Special Issue, Single-cell Analysis: Epistemological Inquiries

Vol. 6, No. 2 (2023)
ISSN: 2532-5876

DOI: 10.13133/2532-5876/18131

Do Single-cell Experiments Challenge the Concept of Cell Type?

Fridolin Gross[a*]

[a] *CNRS UMR5164 ImmunoConcEpT, University of Bordeaux, 146 rue Leo Saignat, Bordeaux, 33076 France*

***Corresponding author:** Fridolin Gross, Email: fridolin.gross@u-bordeaux.fr

Abstract

Recent debates among biologists have highlighted problems with the traditional concept of cell type, which is considered vague and subjective. Single-cell technologies reveal the limitations of the current concept by exposing a high degree of heterogeneity in cell populations. At the same time, some biologists believe that these technologies provide the basis for a more objective and precise concept of cell type that is not dependent on prior theoretical assumptions. In this paper, I explore the impact that single-cell experiments and analyses will have on the concept of cell type. Drawing on the practices of biologists using these methods, but also on more principled arguments, I argue that the idea of a purely theory-free classification is unlikely to be realized. However, single-cell technology may affect the concept of cell type in more subtle ways.

Keywords: ontology, cell type, theory-free classification; pheneticism, single-cell technology

Citation: Gross, F 2023, "Do Single-cell Experiments Challenge the Concept of Cell Type?", *Organisms. Journal of Biological Sciences*, vol. 6, no. 2, pp. 11–23. DOI: 10.13133/2532-5876/18131

Introduction

Biologists use many concepts without worrying too much about clear definitions. Some of these are quite fundamental, such as the concept of gene or even of life itself. This is not necessarily problematic, and it has even been argued that precisely ambiguity and indeterminacy contribute to the fruitfulness of some scientific concepts (Neto 2020). Sometimes, however, a discipline may undergo certain developments that force scholars to think more deeply about a particular concept and clarify it by making some of their tacit assumptions explicit. The concept of cell type is an interesting illustration of this pattern. For more than a century, biological and medical practice has identified and distinguished cell types according to various and often ill-defined sets of criteria related, for example, to morphology, function, location, or developmental origin, without ever converging on an explicit and general account. However, recently introduced experimental techniques, along with computational methods of data analysis, seem to be forcing biologists to clarify their ideas about what they mean when they talk about cell types. In particular, single-cell sequencing experiments provide much more detailed insight into the diversity and heterogeneity of cell populations. This has led some people to argue that biology needs a more principled and possibly more fine-grained classification of cells into types, sub-types, or states. At the same time, many biologists think that the new experimental techniques offer the possibility of achieving a delineation of cell types that, because

purely data-driven, is more objective and precise and thus superior to the subjective, vague, and potentially biased classifications based on the traditional concept of cell type.

In this paper, I investigate what impact such technological advances can be expected to have on the concept of cell type. In particular, I will address the claim that such an approach to classification can be based solely on data-driven methods. Plausibly, such a "theory-free" account may be desirable for a variety of reasons, but it is unclear to what extent scientific concepts and classifications could be based on such foundations alone. Interestingly, philosophers have been discussing very similar questions about classificatory concepts in different contexts, notably with regard to the classification of organisms into species and higher taxa. Therefore, the debate about cell types might benefit from an awareness of some of the problems and arguments that were debated elsewhere in the past.

The paper is structured as follows. In Section 1, I provide an overview of the current debates around cell types, and explain why biologists feel the need to revise or clarify the concept. Section 2 offers some philosophical background on classification and shows how biologists, over time, have endorsed different approaches to the classification of cells that may correspond to different philosophical positions. In particular, the new approaches based on single-cell technologies fit a pheneticist or clustering approach to classification that is familiar in the biological taxonomy debate. In Section 3, I argue that a purely data-driven version of such a pheneticist approach is unlikely to be successful, before wrapping up the matter with some remarks in the Conclusion.

1. Cell types and single-cell experiments

Cell theory, dating back to the 19th century, established the idea that the tissues of animals and plants are made up of basic building blocks, all of which originate from the same fertilized egg cell (Duchesneau 1987; Canguilhem 1995). Although all cells in a multicellular organism contain much the same genetic material, they can differ radically in size, morphology, and the role they play in the context of the organism. Based on early microscopy and staining techniques, cell types were at first distinguished using phenotypic criteria, for example in terms of the functions they carry

