Analysing the extent of cell type information present in Wikidata

This manuscript (permalink) was automatically generated from jvfe/manuscript_panglaodb@e9e294b on August 31, 2020.

Authors

- **João Vitor Ferreira Cavalcante**
  0000-0001-7513-7376 · jvfe
  Bioinformatics Multidisciplinary Environment, Federal University of Rio Grande do Norte

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

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Intellegi quidem, ut propter aliam quampiam rem, verbi gratia propter voluptatem, nos amemus; Nec tamen ille erat sapiens quis enim hoc aut quando aut ubi aut unde? Duo Reges: constructio interrete. Non pugnem cum homine, cur tantum habeat in natura boni; Gloriosa ostentatio in constituendo summo bono. Me igitur ipsum ames oportet, non mea, si veri amici futuri sumus. Cur ipse Pythagoras et Aegyptum lustravit et Persarum magos adiit? Quae animi affectio suum cuique tribuens atque hanc, quam dico. Quis est autem dignus nomine hominis, qui unum diem totum velit esse in genere isto voluptatis? Qui si ea, quae dicit, ita sentiret, ut verba significant, quid inter eum et vel Pyrrhonem vel Aristonem interesset?

Keywords: wikidata, knowledge graph, cell type, ontology.

Introduction

Wikidata

Wikidata is an open, freely editable, knowledge graph database within the semantic web that stores knowledge across a multitude of domains, such as arts, history, chemistry and biology, using an item-property-value linked data model. It is easy to use and edit, by both humans and machines, with a rich web user interface and wrapper packages available in common programming languages such as R and Python. All the data within Wikidata is linked and inherently public domain, thus, it presents a great opportunity to make scientific data more FAIR (Findable, accessible, interoperable and reusable), as well as provides the necessary tools to curate and develop ontologies.

Put figure here illustrating the wikidata data model

Several advances towards biological data integration and biological data analysis in Wikidata have been made before, yielding positive results [1] [2] and showcasing it's potential for bioinformatics-related analyses, such as drug repurposing and ID conversion [2]. Wikidata has been proposed as a unified base to gather and distribute biomedical knowledge, with more than 50 000 human gene items indexed and hundreds of biomedical-related properties [3]. However, as of August 2020, cell type information is still very scarse, with only 264 items being categorized as instances of cell types (Q189118), of those, only nine have a “Cell Ontology ID” (P7963)[4] associated, and most have a varying amount of statements (Table 1).

Table 1: As of August 2020, Wikidata items regarding cell types have a varying amount of information, with most having very few statements.

Cell type Item	Number of statements
red blood cell (Q37187)	48
myocyte (Q428914)	18
mesenchymal cell (Q66568500)	2

PanglaoDB

PanglaoDB [5] is a public database that contains data and metadata on hundreds of single-cell RNA sequencing experiments, providing extensive information on cell types, genes and tissues, as well as manually and community curated cell type markers (Table 2). It also provides a rich web user interface for easy data acquisition, including database dumps for bulk downloads.

Table 2: Database statistics for each species in PanglaoDB, as of 31st August 2020.

	Mus musculus	**Homo sapiens**
Samples	1063	305
Tissues	184	74
Cells	4,459,768	1,126,580
Cell Clusters	8,651	1,748

Objectives

Outline geral, apenas para me lembrar:

- Quantos tipos celulares são descritos na base? Quantos tem, no dia da análise, matches exatos?
- Os que tem matches exatos verbalmente, os matches tão minimamente bem anotados (Cell Ontology ID + instance_of: cell type)
- Os genes, espécies e orgãos tem cobertura no Wikidata?
- Os aliases de genes tem cobertura no Wikidata?

Como categorizar cobertura? Match=True na reconciliação? Cutoff de score da reconciliação?

Como verificar os aliases de genes? - Pegar o altlabel pela interface de serviço

Methodology

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

References

1. **Wikidata: A platform for data integration and dissemination for the life sciences and beyond**
Elvira Mitraka, Andra Waagmeester, Sebastian Burgstaller-Muehlbacher, Lynn M Schriml, Andrew I Su, Benjamin M Good
bioRxiv (2015-11-16) https://doi.org/gg9dk4
DOI: 10.1101/031971

2. **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

3. **Wikidata: A large-scale collaborative ontological medical database**
Houcemeddine Turki, Thomas Shafee, Mohamed Ali Hadj Taieb, Mohamed Ben Aouicha, Denny Vrandečić, Diptanshu Das, Helmi Hamdi
Journal of Biomedical Informatics (2019-11) https://doi.org/gg9dnt
DOI: 10.1016/j.jbi.2019.103292 · PMID: 31557529

4. **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://www.ncbi.nlm.nih.gov/pubmed/27377652
DOI: 10.1186/s13326-016-0088-7 · PMID: 27377652 · PMCID: PMC4932724

5. **PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data**
Oscar Franzén, Li-Ming Gan, Johan LM Björkegren
Database (2019) https://doi.org/ggkzxr
DOI: 10.1093/database/baz046 · PMID: 30951143 · PMCID: PMC6450036