The Cellosaurus, a Cell-Line Knowledge Resource

BAIROCH, Amos Marc

Abstract

The Cellosaurus is a knowledge resource on cell lines. It aims to describe all cell lines used in biomedical research. Its scope encompasses both vertebrates and invertebrates. Currently, information for >100,000 cell lines is provided. For each cell line, it provides a wealth of information, cross-references, and literature citations. The Cellosaurus is available on the ExPASy server (https://web.expasy.org/cellosaurus/) and can be downloaded in a variety of formats. Among its many uses, the Cellosaurus is a key resource to help researchers identify potentially contaminated/misidentified cell lines, thus contributing to improving the quality of research in the life sciences.

Reference

BAIROCH, Amos Marc. The Cellosaurus, a Cell-Line Knowledge Resource. *Journal of Biomolecular Techniques*, 2018, vol. 29, no. 2, p. 25-38

PMID : 29805321
DOI : 10.7171/jbt.18-2902-002

The Cellosaurus, a Cell-Line Knowledge Resource

*Amos Bairoch**

Computer and Laboratory Investigation of Proteins of Human Origin Group, Faculty of Medicine, Swiss Institute of Bioinformatics, University of Geneva, Geneva, Switzerland

The Cellosaurus is a knowledge resource on cell lines. It aims to describe all cell lines used in biomedical research. Its scope encompasses both vertebrates and invertebrates. Currently, information for >100,000 cell lines is provided. For each cell line, it provides a wealth of information, cross-references, and literature citations. The Cellosaurus is available on the ExPASy server (https://web.expasy.org/cellosaurus/) and can be downloaded in a variety of formats. Among its many uses, the Cellosaurus is a key resource to help researchers identify potentially contaminated/misidentified cell lines, thus contributing to improving the quality of research in the life sciences.

Key Words: bioinformatics, computational biology, database

CELL LINES

Cell lines are ubiquitous tools for experimental biomedical research, both in academic and industrial settings. The mouse *L cells* were the first immortalized cell line to be established in 1943,[1] but for almost everyone, the dawn of the era of cell lines is associated with the establishment in 1951 of the *HeLa* cell line from the cervical tumor of Henrietta Lacks by George Otto Gey.[2] Since then, the world of cell lines has grown exponentially, not only in terms of the number of cell lines but also in their variability. Major landmarks were the invention in 1975 by Milstein and Köhler of hybridomas, hybrid cell lines that produce mAb[3]; the derivation of embryonic stem cell (ESC) lines from mice in 1981 and from humans in 1998[4]; and finally, the development of induced pluripotent stem cells (iPSCs) by Takahashi and Yamanaka in 2006.[5]

All of these developments have helped to extend the usefulness of cell lines as reagents in laboratories, and we have recently estimated (unpublished results) that there is a total of ~2 million publications that make use of cell lines. However, as is the case for antibodies,[6] cell lines have been pointed out as one of the culprits in what has been called the reproducibility or replication crisis: the difficulty or even worse, the impossibility of replicating an experiment.

Two different issues are responsible for the contribution of cell lines to the reproducibility crisis. The first one is the problem of cell-line misidentification/contamination. The emergence of this issue is linked with the publication of 2 papers in 1968[7] and 1974.[8] They reported that some of the cell lines that were at that time distributed by the American Type Culture Collection (ATCC; Manassas, VA, USA) and were thought to originate from various types of healthy or cancerous tissues were, in fact, *HeLa* cells. Since then, >100 papers have described all kinds of cell-line misidentifications and contaminations, with the *HeLa* cell line as the largest but far from the sole culprit. It has been estimated that one-third of all cell lines used in the life sciences is misidentified.[9] Yet, the means to fight this problem exist. Seventeen years ago, use of short tandem repeat (STR) profiling, which was first developed for forensic applications to authenticate human cell lines, was proposed[10]. Cell-line collections accordingly are documenting the STR profiles of their cell lines, and many service providers offer to profile the cell lines used by researchers in their studies at a reasonable cost. Unfortunately, as scientists in academia generally operate with a limited budget, they will rarely buy certified stocks of cell lines. Instead, they obtain a cell line from the nearest colleague who is already using it, thus greatly increasing the risk that what they are using is not what they expect to be using. Furthermore, whereas the methodology to authenticate cell lines exists, scientists rarely do so, also for reasons of cost but also because they are often not aware of the severity of this problem.

The second issue that affects cell lines in the context of experimental reproducibility is one that also affects many other aspects of the life sciences: the "naming" issue. There have been very few attempts to publish guidelines on how to name cell lines. One is from 1979 and concerns avian cell lines,[11] and 2 much more recent publications concern ESCs and iPSCs[12, 13] Thus, without appropriate guidelines, scientists have been very inventive in their attempt to name cell lines. Whereas some try to use names that are long

*Address correspondence to: Swiss Institute of Bioinformatics, University of Geneva, CMU, 1, Rue Michel Servet, 1211 Geneva 4, Switzerland. E-mail: amos.bairoch@sib.swiss
doi: 10.7171/jbt.18-2902-002

enough to be unique and thus, unambiguous, many resort to use very short (2–4 characters) names, which are a disaster in terms of specificity. For example, 10 names (C2, CF, DL, K8, ME, OS3, PC-1, PC-3, ST-1, and TK) are associated with 37 different cell lines! So far, we identified 350 cases of identical cell line names, but if one takes into account collisions between the name of 1 cell line and the synonym of another, as well as names that only differ as a result of punctuation (for instance KMH-2 and KM-H2), then the number of nonunique cell-line names rises to slightly above 900 (see ftp://ftp.expasy.org/databases/cellosaurus/cellosaurus_name_conflicts.txt for an up-to-date list of all cell-line name conflicts found so far).

Until recently, both issues were made even more acute as a result of the lack of comprehensive cell-line bioinformatics resources that would report the name of existing cell lines and thus, help a researcher to use a name that is not yet "taken," as well as a centralized compendium of STR profiles to help the community in its efforts to authenticate cell lines. As it will be described in the next section, this is one of the reasons that led to the development of the Cellosaurus.

WHY DEVELOP THE CELLOSAURUS?

In our efforts to annotate with precision the phenotypic effects of protein variations in the context of the development of neXtProt, the knowledge platform on human proteins,[14, 15] we wanted to make use of a cell-line reference resource that would contain a minimal amount of information on all of the cell lines that were used in publications for which we derived annotations. Whereas we found a large number of resources that contain information relevant to cell lines, we could not find one that answered our needs. We thus started to develop, in 2012, what was meant at first to be a simple cell-line thesaurus, hence, the name "Cellosaurus." As we became aware of the acute problem of cell-line misidentification and of the needs of the life sciences community for a comprehensive cell-line-curated resource, the Cellosaurus evolved to become a knowledge resource. To understand the situation that prevailed before the Cellosaurus became available, we describe the "ecosystem" of bioinformatics resources relevant to cell lines.

Cell-line databases

The most similar resource to the Cellosaurus, in terms of its aim and content, is the Cell Line Data Base (CLDB; or HyperCLDB; http://bioinformatics.hsanmartino.it/hypercldb/).[16] The CLDB provides detailed information pages for each cell line stored in the database. Compared with the Cellosaurus, it is more comprehensive in terms of information concerning culture conditions and some biochemical properties, such as the immunologic and cytogenetics profiles, but it lacks slightly more than half of the 37 different data fields present in the Cellosaurus. The main deficiency of the CLDB is its scope: it currently encompasses 6643 cell-line entries, originating from 12 cell-line collections, 6 of them Italian biobanks that are not distributing their cell lines widely. As by design, it provides separate entries if a cell line is distributed by different cell-line collections (for instance, the Hep G2 cell line has 9 different entries), its scope is even smaller, and it provides information on 5400 different cell lines, which corresponds to ~5% of the current number of entries in the Cellosaurus.

Many specialized cell-line databases have been established over the years that cater to a specific taxonomic range or a category of cell lines. Most of these resources are unfortunately no longer maintained and/or available, as is the case for FICELdb (for fish cell lines), CapCellLine (for prostate cancer cell lines),[17] and NISES (for insect cell lines). Currently, we are aware of only 4 active resources: the cell line section of the International Immunogenetics Information System/Human Leukocyte Antigen database (https://www.ebi.ac.uk/ipd/imgt/hla/),[18] which lists cell lines that are used in the context of the description of the alleles of the human major histocompatibility complex, and 3 resources, which are specific to ESCs and iPSCs: the Human Pluripotent Stem Cell Registry (hPSCreg; https://hpscreg.eu),[19] the Stemcell Knowledge & Information Portal (SKIP; https://www.skip.med.keio.ac.jp/en/), and the International Stem Cell Registry (ISCR; https://www.umassmed.edu/iscr/).

