Towards a pragmatic definition of cell type

This manuscript (permalink) was automatically generated from lubianat/technotype@5948372 on September 29, 2020.

Authors

- **Tiago Lubiana**
  0000-0003-2473-2313 · lubianat
  Computational Systems Biology Laboratory, University of São Paulo

- **Helder Takashi Imoto Nakaya (pending approval)**
  0000-0001-5297-9108
  Computational Systems Biology Laboratory, University of São Paulo

Abstract

Introduction

One of the basic classes in any undergraduate major in life sciences is the histology class. The students are tasked with identifying cel 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)[1] that perpetuate the paradigms of what we know about a couple hundred cell types.

Our concept of “cell type”, thus, is still based on century-old histochemical techniques - such as the Golgi-stains of neurons immortalized by Ramon y Cajal[2]. The histological influence is noticeable even in names given to cell types: “erithrocytes”, “eosinophil”, “basophils”, “oxyphilic cell of the thyroid”. The concepts we use are drawn from 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. This may be the reason why attempts to quantify cell types use the scale of “hundreds” of human cell types [3] [4] [5].

New techniques have challenged this anatomy based conceptualization. From flow cytometry to patch clamping, to single-cell RNA-seq, we saw a burst of new categories, and novel cell “subtypes” and “families” popped up in the literature. The bursting intensified explosively in the past few years, with the rise of projects to characterize *all* human cell types [6] HUBMAP [7].

The advances in biology require us to find better answers for how to define a cell type. The concept might not have a “true” meaning, in a philosophical-realistic sense. Nevertheless, we can strive to find nominal, pragmatic definitions for the pragmatic challenges of large scale biology. Otherwise, how can we precisely label single-cell data? How can we formalize the discovery of new cell types? How can we integrate the knowledge from the millions of scientific articles published every year?

The need for conceptual advance is perceived by the community [8] [9] [10], and new perspectives are arising. One core line of thought is evolutionary: the cell type as an evolutionary unit defined by a Core Regulatory Complex (CoRC) of transcription factors. That definition enables the drawing of parallels between the evolution of other biological entities (such as genes, proteins and species) to the evolution of cell types. Models of how multicellular life works greatly benefit from concepts such as “sister types” (cell types that diverged from a single ancestor), “cell type homology,” (cell types in different species that share a common evolutionary origin) and “cell type convergence” (cell types that execute similar functions, but which are not directly evolutionarily related). [11] [12]

However, as much as different concepts of species coexist [13], our quest to define cell types may take various forms. 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 second direction, the groundwork of the Cell Ontology [14] [15] [16] project and the International Workshop on Cells in Experimental Life Sciences[17] [18] are notable. Their contributions base much of the views here and will be discussed in detail throughout the article.

The conceptual quest addressed by this work is one of the research synthesis and is summarized in the following question: which cell type definition can be crafted for rigorously describing biomedical 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 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, researchers presented their views on the "Conceptual Definition of 'Cell Type' in the Context of a Mature Organism?" [8]. The opinions vary, and do not converge to a consensus. Many of the scientists see a core role of cells' functions in defining cell types, a slippery road, as the meaning of "function" in biology is elusive [19].

In one recent attempt to define cell types for single cell RNA-Seq, Aevermann et al came up with a set of needs: "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)." [20] They have great merit in defining clear guidelines for marking a cell type. The requirement of markers is reasonable for the field of single-cell RNA-seq, where marker information is abundant. The Cell Ontology has used markers for defining cell types, an approach specially employed for immune cells [15] [21] [22].

The use of markers, however, leaves us with a conceptual problem: definitions of cell type used by electrophysiologists, or even in the 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, a problem already discussed on the Cell Ontology report of 2011 [15].

Our pragmatic definition of cell type (for eukaryotic, multicellular organisms) consists of 3 + 1 simple rules. A cell type is a class of cells that *must* be:

- Explictly defined
- Theoretically useful
- Identifiable for a defined taxon

And that *should* be:

- Logically related to other cell types

For that specification, *must* represents an "an absolute requirement" and *should* represents that "there may exist valid reasons in particular circumstances to ignore a particular item" (as per RFC 2119 [23])_.

For rule 1, by "explicitly defined", we mean that the cell type needs to be followed by a clear definition, that would allow rational judgements of whether a singular cell belongs or not to the type. The definition *should* be *complete*, providing the necessary and sufficient criteria for classification. An example is a cell type defined by "expression of the proteins CD3 and CD4, but lacking CD8." Even though there is still some ambiguity in that definition (see [21] [22] for longer discussions), it already states clear, reasonable criteria. Any combination of markers (or lack thereof) can define a different *cell type*. This extends to *any* definition, and small differences are enough for constituting new cell types. The degree of rigorousness cannot be decided a priori, as we still do not have a rigorous framework for representing biological knowledge, but we *should* strive to make definitions as rigorous as possibles. Other examples of what could be explicit definitions:

- A "big cell", defined a class of cells with a length of more than 50 micrometers on any axis.

- A “human cortical neuron” is any cell in a human cortex that is able of producing an action potential.
- A “leukocyte” is a class of achromatic cells found in animal blood.

The recognition of multiple valid types of rules is not new. The first Cell Ontology article, in 2005, explicitly acknowledged criteria based on function, histology, lineage and ploidy.[14]. These features were combined in the definitions of “species-neutral” cell types[16], arguably useful for integrating databases, or for teaching biology. Gradually, we are acknowledging that we might need more specific classes to characterize experimental biology, leading to definition of species-specific types defined by granular characteristics. [24] [15].

