecting the top n (where n is user-determined) or 2) finding those genes for which $I(g)$ greater than some user-defined threshold (note that $0 \leq I(g) \leq \log N$, with $I(g) \approx 0.7$ indicating the gene g is expressed in approximately half of cells).

As a measure for use in feature selection, $I(g)$ identifies gene sets largely distinct from those based on Highly Variable Gene selection (the mean overlap of selected genes—for listed data sets excluding *Svensson*—was 0.36, with the number of genes selected determined by `scran`, based on an false discovery rate threshold of 0.05). For instance, $I(g)$ identifies *Hbb-bs*, a marker of erythrocyte maturation, as informative in the Stumpf et al. [41] data set (high $I(g)$ value), whereas `scran` does not, instead attributing the observed



Data Set	Empirical	Randomised Mean	SD	Max
Tian et al. (2019)	199	6.04	0.61	10.1
Zheng et al. (2017)	161	0.796	0.160	1.77
Stumpf et al. (2020)	608	29.3	3.4	44.6

Fig. 2 Information-theoretic single-cell analysis. Recall that *I*(*g*) measures the heterogeneity of a cellular population with respect to the expression of *g*: $I(g) = 0$ when transcripts are expressed uniformly and increases as transcripts are expressed preferentially in a subset of cells, reaching a maximum $I(g) = \log(N)$, where *N* is the number of cells sequenced, when only one cell expresses the gene. **a**–**e** Plots of expression heterogeneity, *I*(*g*) (normalised by the theoretical maximum, $\log(N)$) against log mean expression for the bench-marking sc-Seq data sets described in the main text. In each panel, each point represents a gene profiled. The number of genes associated with large values of *I*(*g*) increases with the number of cell types present in the population profiled, showing *I*(*g*) as a valid measure of cell type diversity. Panel **a** shows data from a technical control [42] (number of cell types, $C = 1$), **b** a mixture of three cancerous cell lines [46] ($C = 3$), **c** FACS sorted immune cells [52] ($C = 4$), **d** a sample of mouse bone marrow [41] ($C = 14$), and **e** a multi-organ mouse cell atlas [44] ($C = 56$). **f**–**h** Biologically meaningful cell annotations are associated with high inter-cluster heterogeneity. Established cell annotations for the **f** *Tian*, **g** *Zheng* and **h** *Stumpf* data are associated with higher inter-cluster heterogeneity than expected by chance (i.e., in randomly permuted clusters; significance is assessed using a one-sided exact test with 10^4 permutations; y axes show $\log_{10}(p + 1)$). In all panels the red line shows $p < 0.05$, false discovery rate corrected for 500 trials [2, 8]. Genes below this threshold are significantly different gene expression patterns across the set of identified cell types. **i** Summary statistics for the total inter-cluster heterogeneity $H_S = \sum_g H_S(g)$ based on established empirical and randomly permuted cell annotations (10^4 random permutations in each case). These statistics show the strong association of high H_S with biologically meaningful groupings of cells. **j** A Uniform Manifold Approximation and Projection (UMAP) [29] plot of the top 500 genes by *I*(*g*) for the *Stumpf* data set; each point is a cell, coloured by its scEC cluster. This shows that *I*(*g*) is able to capture the continuous variation of developing cell types

variance to technical sources (see Additional file 1: Fig. S1)[27]. Critically, information-theoretic heterogeneity is more mathematically precise and computationally simpler to implement than variance-based approaches, making no distributional assumptions (e.g. use of the negative binomial model) and having no free parameters to fit. Furthermore, it may be easily modified to account for the presence of multiple homogeneous cell subpopulations, and thereby act as a measure of cluster quality.

Inter-cluster heterogeneity

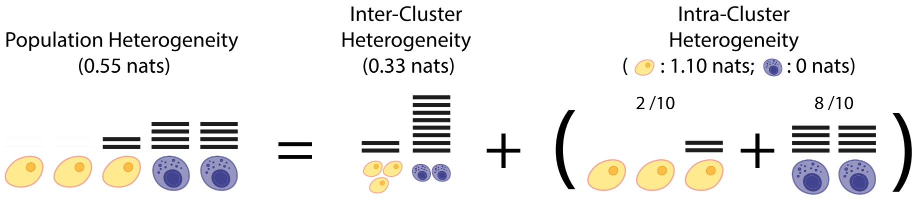
Let S be a discrete clustering of a population of cells—i.e. an assignment of the cells to a finite set of C non-intersecting sub-populations (also known as clusters). In the "Methods" Section, we show that $I(g)$ can be decomposed into two parts: one part that quantifies the extent to which each of the sub-populations defined by S deviate from homogeneity, which we call the *intra-cluster heterogeneity* and write as $h_S(g)$; and one part that quantifies how differently the gene is expressed, in aggregate, between subpopulations, which we call the *inter-cluster heterogeneity* and write as $H_S(g)$. Namely, 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 (within cluster) and global (between cluster) patterns of variation, for any proposed clustering. Full mathematical details of this decomposition, including formulae for $H_S(g)$ and $h_S(g)$, are provided in the "Methods" Section, and an illustration is given in Fig. 3. We also provide an R package for its calculation (see "Methods" Section).

This decomposition can be used to assess cluster quality. For a proposed cluster to meaningfully represent a cell type, it must be associated with differential expression of some subset of marker genes. Because $H_S(g)$ measures the extent to which the expression of gene g deviates from the homogeneous null model it is a simple measure of the extent to which g is differentially expressed between the clusters defined by S. We therefore expect that biologically meaningful clustering based on a marker gene g will result in a high value for $H_S(g)$ and (by necessity) a low value for $h_S(g)$.

This expectation is confirmed for the *Tian*, *Zheng* and *Stumpf* data sets for which we possess high-confidence *a priori* cell type annotations, derived either experimentally (for the *Tian* and *Zheng* data) or from expert annotation of a computational clustering (for the *Stumpf* data). In the set of genes most likely to be differentially expressed between clusters in these data sets (taken to be the top 500 genes by $I(g)$ in each case) the majority



$$I(g) = H_S(g) + (0.2 * h_{S_1}(g) + 0.8 * h_{S_2}(g))$$

$$0.55 = 0.33 + (0.2 * 1.10 + 0.8 * 0)$$

Fig. 3 Heterogeneity is additively decomposable. The heterogeneity of a population of cells (5 cells in this illustration) with respect to the expression of a gene g, $I(g)$, can be decomposed into inter- and intra-cluster heterogeneities for any proposed clustering, S (here, two subpopulations, or clusters, of 3 yellow and 2 purple cells). The inter-cluster heterogeneity $H_S(g)$ is determined by independently aggregating all transcripts (shown as horizontal lines) associated with each sub-population in S and then taking the KLD of the resulting distribution from the uniform distribution of the transcripts over C clusters. It measures the extent to which transcripts are uniformly assigned to clusters. The intra-cluster heterogeneity $h_S(g)$ is determined by taking the weighted sum (with respect to the number of transcripts on each subpopulation) of the heterogeneities of each of the constituent subpopulations, considered independently. It represents the average heterogeneity of the proposed clusters, accounting for disparities in number of transcripts assigned. In this toy example, the overall population heterogeneity of gene g, $I(g) = 0.55$, decomposes as the sum of the inter-cluster heterogeneity $H_S(g) = 0.33$, plus the intra-cluster heterogeneity $h_S(g) = 0.22$. The latter is obtained as the weighted sum (with respect to the number of transcripts in each cluster, here $2/10 = 0.2$ and $8/10 = 0.8$) of the heterogeneities on each subpopulation. Further details and formulae are provided in the "Methods" Section

are associated with substantially higher inter-cluster heterogeneity than expected by chance (i.e., in randomly permuted clusters). Indeed, in *Tian* 97.0%, in *Zheng* 94.4% and in *Stumpf* 99.6% of the tested genes had significantly higher inter-cluster heterogeneity than expected (one-sided exact test with 10^4 permutations, $p < 0.05$, false discovery rate corrected for 500 trials; see Fig. 2f–h for illustration) [2, 8].

