SAPIENZA
UNIVERSITÀ DI ROMA

Special Issue, Single-cell Analysis: Epistemological Inquiries

Vol. 6, No. 2 (2023)
ISSN: 2532-5876
Open access journal licensed under CC-BY
DOI: 10.13133/2532-5876/18129

Single-cell Molecular Analysis: When an Experimental Technique Reveals Conceptual Controversies

András Páldi[ab*] *& Laëtitia Racine*[ab]

[a] *Ecole Pratique des Hautes Etudes - PSL, Paris, France*
[b] *Centre de Recherche Saint-Antoine, Paris, France*

***Corresponding author:** András Páldi, Email: Andras.Paldi@ephe.psl.eu

Abstract

The last decade has witnessed a rapid evolution of highly sensitive single-cell molecular analysis techniques. These techniques allow the simultaneous detection and quantification of mRNA and protein molecules in a large number of individual cells. Some of these methods are already commercialized, making them readily available to any interested lab. While the pitfalls concerning the experimental extraction of biocomponents (mRNA and protein) and analytical bioinformatic methods are widely discussed in the literature, little is known regarding the conceptual difficulties raised by single-cell methodologies. Considered and treated as pure technical difficulties, these issues are rarely discussed explicitly. This is a problem as conceptual difficulties precede technical ones and contribute, to a large extent, to the failure of techniques. Consequently, a new theoretical framework is urgently needed to make sense of the ever-increasing amount of data.

Keywords: ontology, cell type, cell classification, single-cell technology

Citation: Páldi, A, & Racine, L 2023, "Single-cell Molecular Analysis: When an Experimental Technique Reveals Conceptual Controversies", *Organisms. Journal of Biological Sciences*, vol. 6, no. 2, pp. 25–30. DOI: 10.13133/2532-5876/18129

While central to biology, the process of cell differentiation is still not understood. The traditional molecular biology approach to differentiation uses a detailed description of gene expression changes and their correlation to the cell's morphological and physiological characteristics (phenotype). Study of individual cells has become a standard procedure for the investigation of a number of biological questions including differentiation. Single-cell techniques are considered as the best way to discover new and rare cell types, identify their differentiation pathways and the clonal structure of these cell populations (Mincarelli *et al.* 2018).

Multicellular organisms are composed of a large number of phenotypically different cells usually sorted in distinct categories called "cell types". The classification of living organisms and their parts is at the basis of biology as a science. The first classification of biological species was proposed by Carl von Linné in the 18th century in his work *Systema Naturae*. The system was based on hierarchical ranking of the living organisms in classes, orders, genera, species, and varieties. Linné's system, based on the similarity between the entities at each level of the hierarchy, is a perfect application of the essentialist ontology originally proposed by Aristotle and dominant in Western thinking since antiquity. Although Linné's binomial nomenclature is still in use nowadays, the system of classification based on similarity has been questioned by the Darwinian theory of evolution. Darwin proposed a new way of classification based on descent rather than similarities. In this classification different entities belong to the same category if they are derived from the same ancestor. The Darwinian view emphasizes the importance of individuals instead of categories defined on the basis of a set of properties shared by all individuals. Species and higher taxa are reduced to a pragmatic and artificial category made for convenience. As put by Darwin in the Chapter 14 of the *On the Origin of Species*: "In short, we shall have to treat species in the same manner as those naturalists treat genera, who admit that genera are merely artificial combinations made for convenience. This may not be a cheering prospect, but we shall at least be freed from the vain search for the undiscovered and undiscoverable essence of the term species" (Darwin 1859, p. 485). The individuals that are usually classified within the same taxonomic category called species are better characterized by their genealogical proximity rather than their resemblance. As a result, the boundaries between species became blurred. How many generations separate two individuals of two different species? The answer to this question is a matter of convenience, there is no universal rule. We can consider any morphological, functional or genetic characteristics—the result is always circumstantial (Mallet 1995; Mayr 1996).

