ICBO: International Conference on Biomedical Ontology
July 28-30, 2011 · Buffalo, NY, USA

Cell Line Ontology: Redesigning the Cell Line Knowledgebase to Aid Integrative Translational Informatics

Sirarat Sarntivijai[1], Zuoshuang Xiang[1], Terrence F. Meehan[2], Alexander D. Diehl[3],Uma Vempati[4], Stephan Schurer[4], Chao Pang[5], James Malone[5], Helen Parkinson[5], Brian D. Athey[1], Yongqun He[1]

[1]University of Michigan Medical School, Ann Arbor, MI, USA
[2]Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, ME, USA
[3]University at Buffalo School of Medicine and Biomedical Sciences, Buffalo, NY, USA
[4]Miller School of Medicine, University of Miami, FL, USA
[5]EMBL-EBI European Bioinformatics Institute, Hinxton, UK

Abstract. The Cell Line Ontology (CLO) is a community-based ontology in the domain of biological cell lines with a focus on permanent cell lines from culture collections. Upper ontology structures that frame the skeleton of CLO include the Basic Formal Ontology and Relation Ontology. Cell lines contained in CLO are associated with terms from other ontologies such as Cell Type Ontology, NCBI Taxonomy, and Ontology for Biomedical Investigation. A common design pattern for the cell line is used to model cell lines and their attributes, with the Jurkat cell line as an example. Currently CLO contains over 36,000 cell line entries obtained from ATCC, HyperCLDB, Coriell, and by manual curation. The cell lines are derived from 194 cell types, 656 anatomical entries, and 217 organisms. The OWL-based CLO is machine-readable and can be used in various applications.

Keywords: Cell line, Cell Line Ontology, CLO

1 Introduction

A cell line is a colony of cells that is artificially developed and grown under controlled conditions. A cell line typically derives from a multicellular eukaryote. A cell line may derive from a 'normal' or modified/disease tissue. A cell line can be maintained as a stable *permanent* cell lineage for renewable usages, or it may be used as a *primary* cell lineage without a long-term maintenance.

Cell lines have been widely used in research. Information about cell lines is stored in public repositories and/or indexed catalogues available for open access, and cell lines are commercially available or they are transferred between academic laboratories. Information about cell lines has not been well standardized and machine-readable to date. Each commercial provider generates a catalogue, and academic cell lines are not necessarily included. Integration of data from multiple sources is confounded by: lack of consistent naming conventions for cell lines across providers, contamination of cell lines as they are passaged and transferred between laboratories, and provision of the same cell lines by multiple commercial sources but with different biological attributes. To address these issues, we previously produced a normalized catalogue of the Cell Line Knowledgebase (CLKB; http://clkb.ncibi.org/) as a project in the National Center for Integrative Biomedical Informatics (NCIBI) [1]. Since the release of CLKB, biomedical research has rapidly evolved toward integrative translational bioinformatics. In order to support translational research, conform to OBO foundry standards, and produce a resource that can be used in queries and data integration we have transformed the CLKB into an ontology available in OWL format (http://www.w3.org/TR/owl-guide/). Here we present the design patterns, design methodology, and content of the Cell Line Ontology – CLO.

When the Cell Type Ontology (CL) was first introduced to represent *in vivo* cell types [2], primary and permanent cell lines were included in the ontology and no separate cell line ontology existed. The Cell Type Ontology no longer includes primary or permanent cell

lines as the CLO has now become the source ontology for permanent cell lines as agreed by the maintainers of the CL, the Ontology for Biomedical Investigations (OBI), and OBO Foundry. The top-level terminology required for generating a primary cell line is provided by the OBI. The CLO is therefore a collaborative development between the CL, OBI and the CLKB developers at NCIBI and references terms from these and other ontologies in the definitions and modeling of cell lines.

In addition to CLO design and methodology, we also include examples and applications of the CLO in this study.

