Towards a pragmatic definition of cell type

This manuscript ([permalink](https://lubianat.github.io/technotype/v/6dff479b78fea95315619737742bcdf0ea01ed1b/)) was automatically generated from [lubianat/technotype@6dff479](https://github.com/lubianat/technotype/tree/6dff479b78fea95315619737742bcdf0ea01ed1b) on September 30, 2020.

## Authors

* **Tiago Lubiana** [0000-0003-2473-2313](https://orcid.org/0000-0003-2473-2313) · [lubianat](https://github.com/lubianat) Computational Systems Biology Laboratory, University of São Paulo
* **Helder I Nakaya** [0000-0001-5297-9108](https://orcid.org/0000-0001-5297-9108) Computational Systems Biology Laboratory, University of São Paulo

## Abstract

The concept of cell type is key for modeling biology. Recent technological advances are prompting us to rethink what we understand by cell type and how we classify them. There is currently no consensus in the definition of a cell type, which makes it hard to integrate knowledge across life sciences. We propose here that a cell type should represent any class of cell that is explicitly defined; is identifiable within a taxon; and is theoretically useful. We also specified four classes of cell types: *sensu stricto* cell types (to differentiate the concepts applied to a single species), archetypes (from the ones for multiple species), infratypes (populations below the species level), and technotypes (specific experiments). The flexible and rigorous framework we propose can base annotation of single-cell omics data sets, and reconcile knowledge about cells across all different domains of science

# Introduction

One of the basic subjects in any undergraduate major in life sciences is histology. The students are required to identify cell types across various tissues and look for color and shape patterns in hematoxylin-eosin stains. Textbooks, including the one by Junqueira & Carneiro[**???**], work as manuals that perpetuate the paradigms (in the Kuhnian sense)[[1](#bookmark=id.z337ya)] of what we know about a couple of hundred cell types.

Our concept of “cell type,” thus, is still based on centuries-old histochemical techniques, such as the Golgi-stains of neurons immortalized by Ramon y Cajal[[2](#bookmark=id.3j2qqm3)]. The histological influence is noticeable even in the names given to cell types, such as “erythrocytes”, “eosinophils”, “basophils”, and “oxyphilic cell of the thyroid”. The concepts we use are drawn from studies of microanatomy. This connection with anatomy leads us to think about cell types as anatomical entities as if they are clearly dissectible and fixed in an organism. The limits of resolution perpetuated by the histological-anatomical view may be why attempts to quantify cell types use the scale of “hundreds” of human cell types [[3](#bookmark=id.1y810tw)-[5](#bookmark=id.2xcytpi)].

New techniques have challenged this anatomy-based conceptualization. From flow cytometry and patch clamping to single-cell RNA-seq, we have observed an explosion of new categories and novel cell “subtypes” and “families” popping up in the literature. The activity has intensified in the past few years with the rise of projects to characterize all human cell types, such as the Human Cell Atlas (HCA) and the Human BioMolecular Atlas Program (HUBMAP) [[6](#bookmark=id.1ci93xb)] [[7](#bookmark=id.3whwml4)].

The advances in biology require us to find better answers for how to define a cell type. Such concept might not even have a “true” meaning, in a philosophical-realistic sense. Nevertheless, we can strive to find nominal, pragmatic definitions for the real 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 millions of published scientific articles?

The need for a conceptual advance is being perceived by the community [[8](#bookmark=id.2bn6wsx)] [[9](#bookmark=id.qsh70q)] [[10](#bookmark=id.3as4poj)], and new perspectives are rising. One core line of thought is based on the cell type as an evolutionary unit defined by a Core Regulatory Complex (CoRC) of transcription factors. That definition enables the drawing of parallels, from 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](#bookmark=id.1pxezwc)] [[12](#bookmark=id.49x2ik5)].

However, as much as different concepts of species coexist [[13](#bookmark=id.2p2csry)], 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](#bookmark=id.147n2zr)] [[15](#bookmark=id.3o7alnk)] [[16](#bookmark=id.23ckvvd)] project and of the International Workshop on Cells in Experimental Life Sciences[[17](#bookmark=id.ihv636)] [[18](#bookmark=id.32hioqz)] are notable. Their contributions underpin many of the opinions discussed here and are explained in detail throughout the article.

The conceptual quest addressed by this work is one of research synthesis and is summarized in the following question: Which cell type definition can be crafted for rigorously describing biomedical experiments?

Towards that goal, the body of the article is divided into 4 parts. In Part 1, we propose a set of rules that are sufficient for defining cell types. Part 2 offers a small set of names for differentiating the main classes of cell types. In Part 3, we address the logical consequences of the proposed definitions, while Part 4 is a discussion of the pragmatic challenges envisaged in employing such definitions.