It is difficult not to notice the analogy between the concept of species and that of cell types. The fact that a multicellular organism always develops from a single initial cell leaves no doubt about the common origin of all cells of the body. Early studies of the embryo development first identified the three germ layers, ectoderm, mesoderm and endoderm, then the specific structures – the organs – derived from them. From the 19th century until recently, embryologists investigated the origin of the organs, tissues and cell lineages during development. As a result, the classification of the tissues and their cell types was naturally based on origin, rather than on similarity. Embryology textbook illustrations represent germ layers, organs, tissues and cells in a hierarchical graph reminiscent of a genealogical tree. This is a Darwinian way of considering cell types. In parallel, cell biologists, anatomists and physiologists used the well-known classification method based on morphological and functional features. The two visions co-existed and complemented each other until the last decades of the 20th century. When molecular biology became dominant in life sciences, then the situation changed. According to the molecular genetic vision, cells are controlled by a program that is "hard wired" in the genes, and differentiation is a process of this program. Hence, same-type cells must express the same genes and can be identified on the basis of the transcriptional regulator (transcription factors) that they express (Davidson & Erwin 2006).

When flow cytometry, the first single-cell analysis method, was introduced, it was generally admitted that a variation of the mRNA or of the protein expression in same type cells was a simple stochastic fluctuation and a cell type was represented by the average of these parameters (Levsky & Singer 2003). A flow cytometer provides rapid analysis of multiple parameters with physical and chemical characteristics on single cells such as size, granularity and surface protein profile. Usually, it measures the fluorescence intensity emitted by specific surface proteins labelled with a fluorescent tag, generally an antibody. The fluorescence intensity is proportional to the number of molecules on the cells' membrane. This approach allows to label and measure several proteins in a single run, thus obtaining single-cell information from a large number of individual cells. The analysis of the results is typically performed using graphical plots. The most striking systematic observation brought by this technique is the large variation between single-cell values. This means that the amount of any expressed protein varies systematically on an unexpectedly large scale even between cells belonging to the same clonal population. However, it is common to convert data to a logarithmic scale to simplify data representation, which inherently reduces the apparent variation rendering it irrelevant. In fact, most of the actors in the field used to consider (and many still consider) same type cells and same clonal population to be essentially identical. In their opinion, any observed variation comes from measurement noise or size differences due to cell cycle. Groups of cells are defined on a graph using a procedure called "gating". This is mostly guided by the subjective appreciation of the fluorescence intensity of the cells. Even though some procedures based on multi-parametric algorithms exist, these are not widespread, and most experts are still using software like Kaluza or Flowjo in which the gate definition is done by hand in a subjective manner.

These groups are considered as different cell types or subtypes and are subject of further investigation to determine their biological properties. Their analysis usually focuses on the average value of the selected population of cells, so the individual cell-specific information gets lost. Average is considered as a kind of "essence" or "norm" of the cell type that shows how each cell would be on its own if the biological noise was irrelevant to individual variations. Genealogical relationships for the classification of cell types are usually not considered. Finally, the first technique aimed at studying single cell characteristics is used to provide population average, and the cell types defined in this way are approximate categories based on the subjective assessment of similarities between cells. The problem of information loss by the use of averages has already been recognized in biology long time along (Benzer 1953). A detailed discussion of the mathematical inadequacy of using average in biology can be found in (Rauch, Wattis & Bray 2023).