As an analogy, when describing a new species, besides preserving a type specimen, a taxonomist must cover the species *diagnosis* - the ways one can tell a species from others. Even though there are standards for format, the taxonomy codes for botanics and zoology do not restrain which characters should be used, as there is a huge diversity of organisms.[25] In the same way, it might be unrealistic to restrict definitions of cell types to a single class of characters, like expression markers.

Rule number 2 is, thus, an explicit criteria that *must* be followed when talking about cell types scientifically: we need to define the taxons for which a given cell type is expected to manifest. The cell type, then needs to be findable in any individual of the taxon (or taxons) of interest, given the appropriate conditions (e.g. stage of life and biological sex). The set of taxons covered by a cell type is called here a *taxonomy scope* (or just *scope*) of the cell type. Note that, as cell types can be defined by function, and functions can converge, the taxonomix scope is not restricted to monophyletic taxa (“clades”). The definition of taxon used here is liberal, as any class of organisms that any researcher identify explicitly as a unit.

Knowing the scope is important to avoid the pitfalls of extrapolation. One recurrent extrapolation is that theories corroborated by mouse experiments are valid for human cells. 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 to would make inductional claims explicit and enable proper evaluation of claims. 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.

Rule number 3 is a rule of practical concern. There is a massive amount of “explicit definitions” 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 any 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 increased usefulness. Which means that a definition of a cell class is (for research synthesis concerns) of lower 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”. This is presented as a recommendation instead of a requirement as, in practice, it might be an overhead not strictly necessary for claims of discovery of new cell types and similar tasks.

The ontological organization is important for integrating knowledge across studies. A “transcriptomically-defined” cell type and a “electrophysiologically-defined” cell type are not 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 these set of criteria will be discussed further in the following sections.

Naming classes of cell types

In parallel to the criteria raised on part 1,I propose a set of naming conventions for different classes of cell types, to facilitate communication. 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). It is arguably useful to distill these different concepts into their own names.. Given the importance of the concept of species in biological classification [26] [27], I derive a “species-centric” view on the naming of classes of cell types. The three classes I propose are:

- Archetypes, for when the *taxonomic scope* of the type is beyond the level of species. For example, “mammal neutrophils.”
- *Stricto sensu* cell types, for when the *taxonomic scope* of the type corresponds to a single species. For example, “human neutrophils.”
- Infratypes, for when the *taxonomic scope* is below the level of species. For example, considering the mouse strain “C57BL/6J” [28], “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.

In addition, in individual experiments, we work with cells of very specific classes. They are not only infratypes, but very specific infratypes, defined 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 several mornings 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 what we know about cell types. Thus, we benefit from using a name for these very specific classes:

- Technotype: A specific, experimentally defined cell type, which harbors on its definitions the precise conditions of the cells sampled.

Even if really specific, a technotype is still a class. Unless a study used only one single-cell, it likely contained some sampling method. Samples are sampled from a specific population, for which hypothesis are actually tested. This is the most "granular’ cell type in my 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 the claims can be logically combined in “upper” ontological levels for making claims with a higher degree of universality. The propagation of knowledge to upper levels cannot be implicit. (see Yarkoni 2020 for an analogous problem in the psychological sciences [29]). As defended by Popper, knowledge should travel “quasi-inductionally” by fostering hypothesis with higher degrees of generality, which can then be tested for the more universal 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 as a subclass for "cell type". 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 distinction is an important topic for theoretical research, it is not a requirement for representing biomedical experiments.

One example of this entailment is that the class "human cells in metaphase of mitosis" can be considered a cell type, as they can be explictly defined and restricted to a taxon. Even though "metaphase" itself is a biological process, we can describe all cells executing this process as 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 explicit definition used for fibroblast (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.

If 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. This is conflicting with attempts to classify cell-types using single hierarchies in the form of a tree [30] [31] [32]. Cell types need to be represented ontologically (in the computational sense), which can be thought of multiple, intertwining trees, which take 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.

Practical consequences of the definition

In the previous section, we discussed the logical entailments of accepting the proposed rules as valid. Here, we 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 discover of new cell types. With vast amounts of data and loose definition of cell types, it becomes uncannily easy to claim a new cell type. Converserly, if one explictly claims to discover a new "stricto sensu" cell type, 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 evidence for existence 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 2018 [33] [34] about the newly found "ionocyte", a class of cells in the trachea enriched for expression of genes homologous to the *CFTR* gene. Both studies displayed evidence for such a class in both mouse and human samples, corroborating the existence of an archetype. This discovery of an archetype has been denominated by both articles as a discovery of a new cell type.