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

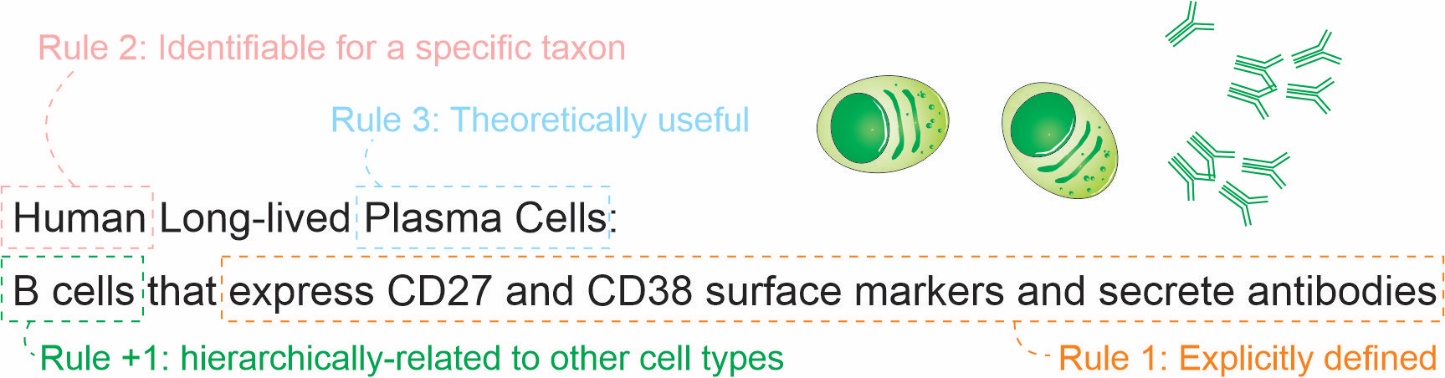
In an opinion article published in Cell Systems in 2017, the researchers presented their views on the conceptual definition of ‘cell type’ in the context of a mature organism [[8](#bookmark=id.2bn6wsx)]. The opinions were varied, and no consensus was achieved. Many of the scientists believed that cell functions have a core role in defining cell types, which is a slippery road, as the very meaning of “function” in biology is elusive [[19](#bookmark=id.1hmsyys)].

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

1. explicitly defined
2. identifiable for a defined taxon
3. theoretically useful

And it should be:

1. logically related to other cell types



**Figure 1: The set of 3 + 1 rules for defining a cell type.**

Here, “must” represents an absolute requirement, whereas “should” suggests that “there may exist valid reasons in particular circumstances to ignore a particular item” (as per RFC 2119 [[20](#bookmark=id.41mghml)]).

For rule 1, we mean that the cell type needs to be followed by a clear definition that would allow rational judgments of whether a singular cell belongs to the type or not. Such definition should provide 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 (see [[21](#bookmark=id.2grqrue)][[22](#bookmark=id.vx1227)] for longer discussions), it already states clear and reasonable criteria. 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 possible. Other examples of what could be explicit definitions are as follows:

* “Big cell” is a class of cells which have a length of more than 50 micrometers on any axis.
* “Human cortical neuron” is a class of cells in a human cortex that are capable of producing an action potential.
* A “leukocyte” is a class of cells found in animal blood which are achromatic cells.

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](#bookmark=id.147n2zr)]. These features were combined in the definitions of “species-neutral” cell types [[16](#bookmark=id.23ckvvd)], 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 the definition of species-specific types defined by granular characteristics [[23](#bookmark=id.3fwokq0)] [[15](#bookmark=id.3o7alnk)].

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 botany and zoology do not limit which characters can be used, as there is a huge diversity of organisms.[[24](#bookmark=id.1v1yuxt)] In the same way, it might be unrealistic to restrict definitions of cell types to a single class of characters like expression markers.

Rule 2 is an explicit criterion that must be followed while discussing cell types scientifically; we need to define the taxa for which a given cell type is expected to manifest. The cell type then needs to be discoverable in any individual of the taxon (or taxa) of interest, given the appropriate conditions (e.g., stage of life and biological sex). The set of taxa covered by a cell type is called here a taxonomic scope (or just scope) of the cell type. Note that, as cell types can be defined by function and functions can converge, the taxonomic scope is not restricted to monophyletic taxa (clades). The definition of taxon used here is liberal and applies to any class of organisms that any researcher identifies explicitly as a unit.

Knowing the scope is important to avoid the pitfalls of extrapolation. A recurrent theme is that theories corroborated by mouse experiments are valid for human cells. Such extrapolation is an instance of the classic problem of induction, which is discussed thoroughly in the “Logic of Scientific Discovery”. The taxonomic scope allows us, researchers, to be clear regarding our claims, and better discern what we claim to be true for a strain, a species or any other class of organisms.

