Towards a pragmatic definition of cell type.

This manuscript (permalink) was automatically generated from lubianat/technotype@db233b9 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 life sciences is the histology class. The students are tasked with identifying cell types across various 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 what we know about a couple of hundred cell types.

New techniques have challenged this anatomy based conceptualization. From multiplexed flow cytometry to patch clamping, single-cell techniques 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. The rise of new cell types has become especially evident in the past few years, with the explosion of single-cell omics and the creation of the Human Cell Atlas and the HUBMAP projects.

We advanced our tools, but our concept of “cell type” is still based on century-old histochemical techniques - such as the Golgi-stains of neurons, immortalized in the drawings of Ramon y Cajal. That implies that the concepts we use are drawn for studies of microanatomy. The connection with anatomy leads to thinking about cell types as anatomical entities as if they were clearly dissectable and fixed in an organism.

The advances in biology require us to find better answers for how to define a cell type is. The concept might not have a “true” meaning, in a material-realistic sense. Nevertheless, 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 several proposals for what a cell type are arising. One core line of though is evolutionary: 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 coexist, our quest to define cell types may take various forms depending on the application. The challenge of representing cell types in the context of evolution is conceptually different from the challenge of representing cell types in biomedical experimentation. In that direction, it is notable the groundwork of the Cell Ontology project and the recurrent ideas proposed at the International Workshop on Cells in ExperimentaL Life Sciences. Their contributions base much of the views here and will be discussed in detail, througout the article.

The conceptual quest addressed by this work is one of the research synthesis. How to find a cell type definition that allows a rigorous description of biomedical knowledge experiments?

For that goal, the body of the article is divided in 4 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 adress 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, researchers presented their views on the “Conceptual Definition of”Cell Type" in the Context of a Mature Organism?" The opinions vary, and do not converge to a consensus. Many of the opinions advocates for a core role of the cells’ function in defining cell types, a slippery road, as the meaning of “function” in biology is elusive. What is clear is that it is not simple to finding a definition that pleases the diverse fields.

Aevermann et al (2018) come up with a set of rules to classifiy 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 þ species).” This conceptual advance is of practical importance, as advances our possibilities to solve many challenges of representing cell types in the era of big data . The Cell Ontology also has used markers in define cell types, specially for immune cells.

The use of markers, however, leaves us with a conceptual problem: definitions of cell type used by electrophysiologists, or even in manuals of histology classes, are not based on markers. Rigorously, this would leave aside a whole part of what we consider biomedical knowledge. Moreover, gene markers are not defined for cell types that span multiple species. Multispecies markers would require us to explicitly consider homology, adding an extra layer of confusion.

My pragmatic definition of cell type (for eukaryotic, multicellular organisms) consists of 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. 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. Such defintions are feasible and, in practice, similar criteria are used in immunology for defining cell types.

Rule number 2 is a speficicaton of a rigourousness: one rigorousness criteria is to define the taxons in which a given cell type is expected to have manifestations. We should be able to identify a cell of the type in any individual of the taxon of interest, given the appropriate conditions (e.g. stage of life and biological sex). I will call this set of taxons the of the cell type. Note that, as cell types can be defined by function, and functions can converge, a global definition cannot restricted to monophyletic taxa (“clades”).

Knowing the scope is important to avoid the pitfalls of extrapolation. One recurrent extrapolation is assuming that theories corroborated with mice experiments are valid for other organisms. This extrapolation is an instance of the classic problem of induction, detailed thoroughly in the Logic of Scientific Discovery. A specification of the scope a researcher is referring would make inductional claims explicit and enable proper evaluation of claims.

Currently, the largest authority on cell type definitions, the Cell Ontology, provides “multispecies” definitions of cell types. These are conceptually important, but their use for labeling studies plainly is dangerous. It is not safe to assume that a “mouse neutrophil” is simply a “neutrophil” that happens to

be in a “mouse”. A definition of scope is essential to tease apart general claims from study-specific claims.

By providing a specific scope, we can make explicit what are currently implicit predictions. If we take a theory that “the scope of neutrophils is mammals,” it explicitly predicts that we expect whales to have neutrophils, for example. If we cannot identify neutrophils after severe tests for any single species, we can agree (by convention) that the claim has been falsified. Then, we would have to consider this previous concept as unreal, and tailor our understanding of neutrophils to new scopes.

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 simplest criteria of usefulness is one based on the individual: a valid cell type is whatever class one individual rationally finds useful.

Rule number 4 is one practical extension of the “usefulness” rule: a cell type has to be logically anchored to other cell types for it to be useful.
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 eukaryotes”.