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 high-throughput single-cell profiling methods is that they allow the simultaneous profiling of thousands of genes in large cell populations.

Our information-theoretic reasoning may be extended to a high-throughput 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 [37]. Thus, we can determine the total information explained by a given clustering S by evaluating the sum $H_S = \sum_g H_S(g)$ over all genes profiled. The total information explained by S is a simple, easily computed measure for cluster evaluation that favours grouping of cells into homogeneous (with respect to gene expression) sub-populations and is maximised at $\sum_g I(g)$ if and only if the proposed clustering divides the population into indivisible perfectly homogeneous sub-populations and thereby accounts for all the heterogeneity contained in the sc-Seq data set. If so, then S is the maximum entropy partition of the cell population into distinct classes, and may therefore be considered as the most parsimonious way of assigning cell identities.

To illustrate the association between total information explained, H_S, and cluster quality, we again considered the established annotations of the *Tian*, *Zheng* and *Stumpf* data sets. As expected, in all cases the observed value of H_S significantly exceeds that of all random label permutations (see Fig. 2i indicating that H_S is strongly associated with cluster quality and that these benchmark cell annotations are associated with homogeneous cell sub-populations. Notably, this association holds true independently of the method used to annotate cell identities (annotations were derived from genotypic information for the *Tian* data, surface protein expression for the *Zheng* data, and unsupervised clustering for the *Stumpf* data), indicating that H_S provides a methodology agnostic, parameter-free, means to quantify cluster quality.

Unsupervised clustering

Building on this empirical association, we investigate the clustering of cells that maximises H_S (and necessarily minimises h_S) as a reasonable approximation to (homogeneous) cell type. That is, this is the cell clustering that explains the most of the gene expression heterogeneity as inter-cluster variability.

We implement H_S-maximisation as the objective function of an unsupervised clustering method, which we call scEC (single-cell Entropic Clustering). We show that our method is comparable to the state-of-the-art Louvain method, and in doing so, validate the underlying definition of heterogeneity and homogeneous cell type.

Namely, we compare the clustering returned by scEC and the established annotations for the *Tian*, *Zheng* and *Stumpf* data sets. The resulting scEC clusterings agreed strongly with the established cell annotations, achieving adjusted Rand Indices (ARI) of 0.99 for the *Tian* annotations, 0.87 for the *Zheng* annotations, and 0.69 for the *Stumpf* annotations (ARIis a measure of similarity of a proposed clustering to a known clustering, Rand [33]). To benchmark these results we also repeated the analysis using the Louvain method, a leading single-cell clustering algorithm, which achieved ARIs of 0.99, 0.99 and 0.35 for the *Tian*, *Zheng*, and *Stumpf* data sets respectively [3, 9, 26, 40].

The scEC objective function, H_S, is naturally built by summation of gene-level measurements (see "Methods" Section), so can be used to directly identify the key drivers of a given clustering without relying on *post-hoc* differential gene expression testing. To demonstrate this feature, we compared the inter-cluster heterogeneity values (i.e., the values $H_S(g)$ for each of the genes profiled) for clusters identified by scEC and by established annotations for the *Tian* & *Stumpf* data sets (see Fig. 4).

We found that the values of $H_S(g)$ associated with scEC clusters are strongly positively correlated with those associated with established annotations for both the *Zheng* and *Stumpf* data sets (Pearson's correlation coefficient of 0.94 for *Zheng* and 0.99 for *Stumpf*; see Fig. 4). Moreover, the key marker genes in each data set were found to have significantly different expression levels across clusters, as measured by $H_S(g)$ (one-sided permutation test on clustering labels, 10^4 permutations, $p < 0.05$, false discovery rate corrected; marker genes for *Zheng*: *CD14*, *CD4*, *CD8* and *NCAM1*, Zheng et al. [52]; and *Stumpf*: *Cd34*, *Kitl*, *Spi1*, *Gata1* and *Pax5*, Stumpf et al. [41]).

This concordance indicates that scEC is generally able to recapitulate known cell clusters in an unbiased and parameter free way. However, despite this concordance, there were differences between scEC clusters and established annotations. To investigate these differences further we selected those genes for which the value of $H_S(g)$ differed substantially $(\Delta H_S(g) = |H_{scEC}(g) - H_{Known}(g)| > 0.1 \text{ nats})$ between scEC clusters and established annotations in each data set for further investigation.

In the *Zheng* data, we found that those genes with expression heterogeneity better explained by the scEC clustering were enriched for cell-cycle related genes (gene ontology enrichment, $p = 1.33 \times 10^{-8}$, false discovery rate corrected). Conversely, those genes with expression heterogeneity better explained by the established annotation were enriched for genes involved in the immune response ($p = 1.49 \times 10^{-3}$, false discovery rate corrected) [1, 2, 10].

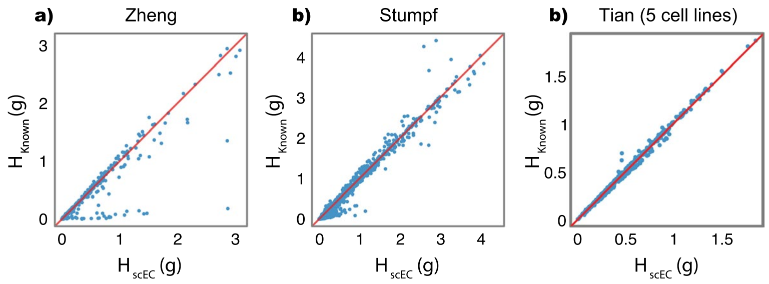


Fig. 4 Comparison of inter-cluster heterogeneity of scEC-generated clusters versus established annotations. Plots of $H_S(g)$ based on an scEC-generated clustering (*x*-axis) and established annotations (*y*-axis) for the **a** *Zheng* and **b** *Stumpf*, and **c** an alternative data set from *Tian* [41, 46, 52]. In all panels each point represents a gene profiled and the red line indicates $H_{scEC}(g) = H_{Known}(g)$. For the genes below the red lines, the scEC clustering is better than the prior annotation at explaining the gene expression heterogeneity as inter-cluster variability, and vice versa for the genes above the red line

In the *Stumpf* data set, we found that those genes with expression heterogeneity better explained by the scEC clustering were enriched for genes involved in erythroblast differentiation and homeostasis ($p = 1.80 \times 10^{-2}$ and p-value $= 6.17 \times 10^{-4}$ respectively, false discovery rate corrected). Conversely, those genes with expression heterogeneity better explained by the established annotation were enriched for genes involved in the immune response ($p = 2.57 \times 10^{-18}$, false discovery rate corrected). In this example, scEC generally identifies additional erythrocyte sub-populations while merging the cell sub-types of the neutrophil lineage (see Fig. 2j for illustration and Additional file 1: Table S1 for contingency table of cluster assignments).

