



Towards a pragmatic definition of cell type.

This manuscript ([permalink](#)) was automatically generated from [lubianat/technotype@4243411](#) on September 9, 2020.

Authors

- **Tiago Lubiana**

 [0000-0003-2473-2313](#) ·  [lubianat](#)

Computational Systems Biology Laboratory, University of São Paulo

Abstract

One of the first classes in any undergraduate major in biomedical sciences is the histology. The students are tasked with identifying cell types across a range of tissues, looking for color and shape patterns in hematoxylin-eosin stains. The text-books, such as the one by Junqueira & Carneiro, work as “manuals” (in the Kuhnian sense) perpetuating our paradigms of this set of a couple hundred cell types that define the cells in human beings.

This anatomy based conceptualization, however, has been constantly challenged by new techniques. The development of techniques to study cells in the individual level, from multiplexed flow cytometry to patch clamping, opened our eyes for an amazing diversity. With a burst of new categories, novel cell “subtypes” and “families” started to pop up in the literature, enriching our view. This has become specially evident in the past few years, with the explosion of single-cell omics, and the rise of the Human Cell Atlas and the HUBMAP projects.

Even though our technologies diversified, our conceptualization of cell types are still based on century-old histochemical techniques, such as the aforementioned Hematoxylin-eosin staining and Golgi-stainings (immortalized in the drawings of Santiago Ramon y Cajal). We refer to cell types as anatomical entities, as if they were clearly dissectable and fixed in an organism. However, there has not yet been one single definition of cell type that could unite views across fields, even if for strictly pragmatic purposes.

The advances in computation and the large scale biology, however, requires us to find better alternatives for the answer of what a cell type is. Even if a ontological definition of a “true” cell type is unrealistic, we can strive to find nominal, pragmatic definitions for the pragmatic challenges of large scale biology. Otherwise, how can we properly label single-cell data? How can we judge claims of discovery of cell types? How can we integrate the knowledge from the millions of scientific articles published every year?

This need for conceptual work is clear, and a number of interesting proposals are arising. Arendt et al has focused on the evolutionary perspective of a cell type: by treating the cell type as an evolutionary unit defined by a Core Regulatory Complex (CoRC) of transcription factors, this view has enabled powerful parallels with the evolution of genes and species. Models of how multicellular life works greatly benefit from concepts such as “sister types”, “cell type homology” and “cell type convergence.”

However, in an analogous way to how different concepts of species serve better different research programs, our quest to define cell types benefits from an orientation. The challenge of representing cell types in the context of evolution is conceptually different from representing cell types in the context of biomedical experimentation. The Cell Ontology project and the International Workshop on Cells in Experimental Life Sciences have been doing important groundwork, that base many of the discussions that will be presented in this article.

In this work, the scope of the quest is the one of research synthesis. How to find a cell type definition that allows rigorous description of biomedical knowledge experiments?

For that goal, this article is divided in the following parts. In Part 1, I will bring a proposal of a set of rules that are sufficient for defining cell types. In Part 2, I'll propose a small set of names for differentiating main classes of cell types. In Part 3, I will address the logical consequences of the proposed definitions. And in Part 4, I will discuss the pragmatic challenges for employing such definitions.

A set of 3 + 1 rules for defining a cell type

In an opinion article published in *Cell Systems* in 2017, prominent researchers presented their views on the “Conceptual Definition of “Cell Type” in the Context of a Mature Organism?” If one consensus could be reached is that the task is widely open. From ditching the concept altogether, to conceptually exploring (a species in which we know the exact number and place of each cell expect in an adult) to ecological definitions of cell types, the work makes it clear how hard it is to find rules that would please all the community. Many advocate for a core role of the cells’ function in defining cell types, a slippery road, as the meaning of “function” in biology is elusive.

Aevermann et al (2018) come up with a set of rules to classify samples in single cell RNA-Seq, based on consisted of “The minimum set of necessary and sufficient marker genes selectively expressed by the cell type”, “A parent cell class in the CL”, and “A specimen source description (anatomic structure & species).” This conceptual advance is practically important, solving many of the issues raised before by Bakken et al (2017) and others. The Cell Ontology also has used gene and protein markers as descriptors for defining cell types.

The use of markers, however, leaves us with a conceptual problem, when we want find a global definitions: definitions of cell type used by electrophysiologists, or even in the Kuhnian manuals of histology classes, are not based on markers. Rigorously, this would leave aside a whole part of what we consider biomedical knowledge. Moreover, markers are not, *strictu sensu*, defined for cell types that span multiple species, requiring us to consider homology. This adds an extra layer of confusion that will be addressed partially in Part 2.