The ontological organization is important for integrating knowledge across studies. A “transcriptomically-defined” cell type and a “electrophysiologically-defined” cell type cannot be rigorously said to be the same, but they can be grouped in a “superclass” that contains cells that match either one or the other criteria. Practically, when describing a cell type, one should make an effort to insert it into the universe of interrelated cell types, even if that implies creating new “superclasses”.

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

Naming classes of cell types

Before analyzing the consequences of the criteria raised on part 1, it is important to make a set of naming conventions for different classes of cell types. The conventions are necessary to avoid confusion. Much of the literature mixes cell types in one species (e.g., when dealing with a cell type as an evolutionary unit) and multispecies (e.g., in the cell ontology). Current advances in the taxonomy of living beings are calling artificial the classifications of “genus,” “families,” “order,” and similar rankings. The level of a “species” is better defined, and useful in practice (with discernible theoretical divergences). In accordance with the PhyloCode’s theoretical views, 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.”
- **Stricto sensu** cell types, for which the scope of the definition is precisely 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”.

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” level: 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 lymph node of 2-month-old chow-fed female C57BL6/J mice from the mouse-house of the Institute of Biochemistry of the University of São Paulo collected in the morning around 10 pm. Albeit really specific, all the mentioned facets (markers, anatomical location, age, biological sex, strain, housing conditions, circadian clock and diet) are known to alter cell types’ behavior. Thus, it is necessary one more name, for clarity:

- Technotype: A specific, experimentally defined cell type, which harbors 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 hypotheses 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 claims 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. For the pragmatic purpose adopted here, I avoid dissecting 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 a definition is an important topic for theoretical research, it is not strictly necessary for the framework presented here.

One example implication of this entailment is that the class “human cells in metaphase of mitosis” can be considered a cell type, as they can be rigorously defined and are restricted to a taxon. Even though “metaphase” itself is still a biological process, we can describe all cells 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 philosophical-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, essential to consider that cells can belong to at least two disjunct classes.

Often, cell types are described taxonomically, as being related in one single hierarchy. Cells can be assigned to disjunct classes. Thus, 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 in the perspective of research synthesis. s

Practical consequences of the definition

In the previous section, I discussed the 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.

By using the set of rules, we can better evaluate claims of discovering nes cell types. With vast amounts of data and loose definition of cell types, it becomes uncannily easy to claim a new cell type. However, suppose one claims to discover a new “stricto sensu” cell type. In that case, one has to provide enough evidence that cells from this class are identifiable across all individuals of a species. A claim of an “archetype” would require provide evidence in more than one species. 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 such a class in both mouse and human samples, corroborating the existence of 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. It describes subclasses of monocytes and dendritic cells in humans, and pragmatically uses markers for their definition. The patients were recruited from “the Boston-based PhenoGenetic project (...) and the Newcastle community.” It is arguable that they did not have a random sample of humanity, and the observed results might not hold for different populations. 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 the evidence, very much we have done for centuries for claims of new animal species.

However, a new problem arises. How to name all these specific cell types? How to we humanely understand with so many “cell types” with such subtle differences?

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 several unique technotypes. This makes nomenclature a nightmare. How should we differ B Cells that were selected by slightly different combinations of markers? I avoid this challenge, focusing on the identification of concepts that are computationally useful.

As described by Sabine Leonelly, the challenges brought up by big data in biology require an advance of our philosophical theories. I argue that the inverse is also true: to advance the theoretical foundations of modern biology, we need to harness the computational tools. The quest for naming cell types becomes simpler when the goal doesn’t require human readability at every step.

Classes in ontologies can have numeric identifiers. By assigning each technotype a Unique Resource Identifier, a URI, and by inserting the URI in a knowledge graph, such as the Cell Ontology or Wikidata, we can start dealing with the complexity of cell types definition across biology. For humans, each

article and each dataset could explicitly declare the class it refers by a name(for example, “neutrophil”), avoiding natural language ambiguities 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 fo deductive logic, whenever you combine two classes (“neutrophil” +“human”) you give rise to a third class (“human neutrophil”). We need to consider several classes: rigorous descriptors, species, anatomical locations, stages of life, biological sexes, strains, stages in the circadian clock and more. The possible number of cell types is of many order of magnitudes higher than currently available. The scale denotes a logistical challenge, as we would require a more significant number of active collaborators in extension and maintenance of such a knowledge base. One way to progress is open systems such as Wikidata, where life scientists can add their ’cell types" with a low entry barrier. The development of such a 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 an ontology of cell types) to define cell types. I proposed 4 namings to clarify discussions on the topic: archetypes (a class with a scope above species level), **stricto sensu** 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).

I dissected some logical entailments of such definition, which, admittedly, might conflict with current views on defining cell types. I do not aim at solving such conflicts but propose a different way 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 the future development of the theoretical basis of cell type definition. The discussion on cell types’ definition 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