CLASTR : the cellosaurus STR similarity search tool - A precious help for cell line authentication

Robin, Thibault; Capes-Davis, Amanda; Bairoch, Amos Marc

How to cite

ROBIN, Thibault, CAPES-DAVIS, Amanda, BAIROCH, Amos Marc. CLASTR : the cellosaurus STR similarity search tool - A precious help for cell line authentication. In: International Journal of Cancer, 2020, vol. 146, n° 5, p. 1299–1306. doi: 10.1002/ijc.32639

This publication URL: https://archive-ouverte.unige.ch//unige:126966
Publication DOI: 10.1002/ijc.32639

IJC
International Journal of Cancer

CLASTR: The Cellosaurus STR similarity search tool - A precious help for cell line authentication

Thibault Robin [1,2,3,4], Amanda Capes-Davis [5] and Amos Bairoch [1,2]

[1]CALIPHO Group, SIB Swiss Institute of Bioinformatics, CMU, Geneva, Switzerland
[2]Microbiology and Molecular Medicine Department, Faculty of Medicine, University of Geneva, Geneva, Switzerland
[3]Proteome Informatics Group, SIB Swiss Institute of Bioinformatics, CMU, Geneva, Switzerland
[4]Computer Science Department, Faculty of Sciences, University of Geneva, Geneva, Switzerland
[5]CellBank Australia, Children's Medical Research Institute, The University of Sydney, Westmead, NSW, Australia

Despite an increased awareness of the problematic of cell line cross-contamination and misidentification, it remains nowadays a major source of erroneous experimental results in biomedical research. To prevent it, researchers are expected to frequently test the authenticity of the cell lines they are working on. STR profiling was selected as the international reference method to perform cell line authentication. While the experimental protocols and manipulations for generating a STR profile are well described, the available tools and workflows to analyze such data are lacking. The Cellosaurus knowledge resource aimed to improve the situation by compiling all the publicly available STR profiles from the literature and other databases. As a result, it grew to become the largest database in terms of human STR profiles, with 6,474 distinct cell lines having an associated STR profile (release July 31, 2019). Here we present CLASTR, the Cellosaurus STR similarity search tool enabling users to compare one or more STR profiles with those available in the Cellosaurus cell line knowledge resource. It aims to help researchers in the process of cell line authentication by providing numerous functionalities. The tool is publicly accessible on the SIB ExPASy server (https://web.expasy.org/cellosaurus-str-search) and its source code is available on GitHub under the GPL-3.0 license.

Introduction

The Cellosaurus[1] is a knowledge resource on cell lines. It aims to describe all cell lines used in biomedical research. It currently contains more than 113,000 cell line entries from 625 species. The majority of the cell lines originate from two species, namely, human (75%) and mouse (17%). For each cell line, a wealth of information is provided, allowing researchers to easily get an idea of how the cell line was generated and its main characteristics. It provides cross-links to 88 different external resources (cell repository catalogs, ontologies, databases) and provides more than 105,000 references to 17,700 publications. The Cellosaurus is updated regularly and is available on the Swiss Institute of Bioinformatics (SIB) ExPASy web server[2] (https://web.expasy.org/cellosaurus) where users can browse the database or download the full set of data in three different formats (structured text, OBO and XLM).

One of the many purposes of the Cellosaurus is to make users aware of cell lines that are known or suspected to be cross-contaminated, misidentified or misclassified. Contaminated cell lines arise from the accidental introduction of foreign cell lines (cross-contamination) or microorganisms (microbial contamination). Misidentified cell lines are the result of errors in their gender or species, while cell lines are defined as misclassified when the tissue type, cell type or disease is incorrect. Cell line cross-contamination was singled out as a major contributor to the reproducibility crisis that has been recently highlighted in life sciences.[3] A recent study[4] estimated that over 30,000 scientific publications were based on data produced using misidentified cell lines. As a consequence, the results of such experiments are partially or totally unreliable. This situation has led journal editors and publishers to ask authors to authenticate the cell lines that they have used prior to publication. The preferred experimental approach to authenticate a cell line is the short-tandem repeat (STR) profiling method (Fig. 1) that had already proven its effectiveness in forensic applications.[5,6] Once the STR profile of a given cell line sample is obtained, it must be compared against a database of reference STR profiles to verify that it does not have an unexpectedly high similarity with an unrelated cell line. If that is the case, it could potentially mean that the cell line sample is either partially or completely contaminated or misidentified.

Author contributions: CLASTR concept: Robin T, Bairoch A and Capes-Davis A. Code writing: Robin T. Project supervision and context development of the Cellosaurus collected the STR profiles used by CLASTR: Bairoch A. Manuscript writing with support from Capes-Davis A: Robin T and Bairoch A.
Key words: authentication, cell lines, cell culture, contamination, misidentification, STR profiling
Conflict of interest: None declared.
DOI: 10.1002/ijc.32639
History: Received 9 Jul 2019; Accepted 14 Aug 2019; Online 23 Aug 2019
Correspondence to: Thibault Robin, Microbiology and Molecular Medicine Department, Faculty of Medicine, University of Geneva, Switzerland, Tel.: +41-22-379-02-40, Fax: +41-22-379-11-34, E-mail: thibault.robin@unige.ch

What's new?