Thanks to the ability to amplify individual nucleic acid molecules by polymerase chain reaction, the resolution of the usual molecular detection techniques has increased more recently. Moreover, numerous new methods emerged and made possible the simultaneous detection and quantification of the whole sets of mRNA molecules, chromatin structural profiles, proteins etc. in a vast number of individual cells. Many authors consider that this technological advance represents an opportunity to redefine and systematically detect cell types (Wagner, Regev, & Yosef 2016; Morris 2019). The amount of single-cell resolution data generated by these experimental techniques is much higher than what is provided by flow cytometry, because they detect more features in a single cell. While flow cytometry can detect the abundance of a limited number of proteins in a single cell, the new technology can detect the approximate number of RNA transcripts of each individual gene in each individual cell simultaneously in a large number of cells. Each cell is described then by as many features as the number of genes, and the resulting data set may contain several hundred million of data points. As it is impossible to analyze huge amount of data by simple visual inspection on a graphical display, sophisticated computational analysis methods are required. However, those modern techniques did not immediately resolve a fundamental question: how to differentiate different cell types? As indicated above, this question is an adaptation to cell biology of a fundamental question of philosophical ontology about entities and identities. Over the last few years the question of cell types has become a subject of intense discussion among biologists (Mincarelli *et al.* 2018; Wagner, Regev, & Yosef 2016; Morris 2019; Han *et al.* 2020; Xia & Yanai I 2019). Surprisingly, however, the nature of the difficulties in answering the question about cell types in most cases is considered technical. For example, some authors explicitly declare that "classification" of cells into discrete types from single-cell profiles is a problem of "unsupervised clustering in high dimensions" (Wagner, Regev, & Yosef 2016). The pre-Darwinian essentialist way of conceptualizing cell types is never questioned. Instead, it is admitted that each individual cell in the organism can be assigned to a well-defined class. This classification is considered as one of the primary objectives of single-cell technologies (Mincarelli *et al.* 2018). A significant effort is made to establish cell catalogues (cell atlases) of various multicellular organisms (for example: www.humancellatlas.org). The cells are grouped on the basis of the similarity of their gene expression patterns, a unique "ID card" for each cell type. In other words, cells belonging to the same type are supposed to share a minimal set of expressed genes. The overall difference between the gene expression patterns of the cells isolated from different organs or tissues of the developing embryo or adult organism is easily distinguishable. However, distinguishing groups of cells with clearly different gene expression patterns from a mixture of cells isolated from the same tissue is far more difficult. Perhaps, the best illustration comes from the study of the human hematopoietic stem cells lineage, that demonstrated the highly variable and continuous nature of mRNA profiles between cells considered as different cell types on the basis of their functional characteristics (Velten *et al.* 2017). A very high number of cells have intermediate gene expression patterns, that is, no minimal set of genes is expressed only in a well-defined group or cluster. Highly likely, a gene expression pattern does not allow identifying the type of a cell randomly picked up from a population. Whatever the mathematical method to cluster the data, some subjective decision is always required and the final result depends on the choice of some key parameters used by the algorithm (p-values, thresholds, filters, the presumed number of clusters one expects, etc.) (Luecken & Theis 2019; Breda, Zavolan, & van Nimwegen 2021). As a result, the number of identifiable cell clusters depends as much on those biased parameters as on data. This procedural subjective component rarely emerges during discussions about cell type analysis methods. In the light of this, it is not surprising that more and more studies report on new rare cell types identified based on single-cell data analysis. There is a high risk that those discoveries are in fact the results of an over interpretation of single-cell data. This problem illustrates the ambiguities of the "cell type" concept defined solely by single cell mRNA profiling. It also shows how the misuse of a concept imposes limitations to our thinking by canalizing the discussions on the technical aspects, leading to the conclusion that collecting more data will solve the difficulties.