Collectively, these differences reflect the preference of scEC for cellular annotations driven by biological processes involving larger tranches of genes. Depending on circumstance, this preference may be a benefit or a drawback. It is generally beneficial because it ensures that scEC annotations are robust to outliers (i.e. anomalous expression patterns of individual genes cannot easily distort scEC cluster assignments). Conversely, in circumstances in which key cellular functions are determined by a small number of genes, scEC may fail to disambiguate important cell populations from similar cell types. However, given the simplicity of the scEC methodology this issue can this issue can be relatively easily addressed, by adjusting the prior, as follows. In the standard formulation of scEC each gene contributes equally to the objective function $H_S = \sum_g H_S(g)$. This corresponds to assuming a uniform prior. However, alternatives may be easily considered that take into account different sources of prior knowledge. In general, each gene can contribute in a weighted way to the overall inter-cluster heterogeneity. In this case the objective becomes $H_S = \sum_g w(g)H_S(g)$, where $w(g)$ is the weight of gene g. These weights can encode known biology—for example, by up-weighting known drivers of cellular identities, down-weighting genes associated with housekeeping processes and/or the cell cycle, or by weighting genes in inverse proportion to the size (i.e. the number of associated genes) of their leading ontology term.

While incorporating such priors into scEC may be beneficial, defining them is a challenge and potentially introduces a source of bias into the resulting clustering and so should be approached with care. An alternative approach—which encodes prior information in a data-led, rather than algorithmic way—is to use a reference data set, such as a cell atlas, to benchmark cell annotations (these atlases can serve as *reference transcriptomes*, in analogy to reference genomes).

Doing so leads to a semi-supervised version of scEC, wherein we cluster a novel test data set based on the known clustering of a given reference data set. We detail the mathematics of semi-supervised clustering in the "Methods" Section, but, in short, the extension to the semi-supervised setting requires no additional mathematical machinery. Such mathematical simplicity affirms information theory as providing a mathematically precise and biological intuitive framework for cellular clustering.

As an example, we cluster the *Tian* data set based on a distinct data set containing the same three cell lines (plus two additional cell lines) [46]. The semi-supervised clustering strongly agreed with the established cell annotation, achieving an ARI of 0.90. Notably, the ARI was reduced compared with the unsupervised version due to the assignment of $\sim 6\%$ of cells to cell types present only in the reference data set, a disadvantage of semi-supervised (and supervised) methods.

Benchmarking and imputation

We benchmark scEC and state-of-the-art methods against a series of data sets consisting of cells sampled from three or five distinct cancerous cell lines and sequenced using various technologies (CEL-Seq2, Sort-seq, 10x, Drop-seq) [15, 28, 31, 46]. We compare each method's clustering against ground truth for each data set (by adjusted rand index), finding that scEC performs subpar to the best of existing methods [13, 21, 27, 34, 40]. Investigating the gene-wise contributions via $H_S(g)$ reveals that scEC-derived clusterings diverge from the ground truth due to genes expressed in small numbers of cells, e.g., SPINK6 (difference in H_{scEC} from ground truth of 0.23 nats; expressed in 1.6% of cells in the five-cell line, 10X data set; see Fig. 4c).

To protect against low cell-count genes, we adopt a data-diffusion step, imputing the expression value of each cell as a weighted average of itself and a small subset of cells with similar expression patterns (i.e., diffusion of observed expression values over a shared nearest-neighbours graph derived from the top 500 genes by $I(g)$, see "Methods" Section) [49]. The smoothing layer shares information between genes, lessening scEC's sensitivity to the expression pattern of individual genes; this raises the performance of scEC clustering to be on par with the state-of-the-art (see Fig. 5a).

We further benchmark scEC (with and without the imputation step) as a feature selection tool, evaluating the ability of $I(g)$ to *a priori* identify differentially expressed genes, assuming differential expression to be the ground truth of feature selection (differentially expressed genes are identified per data set via Wilcoxon Rank Sum testing, selecting those genes with an FDR-adjusted p-value < 0.05 in at least one cluster; Stuart et al. [40]). We find that, without smoothing, scEC is comparable to the state-of-the-art; with smoothing, scEC notably improves, substantially bettering the state-of-the-art in selecting differentially expressed genes from the benchmark data sets (see Fig. 5b). Notably, relatively few genes remain heterogeneous after the imputation step, so the imputed $I(g)$

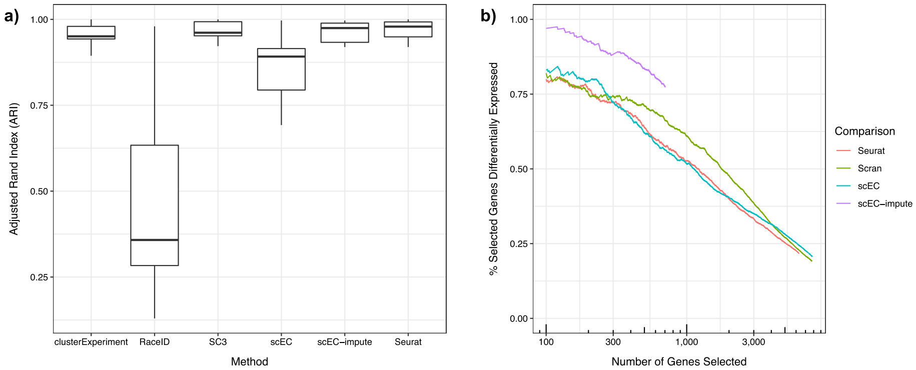


Fig. 5 Benchmarking of scEC performance in unsupervised clustering and feature selection. **a** Adjusted Rand Index of clusterings produced by specified methods against known ground truth for seven data sets, each consisting of three or five cancerous lines sequenced on different platforms. With an additional imputation step, scEC performs on par with other methods. **b** The percent of the top *N* genes by different feature selection metrics that are differentially expressed. Data set is Sc-seq from three cancerous cell lines sequenced by Drop-seq (with 2005 differential expressed genes identified from non-parametric testing for each cell line versus the remaining; Wilcox test, fasle discovery rate corrected p-value < 0.05). The greater ability of scEC-impute to *a priori* select differentially expressed genes is repeated across each benchmark data set, see Additional file 1: Fig. S2. Note that the imputation step in scEC-impute assigns many genes a heterogeneity $I(g)$ of zero, resulting in a low cut-off on total number of selectable genes

identifies a much smaller number of genes than other methods, where an arbitrary cut-off in selected gene number is required.

Conclusion

Traditional unsupervised clustering methods for single cell data are based on the, often implicit, assumption that cell types correspond to regions of high probability density in the joint gene expression distribution [11]. Although useful, this is not a specifically biological assumption (similar assumptions form the basis of numerous clustering algorithms in other different disciplines) and does not have a clear biological basis. Indeed, this assumption does not naturally accord with biological intuition that distinct cell populations should be 'homogeneous' and cells of the same type should be functionally interchangeable. Here, we have taken a different approach that seeks to encode biological intuition about population homogeneity in a mathematical formulation, drawing on tools from information theory.