My pragmatic definition of cell type (for eukaryotic, multicellular organisms) consists in 3+1 simple rules. A cell type is any class of cells which must be: - Rigorously defined - Theoretically useful - Identifiable for a defined taxon

And that should be: - Logically related to other cell types

In the following paragraphs, I will clarify the meaning and motivation of the rules.

For rule 1, by “rigorously defined”, I mean that a cell can be identified as belonging or not to the class given a set of criteria and we should be able to identify a cell from its cell type in any individual that matches the rigorous criteria. An example of such a rigorous criteria is “expression of the CD3 protein, expression of the CD4 protein, lack of expression of the CD8 protein.” This implies the need for using rigorous definitions of what is a “CD3 protein” and what “expression” means, but it is feasible to have nominal definitions. Actually, in practice, similar criteria are used in immunology for defining cell types.

Rule number 2 is a specification of a rigorousness: one rigorousness criteria is to define the taxons in which a given cell type is expected to have manifestations. I will call this set of taxons the of the cell type.

Knowing the scope is importance to avoid the pitfalls of assuming that theories corroborated with mice experiments are valid for other organisms. This is an instance of the classic problem of induction, which subsists much of the core arguments of Logic of Scientific Discovery. A specification of the defined would make inductual claims explicit.

These differences across species are currently left implicit, and it is not uncommon to find scientific abstracts that don’t mention a species at all.

Currently, the largest authority on cell type definitions, the Cell Ontology, provides “multispecies” definitions of cell types. These are conceptually important. However, labeling studies plainly with such concepts is dangerous. It is not safe to assume that a “mouse neutrophil” is simply a “neutrophil” that happens to be in a “mouse”. According to the rules of logic, combinations of “multispecies” cell types and specific taxons make up new concepts.

By providing a specific scope, we can make explicit what are currently implicit predictions. A question “Do whales lack neutrophils?” is falsifiable, and, intuitively, we would imagine that they have cells that could be classified as neutrophils. However, this question is only falsifiable if we have a rigorous definition of “neutrophil” across its scope of taxons. Conversely, if we say that “the scope of neutrophils is mammals” we are making a strong prediction that we should find cells of this category in any mammal. If we would apply severe tests to find neutrophils and whales and recurrently fail to identify one matching cell, we would have a good reason to discuss changing our scope of neutrophils.

Rule number 3 is a rule of practical concern. There is a massive amount of “rigorous classes” that one scientist might come up with, due to the combinatorial nature of classes, far outnumbering the reported number of atoms in the observable universe. For that reason, a criterion of usefulness is necessary for deciding when a class of cells is considered a cell type. The most fair criteria of usefulness is one based on the individual: a valid cell type is whatever class one individual rationally considers useful.

However, we can be more rigorous in defining usefulness, if we consider rule number 4: a cell type has to be anchored to other cell types. Which means that a definition of class is of negligible usefulness if it can't be considered a “subclass” of other cell type. For our practical concerns, all imaginable cell types are subclasses of “cell of eukaryonts”. Note that “human erithrocyte” is a cell type characterized by lack of a nucleus, so depending on your definition of “eukaryotic cell”, it would exclude “human erythrocyte”. Anyways, this anchoring is important for integrating knowledge across studies, because eventhoug a specific “transcriptomically-defined” cell type and a “electrophysiologically-defined” cell type cannot be rigorously said to be the same, they can be grouped in a “superclass” that contains cells that match either one or the other criteria.

The consequences of this set of criteria will be discussed further in the following sections.

Naming classes of cell types

Before we proceed to analyse the consequences of the criteria raised on part 1, it is importante to make a set of naming conventions for different classes of cell types. This is required, as much of the literature consider cell types in the context of one species (e.g. when dealing with a cell type as an evolutionary unit) as well as in the context of multispecies (such as in the case of cell ontology).

Current advances in the taxonomy of living beings are calling artificial the classifications of “genus”, “families”, “order”, and similar rankings. While the level of a “species” is somewhat better defined (with discernible theoretical positions, at least), the other categories comprehen only classes of different levels of coverage. In accordance with the theoretical views of the PhyloCode, I consider only the level of “species” on the naming of classes of cell types. The three classes I propose, are:

- Archetypes, for which the scope of the definition is beyond the level of species. For example, “mammal neutrophils”.
- cell types (or {bona fide} cell types), for which the scope of the definition is exactly one species. For example, “human neutrophils”

- Infratypes, cell types for which the scope is below the level of species. For example, considering the mouse strain “C57BL/6J”, “C57BL/6J neutrophils”.

