Information-theoretic clustering for single-cell transcriptomics

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

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

ABSTRACT

Single-cell sequencing (sc-Seq) experiments are producing increasingly large data sets. However, large data sets do not necessarily contain large amounts of information. Here, we introduce a formal framework for assessing the amount of information obtained from a sc-Seq experiment, which can be used throughout the sc-Seq analysis pipeline, including for feature selection, differential expression analysis, unsupervised clustering and semi-supervised cell annotation. Moreover, we show how this framework leads to a biologically intuitive definition of cell 'type' that can be used in common cell classification tasks. Based on this theory we describe an efficient method for the automatic unsupervised or semi-supervised identification of cell types from sc-Seq data, and provide an R package implementation. Our information-theoretic method performs comparably to the state-of-the-art on a series of benchmark sc-Seq data sets, and additionally provides a simple, biological interpretation of identified clusters.

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 techniques, 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, 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 (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 gene expression heterogeneity, that can be used to precisely measure the amount of information contained in a sc-Seq data set. Our method measures the information obtained from a sc-Seq experiment and decomposes it into that which is explained by putative grouping of cells – for example, by time-point labels or by a proposed clustering into cell types – and that which remains unexplained. This allows us to assess the extent to which known or inferred mechanisms explain observed expression patterns,

or fail to do so. The method is computationally simple and can be used to quickly assess the information content of a sc-Seq data set, identify features of importance, 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.

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 (Kiselev et al. 2019). 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 (Smith & MacArthur 2017). Our approach is motivated by an alternative gene-centric probabilistic view that seeks to more formally specify what is meant by gene 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 (Kullback & Leibler 1951). 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,

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

where $H(g)$ is the entropy of the expression of g in the population (see **Methods** for full details). For technical reasons, explained in the **Methods**, 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

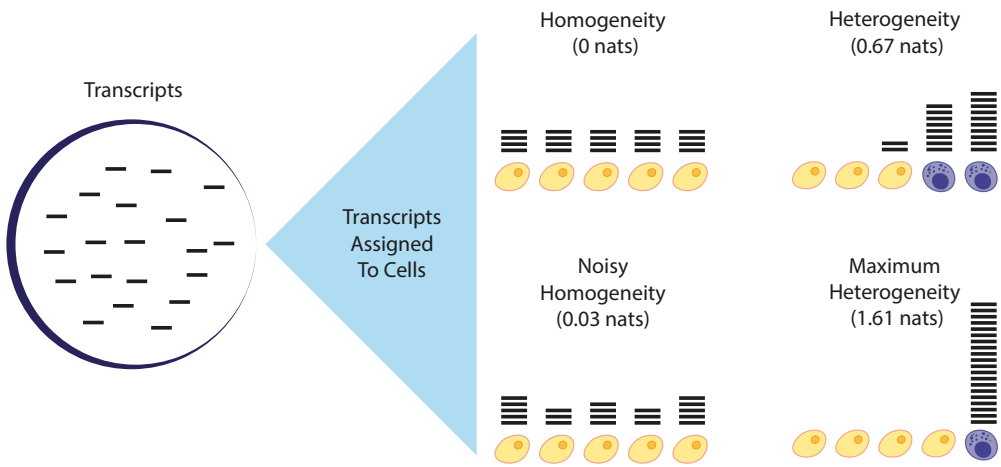


Figure 1. An information-theoretic view of sc-Seq data. Transcripts, or more generally counts, of a given gene are assigned to cells after sequencing. If the cell population is homogeneous with respect to the expression of g, then $I(g)$ will be zero (top left population). In practice, the transcript assignment process is stochastic, and so there will always be some deviation from this ideal (bottom left population). The technical effects of this stochasticity on the information obtained may be reduced by using a shrinkage estimator to estimate the distribution of transcripts (see **Methods**). If the population is heterogeneous, then transcripts are expressed preferentially in a subset of cells and the information obtained from the experiment will be larger (top right population), reaching a maximum at $\log(N)$, where N is the number of cells sequenced, when only one cell expresses the gene (bottom right population, $I(g) = \ln(5) \approx 1.61$).