Rule 3 deals with a practical concern. Rigorously, there is an infinite number of explicit definitions that any scientist might come up with. One simple proof of this infinitude is that size-based cell definitions (as for “big cell” above) may alone consider any of the infinite real numbers. Thus, a cell type “bigger than 7.835 micrometers” will fail this rule. If we want to characterize all human cell types, for example, it is necessary to have a finite number of cell types. Rule 3 could be paraphrased as: a valid cell type is a class of cells that any researcher rationally finds useful for a theoretical perspective of reality. For example, a recent study used single-cell RNA-seq experiments to assign 275,000 Drosophila cells into 200 cell types (<https://www.nature.com/articles/s41586-020-2879-3>). Since these 200 cell types were useful for Özel and colleagues when describing the world, it automatically satisfied rule 3.

Rule 4 is a practical extension of the usefulness rule: a cell type has to be hierarchically-related to other cell types for increased usefulness. This means that a definition of a cell class is (for research synthesis concerns) less useful if it cannot be considered a “subclass” of another cell type. For practical concerns, all imaginable cell types are subclasses of a “eukaryotic cell”. This is presented as a recommendation instead of a requirement as, in practice, it might be an overhead and not strictly necessary for claims of the discovery of new cell types and similar tasks.

Ontological organization is important for integrating knowledge across studies. A cell type that is based on its transcriptome is not the same as one based on its electrophysiology. However, they can be grouped in a superclass of cells that match either one or the other criterion. For example, the green-OFF bipolar cells of the retina and the Syt2-/NK3R+ cells of the retina are considered to be the same cell type (<https://www.cell.com/neuron/fulltext/S0896-6273(12)00883-5>). However, as these features are often measured separately, we have, in fact, two individual classes for which knowledge is produced. These classes, then, can be combined in the superclass “green-OFF OR Syt2- / NK3R cells” for integration of claims across domains. 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.

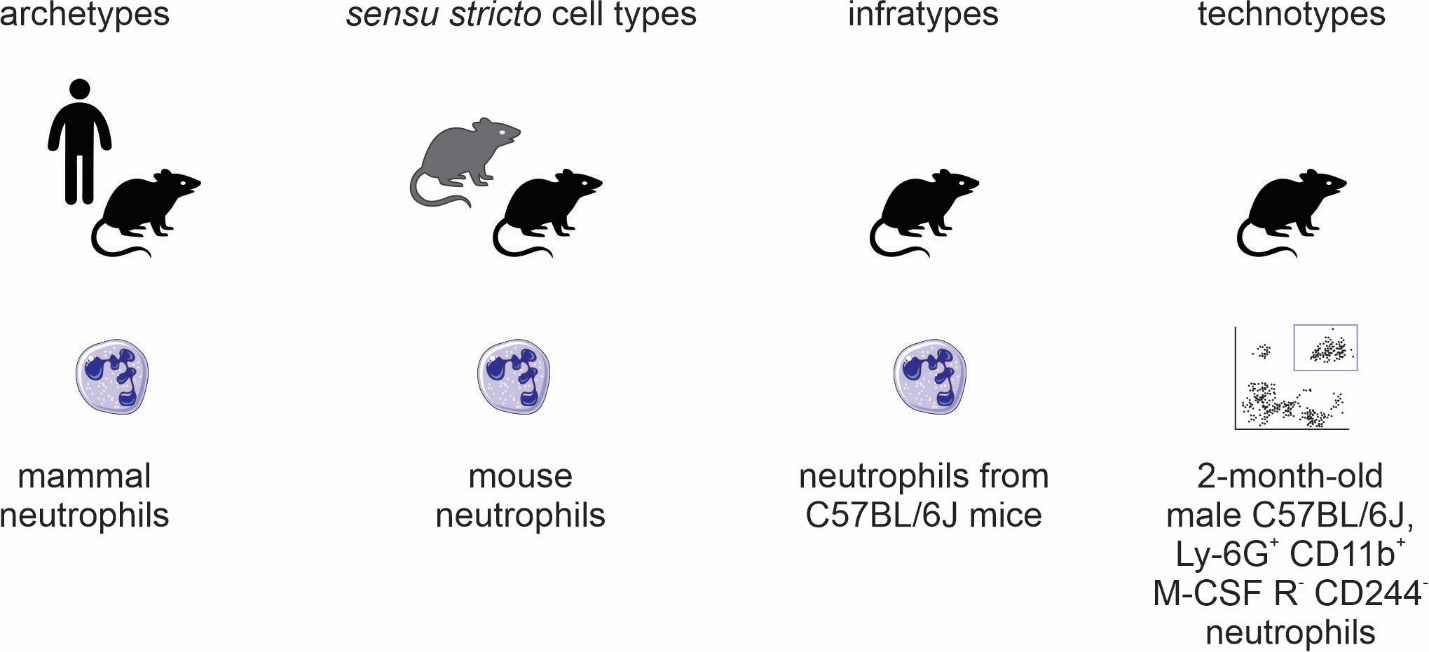
The consequences of these sets of criteria are discussed in the following sections.