To circumvent the problem of rare cell types, one of the most popular *ad hoc* explanations proposed is to further divide the cell type into smaller categories named "cell states", etc. The idea is that single-cell data represent a snapshot of the studied population and rare cell profiles may represent a short-lived transitory cell state. For example, Wagner and colleagues "refer to the more permanent aspects in a cell's identity as its type (e.g., a hepatocyte typically cannot turn into a neuron) and to the more transient elements as its state. Cell types are often organized in a hierarchical taxonomy, where types may be further divided into finer subtypes; such taxonomies are often related to a cell fate map, reflecting key steps in differentiation" (Wagner, Regev, & Yosef 2016). Unfortunately, further dividing a population of cells into "types" or "states" changes nothing to the initial problem since it does not provide any better solution for the classification. This way of categorizing cell types and states into hierarchies merely changes the name of the class from "type" to "state" and suffers from the same conceptual shortcomings as exposed above. As the species concept in population biology, this vision of cell types has operational utility when applied to whole cell populations or whole organisms, but fails when individual cells are considered. Paradoxically, single-cell technologies revealed that the "cell type", contrary to what its name suggests, is a concept that describes the features of a cell population and not that of an individual cell. For purely pragmatic reasons, cell populations but no single cells can be grouped on the basis of their gene expression profiles. The concept of cell type as mentioned above is the closest biological analogy of the concept of "species". Cell "type" can fit a group of cells based on their biological function. It can emerge from the characteristics of individual cells but it is inapplicable to them in the same way as the concept of "pressure" describes a gas but cannot be applied to individual gas molecules. The way the cell type is inferred from the single cell mRNA data captures in fact some kind of "average", a rather statistical reflection of a more or less arbitrary chosen population of cells. Contrary to what is asserted by many authors, the average does not describe the intrinsic, functional and context-independent biological features of individual cells. Claims such as: "it is possible to practically define cell types according to their expressed transcription factors (TFs)" (Xia & Yanai 2019) are simply not supported by observation (Weinreb, Rodriguez-Fraticelli, Camargo, & Klein 2020). We do not know yet how individual cells behave and to what extent they can change their function and morphology (what we call "phenotype"). Single-cell mRNA profiling is a simple "cross section" of a temporal process at a given time-point; alone it cannot provide the information many experts expect without taking into account the temporal character of the cell, her lineage history and environment.

Therefore, calling for the revision of the "cell type" concept is one of the unrecognized but important contributions of single-cell technologies. Such a revision is, however, impossible without rethinking another key concept, i.e. "cell identity". The identity of an individual cell—its phenotype—is not simply an intrinsic property of the cell that can be deduced from its molecular composition. The cell is continuously interacting with the biological (the other cells), physical (intracellular matrix) and chemical (available nutrients, oxygen, pH, etc.) micro- and macro-environments. These interactions act as extrinsic constraints; their changes promote and canalize the phenotypic change of the cells. In turn, the cell also modifies its environment, forming in this way a complex interacting system. On the other hand, the phenotype is also constrained by the cell's own life history and genealogy, conveyed by what is usually called cellular or epigenetic memory. Cellular memory represents an intrinsic limitation to the change by restraining the repertoire of genes that can be easily expressed (Páldi 2020). As a result, at any moment, the cell phenotype is determined by the outcome of the interplay between intrinsic and extrinsic constraints and reflects a dynamic equilibrium of rapid change-promoting and inhibiting processes. The phenotype encompasses the whole life cycle of the cell and it is impossible to specify the exact moment when the "true identity" appears. Therefore, the phenotype or "identity" of a cell is better described as a dynamic equilibrium of many different and frequently opposing processes than as a static state (Dupré & Nicholson 2018). Recent observations suggest that the transition of the cell toward a new phenotype, usually called fate choice or differentiation, is indeed highly dynamic and not a simple switch as previously thought (Moussy *et al.* 2017; Parmentier *et al.* 2022). It is more like a trial-and-error process based on the permanent dynamic exchange between the cell and the micro-environment. From the point of view of the "process", the capacity to change requires no specific explanation as variation is its true nature. What requires explanation however, is the lack of change, i.e., stability, the equilibrium of the antagonistic processes. Only if the equilibrium is maintained for an extended period, then the cell morphology and function appear stable. It may be tempting to consider the cell's appearance during this period as the "true" phenotype. Nevertheless, a simple snapshot is unsuitable to determine the stability of a cell phenotype. This is only possible based on continuous observation over a period of time or using a time series of snapshots of the same cells. It is worth remembering however, that "stable" or "transient" depends entirely on the time scale of these observations. There is no privileged time scale. If the frequency of the snapshots is lower and the period of observation is longer, the rapid changes are not detected and the proportion of different morphologies or gene expression patterns will appear constant in a cell population (Brock, Chang, & Huang 2009). Current single-cell mRNA detection technologies provide only a single snapshot for an individual cell because they are invasive to the point of destroying the cells during the analysis. Although there are promising attempts to overcome this limitation (Chen *et al.* 2022; Boersma *et al.* 2019; Lyon, Aguilera, Morisaki, Munsky & Stasevich 2019), it is currently impossible to repeat the same measurement on the same cell or repeat the analysis of the same cell at a later point. Taken together, these considerations suggest that single-cell molecular approaches, as they stand today, can only be used to follow the general trend of changes if applied to a time series of pre-defined groups or cell populations. These general trends tell us little about the trajectory of individual cells; they only allow for conjectures.