Despite increased awareness, cell line cross-contamination and misidentification remain a major source of erroneous experimental results in biomedical research. Nowadays, researchers performing experiments on cell lines are thus expected to ensure their authenticity using short-tandem repeat (STR) profiling. The Cellosaurus, which compiles all publicly available STR profiles, has become a valuable knowledge resource for this purpose. However, the database lacked a dedicated tool allowing a similarity search for a query STR profile. Here, the authors present CLASTR, the Cellosaurus STR similarity search tool that aims to facilitate the authentication process and the detection of potentially cross-contaminated or misidentified cell lines.

All currently available cell line STR search tools are restricted in the number and scope of the STR profiles that are available in their database (see Table 1, next section), and a given sample would have to be tested against multiple data sets to ensure its authenticity. By compiling all publicly available STR profiles, the Cellosaurus provides a remedy for this issue. However, the database lacked a dedicated tool allowing a similarity search for a query STR profile until recently. Here we present CLASTR (Cell Line Authentication using STR), the Cellosaurus STR similarity search tool, which aims to provide a large panel of functionalities to facilitate the similarity search process.

Materials and Methods

STR profiles

The current Cellosaurus release (release July 31, 2019) contains STR profiles for 6,556 distinct cell lines (6,474 human, 46 mouse and 36 dog cell lines) from 444 different sources providing, by far, the largest publicly available data set in terms of the number of STR profiles (Table 1). More than two-thirds of these human cell lines have only a single source for their STR profile (Fig. 2). For cell lines with more than one source, these frequently disagree on the exact allele value of a given STR marker. Such "conflicting" markers have their different alleles and corresponding sources clearly labeled on the STR profile section of a Cellosaurus entry. Collectively, individual scientific publications constitute the largest source for the Cellosaurus STR profiles. Cell line collections such as ATCC, CLS, DSMZ, ECACC, JCRB, KCLB, RCB and TKG, and initiatives such as the COSMIC cell line project[7] are also major contributors (Fig. 3).

STR markers

The first multiplex STR amplification kits were limited in terms of the number of amplified STR markers, usually containing

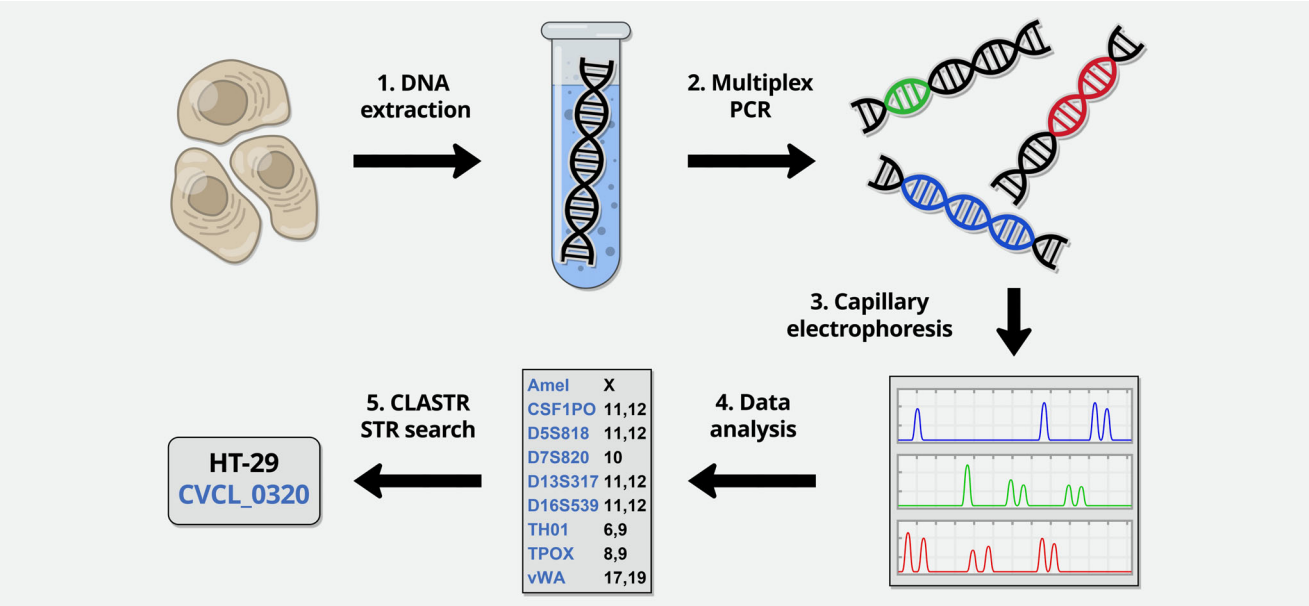

Figure 1. Workflow of cell line authentication by STR profiling. The process of the authentication of a cell line using STR profiling can be summarized by the following steps: (1) DNA is extracted from a cell sample; (2) fluorescent-labeled primers targeting specific STR loci are added and the corresponding DNA sequences are amplified simultaneously though Multiplex polymerase chain reaction (Multiplex PCR); (3) capillary electrophoresis is used to separate the amplified DNA fragments and the fluorescence is recorded to produce an electropherogram; (4) the electropherogram is interpreted as a STR profile by converting the size of each amplified fragment to the number of repetitions at each locus using specific software and controls, with manual validation as required; (5) the STR profile is searched against the Cellosaurus using CLASTR, allowing to know the identity of the cell sample and detect a potential cross-contamination. [Color figure can be viewed at wileyonlinelibrary.com]