# 

# Naming classes of cell types

To facilitate communication among life scientists, we propose a set of naming conventions for different classes of cell types. Much of the literature mix cell types in one species (e.g., when dealing with a cell type as an evolutionary unit) or in 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 [[25](#bookmark=id.4f1mdlm)] [[26](#bookmark=id.2u6wntf)], we derive a species-centric view on the naming of classes of cell types. The four classes (Figure 2) we propose are as follows:

* archetypes, for when the taxonomic scope of the type is beyond the level of species; for example, “mammal neutrophils.”
* *sensu stricto* cell types, for when the *taxonomic scope* of the type corresponds to a single species; for example, “mouse neutrophils.”
* infratypes, for when the taxonomic *scope* is below the level of species; for example, considering the mouse strain “C57BL/6J” [[27](#bookmark=id.19c6y18)], “neutrophils from C57BL/6J mice”.
* technotype: A specific, experimentally defined cell type that harbors in its definition the precise conditions of the cells sampled; “2-month-old male C57BL/6J, Ly-6G+ CD11b+ M-CSF R- CD244- neutrophils”.



**Figure 2. Names for classes of cell types.**

By adopting a precise vocabulary, we can avoid misunderstandings and communicate more 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 to, once again, avoid failing implicitly at 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 on several mornings around 10 pm. Although quite 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: a technotype.

Even if it is specific, a technotype is still a class. Unless a study used only one single-cell, it likely contained some sampling method. Samples are from a specific population for which hypotheses are actually tested. This is the most granular cell type, in our considered view, for research synthesis. This is the type that can be strictly annotated in single-cell RNA-seq analysis, for example.

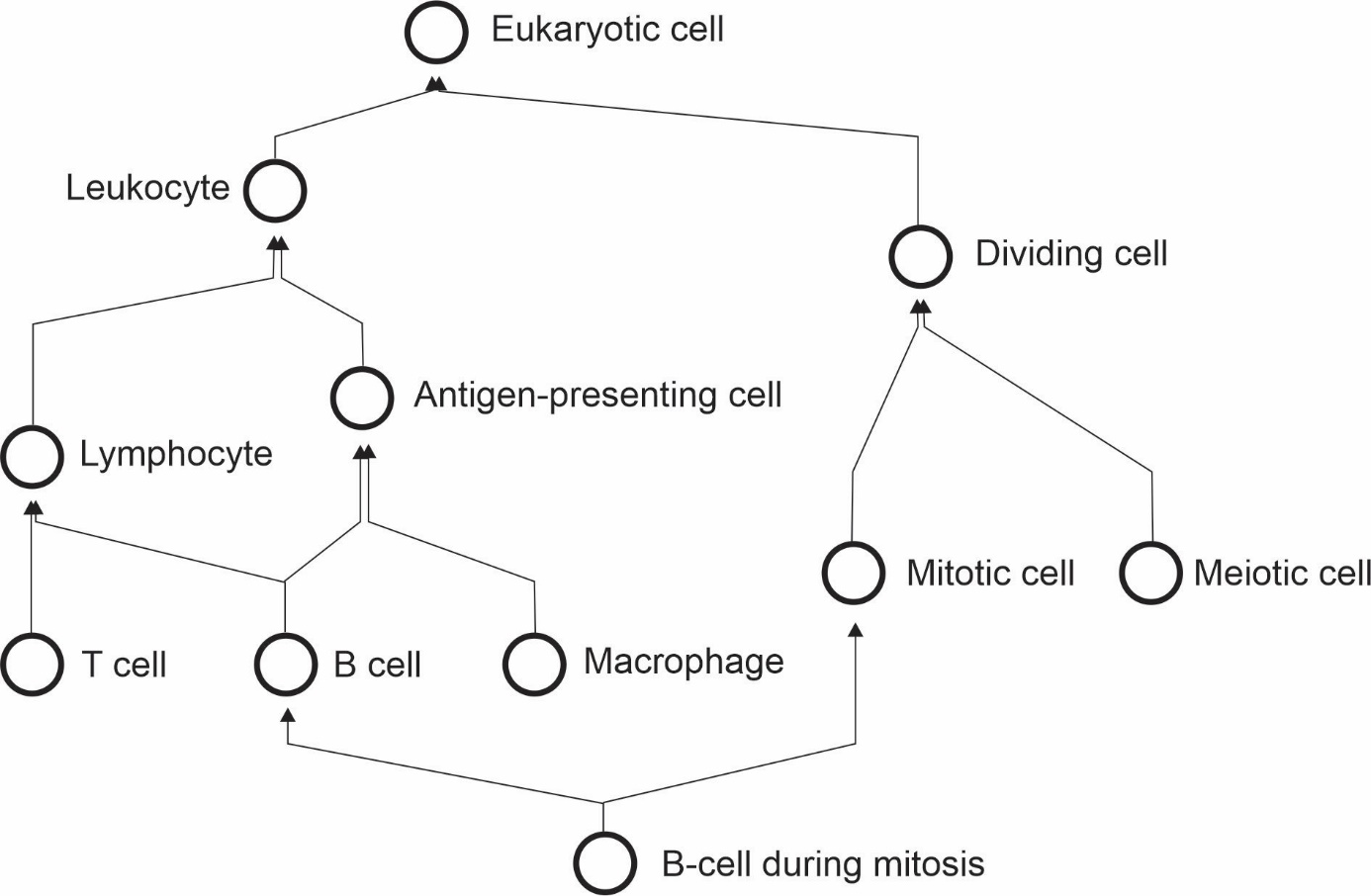
Single claims are made and tested for technotypes, and the claims can be logically combined in “upper” ontological levels for reaching a higher degree of universality. The propagation of knowledge to upper levels cannot be implicit. (see Yarkoni 2020 for an analogous problem in the psychological sciences [[28](#bookmark=id.3tbugp1)]). As Popper defends, knowledge should travel “quasi-inductionally” by fostering hypotheses with higher degrees of generality, which can then be tested for the more universal class [**???**].