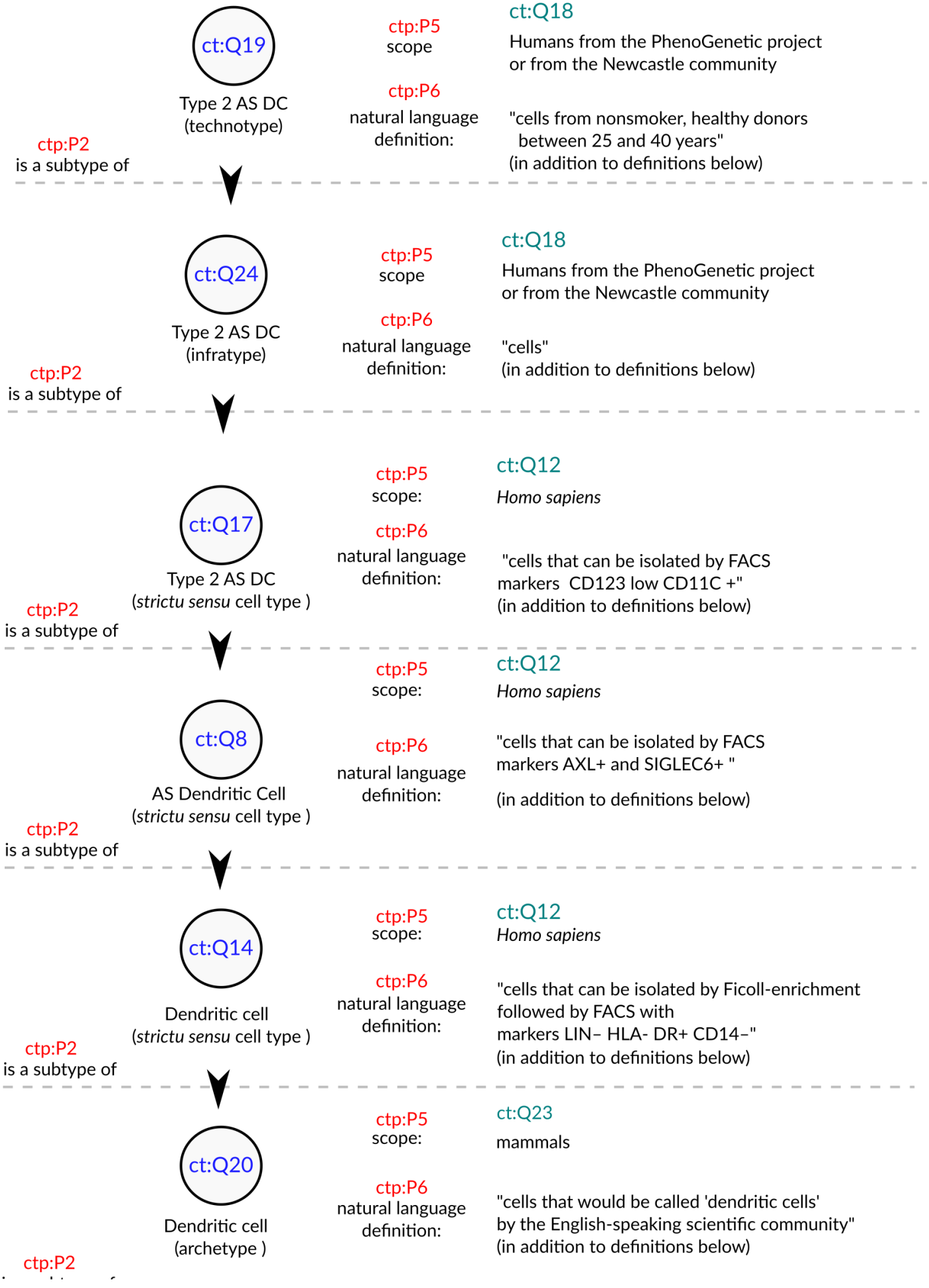
Another example of cell type discovery is a pioneer article by Villani et al [35]. It describes subclasses of monocytes and dendritic cells in humans, and pragmatically uses markers for their definition. The patients were recruited from "the Boston-based PhenoGenetic project (...) and the Newcastle community." 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”.

An example from the article is the discovery of the “AS Dendritic cell” (and two subpopulations of it), characterized by expression of the antigens for the proteins AXL and SIGLEC6. This and other cell types are presented in the article as part of a “Human dendritic cell atlas”, generalizing the theory for the whole humanity. However, it is not clear if the population sampled included individuals from different human background. Thus, it is technically possible that the existence of the “cell type” exactly as described might be restricted to some human groups. The jump from technotype (which takes into consideration also descriptors like “healthy” and “age between 25 and 40 years”) to infratype (“all humans in this population scope”) to cell type *strictu sensu*(all humans) is depicted in the figure 1 and exemplifies the logical flow.

“Dendritic cells” is one of the cell types most thoroughly modelled by Cell Ontology. [36] [37]. The current natural language definition of dendritic cell (CL_0000451) states that a dendritic cell is “A cell of hematopoietic origin, typically resident in particular tissues, specialized in the uptake, processing, and transport of antigens to lymph nodes for the purpose of stimulating an immune response via T cell activation. These cells are lineage negative (CD3-negative, CD19-negative, CD34-negative, and CD56-negative).” The structured definitions are derived from the leukocyte (CL_0000738) definition, which defines such cells as “achromatic cell of the myeloid or lymphoid lineages capable of ameboid movement.” These definitions are not reconcilable to the “dendritic cells” studied by Villani et al’s . We have no way of knowing if the cells in their work are “typically resident in particular tissues”, “achromatic” or “capable of ameboid movement”. That might sound pedantic, but the logical requirements of computational systems leads to both biocurators and computers being seen as pedantic. This high level of precision is necessary to accurately depict not only the complexities of cell types, but the complexities of research settings.

@prefix ct : <https://celltypes.wiki.opencura.com/wiki/Item:> .
@prefix ctp : <https://celltypes.wiki.opencura.com/wiki/Property:> .