Table 1. Comparison of the publicly available human STR profile data sets

Database	ATCC STR profile database	Cellosaurus	CLIMA	COG single record STR database	DSMZ STR profile database	NCBI BioSample
Number of human cell line with a STR profile	1,626	6,474	4,354	3,380	2,455	3,083

only four markers plus amelogenin used for gender determination. Although these studies continue to be informative, the number of STR loci used was insufficient to discriminate between cell lines from different donors. Over time, their limitations became clear, leading to the publication of an American National Standards Institute standard (ANSI ASN-0002-2011) for the authentication of human cell lines by STR profiling[8] in 2011. This standard requires the use of eight STR markers (CSF1PO, D13S317, D16S539, D5S818, D7S820, TH01, TPOX and vWA) plus amelogenin. In recent years, a large panel of multiplex STR systems has become commercially available, with a great variability in the STR markers and their numbers. The largest kits can amplify up to 27 STR markers at a time.

As a result, the STR markers constituting the STR profiles contained in the Cellosaurus can vary drastically between two entries. Currently, 32 distinct human STR markers (including amelogenin) are allowed in the database, although only 17 of them are commonly used in cell line STR genotyping.

It should also be noted that there have been some efforts recently to develop sets of STR markers for nonhuman species, specifically for dog[9,10] and mouse.[11,12]

Results

CLASTR, the Cellosaurus STR similarity search tool provides an intuitive and reliable platform to perform similarity searches on the human STR profiles contained in the Cellosaurus resource. It allows the efficient identification of cell lines and thus helps to detect contaminated and/or misidentified samples. The tool provides a wide range of search options for the users to choose from while implementing many useful functionalities. CLASTR is freely accessible on the ExPASy web server (https://web.expasy.org/cellosaurus-str-search). As the current number of mouse (44) and dog (28) cell lines with a STR profile is very limited we did not implement the option to search for similarity across samples from these species, but this may change if, as expected, the number of such authenticated nonhuman cell lines significantly increases over time.

Comparison to similar tools

Five bioinformatics tools enabling the pairwise comparison of STR profiles have been developed over time. These are the ATCC STR Profiling Analysis (https://www.atcc.org/STR%20Database.aspx), CLIMA,[13] the COG Single Record STR Database Search (https://strdb.cccells.org), the DSMZ Online STR matching analysis (OSTRA)[14] and the Search Program for the STR profile database of the JCRB Cell Bank (https://cellbank.nibiohn.go.jp/legacy/str2/top.html). All these tools were designed to support a specific set of STR profiles and, consequently, cannot be applied on other data sets. Furthermore, their source code is not publicly available. As the Cellosaurus has

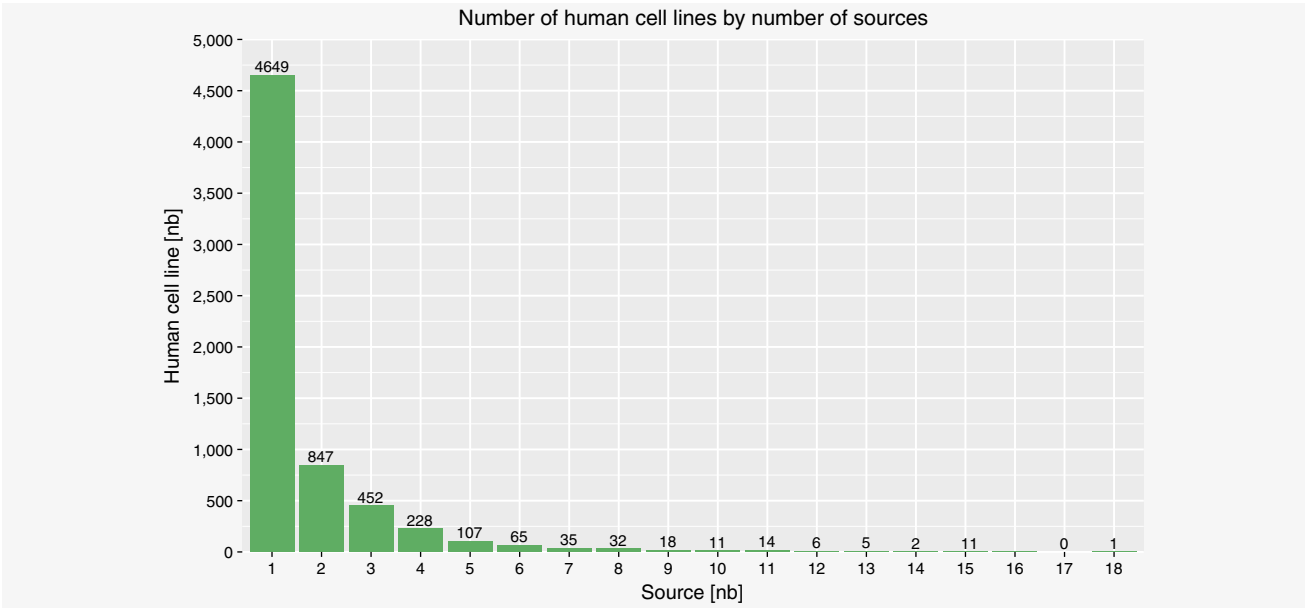

Figure 2. Number of human cell lines by number of sources. [Color figure can be viewed at wileyonlinelibrary.com]