2 Method

2.1 Cell Line Data Sources for CLO Development

The CLO uses data from multiple sources, which are described in Table 1. The CLO cell line data were first drawn from CLKB entries, which consist of 8740 cell lines stored in ATCC (http://www.atcc.org/) and HyperCLDB (http://bioinformatics. istge.it/cldb/). CLKB will be kept as a backup source but will become obsolete at the release of the new CLO. Additional 27,000 permanent cell lines are obtained from European Bioinformatics Institute Coriell Catalogue Ontology that models cell lines from the Coriell cell repository (http://ccr.coriell.org/), and cell lines (both primary and permanent) provided by the Bioassay Ontology (BAO; http://bioassay ontology.org/) development team. Cell lines that are listed in multiple repositories contain cross-reference pointers to these repositories. Cell line names can be misleading. Similar or synonymous names do not guarantee identical cell lines. Automatic mapping and manual annotation have been combined to ensure correct cell line annotation in CLO.

2.2 Importing External Ontology Terms by OntoFox

CLO imports the whole Basic Formal Ontology (BFO) [3] as its upper level ontology and the Relation Ontology (RO) [4] as its core relations. The use of these ontologies promotes integration as these resources are used by many biomedical ontologies. We used OntoFox [5] - a technology for merging ontologies to integrate external ontologies such as NCBI_Taxon and Cell Type Ontology into the CLO. All namespaces are preserved for these ontology terms.

2.3 Definition and Annotation of CLO-Specific Ontology Terms

All cell lines and cell line-specific terms are given unified CLO IDs. The cell line data from the Coriell Cell Line ontology (http://bioportal. bioontology.org/ontologies/45331) have been merged to CLO with newly assigned CLO IDs. The BioAssay Ontology (BAO) has also provided a list of cell lines for inclusion in CLO. In these two cases a namespace is not preserved. When a cell line term is imported from the Coriell Cell Line ontology, we have provided a cross reference to the ontology using the *seeAlso* annotation property. Using the annotation property *comment*, BAO is noted as the source for those cell lines coming from BAO. A cell line design pattern is developed to make generic pattern between CLO cell lines and other ontology terms.

2.4 CLO Editing and Access

The development of CLO follows the OBO Foundry principles [6]. Specifically, we use unique IDs, and provide text definition for each cell line. The Web Ontology Language (OWL) is used as the default CLO format. CLO is edited using Protégé 4 Ontology Editor (http://protege.stanford.edu). The latest CLO is available for public view and download at http://sourceforge.net/projects/clo-ontology/. The latest version of CLO is also available for visualization and download from NCBO BioPortal: http://purl.bioontology.org/ontology/CLO.

3 Results

3.1 CLO Top Structure and Statistics

The key top level classes in CLO are shown in Fig. 1. To support data integration and automated reasoning, CLO imports many terms from existing ontologies as upper level terms (e.g., *material_entity* from BFO) or terms needed for association (e.g., *cell* in CL). Cell line-specific terms are assigned with CLO

IDs (Fig. 1). The CLO-specific class *cell line* is the parent class for all specific cell lines in the CLO. The classes *permanent cell line* and *primary cell line* are the major differentia based on culture for cell lines in the CLO at present. The majority of cell lines in the CLO are permanent cell lines. Normalized cell line entries are entered as asserted CLO classes under these two subclasses. A cell line can be cultured or modified, and supplied or managed by a cell line repository (e.g., ATCC) (Fig. 1). The detailed relations among these terms are described in our cell line design pattern (Fig. 2).

Currently CLO contains 8797 cell line-specific terms with unique CLO identifiers. In total, CLO contains 38172 terms (Table 1). The Coriell cell line records were integrated and assigned CLO-specific identifiers.