Over the past decade, single-cell molecular technologies have produced a huge amount of data. Although this gives us the illusion of knowledge, only a small fraction of such information is really exploited to improve our understanding of the process of cell differentiation. What we really need now is a new interpretation framework based on solid theoretical ground to develop analytic methods and go beyond the calculation of gene expression profile and resemblance between groups of cells. Such a method should establish a true association between the single-cell gene expression pattern and the individual cell's phenotype that can be used for functional studies. A promising way to build a new paradigm is to capitalize on the organicist tradition of the pre-molecular biology period, as suggested by several authors (Dupré & Nicholson 2018). As Paul Weiss put it: "Life is a dynamic process. Logically, the elements of a process can be only elementary processes, and not elementary particles or any other static units"... Life "can never be defined in terms of a static inventory of compounds, however detailed, but only in terms of their interactions" (Allen 1962).

References

Allen, JM (ed.) 1962, *The Molecular Control of Cellular Activity*. New York - Toronto - London: McGraw Hill. Available from: https://archive.org/details/molecularcontrol0000alle/mode/2up.

Benzer, S 1953, "Induced synthesis of enzymes in bacteria analyzed at the cellular level", *Biochimica et Biophysica Acta*, vol. 11, no. 3, pp. 383–395. Available from: https://doi.org/10.1016/0006-3002(53)90057-2.

Boersma, S, *et al.* 2019, "Multi-color single-molecule imaging uncovers extensive heterogeneity in mRNA decoding", *Cell*, vol. 178, no. 2, pp. 458–472.e19. Available from: https://doi.org/10.1016/j.cell.2019.05.001.

Breda, J, Zavolan, M, & van Nimwegen, E 2021, "Bayesian inference of gene expression states from single-cell RNA-seq data", *Nature Biotechnology*, vol. 39, no. 8, pp. 1008–1016. Available from: https://doi.org/10.1038/s41587-021-00875-x.

Brock, A, Chang, H, & Huang, S 2009, "Non-genetic heterogeneity—a mutation-independent driving force for the somatic evolution of tumours", *Nature Reviews. Genetics*, vol. 10, no. 5, pp. 336–342. Available from: https://doi.org/10.1038/nrg2556.

Chen, W, *et al.* 2022, "Live-seq enables temporal transcriptomic recording of single cells", *Nature*, vol. 608, no. 7924, art. 7924. Available from: https://doi.org/10.1038/s41586-022-05046-9.

Darwin, C 1859, *On the Origin of Species*, London: John Murray.

Davidson, EH, & Erwin, DH 2006, "Gene regulatory networks and the evolution of animal body plans", *Science*, vol. 311, no. 5762, pp. 796–800. Available from: https://doi.org/10.1126/science.1113832.

Dupré, J & Nicholson, DJ 2018, "A manifesto for a processual philosophy of biology", in: Nicholson, DJ, & Dupré, J (eds.) *Everything Flows: Towards a Processual Philosophy of Biology*, Oxford - New York: Oxford University Press, pp 3–46. Available from: https://doi.org/10.1093/oso/9780198779636.003.0001.