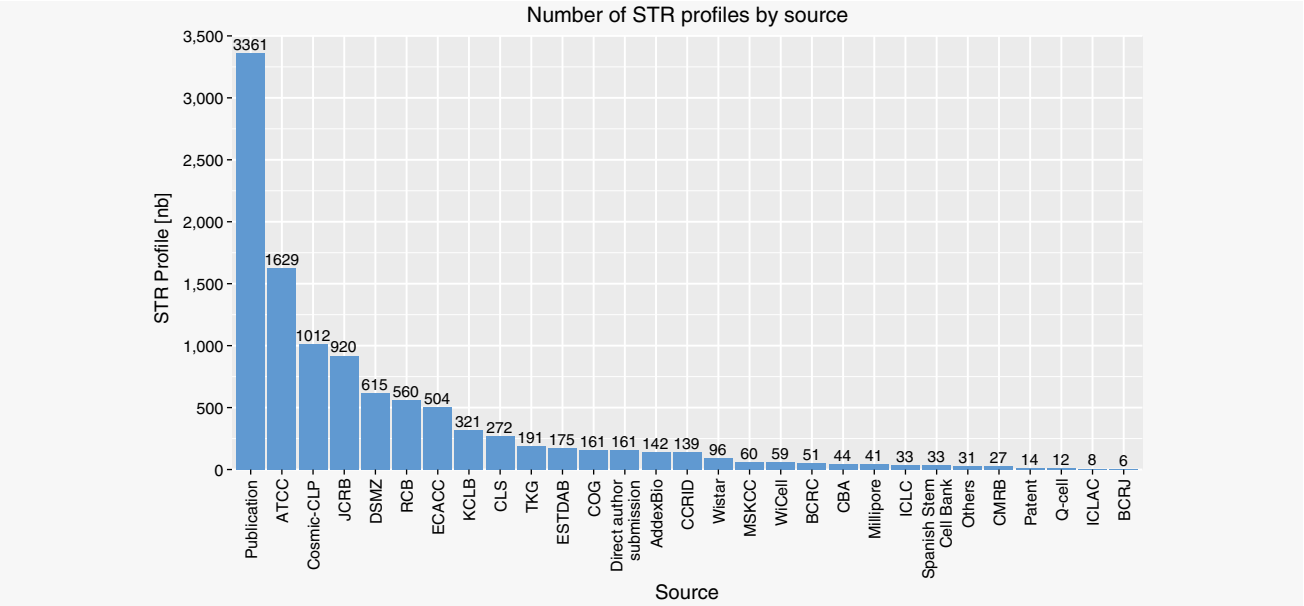

Figure 3. Number of STR profiles by source. [Color figure can be viewed at wileyonlinelibrary.com]

grown to become the largest public repository of human cell line STR profiles, the need for a tool enabling searching for similar STR profiles became more pressing and many users requested this feature. This prompted the development of CLASTR, which implements numerous features to facilitate and automate the search process.

While all previously available STR similarity search tools are based on the same core principle and share similar scoring algorithms, they differ from CLASTR in terms of user experience and interactivity (Table 2). The large majority of these tools do not allow the generated result table to be exported directly, preventing the storage of matching cell line information. In all these tools, allele information can only be entered manually, which can be tedious and prone to errors. Moreover, the absence of a public API or any other means to perform batch searches makes the process challenging and time-consuming when analyzing more than a few STR profiles. CLASTR is currently the software with the most features available and the added advantage of accessing the largest data set of human cell line STR profiles. The tool is designed to provide an easy, intuitive experience for experimental researchers and bioinformaticians.

Scoring algorithms

Two similar algorithms are frequently used to perform the pairwise comparison between a "query" and a "reference" STR profile: the Masters algorithm[15] and the Tanabe algorithm[16] (also known as the Sørensen–Dice coefficient). Both algorithms are based on the same concept, where a ratio is calculated based on the total number of alleles in each sample and the number of alleles that are shared by both samples. The Masters algorithm consists of the ratio between the number of shared alleles and the total number of alleles in the query (or in the reference for its modified version), while the Tanabe algorithm consists of the ratio between twice the number of shared alleles and the sum of the alleles in the query and reference. While these algorithms are fairly simple, they have been shown to be sufficiently

Table 2. Comparison of the features of the previously existing STR similarity search tools with CLASTR