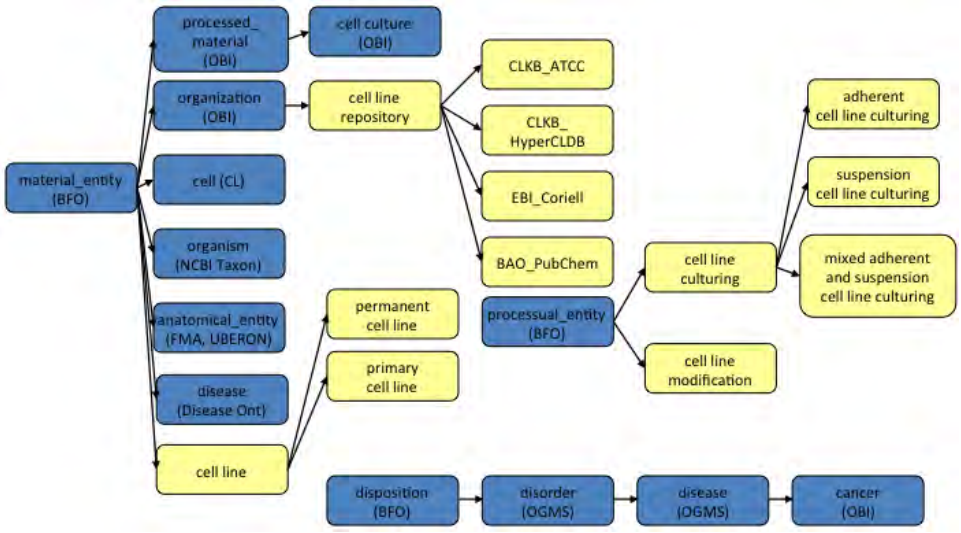

Figure 1. The top level CLO hierarchical structure of ontology terms. The terms in light blue boxes are imported from existing ontologies. The terms shown in light yellow boxes are terms with CLO unique IDs.

Ontology	Classes	Object Properties	Datatype Properties	*Total*
CLO (Cell Line Ontology) specific	36879	14	0	36893
Imported full ontologies				
BFO (Basic Formal Ontology)	39	0	0	39
RO (Relation Ontology)	6	25	0	31
IAO (Information Artifact Ontology)	102	14	5	121
Imported terms from other external ontologies				
OBI (Ontology for Biomedical Investigation)	15	6	0	21
CL (Cell Type Ontology)	194	0	0	194
UBERON	622	34	0	656
NCBITaxon (NCBI Taxonomy)	217	0	0	217
Total	38074	93	5	38172

Table 1. Summary of ontology terms in CLO and source ontologies used in CLO.

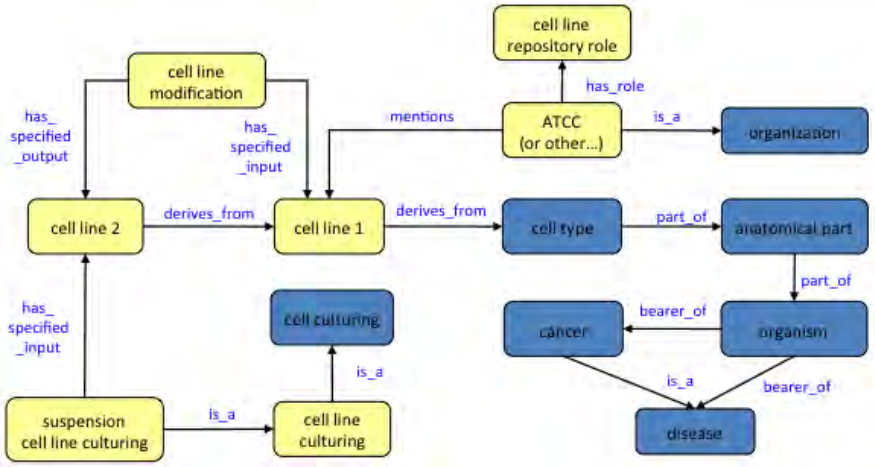

Figure 2. Basic design pattern for representing cell lines in CLO. Components shown in yellow boxes are specific to CLO, while those in blue boxes signify classes imported from other ontologies. Depending on the cell line being described, *suspension cell line culturing* and *ATCC* can be replaced with another cell line culturing process and another cell line repository, respectively.

3.2 The CLO Cell Line Design Pattern

The CLO design pattern supports representation of anatomy, cell types, disease and pathology, source information in the form of ownership and derivation where cell lines are related, and technical information such as culture conditions. Fig. 2 depicts the design pattern developed to model this information retrieved from data sources. Briefly, a cell line is originally derived from a cell type that is part of an anatomical part (e.g., liver) of a specific organism (e.g., human) having a cancer (e.g., lymphoma) or some other disease. A cell line can be derived from another cell line through a particular cell line modification. A cell line is cultured differently (e.g., *suspension cell line culturing*), which reflects a particular culturing condition or growth mode (e.g., suspension). A cell line is supplied, owned, or managed by a specific organization such as *ATCC* that has a *cell line repository role*. Since relation terms such as *supply*, *own*, or *manage* do not exist in any ontology, we use a similar relation, *mentions* (Fig. 2).