We have proposed a formal measure of gene expression heterogeneity and shown that this measure captures biologically relevant heterogeneity arising from differential expression between established cell types. We formalised the additive decomposition of heterogeneity with respect to a given grouping of cells (in "Methods" Section), and tested the association between high inter-cluster heterogeneity and grouping by cell type. Finally, we used this measure as a basis of a biologically-motivated clustering procedure, which we call single-cell entropic clustering (scEC), which identifies homogeneous cell types. The scEC method is free of tunable hyper-parameters and its performance is comparable to a state-of-the-art clustering algorithm on a series of benchmark sc-Seq data sets, suggesting that this underlying biological basis is justified.

While our method represents a mathematically rigorous definition of heterogeneity, the stochasticity inherent in sc-Seq data does limit our resolution. For instance, nearly every gene is associated with some non-zero level of heterogeneity and, over thousands of genes, even low levels of stochastically-induced heterogeneity will accumulate (a version of the so-called 'curse of dimensionality'). This limits our ability to directly interpret the genome-wide heterogeneity (e.g. the total heterogeneity of the technical control *Svensson* exceeds that of *Tian*, both absolute and per gene, although the trend reverses when excluding those genes with fewer than 100 total transcripts).

Aside from these technical limitations, our approach demonstrates the benefit of a biologically grounded, mathematically formal approach to understanding sc-Seq data. As these data sets grow in complexity and size, there will be a greater need for biological interpretability and mathematical rigour in analysis. We suggest information theory as the appropriate mathematical language for describing gene expression and its heterogeneity. Indeed, we anticipate that information theory will become an established part of the quantitative biologist's toolkit.

Methods

Data collection

Count matrices for each data set were downloaded from their respective repositories (see Availability of Data and Materials). For the *Zheng* and *Tabula Muris* data sets, the count matrices of the various cell types/tissues were concatenated using the Matrix (v1.3-2) package in R (v4.0.3) [32].

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. For the *Zheng* data set, FACS identities were taken from the repository metadata, with the CD4+ Helper T-cells, CD4+/CD25+ Regulatory T-cells, CD4+/CD45RA+/CD25- Naïve T cells, CD4+/CD45RO+ Memory T-cells, CD8+ Cytotoxic T cells and CD8+/CD45RA+ Naïve Cytotoxic T-cells merged into the single identity of T-cells.

Normalisation

Since sc-Seq data is noisy, rather than working with the experimentally observed proportions of transcripts assigned to each cell, we instead adopt a Bayesian approach to estimate the cellular frequencies using the James–Stein estimator, as explained below, and use this estimator in all subsequent calculations.

We write m_i^g for the number of transcripts of gene g associated with cell i in a population of N cells. Let $\sum_{i=1}^N m_i^g = M^g$ be the total number of transcripts of gene g, and $p_i^g = m_i^g/M^g$ the fraction of transcripts of gene g expressed by cell i, for each $1 \leq i \leq N$. Thus, $\sum_{i=1}^N p_i^g = 1$ and we can think of p_i^g as the probability that a randomly selected transcript of gene g is associated with cell i. To take under-sampling into account, we adjust the observed probabilities $p_i = p_i^g$ as follows (let us drop the superscript g for simplicity). We define $X = X^g$ as the discrete random variable on the set $\{1, 2, \ldots N\}$ with probabilities

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

where λ' is the shrinkage intensity given by

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

which is the James–Stein estimator, with the uniform distribution as the shrinkage target [16]. Shrinkage approaches deal with under-sampling and so are able to correct for the substantial sparsity observed in single-cell RNA-sequencing data. The shrinkage probabilities x_i given in Eq. 3 are a compromise between the observed probabilities p_i, which are unbiased but with a high variance, and the target uniform probabilities $1/N$, which are biased but with low variance [6, 16].

Feature selection

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 $0 \cdot \log 0 = 0$. For the logarithm, we take base e for ease in differentiation (see later) rather than the usual base 2 (recall that $\log_e(x) = \log_2(e)\log_2(x)$ for all $x > 0$), and thus $H(X)$, and $I(g)$ below, are measured in nats.

The Shannon entropy is a measure of the uncertainty in the outcomes of the random variable X. It has a minimum value of zero, when $x_i = 1$ for some i (e.g. if the gene g is expressed in only one cell of the population) and has a maximum value of $\log(N)$, when $x_i = 1/N$ for all i (e.g. if 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 (note that the association of high entropy with homogeneity is counter to the usual intuition: it occurs because we are concerned with the entropy of the generative process of assigning transcripts to cells, rather than transcript count distribution in the cell population, as is more usual. A high entropy transcript assignment process gives rise to a sharp transcript distribution, hence its association with homogeneity). By contrast, the quantity $\log(N) - H(X)$ also ranges between zero and $\log(N)$, yet is minimised 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

$$\begin{aligned} I(g) =& \log(N) - H(X) = \sum_{i=1}^{N} x_i \log(N) + \sum_{i=1}^{N} x_i \log(x_i) = \sum_{i=1}^{N} x_i \log(Nx_i) \\ =& \sum_{i=1}^{N} x_i \log\left(\frac{x_i}{1/N}\right). \end{aligned} \tag{6}$$

The Kullback–Leibler divergence (KLD), or relative entropy, is a measure of the information encoded in the difference between two probability distributions [22]. The relative entropy of a discrete probability distribution $p_1, \ldots, p_N$ from a discrete probability distribution $q_1, \ldots, q_N$ is, by definition,

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

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 and Eq. 6, it is clear that our measure of heterogeneity is simply the relative entropy of the observed expression distribution from the uniform distribution U. Thus,

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

where U denotes to the uniform distribution on the set $\{1, 2, \ldots N\}$ with probabilities $u_i = 1/N$ for all $1 \leq i \leq N$.

Cluster-level heterogeneity

A crucial property of the relative entropy is that it is additively decomposable with respect to arbitrary groupings [38]. Informally, this means that if we have a clustering of the cells into disjoint groups, then $I(g)$ can be reconstructed from inter-cluster and intra-cluster heterogeneities. Next, we formalise this additive decomposition and give a self-contained derivation (cf. Theil [45]).

Consider a clustering $S = \{S_1, \ldots, S_C\}$ 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 the shrinkage estimator. That is,

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

with x_i defined by Eq. 3. 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 $i \in S_k$. Note that

$$\sum_{i \in X_k} z_i = \sum_{i \in X_k} \frac{x_i}{y_k} = y_k \sum_{i \in X_k} x_i = 1 \tag{10}$$

so Z_k is a random variable on the set (cluster) S_k.