Ontologies

There are a number of ontologies that cater either primarily or partially to cell lines. Before describing them, it must first be emphasized that the purpose of a life science ontology is not identical to that of a knowledge base. The primary goal of an ontology is to alleviate terminological ambiguities by categorizing and defining the objects that it describes, whereas a knowledge base attempts to capture as much knowledge as possible on its subject of interest. Therefore, whereas all of the ontologies described hereinafter are useful in the context of the issue of the precise identification of cell lines, they only provide minimal information on these lines, *i.e.*, their name, their species of origin, the precise description of the category of cell line, and in general, a single reference to a publication or a cell-line collection catalog number.

The Cell Line Ontology (CLO; http://www.clo-ontology.org/)[20] is currently the sole cell-line-specific ontology. It contains 37,317 terms that describe 36,165 different cell lines (as a result of some redundancy). In terms of scope, it encompasses all of the cell lines described in CLDB, as well as those distributed by the Coriell Institute for Medical Research (https://www.coriell.org; but not all subcollections) and by Riken. After the Cellosaurus, CLO is therefore the resource that describes the greatest number of cell lines.

The Braunschweig Enzyme Database (BRENDA) Tissue and Enzyme Source Ontology (BTO; https://bioportal.bioontology.org/ontologies/BTO)[21] was created to facilitate the annotation of the enzyme data stored in BRENDA with a structured network of source tissues, cell types, and cell lines. It currently contains 6000 terms, 2147 of which describe cell lines. Unlike CLO, the cell-line entries are nonredundant. It contains many cell lines that were not, until we created the Cellosaurus, described in other resources, as they were entered by BRENDA curators while annotating papers describing enzymes.

The third ontology, with respect to the number of cell-line terms, is the Experimental Factor Ontology (EFO; https://www.ebi.ac.uk/efo/),[22] which was developed for the purpose of helping to annotate the metadata associated with experimental data captured in European Bioinformatics Institute (EBI) resources. EFO currently contains 1300 cell-line terms. There is a significant overlap between EFO and CLO, and recently,[23] the 2 groups published a strategy (based on our mappings to these 2 ontologies) to merge the cell lines unique to EFO into CLO.

Molecular Connection distributed in 2010 a first version of an ontology (MCCL; https://bioportal.bioontology.org/ontologies/MCCL) describing 505 cell lines, but it did not sustain this development. There are other ontologies that include cell-line terms, but as the number of their cell-line terms is rather small (no more than 50), we will not describe them here.

Cell-line collections

According to our analysis, there are ~50 organizations that are major players in the distribution of cell lines (*i.e.*, they each distribute >10 different cell lines). Whereas they cannot be considered as bioinformatics resources, they constitute, de facto, a source of information for the cell lines that they distribute. Unfortunately, they are rarely savvy in terms of good information management practices and do not offer any tools that allow the integration of their data into other resources. Furthermore, the information they provide in their product pages is very heterogeneous, in free text, rarely standardized, and often contains errors. As will be discussed in the section on annotation strategies, we are collaborating with many cell-line collections, and they are slowly but gradually improving their practices. Another drawback of these resources is that, by design, they only provide information on the cell lines that they distribute, thus fragmenting the information space on cell lines in as many silos as there are cell-line collections.

Experimental portals/repositories

There are an increasing number of portals that are built around initiatives that perform large-scale omics experiments on cell lines. Four of them were developed specifically in the context of projects that study large panels of human cancer cell lines; these are the following: the Cancer Cell Line Encyclopedia (CCLE; https://portals.broadinstitute.org/ccle),[24] which provides gene expression, chromosomal copy number, and exome sequencing data for ~1000 lines; the Genomics of Drug Sensitivity in Cancer (GDSC) Project (https://www.cancerrxgene.org),[25] which reports the response of ~1000 lines to 266 anticancer drugs; the MD Anderson Cell Lines Project (http://tcpaportal.org/mclp/#/),[26] which measures the level of protein expression using reverse-phase protein arrays in 650 lines; and the Sanger Institute cancer cell-line project (http://cancer.sanger.ac.uk/cell_lines), which provides exome and copy-number variation data for a little more than 1000 lines. It must be noted that whereas each of these 4 projects encompasses ~1000 cell lines (650 for GDSC), they are not studying the same sets of cell lines. With the combination of the targets of these projects, we count a total of 1653 different cell lines. A fifth portal—the Human iPSCs Initiative (http://www.hipsci.org)[27]—is performing and reporting extensive molecular characterization (genomics, transcriptomics, and proteomics) on a large set (currently 859) of iPSCs from healthy and diseased donors.

Generic experimental repositories also capture high-throughput information relevant to cell lines. This is the case of the National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO; https://www.ncbi.nlm.nih.gov/geo/),[28] which stores microarray, next-generation sequencing, and other forms of high-throughput functional genomic data, and of 3 EBI resources: ArrayExpress (https://www.ebi.ac.uk/arrayexpress/),[29] whose scope is similar to that of GEO, MetaboLights (https://www.ebi.ac.uk/metabolights/)[30] for metabolomics experiments, and the Proteomics Identifications database (PRIDE, https://www.ebi.ac.uk/pride/archive/),[31] which captures mass spectrometry-based proteomics data. Whereas the cell-line-specific experimental repositories are generally consistent in terms of the identification and naming of the cell lines that they have analyzed, this is not the case of these generic repositories that are highly heterogeneous in the quality of the metadata that they capture. It is currently impossible to query reliably these repositories to obtain the experimental data originating from a particular cell line, as this information is either in free text with many spelling variations and misspellings or is totally absent and can only be identified by going back to the original publications.

Integrative databases/portals

These are resources that pull together various publicly available omics data sets so as to present an integrated view of the status of a specified subset of cell lines. We have identified the following 4 such resources, all developed around human cancer cell lines:

- CellFinder (http://www.cellfinder.org/)[32] is a repository of microscopic and anatomic images, expression profiles from RNA Sequencing, microarrays, and protein expressions profiles for cell lines and tissues.
- CellMiner (https://discover.nci.nih.gov/cellminer/home.do)[33] is a web portal that integrates the results of the wealth of omics experiments performed on the National Cancer Institute (NCI)-60 cell line panel (see Auditing/classification comments).
- The Colorectal Cancer Atlas (http://www.colonatlas.org)[34] integrates information from CCLE, Catalogue of Somatic Mutations in Cancer (COSMIC), and a number of proteomic studies for 179 colorectal cancer cell lines.
- The Integrated Genomic Resources of human Cell Lines for Identification (http://igrcid.ibms.sinica.edu.tw/cgi-bin/index.cgi)[35] integrates microarray expression data and somatic mutations from COSMIC (release v47) and from a TP53 mutation database for 520 cell lines.

STR profile sites

As mentioned in the introduction, STR profiling is a powerful method to authenticate cell lines. Therefore, it is important to make available the reference STR profiles for as many cell lines as possible. Until the Cellosaurus started incorporating this information, it was not easily found in any resource. Some cell-line collections display the STR profile of the lines that they distribute, but many others do not. However, there are a number of web sites where users can compare the STR profile of their cell line with an internal database of profiles. Unfortunately, these sites do not provide their profile database as a downloadable file. The following 2 cell-line collections offer this search option: ATCC (https://www.atcc.org/STR_Database.aspx) and Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ) (https://www.dsmz.de/services/services-human-and-animal-cell-lines/online-str-analysis.html).[36] The CLDB includes what it terms the Cell Line Integrated Molecular Authentication Database (http://bioinformatics.hsanmartino.it/clima2/), and the Childhood Cancer Repository has an STR profile search option (https://strdb.cogcell.org/search_strname.php).

As mentioned in Biologic samples resources, NCBI BioSample contains ~3000 cell-line entries associated with STR profiles.

Databases focusing on specific properties of cell lines

There are a few resources that are providing information on specific characteristics of cell lines. These include the following:

- The TOKU-E Cell-culture database (http://cell-lines.toku-e.com), for which its scope includes culture media and concentrations and combinations of antibiotics for selection and transfection experiments for ~440 cell lines.
- The Comparative Genomic Hybridization (CGH) Data Base (http://www.cghtmd.jp/CGHDatabase/index_e.jsp) provides information on genomic aberrations detected through CGH for 511 human cancer cell lines.
- Pawefish (http://www.pawefish.path.cam.ac.uk/index.html), a collection of spectral karyotyping (SKY) and molecular cytogenetic data for 91 human epithelial cancer cell lines.
- The SKY/Multiplex Fluorescence *In Situ* Hybridization (M-FISH) & CGH Data Base (https://www.ncbi.nlm.nih.gov/dbvar/studies/nstd136)[37] provides cytogenetics data obtained by CGH, SKY, and M-FISH on ~170 human and mouse cancer cell lines. This database is no longer maintained, but its data can still be downloaded.