The basic cell line design pattern is followed in our CLO development. In many cases, we have also extended this design pattern by adding more content. For example, we may add a sex (female or male) quality to the organism. The pathology of the cell lines in Coriell represents one of the most important aspects to users of these artifacts and many are used as models for a particular disease. A cell line may derive from an organism that has a specific disease (Fig. 2). The *is_model_for* relation is used to link a disease to a cell line. This relation has been created as a shortcut relation to represent the association between a cell line and a disease. In the case that a cell line derives from a normal tissue, the information of disease is omitted.

A deep understanding of the cell line design pattern that portrays a true composite architecture of cell lines requires more discussion and explanation on the relationship between CLO, CL and NCBI Taxonomy. More information is provided below.

A cell line is derived from a cell type (Fig. 2). The CL developers have been working with the CLO to ensure adequate representation of cell types from which cell lines originate. This allows mappings between the CLO and the CL using the *derives_from* relationship. Such integration also promotes error detection. To enhance interoperability with other OBO Foundry ontologies, CL-CLO mapping associates cell-types with anatomical structures using the species-neutral UBERON ontology. Thus, mappings between CLO and CL allow for associations from cell lines to anatomical structures. Sometimes a cell line cannot be mapped directly to CL as the cell line may contain multiple cell types, which can be a case of anatomical part + cell type (e.g., HCC cell line is annotated as having tissue type '*mammary gland, epithelial*'), or cell type + pathological description (e.g., AtT-20 cell line is annotated as having tissue type *pituitary tumor, small, rounded*), or multiple cell types

(e.g., p53NiS1 cell line has annotated tissue type *fibrous histocytoma, fibroblast*). In this case, this cell line is related to all associated cell types using the same *derives_from* relation. According to the original repository, a cell line may derive from a cell type named by its associated anatomical part (e.g., *liver cell, peripheral blood*). These anatomy associated cell type terms have been added to CL as new CL terms to support this design. In total, 194 CL terms and 656 UBERON terms are imported to CLO (Table 1) and CLO development has expanded the CL.

The NCBI Taxonomy is the source ontology for CLO to import organism information associated with individual cell lines (Table 1). A cell line may be listed as a hybrid from multiple organisms and therefore organism and not species is modeled. In this situation, the cell line will be linked to multiple organisms. One exception of this mapping occurs when a cell line is recorded as being part of mouse/rat hybrid as there exists a class named *Mus musculus x Rattus norvegicus* as a special class of the taxonomy. Investigation of the NCBI Taxonomy also reveals that a few classes relating to those of a cell line have a place holder within NCBI Taxon, such as *mouse/rat hybrid cell lines being* classified under parent term *unclassified Muridae*. However, since the primary purpose of importing NCBI Taxon terms to CLO is to use the information to define organism classes, and not to redefine cell lines, we do not import these terms to CLO. A few organism values that could not be mapped to NCBI_Taxon appeared to be the result of typographical errors or spelling variants. For example, there are a few cell line entries with annotated organism 'Agrothis segetum', which is believed to be a spelling variation of 'Agrotis segetum' (NCBITaxon: 47767). We do not omit or modify these original values, keeping 'Agrothis segetum' as obtained from the source (e.g., ATCC), and putting a remark in the cell line class' comment with the information pointing to NCBITaxon: 47767.

3.3 Describing Cell Line with CLO: The Example of Jurkat

We have modeled the Jurkat Clone E6-1 cell line (ATCC # TIB-152) and its derived cell line J.Cam1.6 (ATCC #CRL-2063) as a demonstration of our cell line design pattern usage (Fig. 3). Jurkat Clone E6-1 is a clone of an immortalized line (Jurkat) of T lymphocyte cells that was established in the late 1970s from the peripheral blood of a 14 year old boy with T cell leukemia ([7]). The J.CaM1.6 cell line is a derivative mutant of Jurkat E6-1 by treatment with ethylmethanesulfonate (EMS). J.CaM1.6 cells are deficient in Lck kinase activity and miss exon 7 in their lck mRNA.

It is noted that J.CaM1.6 cell line is not a child term of Jurkat Clone E6-1 cell line in CLO. Cell lines derived from one base cell line (e.g., J.CaM1.6 cell line deriving from Jurkat Clone E6-1) is by definition not an *is_a* relation to the base cell line but rather a *derives_from* relation. Based on this *derives_from* relation, we generate a term *Jurkat derivative cell line*, and J.CaM1.6 cell line can be inferred to be a *Jurkat derivative cell line*.