These are common uses of “cell type” which are, nevertheless, mixed in practice. By adopting a more precise vocabulary we can flesh out misunderstandings and communicate clearly. At the level of individual scientific experiments we usually work at the “infratype”: the samples come only from a subpopulation of the species of interest, and cannot be assumed to be “randomly sampled” from all individuals. This has important practical considerations for, once more, avoiding failing implicitly for the problem of induction.

Moreover, in individual experiments, we not only work with infratypes, we work with very specific infratypes, guided by non-random research setups and pragmatic choices. For example, we might call “CD4 T cells” what are actually CD3+, CD4+, CD8+ CELLS from the axillary lymphnode of 2 month old female C57BL/6J mice from the mouse-house of the Institute of Biochemistry of the University of São Paulo collected in the morning around 10 pm for chow-fed ad libitum mice. Albeit really specific, all the mentioned facets (markers, anatomical location, age, biological sex, strain, housing conditions, circadian clock and diet) are known to alter the behaviour of cell types. Thus, it is necessary one more name, for clarity:

- Technotype: A specific, experimentally defined cell type, which harbours on its definitions the exact conditions from which the cell types were sampled.

Note: a technotype is still a class. Unless a study used only one single-cell, it likely contained some sampling method, which is the class for which hypothesis are actually tested, for example. This is the most “granular” cell type in our pragmatic view for research synthesis. This is the type that can be strictly annotated in single-cell RNA-seq analysis, for example.

Single claims are made and tested for technotypes, and this claim can be logically combined in “upper” ontological levels for making claims with a higher degree of universality. This propagation of knowledge to upper levels, however, should not be implicit. (See Yarkoni2020 for an analogous problem in the psychological sciences). Knowledge should travel “quasi-inductionally” by fostering hypothesis with higher degrees of generality, which can then be tested for the expanded class.

Logical consequences of the definition

One notable logical consequence of the proposed set of criteria is that the definition of a “cell state” is left aside altogether. For the pragmatic definition adopted here, we explicitly leave aside the dissection of the differences between persistent classes of cells (which I refer to as “traditional cell types”) or the transient, fugacious classes of cells (which I refer to as the “traditional cell state”. Even though such an explicit definition is definitely an important topic of theoretical research, it is not strictly necessary for the framework presented here.

One practical example is that, by the rules here proposed, the class “human cells in metaphase of mitosis” can be considered a cell type, as they can be rigorously defined, are restricted to a taxon and are theoretically useful. Even though “metaphase” itself is still a biological process, we can describe all cells that are executing this process as from a single cell type.

However, does a dividing fibroblast stop being a fibroblast, even if temporarily? Again, I do not aim to answer this in a physico-ontological sense. Pragmatically, if the rigorous definition used (e.g. expression of a marker) still holds during duplication, this cell can be assigned to two disjunct classes: “fibroblasts” and “doubling cells”. It is, thus, important to consider that cells can belong to at least 2 disjunct classes.

This has practical importance. Often, cell types are described taxonomically, as being related in one single hierarchy. However, if we considered that cells can be assigned to disjunct classes, it is not possible to annotate cell types with a single identifier using a taxonomic tree, in which each concept is represented by a single node with one (and only one) direct parent node. Cell types need to be represented ontologically (in the computational sense), which can be thought of multiple, intertwining taxonomies, taking into account different ways of classifying cells.

Another logical consequence of the definition is that concepts of “subtype” become redundant with “cell type”. A “subtype” then, is a concept that only makes sense when talking about classes with different degrees of universality. Thus, claims to discover new cell “subtypes” or “types” differ only stylistically and can be considered indistinguishable from the perspective of research synthesis.

Practical consequences of the definition

In the previous section, I discussed logical entailments of accepting the set proposed set of rules as valid. Here, I will extend the pragmatic considerations on using such a system for real world applications.

A first pragmatic application, already mentioned previously, is that cell type discovery claims can be organized. With vast amounts of data and loose definition of cell types, it becomes uncannily easy to claim to have discovered a new cell type. If one claims to discover a new “*stricto sensu*” cell type, one should provide enough evidence that cells from this class are identifiable across all organisms of a species. A claim of an “archetype” should provide evidence in more than one species. And, consequently, experiments that only use a specific strain of mice, can only claim to discover an infratype.

An example of the discovery of a new “archetype” is the pair of articles published in *Nature* in 2019 about the newly found “ionocyte”, a class of cells in the trachea enriched for expression of genes homologous to the gene. Both studies displayed evidence for the existence of such a class in both mouse and human samples, thus corroborating the existence of such an archetype. This discovery has been denominated by both articles as a discovery of a new cell type.