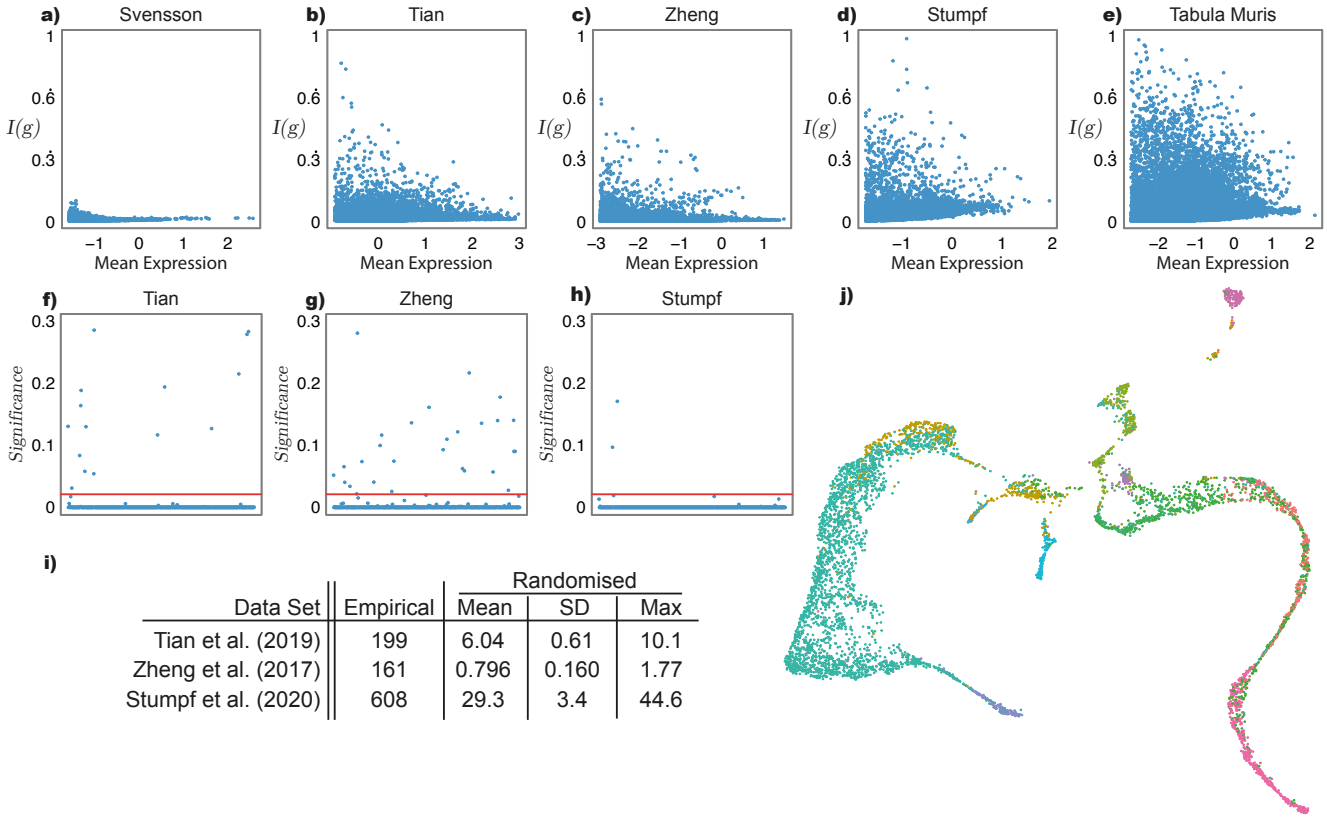
basis for numerous aspects of the sc-Seq pipeline, including feature selection, differential expression analysis, and unsupervised clustering.

To validate these uses, we considered a series of single-cell RNA-sequencing bench-marking data sets: *Svensson*, a technical control (Svensson et al. 2017); *Tian*, a mixture of three cancerous cell lines (Tian et al. 2019); *Zheng*, a set of FACS separated peripheral blood mononuclear cells (PMBCs); *Stumpf*, a sampling of cells from mouse bone marrow (Stumpf et al. 2020); and the *Tabula Muris*, a mouse cell atlas with cells from twelve organs (Tabula Muris Consortium 2018). **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) is a simple measure of quality for technical controls, which can be used to select potentially informative features for further study. Moreover, it may be easily modified to account for the presence of multiple homogeneous cell sub-populations, and thereby act as a measure of cluster quality.