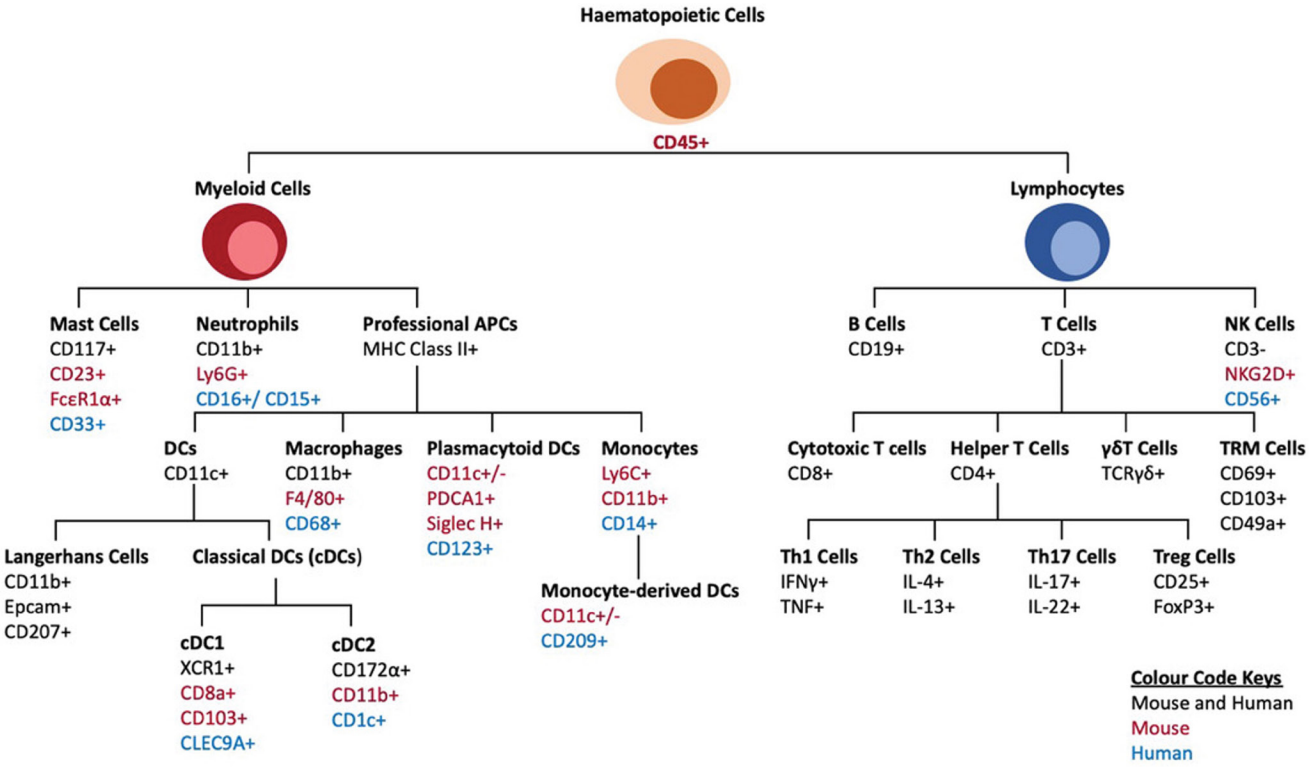

Figure 1: Lineage tree of human hematopoietic cells. Adapted from Murphy et al. (2022) (CC BY-NC-ND 4.0).

out within the organism or according to morphological features such as size or shape.

Ramon y Cajal and Camillo Golgi for the study of the fine structure of the nervous system (Jones 1999) and Alexander Maximov for the discrimination of different types of blood cells (Novik *et al.* 2009) are among the pioneers in applying these methods. Over the years these techniques were refined, often by linking cell types to specific "marker" molecules (Baskin 2015). Notably in the context of immunology, a sophisticated system of such markers, known as "cluster of differentiation" (CD), was developed. It allows biologist to detect and distinguish different cell types using techniques such as immunohistochemistry and flow cytometry that are based on the detection of specific molecular patterns on the cell surface (Chan *et al.* 1988). As an example, Figure 1 shows the lineage tree of hematopoietic cells (i.e., white and red blood cells) along with common markers.

Overall, however, these approaches have remained largely qualitative and context-dependent. And while they have given rise to very precise methods of detection, and thus to operational definitions of specific cell types, they have not led to a general agreed-upon conceptual definition that can be straightforwardly applied across different biological sub-disciplines (Clevers *et al.* 2017). In line with this, one does not find much explicit discussion of the concept of cell type by biologists up until very recently. For instance, a standard textbook, such as (Alberts *et al.* 2015) does not contain any clear definition, nor does it its historical equivalent from 1924 (Cowdry 1924).

Despite this lack of a clear definition, there seems to be a shared intuitive understanding. Among the main criteria commonly alluded to in the discussions around cell types, we find *structure*, *function*, and *lineage* (Clevers *et al.* 2017). Structural criteria classify cells according to differences in the arrangement of and the relations between their parts. This includes broad features, such as shape, size and morphology, but also comprises details that are more specific, such as distinctive expressed molecules, or the presence of particular cellular substructures. Functional criteria, by contrast, classify cells according to the role that they carry out in the context of the organism. For example, fibroblasts are sometimes defined as cells that contribute to the formation of connective tissue by secreting collagen proteins (National Human Genome Research Institute 2022). Finally, lineage-based criteria classify cells according to their developmental ancestry. This means that we identify a type of cell in terms of its position in a lineage tree such as the one shown in Figure 1. Much of biological research seems to be based on the tacit assumption that these criteria neatly coincide and yield one objective classification scheme. However, it is by no means obvious that this is the case, and the assumption that different perspectives on a system lead to matching ways of decomposing it into parts may reflect a serious underestimation of its actual complexity (Wimsatt 2007).

Such complexity is revealed by recent research and advances in experimental methods. On the one hand, observations of cellular plasticity, dedifferentiation, transdifferentiation, and especially the "reprogramming" of terminally differentiated cells to a pluripotent state have led to a fundamental rethinking of some of the basic assumptions of the field (Andrews 2002; Sánchez Alvarado and Yamanaka 2014; Laplane and Solary 2019). The idea of cell types as clearly demarcated and irreversibly committed end points of differentiation has been put into question and given way to a more fluid picture.

In parallel, advances in genomics have led to the realization that the activities of cells are based on a complex and dynamic orchestration of genetic and epigenetic factors that influence their development as well as their morphology and functional properties. Biologists have revisited earlier ideas from Conrad Waddington who in the 1950s coined the metaphor of the epigenetic landscape, which compares the differentiation of cells and tissues to a marble rolling down an inclined surface (Waddington 1957). The particular shape of the surface, with hills and valleys, creates preferred paths and branching points for the marble, corresponding to developmental trajectories and decision points that eventually lead the developing system towards one of several possible ends or 'fates.' Drawing in particular on the work by (Kauffman 1974), it has been proposed that cell types should be understood as different "attractor states" of the complex dynamical system constituted by the gene regulatory network that is shared by all cells of an organism (Kauffman 2004; Huang 2009).

Finally, the advent of next-generation sequencing techniques, particularly single-cell sequencing has enabled biologists to measure the diversity of cell populations with an unprecedented level of detail.

For example, single-cell RNA sequencing provides a snapshot of the simultaneous expression of thousands of genes in individual cells, while single-cell ATAC sequencing captures the accessibility and therefore the regulatory state of the genome at the single-cell level (Van den Berge *et al.* 2019; Yan *et al.* 2020). Sophisticated computational techniques are required to convert the resulting high-dimensional data sets into representations that can be meaningfully analyzed, notably using various clustering algorithms along with methods to annotate these clusters to known cell types. The picture that emerges from this type of analysis suggests that cell populations previously considered to be of the same type are often much more heterogeneous than expected and appear to contain different sub-populations or cell states (Trapnell 2015).

Interestingly, there have recently been a number of articles by biologists that explicitly raise the question of how the cell type concept should be defined given the recent scientific and technological developments (Arendt *et al.* 2016; Clevers *et al.* 2017; Fishell & Heintz 2013; Morris 2019; Zeng 2022). Some of them suggest that the new techniques will allow a more formal and objective classification of cells into types and states analogous to the classification of chemical elements and their isotopes in the periodic table (Xia & Yanai 2019), while others are concerned that the observed degree of plasticity and heterogeneity will render any attempt to classify them into discrete types entirely subjective (Clevers *et al.* 2017). Overall, the impression is that intuitions about what constitutes a cell type vary considerably among biologists and seem to include different ways of capturing the relationships between the three criteria of structure, function, and lineage mentioned above. Corresponding to this is a lack of standardization in the field and a variety of often conflicting methods by which cell classifications are performed in practice based on the new experimental techniques, a problem well summarized in the following quote from a recent article in Cell, one of the leading biology journals:

> "Single-cell biology is facing a crisis of sorts. Vast numbers of single-cell molecular profiles are being generated, clustered and annotated. However, this is overwhelmingly ad hoc, and we continue to lack a principled, unified, and well-moored system for defining, naming, and organizing cell types" (Domcke & Shendure 2023, p. 1103).