The sharing of tissue, tumor, and organism can be used to group different cell lines, such as Jurkat and Jurkat Clone E6-1. The original value *'peripheral blood'* obtained from source is mapped to anatomical term *'blood'* that best fits this term mapping as there are no such terms in FMA or UBERON that describe peripheral blood. Furthermore, a cell line deriving from T cell such as Jurkat is potentially problematic as T cells scatter throughout the body. Not all T cells are *part_of* some blood. Jurkat was extracted from an instance of lymphoma that was in blood. But CL's definition of lymphocytes does not restrict all lymphocytes to blood tissue. This information is however described by *'isolation'* (an OBI term used to associate tissue and cell type) inside CL. A cell line's description embedded in CLO is specific to each individual cell line being conceptualized in each class. Reasoning with knowledge obtained from CLO and CL can capture this issue of specificity.

Many CLO specific terms (e.g., *peripheral blood cell line*) have been generated. A reasoner can be used to infer what cell lines belong to such CLO terms. Such terms are needed for many applications. For example, the ArrayExpress staff needs to know all the blood-derived cell lines and cell types for a meta-analysis of gene expression data on blood. Without such defined classes, it is difficult to obtain the results.

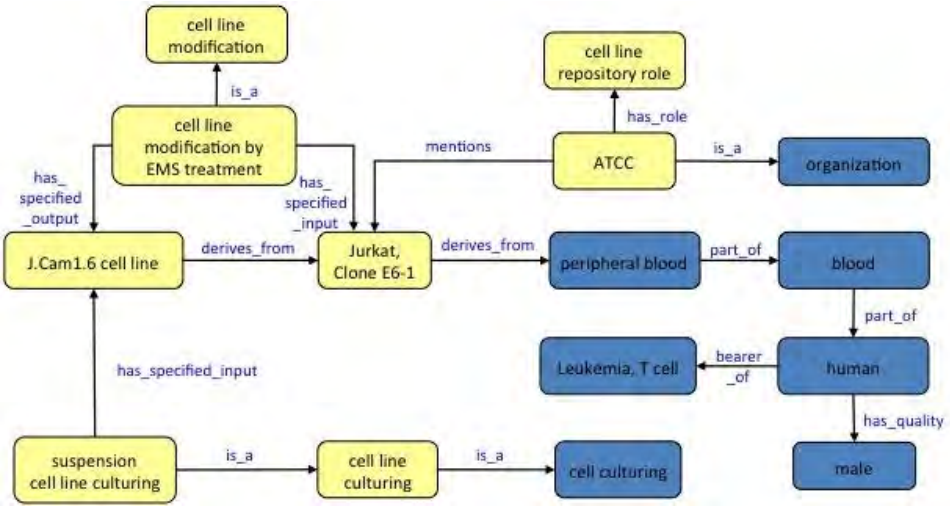

Figure 3. Modeling Jurkat and its derivative Jurkat Clone J.Cam1.6 using CLO.

3.4 CLO Application Use Case: Application of CLO in Bioassay Data Analysis

The Bioassay Ontology (BAO – http://bioassay ontology.org/) describes bioassays and results obtained from small molecule perturbations, such as those in the PubChem database [8]. Integrating a formal representation of cell lines will benefit researchers in interpreting and analyzing cell-based screening results. It will also enable linking PubChem assays to other types of information (such as diseases and pathways). Moreover, formally described cell lines can help researchers in the design of novel assays, for example with respect to choosing the best cellular model system, and also in identifying which modified cell lines are available and which ones work best in existing assays.

To describe and annotate cell-based PubChem assays and screening results comprehensively, BAO is being extended through collaborative development of the CLO. By integrating BAO with CLO, those cell lines that are typically used in cellular assays are added into CLO. Based on the demands of BAO bioassay modeling, extended parameters are being added to CLO, including different sources of cell lines (normal/healthy tissue, pathological tissue, or tumor), cell modification methods (plasmid transfection, viral transduction, cell fusion, *etc.*), culture condition (composition of culture medium), morphology (epithelial, lymphoblast, *etc.*), growth properties (adherent or suspension), short tandem repeat (STR) profiling and other properties that are relevant for cellular screening.