Software	Available STR markers	Algorithms	Filters	Import file	Batch queries	Export formats	API
CLASTR	31	Tanabe and Masters	Score, min markers and max results	Yes	Yes	Excel (XLSX), CSV and JSON	REST
ATCC STR Profiling Analysis	8	Masters[1]	Score	No	No	CSV	No
CLIMA	8	Masters[1]	Score	No	No	None	No
COG single record STR database search	15	Masters[1]	Score	No	No	None	No
DSMZ Online STR matching analysis	8	Tanabe	Score	No	No	None	No
Search Program for the STR profile database of the JCRB Cell Bank	8	?[2]	?[2]	No	No	?[2]	No

[1]Modified version of the Masters algorithm only.
[2]This tool is no longer maintained and is nonfunctional, thus we could not check the status of these features.

effective to discriminate related from unrelated cell lines, provided that enough STR markers were compared.[17] Both the Masters and Tanabe algorithms are implemented in CLASTR and can be selected when performing a search. As the Tanabe algorithm is symmetrical and produces the same score if the query and reference cell lines are swapped, we have selected it as the default algorithm.

Scoring modes

Although the scoring algorithms are precisely described in terms of the score computation itself, they do not define a default behavior in the case of missing allele data for one of the two STR profiles to be compared. Such a problem is particularly relevant to the Cellosaurus data set because it contains STR profiles originating from many different sources that vary in the number and extent of the analyzed STR markers. To address this problem, we implemented different scoring modes. By default, only the STR markers for which both the query and reference have allele data are included in the score computation. However, as an option, the user can choose to compute the score based on all the query markers, even if the reference is lacking allele data for some of the markers. The reverse option is also available.

It is important to note that both algorithms do not define if the homozygous STR loci should count as one or two alleles in the score computation. Based on feedback from members of the International Cell Line Authentication Committee (ICLAC) and from experts in cell line STR profiling we decided to count homozygous STR loci as one allele. This choice is motivated by the fact that many cell lines present abnormal karyotypes and are no longer diploid. The chromosome counts can vary even between cells of the same culture. By counting homozygous loci as one, we do not falsely imply a specific number of chromosomes that could turn out to be often erroneous.

Conflicting STR profiles

Another consequence of the great diversity of sources of STR profiles in the Cellosaurus is that they may disagree on the exact allele value of a given STR marker. Since the similarity search is performed as a pairwise comparison, only one version of a given conflicting STR marker can be searched at a time. Cell lines that contain one or more conflicting STR marker values need to be handled differently than those without any conflicts. By default, our tool will try to resolve the conflicts by grouping the alleles of conflicted STR markers in distinct STR profiles based on their common sources. If this step cannot be properly completed because the set of sources differ between the STR markers, all the possible combinations of the alleles (up to a maximum of 150) of conflicted STR marker are then computed and stored as "virtual" STR profiles. Since it would not be manageable to report all these STR profiles in the results, only those with the best and worst scores are displayed so as to represent the extremes of the range of computed STR profiles.

Problematic cell lines

As noted in the Introduction, the contamination and misidentification of cell lines is a critical issue directly affecting the reproducibility of scientific research. Hence, one of the top priorities for the Cellosaurus is to clearly list and label all cell lines deemed to be problematic. The development of CLASTR was initiated so as to further endorse the objective of warning investigators about potential problems regarding the cell lines they are using in their research. As the identification of problematic cell lines is one of the key features of the tool, special care was taken to ensure that these cell lines would be properly flagged in both the web interface and export formats.

Web interface

The CLASTR web interface was designed as a single-page application composed of two key components: the input form and the result table. The input form consists of a main panel containing the STR marker input fields and a side panel containing the different parameters and possible actions (Fig. 4). The STR markers on the main panel are divided into two columns: the left one representing Amelogenin and the most common STR markers (CSF1PO, D2S1338, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D19S433, D21S11, FGA, Penta D, Penta E, TH01, TPOX and vWA) and the right one representing the less common STR markers (D1S1656, D2S441, D6S1043, D10S1248, D12S391, D22S1045, DXS101, DYS391, F13A01, F13B, FESFPS, LPL, Penta C and SE33). A cross-reference to the STRBase database[18] is also provided, which acts as a source of further information for all STR loci having a corresponding entry page in this resource. By default, only the most common STR markers are included in the result table. The less common STR markers need to be checked prior to the search so as to be displayed in the results. At the right of the markers input section, there is a panel that contains (from top to bottom): (*i*) choices for scoring algorithms and modes along with the possibility to include amelogenin in the score computation; (*ii*) the option to filter the search results by a minimum score, by a minimum number of common STR markers and by a maximum number of Cellosaurus entries returned; and (*iii*) buttons to initiate the main actions available and to be directed to the Help and About pages.