The crisis described motivates the guiding questions of this paper: Are single-cell methods by themselves sufficient to enable a more coherent classification of cells? If not, what is their role in improving the traditional concept, which is considered deficient in important aspects? Before addressing these questions directly, I will provide some philosophical background on classification that will help illuminate some of the conceptual issues involved.

1. Cell types and the philosophy of classification

Creating a scheme according to which cells are assigned to specific types means creating a classification. Philosophers have been thinking about classification and related practices for millennia, and so it might be useful to look at some of this work to see if some insights might illuminate the search for the right cell type concept. In particular, the debates among biologists and philosophers of biology about the classification of organisms into a system of taxonomy have interesting parallels with the case of cell types.

A first distinction can be made between classifications that are arbitrary or simply based on human interests and classifications that in some way reflect actual patterns in the world. A common traditional way of thinking about such "natural" classifications is that objects in the world belong to the same class if they share certain basic properties or *essences*. For example, all water molecules share a common molecular structure described by the chemical formula H_2O. In general, the real essence of a class may not be known, just as the molecular structure of water was not known until relatively recently. John Locke thought that most actual classifications used by humans are based on nominal essences, by which he meant that they are based on observable macroscopic properties that do not necessarily coincide with the underlying and unknown real essences. While essentialism may be a defensible position with respect to chemical elements and molecules, it has been largely discarded in biological debates about the classification of organisms into species and higher taxa. The insights of evolutionary biology have shown that species are not static entities but are subject to change over time, and that the organisms within a species exhibit significant variation at any point in time. It has been doubted, therefore, whether in general any fixed set of

properties can be found that is shared by all and only the members of a given taxon (Hull 1965).

Following Ereshefsky (2000), two main approaches to classification have been proposed as an alternative to essentialism: cluster approaches and historical approaches. Cluster approaches are similar to essentialism in that they are also based on the properties of the objects being classified, but they are less rigid in that they do not require properties to be shared by all members of a class. Instead, membership is based on overall similarity, that is, the degree to which objects share a set of properties. A prominent example of such an approach in the context of biological taxonomy is pheneticism. Pheneticists hold that biologists should record as many properties of individual organisms as possible, usually in morphology or other observable traits, and then classify them according to a measure of distance in the "phenotypical space" constituted by these properties. Phylogeny or other evolutionary relationships are deliberately ignored. Historical approaches, by contrast, are not based on the shared properties of objects at all, the criterion instead is whether they share a causal history. In the context of biological taxa this causal history is provided by common evolutionary descent. David Hull's account of species as individuals is an example of a historical approach (Hull 1976).

It is interesting to see how different stages in the history of cell type classification can be reconciled with different philosophical approaches to classification, without this necessarily being in the minds of the biologists involved. I will admit right away that such an assignment is based on a sketchy and probably somewhat caricatured representation of the actual history. It should also be noted that this account is not intended to represent a purely conceptual development but is clearly determined to a considerable extent by the experimental technologies available for the study of cells. Nevertheless, I think this perspective illuminates some of the conceptual issues surrounding the problem of cell type classification.

The early investigations based on light microscopy and staining techniques can be interpreted as classifications based on nominal essences, in Locke's sense. Cells were identified and distinguished based on readily observable, macroscopic features, such as morphology, size, or color after staining. Already in the late 19th century, biologists suspected that chromatin, a stainable nuclear substance, was involved in cellular differentiation, assuming that stem cells preserve and pass on the complete chromatin of the fertilized egg, while differentiated somatic cells preserve only specific parts of it (Maehle 2011). However, these ideas could at the time not be linked to specific experimental measurements, and therefore the presumed "real essences" of cell types were out of reach.

At the same time, comparative embryologist studies revealed the developmental relationships of differentiating cells. Studying the formation of blood cells, researchers such as Artur Pappenheim and Alexander Maximow revealed complex 'stem trees' that displayed the genealogical relationships between different types of blood cells. These studies suggest an alternative criterion for the classification of cells into types according to their developmental ancestry, which is in analogy to the historical approach to classifying organisms according to phylogenetic relationships (Lancaster 2017).

The first half of the 20th century saw tremendous advances in how the genetic material affects the properties of cells and organisms, especially with the transition from classical genetics to molecular genetics. According to the central dogma of molecular biology, which can be considered as the culmination of these developments, the information-bearing part of chromatin is DNA, and this information is transferred from nucleic acids to proteins, determining phenotypic characteristics by specifying functionally active molecules. Consistent with this picture, cell types were conceptualized as endpoints of unidirectional and irreversible differentiation pathways, during which cells acquire the ability to produce specific types of proteins that enable them to carry out their respective functions in the organism. Historian of biology, Richard Burian summarizes this view as follows:

> "The underlying hypothesis was that differentiation is an irreversible commitment of a cell lineage to the manufacture of a coordinated set of "luxury" proteins—i.e., specialized proteins not needed to maintain the life of the cell. Thus, the primary differences among nerve, kidney, skin, and blood cells were thought to depend on the specialized sets of proteins that they make, which, in turn, affect their morphologies, interactions with other cells, and responses to biological signals and stimuli" (Burian 1993, p. 391).

We may interpret this view as the confirmation of an essentialist position, in which real essences are identified with the unique set of molecules that characterize the phenotype and function of a differentiated cell. In particular, in the context of immunology, it was assumed that cell types could be characterized in terms of their "surface phenotype" of specific proteins expressed on the cell surface and measured by flow cytometry (e.g., Lanier *et al.* 1983).