As a demonstration of the use of CLO in BAO bioassay modeling, we have modeled the HeLa cell line in the context of a PubChem assay (AID 1611) (Fig. 4). HeLa is an immortal cell line established from cervical adenocarcinoma of a patient in 1951 [9] and available from the ATCC (catalog # CCl-2). In the PubChem assay, HeLa cells were modified by stable transfection with a heat shock promoter driven-luciferase reporter gene construct. In this assay, the modified HeLa cells were used to screen for compounds that could induce heat shock transcriptional response as a potential therapeutic for Huntington's disease and amytrophic lateral sclerosis (ALS).

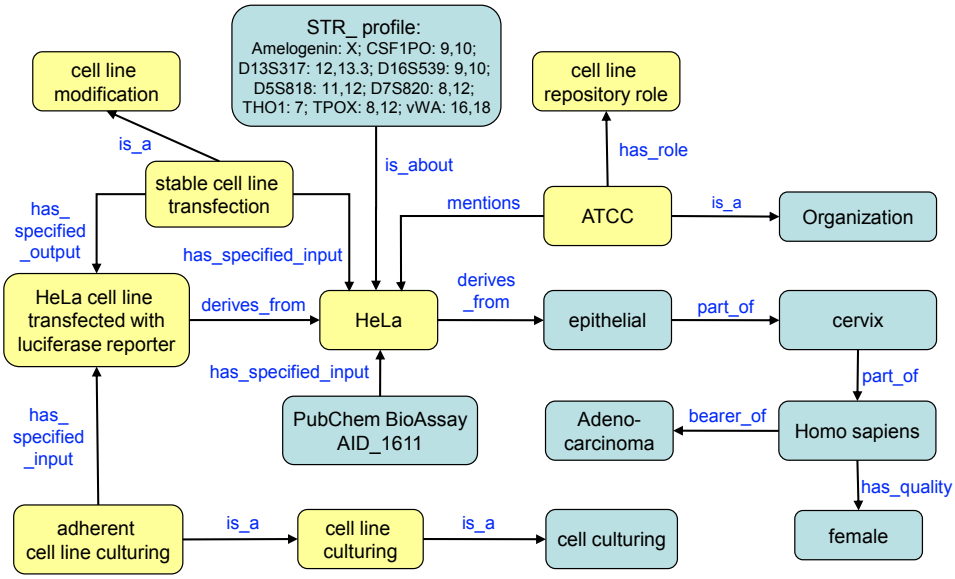

Figure 4. Application of CLO in bioassay data integration and analysis.

HeLa as demonstrated in figure 4 can be described as a *permanent cell line*, which is 'part_of' cervix and is 'derived_from' *Homo sapiens* that is 'bearer_of' cervical carcinoma. HeLa is an *epithelial cell of cervix* (CL:0002535), whose *growth mode* is *adherent*. Describing the other details of the assay are out of the scope of this paper, as they require concepts from BAO.

Many other CLO applications are being studied. For example, cell line knowledge can be used for microarray data analysis. A separate paper has been submitted to the International Conference on Biomedical Ontology (http://icbo.buffao.edu/) that provides more details of how the Coriell Cell Line Ontology, which has been merged to CLO, is used for ArrayExpress microarray data analysis.

4 Discussion

The availability of cellular assays and the ability to sequence DNA and RNA from single cells has promoted the use of cell lines in research and highlighted the role of cell lines in biomedical research. The release of CLO is therefore timely and will support many applications in biomedical informatics. First, CLO can be used as a tool to facilitate the data entry process for public cell line repositories (e.g., ATCC) and the referencing of these by resources such as archival repositories (e.g., ArrayExpress) Cross-referencing with other source ontologies that are imported to CLO will allow a standard controlled vocabulary to be utilized at the data-entry point to avoid typographic errors and aid better annotation, while the depositor can also verify if the cell line being deposited already exists in the ontology, thus eliminating redundant data. Although there are currently no central authorities to assess cell line nomenclature and a cell line name is often assigned by the lab of origin, utilizing the CLO structure to frame the process will help reduce the use of duplicate names for different cell lines. It is our plan to solicit directions of new cell lines to CLO through a community-based agreement. This is also a crucial step to achieve an efficient cell line authentication process.