Specialized databases

There are 2 large, curated resources, respectively, in the fields of cancer and chemistry that contain a wealth of information obtained from biologic samples, some of which are cell lines.

COSMIC (http://cancer.sanger.ac.uk/cosmic)[38] is a resource that catalogs information on somatic mutation in human cancer. In its current release (v.84), COSMIC contains ~5.5 million somatic mutations originating from 1.4 million samples, and it has curated >25,000 papers. A small (17,000) but non-negligible number of samples are cancer cell lines. However, as COSMIC assigns a new sample identification number to every sample listed in a paper or integrated from an external resource, and it does not attempt to standardize cell line names, the process of mapping cell lines to COSMIC samples is not trivial. For example, there are 79 COSMIC samples that correspond to the HCT 116 cell line, and these samples are associated with 3 small variations of its name (HCT 116, HCT-116, and HCT116). As we are continuously mapping COSMIC samples to Cellosaurus entries, we can estimate that it currently contains information relevant to ~5000 different cell lines.

ChEMBL (https://www.ebi.ac.uk/chembl/),[39] which contains binding, functional, and Absorption, Distribution, Metabolism, Excretion, Toxicity information for drug-like bioactive compounds, annotates in which cell lines an assay has been implemented. Currently, its cell-line table contains

1624 entries, of which ~1300 can be mapped to a precise cell line. The remaining terms are relevant to primary cells or groups of cell lines (for example, "Panel NCI-60") or are too ambiguous to be mapped to a known line.

Biologic sample resources

In recent years, a number of bioinformatics resources have been developed that contain the descriptions of biologic materials used in experimental assays. Cell lines are included in the scope of these resources. Three of the following resources are generic and accept submissions from a variety of experimental projects, including individual laboratories:

- The NCBI BioSample (https://www.ncbi.nlm.nih.gov/biosample/)[40] and the cognate EBI BioSamples (https://www.ebi.ac.uk/biosamples/)[41] each collects information on the biologic samples used in submission to their respective nucleotide-sequence archives and repositories, as well as other projects.
- eagle-i (https://www.eagle-i.net/)[42] is a project to which 40 academic and not-for-profit research institutions in the United States currently participate and submit information on their samples (here, called resources).

It is difficult to estimate the number of cell lines that are represented in these resources, as they are evolving quite rapidly, and they are quite heterogeneous in terms of how they store and represent cell lines. What must be noted is that the NCBI BioSample contains 2 specific sets of entries that are relevant to cell lines: misidentified cell lines (451) and human cell lines with an STR profile (3039). eagle-i describes ~2000 different iPSCs.

The following 4 large research consortiums have created data portals that provide information on all of the biologic samples used in experiments carried out in the context of their projects:

- The Encyclopedia of DNA Elements (ENCODE; https://www.encodeproject.org).[43] Currently, it has 5400 cell line "biosamples" that map back to ~450 different cell lines.
- The International Genome Sample Resource (IGSR; http://www.internationalgenome.org) contains information on all of the Epstein-Barr virus-immortalized cell lines (so far, ~3400) that have been the target of this large human genome-sequencing project.
- The Library of Integrated Network-based Cellular Signatures (LINCS) program (http://lincsportal.ccs.miami.edu/dcic-portal/),[44] which aims to characterize how a variety of human cells and tissues responds to perturbations by drugs and other factors; it has a data portal that describes 1127 cell lines.
- The recently created 4D Nucleome Data Portal (https://data.4dnucleome.org),[45] which is targeted toward the study of nuclear organization in space and time, contains ~50 cell lines "biosources" that correspond to as many human and mouse cell lines.

SCOPE OF THE CELLOSAURUS

The Cellosaurus provides information on immortalized cell lines (*e.g.*, transformed or cancer cell lines), naturally immortal cell lines (*e.g.*, stem cell lines), as well as cell lines with a finite lifespan when these are distributed and used widely. It does not encompass primary cells.

In terms of species, the Cellosaurus covers both vertebrate and invertebrate (insects and ticks) cell lines. It does not include plant cell lines.

CONTENT OF THE CELLOSAURUS

For each cell line that it describes, the Cellosaurus provides a wealth of information. Some information items are mandatory, whereas others are optional.

Mandatory fields

- A recommended name. This is most frequently (but not always) the name provided in the original publication.
- A unique primary accession number. This accession number should be used to reference unambiguously a specific cell line. Cellosaurus accession numbers make use of the name space "CVCL." The name space is followed by an underscore and 4 alpha-numerical characters (*e.g.*, CVCL_0E45).
- The species of origin. We use the NCBI Taxonomy database (https://www.ncbi.nlm.nih.gov/taxonomy)[46] as our reference taxonomic resource, and we provide cross-reference to that resource. Note that a hybrid cell line or an hybridoma can originate from >1 species.
- The category to which a cell line belongs. Currently, this field can take 1 of 14 values: adult stem cell, cancer cell line, conditionally immortalized cell line, ESC, factor-dependent cell line, finite cell line, hybrid cell line, hybridoma, iPSC, spontaneously immortalized cell line, stromal cell line, telomerase-immortalized cell line, transformed cell line, and undefined cell line type.

Descriptive fields that are not mandatory but very often present

- Synonyms: we try to list all of the different synonyms for a cell line, including alternative use of lower- and uppercase characters. It should be noted that misspellings are not included in synonyms; they are stored in a specific structured comment that is described later.
- Sex (gender) of the individual from which the cell line was derived: this field can take 1 of 5 values: female, male, mixed sex, sex ambiguous, or sex unspecified.
- The age of the individual from which the cell line has been derived (at the time of "sampling").

Relationships fields

- If a cell line originates from another one, then this is captured in a "hierarchy" field that provides a cross-reference to the parent cell line. Note that parent cell lines do not explicitly contain the list of their children, as this information can be inferred automatically.
- If a cell line originates from the same individual as other cell line(s) (often termed "autologous" or "sister" cell lines), then cross-references to these autologous cell line(s) are provided.

Disease information

Many cell lines were established either from cancerous tumors in patients or animals or from individuals suffering from a monogenic genetic disorder. Therefore, we report the name(s) of the relevant disease(s) (an individual can be affected by >1 disease), and we provide a link to the disease definition in the NCI Thesaurus (https://ncit.nci.nih.gov/ncitbrowser/).[47] We chose to use the NCI Thesaurus over other disease ontologies, as it is comprehensive in terms of cancer terms for human, mouse, and rat. Whereas it is less complete in terms of human genetic disorders and for cancers in other species, we have established a very fruitful collaboration with the team developing this resource, and they have added all of the terms (>500 so far) that we required for a complete coverage of the diseases represented in the Cellosaurus.

We have recently started to capture important sequence variations (compared with the reference genome of the species). Most of these variants belong to 1 of the following 2 categories: somatic mutations in oncogenic genes in cancer cell lines and inherited or *de novo* genomic mutations in genetic-disorder cell lines. We store this information in what we term a structured comment, called sequence variation. There are many different types of structured comments that provide specific information. They are described later and have been categorized as follows.

Structured comments relevant to general characteristics of a cell line

Characteristics, which as its name implies, serve to indicate some specific characteristics of a cell line, such as its intended use or some interesting property (*e.g.*, the dependence on a growth factor, the permissiveness or known infection by a virus, the capacity to differentiate, *etc.*).

Biotechnology is used to describe the use of a cell line in a biotechnological context (such as its use for a specific bioassay or the production of a vaccine).

Breed/subspecies is used to specify from which breed, strain, or subspecies an animal or insect cell line was derived.

Doubling time is used to store the population doubling time of a cell line. Curiously enough, although this information is quite useful to researchers who are going to cultivate a cell line in their laboratory, it is provided by only a handful of cell collections and often needs to be retrieved from publications.

Microsatellite instability is used to report the status of a cell line in terms of its genetic hypermutability that results from impairment in the DNA mismatch repair pathway. Cell lines can be reported to be microsatellite stable or microsatellite instable-high or -low.

To specify if a cancer cell line originates from a site that is different from that of the original tumor, we use 2 different comments: "derived from metastatic site" and "derived from sampling site"; the latter is used principally for hematopoietic and lymphoid malignancies.

Omics is used to indicate if a cell line has been the target of an "omics" experimental study. For a cell line, there are as many omics comments as there are different types of omics experiments that have been carried out on that cell line. Examples of omics comments include the following: genome sequenced, deep RNA Sequencing analysis, single nucleotide polymorphism array analysis, transcriptome analysis, *etc.*

Structured comments for cell-line engineering/transformation

The transfected comment is used to indicate what foreign genes have been inserted into a cell line so as to be expressed either constitutively or by induction by a specific stimulus. If possible, we indicate the exact nature of the gene introduced in the cell line by cross-referencing to 1 of the following gene/protein resources: FlyBase (http://flybase.org/)[48] for Drosophila genes, HUGO Gene Nomenclature Committee (https://www.genenames.org/)[49] for human genes, Mouse Genome Informatics (http://www.informatics.jax.org/)[50] for mouse genes, Rat Genome Database (https://rgd.mcw.edu/)[51] for rat genes, and Universal Protein Resource Knowledgebase (UniProtKB; http://www.uniprot.org/)[52] for all other species.