# Logical consequences of the definition of a cell type

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, we avoid the dissection of the differences between persistent classes of cells (which we refer to as traditional cell types) or the transient, fugacious classes of cells (which we refer to as traditional cell states). 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 explicitly 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, we 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 classes that are not hierarchically related: “fibroblasts” and “doubling cells”. If cells can be assigned to multiple classes that are not hierarchically related, 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 in conflict with attempts to classify cell-types using single hierarchies in the form of a tree [[29](#bookmark=id.28h4qwu)] [[30](#bookmark=id.nmf14n)] [[31](#bookmark=id.37m2jsg)]. Cell types need to be represented ontologically (in the computational sense), which can be thought of as multiple, intertwining trees that take into account different ways of classifying cells (Figure 3).



**Figure 3 - Not a tree Legend here!!.**

Another logical consequence of the definition is that the concept of subtype becomes redundant with the concept of cell type. The notion of subtype, then, only makes sense when discussing classes with different degrees of universality. Thus, claims to discovery of new cell “subtypes” or “types” differ only stylistically and can be considered indistinguishable from the perspective of research synthesis.

# Practical consequences of the definition of a cell type

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

In a 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 (Cell Ontology)”, and “A specimen source description (anatomic structure þ species)” [[32](#bookmark=id.1mrcu09)]. Their approach has 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 employed in particular for immune cells [[15](#bookmark=id.3o7alnk)] [[21](#bookmark=id.2grqrue)] [[22](#bookmark=id.vx1227)].

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 adopted, this requirement would leave aside an entire segment of what we consider biomedical knowledge. Moreover, gene markers are not defined for cell types that span multiple species, a problem already discussed in the Cell Ontology report of 2011 [[15](#bookmark=id.3o7alnk)].