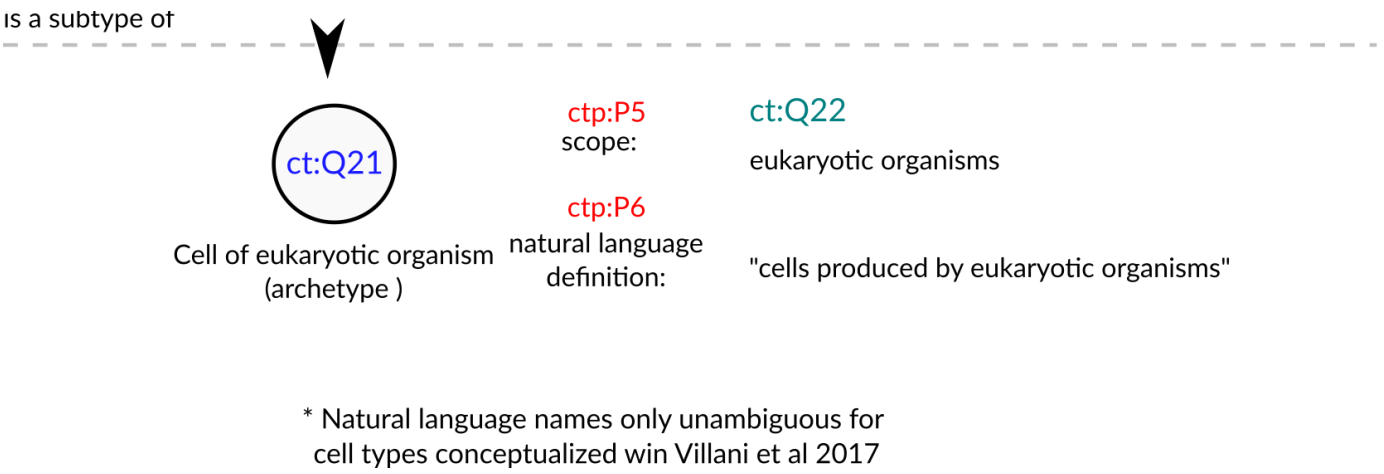


Figure 1: Conceptualization of a set of the cell types in Villani et al, 2017 [35]. The depicted cell types were manually curated from the article, where they are either implicitly or explicitly mentioned. The set of cell types is not comprehensive, and represent only a small fraction of the concepts handled by the authors. Identifiers for cell types are written in pseudocode based on the Turtle serialization for RDF knowledge graphs (https://www.w3.org/TR/turtle/) and represent valid URIs (described in the database https://celltypes.wiki.opencura.com/wiki/Main_Page). URI: Universal Resource Identifier; RDF: Resource Description Framework.

Even if we are not yet able to formally represent all the aspects that go into a cell type definition, we can use an explicit "natural language definition" property to define cell types. As David Osumi-Sutherland puts in a 2017 article about cell type classification, there is a "*mismatch between quantified logic, which records assertions about all members of a class, and the messy, noisy reality of biology and the data we collect about it.*." [24]. We do not need complex assertions to create a logically valid ontology of cell types. Taking as example in figure 1, all cell types treated as "dendritic cells" in the literature are valid subclasses of the dendritic cell archetype (ct:Q20). Such a subclassing system might lack the power to computationally check the validity of definitions. However, by the principle of minimal commitment [**???**], it could already be be a suitable scaffold for representing experimental data (e.g from single-cell transcriptomics) and allow logically robust data integration.

However, a new problem arises. How to name all these specific cell types? How to humanely understand so many "cell types" with such subtle differences? Which names should we use to differ cells that were selected by slightly different combinations of markers?

For accurately classifying cell types from the perspective of research synthesis, we need explicit definitions, and they should be as rigorous as possible. This makes the task of finding common names specially hard. We avoid this challenge, focusing on the identification of concepts that are computationally useful. Common names can be agreed in a context by referencing to identifiers, similarly to common names and scientific names of species.

Sabina Leonelli stated that the challenges brought up by big data in biology require an advance of our philosophical theories [38]. We agree, and argue that the inverse is also true: to advance the theoretical foundations of modern biology, we need to harness the power of computational tools. Computational ontologies provide a solution for dealing with complex concepts. Classes in ontologies can have alpha-numeric identifiers. We can, thus assign each technotype a Unique Resource Identifier, a URI, similar to the Cell Ontology (CL) [14] [15] [16] or Wikidata [39] [40].The quest for naming cell types becomes simpler when the goal doesn't require human readability at every step. Instead, by harnessing the computer power to record the identifiers and their explicit definitions, we can focus on higher level abstractions

The idea of "technotype" can theoretically solve labeling cell types in single-cell experiments. The non-existence of exact matches in CL (even when combined with other ontologies) renders it theoretically impossible to annotate articles and datasets without incurring in induction issues. The "technotype" avoids that imprecision by giving power to every researcher to craft their "cell type" of interest. As

mentioned in the previous section, by having a knowledge connecting the concepts, we would still be able to compare results from different researchers, but now explicitly stating the level of abstraction in which they can be compared.