The aim of the knockout cell comment is the reverse of the transfected comment: it is used to describe genes that have been partially or completely knocked out from a cell line. As for transfected, the knockout cell comment can be cross-referenced to FlyBase, HUGO Gene Nomenclature Committee, Mouse Genome Informatics, Rat Genome Database, or UniProtKB. In addition to the gene name and cross-reference, it indicates the methodology that has been used to disable the gene (*e.g.*, clustered regularly interspaced short palindromic repeats/clustered regularly interspaced short palindromic repeats-associated protein 9, gene trap, homologous recombination, knockout mouse, short hairpin RNA knockdown, *etc.*).

The selected for resistance to comment: resistance to anticancer drugs is a major problem in the oncology field, thus leading to the development of many drug-resistant cell lines that are used to identify drug-resistance mechanisms.[53] For such cell lines, we indicate to which compound they have been made resistant. In addition to the name of the chemical compound, we provide a cross-reference to the Chemical Entities of Biologic Interest (ChEBI) database (https://www.ebi.ac.uk/chebi/).[54] In the case of resistance to a large molecule, such as a therapeutic mAb, we cross-reference to DrugBank (https://www.drugbank.ca/)[55] and for protein toxins, to UniProtKB.

The transformant comment serves to indicate the agent that was responsible for the transformation of a normal, finite life cell into an immortal cell line. We use this comment both for artificially transformed cell lines and for cancer cell lines that have arisen through viral carcinogenesis. When possible, we cross-reference to the NCBI Taxonomy database for viruses, to ChEBI for chemical compounds, and to the NCI Thesaurus for all forms of irradiation.

Warning comments

The problematic comment conveys a very important information item—that a cell line is known or suspected to be contaminated or misidentified. In this comment, we describe what the cell line was originally thought to be and what it really is, as well as the source of this information. On the ExPASy server version of the Cellosaurus (see THE CELLOSAURUS ON ExPASy), this information is displayed in red so as to highlight it.

The caution comment warns users of potential problems, ambiguities, or discrepancies in the information provided in a cell-line entry, for example, if 1 source states that a cell line originates from a male patient, whereas another states that it is from a female patient.

The misspelling comment, as its names implies, serves to record incorrect spellings of cell-line names. Such information is especially useful in the context of literature text-mining activities. It is also important when a misspelled name is used in an external resource to which the Cellosaurus cross-reference helps users to understand the discrepancy between the name used in that resource and the ones provided in the Cellosaurus.

The discontinued comment is used to indicate if a cell line has been discontinued from a cell catalog. This is important information that is unfortunately rarely available from cell-line collections and companies distributing cell lines.

Auditing/classification comments

The group comment: to help some research communities, we explicitly “group” together some cell lines so that they can be retrieved easily without the need to perform complex queries. Many of the groups are there to help find cell lines that belong to a specific taxonomic range. These include the following: amphibian, bat, bird, cetacean, crustacean, fish, insect, marsupial, mollusk, nonhuman primate, reptilian, and tick. Other groups are targeted toward the intended biotechnological use of a cell line. These include the following: clinical-grade human ESC, human/rodent somatic cell hybrids, hybridoma fusion partner, recombinant protein production, serum/protein-free medium, adenovirus packaging, retrovirus packaging, and vaccine production. Four groups do not belong to the above categories; these are the following: cancer stem cell, triple-negative breast cancer, haploid karyotype, and endangered species/breed. We believe that the last group will become associated with an increasing number of cell lines, as conservation groups are embarking on initiatives[56] that plan to use iPSC technology to preserve or rescue endangered species.

The part-of comment is used to indicate if a cell line belongs to a specific cell panel or collection. The most well-known example of such a type of panel is the NCI-60 cancer cell line panel (https://dtp.cancer.gov/discovery_development/nci-60/) that has been used since the mid-1990s to test for anticancer compounds. However, the NCI-60 is the only one of almost 80 panels that we have identified so far.

The registration comment is used to specify if a cell line has been entered in an official registry, along with its identifier in that registry. Three types of registry are important in the realm of cell lines. *1*) Those that keep track of ESC lines for which use is approved in a specific country. The most well-known one is the NIH Human Embryonic Stem Cell Registry (https://grants.nih.gov/stem_cells/registry/current.htm) that lists the ESC lines that are eligible for use in NIH-funded research. *2*) The International Cell Line Authentication Committee (ICLAC) register of misidentified cell lines that we will describe in THE CELLOSAURUS AND ICLAC, which exemplifies our collaboration with ICLAC. *3*) Patent-related registries; cell lines that are described in patents fall under the auspice of the Budapest Treaty (http://www.wipo.int/treaties/en/registration/budapest/), which regulates the deposition of microorganisms (cell lines are included in this broad category) in 1 of the recognized International Depositary Authority. We capture this information by providing the name of the International Depositary Authority in which the cell line has been deposited and its registration number.

The from comment is used to indicate the research and/or the institution that has established a cell line. It is used either when there is no publication from the originator of the cell line or in the case of ESCs and iPSCs, to indicate the institution responsible for the establishment of the cell line.

Hybridoma-specific comments

For hybridomas, we provide 2 specific, structured comments: mAb target, which allows the description of the target of the produced mAb, and mAb isotype, which is used to indicate the isotype of the produced mAb (*e.g.*, "IgG2b, κ"). The target of a mAb can be a protein, chemical compound, bacteria, or virus. In cases where a mAb binds to a precisely defined protein or chemical compound, the mAb target comment is cross-referenced to a UniProtKB or a ChEBI accession number, respectively.

Other comments

Population is used for cell lines that are applied in a population studied to indicate to which ethnic group the donor belongs. Anecdotal is used for anecdotes concerning the establishment of a cell line or on the donor of that cell line. Miscellaneous is used for anything not falling into the scope of the other defined comment types.

STR markers

A standard for the authentication of human cell lines by STR profiling was established by a working group convened by ATCC and was approved in 2011 by the American National Standards Institute (ASN-0002-2011).[57, 58] It requires the use of 8 STR markers (CSF1PO, D13S317, D16S539, D5S818, D7S820, TH01, TPOX, and vWA), plus amelogenin, which is used for gender determination. Whereas cell-line collections generally use these 8 + 1 loci, currently, many authentication services test for up to 18 loci (the 9 additional markers are the following: D18S51, D19S433, D21S11, D2S1338, D3S1358, D8S1179, FGA, Penta D, and Penta E). Additional loci have been used occasionally by some groups over the years (we have identified 14 of them), but their use is much less prevalent than that of the 18 loci listed above. It must also be noted that whereas there are discussions to increase the number of STR loci in a revised American National Standards Institute standard, such discussions need to balance on 1 hand, the respect of the donor privacy and on the other hand, the usefulness of this information to ensure cell-line authentication. For the authentication of dog cell lines, a panel of 10 STR markers (FHC2010, FHC2054, FHC2079, PEZ1, PEZ3, PEZ5, PEZ6, PEZ8, PEZ12, and PEZ20) has been proposed.[59]

In the Cellosaurus that we decided to store for both human and dog cell lines, the results of the STR profiling originated from as many independent sources as possible. This ensures the largest possible coverage of loci, but more crucially, it allows the representation of conflicting results among different profile sources. In general, most of the conflicts are minor and are probably a result of genetic drift,[60] but some are caused by various reasons that include clerical errors, as well as mix-ups between the results of 2 different loci.

We believe that this section of the Cellosaurus is highly important in the context of cell-line authentication and helps to alleviate the problem of cell-line misidentification.

Literature references

We provide the references for publications describing the establishment of a cell line, its characterization, as well as relevant reports of high-throughput omics experiments. Publications can belong to 1 of the following 4 categories: published papers, book chapters, patents, and theses (Doctor of Philosophy, Doctor of Medicine, Doctor of Veterinary Medicine, Master of Science, or Bachelor of Science). Each reference is uniquely identified by 1 of the following 4 types of identifiers: a PubMed identifier, a digital object identifier, a patent number, or an internal identifier, for what we term CelloPub references. These internal identifiers are assigned to references that cannot be associated with 1 of the first 3 types of identifiers. As it is difficult to find information on these CelloPub references, we have compiled a file containing all of the necessary information [author(s), title, journal name, web link, and if it is available, the abstract]. This file is available for download by file transfer protocol (FTP), and its information is integrated in the cell-line view on ExPASy (see The Cellosaurus on ExPASy).

