

# Recent developments of the Cell Line Integrated Molecular Authentication Database

Paolo Romano<sup>1</sup>, Paola Visconti<sup>2</sup>, Barbara Parodi<sup>2</sup>

<sup>1</sup> Bioinformatics Lab, IRCCS AOU San Martino IST, Genoa, Italy

<sup>2</sup> Biological Resources Center, IRCCS AOU San Martino IST, Genoa, Italy

## Motivation

Cross-contamination of human and animal cell lines is a frequent event (1). For this reason, the results obtained with the same cell lines by different research groups are often not fully comparable (2,3), this leading to main reproducibility issues. The short tandem repeat (STR) profile has been proposed as a molecular method for cell line authentication (4). STR profile standard data sets for human cell lines were proposed by some of the leading cell banks worldwide which also have made the results of STR profiling of their cell lines available on-line.

We have built the Cell Line Integrated Molecular Authentication Database (CLIMA) (5) as a reference portal where authentication data are made available to the scientific community. This prototype system, although already largely utilized by researchers from all over the world, presented some limitations and only included a limited amount of STR profiles. Here, we present its most recent developments and discuss about some current challenges towards a better identification system for human cell lines.

## Methods

CLIMA is the result of an integration effort aimed at including all certified STR profiles of human cell lines in a unique database whose schema was revised for optimization of query performances. In the database schema, the following main entities are now taken into account: STR loci, STR profiling kits provided by companies (described on the basis of the loci they consider), data sets (in a one-to-many relation with cell banks, and with a specification of the adopted kit), STR profiles (which are linked to a unique cell line in a given cell bank, using a given kit). Due to the format heterogeneity of available profiles, these are converted according to an automatic procedure before insertion. Updates are implemented as integral substitution (delete followed by insert) of data sets.

While the searches by cell line name and by locus value were implemented as simple, and effective, relational queries, the feature related to the identification of cell lines was implemented according to the standard ANSI/ATCC ASN-0002-2011 (6), which makes reference to an STR profiling kit including the following genomic loci: D5S818, D13S317, D7S820, D16S539, VWA, TH01, Amelogenin, TPOX, CSF1PO. For the identification purposes, the number of matches, i.e. of identical values for the same locus, is computed for the entered profile (query) and each profile in the database (target). For each pair of profiles, the Standard Percent Match (StPM), that is the ratio between the number of matches and the number of distinct values in the target profile in the database, is computed. Identification is then achieved for cell lines having an StPM  $\geq 0.80$ . The system was built in a LAMP (Linux, Apache, MySQL, PHP) environment and is available on-line to all interested researchers.

## Results


The new version of CLIMA presents two main additions: more cell banks and profiles are included and a new identification tool is available.

STR profiles from human cell lines collected by ATCC (USA), JCRB (Japan), ICLC (Italy), DSMZ (Germany), and GOG (USA) were gathered and integrated in the database schema along with data included in the paper from Masters et al. (4). Currently, 5,450 STR profiles, representing 4,354 distinct cell line names, are included in the database.

Using the search engine is straightforward and results are displayed in a simple format. CLIMA 2 can be used to identify cell lines and to query the database either by name or by locus. A summary of queries and results is returned to the user by email when a valid address is provided.

The identification tool in CLIMA allows users to submit the STR profile of a cell line they want to identify. This profile is checked against all profiles in CLIMA that were defined by using a kit having the same loci. All profiles having an StPM greater than or equal to a desired value are returned, ranked by

descending StPM. The first row of the resulting table reports the queried profile. Each following row includes the name of the cell line, the data set it belongs to, and the catalogue code of the line, when it exists, the StPM for the queried profile, and all STR profile values (first the loci of the standard kit, then all remaining loci), where matching values are shown in red (see figure 1).



**CLIMA 2.1: the Cell Line Integrated Molecular Authentication Database 2.1**  
Version 2.1.201505  
This system is under active development. Please forgive us for errors and send us your comments, criticisms and congratulations.  
For information: Paolo Romano, Bioinformatics, IRCSS AOU San Martino IST ([paolo.romano@hsanmartino.it](mailto:paolo.romano@hsanmartino.it))

**CLIMA 2.1: Cell line identification: results**

You submitted the following information  
D5S818: 11,12; D13S317: 12,13,3; D7S820: 8,12; D16S539: 9,10; VWA: 16,18; TH01: 7; Amelogenin: X; TPOX: 8,12; CSF1PO: 9,10;  
Lower percent match to show: 80%