Another example of cell type discovery is an pioneer article by Villani et al that describes subclasses of monocytes and dendritic cells in humans, and pragmatically use markers for their definition. The patients were recruited from “the Boston-based PhenoGenetic project (...) and the Newcastle community.” Strictly speaking, it is arguable that they did not have a random sample of humanity, and the observed results might not hold for different populations. Moreover, it is clear that the different “types” described are not being purposed as archetypes. This discovery of infratypes has also been described as a discovery of a new “cell type”.

Thus, by stating the scope of the cell type discovered, and rigorously specifying its characteristics, a claim of discovery can be compared in the light of evidence, very much like it is done for centuries for animal species. However, how can we humanely understand with so many “cell types” with such subtle differences?

As described by Sabine Leonelly, the challenges brought up by big data in biology require an advance of our philosophical ground rules. That is a motivation of this work: the new challenges of cell type representation led to the set of rules presented here. I argue that the opposite way is also true: to advance the theoretical foundations of biology, we need to harness the computational advances. This can be exemplified by the quest of naming specific cell types.

For accurately, pragmatically classifying cell types from the perspective of research synthesis, we need rigorous definitions. These are very specific in nature, and every single empirical article might include a number of unique technotypes. This makes nomenclature a nightmare. How should we differ B Cells that were selected by a slightly different combinations of markers? Well, the task of finding the correct labels is important for human communication, but can be more easily addressed for describing research.

Classes in ontologies can have numeric identifiers, which are, in practice, what matters from a computational perspective. By assigning each technotype a Unique Resource Identifier, an URI, and by inserting the URI in a knowledge graph, such as the cell ontology, we can start dealing with the complexity of cell types definition across biology. Each article and each dataset, then, can explicitly declare once which class they refer by a term (for example, "neutrophil"), so to avoid the ambiguities of natural language, while maintaining readability.

The Cell Ontology currently holds less than 4.000 cell types. The number of rigorous and useful cell types, however, is considerably larger. By the rules of deductive logic, whenever you combine two classes ("neutrophil" + "human", for example) you give rise to a third class ("human neutrophil", for example). If we consider the possible rigorous descriptors, species, anatomical locations, stages of life, biological sexes, strains, stages in the circadian clock and so on, the possible number of cell types is of many orders of magnitude higher. This presents a logistical challenge, which would require a greater number of active collaborators in extension and maintenance of such knowledge base. One solution would be open systems such as Wikidata, where life scientists could add their technotypes with a low entry barrier. The development of such system is a direction for future research to operationalize the descriptions here.

The idea of "technotype", if coupled to the possibility for every researcher to craft their "cell type" of interest, solves a major problem of correctly labeling cell types in single cell experiments. The non-existence of exact matches in CL (even when combined with other ontologies) render it impossible to accurately annotate articles and datasets without incurring in induction issues. As mentioned previously, this would still allow to cross results from different setups, with the gain of explicitly.

A branch of computational single-cell development has dedicated itself to find tools for labeling single cell experiments. Some of the approaches have aimed at anchoring the classes used to definitions by the Cell Ontology, both via manual match based on "expert comparison" or via algorithms such as BLAST2CO. By considering technotypes, however, the task conceptually changes not to find a class that "matches" the current experiment, but a candidate class for a "parent" of the described technotype. Albeit subtle, taking this extra layer of consideration can allow more precise conceptualizations of labeling mechanisms and, perhaps, improving the overall performance of such approaches.

Final remarks

In this article, I proposed a set of 3 rules (rigorous description, taxon scope restriction and theoretical usefulness) and 1 recommendation (link to ontology of cell types) to pragmatically define classes of cells (cell types). I then followed to propose 4 namings to clarify discussions on the topic: archetypes (a class with a scope above species level), cell types (a class with scope equal to one species), infratypes (a class with scope below the species level) and technotypes (the exact cell type defined for an experimental setup).

In the following sections I dissected some logical entailments of such definition, which might conflict with current view on definitions of cell types. I do not aim at solving such conflicts, but merely propose this as one, out of many possible ways, of organizing our knowledge about cell types. Of note, I would

like to highlight the concept of the “technotype” as representing the unit for cell classification, in an analogous way of how the “species” is the unit for evolutionary biology.

This article is intended to clarify some of the meanings, and provide directions to future development of the theoretical basis of cell type definition. The discussion on the definition of cell types is still on its infancy, and we need human power to tackle the huge theoretical challenges. Biologists, philosophers and computer scientists ought to distill the details of defining cell types, powering the Human Cell Atlas and the life sciences research enterprise of this century.

References