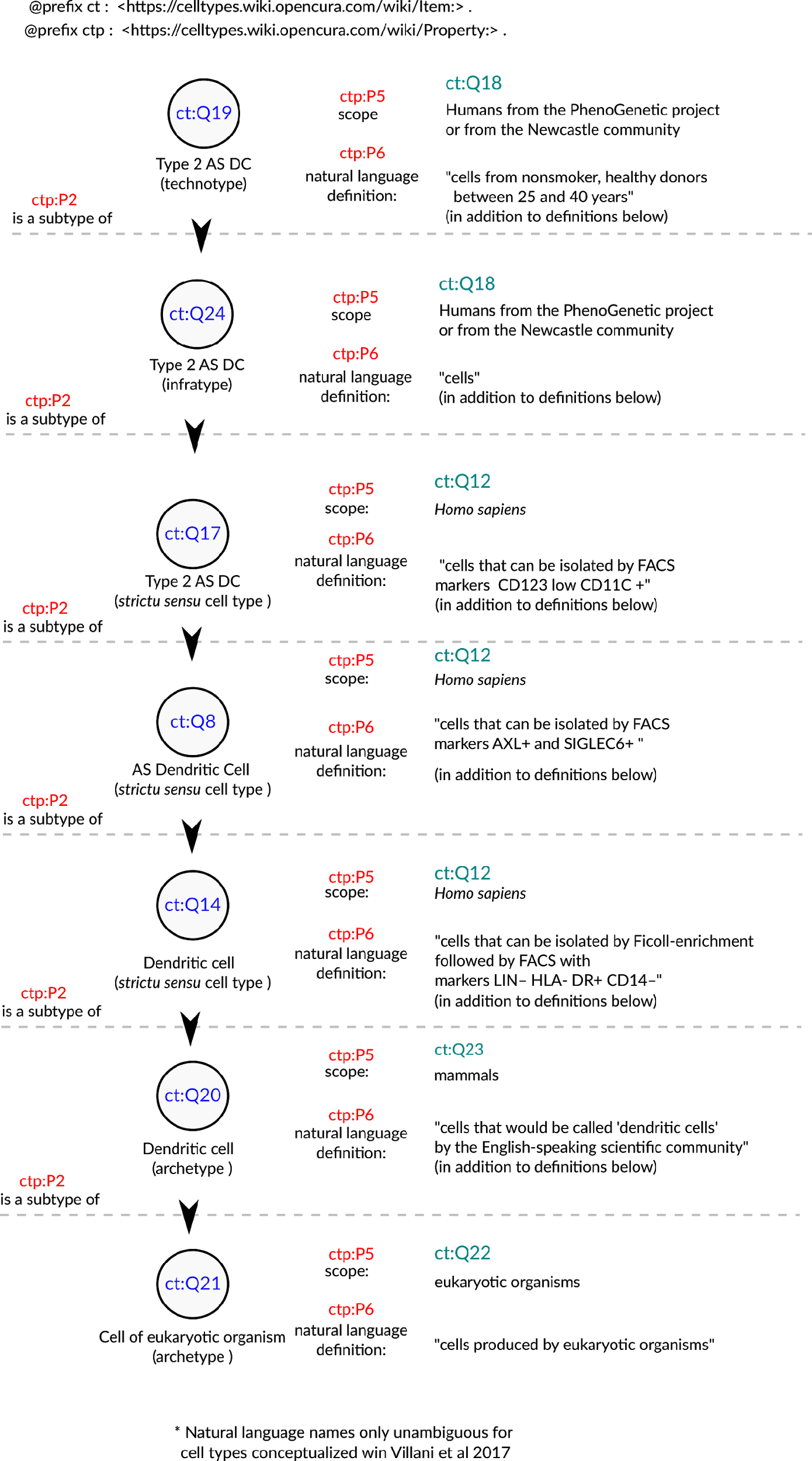
By using our less stringent set of rules, we can better evaluate claims of discovery of new cell types. With vast amounts of data and loose definition of cell types, it becomes uncannily easy to claim a new cell type. Conversely, if one explicitly claims to have discovered a new *sensu stricto* cell 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 of existence in more than one species. Consequently, experiments that only use a specific strain of mice have a stronger claim if the expectation is limited to the infratype.

An example of the discovery of a new archetype is the pair of articles published in Nature in 2018 [[33](#bookmark=id.46r0co2)] [[34](#bookmark=id.2lwamvv)] about the newly found “ionocyte”, a class of cells in the trachea enriched for the 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 pioneering article by Villani et al [[35](#bookmark=id.111kx3o)]. 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.” Arguably, 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 the discovery of a new cell type.

An example from the article is the discovery of the “AXL+ SIGLEC6+ AS Dendritic cell”. This and other cell types are presented in the article as part of a “Human dendritic cell atlas”, generalizing the theory for the whole of humanity. 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 *sensu stricto* (all humans) is depicted in Figure 4 and exemplifies the logical flow.

“Dendritic cells” are one of the cell types most thoroughly modeled by the Cell Ontology [[36](#bookmark=id.3l18frh)] [[37](#bookmark=id.206ipza)]. The current definition of the dendritic cell ([CL\_0000451](http://purl.obolibrary.org/obo/CL_0000451)) is downstream of the leukocyte ([CL\_0000738](http://purl.obolibrary.org/obo/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 with the “dendritic cells” studied by Villani et al. We have no way of knowing if the cells in their work are achromatic or capable of ameboid movement. That might sound pedantic and might, unfortunately, be so, but the logical requirements of computational systems lead to both [biocurators](https://www.biocuration.org/community/biocuration-generic-job-description/) and [computers](https://eloquentjavascript.net/00_intro.html) being seen as pedantic. This high level of precision is necessary to accurately depict not only the complexities of cell types but also of research settings.



***Figure 4: Conceptualization of a set of the cell types in Villani et al., 2017 [***[**35**](#bookmark=id.111kx3o)***].*** *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 represents 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, in practice, an explicit “natural language definition” property. As David Osumi-Sutherland puts in his 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*” [[23](#bookmark=id.3fwokq0)]. We do not need to have all the biology classified before we deal with cell types. Taking the example in Figure 4, 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 a suitable scaffold for representing experimental data (e.g., from single-cell transcriptomics) and may allow logically robust data integration.

However, a new problem arises. How to name all these specific cell types? How to humanely comprehend so many cell types with such subtle differences? Which names should we use to differentiate 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 especially hard. We avoid this challenge, focusing on the identification of computationally useful concepts. Common names can be agreed upon in a context by referencing identifiers as with regard to the common and scientific names of species.