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

$$I(g) = \log(N) - \sum_{i=1}^{N} x_i \log\left(\frac{1}{x_i}\right), \tag{11}$$

$$= \log(N) - \sum_{k=1}^{C} \sum_{i \in S_k} x_i \log\left(\frac{1}{x_i}\right) \tag{12}$$

$$= \log(N) - \sum_{k=1}^{C} y_k \sum_{i \in S_k} \frac{x_i}{y_k} \left(\log\left(\frac{1}{x_i / y_k}\right) + \log\left(\frac{1}{y_k}\right)\right), \tag{13}$$

$$= \log(N) - \sum_{k=1}^{C} y_k \underbrace{\sum_{i \in S_k} \frac{x_i}{y_k} \log\left(\frac{1}{x_i / y_k}\right)}_{H(Z_k)} - \sum_{k=1}^{C} y_k \sum_{i \in S_k} \frac{x_i}{y_k} \log\left(\frac{1}{y_k}\right), \tag{14}$$

$$= \log(N) - \sum_{k=1}^{C} y_k H(Z_k) - \underbrace{\sum_{k=1}^{C} \log\left(\frac{1}{y_k}\right) \overbrace{\sum_{i \in S_k} x_i}^{y_k}}_{H(Y)}, \tag{15}$$

$$= \log(N) - \sum_{k=1}^{C} y_k H(Z_k) - H(Y), \tag{16}$$

$$= \underbrace{\log(N) - H(Y) - \sum_{k=1}^{C} y_k \log(N_k)}_{A} + \underbrace{\sum_{k=1}^{C} y_k \log(N_k) - \sum_{k=1}^{M} y_k H(Z_k)}_{B}. \tag{17}$$

Expression A may be rewritten as:

$$A = \log(N) - H(Y) - \sum_{k=1}^{C} y_k \log(N_k), \tag{18}$$

$$= \sum_{k=1}^{C} y_k \log(N) - \sum_{k=1}^{C} y_k \log\left(\frac{1}{y_k}\right) - \sum_{k=1}^{C} y_k \log(N_k), \tag{19}$$

$$= \sum_{k=1}^{C} y_k \log\left(\frac{y_k}{N_k/N}\right), \tag{20}$$

$$= D(Y||U_{group}). \tag{21}$$

This is the relative entropy 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 the cluster S_k, it measures the deviation from the assumption that the clusters are homogeneous in their expression of g (i.e. each cluster expresses g at the same level). Since it is a measure of the extent to which the population deviates from homogeneity between clusters, we will term this contribution the *inter-cluster heterogeneity* of g with respect to the clustering S, denoted H_S. Informally, it is a measure of the extent to which the gene g is differentially expressed between clusters.

Expression B may be rewritten as:

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

$$= \sum_{k=1}^{C} y_k \big(\log(N_k) - H(Z_k)\big), \tag{23}$$

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

$$= \sum_{k=1}^{C} y_k \sum_{i \in S_k} \frac{x_i}{y_k} \log\left(\frac{x_i/y_k}{1/N_k}\right), \tag{25}$$

$$= \sum_{k=1}^{C} y_k D(Z_k||U_k). \tag{26}$$

This is the weighted sum of the relative entropies of the empirical distributions 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 deviation from 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 homogeneity within clusters, we will term this contribution the *intra-cluster 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 local and global properties of the expression distribution of g with respect to a given clustering $S = \{S_1, \ldots, S_C\}$:

$$I(g) = A + B = H_S(g) + h_S(g). \tag{27}$$

The relative entropy 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 cluster (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 cluster (i.e. $C = 1$) then $Y = U_S$ and thus $H_S(g) = 0$. In this case, the population heterogeneity is equivalent to the intra-cluster heterogeneity of the trivial clustering.

Unsupervised clustering

The above measures of heterogeneity can be easily extended from one gene to many. In this case, the homogeneous null model is obtained by assuming that each gene is expressed homogeneously and independently. Because entropy from independent sources is additive [7], the total heterogeneity of a single-cell RNA-sequencing data set under a given clustering is given by the sum:

$$I = \sum_g I(g) = \sum_g H_S(g) + \sum_g h_S(g) = H_S + h_S. \tag{28}$$

Here $H_S = \sum_g H_S(g)$ and $h_S = \sum_g h_S(g)$ represent the total inter-cluster and intra-cluster heterogeneity of the data with respect to a clustering $S = \{S_1, \ldots, S_C\}$.

Naturally, we want to identify a clustering (e.g. an assignment of cells to types) that produces maximally homogeneous sub-groups and thereby explains most of the expression heterogeneity for most genes in terms of inter-cluster heterogeneity. Mathematically, this means we want to find a clustering S that maximises H_S or, equivalently, minimises h_S. A brute force approach is not feasible, as the number of partitions of a set with n elements into two or more subsets grows exponentially with n.

To approach the problem we therefore converted the problem from one of assigning discrete identities (i.e., a hard clustering problem) to one of assigning continuous identities (i.e., a fuzzy clustering problem). The advantage of this approach is that it defines the problem in terms of a continuous, differentiable objective function which can be approached with more efficient optimisation routines.

Fuzzy clustering

We adopt a fuzzy conception of clustering in which cells are assigned to C (possibly) overlapping, fuzzy clusters, $S_1, \ldots, S_C$. Each cell i has C corresponding membership functions

$\mu_{ik} \in [0,1]$ for $k = 1, 2, \ldots, C$, which assess the probability that cell i belongs to cluster S_k. Thus,

$$\sum_{k=1}^{C} \mu_{ik} = 1 \text{ with } \mu_{ik} \geq 0, \tag{29}$$

Our information-theoretic framework can be adapted to this fuzzy setting. Population heterogeneity, *I*(*g*), is independent of the chosen clustering, so it remains unaffected by the adoption of fuzzy clusters, i.e. the total gene expression heterogeneity is unaffected by the choice of discrete or fuzzy cluster memberships. For the calculations of inter-cluster and intra-cluster heterogeneities, we extend the discrete random variables Y^g and Z_k^g to the fuzzy setting, where Y^g now measures the expression distribution of the gene *g* across the *C* fuzzy clusters, and Z_k^g measures the expression distribution of the gene *g* within fuzzy cluster S_k, as follows.

We define the discrete random variable Y^g on the set of fuzzy clusters $S = \{S_1, \ldots, S_C\}$ with probabilities y_k^g given by

$$y_k^g = \sum_{i=1}^{N} \mu_{ik}\, x_i^g. \tag{30}$$

We also define, for each $S_k \in S$, a discrete random variable Z_k^g on the set of cells, $i = 1, \ldots, N$, with probabilities z_{ik}^g given by,

$$z_{ik}^g = \mu_{ik} \frac{x_i^g}{y_k^g}. \tag{31}$$

We can rewrite *I*(*g*) (which is independent of the clustering) in terms of Y^g and Z_k^g (which depend on the clustering) as

$$I(g) = \log(N) - H(Y^g) - \sum_{k=1}^{C} y_k^g H(Z_k^g), \tag{32}$$

where,

$$H(Y^g) = \sum_{k=1}^{C} y_k^g \log\left(\frac{1}{y_k^g}\right) \tag{33}$$

and

$$H(Z_k^g) = \sum_{i=1}^{N} \mu_{ik} \frac{x_i^g}{y_k^g} \log\left(\frac{1}{x_i^g / y_k^g}\right). \tag{34}$$

Note that Eq. 33 is the entropy of the random variable Y^g, however $H(Z_k^g)$ is not quite the entropy of Z_k^g (there is a missing $1/\mu_{ik}$ factor inside the logarithm).

From Eq. 32, we can obtain a second decomposition,

$$I(g) = H(g) + h(g) \tag{35}$$

where

$$H_S(g) = \sum_{k=1}^{C} y_k^g \log\left(\frac{y_k^g}{N_k/N}\right), \tag{36}$$

$$h_S(g) = \sum_{k=1}^{C} y_k^g \sum_{i=1}^{N} \frac{\mu_{ik}\, x_i^g}{y_k^g} \log\left(\frac{x_i^g / y_k^g}{1/N_k}\right), \tag{37}$$

$$N_k = \sum_{i=1}^{N} \mu_{ik}. \tag{38}$$

Here $H(g)$ and $h(g)$ represent the inter-cluster heterogeneity and intra-cluster heterogeneity of gene g with respect to the fuzzy clustering S. Equation 36 is also the KLD of the distribution of Y^g from the uniform distribution U_group given by $u_k = N_k/N$ for $k = 1, 2, \ldots, C$ (note that $\sum_{k=1}^{C} u_k = 1$). Similarly, Eq. 37 is the KLD of the distribution of Z^g from the uniform distribution U_k given by $p_i = \mu_{ik}/N_k$ for $i = 1, 2, \ldots, N$.