Specifically, for single-cell transcriptomics, the technotype refines our model for labeling cells (and, consequently, cell clusters). A branch of computational single-cell development has dedicated itself to find tools for labeling single-cell experiments. While some approaches ignore ontologies [30] [41], others aim at finding the best class among the Cell Ontology [42] or MeSH IDs[43] [44]. Manual matches have been fed to algorithms such as BLAST2CO [44] to predict best "matches" for a single-cell cluster. However, unless the cells were sampled in the same way across articles, and drawn at random from the same population of individuals, they represent strictly different classes, even if very similar. Thus, we must change the task from finding a "match" to the cells in a given current experiment to finding a "point of insertion" in an ontological network. By acknowledging these real differences, we can have precise metadata, enabling precise statements and facilitating valid reuse of publically available data.

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). The concept of the "technotype" can be harnessed as the unit for classifying cells, in an analogous way of how the "species" is the conventional unit for classifying organisms into higher-order taxa.

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,

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.

Supplementary notes

A set of notes that may be incorporated into the final text, or become appendices, or end up as blog posts somewhere.

What to do when two researchers disagree on a definition?

Cell type names are notoriously ambiguous and one definition might collide with an other, specially regarding the natural language name used to described. There are many different, equally valid definitions of a "dendritic cell." I do not aim to solve this problem from a societal standpoint. However, from a computational-ontology standpoint, there is one simple solution: split the concept.

This approach is similar to King Solomon's solution in a famous bible story, called the Judgement of Solomon. In a dispute between two women that claimed to be the mothers of a child, the solution of

the king was simple: split the baby. However, babies are notoriously indivisible, and the true mother did not really like the idea.

It may be that some scientists are attached to their names, as mothers are to their babies. However, unlike babies, namings can be divided. Each of the scientists gets to name their specific conceptualization however they choose. Many names might "collide" in that way, and that is okay. Under the hood, however, the names refer to different identifiers. Computationally there would be no ambiguity. Then, it is just a matter of the researcher to respect the choice of their peers of calling something by the *wrong* name, as long as the identifier is correct.

Splitting concepts upon conflicts in the end is more the multiplication of bread and fish in the Feeding the multitude episode, and everyone gets to eat.

But ontologies are different from ordinary babies and magical fish. The splitting of concepts would not only create new concepts, but leave a trace. They would be immediate subclasses of their conjunction. An equally valid superclass that can be defined by "a cell containing characteristics of any of their subclasses".

In a parallel with text-book mitosis, the concept gets divided in two new, equally real concepts. And as we can trace cells in an animal to a single zygote, we can keep track of concepts while they keep dividing, whenever a new conflict pops up.

The big assumption of continuity in time

One assumption that underlies the validity of the models proposed here is that taxons preserve their characteristics throughout time.

In Popper's Logic of Scientific Research, he states that he has a metaphysical faith on the continuity of laws of nature through time.

We have no way of testing this metaphysical faith, and it is absolutely necessary for the scientific endeavour as we understand.

While in physics this assumption seems to be reasonable, evolution makes biology quite more complicated. Statements that we have about the human species, for example, might be valid today, but were not valid 2000 years ago, and vice-versa.

When I talk about sub-species taxons, which might be a local population of a town, for example, this unit is not immutable. The population of Newcastle, as per the example, might change in time, with imigration and emigration, mutation, natural selection, neutral evolution and the many forms of modifications of a gene pool.

It is, thus, and heuristic, to call "population of Newcastle" a class. We could specify a period in time for which we expect the information to be valid. For example, we may say we are sampling from "the Newcastle population in the years 2019-2020." This would be a valid statement, but it would not be falsifiable, as by 1st of January 2021, no independent tests of the theory can be done.

It is technically possible to have a technotype so precise as to have a scope with a time constraint. In fact, that might be the right way of representing information, if we want to compare experiments done in evolving populations.

While evolutionary definitions take this dimension into account, they are fit to theoretical research, but still lack the rigour for explaining real world experiments.

All research that uses human samples are subject to strong influence of time. I started this as a note, but it is so important that I'll have to add to the main text.

Clusters are not cells

In the era of large-scale omics, we are starting to see declarations of cell types that are not based on pre-selected criteria, but derived from unsupervised clustering followed by labelling.

This is a powerful exploratory approach, which, as mentioned in the main text, has led to discoveries of ionocytes and new classes of dendritic cells, for example.

Many algorithms and "expert-based" annotation protocols focus on labeling *clusters* instead of labeling *cells*.

Cells in a cluster are arbitrarily similar (as determined by the clustering algorithm) and so they will, by definition, differ from other cells in the sample.

For single-cell RNA-seq, one usually checks which genes are differentially expressed when comparing the cells in a cluster with cells in other clusters. These genes are called "markers" and used for labeling a cell cluster.

What does it mean to label a cell cluster, though? Does it mean that *all* cells there conform to the cell type? Does it mean that *most* cells there conform to the cell type? Does it mean that cells from other clusters in the dataset *definitely do not* conform to the cell type?

So far, I haven't seem a clear, explicit, coherent definition for a cluster label. Not even once.

Marker-based definitions are assumed for the group as a whole, but in current pipelines, nothing blocks one cell in a cluster to lack the expression of a "name-giver" marker.

The classification scheme proposed here works to classify *cells*, but is not sufficient for labeling unsupervisedly-defined *cell clusters*.

What is possible, though, is to use clustering for data exploration. From then on, strict patterns can be decided (ex: a cell that expresses A and B, but not C) and then apply this pattern to the whole dataset. For clusters with consistent markers, this approach should be roughly equivalent to the previously described.

