Cellosaurus newsletter 9 of October 2021

1) Summary of what changed from release 34 to release 39

Since our last newsletter in March 2020, we have deployed five Cellosaurus releases; the latest one is release 39 of September 12, 2021 which contains information on 130,952 cell lines from 741 species. During this 18 months period we added about 7,800 new entries, but also updated the information contained in more than 80% of the existing entries. Such a high rate of update is due to many concurrent ongoing “retrofitting” tasks in which we add to existing entries fields that did not exist in earlier versions of the resource. This is the case of the age at sampling, the sampling site (organ, tissue, cell type), HLA typing, karyotypic information and, as described in the next section, the sequence variations.

The correlate of such in-depth biocuration activities is the rapid expansion of the number of references cited in Cellosaurus entries. We have added almost 3,000 references in these 18 months to reach a grand total of 22,861 distinct references (21,255 journal publications, 1,448 patents, 221 book chapters and 98 theses).

In term of STR profile information we have now almost 8,000 entries that contain such information, an increase of about 1,000 entries since release 34.

The Cellosaurus is now cross-referenced to 95 distinct resources as we have added cross-references to three more cell line collections: Kerafast, Horizon Discovery and PerkinElmer, for which we had web links that we managed to convert into cross-references. We also added cross-references to three additional resources: Liver Cancer Model Repository (LIMORE), Progenetix and the SUM Breast Cancer Cell Line Knowledge Base (SLKBase)

2) The annotation and restructuration of sequence variations

We doubled the number of cell line entries that contain sequence variation information. We currently have 22,400 entries that contain at least one somatic or genomic mutation description. In total we have about 43,000 sequence variation lines representing over 7,000 distinct variations. Thanks to the help of Rahel Paloots from the group of Michael Baudis in Zurich, we mapped about 3,000 of these variations to ClinVar (the remaining variations are either not represented in ClinVar or are found in non-human cell lines).

We completely restructured the sequence variation lines which now contain cross-references to HGNC, MGI, RGD, UniProtKB and to VGNC to indicate the variation target gene and to ClinVar or dbSNP when the variation is described in one of these resources. Here below are two examples of the evolution of the variation information in the text version:

```
Sequence variation: Heterozygous for SLC2A1 p.Pro485Leu (c.1454C>T)
(PubMed=30197081).
```

is now:

```
Sequence variation: Mutation; HGNC; 11005; SLC2A1; Simple; p.Pro485Leu (c.1454C>T);
ClinVar=VCV000871442; Zygosity=Heterozygous (PubMed=30197081).
```

```
Sequence variation: WWC1-ADRBK2 in-frame gene fusion (PubMed=22032724).
```

is now:

```
Sequence variation: Gene fusion; HGNC; 290; GRK3 + HGNC; 29435; WWC1; Name(s)=WWC1-
GRK3, WWC1-ADRBK2; Note=In frame (PubMed=22032724).
```

The XML format was also updated to accommodate these major changes.

3) Recognition of the Cellosaurus importance

In the last months, three important developments have taken place which are relevant to the importance of the Cellosaurus as an integral part of the ecosystem of high quality knowledge resources in the Life Sciences.

- In December 2020, the Cellosaurus became an ELIXIR Core Data Resource (CDR) (https://elixir-europe.org/platforms/data/core-data-resources). The ELIXIR CDRs are "a set of European data resources of fundamental importance to the wider life-science community and the long-term preservation of biological data".
- In July 2021, the International Rare Diseases Research Consortium (IRDiRC) included the Cellosaurus into its list of IRDiRC Recognized Resources (https://irdirc.org/resources/), thus recognizing the usefulness to rare disease research of curation work carried out in the last three on entries describing cell lines relevant to rare diseases (the addition of cross-references to the Orphanet ORDO ontology described in newsletter 8 and the sequence variation annotation activities described in section 2 above).
- Also in July 2021, the SIB - Swiss Institute of Bioinformatics, has included the Cellosaurus in its portfolio of SIB funded resource. Thanks to the funding that has been allocated we will be able to initiate important software developments such as the creation of an API, a semantic triples RDF version of the Cellosaurus, a SPARQL endpoints as well as user entry forms to submit information on new cell lines.

4) The Cellosaurus on ExPASy

The traffic toward the Cellosaurus on ExPASy is continuing to increase. Since it was made available in May 2015, it has been visited 3 million times by almost 1.6 million distinct users that have browsed 9 million pages.

5) Support for SARS-CoV-2 research

To provide support to researchers working on SARS-CoV-2 we have performed a number of specific curation tasks. These are: a) the annotation of information on the susceptibility or lack of susceptibility of cell lines regarding infection by SARS-CoV-2; b) the annotation of cell lines specifically engineered (by transfection, transduction, KO or gene edition) to be useful for the study of the virus (in this context we created a new group "SARS-CoV-2 research cell line" that allows to easily retrieve such cell lines); c) the annotation of mouse hybridomas against the spike protein; and d) the annotation of cell lines used in the production of COVID-19 vaccines. We added on the home page a link to a web page (https://web.expasy.org/cellosaurus/sars-cov-2.html) which serves as a portal to the cell lines relevant to SARS-CoV-2 research.

6) New steps in the integration of Cellosaurus information in Wikidata

In 2018, Lelia Debornes, as part of her M.S thesis wrote the code for a "bot" that adds Cellosaurus information in Wikidata. Due to changes in both the content and structure of the Cellosaurus and of Wikidata, this bot was no longer functional. Thankfully, since May 2020, Tiago Lubiana Alves has taken upon himself the task of updating the bot code and running it at each release of the Cellosaurus. Two new Wikidata properties have been created that are used in cell line concepts: "derived from organism type" which is more appropriate than the property "found in taxon" that was originally used by the bot and "hPSCreg cell line ID", proposed by Daniel Mietchen, which is used to add the cross-references to hPSCreg in the relevant Wikidata cell line entries.

PS: Do not forget to subscribe to our Twitter page (https://twitter.com/Cellosaurus) for tweets about new developments regarding the Cellosaurus and the universe of cell lines.