As mentioned above, more recent research has undermined this view by revealing, on the one hand, that cell fates are much less static and irreversible than previously thought. Dynamic transitions between cell fates can be induced experimentally and may occur also under physiological conditions, notably the "reprogramming" of terminally differentiated cells into a pluripotent state. On the other hand, genome-wide single cell sequencing methods have enabled a much more detailed exploration of the heterogeneity of cellular population and notably have provided a refined idea of how gene expression relates to cellular phenotypes and functions. The view that well-delineated sets of genes are switched either on or off to determine the characteristics of cell types now appears simplistic. Instead, the idea of basing cell type classifications on exhaustive molecular measurements seems much more solid. In addition, it seems that computational methods of identifying clusters in these high-dimensional datasets are nowadays available to make classifications easily available. This development can be understood as a move towards a pheneticist conception of cell types, and many of the arguments put forward in favor of such an approach resemble the arguments that pheneticists have formulated against alternative views on the taxonomy of organisms. One common argument, in particular, is that a pheneticist account is desirable because it would make classifications independent of any theoretical assumptions. In the following section we will take a closer look at the prospects of such a view in the context of cell type classification.

2. A theory-free account of cell types?

When looking at the recent discussions among biologists concerning the definition of cell types, a recurring motif is the idea that such a definition should be independent of theoretical assumptions, and that single-cell techniques can provide the basis for such a definition. Developmental biologist Samantha Morris, for instance, points out:

> "These methods enable the capture of many thousands of features, without the requirement for experimental cell enrichment, thus generating a rigorous and unbiased picture of the range of cell phenotypes that exists within any given tissue" (Morris 2019, p. 2).

In the context of neuroscience, Hongkui Zeng makes a similar case for data-driven classification:

> "To untangle this complexity, it is necessary to adopt approaches that provide comprehensive, unbiased, quantitative, and standardizable measurements and are scalable to densely sample a sufficient number of cells within a brain region or tissue organ as well as across the entire brain and body to eventually reach completeness, and then perform data-driven computational clustering and analysis to obtain cell type classification" (Zeng 2022, pp. 2739–2740).

Similarly, the neuroscientist Ed Lein emphasizes the superiority of those approaches to traditional ways of classifying cells:

> "...traditional approaches to neuronal classification rely on single-cell anatomy and physiology, which are typically qualitative and under-sampled. Transcriptomics has recently offered an unbiased, quantitative, and high-throughput alternative" (Clevers *et al.* 2017, p. 256).

And the authors of the article observing a "crisis" of single-cell biology, already quoted above, explicitly mention this as one of the desiderata for a successful cell type classification:

> "In our opinion, we should be pushing for a cell type nomenclature that meets some of the same key criteria as Linnaean taxonomy, as well as additional ones, including: (1) accommodating all cells arising during the life cycle of a given organism; (2) accommodating inter-individual variation, both normal and disease-related; (3) relating cell types to one another in a biologically meaningful way; (4) being stable to the incorporation of new data or new data types; and (5) *being constructed in a largely, if not entirely, data-driven manner* (Domcke & Shendure 2023, p. 1104, emphasis added).

Thus, there seems to be a common understanding that a purely data-driven classification of cells is both

desirable and feasible. In the remainder of this section, I will challenge this common understanding, drawing in particular on lessons learned from the debate on taxonomy. I start by providing some necessary background on single-cell experiments and analyses. This is followed by two lines of argument. First, I argue that practice shows that biologists do not believe that these types of experiments provide sufficient evidence to refute or justify any typology classification claim. Instead, such claims are always validated by more conventional and "theory-based" methods. Second, I provide more principled reasons for why a purely data-driven account of classification is destined to fail. These are analogous to some of the arguments that have been put forward against pheneticism in the context of taxonomy.

To avoid possible misunderstandings, I would like to point out at the outset that when I speak of "theory" in this context, I do not mean the traditional narrow sense of an axiomatic system based on laws of nature. Rather, the term "theory" here refers to any prior assumptions about underlying biological processes and mechanisms. Thus, a theory-free classification is one whose criteria depends solely on regularities in the observed data and do not presuppose any domain-specific knowledge. This corresponds to the use of "theory" and "theory-free" in the debates about the classification of organisms (see Ereshefsky 2000) and seems to capture the sense that contemporary biologists such as those cited above have in mind when they speak of "data-driven" or "unbiased" approaches.

3.1. A primer on single-cell experimentation and analysis

Progress in sequencing technology in the past two decades has enabled the quantification of gene expression on a genome-wide scale. To determine gene expression based on sequencing, the RNA isolated from a tissue is fragmented into small pieces that are afterwards sequenced in parallel. Counting the number of fragments that can be aligned to a particular gene sequence provides a quantitative proxy for the expression of that gene. Traditional RNA-sequencing (or "bulk" sequencing) experiments are based on mixed samples of thousands of cells and therefore provide an idea of the average expression of a gene in the sample. However, they do not provide information about the composition of the sample and about differences between individual cells. Single-cell sequencing technologies circumvent this problem by isolating single-cells in tiny droplets in a microfluidic device and adding a unique "barcode" sequence to each of them that allows the assignment of each RNA molecule to its cell of origin. Single-cell experiments thus provide a much higher-resolution image of gene expression in a population of cells. The result of a single-cell experiment is typically represented in the form of a large count matrix in which columns correspond to the individual cells and rows correspond to genes. Thus, each entry in this matrix indicates the number of reads (i.e., sequence fragments) of a particular gene in a particular cell. However, due to the small amounts of starting material, the resulting data are extremely noisy and sparse, which is to say that for any given cell in the sample a large fraction of genes will not be detected and appear as zeros in the count matrix. Therefore, perhaps paradoxically, single cell experiments cannot generally be used to obtain meaningful information about individual cells. However, they do provide information about the detailed structure of a cell population, which can be used to answer a range of biological questions.

The data analysis necessary to identify cell types based on single cell experiments consists of several steps. For the sake of brevity, I will focus on only two of them: dimensionality reduction and clustering. Other steps such as quality control, imputation, or normalization are part of the overall pipeline, but they are less directly related to the conceptual question at stake in this paper. In principle, each cell can be thought of as a data point in gene expression space, with each gene corresponding to one dimension of the space. Importantly, data analysis typically includes a step of dimensionality reduction. This means that the data are not analyzed and represented directly in the full gene expression space, but in a lower-dimensional space whose dimensions correspond to appropriate combinations of genes that capture important structural information in the given data set. Dimensionality reduction mitigates both the problem of noise and sparseness of data and the more fundamental "curse of dimensionality", which refers to the fact that as the number of dimensions increases, the distances between data points become more similar and thus less informative (Kiselev *et al.* 2019). Finally, dimensionality reduction makes subsequent analyses computationally more tractable. It corresponds to a

