Current OBO ontologies are not sufficient to annotate COVID-19-related cell types

Tiago LUBIANA [a,1]
[a] *University of São Paulo, Brazil*

Abstract. The OBO Foundry ontologies are outstanding resources for classifying and curating concepts in the life sciences. However, the use of annotated texts, figures, and data is far from widespread. In this work, I curated 360 cell type mentions in five different single-cell RNA-seq, COVID-19 related scientific articles. These mentions were gathered from the main figures, alongside mentions of tissue, taxon, life stage, and gender of the samples. I manually matched the terms used to NCBITaxon, UBERON, MMUSDV, HSAPDV, PATO, UBERON, and CL ontologies where appropriate. Only 130/360 cell type mentions (36%) could be matched (based on synonyms) to UBERON and CL. Strikingly, none of the 360 mentions cell types could be completely described by OBO ontologies. These results provide evidence that there is an urgent need to improve the reporting of cell-type-related results and enhance ontology generation systems to facilitate completeness in the light of the rise of novel definitions.

Keywords. Cell Ontology, cell types, COVID-19, single-cell RNA-seq

The Cell Ontology (CL) is a long ongoing effort and an outstanding resource for the scientific community. It focuses on general cell types (e.g. "neuron", without species restriction) and it is interoperable with the other OBO Foundry ontologies for taxons, tissues, stages of life, and more. It is expected, then, that a combination of these ontologies should be enough to annotate cell types described in scientific articles. The COVID-19 pandemic has appeared in the moment of the rising use of single-cell RNA sequencing. Cell types are at the core of single-cell transcriptomics. Articles that sequence individual cells in the contexto of COVID-19 are rapidly being published, and annotation of data and results is important to ensure interoperability and data reuse. This work explores the challenges of annotating these articles using current OBO ontologies.

I selected five articles about COVID-19 that use single-cell transcriptomics as the main experimental approach to investigate this. Two of them [1][2] studied airway cells, the other two investigated peripheral blood immune cells [3][4] and a fifth investigated cells in 13 different human tissues[5]. Cell types were identified on main text figures, alongside four core descriptors (tissue of origin, taxon, age, and biological sex of samples). Data was curated in a spreadsheet and were manually searched in Ontobee in the NCBITaxon, UBERON, MMUSDV, HSAPDV, PATO, UBERON, and CL ontologies where appropriate.

I identified 360 mentions to cell types spread across 19 figures in the 5 papers. From these 360 mentions, I was able to find matches on CL for only 238 (66%). In other words,

[1]Corresponding Author: Tiago Lubiana, University of São Paulo, Brazil. E-mail:tiago.lubiana.alves@usp.br

roughly a third of cell types in the dataset are not represented in CL, if we reconcile both sources by using synonyms.

The situation is worsened because the definitions of cell types on the Cell Ontology do not allow precise classification based solely on transcriptomic characteristics. Rigorously, none of the cell types claimed in these studies are properly captured by CL. For example, the definition of a dendritic cell (CL_0000451) starts by "A cell of hematopoietic origin, typically resident in particular tissues (...)". All five articles use "dendritic cell" (or one of its subclasses) to label single-cell clusters, and none of those clusters can be rigorously matched to CL, as the articles do not trace their origin.

We know that cell types behave differently depending on tissue, biological sex, and life stage. I could label both the tissue and cell line for 130/360 cell type mentions (36%). Only 35/360 (9%) had statements that could be matched to life stage ontologies (MMUSDV, HSAPDV). None of the 360 cell types mentions had annotations regarding the biological sex of the subjects. This lack of consideration of biological sex, and little consideration of age is striking. On the bright side, all 360 mentions included the taxon of interest, allowing precise matching to NCBITaxon ontology.

In light of the Human Cell Atlas [6] and the considerable effort to characterize cell types, it becomes urgent to advance the use of ontologies to annotate these texts. Due to a lack of accuracy in reporting, the gaps in the Cell Ontology (and other ontologies), and the challenges of defining cell types, it is not possible to accurately annotate single-cell transcriptomic studies. The intent of this work is to bring these issues into attention, as annotation would be vital for maximizing what we can get from the data.

Systems such as Wikidata might complement the OBO Foundry effort by allowing public edits and avoiding formal definitions. By using *descriptions* instead of definitions, we can work with cell types that are rigorously different, but in practice treated as the same. The putative lack of rigor of such systems might be outweighed by the benefits of making annotations feasible. Also, by providing user-friendly graphic interfaces, Wikidata might make it easier for more biologists to get initiated in ontologies, possibly improving, in the long run, the precision of the terms used in scientific reporting.

References

[1] Muus C, Luecken MD, Eraslan G, Waghray A, Heimberg G, Sikkema L, Kobayashi Y, Vaishnav ED, Subramanian A, Smilie C, Jagadeesh K. Integrated analyses of single-cell atlases reveal age, gender, and smoking status associations with cell type-specific expression of mediators of SARS-CoV-2 viral entry and highlights inflammatory programs in putative target cells. BioRxiv. 2020 Jan 1.

[2] Liao M, Liu Y, Yuan J, Wen Y, Xu G, Zhao J, Cheng L, Li J, Wang X, Wang F, Liu L. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. Nature Medicine. 2020 May 12:1-3.

[3] Wen W, Su W, Tang H, Le W, Zhang X, Zheng Y, Liu X, Xie L, Li J, Ye J, Dong L. Immune cell profiling of COVID-19 patients in the recovery stage by single-cell sequencing. Cell Discovery. 2020 May 4;6(1):1-8.

[4] Wilk AJ, Rustagi A, Zhao NQ, Roque J, Martinez-Colon GJ, McKechnie JL, Ivison GT, Ranganath T, Vergara R, Hollis T, Simpson LJ. A single-cell atlas of the peripheral immune response to severe COVID-19. medRxiv. 2020 Jan 1.

[5] Qi F, Qian S, Zhang S, Zhang Z. Single cell RNA sequencing of 13 human tissues identify cell types and receptors of human coronaviruses. Biochemical and biophysical research communications. 2020 Mar 19.