A useful feature of the web interface is the ability to load STR profile data from a file into the input form for subsequent searches. The input table file can either be a plain text file (CSV, TSV, TXT) format, a Microsoft Excel file (XLS, XLSX) format or a file produced by the GeneMapper ID-X software (https://www.thermofisher.com/ch/en/home/industrial/human-identification/genemapper-id-x-software.html). The table needs to include a column representing the sample name (labeled as "Name," "Sample" or "Sample Name") and a range of columns representing the STR markers. Note that the order of the columns does not matter and additional columns will be ignored. As an option, a "batch query" can be performed, performing iteratively the similarity search on all the samples contained in the

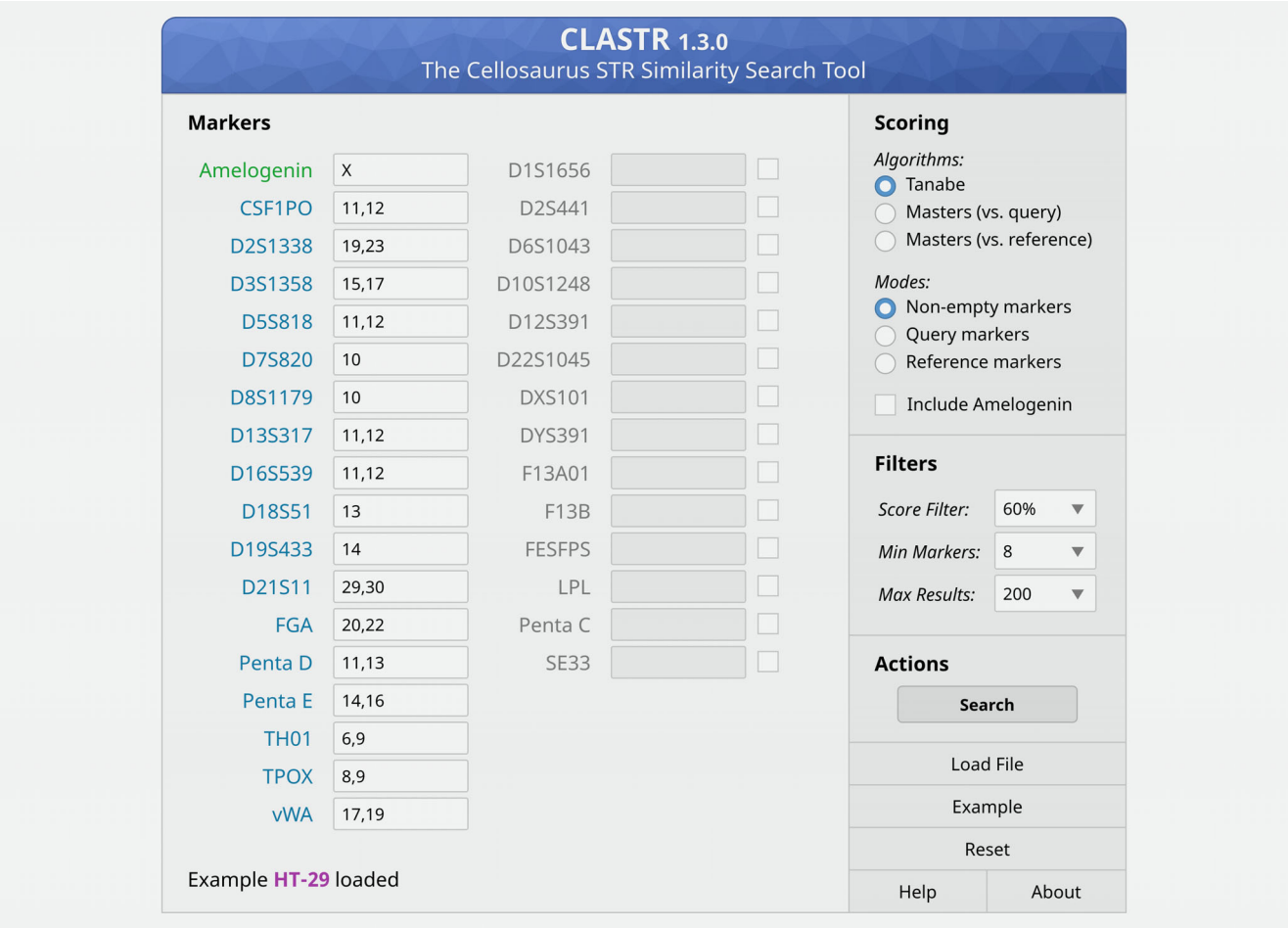

Figure 4. Screenshot of the input form. [Color figure can be viewed at wileyonlinelibrary.com]

table input file and directly returning all the results in XSLX, JSON or CSV formats. Note that in the case of the CSV format, one file per submitted sample is generated; these are returned as a single compressed archive in ZIP format.

The result table is a dynamic and sortable HTML table that is displayed once the similarity search is complete. The first row of the table always represents the query that was submitted to be searched against the STR profiles in the Cellosaurus. The first column labeled "Accession" provides the Cellosaurus accession number of the cell line along with a link toward its corresponding Cellosaurus entry page on ExPASy. In the case of problematic cell lines, the accession number is displayed in red and is associated with a tooltip describing the problem. In the case of a cell line with one or more STR allele value conflicts, an additional keyword is inserted to specify if the tested STR profile corresponds to the "best" or "worst" result (see section on conflicting STR profiles). The second column labeled "Name" provides the name of the cell line. The third column labeled "N° Markers" provides the number of STR markers that were used in the score computation, which depends on the scoring mode selected (see section on scoring modes). The fourth column labeled "Score" provides the computed score from the pairwise comparison based on the selected scoring algorithm and related parameters. The left border of the table is color coded so as to conveniently indicate if the cell line is highly related with the query or not. The following columns provide the alleles of the STR markers. The alleles that do not match the ones of the query are displayed in red. In the case of conflicted STR markers, the alleles are underlined and tooltips indicating the corresponding sources are provided.