Using such "regular expressions" might lead to a cell being assigned to multiple clusters. Even if we assume that the sample is free of doublets, that cannot be a problem. Cells may have multiple functions. As argued in the main text, each cell can be labeled by multiple standards.

We may avoid multiple labeling if we really need in practice, though, and make preferential claims (if a cell matches defintiions X and Y, it is assigned only to X, for example).

By having explicit "regular expression" patterns for cell definitions in single-cell datasets, the "cell-type assesment" problem becomes trivial: a cell in a new dataset is of the *exact* same type if (and only if) it matches the *exact* definition.

When that is not the case, current algorithms for reconciling single cell datasets can still be successfully employed. But instead of propagating a label, it would propagate a parent class, looking for cells of a similar, sister class.

What this work is not

This is not an attempt to substitute the Cell Ontology (CL) or contradict it in any way. CL is an amazing resource, built by a community of wonderful researchers. Its relation with CL is coexistential, and topics discussed here might be or might not be of interest of CL, and that is OK.

This is not an attempt to create an ontology itself, or a system that allows reasoning. It is a set of suggestions that can be taken into consideration for building a coherent ontology. The Cell Type Wikibase is an experimental ontology, and far from ready to use.

This is not an attempt to have a one-size-fits-all definition of cell type. It is built as a theoretical solution for one cell-type related task. Similar to species definitions, we need an ecosystem of cell-type and cell state definitions that better suit different areas.

This is not an attempt to claim anything about the "true" nature of cell types, in the biological sense. It is a proposal of pratical guidelines to represent research data.

This is not an attempt to solve *all* problems for cell type data annotation. It is the introduction of alternatives that need to be further developed and discussed.

Additional notes

Immune Epithope DB

The Immune Epitope Database and Analysis Resource (IEDB) announced it in 2006 [doi:10.1371/journal.pcbi.0020125]:

"the goal of the IEDB is to present as much information as possible without subjective interpretation, we can never presume any information, but rather we must try to capture the data exactly as presented in the reference, while maintaining the conclusions of the reference in a uniform manner. For example, if all experiments are performed with a whole cell population, but the authors attribute the response to a particular cell type without any evidence, we must capture the effector cells as the entire population."

That is the kind of challenge that the "technotype" solves in theory, as it gets the objective population sampled in any article.

Phenetic Species Concept

Species are the smallest groups that are consistently and persistently distinct and distinguishable by ordinary means (Cronquist 1978; DuRietz 1930; Sokal 1973; Doyen and Slobodchikoff 1974)

Maybe that is something crosslinkable.

References

1. **The structure of scientific revolutions**
   Thomas S. Kuhn, Ian Hacking
   The University of Chicago Press (2012)
   ISBN: 9780226458113

2. **Neuronal cell types**
   Richard H Masland
   Current Biology (2004-07) https://doi.org/dg84br
   DOI: 10.1016/j.cub.2004.06.035 · PMID: 15242626

3. **Search BioNumbers - The Database of Useful Biological Numbers**
   https://bionumbers.hms.harvard.edu/search.aspx

4. **How Many Types of Cells Are in the Human Body?**
   ibswit
   (2017-05-17) https://askabiologist.asu.edu/questions/human-cell-types

5. **What Is Your Conceptual Definition of "Cell Type" in the Context of a Mature Organism?**
   Cell Systems
   (2017-03) https://doi.org/d38b
   DOI: 10.1016/j.cels.2017.03.006 · PMID: 28334573

6. **The Human Cell Atlas**
   Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, … Human Cell Atlas Meeting Participants
   eLife (2017-12-05) https://doi.org/gcnzcv
   DOI: 10.7554/elife.27041 · PMID: 29206104 · PMCID: PMC5762154

7. **The human body at cellular resolution: the NIH Human Biomolecular Atlas Program.** *Nature*
   (2019-10-09) https://www.ncbi.nlm.nih.gov/pubmed/31597973
   DOI: 10.1038/s41586-019-1629-x · PMID: 31597973 · PMCID: PMC6800388

8. **What Is Your Conceptual Definition of "Cell Type" in the Context of a Mature Organism?**
   Cell systems
   (2017-03-22) https://www.ncbi.nlm.nih.gov/pubmed/28334573
   DOI: 10.1016/j.cels.2017.03.006 · PMID: 28334573

9. **A periodic table of cell types**
   Bo Xia, Itai Yanai
   Development (2019-06-15) https://doi.org/ggctwf
   DOI: 10.1242/dev.169854 · PMID: 31249003 · PMCID: PMC6602355

10. **Exciting times to study the identity and evolution of cell types**
    Maria Sachkova, Pawel Burkhardt
    Development (2019-09-19) https://doi.org/ghdb9v
    DOI: 10.1242/dev.178996 · PMID: 31537583

11. **The evolution of cell types in animals: emerging principles from molecular studies.**
Detlev Arendt
Nature reviews. Genetics (2008-11) https://www.ncbi.nlm.nih.gov/pubmed/18927580
DOI: 10.1038/nrg2416 · PMID: 18927580

12. **The origin and evolution of cell types**
Detlev Arendt, Jacob M. Musser, Clare V. H. Baker, Aviv Bergman, Connie Cepko, Douglas H. Erwin, Mihaela Pavlicev, Gerhard Schlosser, Stefanie Widder, Manfred D. Laubichler, Günter P. Wagner
Nature Reviews Genetics (2016-11-07) https://doi.org/f9b62x
DOI: 10.1038/nrg.2016.127 · PMID: 27818507