Furthermore, information stored in CLO can potentially validate other existing cell culture information in various sources. Gene expression data that contain the information of cell line used can be analyzed to observe if there is any data inconsistency when compared back to the information received in CLO based on the same cell line. Inconsistency in the record's attributes such as organism, tissue, tumor, or genetic mutations based on the cell line's modification may signify the possibility of cross contamination.

Cell line contamination occurs easily. It is

reported that 15% of the times cell lines being used are not what they are assumed to be [10]. Contamination also leads to issue of misidentification and mislabelling. To address this issue of contamination and mislabelling and improve cell line authentication, the American Type Cell Culture: Standards Development Organization (ATCC SDO) has proposed to establish a community-supported central authority and to use short tandem repeats (STR) as one method of verification. As a normalized indexed catalogue with ontological structure and semantics, the CLO will play a critical role in standardizing and representing cell lines and properly addressing the issue of cell line contamination. CLO can also be further expanded to link out to this STR verification information of each cell line. CLO is currently being studied for use in the ATCC SDO's authentication process.

Normalized cell line data and additional features in CLO also support applications in translational informatics such as cell line-disease association analysis, annotations of complex organ/tissue in cell cultures, and combined studies of cell culture and bioassay data.

The creation of international BioSamples databases at the EBI (http://www.ebi.ac.uk/biosamples) and NCBI (http://www.ncbi.nlm.nih.gov/biosample) provides a strong use case in that storage of non-standardized data on thousands of cell lines is not useful for high level query purposes and queries such as 'retrieve data on all ENCODE cell lines' or 'all Drosophila cell lines' will be facilitated by the addition of defined classes to the CLO, and the submission process to archival repositories will be easier if the users are able to query using the CLO ontology to retrieve validated cell line information instead of providing all this information again.

Future work of the CLO development includes the insertion of more cell lines and cell line-associated attributes. Additional CLO applications are under investigation.

Acknowledgments

We acknowledge and appreciate the following support: NIH grants 1R01AI081062, U54-DA-021519 for the National Center for Integrative Biomedical Informatics (NCIBI), NHGRI ARRA Administrative grant HG002273-09Z (CL), RC2 HG005668 (BAO), and Gen2Phen EMBL contract number 200754 (EBI).

References

1. Sarntivijai, S., Ade, A.S., Athey, B.D., States, D.J.: A Bioinformatics analysis of the cell line nomenclature. J. Bioinformatics 24(23), 2760--2766 (2008)
2. Bard, J., Rhee, S.Y., Ashburner, M.: An ontology for cell types. Genome Biol. 6, R21 (2005)
3. Arp, R., Smith, B. Function, Role, and Disposition in Basic Formal Ontology. Nat. Preced. hdl:10101/npre.2008.1941.1 (2008)
4. Smith, B., Ceusters, W., Klagges, B., Kohler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A.L., Rosse, C.: Relations in biomedical ontologies. Genome Biol. 6(5): R46 (2005)
5. Xiang, Z., Courtot, M., Brinkman, R.R., Ruttenberg, A., He, Y.: OntoFox: web-based support for ontology reuse. BMC Res. Notes 22(3), 175 (2010)
6. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J.; OBI Consortium, Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuemann, R.H., Shah, N., Whetzel, P.L, Lewis, S.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol. 25(11), 1251-1255 (2007)
7. Weiss, A., Wiskocil, R.L., Stobo, J.D.: The role of T3 surface molecules in the activation of human T cells; a two-stimulus requirement for IL2 productions reflects events occurring at a pre-translational level. J. Immunol. 133(1), 123-128 (1984)
8. Schürer, S.C., Vempati, U., Smith, R., Southern, M., and Lemmon, V.: BioAssay Ontology Annotations Facilitate Cross-Analysis of Diverse High-Throughput Screening Data Sets. J Biomol Screen 16, 415-426 (2011)
9. Scherer, W.F., Syverton, J.T., Gey, G.O.: Studies on propagations in vitro of poliomyelitis viruses. IV. Viral multiplication in a stable strain of human malignant epithelial cells (strain HeLa) derived from an epidermoid carcinoma of the cervix. J. Exp. Med. 97(5), 695-710 (1953)
10. Drexter, H.G., Quentmeier, H., Dirks, W.G., Uphoff, C.C., MacLeod, R.A.: DNA profiling and cytogenetic analysis of cell line WSU-CLL reveal cross-contamination with cell line REH (pre B-ALL). Leukemia 16(9), 1868-1870 (2002)