Cross-references

To help users explore a maximum of information relevant to cell lines, we provide an extremely large number of cross-references to 75 different resources:

- Cell-line collections/catalogs: AddexBio, ATCC, BCRC, BCRJ, CBA, CCLV, Cell_Biolabs, CLS, Coriell, DGRC, DSMZ, DiscoverX, ECACC, ICLC, Imanis, IZSLER, JCRB, KCB, KCLB, Millipore, MMRRC, NCBI_Iran, NCI-DTP, NHCDR, NIH-ARP, NISES, RCB (Riken), RSCB, TCB, TKG, TNGB, WiCell, and Ximbio.
- Ontologies: BCGO, BTO, MCCL, CLO, EFO, MeSH, and Wikidata.
- Cell-line databases/resources: CCLE, CCRID, CHG-DB, CLDB, Colorectal Cancer Atlas, Cosmic-CLP, dbMHC, ESTDAB, FlyBase, GDSC, hPSCreg, IHW, IMGT/HLA, IGSR, ISCR, LINCS_HDP, LINCS_HMS, Lonza, SKY/M-FISH/CGH, SKIP, and TOKU-E.
- Resources that list cell lines as samples: 4DN, BioSample, BioSamples, ChEMBL (cells and targets), Cosmic, eagle-I, ENCODE, GEO, International Genome Sample Resource, MetaboLights and PRIDE.

Some of these resources have recently started to link back to us. This is currently the case of BioSample, ChEMBL, ENCODE, hPSCreg, JCRB, Ximbio, and Wikidata.

Web links

Whenever this is necessary, we provide links to pages on the web that provide information on a cell line. These pages include links to some commercial cell-line providers for which it was not possible to create a cross-reference (as these require a mechanism to create automatically a valid uniform resource locator (URL), based on the catalog number of the cell line). We also include links to cell line-relevant pages in sites, such as those of technology transfer offices and research laboratories and in Wikipedia.

There is quite a lot of volatility in terms of web pages; they often move and unfortunately, in some cases, disappear completely. When this happens, we try to find the last archived version of that page in the Internet Archive (http://web.archive.org/), and we link back to the archived page.

ANNOTATION STRATEGY AND SOME CONSIDERATIONS CONCERNING THE SOURCES OF INFORMATION USED IN THE CELLOSAURUS

We use a 3-tiered approach to populate the Cellosaurus with information concerning cell lines, using data provided by on-line cell-line catalogs/product-information pages, publications that describe the establishment of cell lines, or some relevant properties of already-established lines. We integrate unpublished information submitted by researchers. Each of these 3 strategies is associated with interesting sociological and technical considerations.

As we mentioned earlier, cell-line collections are rarely savvy in terms of good information-management practices. The deficiencies that we encountered quite frequently are the following: they do not provide stable URLs for their products, and these URLs are often not based on the catalog number of the cell line. Whereas they often (but not always) indicate new "offerings," they never provide a list of cell lines for which they have discontinued the distribution. Finally, some of them are reluctant to provide the STR profiles for their human cell lines. To balance this negative view, we emphasize that we are in contact with many of these organizations and that most of them have been very responsive to our criticisms. This is especially the case in terms of error corrections. What remains true is that from a technological point of view, extraction of information from these resources is time consuming, as it generally entails manual web scraping.

We have curated already >15,000 publications relevant to the cell lines in the Cellosaurus. These publications consist of articles (14,366), patents (738), book chapters (125), and theses (43). As can be seen in **Fig. 1**, the oldest publications cited in the Cellosaurus were published in the late 1940s. The number of curated publications peaks in the 1990s (almost 500/yr) and after decreasing for ~10 yr, it is currently on a new, upward slope (probably as a result of the contribution of publications concerning ESCs and iPSCs).

Whereas thanks to agreements between scientific publishers and Swiss academic libraries, we can get a hold of many articles that are not open access, these agreement do not cover all publications and in many instances, do not cover access to "old" papers. The problem of lack of free access to the results of scientific research is an issue that is widely debated (see, for example, the United Nations Educational, Scientific and Cultural Organisation's position on open access to scientific publications: http://www.unesco.org/new/en/communication-and-information/access-to-knowledge/open-access-to-scientific-information/), but it is an especially acute problem for curated, knowledge resources, such as the Cellosaurus, whose staff needs to peruse a high number of scientific publications yet does not have the budget to pay the publishers' article fees. This situation is especially frustrating when one realizes that knowledgebases are providing links back to publications and are, in effect, increasing the traffic toward publisher sites. Thus, in the context of the establishment of the Cellosaurus, it was necessary to use a variety of nonstandard approaches to access the full-text version of papers behind paywalls.

In terms of perusing unpublished information submitted by researchers, the main issue that we encounter is the lack of responsiveness of many individuals. In the last

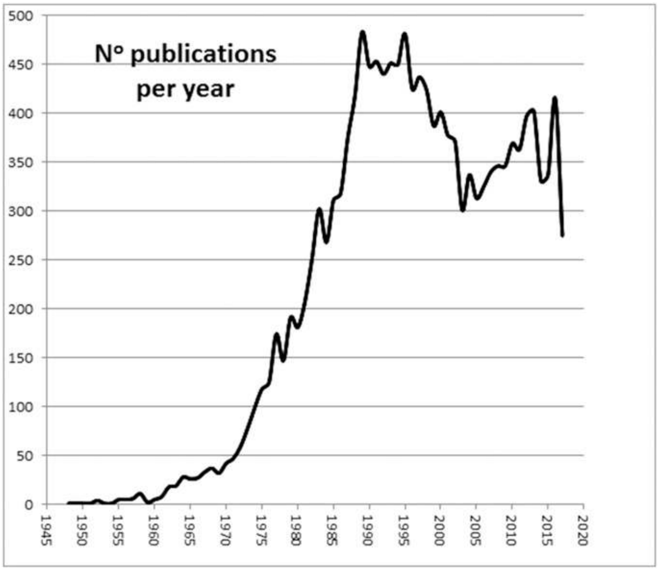


FIGURE 1

Evolution of the number of curated publications in the Cellosaurus, according to the year in which they were published.

3 yr, we have emailed ~1 thousand requests for information, and the response rate (after at least 2–3 separate attempts at 1-mo intervals) is <30%. Granted, some of these emails may have landed in spam folders or have been blocked by various filters, but we believe that in most cases, they were ignored because of a lack of time and/or interest by the recipients. It did not escape our attention that the response rate was lower than average when the email contained questions relevant to errors or inconsistencies in the papers authored by the scientists that we were trying to contact!

STATISTICS OF THE CURRENT RELEASE

The current Cellosaurus release (release 25 of March 2018) describes 101,528 cell lines from 590 species. In terms of species distribution, the first 10 species comprise 97.5% of all of the entries: 74,534 (82%) from human, 19,137 (19%) from mouse, 1908 (1.8%) from rat, 1580 (1.5%) from Chinese hamster, 571 (0.5%) from dog, 357 from chicken, 272 from bovine, 244 from *Drosophila melanogaster*, 230 from chimpanzee, and 183 from pig. The great majority of species is only represented by very few cell lines: 225 species have a single cell line, 114 have 2, 118 have from 3 to 5, 58 from 6 to 10, and 62 from 11 to 100, and only 13 are associated to >100 cell lines. In terms of groups of species, many papers, reviews, or web sites report outdated and underestimated statistics concerning the number of insect and fish cell lines that have been established; it is therefore beneficial to report that there are 956 insect and 551 fish cell lines in the current Cellosaurus release.

The 6 most represented cell-line categories are the following: transformed cell lines (43,853; 43%), cancer cell lines (18,284; 18%), ESCs (13,479; 13%), finite cell lines (8553; 8%), iPSCs (7234; 7%), and hybridomas (4264; 4%). The high number of transformed cell lines is primarily a result of the establishment in the last 40 yr by the Coriell Institute for Medical Research of >30,000 Epstein-Barr virus-transformed peripheral blood B-lymphocyte cell lines.

Release 25 contains 56,610 synonyms, 83,338 references to 15,116 publications, 203,181 cross-references to 73 distinct resources (cell line catalogs, ontologies, and databases), and 12,928 web links.

In terms of STR profiles, these are available for 5401 human and 28 dog cell lines, originating from 273 distinct sources (cell-line collections, publications, submissions, *etc.*). Whereas this is, by far, the largest publicly available collection of STR profiles, it is also exemplified how much still remains to be done to obtain a comprehensive coverage of human cell lines.