Information-theoretic clustering

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**, 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 sub-populations, which we call the *inter-cluster heterogeneity* and write as $H_S(g)$. Namely, we show that

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



i)

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

Figure 2. Information-theoretic single-cell analysis. a-e) Plots of expression heterogeneity, $I(g)$ (normalised by the theoretical maximum, $\log(N)$, where N is the number of cells) 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. Panel **a)** shows data from a technical control (Svensson et al. 2017) (number of cell types, $C = 1$), **b)** a mixture of three cancerous cell lines (Tian et al. 2019) ($C = 3$), **c)** FACS sorted immune cells (Zheng et al. 2017) ($C = 4$), **d)** a sample of mouse bone marrow (Stumpf et al. 2020) ($C = 14$), and **e)** a multi-organ mouse cell atlas (Tabula Muris Consortium 2018) ($C = 56$). **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 (Benjamini & Hochberg 1995, Fisher 2017). Genes below this threshold are significantly differentially expressed between cell types as measured by $H_S(g)$. **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). **j)** A Uniform Manifold Approximation and Projection (UMAP) (McInnes et al. 2018) plot of the top 500 genes by $I(g)$ for the *Stumpf* data set. Each point is a cell, coloured by its scEC cluster.

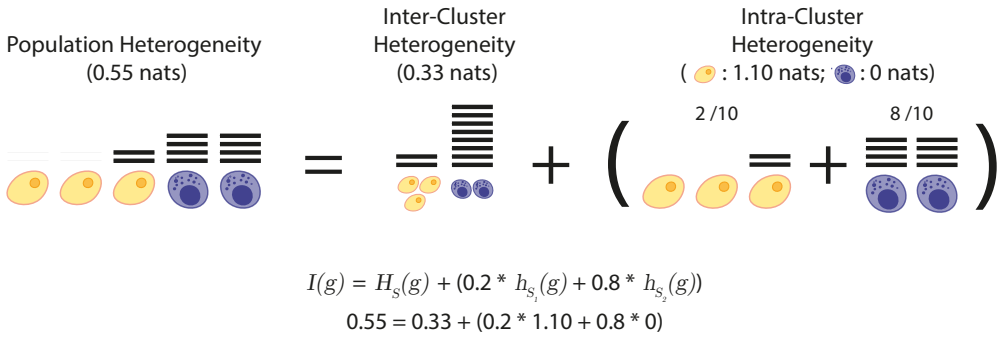


Figure 3. Heterogeneity is additively decomposable. The heterogeneity of a population of cells with respect to the expression of a gene g can be decomposed into inter- and intra-cluster heterogeneities for any proposed clustering, S. The inter-cluster heterogeneity $H_S(g)$ is determined by independently aggregating all transcripts associated with each sub-population in S and then taking the KLD of the resulting distribution from the uniform distribution on C letters. 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 of the heterogeneities of each of the constituent sub-populations, considered independently. It represents the average heterogeneity of the proposed clusters, accounting for disparities in cluster sizes. Further details and formulae are provided in the **Methods**.

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**, and an illustration is given in **Fig. 3**. We also provide an `R` package for its calculation (see **Methods**).

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 in 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 in these data sets (taken to be the top 500 genes by $I(g)$ in each case) the majority 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) (Benjamini & Hochberg 1995, Fisher 2017).

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 (Shannon 1948). 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 sub-populations and is maximised at $\sum_g I(g)$ if and only if

the proposed clustering divides the population into indivisible 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.

Building on this empirical association, we reasoned that the most accurate cell type annotation processes should maximise H_S for a given cell population. Based on this assumption, we established an automated process for cell type annotation from sc-Seq data that assigns each cell to one of C non-intersecting discrete identities such that H_S is maximised. This optimisation constitutes a novel unsupervised clustering method developed explicitly for sc-Seq data, based on a biologically-motivated notion of expression heterogeneity. We refer to this method as scEC (single-cell Entropic Clustering).

To determine the efficacy of scEC we assessed the similarity of 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 (the adjusted Rand index is a measure of similarity of a proposed clustering to a known clustering, Rand (1971)). 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 (Blondel et al. 2008, Luecken & Theis 2019, Freytag et al. 2018, Stuart et al. 2019).