13. **Species Concepts and Species Delimitation**
Kevin De Queiroz
Systematic Biology (2007-12) https://doi.org/c34kzf
DOI: 10.1080/10635150701701083 · PMID: 18027281

14.**:{unav)**
Jonathan Bard, Seung Y Rhee, Michael Ashburner
Genome Biology (2005) https://doi.org/dfxc74
DOI: 10.1186/gb-2005-6-2-r21 · PMID: 15693950 · PMCID: PMC551541

15. **Logical Development of the Cell Ontology**
Terrence F Meehan, Anna Maria Masci, Amina Abdulla, Lindsay G Cowell, Judith A Blake, Christopher J Mungall, Alexander D Diehl
BMC Bioinformatics (2011-01-05) https://doi.org/c7kw6x
DOI: 10.1186/1471-2105-12-6 · PMID: 21208450 · PMCID: PMC3024222

16. **The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability**
Alexander D. Diehl, Terrence F. Meehan, Yvonne M. Bradford, Matthew H. Brush, Wasila M. Dahdul, David S. Dougall, Yongqun He, David Osumi-Sutherland, Alan Ruttenberg, Sirarat Sarntivijai, … Christopher J. Mungall
Journal of Biomedical Semantics (2016-07-04) https://doi.org/gg99b9
DOI: 10.1186/s13326-016-0088-7 · PMID: 27377652 · PMCID: PMC4932724

17. **Cells in experimental life sciences - challenges and solution to the rapid evolution of knowledge**
Sirarat Sarntivijai, Alexander D. Diehl, Yongqun He
BMC Bioinformatics (2017-12-21) https://doi.org/gg99b7
DOI: 10.1186/s12859-017-1976-2 · PMID: 29322916 · PMCID: PMC5763506

18. **Cells in ExperimentaL Life Sciences (CELLS-2018): capturing the knowledge of normal and diseased cells with ontologies**
Sirarat Sarntivijai, Yongqun He, Alexander D. Diehl
BMC Bioinformatics (2019-04-25) https://doi.org/gg99b8
DOI: 10.1186/s12859-019-2721-9 · PMID: 31272374 · PMCID: PMC6509796

19. **The meanings of "function" in biology and the problematic case of de novo gene emergence**
Diane Marie Keeling, Patricia Garza, Charisse Michelle Nartey, Anne-Ruxandra Carvunis
eLife (2019-11-01) https://doi.org/ggjnmv
DOI: 10.7554/elife.47014 · PMID: 31674305 · PMCID: PMC6824840

20. **Cell type discovery using single-cell transcriptomics: implications for ontological representation.**

Brian D Aevermann, Mark Novotny, Trygve Bakken, Jeremy A Miller, Alexander D Diehl, David Osumi-Sutherland, Roger S Lasken, Ed S Lein, Richard H Scheuermann
Human molecular genetics (2018-05-01) https://www.ncbi.nlm.nih.gov/pubmed/29590361
DOI: 10.1093/hmg/ddy100 · PMID: 29590361 · PMCID: PMC5946857

21. **Reporting and connecting cell type names and gating definitions through ontologies**
James A. Overton, Randi Vita, Patrick Dunn, Julie G. Burel, Syed Ahmad Chan Bukhari, Kei-Hoi Cheung, Steven H. Kleinstein, Alexander D. Diehl, Bjoern Peters
BMC Bioinformatics (2019-04-25) https://doi.org/ghbk9r
DOI: 10.1186/s12859-019-2725-5 · PMID: 31272390 · PMCID: PMC6509839

22. **flowCL: ontology-based cell population labelling in flow cytometry**
Mélanie Courtot, Justin Meskas, Alexander D. Diehl, Radina Droumeva, Raphael Gottardo, Adrin Jalali, Mohammad Jafar Taghiyar, Holden T. Maecker, J. Philip McCoy, Alan Ruttenberg, … Ryan R. Brinkman
Bioinformatics (2015-04-15) https://doi.org/f7cc46
DOI: 10.1093/bioinformatics/btu807 · PMID: 25481008 · PMCID: PMC4393520

23. https://www.ietf.org/rfc/rfc2119.html

24. **Cell ontology in an age of data-driven cell classification**
David Osumi-Sutherland
BMC Bioinformatics (2017-12-21) https://doi.org/ghcbdk
DOI: 10.1186/s12859-017-1980-6 · PMID: 29322914 · PMCID: PMC5763290

25. **Summary for Policymakers**
Cambridge University Press
(2014-06-09) https://doi.org/bwnm
DOI: 10.1017/cbo9781107415324.004

26. ***PhyloCode***
Wikipedia
(2020-07-10) https://en.wikipedia.org/w/index.php?title=PhyloCode&oldid=967070715

27. **PhyloCode: Division I. Principles** http://phylonames.org/code/divisions/1/

28. **000664 - C57BL/6J** https://www.jax.org/strain/000664

29. **The Generalizability Crisis**
Tal Yarkoni
(2019-11-22) https://doi.org/ggdf7h
DOI: 10.31234/osf.io/jqw35

30. **CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing**
Jurrian K de Kanter, Philip Lijnzaad, Tito Candelli, Thanasis Margaritis, Frank CP Holstege
Nucleic Acids Research (2019-09-19) https://doi.org/gg99dp
DOI: 10.1093/nar/gkz543 · PMID: 31226206 · PMCID: PMC6895264