Sabina Leonelli stated that the challenges thrown up by big data in biology require advancement of our philosophical theories [[38](#bookmark=id.4k668n3)]. We agree and argue that the converse 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](#bookmark=id.147n2zr)] [[15](#bookmark=id.3o7alnk)] [[16](#bookmark=id.23ckvvd)] or Wikidata [[39](#bookmark=id.2zbgiuw)] [[40](#bookmark=id.1egqt2p)].The quest for naming cell types becomes simpler when the goal does not require human readability at every step. Instead, by harnessing computer power to record the identifiers and their explicit definitions, we can focus on higher-level abstractions

The idea of “technotype” can theoretically solve issues with 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 with perfection. 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 knowledge that connect the concepts, we would still be able to compare results from different researchers but not explicitly state the level of abstraction at 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 finding tools for labeling single-cell experiments. While some approaches ignore ontologies [[29](#bookmark=id.28h4qwu)] [[41](#bookmark=id.3ygebqi)], others aim at finding the best class among the Cell Ontology [[42](#bookmark=id.2dlolyb)] or MeSH IDs[[43](#bookmark=id.sqyw64)] [[44](#bookmark=id.1rvwp1q)]. Manual matches have been fed to algorithms such as BLAST2CO [[44](#bookmark=id.1rvwp1q)] 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 publicly available data.

# Final remarks

In this article, we have proposed a set of three rules (rigorous description, taxon scope restriction, and theoretical usefulness) and one recommendation (link to an ontology of cell types) to define cell types. We have also proposed four types of naming to clarify discussions on the topic: archetypes (a class with a scope above species level), *sensu* *stricto* 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 a manner analogous to how the “species” is the conventional unit for classifying organisms into higher-order taxa.

We have dissected some logical entailments of such definition, which admittedly might conflict with current views on defining cell types. We do not aim to solve such conflicts or negate the other perspectives but only to propose a unique way of organizing our knowledge on cell types.

This article clarifies some of the meanings and provides directions for the future development of the theoretical basis of cell type definition. The discussion on cell types’ definition is still in its infancy, and we need human power to tackle these 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

1. **The structure of scientific revolutions**   
Thomas S. Kuhn, Ian Hacking  
*The University of Chicago Press* (2012)   
ISBN: [9780226458113](https://worldcat.org/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](https://doi.org/10.1016/j.cub.2004.06.035) · PMID: [15242626](https://www.ncbi.nlm.nih.gov/pubmed/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](https://doi.org/10.1016/j.cels.2017.03.006) · PMID: [28334573](https://www.ncbi.nlm.nih.gov/pubmed/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](https://doi.org/10.7554/elife.27041) · PMID: [29206104](https://www.ncbi.nlm.nih.gov/pubmed/29206104) · PMCID: [PMC5762154](https://www.ncbi.nlm.nih.gov/pmc/articles/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](https://doi.org/10.1038/s41586-019-1629-x) · PMID: [31597973](https://www.ncbi.nlm.nih.gov/pubmed/31597973) · PMCID: [PMC6800388](https://www.ncbi.nlm.nih.gov/pmc/articles/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](https://doi.org/10.1016/j.cels.2017.03.006) · PMID: [28334573](https://www.ncbi.nlm.nih.gov/pubmed/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](https://doi.org/10.1242/dev.169854) · PMID: [31249003](https://www.ncbi.nlm.nih.gov/pubmed/31249003) · PMCID: [PMC6602355](https://www.ncbi.nlm.nih.gov/pmc/articles/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](https://doi.org/10.1242/dev.178996) · PMID: [31537583](https://www.ncbi.nlm.nih.gov/pubmed/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](https://doi.org/10.1038/nrg2416) · PMID: [18927580](https://www.ncbi.nlm.nih.gov/pubmed/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](https://doi.org/10.1038/nrg.2016.127) · PMID: [27818507](https://www.ncbi.nlm.nih.gov/pubmed/27818507)