There are 810 (0.8%) cell lines that are tagged as being problematic (misidentified or contaminated). This is probably the tip of the iceberg as the above-mentioned dearth of STR profiles for human cell lines and their absence in other species (except for dog) are likely to hide many contenders for future inclusion in this category.

Miscellaneous, other interesting statistics include the following:

● Cell lines [43,113 (42%)] are linked to 1 or more disease(s) (cancers or genetic disorders), and 5170 (5%) entries contain information on sequence variations; the latter number is expected to increase significantly in the near future, as we have only recently started to annotate these variations.

● Cell lines [3804 (3.7%)] have information on their population doubling time.

● We have managed to track down 5593 discontinued catalog numbers from cell-line collections/organizations.

THE CELLOSAURUS ON EXPASY

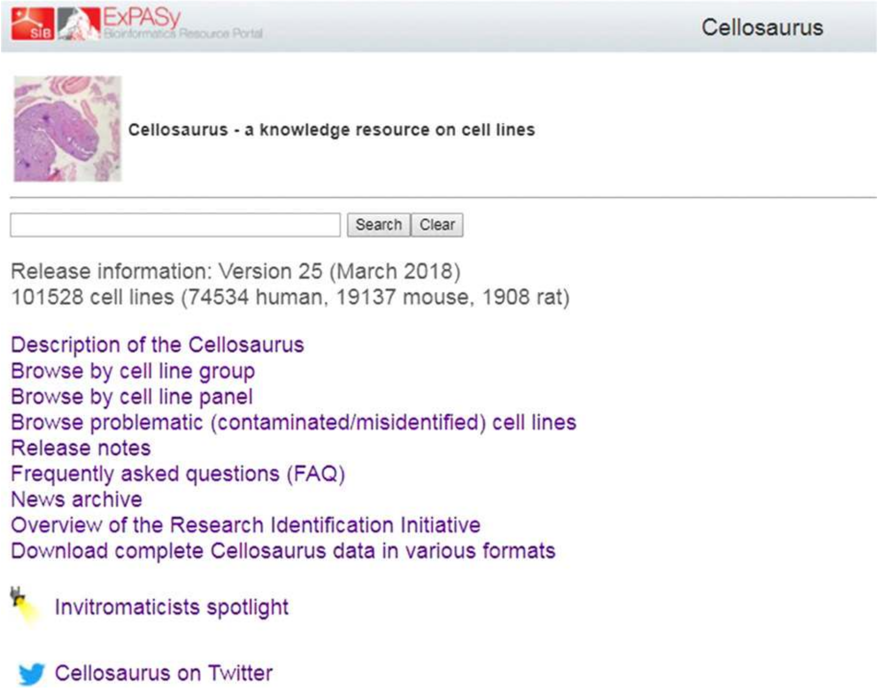
Since mid-2015, the Cellosaurus has been available on the Swiss Institute of Bioinformatics ExPASy web server[61] (https://web.expasy.org/cellosaurus/). As seen in **Fig. 2**, the content of the home page of the Cellosaurus on ExPASy is quite simple. It consists of a search bar, as well as links to a number of specific pages, some of which are described hereunder.

Currently, the search function of the ExPASy implementation of Cellosaurus is based on a full text search engine using the PERL Comprehensive Perl Archive Network Library Lucy (http://search.cpan.org/~nwellnhof/Lucy-0.6.2/lib/Lucy.pod) using Apache Lucene technology. Results of the search are displayed in a 3-column output, namely, the cell-line accession number, name, and species of origin (**Fig. 3**). The clicking on 1 of the accession numbers opens a dynamic hypertext markup language-based entry view (*e.g.*, https://web.expasy.org/cellosaurus/CVCL_0033) that presents all of the information available in the Cellosaurus for a given cell line in an attractive yet compact layout. These cell-line pages are dynamically linked to all external resources that are cited in the Cellosaurus (taxonomy, references, cross-references, web links, ChEBI, NCI Thesaurus, and genes/proteins resources).

The home page of the Cellosaurus on ExPASy also includes links to 3 pages that, respectively, allow users to browse the Cellosaurus by cell-line groups, cell-line panels, and problematic cell lines. A link to a page providing answers to frequently asked questions is also available.

Users who want to contact us directly from the web site can do so by clicking on the “Contact” link on the top right side of all of the Cellosaurus pages.

Currently, the Cellosaurus is accessed on ExPASy at an average of >1500 sessions/d. Since its inception in May 2015, it has served 350,000 users with a cumulative total of 570,000 sessions and >2 million page views.



Cellosaurus - a knowledge resource on cell lines

Release information: Version 25 (March 2018)
101528 cell lines (74534 human, 19137 mouse, 1908 rat)

Description of the Cellosaurus
Browse by cell line group
Browse by cell line panel
Browse problematic (contaminated/misidentified) cell lines
Release notes
Frequently asked questions (FAQ)
News archive
Overview of the Research Identification Initiative
Download complete Cellosaurus data in various formats

Invitromaticists spotlight

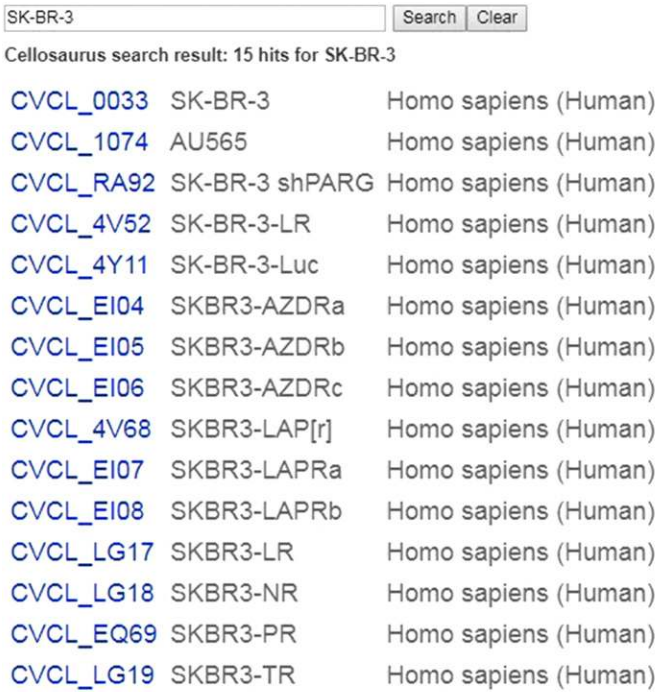
Cellosaurus on Twitter

FIGURE 2

Home page of the Cellosaurus on ExPASy.

DOWNLOADING THE CELLOSAURUS

The Cellosaurus is distributed under the Creative Commons Attribution-NoDerivs License (https://creativecommons.org/licenses/by-nd/3.0/) and can be downloaded by FTP (ftp://ftp.expasy.org/databases/cellosaurus). It is available in 3 different formats: structured flat file, Open Biomedical Ontologies (OBO), and XML (see Supplemental Figs. S1–S3 for examples of an entry in these 3 formats).



Cellosaurus search result: 15 hits for SK-BR-3

CVCL_0033	SK-BR-3	Homo sapiens (Human)
CVCL_1074	AU565	Homo sapiens (Human)
CVCL_RA92	SK-BR-3 shPARG	Homo sapiens (Human)
CVCL_4V52	SK-BR-3-LR	Homo sapiens (Human)
CVCL_4Y11	SK-BR-3-Luc	Homo sapiens (Human)
CVCL_EI04	SKBR3-AZDRa	Homo sapiens (Human)
CVCL_EI05	SKBR3-AZDRb	Homo sapiens (Human)
CVCL_EI06	SKBR3-AZDRc	Homo sapiens (Human)
CVCL_4V68	SKBR3-LAP[r]	Homo sapiens (Human)
CVCL_EI07	SKBR3-LAPRa	Homo sapiens (Human)
CVCL_EI08	SKBR3-LAPRb	Homo sapiens (Human)
CVCL_LG17	SKBR3-LR	Homo sapiens (Human)
CVCL_LG18	SKBR3-NR	Homo sapiens (Human)
CVCL_EQ69	SKBR3-PR	Homo sapiens (Human)
CVCL_LG19	SKBR3-TR	Homo sapiens (Human)

FIGURE 3

Example of a search for a cell line in the Cellosaurus on ExPASy.

The structured flat-file format is similar in its principles to that used by UniProtKB. The data are stored in different line types, prefixed by a 2-character line code [*e.g.*, "SY" for synonym(s), "DI" for disease(s), *etc.*]. The line code is followed by 3 spaces, and the relevant data make up the rest of the line. As it is the case for UniProtKB, an entry starts with an ID line (identification, containing the recommended name of the cell line) and ends with a "//" line. There is no defined maximum line length. The full records for the references are stored in a separate file from that of the cell-line entries that contain the identifiers of the references. A third file describes how to implement web links for the external resources to which the Cellosaurus cross-references.