Either $H_S = \sum_g H_S(g)$ or $h_S = \sum_g h_S(g)$ can serve as our objective function, *f*, for optimisation: optimising either quantity simultaneously identifies the clustering with the greatest heterogeneity between clusters (i.e. the greatest differential gene expression between clusters) and the least heterogeneity within clusters (i.e. in which the cells within each cluster are closest to being interchangeable with respect to their patterns of gene expression).

Choosing inter-cluster heterogeneity H_S, we solve the optimisation problem

$$\max_S H_S, \tag{39}$$

where

$$H_S = \sum_{g=1}^{G} H_S(g) = \sum_{g=1}^{G} \sum_{k=1}^{C} y_k^g \log\left(\frac{y_k^g}{N_k/N}\right), \tag{40}$$

using the limited-memory, box constrained BFGS (*L-BFGS-B*) optimisation algorithm from the `Python3` (v3.8.2) package `SciPy` (v1.5.3) [5, 50, 51, 53].

The *L-BFGS-B* algorithm is an efficient non-linear local optimisation method developed for solving large, dense problems such as this [53] but its speed and the quality of the solution produced depends on the availability of an explicit formulation of the gradient of the objective function. Therefore, we derive the required gradient by partially differentiating the total inter-cluster heterogeneity, H_S, with respect to fuzzy cluster membership, i.e. the $N \cdot C$ membership functions μ_{rs} $(1 \leq r \leq N, 1 \leq s \leq C)$.

Beginning by differentiating individual elements of *f* with respect to μ_{rs}, from Eqs. 30 and 38, we have

$$\frac{\partial}{\partial \mu_{rs}}\left(y_k^g\right) = \begin{cases} x_r^g & k = s \\ 0 & k \neq s \end{cases} \tag{41}$$

$$\frac{\partial}{\partial \mu_{rs}}(N_k) = \begin{cases} 1 & k = s \\ 0 & k \neq s \end{cases} \tag{42}$$

Using the product, chain and quotient rules, we can differentiate the g-summand, call it $f^g\left(= H_S(g)\right)$, of the objective function f:

$$\frac{\partial f^g}{\partial \mu_{rs}} = \frac{\partial}{\partial \mu_{rs}}\left(\sum_{k=1}^{C} y_k^g \log\left(\frac{y_k^g}{N_k/N}\right)\right) \tag{43}$$

$$= \sum_{k=1}^{C} \frac{\partial}{\partial \mu_{rs}}\left(y_k^g\right) \log\left(\frac{y_k^g}{N_k/N}\right) + \sum_{k=1}^{C} y_k^g \frac{\partial}{\partial \mu_{rs}}\left(\log\left(\frac{y_k^g}{N_k/N}\right)\right) \tag{44}$$

$$= x_r^g \log\left(\frac{y_s^g}{N_s/N}\right) + \sum_{k=1}^{C} y_k^g \frac{N_k/N}{y_k^g} \frac{\partial}{\partial \mu_{rs}}\left(\frac{y_k^g}{N_k/N}\right) \tag{45}$$

$$= x_r^g \log\left(\frac{y_s^g}{N_s/N}\right) + \sum_{k=1}^{C} \frac{N_k}{N} N \frac{\partial}{\partial \mu_{rs}}\left(\frac{y_k^g}{N_k}\right) \tag{46}$$

$$= x_r^g \log\left(\frac{y_s^g}{N_s/N}\right) + \sum_{k=1}^{C} N_k \frac{1}{N_k^2}\left(\frac{\partial}{\partial \mu_{rs}}\left(y_k^g\right) N_k - y_k^g \frac{\partial}{\partial \mu_{rs}}(N_k)\right) \tag{47}$$

$$= x_r^g \log\left(\frac{y_s^g}{N_s/N}\right) + \frac{1}{N_s}\left(x_r^g N_s - y_s^g\right) \tag{48}$$

$$= x_r^g \left(\log\left(\frac{y_s^g}{N_s}\right) + \log(N) + 1\right) - \frac{y_s^g}{N_s}. \tag{49}$$

The membership functions μ_{rs} are not independent, since $\sum_{k=1}^{C} \mu_{ik} = 1$ for all i, by Eq. 29. To incorporate this constraint, and the constraint $\mu_{ik} \geq 0$ for all i, k, we introduce variables w_{ik} given by

$$\mu_{ik} = \frac{e^{w_{ik}}}{\sum_{l=1}^{C} e^{w_{il}}}, \tag{50}$$

and make the objective function f a function of the w_{ik}. By the chain rule,

$$\frac{\partial f}{\partial w_{ij}} = \sum_{r=1}^{N} \sum_{s=1}^{C} \frac{\partial f}{\partial \mu_{rs}} \frac{\partial \mu_{rs}}{\partial w_{ij}}. \tag{51}$$

Determining $\frac{\partial \mu_{rs}}{\partial w_{ij}}$, let us write $M_i = \sum_{l=1}^{C} e^{w_{il}}$; then, Eq. 50 can be written as $\mu_{ik} = e^{w_{ik}}/M_i$. By the quotient rule,

$$\frac{\partial \mu_{rs}}{\partial w_{ij}} = \frac{\partial}{\partial w_{ij}}\left(\frac{e^{w_{rs}}}{M_r}\right) = \begin{cases} \frac{e^{w_{ij}}M_i - e^{2w_{ij}}}{M_i^2} & \text{if } r = i \text{ and } s = j, \\ \frac{-e^{w_{is}}e^{w_{ij}}}{M_i^2} & \text{if } r = i \text{ and } s \neq j, \\ 0 & \text{if } r \neq i. \end{cases} \tag{52}$$

Numerical optimisation

The objective function H_S, as formulated in Eq. 40, was optimised with respect to cluster memberships, μ_{ik} using the *L-BFGS-B* optimisation algorithm from the `Python3` (v3.8.2) package `SciPy` (v1.5.3) [5, 50, 51], with random initial starting points (initial cluster memberships were defined by random sampling of a uniform distribution centred on zero). The gradient function, as formulated in Eq. 51, was supplied to the optimisation routine.

Because initial conditions are randomly generated, and the *L-BFGS-B* algorithm only identifies local optima, the specific clustering found by the algorithm may depend on the choice of initial vector [5, 53]. To make the implementation robust a multi-start process is adopted, in which the optimisation is repeated multiple times with different initial vectors, and the clustering S with the greatest inter-cluster heterogeneity H_S is chosen.

Permutation testing

Annotations for the *Tian*, *Zheng* and *Stumpf* data were taken from the respective repositories' metadata. Random clusterings were generated by randomly permuting the established annotations 10,000 times each. Comparison of true and shuffled annotations was formulated as an exact one-side hypothesis test, generating exact *p*-values [8]. We control the false discovery rate arising from multiple testing for each data set using the `R` function *p.adjust* [2].