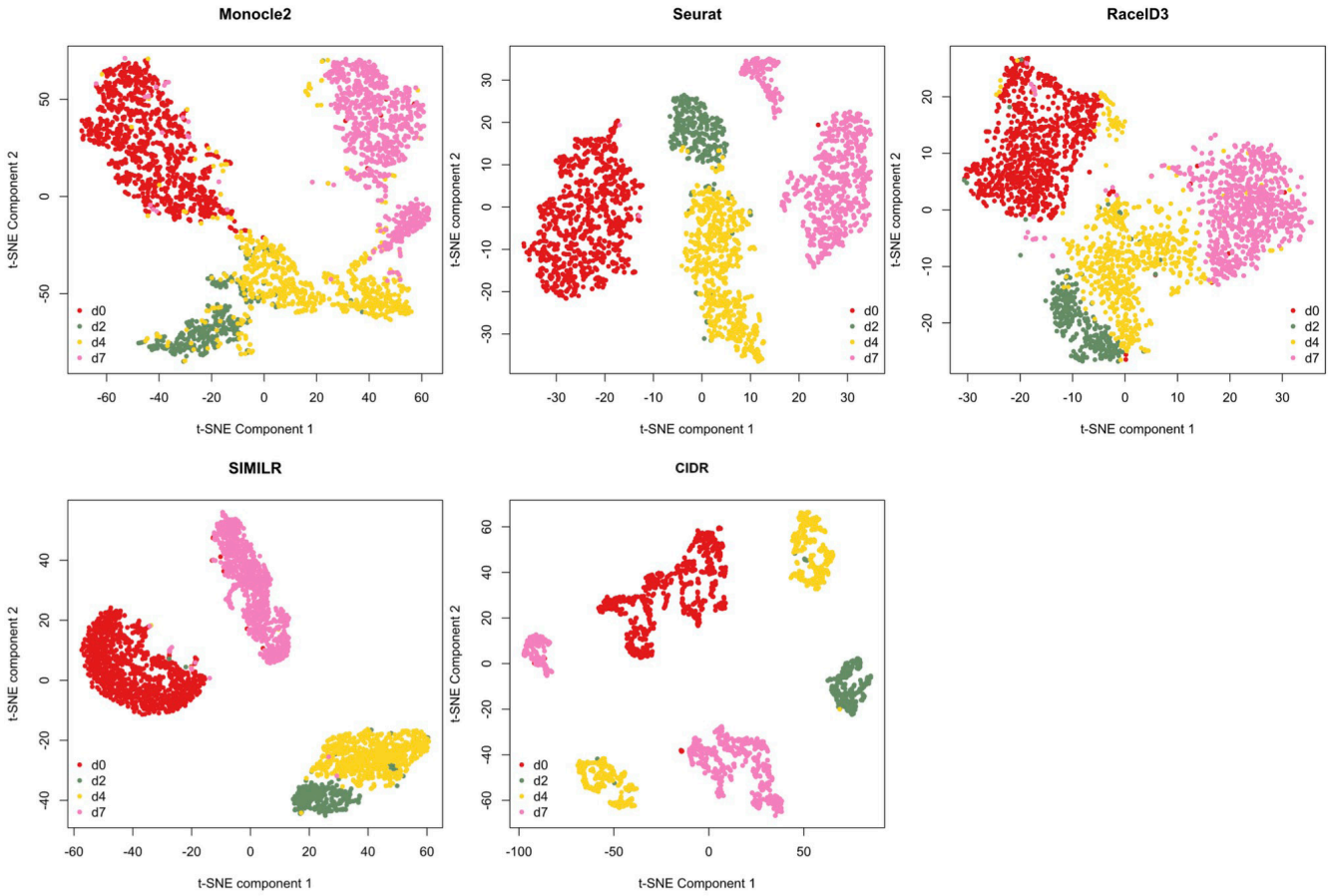

Figure 2: Comparison of different analysis pipelines for the identification of cell types based on a test dataset. Colors correspond to the "ground truth" annotations. Adapted from Zhang *et al.* (2023) (CC BY-NC 4.0)..

complex mathematical transformation of the data, which can be performed by various algorithms that may differ considerably in their results (Sun *et al.* 2019).

After dimensionality reduction, cluster analysis is used to identify cell types or other subsets of cells. In general, clustering methods are based on a measure of similarity between the objects to be clustered and then group the objects so that the objects in the same cluster are more similar to each other than the objects in other clusters. As with dimension reduction, there are a variety of different methods for performing this task, which may produce different results. Examples of widely used methods are k-means clustering and hierarchical clustering.

3.2. Are classifications based on single-cell methods really theory-free?

If one considers the current practice of biologists in establishing and using single-cell experiments to identify cell types, it quickly becomes apparent that they generally do not consider these experiments to provide a definition of cell types or to form a sufficient basis for classification. Instead, these methods are evaluated and calibrated based on previous biological knowledge. Thus, the choice of the appropriate clustering method and specific parameters is not imposed on scientists by the properties of the data alone. This means that the classifications resulting from these methods are not—strictly speaking—theory-free or unbiased, even though they are based on comprehensive and unsupervised methods of data analysis. The following statement from a recent review summarizes this current state of affairs:

> "Although considerable progress has been made in terms of clustering algorithms over the past few years, a number of questions remain unanswered. In particular, there is no strong consensus about what is the best approach or how cell types can be defined based on scRNA seq data" (Kiselev *et al.* 2019, p. 273).

One problem is that such methods usually base classifications on a specific class of molecular constituents, typically RNA, and neglect other potentially relevant classes (e.g. proteins or metabolites). However, the idea that a data set constrained in this way can lead to successful classifications amounts to an important theoretical assumption in itself that is not necessarily justified.

Another problem is due to the variety of methods that are available for important parts of the computational analysis, such as dimensionality reduction and cluster analysis. A theory-free account could be salvaged by assuming that all these methods lead to essentially the same classification. However, this does not seem to be the case. Figure 2 shows results from a study that compared different data analysis pipelines on the same data set. While there is clearly some agreement between the methods, the differences between results are perhaps even more striking. In particular, one can observe that cell groups, which are clearly separated by one method end up mixed or overlapping when another method is used.

Furthermore, even when focusing on one particular method alone, biologists are confronted with various choices. Dimensionality reduction obviously requires a decision on the dimension of the reduced space, which in turn affects the results of subsequent cluster analysis (Sun *et al.* 2019). Both too many and too few dimensions will lead to unsatisfactory results. An additional problem for some of these methods is that they are non-deterministic. For example, the widely used t-distributed stochastic neighbor embedding (tSNE) method is based on a non-deterministic algorithm, which means that different runs on the same dataset and with the same settings will lead to different lower-dimensional representations of the data (Zhang *et al.* 2023).