An export table button located at the top left of the table provides the ability to directly export the generated table in XSLX, CSV or JSON formats. Note that all the files generated by the tool contain metadata, making it possible to keep track of the relevant search information (version of the Cellosaurus, version of CLASTR, run date and parameters) regardless of the selected format.

Additionally, each Cellosaurus entry page associated with a human cell line STR profile is directly linked to the CLASTR home page. The link encapsulates the STR profile information

as URL parameters thus allowing the tool to load the corresponding allele data automatically. This is particularly useful to search a specific cell line and identify the cell lines that have similar STR profiles.

RESTful API

The CLASTR RESTful API allows STR profile searches to be performed without needing to use the web interface. Two main distinct public API resources are available: "single entry mode query" and "batch mode query." The single entry mode query provides the ability to search a single STR profile using a GET or POST HTTP method and retrieve the response content in XLSX, CSV or JSON formats. The batch mode query provides the ability to search several STR profiles at the same time using a POST HTTP method and retrieve the response content in XLSX, CSV or JSON formats. More information about the RESTful API is available in the online help page (https://web.expasy.org/cellosaurus-str-search/help.html).

Source code

The CLASTR source code is composed of three main parts: the front end, the back end and the web app. The front end handling the web interface is written in HTML/CSS and JavaScript with the jQuery and jQuery UI libraries. The back end performing the similarity search is written in Java 8. The web app integrating the back end and linking it to the front end, while also managing the RESTful API, is written in Java 8 and is deployed as a web application using the Apache Tomcat application server. All source code is publicly available on GitHub (https://github.com/calipho-sib/cellosaurus-STR-similarity-search-tool) under the GPL-3.0 license. Python 3 scripts showcasing the use of the RESTful API are also available.

Data privacy

Data privacy is an important concern when it comes to genomic data.[19,20] However, this has not hindered the growth in the number of publicly available DNA sequences in recent years.[21] Although a STR profile represents only a fraction of the total genomic information, it still has the potential to identify the individual from which it originated. To address this issue, we made sure that no query data is kept on the server once the similarity search is completed. Moreover, all connections are encrypted using the HTTPS security protocol in order to ensure data confidentiality during transfers.

Discussion

Cell line authentication remains the best approach to address the problem of cross-contamination and misidentification. Although it does not directly prevent contamination cases, it allows researchers to verify the authenticity of a cell line before performing any experiments on it. Because contamination has numerous causes, including poor technique (e.g., sharing media between cell lines),[22] we expect that new cases will continually arise despite the implementation of stricter requirements. Consequently, contamination is expected to be an enduring issue in biomedical research, and bioinformatics will need to play an important role to limit its prevalence. Increased knowledge about contaminated cell lines needs to be gathered and the authentication process needs to be facilitated for all actors involved with cell lines, aims which are being addressed by the Cellosaurus and by CLASTR, respectively.

A big advantage of using the Cellosaurus STR data set is that its extensiveness increases the probability of detecting potential contamination cases. As all other available STR similarity search tools are based on their own specific data sets, which are restricted in scope, a number of STR profiles are never compared against each other and some problematic cases can be missed. This is especially relevant since the majority of the STR profiles contained in the Cellosaurus come from individual publications, as mentioned in the data section.

For each cell line entry, the Cellosaurus reports its hierarchy (i.e., if it has parents or children) and if other cell lines originate from the same individual. In the case of a contaminated cell line entry, its hierarchy is modified to indicate the contaminating cell line as parent. This type of information allows one to know if two given cell lines are annotated as related and are thus expected to have similar STR profiles. If two unrelated cell lines turn out to have a high similarity, further investigations are required to determine if they are actually related or if contamination is involved. To automate this auto-validation process of the Cellosaurus, CLASTR was adapted to run as a procedure in which a search is performed against all STR profiles. This workflow enables the investigation of any cell line pair that has an unexpectedly high similarity. Over time, new problematic cases will regularly be reported to ICLAC (https://iclac.org) to be investigated and the corresponding Cellosaurus entries will be updated accordingly to reflect their problematic status.

Originally, eight core STR markers were believed to be sufficient to be able to discriminate related from unrelated cell lines[17] fully. However, at the time, studies were already pointing out that this number may be too limited and that at least 13 core STR markers should be compared to identify cell lines with a high confidence.[23] While the cell line community seems at present to agree that using only eight STR markers is too limiting, few recent studies explore in detail how beneficial it is to include more STR markers and what is the preferred number of markers to be used. With the large amount of STR profiles that the Cellosaurus provides and the flexibility brought by CLASTR, we expect that it could also help in the process of establishing new standards and guidelines.