The OBO format version of the Cellosaurus is provided for users who want to integrate core cell-line information (name, accession number, synonym, category, gender, hierarchy, *etc.*) within an ontological framework. It is not intended to contain all of the information available in the Cellosaurus (it does not contain STR profiles nor the age of the donor).

The XML format of the Cellosaurus is described in an associated schema definition file. All of the data in the Cellosaurus is available in the XML format. We encourage users who want to implement locally a full version of the knowledgebase to use the XML format version.

The FTP site also contains additional files, such as the list of deleted accession numbers, a file containing frequently asked questions, or one listing cell lines with identical names. The complete records for the CelloPub references are also available in the file "cellopub.txt." A "readme" file describes the full complement of the downloadable files.

The Cellosaurus is also available for download from GitHub (https://github.com/calipho-sib/cellosaurus). As a result of the current GitHub maximum file size limit of 100 Mb, the XML version is not stored on that platform. As GitHub acts as a version-control system, it can be used by users who want to retrieve an old version of the knowledgebase (the oldest version available on GitHub is release 11 of November 2014).

UPDATING SCHEDULE AND INFORMATION ABOUT RELEASES

The Cellosaurus is updated ~4–6 times/yr. New releases are announced on Twitter (@cellosaurus) and on ResearchGate (https://www.researchgate.net/project/Cellosaurus). We also use our Twitter channel to highlight specific Cellosaurus features and publicize publications reporting the establishment of new cell lines, as well as other relevant topics. At each release, release notes are prepared that describe all of the format and scope changes compared with the previous release. These release notes are available on the FTP site and from the ExPASy Cellosaurus home page (https://web.expasy.org/cellosaurus/cellosaurus_relnotes.txt). Twice each year, we email to our users a short newsletter. These newsletters are archived on ExPASy (https://web.expasy.org/cellosaurus/news_archive/).

THE CELLOSAURUS AND ICLAC

ICLAC (http://iclac.org/) was set up in 2012 by volunteers from research laboratories, cell-line collections, the pharmaceutical industry, and cell-line authentication service providers to make the community aware of the extent of the problem of cell-line contamination/misidentification. It also promotes authentication testing as effective ways to combat this problem. ICLAC maintains a "Register of Misidentified Cell Lines," as well as various guideline documents pertinent to related issues, such as how to help a scientist find a unique name for a new cell line. We work in close collaboration with ICLAC (to which we are a member) so as to integrate in the Cellosaurus all of the information relevant to what we call problematic cell lines, as well as to inform ICLAC of potential additions or modifications to the register. The ICLAC Register and Cellosaurus are bidirectionally cross-referenced.

THE CELLOSAURUS AND THE RESOURCE IDENTIFICATION INITIATIVE

The Resource Identification Initiative[62] (RRI) aims to "promote research resource identification, discovery, and reuse." The initiative introduced the concept of Research Resource Identifiers (RRIDs), persistent and unique identifiers for referencing a research resource. A critical goal of this initiative is the widespread adoption of RRIDs to cite biologic resources, such as antibodies, cell lines, organisms, or tools, in the biomedical literature and other places that reference their generation or use. RRIDs reuse established community identifiers where they exist. The Cellosaurus is the cell-line resource for this important initiative.

To ensure that they are recognizable, unique, and traceable, identifiers are prefixed with "RRID," followed by a repository-specific prefix that indicates the source authority that provided it. For the Cellosaurus, this is CVCL. In research papers, authors are thus encouraged to cite cell lines using sentences, such as "we have used HeLa (RRID: CVCL_0030) obtained from ATCC (catalog number CCL-2)."

The RRI has put in place a portal (https://scicrunch.org/resources) to search for these RRIDs. All cell lines in the Cellosaurus are integrated in this portal and are linked back to the relevant entry on the ExPASy server. If a cell line is not yet represented in the Cellosaurus, then authors are encouraged to ask us to create a new entry, and we will swiftly provide them with the corresponding RRID.

At the end of March 2018, ~2500 cell lines had been referenced using RRIDs in 870 articles from 86 journals. These numbers are expected to rise quickly, as an increasing number of publishers and journal editors are requesting that authors use RRIDs in their articles.

THE CELLOSAURUS IN WIKIDATA

Wikidata (https://www.wikidata.org) is a free and collaboratively edited knowledge base hosted by the Wikimedia Foundation. The life-science community is interested[63] in using this platform as a structured, semantic web-compatible integration hub for biologic and medical data. In this context, we initiated a project to enter a minimal set of information regarding all Cellosaurus cell lines in Wikidata. We have defined a number of "properties" relevant to cell lines (https://www.wikidata.org/wiki/Q27968522), including 1 to link back to the Cellosaurus using its accession numbers (Cellosaurus ID: https://www.wikidata.org/wiki/Property:P3289). We have seeded Wikidata with a number of example cell lines (https://goo.gl/yyGFL3). The next step is to write a "bot" (https://www.wikidata.org/wiki/Wikidata:Bots) to enter and update cell-line information in Wikidata.

INVITROMATICISTS SPOTLIGHT

Recently, 3 new terms—invitromatics, invitrome, and invitroomics—were introduced by Bols and coworkers.[64] At the same time, we were interested in providing to Cellosaurus users short biographical sketches concerning researchers that have played a major role in establishing 1 or more cell lines.

Niels Bols coined a fourth term, invitromaticists, for such scientists, and together with Lucy Lee, they agreed to write the first 2 installments. Since then, we have regularly added other invitromaticist profiles. These are all accessible from a page on the ExPASy Cellosaurus site (https://web.expasy.org/cellosaurus/invitromaticists/). We welcome new submissions and especially encourage former colleagues of deceased invitromaticists to contribute articles about their mentors.

FUTURE DEVELOPMENTS

There are many things that we plan to do in the coming years to increase the usefulness of the Cellosaurus to the scientific community. In terms of scope, we are planning to add patient-derived xenografts. In terms of the depth of information, we are considering addition of the tissue of origin of a cell line using Uberon[65] as the underlying anatomy ontology. We also want to add information regarding integrated viruses in cell lines, translocations in cancer cell lines, and the biosafety level of a cell line. In terms of format and tools, we would like to provide a resource description framework version of the Cellosaurus; this development would allow us to offer an advanced search tool based on the SPARQL technology. Last but not least, in the context of the issue of cell-line contamination, a tool to search and compare STR profiles is high on our list of priorities.

ACKNOWLEDGMENTS

The authors thank Alain Gateau for developing the Scala utility celloparser that performs error checking on the Cellosaurus and converts the flat-file format into OBO and XML, Elisabeth Gasteiger for implementing the Cellosaurus on ExPASy, Amanda Capes-Davis for many useful suggestions and for using her skills as a cell-line detective to untangle many cell-line mysteries, Monique Zahn for critical reading of the manuscript, and the hundreds of scientists that have provided us with information, corrections, and updates. The author declares no conflicts of interest.