31. **Mapping the transcriptional diversity of genetically and anatomically defined cell populations in the mouse brain**
Ken Sugino, Erin Clark, Anton Schulmann, Yasuyuki Shima, Lihua Wang, David L Hunt, Bryan M Hooks, Dimitri Tränkner, Jayaram Chandrashekar, Serge Picard, … Sacha B Nelson

eLife (2019-04-12) https://doi.org/ghbc3p
DOI: 10.7554/elife.38619 · PMID: 30977723 · PMCID: PMC6499542

32. **How Single-Cell Genomics Is Changing Evolutionary and Developmental Biology**
John C. Marioni, Detlev Arendt
Annual Review of Cell and Developmental Biology (2017-10-06) https://doi.org/ggb632
DOI: 10.1146/annurev-cellbio-100616-060818 · PMID: 28813177

33. **A revised airway epithelial hierarchy includes CFTR-expressing ionocytes**
Daniel T. Montoro, Adam L. Haber, Moshe Biton, Vladimir Vinarsky, Brian Lin, Susan E. Birket, Feng Yuan, Sijia Chen, Hui Min Leung, Jorge Villoria, … Jayaraj Rajagopal
Nature (2018-08-01) https://doi.org/gdwskh
DOI: 10.1038/s41586-018-0393-7 · PMID: 30069044 · PMCID: PMC6295155

34. **A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte**
Lindsey W. Plasschaert, Rapolas Žilionis, Rayman Choo-Wing, Virginia Savova, Judith Knehr, Guglielmo Roma, Allon M. Klein, Aron B. Jaffe
Nature (2018-08-01) https://doi.org/gdwsj7
DOI: 10.1038/s41586-018-0394-6 · PMID: 30069046 · PMCID: PMC6108322

35. **Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors**
Alexandra-Chloé Villani, Rahul Satija, Gary Reynolds, Siranush Sarkizova, Karthik Shekhar, James Fletcher, Morgane Griesbeck, Andrew Butler, Shiwei Zheng, Suzan Lazo, … Nir Hacohen
Science (2017-04-20) https://doi.org/f94x5t
DOI: 10.1126/science.aah4573 · PMID: 28428369 · PMCID: PMC5775029

36. **An improved ontological representation of dendritic cells as a paradigm for all cell types**
Anna Masci, Cecilia N Arighi, Alexander D Diehl, Anne E Lieberman, Chris Mungall, Richard H Scheuermann, Barry Smith, Lindsay G Cowell
BMC Bioinformatics (2009) https://doi.org/cpxdhs
DOI: 10.1186/1471-2105-10-70 · PMID: 19243617 · PMCID: PMC2662812

37. **Hematopoietic cell types: Prototype for a revised cell ontology**
Alexander D. Diehl, Alison Deckhut Augustine, Judith A. Blake, Lindsay G. Cowell, Elizabeth S. Gold, Timothy A. Gondré-Lewis, Anna Maria Masci, Terrence F. Meehan, Penelope A. Morel, Anastasia Nijnik, … Christopher J. Mungall
Journal of Biomedical Informatics (2011-02) https://doi.org/c6dmmh
DOI: 10.1016/j.jbi.2010.01.006 · PMID: 20123131 · PMCID: PMC2892030

38. **The challenges of big data biology**
Sabina Leonelli
eLife (2019-04-05) https://doi.org/gfzw8q
DOI: 10.7554/elife.47381 · PMID: 30950793 · PMCID: PMC6450665

39. **Wikidata as a knowledge graph for the life sciences**
Andra Waagmeester, Gregory Stupp, Sebastian Burgstaller-Muehlbacher, Benjamin M Good, Malachi Griffith, Obi L Griffith, Kristina Hanspers, Henning Hermjakob, Toby S Hudson, Kevin Hybiske, … Andrew I Su
eLife (2020-03-17) https://doi.org/ggqqc6
DOI: 10.7554/elife.52614 · PMID: 32180547 · PMCID: PMC7077981

40. **Wikidata** https://www.wikidata.org/wiki/Wikidata:Main_Page

41. **Probabilistic gene expression signatures identify cell-types from single cell RNA-seq data**
Isabella N. Grabski, Rafael A. Irizarry
bioRxiv (2020-01-23) https://doi.org/gg99dq
DOI: 10.1101/2020.01.05.895441

42. **ontoProc: Ontology interfaces for Bioconductor, with focus on cell type identification**
https://www.bioconductor.org/packages/release/bioc/vignettes/ontoProc/inst/doc/ontoProc.html#conceptual-overview-of-ontology-with-cell-types

43. **CellMeSH: Probabilistic Cell-Type Identification Using Indexed Literature**
Shunfu Mao, Yue Zhang, Georg Seelig, Sreeram Kannan
Cold Spring Harbor Laboratory (2020-05-31) https://doi.org/gg99dr
DOI: 10.1101/2020.05.29.124743

44. **Searching large-scale scRNA-seq databases via unbiased cell embedding with Cell BLAST**
Zhi-Jie Cao, Lin Wei, Shen Lu, De-Chang Yang, Ge Gao
Nature Communications (2020-07-10) https://doi.org/gg4mm3
DOI: 10.1038/s41467-020-17281-7 · PMID: 32651388 · PMCID: PMC7351785