These results indicate that scEC is comparable in performance to the state-of-the-art in unsupervised single cell clustering. However, scEC has two additional advantages. First, unlike most alternatives, scEC is not a purely statistical process, but rather is anchored on a biologically-motivated definition of expression heterogeneity, and is specifically tailored to deciphering cell types from single cell expression data, rather than clusters in data more generally. Consequently, scEC has no free parameters, and can be implemented in a simple, robust way. Second, because it is based on a biologically-motivated definition of heterogeneity scEC is biologically intuitive and is, consequently, easily interpretable. For instance, because the scEC objective function, H_S, is naturally built by summation of gene-level measurements, it 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 established annotations for the *Tian* & *Stumpf* data sets (see **Fig. 4** for an illustration).

We found that the values of $H_S(g)$ associated with scEC clusters was 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 values associated with scEC clusters for the key marker genes used in each data set (*Zheng* : *CD14*, *CD4*, *CD8* and *NCAM1*, Zheng et al. (2017); *Stumpf*: *Cd34*, *Kitl*, *Spi1*, *Gata1* and *Pax5*, Stumpf et al. (2020)) were significantly higher than expected by chance (one-sided exact test, 10^4 permutations, $p < 0.05$, false discovery rate corrected).

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 between scEC clusters and established annotations in each data set for further investigation. To illustrate we chose those genes for which

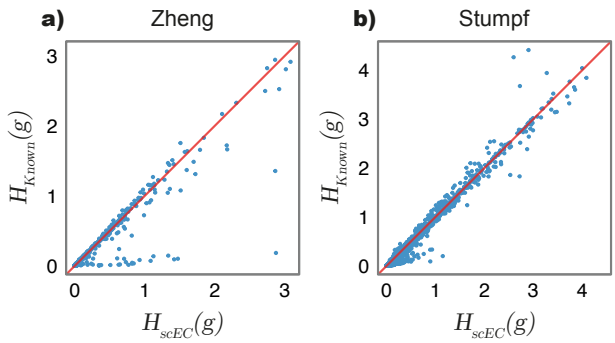


Figure 4. Comparison of inter-cluster heterogeneity of scEC-generated clusters and established annotations. Plots of $H_S(g)$ based on an scEC-generated clustering (x-axis) and established annotations (y-axis) for the **a)** *Zheng* and **b)** *Stumpf* data sets (Zheng et al. 2017, Stumpf et al. 2020). In both panels each point represents a gene profiled and the red line indicates $H_{scEC}(g) = H_{Known}(g)$. Genes below the red line are better explained by the scEC clustering, while those above the line are better explained by the prior annotation.

$\Delta H_S(g) = |H_{scEC}(g) - H_{Known}(g)| > 0.1$ nats.

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) (Ashburner et al. 2000, Gene Ontology Consortium 2021, Benjamini & Hochberg 1995).

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 erythrocyte 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 **Supplementary Table** for contingency table of cluster assignments).

Collectively, these differences reflects the preference of scEC for cellular annotations driven by biological processes involving larger tranches of genes. Depending on circumstance, this preference may be both 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, because of the simplicity of the scEC methodology this issue can this issue can be relatively easily addressed. 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. Doing so leads to a semi-supervised version of scEC, which we explain below.

Semi-supervised cell classification

Taken together, these results indicate that our information-theoretic clustering method can be used to successfully identify clusters, corresponding to distinct homogeneous cell sub-populations, from sc-Seq data. However, scEC can be insensitive to important cellular processes driven by small numbers of genes. Moreover, annotating the recovered clusters (i.e. assigning biologically meaningful identities, for example based on marker gene expression patterns) still requires time-consuming manual inspection and differential gene expression analysis. We therefore sought to develop a semi-supervised version of scEC that combines clustering and annotation, by comparing experimentally derived cell profiles with those in a relevant reference database, such as a cell atlas.

Assuming that individual cells in the reference data set have been clustered and correctly and unambiguously annotated, the semi-supervised implementation of scEC assigns cells of a novel, test data set, to the pre-identified clusters of the reference data set such that the total inter-cluster heterogeneity H_S of the combined data set is maximised. The cells of the test data set then inherit the annotation of their assigned reference cluster, thereby assigning a biological identity without laborious manual inspection. A full mathematical description of this semi-supervised classification process is provided in the **Methods**.