Han, X, *et al.* 2020, "Construction of a human cell landscape at single-cell level", *Nature*, vol. 581, no. 7808, pp. 303–309. Available from: https://doi.org/10.1038/s41586-020-2157-4.

Levsky, JM, & Singer, RH 2003, "Gene expression and the myth of the average cell", *Trends in Cell Biology*, vol. 13, no. 1, pp. 4–6. Available from: https://doi.org/10.1016/s0962-8924(02)00002-8.

Luecken, MD, & Theis, FJ 2019, "Current best practices in single-cell RNA-seq analysis: A tutorial", *Molecular Systems Biology*, vol. 15, no. 6. Available from: https://doi.org/10.15252/msb.20188746.

Lyon, K, Aguilera, LU, Morisaki, T, Munsky, B,& Stasevich, TJ 2019, "Live-cell single RNA imaging reveals bursts of translational frameshifting", *Molecular Cell*, vol. 75, no. 1, pp. 172–183.e9. Available from: https://doi.org/10.1016/j.molcel.2019.05.002.

Mallet, J 1995, "A species definition for the modern synthesis", Trends in Ecology & Evolution, vol. 10, no. 7, pp. 294–299. Available from: https://doi.org/10.1016/0169-5347(95)90031-4.

Mayr, E 1996, "What is a species, and what is not?", *Philosophy of Science*, vol. 63, no. 2, pp. 262–277. Available from: https://doi.org/10.1086/289912.

Mincarelli, L, Lister, A, Lipscombe, J, & Macaulay, IC 2018, "Defining cell identity with single-cell omics", *Proteomics*, vol. 18, no. 18, p. 1700312. Available from: https://doi.org/10.1002/pmic.201700312.

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

Moussy, A, *et al.* 2017, "Integrated time-lapse and single-cell transcription studies highlight the variable and dynamic nature of human hematopoietic cell fate commitment", *PLOS Biology*, vol. 15, no. 7, art. e2001867. Available from: https://doi.org/10.1371/journal.pbio.2001867.

Páldi, A 2020, "Random walk across the epigenetic landscape", in: Levine, H, Jolly, MK, Kulkarni, P, & Nanjundiah, V (eds.) *Phenotypic Switching: Implications in Biology and Medicine*, London - San Diego - Cambridge - Kidlington: Academic Press - Elsevier, pp. 53–76. doi: 10.1016/B978-0-12-817996-3.00008-6.

Parmentier, R, *et al.* 2022, "Global genome decompaction leads to stochastic activation of gene expression as a first step toward fate commitment in human hematopoietic cells", *PLOS Biology*, vol. 20, no. 10, art. e3001849. Available from: https://doi.org/10.1371/journal.pbio.3001849.

Rauch, C, Wattis, J, & Bray S 2023, "On the meaning of averages in genome-wide association studies: What should come next?", *Organisms*, vol. 6, no. 1, pp. 7–22. Available from: https://doi.org/10.13133/2532-5876/17811.

Velten, L, *et al.* 2017, "Human haematopoietic stem cell lineage commitment is a continuous process", *Nature Cell Biology*, vol. 19, no. 4, pp. 271–281. Available from: https://doi.org/10.1038/ncb3493.

Wagner, A. Regev, A, & Yosef, N 2016, "Revealing the vectors of cellular identity with single-cell genomics", *Nature Biotechnology*, vol. 34, no. 11, pp. 1145–1160. Available from: https://doi.org/10.1038/nbt.3711.

Weinreb, C, Rodriguez-Fraticelli, A, Camargo, FD, & Klein, AM 2020, "Lineage tracing on transcriptional landscapes links state to fate during differentiation", *Science*, vol. 367, no. 6479, art. eaaw3381. Available from: https://doi.org/10.1126/science.aaw3381.

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