Clustering comparison

We compare our clustering results to two standard clustering methods: `Seurat` (v3) clustering (Louvain community detection), and UMAP projection, both with default parameters, with the exception of the resolution, as described in Stuart et al. [40] [3, 29]. The Adjusted Rand Index (ARI) was calculated for each data set using the associated function in the `R` package `mclust` (v5.4.7) [33, 36].

Semi-supervised classification

Consider a pair of count matrices, $T = \{T_{ref}, T_{test}\}$, where one count matrix, which we term the reference, T_{ref}, has a known cluster structure (and therefore each cell has a unique, discrete cell type classification). The goal of reference mapping is to classify cells in the test matrix, T_{test}, based on cellular classifications of the reference matrix.

We normalise each count matrix separately as described in **Normalisation**, using the James–Stein-type estimator so that $\sum_{i \in T_k} x_i^g = 1$ [16]. We assume that the same set of genes are profiled in T_{ref} and T_{test} and derive the normalised combined count matrix, X_{mix}, via the weighted concatenation:

$$X_{mix} = \left[\frac{N_{ref}}{N}X_{ref}, \frac{N_{test}}{N}X_{test}\right], \tag{53}$$

where N_{ref} is the number of cells in the reference data set, N_{test} is the number of cells in the unclassified data set, and N is the total number of cells across both data sets.

As in the unsupervised setting, we cluster the combined data set, X_{mix}, by maximising H_S of the combined data set with respect to a fuzzy clustering S, see Eq. 56. However, unlike in the unsupervised setting, a subset of cellular cluster memberships are known *a priori*. Let $R = R_1, \ldots, R_C$ be the discrete clustering of the reference data set, with cluster sizes $n_1, \ldots, n_C$, where $\sum_{k=1}^{C} n_k = N_{ref}$. As the cluster identities of cells originating from the reference data set are fixed, each cluster in the reference data set constitutes a subset of a cluster in the mixed data set, $R_k \in S_k$, where $S = S_1, \ldots, S_C$ is the clustering of the mixed data set. Note that number of clusters in the reference and in the combined data set is the same, i.e. C.

Based on the co-mixing of known and unknown cellular identities, y_k^g and N_k of the combined data set X_{mix} are given by

$$y_k^g = \sum_{i=1}^{N_{test}} \mu_{ik}\, x_i^g + \sum_{i \in R_k} x_i^g, \text{ and} \tag{54}$$

$$N_k = \sum_{i=1}^{N_{test}} \mu_{ik} + n_k. \tag{55}$$

Based on these formulations of y_k^g and N_k, the definition of H_S follows as before,

$$H_S = \sum_{g=1}^{G} H_S(g) = \sum_{g=1}^{G}\sum_{k=1}^{C} y_k^g \log\left(\frac{y_k^g}{N_k/N}\right). \tag{56}$$

Similarly, the derivative of H_S follows as before, except that the membership function, μ_{rs}, is only defined with respect to cells of the test data set, $1 \leq r \leq N_{test}$ & $1 \leq i \leq N_{test}$,

$$\frac{\partial f}{\partial w_{ij}} = \sum_{r=1}^{N_{test}}\sum_{s=1}^{C} \frac{\partial f}{\partial \mu_{rs}}\frac{\partial \mu_{rs}}{\partial w_{ij}} = \sum_{r=1}^{N_{test}}\sum_{s=1}^{C}\sum_{g=1}^{G}\left[x_r^g\left(\log\left(\frac{y_s^g}{N_s}\right) + \log(N) + 1\right) - \frac{y_s^g}{N_s}\right]\frac{\partial \mu_{rs}}{\partial w_{ij}}. \tag{57}$$

Numerical optimisation follows as in the unsupervised setting, using the *L-BFGS-B* algorithm, except that the known cellular identities of the reference mean that only a single run of the algorithm is required, i.e. no multiple starts.

Imputation

We rely on a cell-by-cell adjacency matrix for imputation, encoding which cells have similar gene expression profiles. Specifically, we use the shared-nearest-neighbours matrix produced by Seurat, using default settings [40].

We begin with the expression matrix X, keeping only the top G genes by $I(g)$ (default is 500). Then, briefly, the Seurat method follows the following steps: data is scaled and normalised using a variance stabilising transform; the transformed data then undergoes principal components, and a cell-cell Euclidean distance matrix is calculated on the first 10 principal components. Next, a k-nearest-neighbours graph is constructed, where k is by default 20; a shared nearest-neighbours graph is constructed by taking the Jaccard Index of the overlap in neighbourhoods of each pair of cells [18].

From this adjacency matrix, A, we compute the imputed data via a one-hop smoothing operator:

$$Q_{ij} = \frac{A_{ij}}{\sum_j A_{ij}} \tag{58}$$

$$\widetilde{X}_i^g = \sum_j Q_{ij} X_j^g, \tag{59}$$

where Q is the row-normalised stochastic matrix, i.e. $\sum_j Q_{ij} = 1 \; \forall \; i \in 1, \ldots, N$, $X^g = (X_j^g)$ is the g-column of X, and $\widetilde{X}^g = (\widetilde{X}_i^g)$ are the imputed values for gene g [30]. The imputed expression value of gene g in cell i is a weighted mean of observed expression values in cell i and its neighbourhood in A. This is an example of a data diffusion imputation, first applied to Sc-seq data in Van Dijk et al. [49].

The vector $\widetilde{X}^g$ can then be used in place of X^g in any of the above described work. In the main text, we select the top 500 genes by $I(g)$ based on $\widetilde{X}^g$, before clustering on $\widetilde{X}^g$. Accordingly, use of the imputed expression matrix requires two rounds of feature selection: first in the construction of the adjacency matrix, A, then for the use of $\widetilde{X}^g$ in clustering.

For the benchmarking of the imputation, we recapitulated the set of data sets and methods tested in Tian et al. [46]. However, we were unable to rerun RCA [23].

Abbreviations

ARI	Adjusted rand index
KLD	Kullback–Leibler divergence
scEC	Single-cell entropic clustering
scSeq	Single-cell sequencing

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-023-05424-8.

Additional file 1. Two additional figures and one additional table: Expression of *Hbb-bs* in the *Stumpf* data set (Supplementary Figure 1); Feature selection benchmarking of scEC and scEC-impute (Supplementary Figure 2); and Contingency table between cell annotations provided in the *Stumpf* data set and scEC labels.

Acknowledgements

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

Author contributions

Conceptualization, MJC, RSG, BDM; Software, MJC; Investigation, MJC, JF, RJSG and BDM; Writing—original draft, MJC, RSG and BDM; Writing—review and editing, MJC, JF, RSG and BDM; Visualization, MJC; Supervision, RSG and BDM. All authors have seen and approved the manuscript.

Funding

RSG and BDM were supported by The Alan Turing Institute under the EPSRC Grant EP/N510129/1. RSG was supported by the Alan Turing Institute—Roche strategic partnership under the project 'Structured missingness in heterogeneous data'.

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 sets were downloaded from https://github.com/LuyiTian/sc_mixology The ***Zheng*** data was downloaded as set of files "Gene / cell matrix (filtered)" in section "Single Cell 3' Paper: Zheng et al. 2017" from https://support.