To illustrate this method, we again considered the *Tian* data set, which consists of sc-Seq data obtained from a mixture of cells drawn from three distinct cancerous cell lines (known as H1975, H2228 and HCC82). In addition to this data set, Tian and co-workers also generated a separate data set consisting of sc-Seq data obtained from a mixture of cells drawn from the original three cell lines and from a further two cancerous cell lines (known as A549 and H838) (Tian et al. 2019). Since true cell annotations obtained by genotyping are available in both cases, we reasoned that these two data sets would collectively provide a good benchmark to assess the accuracy of our semi-supervised methodology. To test this possibility we used the *Tian5* data set as a reference against which to compare the cell profiles in the *Tian* data.

In this example, the semi-supervised cell annotation process proceeds in two steps. First the *Tian* data *Tian5* data are concatenated to create a merged data set, and each cell of the reference *Tian5* data set is assigned its known cellular identity, while each cell of the test *Tian* data set is assigned a fuzzy cluster identity, with uniform weighting across the set of five reference clusters (i.e. each cell i in the test set is assigned to cluster j with probability p_{ij}, where, for each i elements of p_{ij} are chosen randomly and uniformly). Second, the total inter-cluster heterogeneity $H_S = \sum_g H_S(g)$ of the merged data set is maximised, allowing the annotations of the cells in the test data set of vary, while keeping the identities of the reference cells fixed (details of the optimisation routine are provided in the **Methods**). This semi-supervised assignment process thereby automatically annotates each cell of the test data set using the reference labels.

The results of this semi-supervised assignment process are given in **Fig. 5**. Cellular annotations obtained by comparing the *Tian* data with the reference *Tian5* substantially agreed with their genotype annotations (ARI of 0.90), indicating that identities of cells in the *Tian* data were robustly recovered by our semi-supervised process. Notably, while it is possible for cells in the test data to be mis-assigned to the A549 and H838 identities – which are present in the *Tian5* reference data but not in the *Tian* test data – this occurred only very rarely (only 6% of cells were mis-assigned in this way) despite considerable similarities in these cell lines. This example illustrates that our information-theoretic semi-supervised cell annotation process is able to reliably identify cell types in heterogeneous samples, even when differences between cell types are subtle.

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 (Greulich et al. 2020). Although useful, this is not a specifically biological assumption (similar assumptions form the basis of numerous clustering algorithms in

	H1975	H2228	HCC827
H1975	276	2	0
H2228	0	308	1
HCC827	0	0	258
A549	29	5	15
H838	8	0	0

Figure 5. Semi-supervised classification of cancer cell lines. Confusion matrix for semi-supervised classification of cells in the *Tian* data, using the *Tian5* data as a reference (Tian et al. 2019). Columns represent the established (i.e., true) experimentally-derived identities (H1975, H2228 and HCC82) while rows give predicted cell identities derived from the reference *Tian5* data set (H1975, H2228, HCC82, A549 and H838), based on our semi-supervised process. In this case, the adjusted Rand index between the semi-supervised classification and true identities is 0.90.

other different disciples) 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 in a mathematical formulation, drawing on tools from information theory. We have proposed a formal measure of expression heterogeneity and have used this measure to establish a biologically-motivated clustering procedure, which we call single cell entropic clustering (scEC), that is specifically tailored for single-cell expression data.

While successful and well-established clustering methods, such as the popular Louvain method, may be more general, scEC is more biologically intuitive and has an interpretation that will be familiar to many experimentalists. Importantly, the performance of scEC is comparable to a state-of-the-art on a series of benchmark sc-Seq data sets, suggesting that this underlying biological basis is justified. With sc-Seq data sets continually growing in size and complexity, there is a need for biologically grounded – rather than purely statistical – approaches to understanding them. scEC provides an efficient, parameter-free method for doing this. An `R` package for the implementation of scEC available at `https://github.com/mjcasy/scEC`.

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) (R Core Team 2020).

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 (Hausser & Strimmer 2009). 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 **Eqn. 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 (Hausser & Strimmer 2009, Chan et al. 2017).

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

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

The Kullback-Leibler divergence (KLD), or relative entropy, is a measure of the information encoded in the difference between two probability distributions (Kullback & Leibler 1951). 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 **Eqn. 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 $1 \leq i \leq N$.