Name	Dataset	Cat. No.	% Match	D5S818	D13S317	D7S820	D16S539	VWA	TH01	AMG	TPOX	CSF1PO	D18S51	D19S433	D21S11	D2S1338	D3S1358	D8S1179	FGA	
<b>Query</b>				11,12	12,13,3	8,12	9,10	16,18	7	X	8,12	9,10								
HEp-2	ATCC	CCL-23	100.00%	11,12	12,13,3	8,12	9,10	16,18	7	X	8,12	9,10								
Intestine 407	ATCC	CCL-6	100.00%	11,12	12,13,3	8,12	9,10	16,18	7	X	8,12	9,10								
GIRARDI HEART C7	DSMZ	ACC-121	100.00%	11,12	12,13,3	8,12	9,10	16,18	7	X	8,12	9,10								
HeLa229	JCRB	JCRB9086	100.00%	11,12	12,13,3	8,12	9,10	16,18	7	X	8,12	9,10								
Chang Liver	JCRB	IFO50016	100.00%	11,12	12,13,3	8,12	9,10	16,18	7	X	8,12	9,10								
HeLa-P3	JCRB	JCRB0649	100.00%	11,12	12,13,3	8,12	9,10	16,18	7	X	8,12	9,10								
KB-V1	DSMZ	ACC-149	100.00%	11,12	12,13,3	8,12	9,10	16,18	7	X	8,12	9,10								
WRL 68	ATCC	CL-48	100.00%	11,12	12,13,3	8,12	9,10	16,18	7	X	8,12	9,10								
KB	JCRB	JCRB9027	100.00%	11,12	12,13,3	8,12	9,10	16,18	7	X	8,12	9,10								
J-111	JCRB	JCRB0073	100.00%	11,12	12,13,3	8,12	9,10	16,18	7	X	8,12	9,10								
HELA-S3	DSMZ	ACC-161	100.00%	11,12	12,13,3	8,12	9,10	16,18	7	X	8,12	9,10								
GIRARDI HEART C2	DSMZ	ACC-116	100.00%	11,12	12,13,3	8,12	9,10	16,18	7	X	8,12	9,10								
HeLa	DSMZ	CRM-CCL-2	100.00%	11,12	12,13,3	8,12	9,10	16,18	7	X	8,12	9,10								
Clone 1-5c-4	ATCC	CCL-20.2	100.00%	11,12	12,13,3	8,12	9,10	16,18	7	X	8,12	9,10								
H1-HeLa	ATCC	CRL-1958	100.00%	11,12	12,13,3	8,12	9,10	16,18	7	X	8,12	9,10								
HeLa 229	ATCC	CCL-2.1	100.00%	11,12	12,13,3	8,12	9,10	16,18	7	X	8,12	9,10								

**Figure 1:** result page for cell line identification. See text for details.

In all pages, links are established to the description of the genomic loci that is provided by the Short Tandem Repeat DNA Internet DataBase (STRBase) (7). Links to the involved cell banks are also provided in all results pages, so that the profiles are connected to the actual human cell line in the respective cell bank. Other features, described in (5) and documented on-line, left unchanged. The tool is available on-line at <http://bioinformatics.hsanmartino.it/clima2/>.

## References

1. The International Cell Line Authentication Committee. New Resources for an Old Problem: Cell Line Cross-Contamination. *In Vitro Report* 2014, 48.3 <https://sivb.org/InVitroReport/48-3/cross-contamination.html>
2. Freshney RI. Authentication of cell lines: ignore at your peril! *Expert Rev. Anticancer Ther.* 2008, 8(3):311-314. <http://dx.doi.org/10.1586/14737140.8.3.311>; <http://www.ncbi.nlm.nih.gov/pubmed/18366279>
3. Lacroix M. Persistent use of 'false' cell lines. *Int. J. Cancer* 2008, 122(1):1-4. <http://dx.doi.org/10.1002/ijc.23233>; <http://www.ncbi.nlm.nih.gov/pubmed/17960586>
4. Masters JR, Thomson JA, Daly-Burns B, Reid YA, Dirks WG, Packer P, Toji LH, Ohno T, Tanabe H, Arlett CF, Kelland JR, Harrison M, Virmani A, Ward TH, Ayres KL, Debenham PG. Short tandem repeat profiling provides an international reference standard for human cell lines. *PNAS* 2001, 98(14):8012-8017. DOI: <http://dx.doi.org/10.1073/pnas.121616198>; <http://www.ncbi.nlm.nih.gov/pubmed/11416159>
5. Romano P, Manniello MA, Aresu O, Armento M, Cesaro M, Parodi B. Cell Line Data Base: structure and recent improvements towards molecular authentication of human cell lines. *Nucl Acids Res.* 2009, 37(Database issue):D925-D932. <http://dx.doi.org/10.1093/nar/gkn730>; <http://www.ncbi.nlm.nih.gov/pubmed/18927105>
6. ATCC Standards Development Organization Workgroup ASN-0002: Alston-Roberts C, Barallon R, Bauer SR, Butler J, Capes-Davis A, Dirks WG, Elmore E, Furtado M, Kerrigan L, Kline MC, Kohara A, Los GV, MacLeod RAF, Masters JR, Nardone M, Nims RW, Price PJ, Reid YA, Shewale J, Steuer AF, Storts DR, Sykes G, Taraporewala Z, Thomson J (2011). Authentication of Human Cell Lines: Standardization of STR Profiling. ANSI/ATCC ASN-0002-2011. Copyrighted by ATCC and the American National Standards Institute (ANSI). ANSI eStandards Store: <http://webstore.ansi.org/RecordDetail.aspx?sku=ANSI%2fATCC+ASN-0002-2011>
7. Ruitberg CM, Reeder DJ, Butler JM. STRBase: a short tandem repeat DNA database for the human identity testing community. *Nucl Acids Res* 2001, 29(1):320-322. <http://dx.doi.org/10.1093/nar/29.1.320>; <http://www.ncbi.nlm.nih.gov/pubmed/11125125>