10xgenomics.com/single-cell-gene-expression/datasets The ***Stumpf*** data was downloaded as file "RData.zip" from http://dx.doi.org/10.17632/csvm3kpkxd.1 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

An R package for the implementation of the described methods is available on github https://github.com/mjcasy/scEC

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 22 March 2022 Accepted: 18 July 2023
Published online: 12 August 2023

References

1. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. Nate Genet. 2000;25(1):25–9.
2. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc: Ser B (Methodol). 1995;57(1):289–300.
3. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech: Theory Exp. 2008;2008(10):P10008.
4. Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, Baying B, Benes V, Teichmann SA, Marioni JC, et al. Accounting for technical noise in single-cell rna-seq experiments. Nat Meth. 2013;10(11):1093–5.
5. Byrd RH, Lu P, Nocedal J, Zhu C. A limited memory algorithm for bound constrained optimization. SIAM J Sci Comput. 1995;16(5):1190–208.
6. Chan TE, Stumpf MP, Babtie AC. Gene regulatory network inference from single-cell data using multivariate information measures. Cell Syst. 2017;5(3):251–67.
7. Cover TM, Thomas JA. Elements of information theory. Wiley; 2012.
8. Fisher R. Statistical methods for research workers. Gyan Books; 2017.
9. Freytag S, Tian L, Lönnstedt I, Ng M, Bahlo M. Comparison of clustering tools in r for medium-sized 10x genomics single-cell RNA-sequencing data. F1000Research. 2018;7:1297.
10. Consortium Gene Ontology. The gene ontology resource: enriching a gold mine. Nucl Acids Res. 2021;49(D1):D325–34.
11. Greulich P, Smith R, MacArthur BD. The physics of cell fate. In: Levine H, Jolly MK, Kulkarni P, Nanjundiah V, editors. Phenotypic switching. NewYork: Academic Press; 2020. p. 189–206.
12. Grün D, Kester L. Van, Oudenaarden A. Validation of noise models for single-cell transcriptomics. Nat Meth. 2014;11(6):637–40.
13. Grün D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, Clevers H. Van, Oudenaarden A. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature. 2015;525(7568):251–5.
14. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. 2019;20(1):1–15.
15. Hashimshony T, Senderovich N, Avital G, Klochendler A, De Leeuw Y, Anavy L, Gennert D, Li S, Livak KJ, Rozenblatt-Rosen O, et al. Cel-seq2: sensitive highly-multiplexed single-cell RNA-seq. Genome Biol. 2016;17:1–7.
16. Hausser J, Strimmer K. Entropy inference and the James–Stein estimator, with application to nonlinear gene association networks. J Mach Learn Res. 2009;10(7):1469–84.
17. Hicks SC, Townes FW, Teng M, Irizarry RA. Missing data and technical variability in single-cell RNA-sequencing experiments. Biostatistics. 2018;19(4):562–78.
18. Jaccard P. The distribution of the flora in the alpine zone. 1. NewPhytol. 1912;11(2):37–50.
19. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Meth. 2014;11(7):740–2.
20. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet. 2019;20(5):273–82.
21. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, Natarajan KN, Reik W, Barahona M, Green AR, et al. Sc3: consensus clustering of single-cell RNA-seq data. Nat Meth. 2017;14(5):483–6.
22. Kullback S, Leibler RA. On information and sufficiency. Ann Math Stat. 1951;22(1):79–86.
23. Li H, Courtois ET, Sengupta D, Tan Y, Chen KH, Goh JJL, Kong SL, Chua C, Hon LK, Tan WS, et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. Nat Genet. 2017;49(5):708–18.
24. Liu B, Li C, Li Z, Wang D, Ren X, Zhang Z. An entropy-based metric for assessing the purity of single cell populations. Nat Commun. 2020;11(1):1–13.

25. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with deseq2. Genome Biol. 2014;15(12):1–21.
26. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol. 2019;15(6): e8746.
27. Lun AT, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with bioconductor. F1000Research. 2016;5:2122.
28. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, Tirosh I, Bialas AR, Kamitaki N, Martersteck EM, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015;161(5):1202–14.
29. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426. 2018
30. Ortega A. Introduction to graph signal processing. Cambridge University Press; 2022.
31. Peterman N, Levine E. Sort-seq under the hood: implications of design choices on large-scale characterization of sequence-function relations. BMC Genom. 2016;17:1–17.
32. R Core Team R: a language and environment for statistical computing, R foundation for statistical computing, Vienna. https://www.R-project.org/, 2020
33. Rand WM. Objective criteria for the evaluation of clustering methods. J Am Stat Assoc. 1971;66(336):846–50.
34. Risso D, Purvis L, Fletcher RB, Das D, Ngai J, Dudoit S, Purdom E. Clusterexperiment and rsec: a bioconductor package and framework for clustering of single-cell and other large gene expression datasets. PLoS Comput Biol. 2018;14(9): e1006378.
35. Robinson MD, McCarthy DJ, Smyth GK. edger: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
36. Scrucca L, Fop M, Murphy TB, Raftery AE. mclust 5: clustering, classification and density estimation using gaussian finite mixture models. R J. 2016;8(1):289.
37. Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948;27(3):379–423.
38. Shorrocks AF. The class of additively decomposable inequality measures. Econom: J Econom Soc. 1980;48(3):613–25.
39. Smith RC, MacArthur BD. Information-theoretic approaches to understanding stem cell variability. Curr Stem Cell Rep. 2017;3(3):225–31.
40. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM III, Hao Y, Stoeckius M, Smibert P, Satija R. Comprehensive integration of single-cell data. Cell. 2019;177(7):1888–902.
41. Stumpf PS, Du X, Imanishi H, Kunisaki Y, Semba Y, Noble T, Smith RC, Rose-Zerili M, West JJ, Oreffo RO, et al. Transfer learning efficiently maps bone marrow cell types from mouse to human using single-cell RNA sequencing. Commun Biol. 2020;3(1):1–11.
42. Svensson V, Natarajan KN, Ly L-H, Miragaia RJ, Labalette C, Macaulay IC, Cvejic A, Teichmann SA. Power analysis of single-cell RNA-sequencing experiments. Nat Meth. 2017;14(4):381.
43. Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. Nat Protoc. 2018;13(4):599–604.
44. Consortium Tabula Muris. Single-cell transcriptomics of 20 mouse organs creates a tabula muris. Nature. 2018;562(7727):367.
45. Theil H. Economics and Information, theory studies in mathematical and managerial economics. North-Holland Publishing Company; 1967.
46. Tian L, Dong X, Freytag S, Lê Cao K-A, Su S, JalalAbadi A, Amann-Zalcenstein D, Weber TS, Seidi A, Jabbari JS, et al. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. Nat Meth. 2019;16(6):479–87.
47. Townes FW, Hicks SC, Aryee MJ, Irizarry RA. Feature selection and dimension reduction for single-cell RNA-seq based on a multinomial model. Genome Biol. 2019;20(1):1–16.
48. Trapnell C. Defining cell types and states with single-cell genomics. Genome Res. 2015;25(10):1491–8.
49. Van Dijk D, Sharma R, Nainys J, Yim K, Kathail P, Carr AJ, Burdziak C, Moon KR, Chaffer CL, Pattabiraman D, et al. Recovering gene interactions from single-cell data using data diffusion. Cell. 2018;174(3):716–29.
50. Van Rossum G, Drake FL. Python 3 reference manual. Scotts Valley: CreateSpace; 2009.
51. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. Nat Meth. 2020;17(3):261–72.
52. Zheng GX, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8(1):1–12.
53. Zhu C, Byrd RH, Lu P, Nocedal J. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. ACM Trans Math Softw (TOMS). 1997;23(4):550–60.

Publisher's Note