While clustering methods are usually considered as unsupervised methods, i.e., they identify features or structure in data without directly relying on prior information, they do rely on the choice of important parameter values. For example, k-means clustering requires the number of desired clusters (k) to be specified in advance, and most clustering methods rely on a distance measure between cells (when represented as data points in the reduced space), for which there are various possible choices. An obvious way out is to make the choice of these parameters automatic and data-driven as well, but then one has to choose the corresponding property to be optimized, for which again there are several possibilities.

A further issue is that clustering methods cannot be considered as completely devoid of biologically relevant assumptions. For example, k-means clustering relies on the assumption that there are discrete groups of cells in the first place, an assumption that is of course difficult to assess if one has no prior idea of the structure of the underlying cell population. Moreover, it tends to identify spherical clusters, which amounts to a strong assumption about the way in which cells of one type differ in their gene expression patterns. These assumptions can either lead to the failure to detect biologically relevant subpopulations (e.g., rare cell types) or, conversely, to the detection of spurious clusters.

The most important point, however, is that clustering methods in practice are evaluated based on a "ground truth", which consists in pre-labeled data sets (Zhang *et al.* 2023). As highlighted in the review cited above:

> "Perhaps the most challenging aspect of scRNA seq analysis (and this is not restricted to clustering) is how to validate a computational analysis method. The best strategy currently available is to have a setup where the cell types are known through other means, for example, by selecting cells from distinct cell lines, using tissues that are very well studied and understood (...), or considering cells taken from the earliest stages of embryonic development" (Kiselev *et al.* 2019, p. 278).

Thus, the choice of method and the specific settings are not determined based on "theory-free" considerations alone. Instead, it is importantly driven by the concern to reproduce previously accepted cell type classifications. If data-driven methods were indeed considered constitutive of the cell type concept, then the idea of an assessment based on a previously established baseline data set would not make sense, and other non-theoretical considerations would have to determine which method and settings should be used to identify and classify cells.

While it is possible that accepted single-cell based methods may subsequently be used to discover new cell types or even to correct and refine previous annotations, it seems inappropriate to refer to them as "theory-free" or purely data-driven as this would ignore the clearly theory-guided process of method selection.

It is also telling in this respect, that biologists usually do not accept the discovery of a new cell type based on single-cell experiments alone:

> "...for a new cell type to be accepted, it is necessary to go beyond characterization of the transcriptome. Researchers must demonstrate that the newly identified cluster is also functionally distinct. There are no universally applicable rules that can be applied here, and which assay is appropriate depends on the biological context" (Kiselev *et al.* 2019, p. 280).

This quote shows at the same time that some biologists seem to think that functional considerations that cannot be captured by gene expression data alone are relevant for cell type classifications, a point to which I will return later.

For the purpose of this paper, I can only hint at the full complexity of single-cell analysis, and I have neglected many aspects that may be considered equally relevant to the problem of identifying cell types. However, I think it has become clear that current scientific practices do not easily support the idea of a theory-free account of cell types.

3.3. General problems of a pheneticist approach

My previous arguments do not preclude the development in the future of methods based on single-cell experiments that can be considered theory-free in the relevant sense and accepted by biologists as truly constitutive of cell type classifications. In particular, an objection to the line of argument put forward in the previous section might be that it relies on the contingent imperfections of current single-cell technologies. Perhaps, an ideal single-cell experiment, unaffected by the noise and incompleteness of existing methods, could serve to build a satisfactory account of cell types. Consistent with this idea, some biologists have argued that the desired classification must be based on the integration of many different data modalities beyond gene expression, such as proteomic analysis and genome accessibility (e.g., Zeng 2022; Domcke & Shendure 2023). The underlying thought is that the more comprehensive data become, the more data-driven methods will approach the "natural" classification of cells into types.

In this section I will therefore move to some more principled reasons for doubting that this will be possible in a straightforward way. In particular, I will discuss some arguments that can be put forward against theory-free approaches in general, and in particular to the pheneticist approach to taxonomy. Further points take into account some specific features of the particular context of cell type classification.

One common argument against pheneticism is that the idea of "overall similarity" between the objects to be classified is not well-defined. Similarity is usually understood in terms of shared properties, but there is potentially an infinite number of properties that may be used for this assessment, and depending on the properties one chooses and how one weights their relative importance, one may arrive at very different and even diametrically opposed outcomes (Goodman 1972). The idea that this problem can be solved simply by measuring as many properties as possible rests on the tenuous "asymptote hypothesis". It states that, as the number of measured properties increases, the similarity converges to a constant value (Sneath 1995). The discussed "curse of dimensionality" illustrates the difficulties with this hypothesis. In defense of pheneticism, one might argue that the threat to a coherent notion of overall similarity is based on the mistaken idea that there is no restriction for allowed candidate properties, and that there is instead a set of "natural properties" on which a measure of similarity can be based (Lewens 2012). This latter move, however, presupposes prior ideas about which properties are biologically relevant; and while it might lead to a respectable version of pheneticism, clearly it would not be theory-free.

Another objection against pheneticism is that it is mistaken about the goals of taxonomy. The idea is that phenetic criteria of clustering organisms according to overall similarity will not pick out the evolutionarily salient actors. For instance, Ereshefsky (2000) points out that a pheneticist account would assign different developmental stages of the same organism or males and females of the same species to different groups. Similar considerations can be made for the case of cell types. It is conceivable that small differences in the expression of only a few genes can cause large phenotypic differences. On the other hand, there might be considerable differences in the transcriptomes of closely related cells due to stochastic variations or to transient differences (e.g. cell cycle stages). In such cases, it would be quite misleading to rely on a measure

of overall similarity which weights every feature equally. In response to such arguments, Lewens (2012) thinks that pheneticism should not be construed as a proposal that replaces other approaches to taxonomy that pursue specific goals. Rather, pheneticism provides a general-purpose taxonomy that allows for the investigation of more specific hypotheses regarding a variety of scientific problems. If general-purpose taxonomy clashes with groupings established by different means, this can be taken as a reason for refining the former. In the context of cell types, this might be an attractive option, notably in light of the fact that many biologists explicitly strive for an account of cell types that is universally applicable across all biological contexts (as manifested in the attempts to build comprehensive reference classifications, such as the human cell atlas). However, it should be clear that a classification obtained as a result of such a process of iterative refinement will not itself be theory-free. In addition, one may ask whether in the context of cell types there is a similar plurality of purposes as in taxonomy. One recurring idea in recent literature is that cell types should ultimately be defined in terms of their function (Clevers *et al.* 2017). Thus, if there is indeed overwhelming consensus that cell type classifications should track functional differences, then the argument of mistaken goals regains at least some of its bite. While it is plausible that divisions based on functional differences will roughly coincide with structurally defined differences, conflict between the two approaches is not at all excluded. Why then should biologists focus so much on a theory-free classification approach if that approach misses the central goal that cell type classifications are meant to achieve?