REFERENCES

1. Earle WR, Schilling EL, Stark TH, Straus NP, Brown MF, Shelton E. Production of malignancy in vitro. IV. The mouse fibroblast cultures and changes seen in the living cells. *J Natl Cancer Inst* 1943;4:165–212.
2. Skloot R. 2010. *The Immortal Life of Henrietta Lacks*. Random House, New York.
3. Milstein C. The hybridoma revolution: an offshoot of basic research. *BioEssays* 1999;21:966–973.
4. Solter D. From teratocarcinomas to embryonic stem cells and beyond: a history of embryonic stem cell research. *Nat Rev Genet* 2006;7:319–327.
5. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 2006;126:663–676.
6. Baker M. Reproducibility crisis: blame it on the antibodies. *Nature* 2015;521:274–276.
7. Gartler SM. Apparent Hela cell contamination of human heteroploid cell lines. *Nature* 1968;217:750–751.
8. Nelson-Rees WA, Flandermeyer RR, Hawthorne PK. Banded marker chromosomes as indicators of intraspecies cellular contamination. *Science* 1974;184:1093–1096.
9. Lorsch JR, Collins FS, Lippincott-Schwartz J. Cell biology. Fixing problems with cell lines. *Science* 2014;346:1452–1453.
10. Masters JR, Thomson JA, Daly-Burns B, et al. Short tandem repeat profiling provides an international reference standard for human cell lines. *Proc Natl Acad Sci USA* 2001;98:8012–8017.
11. Kato S, Calnek BW, Powell PC, Witter RL. A proposed method for designating avian cell lines and transplantable tumours. *Avian Pathol* 1979;8:487–498.
12. Luong MX, Auerbach J, Crook JM, et al. A call for standardized naming and reporting of human ESC and iPSC lines. *Cell Stem Cell* 2011;8:357–359.
13. Kurtz A, Seltmann S, Bairoch A, et al. A standard nomenclature for referencing and authentication of pluripotent stem cells. *Stem Cell Reports* 2018;10:1–6.
14. Gaudet P, Michel PA, Zahn-Zabal M, et al. The neXtProt knowledgebase on human proteins: current status. *Nucleic Acids Res* 2015;43:D764–D770.
15. Hinard V, Britan A, Schaeffer M, et al. Annotation of functional impact of voltage-gated sodium channel mutations. *Hum Mutat* 2017;38:485–493.
16. Romano P, Manniello A, Aresu O, Armento M, Cesaro M, Parodi B. Cell Line Data Base: structure and recent improvements towards molecular authentication of human cell lines. *Nucleic Acids Res* 2009;37:D925–D932.
17. Sobel RE, Sadar MD. Cell lines used in prostate cancer research: a compendium of old and new lines–part 1. *J Urol* 2005;173:342–359.
18. Robinson J, Soormally AR, Hayhurst JD, Marsh SG. The IPD-IMGT/HLA Database - new developments in reporting HLA variation. *Hum Immunol* 2016;77:233–237.
19. Seltmann S, Lekschas F, Müller R, et al. hPSCreg–the human Pluripotent Stem Cell Registry. *Nucleic Acids Res* 2016;44: D757–D763.
20. Sarntivijai S, Lin Y, Xiang Z, et al. CLO: the Cell Line Ontology. *J Biomed Semantics* 2014;5:37.
21. Gremse M, Chang A, Schomburg I, et al. The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res* 2011;39: D507–D513.
22. Malone J, Holloway E, Adamusiak T, et al. Modeling sample variables with an experimental factor ontology. *Bioinformatics* 2010;26:1112–1118.
23. Ong E, Sarntivijai S, Jupp S, Parkinson H, He Y. Comparison, alignment, and synchronization of cell line information between CLO and EFO. *BMC Bioinformatics* 2017; 18(Suppl 17):557.
24. Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 2012;483:603–607.
25. Yang W, Soares J, Greninger P, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res* 2013;41: D955–D961.
26. Li J, Zhao W, Akbani R, et al. Characterization of human cancer cell lines by reverse-phase protein arrays. *Cancer Cell* 2017;31:225–239.
27. Streeter I, Harrison PW, Faulconbridge A, et al. The human-induced pluripotent stem cell initiative-data resources for cellular genetics. *Nucleic Acids Res* 2017;45:D691–D697.
28. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res* 2013; 41:D991–D995.
29. Kolesnikov N, Hastings E, Keays M, et al. ArrayExpress update–simplifying data submissions. *Nucleic Acids Res* 2015; 43:D1113–D1116.
30. Kale NS, Haug K, Conesa P, et al. MetaboLights: an open-access database repository for metabolomics data. *Curr Protoc Bioinformatics* 2016;53:14.13.1–14.13.18.

31. Vizcaíno JA, Csordas A, Del-Toro N, et al. 2016 Update of the PRIDE database and its related tools. *Nucleic Acids Res* 2016; 44:11033.
32. Stachelscheid H, Seltmann S, Lekschas F, et al. CellFinder: a cell data repository. *Nucleic Acids Res* 2014;42:D950–D958.
33. Reinhold WC, Sunshine M, Liu H, et al. CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. *Cancer Res* 2012;72:3499–3511.
34. Chisanga D, Keerthikumar S, Pathan M, et al. Colorectal cancer atlas: an integrative resource for genomic and proteomic annotations from colorectal cancer cell lines and tissues. *Nucleic Acids Res* 2016;44:D969–D974.
35. Shiau CK, Gu DL, Chen CF, Lin CH, Jou YS. IGRhCellID: integrated genomic resources of human cell lines for identification. *Nucleic Acids Res* 2011;39(Suppl 1):D520–D524.
36. Dirks WG, MacLeod RA, Nakamura Y, et al. Cell line cross-contamination initiative: an interactive reference database of STR profiles covering common cancer cell lines. *Int J Cancer* 2010;126:303–304.
37. Knutsen T, Gobu V, Knaus R, et al. The interactive online SKY/M-FISH & CGH database and the Entrez cancer chromosomes search database: linkage of chromosomal aberrations with the genome sequence. *Genes Chromosomes Cancer* 2005;44:52–64.
38. Forbes SA, Beare D, Boutselakis H, et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res* 2017;45: D777–D783.
39. Gaulton A, Hersey A, Nowotka M, et al. The ChEMBL database in 2017. *Nucleic Acids Res* 2017;45:D945–D954.
40. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2016;44:D7–D19.
41. Faulconbridge A, Burdett T, Brandizi M, et al. Updates to BioSamples database at European Bioinformatics Institute. *Nucleic Acids Res* 2014;42:D50–D52.
42. Vasilevsky N, Johnson T, Corday K, et al. Research resources: curating the new eagle-i discovery system. *Database* 2012;2012: Bar067.
43. Davis CA, Hitz BC, Sloan CA, et al. The Encyclopedia of DNA Elements (ENCODE): data portal update. *Nucleic Acids Res* 2018;46:D794–D801.
44. Koleti A, Terryn R, Stathias V, et al. Data portal for the Library of Integrated Network-Based Cellular Signatures (LINCS) program: integrated access to diverse large-scale cellular perturbation response data. *Nucleic Acids Res* 2018;46:D558–D566.
45. Dekker J, Belmont AS, Guttman M, et al. The 4D Nucleome project. [Corrigendum: Nature 2017;552:278.] *Nature* 2017; 549:219–226.
46. Sayers EW, Barrett T, Benson DA, et al. Database resources of the National Center for Biotechnology Information. [Erratum: Nucleic Acids Res 2009;37:3124.] *Nucleic Acids Res* 2009;37:D5–D15.
47. Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu WL, Wright LW. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *J Biomed Inform* 2007;40:30–43.
48. Marygold SJ, Crosby MA, Goodman JL; ; FlyBase Consortium. Using FlyBase, a database of Drosophila genes and genomes. *Methods Mol Biol* 2016;1478:1–31.
49. Gray KA, Yates B, Seal RL, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res* 2015;43:D1079–D1085.
50. Eppig JT, Richardson JE, Kadin JA, Ringwald M, Blake JA, Bult CJ. Mouse Genome Informatics (MGI): reflecting on 25 years. *Mamm Genome* 2015;26:272–284.
51. Shimoyama M, De Pons J, Hayman GT, et al. The Rat Genome Database 2015: genomic, phenotypic and environmental variations and disease. *Nucleic Acids Res* 2015;43: D743–D750.
52. UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2018;46:2699.
53. Xavier CP, Pesic M, Vasconcelos MH. Understanding cancer drug resistance by developing and studying resistant cell line models. *Curr Cancer Drug Targets* 2016;16:226–237.
54. Hastings J, Owen G, Dekker A, et al. ChEBI in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Res* 2016;44:D1214–D1219.
55. Law V, Knox C, Djoumbou Y, et al. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res* 2014;42: D1091–D1097.
56. Saragusty J, Diecke S, Drukker M, et al. Rewinding the process of mammalian extinction. *Zoo Biol* 2016;35:280–292.
57. American Type Culture Collection Standards Development Organization Workgroup ASN-0002. Cell line misidentification: the beginning of the end. *Nat Rev Cancer* 2010;10: 441–448.
58. Almeida JL, Cole KD, Plant AL. Standards for cell line authentication and beyond. *PLoS Biol* 2016;14:e1002476.
59. O'Donoghue LE, Rivest JP, Duval DL. Polymerase chain reaction-based species verification and microsatellite analysis for canine cell line validation. *J Vet Diagn Invest* 2011;23:780–785.
60. Capes-Davis A, Reid YA, Kline MC, et al. Match criteria for human cell line authentication: where do we draw the line? *Int J Cancer* 2013;132:2510–2519.
61. Artimo P, Jonnalagedda M, Arnold K, et al. ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res* 2012;40: W597–W603.
62. Bandrowski A, Brush M, Grethe JS, et al. The resource identification initiative: a cultural shift in publishing. *F1000 Res* 2015;4:134.
63. Burgstaller-Muehlbacher S, Waagmeester A, Mitraka E, et al. Wikidata as a semantic framework for the Gene Wiki initiative. *Database* 2016;2016:baw015.
64. Bols NC, Pham PH, Dayeh VR, Lee LE. Invitromatics, invitrome, and invitroomics: introduction of three new terms for in vitro biology and illustration of their use with the cell lines from rainbow trout. *In Vitro Cell Dev Biol Anim* 2017;53: 383–405.
65. Haendel MA, Balhoff JP, Bastian FB, et al. Unification of multi-species vertebrate anatomy ontologies for comparative biology in Uberon. *J Biomed Semantics* 2014;5:21.