Differential expression analysis

A crucial property of the relative entropy is that it is additively decomposable with respect to arbitrary groupings (Shorrocks 1980). Informally, this means that if we have a clustering of the cells into disjoint groups, then $I(g)$ can be reconstructed from inter-cluster and intra-cluster heterogeneities. Next, we formalise this additive decomposition and give a self-contained derivation (cf. Theil (1967)).

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 **Eqn. 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:

$$\begin{aligned}
I(g) &= \log(N) - \sum_{i=1}^{N} x_i \log\left(\frac{1}{x_i}\right), && (11)\\
&= \log(N) - \sum_{k=1}^{C} \sum_{i \in S_k} x_i \log\left(\frac{1}{x_i}\right), && (12)\\
&= \log(N) - \sum_{k=1}^{C} y_k \sum_{i \in S_k} \frac{x_i}{y_k} \left(\log\left(\frac{1}{x_i/y_k}\right) + \log\left(\frac{1}{y_k}\right)\right), && (13)\\
&= \log(N) - \sum_{k=1}^{C} y_k \underbrace{\sum_{i \in S_k} \frac{x_i}{y_k} \log\left(\frac{1}{x_i/y_k}\right)}_{H(Z_k)} - \sum_{k=1}^{C} y_k \sum_{i \in S_k} \frac{x_i}{y_k} \log\left(\frac{1}{y_k}\right), && (14)\\
&= \log(N) - \sum_{k=1}^{C} y_k H(Z_k) - \underbrace{\sum_{k=1}^{C} \log\left(\frac{1}{y_k}\right) \overbrace{\sum_{i \in S_k} x_i}^{y_k}}_{H(Y)}, && (15)\\
&= \log(N) - \sum_{k=1}^{C} y_k H(Z_k) - H(Y), && (16)\\
&= \underbrace{\log(N) - H(Y) - \sum_{k=1}^{C} y_k \log(N_k)}_{A} + \underbrace{\sum_{k=1}^{C} y_k \log(N_k) - \sum_{k=1}^{M} y_k H(Z_k)}_{B}. && (17)
\end{aligned}$$

Expression A may be rewritten as:

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

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

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

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

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:

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

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

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

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

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

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 (Cover & Thomas 2012), 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 **Eqn. 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 **Eqn. 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. **Eqn. 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, **Eqn. 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) (Byrd et al. 1995, Zhu et al. 1997, Van Rossum & Drake 2009, Virtanen et al. 2020).

The *L-BFGS-B* algorithm is an efficient non-linear local optimisation method developed for solving large, dense problems such as this (Zhu et al. 1997)) 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 **Eqn. 30** and **38**, we have

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

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

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

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

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

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

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

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

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

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

The membership functions μ_{rs} are not independent, since $\sum_{k=1}^{C} \mu_{ik} = 1$ for all i, by **Eqn. 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, **Eqn. 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 **Eqn. 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) (Byrd et al. 1995, Van Rossum & Drake 2009, Virtanen et al. 2020), 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 **Eqn. 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 (Byrd et al. 1995, Zhu et al. 1997). 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 (Fisher 2017). We control the false discovery rate arising from multiple testing for each data set using the `R` function *p.adjust* (Benjamini & Hochberg 1995).

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. (2019) (Blondel et al. 2008, McInnes et al. 2018). The Adjusted Rand Index (ARI) was calculated for each data set using the associated function in the `R` package `mclust` (v5.4.7) (Rand 1971, Scrucca et al. 2016).

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 $\sum_{i \in T_k} x_i^g = 1$ (Hausser & Strimmer 2009). 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 **Eqn. 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.

Declarations

Authors' contributions

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

Acknowledgements

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

Funding

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

Availability of data and materials

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

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

The *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 https://data.mendeley.com/datasets/csvm3kpkxd/1?fbclid=IwAR3j_hvq5Zt0cdxBBM72Fqr8zXwVC6XRpkY2JtVwbcj1gzyNLMVsdoFYWgQ

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

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T. et al. (2000), 'Gene ontology: tool for the unification of biology', *Nature genetics* **25**(1), 25–29.

Benjamini, Y. & Hochberg, Y. (1995), 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', *Journal of the Royal statistical society: series B (Methodological)* **57**(1), 289–300.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. (2008), 'Fast unfolding of communities in large networks', *Journal of statistical mechanics: theory and experiment* **2008**(10), P10008.