Variations in STR profiles obtained from the same cell line are an important issue when interpreting data from CLASTR searches. These variations arise due to biological and technical factors. Many cell lines were established from malignant tissue, which is inherently heterogeneous. Clonal populations are present that evolve as the culture is passaged and when derivatives are established. Laboratories with expertise in cell line

authentication have developed match criteria for interpretation of variable STR profiles, which are discussed elsewhere.[15–17,23] However, it is important to minimize such variation by "banking" cell lines at early passage. This is an important part of good cell culture practice for all laboratories, along with the need to perform authentication testing before further work commences.[24] Technical factors may also arise, due to variations in STR profiling procedures and data interpretation. All laboratories that perform STR profiling must do so using a standardized approach. Standards have been developed for authentication of human and nonhuman cell lines that set out consensus requirements.[8] Adherence to these technical standards, and the use of early passage material for testing, will ensure that published STR profiles are fit for use in CLASTR searches.

In its initial testing phase, CLASTR received an overwhelmingly positive response from beta-testers with expertise in cell lines and STR profiling; it was significantly improved thanks to their feedback. CLASTR was made publicly available on ExPASy on March 22, 2019. Since then it has been used ~3,500 times. In approximately half of the cases, it has been accessed from its home (input) page while the second half consist of users accessing it from a Cellosaurus cell line entry. With the progressive increase in the number of STR profiles stored in the Cellosaurus, we expect that the popularity of CLASTR will grow concomitantly over time.

Acknowledgements

We are very grateful to Richard Neves and Gregory Sykes for actively beta-testing CLASTR and for providing extremely useful feedback. We thank Elisabeth Gasteiger for installing the tool on the ExPASy server and for enabling a Cellosaurus entry with a STR profile to be directly linked to CLASTR. We are also thankful to Monique Zahn for careful proofreading of this manuscript.

References

1. Bairoch A. The Cellosaurus, a cell-line knowledge resource. *J Biomol Tech* 2018;29:25–38.
2. Artimo P, Jonnalagedda M, Arnold K, et al. ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res* 2012;40:W597–603.
3. Freedman LP, Gibson MC, Ethier SP, et al. Reproducibility: changing the policies and culture of cell line authentication. *Nat Methods* 2015;12:493–7.
4. Horbach SPJM, Halffman W. The ghosts of HeLa: how cell line misidentification contaminates the scientific literature. *PLoS One* 2017;12:e0186281.
5. Clayton TM, Whitaker JP, Maguire CN. Identification of bodies from the scene of a mass disaster using DNA amplification of short tandem repeat (STR) loci. *Forensic Sci Int* 1995;76:7–15.
6. Holt CL, Stauffer C, Wallin JM, et al. Practical applications of genotypic surveys for forensic STR testing. *Forensic Sci Int* 2000;112:91–109.
7. Forbes SA, Beare D, Boutselakis H, et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res* 2017;45:D777–83.
8. Almeida JL, Cole KD, Plant AL. Standards for cell line authentication and beyond. *PLoS Biol* 2016;14:e1002476.
9. Fowles JS, Dailey DD, Gustafson DL, et al. The Flint animal cancer center (FACC) canine tumour cell line panel: a resource for veterinary drug discovery, comparative oncology and translational medicine. *Vet Comp Oncol* 2017;15:481–92.
10. Berger B, Berger C, Hecht W, et al. Validation of two canine STR multiplex-assays following the ISFG recommendations for non-human DNA analysis. *Forensic Sci Int Genet* 2014;8:90–100.
11. Almeida JL, Hill CR, Cole KD. Mouse cell line authentication. *Cytotechnology* 2014;66:133–47.
12. Almeida JL, Dakic A, Kindig K, et al. Interlaboratory study to validate a STR profiling method for intraspecies identification of mouse cell lines. *PLoS One* 2019;14:e0218412.
13. Romano P, Manniello A, Aresu O, et al. Cell line Data Base: structure and recent improvements towards molecular authentication of human cell lines. *Nucleic Acids Res* 2009;37:D925–32.
14. Dirks WG, MacLeod RAF, 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;126:303–4.
15. 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;98:8012–7.
16. Tanabe H, Takada Y, Minegishi D, et al. Cell line individualization by STR multiplex system in the cell bank found cross-contamination between ECV304 and EJ-1/T24. *Tissue Cult Res Commun* 1999;18:329–38.
17. 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;132:2510–9.
18. Ruitberg CM, Reeder DJ, Butler JM. STRBase: a short tandem repeat DNA database for the human identity testing community. *Nucleic Acids Res* 2001;29:320–2.
19. Naveed M, Ayday E, Clayton EW, et al. Privacy in the genomic era. *ACM Comput Surv* 2015;48:1–44.
20. Kaye J. The tension between data sharing and the protection of privacy in genomics research. *Annu Rev Genomics Hum Genet* 2012;13:415–31.
21. Poon H, Quirk C, DeZiel C, et al. Literome: PubMed-scale genomic knowledge base in the cloud. *Bioinformatics* 2014;30:2840–2.
22. Capes-Davis A, Theodosopoulos G, Atkin I, et al. Check your cultures! A list of cross-contaminated or misidentified cell lines. *Int J Cancer* 2010;127:1–8.
23. Bady P, Diserens A-C, Castella V, et al. DNA fingerprinting of glioma cell lines and considerations on similarity measurements. *Neuro-oncology* 2012;14:701–11.
24. Geraghty RJ, Capes-Davis A, Davis JM, et al. Guidelines for the use of cell lines in biomedical research. *Br J Cancer* 2014;111:1021–46.