The Cell Line Integrated Molecular Authentication (CLIMA) database

Romano P[(1)], Aresu O[(2)], Manniello MA[(2)], Cesaro M[(2)], Parodi B[(2)]

[(1)] Bioinformatics, National Cancer Research Institute, Genova - Italy
[(2)] Cell Bank, National Cancer Research Institute, Genova - Italy

Motivation

Cross-contamination of human and animal cell lines is a repeated and frequent cause of scientific misrepresentation. Assumption that results obtained with the same cell lines in different laboratories are fully comparable is often not true. Molecular methods for cell line authentication, such as fingerprinting, STR profile and SNPs analysis, have been proposed in order to overcome this problem. A reference standard for STR profiles for human cell lines was recently proposed by authoritative cell banks and research institutes, who tested 253 human cell lines. The American Type Culture Collection (ATCC) and the Japanese Collection of Research Bioresources (JCRB) published STR profiles of the cell lines of their catalogues. STRdb, a database on main kits for STR profiling and related loci was also built. Now, the challenge is to build a public reference system able to link real cell lines maintained by different banks with their authoritative molecular characterizations. In this paper, we present the Cell Line Integrated Molecular Authentication database (CLIMA), that allows to link available authentication data to real cell lines distributed by cell banks.

Methods

CLIMA was designed with the aim of representing validated molecular authentications of human cell lines independently from the platforms (laboratory kits and/or sets of STR loci) used. It therefore includes STR profiling obtained by using different platforms and end users can retrieve data on cell line authentications performed by different cell banks. The database includes data tables, where actual authentication data is stored, and metadata tables, where information on platforms and datasets are stored. Data tables consist in a general table, where information on all cell line names for which a molecular characterization exists are stored, and in one further table for each dataset including related loci values. Each dataset may have a distinct set of loci and may include different additional information: it thus has a unique data structure. We set up one distinct table for each dataset also because this simplifies data updates. Metadata tables include tables for the description of datasets, kits and related loci, and bibliographic information. It is used by applications for building query forms and carrying out searches. Two platforms, sharing a limited number of loci, have already been included. The first is SGM, a silica-gel-based purification kit from Qiagen including: human tyrosine hydroxylase (TH01, 11p15.5), human von Willebrand factor (vWF, 12p-12pter), D8S1179 (8), D21S11 (21-q11.2–21q21), human a Fibrinogen (FGA, 4q28) and D18S51 (18q21.3), human amelogenin (Xp22.1–22.3 and Yp11.2). The second is the Promega PowerPlex® 1.2 system, including the amelogenin gene and eight STR loci: D16S539 (16q24–qter), D7S820 (7q11.21–22), D13S317 (13q22–q31), D5S818 (5q23.3–32), human c-fms proto-oncogene for CSF-1 receptor gene (CSF1PO, 5q33.3–34), human thyroid peroxidase gene (TPOX, 2p24–2pter), human tyrosine hydroxylase gene (TH01, 11p15.5), human von Willebrand factor gene (vWF, 12p12–pter). For both platforms two datasets are available. Public datasets were downloaded from ATCC and JBRC. Data from IST Cell Bank was already available locally. Profiles of cell lines checked in the development of the reference standard for STR profiles for human cell lines was manually inserted. Metadata information was limited to the description of above datasets and platforms.

Results

CLIMA includes information on 1,294 cell lines names and for 1,737 authentications. Four datasets are available. The first contains profiles of the standard and includes data on 223 cell lines. The second dataset contains profiles made available by ATCC on 670 cell lines. The third includes profiles made available by JBRC on 828 cell lines. The fourth dataset contains profiles obtained at IST Cell Bank on 16 cell lines. CLIMA can be searched on-line by name and by locus value. Search by name is case insensitive and retrieve all STR data that refers to cell lines whose name matches the query according to one of the following criteria: exact match, truncation, soundex (match “by sound”), and punctuation removal (only numbers and letters are used). Search by locus retrieves all authentications matching searched values: up to four values per

locus can be searched, but Amelogenin that only has two distinct values, and then combined by AND and OR. Both searches return, for each dataset, a table including profiles and cell line catalogue codes. Links to the site from where information was taken and to cell line description in the catalogue is also provided, when available. We plan to add further platforms and datasets and to build a list of misidentified (and possibly misidentified) cell lines. Authentication of non-human cell lines could also be performed by STR profiles from different species. Access to CLIMA is provided through the following URL:
http://bioinformatics.istge.it/clima/ .

Contact : paolo.romano@istge.it