Byrd, R. H., Lu, P., Nocedal, J. & Zhu, C. (1995), 'A limited memory algorithm for bound constrained optimization', *SIAM Journal on scientific computing* **16**(5), 1190–1208.

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

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

Fisher, R. (2017), *Statistical Methods For Research Workers*, Gyan Books.

Freytag, S., Tian, L., Lönnstedt, I., Ng, M. & Bahlo, M. (2018), 'Comparison of clustering tools in r for medium-sized 10x genomics single-cell rna-sequencing data', *F1000Research* **7**.

Gene Ontology Consortium (2021), 'The gene ontology resource: enriching a gold mine', *Nucleic Acids Research* **49**(D1), D325–D334.

Greulich, P., Smith, R. & MacArthur, B. D. (2020), The physics of cell fate, *in* H. Levine, M. K. Jolly, P. Kulkarni & V. Nanjundiah, eds, 'Phenotypic Switching', Academic Press, pp. 189–206.

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

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

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

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

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

Luecken, M. D. & Theis, F. J. (2019), 'Current best practices in single-cell rna-seq analysis: a tutorial', *Molecular systems biology* **15**(6), e8746.

McInnes, L., Healy, J. & Melville, J. (2018), 'Umap: Uniform manifold approximation and projection for dimension reduction', *arXiv preprint arXiv:1802.03426* .

R Core Team (2020), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

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

Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. (2016), 'mclust 5: clustering, classification and density estimation using gaussian finite mixture models', *The R journal* **8**(1), 289.

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.

Stumpf, P. S., Du, X., Imanishi, H., Kunisaki, Y., Semba, Y., Noble, T., Smith, R. C., Rose-Zerili, M., West, J. J., Oreffo, R. O. et al. (2020), 'Transfer learning efficiently maps bone marrow cell types from mouse to human using single-cell rna sequencing', *Communications biology* **3**(1), 1–11.

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.

Theil, H. (1967), *Economics and Information Theory*, Studies in mathematical and managerial economics, North-Holland Publishing Company.

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

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

Van Rossum, G. & Drake, F. L. (2009), *Python 3 Reference Manual*, CreateSpace, Scotts Valley, CA.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J. et al. (2020), 'Scipy 1.0: fundamental algorithms for scientific computing in python', *Nature methods* **17**(3), 261–272.

Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J. et al. (2017), 'Massively parallel digital transcriptional profiling of single cells', *Nature communications* **8**(1), 1–12.

Zhu, C., Byrd, R. H., Lu, P. & Nocedal, J. (1997), 'Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization', *ACM Transactions on Mathematical Software (TOMS)* **23**(4), 550–560.

Supplementary Table

Erythroblasts	2436	171	0	0	0	1	0	0	1	0	233	0	3	0
Myeloblasts	39	4	313	0	0	0	0	0	0	0	0	245	2	0
Pro-B	38	0	0	187	0	0	0	0	0	0	0	0	0	0
Pre-B	0	0	0	58	0	0	0	0	2	0	0	2	0	0
Monocytes	0	0	0	23	165	0	0	0	4	2	0	23	0	0
Basophils	0	0	0	0	0	130	0	0	0	0	2	1	0	0
Pericytes	0	0	0	0	0	0	110	0	0	2	0	0	0	0
Megakaryocytes	0	0	0	0	0	0	0	40	0	0	21	0	0	0
T-NK	3	0	0	0	0	0	0	0	56	0	0	1	0	0
Endothelial Cells	0	0	0	0	1	0	4	0	0	41	4	0	0	0
HSPCs	49	4	3	5	1	7	0	0	2	0	182	50	0	0
Monoblasts	7	0	16	7	66	0	1	0	0	0	3	200	0	0
Neutrophils	0	0	0	1	1	1	0	0	1	0	0	8	255	0
Myelocytes	2	0	4	3	0	0	0	0	0	0	0	93	164	0

Contingency table between cell annotations provided in (Stumpf et al. 2020) and scEC labels. Note that while the number of clusters for scEC was fixed at 14, only 13 clusters were realised; generally, the number of clusters returned will match the number specified, except in cases where any further splits in clusters provide only a minor increase in inter-cluster heterogeneity, so are harder to find by numerical optimisation.