13. **Species Concepts and Species Delimitation**   
Kevin De Queiroz  
*Systematic Biology* (2007-12) <https://doi.org/c34kzf>   
DOI: [10.1080/10635150701701083](https://doi.org/10.1080/10635150701701083) · PMID: [18027281](https://www.ncbi.nlm.nih.gov/pubmed/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](https://doi.org/10.1186/gb-2005-6-2-r21) · PMID: [15693950](https://www.ncbi.nlm.nih.gov/pubmed/15693950) · PMCID: [PMC551541](https://www.ncbi.nlm.nih.gov/pmc/articles/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](https://doi.org/10.1186/1471-2105-12-6) · PMID: [21208450](https://www.ncbi.nlm.nih.gov/pubmed/21208450) · PMCID: [PMC3024222](https://www.ncbi.nlm.nih.gov/pmc/articles/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](https://doi.org/10.1186/s13326-016-0088-7) · PMID: [27377652](https://www.ncbi.nlm.nih.gov/pubmed/27377652) · PMCID: [PMC4932724](https://www.ncbi.nlm.nih.gov/pmc/articles/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](https://doi.org/10.1186/s12859-017-1976-2) · PMID: [29322916](https://www.ncbi.nlm.nih.gov/pubmed/29322916) · PMCID: [PMC5763506](https://www.ncbi.nlm.nih.gov/pmc/articles/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](https://doi.org/10.1186/s12859-019-2721-9) · PMID: [31272374](https://www.ncbi.nlm.nih.gov/pubmed/31272374) · PMCID: [PMC6509796](https://www.ncbi.nlm.nih.gov/pmc/articles/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](https://doi.org/10.7554/elife.47014) · PMID: [31674305](https://www.ncbi.nlm.nih.gov/pubmed/31674305) · PMCID: [PMC6824840](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6824840)

20. <https://www.ietf.org/rfc/rfc2119.html>

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](https://doi.org/10.1186/s12859-019-2725-5) · PMID: [31272390](https://www.ncbi.nlm.nih.gov/pubmed/31272390) · PMCID: [PMC6509839](https://www.ncbi.nlm.nih.gov/pmc/articles/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](https://doi.org/10.1093/bioinformatics/btu807) · PMID: [25481008](https://www.ncbi.nlm.nih.gov/pubmed/25481008) · PMCID: [PMC4393520](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4393520)

23. **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](https://doi.org/10.1186/s12859-017-1980-6) · PMID: [29322914](https://www.ncbi.nlm.nih.gov/pubmed/29322914) · PMCID: [PMC5763290](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5763290)

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

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

26. **PhyloCode: Division I. Principles** <http://phylonames.org/code/divisions/1/>

27. **000664 - C57BL/6J** <https://www.jax.org/strain/000664>

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

29. **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](https://doi.org/10.1093/nar/gkz543) · PMID: [31226206](https://www.ncbi.nlm.nih.gov/pubmed/31226206) · PMCID: [PMC6895264](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6895264)

30. **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](https://doi.org/10.7554/elife.38619) · PMID: [30977723](https://www.ncbi.nlm.nih.gov/pubmed/30977723) · PMCID: [PMC6499542](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6499542)

31. **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](https://doi.org/10.1146/annurev-cellbio-100616-060818) · PMID: [28813177](https://www.ncbi.nlm.nih.gov/pubmed/28813177)

32. **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](https://doi.org/10.1093/hmg/ddy100) · PMID: [29590361](https://www.ncbi.nlm.nih.gov/pubmed/29590361) · PMCID: [PMC5946857](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5946857)

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](https://doi.org/10.1038/s41586-018-0393-7) · PMID: [30069044](https://www.ncbi.nlm.nih.gov/pubmed/30069044) · PMCID: [PMC6295155](https://www.ncbi.nlm.nih.gov/pmc/articles/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](https://doi.org/10.1038/s41586-018-0394-6) · PMID: [30069046](https://www.ncbi.nlm.nih.gov/pubmed/30069046) · PMCID: [PMC6108322](https://www.ncbi.nlm.nih.gov/pmc/articles/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](https://doi.org/10.1126/science.aah4573) · PMID: [28428369](https://www.ncbi.nlm.nih.gov/pubmed/28428369) · PMCID: [PMC5775029](https://www.ncbi.nlm.nih.gov/pmc/articles/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](https://doi.org/10.1186/1471-2105-10-70) · PMID: [19243617](https://www.ncbi.nlm.nih.gov/pubmed/19243617) · PMCID: [PMC2662812](https://www.ncbi.nlm.nih.gov/pmc/articles/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](https://doi.org/10.1016/j.jbi.2010.01.006) · PMID: [20123131](https://www.ncbi.nlm.nih.gov/pubmed/20123131) · PMCID: [PMC2892030](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2892030)

38. **The challenges of big data biology**   
Sabina Leonelli  
*eLife* (2019-04-05) <https://doi.org/gfzw8q>   
DOI: [10.7554/elife.47381](https://doi.org/10.7554/elife.47381) · PMID: [30950793](https://www.ncbi.nlm.nih.gov/pubmed/30950793) · PMCID: [PMC6450665](https://www.ncbi.nlm.nih.gov/pmc/articles/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](https://doi.org/10.7554/elife.52614) · PMID: [32180547](https://www.ncbi.nlm.nih.gov/pubmed/32180547) · PMCID: [PMC7077981](https://www.ncbi.nlm.nih.gov/pmc/articles/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](https://doi.org/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](https://doi.org/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](https://doi.org/10.1038/s41467-020-17281-7) · PMID: [32651388](https://www.ncbi.nlm.nih.gov/pubmed/32651388) · PMCID: [PMC7351785](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7351785)