Finally, it should be noted that there are important differences between the questions faced by evolutionary biologists and those faced by biologists interested in classifying cell types. For example, many of the debates between different approaches to the taxonomy of organisms reflect the difficulty of inferring phylogenetic relationships because of incomplete evidence about past evolutionary events. Therefore, a pheneticist approach is attractive because it does not make any assumptions about unobservable events and processes. This problem is less severe in the case of ontogenetic relationships between cells because it is possible, at least in principle, to directly study the events involved in cellular differentiation and organismal development. The concern about independence from "theory", therefore, has a different urgency in evolutionary contexts because such theory usually involves weak hypotheses that likely will be overturned by new evidence.

All these considerations lead to think that even in the long run, single-cell technologies will not be able to provide a purely theory-free classification of cell types.

Conclusion

In this contribution, I considered the question whether and to what extent recent single-cell technologies challenge the notion of cell type. There is an intuitive concept of cell type that is based in some way on a combination of structural, functional, and developmental criteria. I have suggested that the cell type concept at different historical stages can be aligned with different approaches to classification. In particular, the idea of grounding cell classifications in the unbiased and theory-free clustering of single-cell data can be understood as the application of pheneticism to the context of cells. I have provided arguments to question that such a theory-free account can be achieved, both based on current scientific practice and on more principled grounds. It is interesting to see that concrete proposals of how cell types should be classified based on single-cell experiments are clearly theory-based in important ways. For example, the "periodic table" of cell types presented by Xia and Yanai (2019) does not use comprehensive gene expression, but relies on the idea of "core regulatory complexes" to provide the subsets of genes that are relevant for comparison. Similarly, Domcke and Shendure (2023) argue that a satisfactory description of cell identity must go beyond static molecular profiles and include information about ontogeny, i.e., the lineage tree of cells that corresponds to the development of the organism. Bioinformaticians are working on techniques to estimate phylogenetic relationships based solely on single-cell data (e.g., Farrell *et al.* 2018), but I strongly suspect that upon closer inspection these methods will not prove to be theory-free in the sense discussed in this paper either. I will save a more detailed discussion of this topic for a later occasion.

Does this mean that single-cell experiments do not affect cell type classifications at all? This does not seem plausible. However, overemphasizing the idea that a respectable approach to cell classification must be theory-free is wrong. One way for single-cell

experiments to affect cell classifications is by correcting particular assignments of cells to certain types. They simply provide additional information that may lead biologist to reconsider assignments they have made based on an incomplete evidence. The more interesting question, however, is whether single-cell experiments will affect classification criteria. This is less clear, but could be envisaged if one drops the requirement that classification should be theory-free. In particular, we could think of single-cell experiments as a way to iteratively refine the traditional concept, rather than replace it. Once an analysis pipeline has been validated based on test data of prior classifications, it can be used to make predictions on unseen data. While in case of mismatch previous classifications or biologists' intuitions might initially be given more weight, in the long run one may end up with a "reflective equilibrium" that represents the best compromise between fit and certain theoretical desiderata. The analysis method would then effectively be a theory that embodies, extends, and systematizes biologists' prior intuitions about what a cell type is.

References

Andrews, PW 2002, "From teratocarcinomas to embryonic stem cells", *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences*, vol. 357, no. 1420, pp. 405–417. https://doi.org/10.1098/rstb.2002.1058.

Arendt, D, Musser, JM, Baker, CVH, Bergman, A, Cepko, C, Erwin, DH, Pavlicev, M *et al.* 2016, "The origin and evolution of cell types", *Nature Reviews Genetics*, vol. 17, no. 12, pp. 744–757. doi: 10.1038/nrg.2016.127.

Baskin, DG 2015, "A historical perspective on the identification of cell types in pancreatic islets of Langerhans by staining and histochemical techniques", *Journal of Histochemistry & Cytochemistry*, vol. 63, no. 8, pp. 543–558. https://doi.org/10.1369/0022155415589119.

Bruce, A, Johnson, A, Lewis, J, Morgan, D, Raff, M, Roberts, K, & Walter, P 2015, *Molecular biology of the cell*, 5th ed. New York: Garland Science.

Burian, RM 1993, "Technique, task definition, and the transition from genetics to molecular genetics: Aspects of the work on protein synthesis in the laboratories of J. Monod and P. Zamecnik", *Journal of the History of Biology*, vol. 26, no. 3, pp. 387–407. https://doi.org/10.1007/BF01062055.

Canguilhem, G 1995, *Études d'histoire et de philosophie des sciences*, 5th ed. Paris: Vrin.

Chan, JKC, Ng, CS, & Hui, PK 1988, "A simple guide to the terminology and application of leucocyte monoclonal antibodies", *Histopathology*, vol. 12, no. 5, pp. 461–480. https://doi.org/10.1111/j.1365-2559.1988.tb01967.x.

Clevers, H, Rafelski, S, Elowitz, M & Lein, E 2017, "What is your conceptual definition of 'cell type' in the context of a mature organism?", *Cell Systems*, vol. 4, no. 3, pp. 255–259. https://doi.org/10.1016/j.cels.2017.03.006.

Cowdry, EV, ed. 1924. *General cytology: A textbook of cellular structure and function for students of biology and medicine*, Chicago: University of Chicago Press.

Domcke, S, & Shendure, J 2023, "A reference cell tree will serve science better than a reference cell atlas", *Cell*, vol. 186, no. 6, pp. 1103–1114. https://doi.org/10.1016/j.cell.2023.02.016.

Duchesneau, François. 1987. *Genèse de la théorie cellulaire*. Paris: Vrin.

Ereshefsky, M 2000, *The poverty of the Linnaean hierarchy: A philosophical study of biological taxonomy*, Cambridge: Cambridge University Press.

Farrell, JA, Wang, Y, Riesenfeld, SJ, Shekhar, K, Regev, A, & Schier, AF, 2018, "Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis", *Science*, vol. 360, no. 6392, pp. eaar3131. https://doi.org/10.1126/science.aar3131.

Fishell, G, & Heintz, N 2013, "The neuron identity problem: Form meets function", *Neuron*, vol. 80, no. 3, pp. 602–612. https://doi.org/10.1016/j.neuron.2013.10.035.

Goodman, N 1972, *Problems and projects*, Indianapolis: Bobbs-Merrill.

Huang, S 2009. "Reprogramming cell fates: Reconciling rarity with robustness", *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, vol. 31, no. 5, pp. 546–560. https://doi.org/10.1002/bies.200800189.

Hull, DL 1965, "The effect of essentialism on taxonomy—Two thousand years of stasis", *The British Journal for the Philosophy of Science*, vol. 15, no. 60, pp. 314–326. https://doi.org/10.1093/bjps/XV.60.314.

Hull, DL 1976, "Are species really individuals?", *Systematic Biology*, vol. 25, no. 2, pp. 174–191. https://doi.org/10.2307/2412744.

Jones, EG 1999, "Golgi, Cajal and the neuron doctrine", *Journal of the History of the Neurosciences*, vol. 8, no. 2, pp. 170–178. https://doi.org/10.1076/jhin.8.2.170.1838.

Kauffman, SA 1974, "The large scale structure and dynamics of gene control circuits: An ensemble approach", *Journal of Theoretical Biology*, vol. 44, no. 1, pp. 167–190. https://doi.org/10.1016/S0022-5193(74)80037-8.

Kauffman, SA 2004, "A proposal for using the ensemble approach to understand genetic regulatory networks", *Journal of Theoretical Biology, Special Issue in Honour of Arthur T. Winfree*, vol. 230. No. 4, pp. 581–590. https://doi.org/10.1016/j.jtbi.2003.12.017.

Kiselev, VY, Andrews, TS, & Hemberg, M 2019, "Challenges in unsupervised clustering of single-cell RNA-seq data", *Nature Reviews Genetics*, vol. 20, no. 5, pp. 273–282. https://doi.org/10.1038/s41576-018-0088-9.

Lancaster, C 2017, *A history of embryonic stem cell research: Concepts, laboratory work, and contexts*, Doctoral thesis, Durham University.

Lanier, LL, Engleman, EG, Gatenby, P, Babcock, GF, Warner, NL & Herzenberg, LA, 1983 "Correlation of functional properties of human lymphoid cell subsets and surface marker phenotypes using multiparameter analysis and flow cytometry", *Immunological Reviews*, vol. 74, no. 143, pp. 143-160.

Laplane, L, & Solary, E 2019, "Towards a classification of stem cells", *eLife*, vol. 8, article e46563. https://doi.org/10.7554/eLife.46563.

Lewens, T 2012, "Pheneticism reconsidered", *Biology & Philosophy*, vol. 27, no. 2, pp. 159–177. https://doi.org/10.1007/s10539-011-9302-2.

Maehle, AH 2011, "Ambiguous cells: The emergence of the stem cell concept in the Nineteenth and Twentieth centuries", *Notes and Records of the Royal Society of London*, vol. 65, no. 4, pp. 359–378.

Morris, SA 2019, "The evolving concept of cell identity in the single cell era", *Development*, vol. 146, no. 12, article dev169748. https://doi.org/10.1242/dev.169748.

Murphy, PR, Narayanan, D, & Kumari, S 2022, "Methods to identify immune cells in tissues with a focus on skin as a model", *Current Protocols*, vol. 2, no. 7, article e485. https://doi.org/10.1002/cpz1.485.

Neto, C 2020. "When imprecision is a good thing, or how imprecise concepts facilitate integration in biology", *Biology & Philosophy*, vol. 35, no. 6, p. 58. https://doi.org/ 10.1007/s10539-020-09774-y.

National Human Genome Research Institute 2022, "Fibroblast". Available from: www.genome. gov/genetics-glossary/Fibroblast.

Novik, AA, Ionova, TI, Gorodokin, G, Smolyaninov, A, & Afanasyev, BV 2009. "The Maxow 1909 centenary: A reappraisal", *Cellular Therapy and Transplantation*, vol. 1, no. 3, pp. 31–34.

Sánchez Alvarado, A, & Yamanaka, S 2014, "Rethinking differentiation: Stem cells, regeneration, and plasticity", *Cell*, vol. 157, no. 1, pp. 110–119. https://doi.org/10.1016/j.cell.2014.02.041.

Sneath, PHA 1995, "Thirty years of numerical taxonomy", *Systematic Biology*, vol. 44, no. 3, pp. 281–298. https://doi.org/10.1093/sysbio/44.3.281.

Sun, S, Zhu, J, Ma, Y, & Zhou, X 2019, "Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis", *Genome Biology*, vol. 20, no. 1, p. 269. https://doi.org/10.1186/s13059-019-1898-6.

Trapnell, C 2015, "Defining cell types and states with single-cell genomics", *Genome Research*, vol. 25, no. 10, pp. 1491–1498. https://doi.org/10.1101/gr.190595.115.

Van den Berge, K, Hembach, KM, Soneson, C, Tiberi, S, Clement, L, Love, MI, Patro, R, & Robinson, MD 2019, "RNA sequencing data: Hitchhiker's guide to expression analysis", *Annual Review of Biomedical Data Science*, vol. 2, no. 1, pp. 139–173. doi:10.1146/annurev-biodatasci-072018-021255.

Waddington, CH 1957, *The strategy of the genes: A discussion of some aspects of theoretical biology*. London: Allen & Unwin.

Wimsatt, WC 2007, *Re-engineering philosophy for limited beings: Piecewise approximations to reality*. Cambridge, MA: Harvard University Press.

Xia, B, & Yanai, I 2019, "A periodic table of cell types", *Development*, vol. 146, no. 12, article dev169854. https://doi.org/10.1242/dev.169854.

Yan, F, Powell, DR, Curtis, DJ, & Wong, NC 2020, "From reads to insight: A hitchhiker's guide to ATAC-seq data analysis", *Genome Biology*, vol. 21, no. 1, p. 22. https://doi.org/10.1186/s13059-020-1929-3.

Zeng, H 2022, "What is a cell type and how to define it?", *Cell*, vol. 185, no. 15, pp. 2739– 2755. https://doi.org/10.1016/j.cell.2022.06.031.

Zhang, S, Li, X, Lin, J, Lin, Q, & Wong, KC 2023, "Review of single-cell RNA-seq data clustering for cell-type identification and characterization", *RNA*, vol. 29, no. 5, pp. 517–530. https://doi.org/10.1261/rna.078965.121..