An interactive tool enabling a comparative analysis of STR profiles

G. Ukmar, R. Bosotti, A. Somaschini, J. Malysko, L. Raddrizzani, G. Masetti, E. Scacheri, A. Isacchi, A. Nuzzo✉

Business Unit Oncology, Nerviano Medical Sciences, Milan, Italy

Motivations

Cell line misidentification is a critical issue in molecular experiments that use normal and/or tumor cell lines as tools for disease characterization and pharmaceutical treatments. Thus, confirmation of cell line genetic identities is crucial to validate the obtained results. DNA fingerprinting of Short Tandem Repeats (STR) has become a standard technique used to identify the unique genetic profile of a cell line, based on the comparison of its microsatellite loci pattern with a known profile. The efficiency of this approach depends on the number of loci analysed, as well as on the existence of a large reference dataset. We applied a large scale STR characterization of 16 loci to a panel of about 300 commercially available tumor cell lines, generating an unprecedented dataset with uniform characterization. In order to maximize its exploitation by the scientific community, we developed an interactive software tool which allows to easily perform the automated comparison of the STR profile obtained for a cell line of interest against Nerviano Medical Sciences cell line database, facilitating the identification of the cell line and the potential discovery of unreported similarities.

Methods

We designed and developed an easy-to-use software tool which allows to quickly retrieve STR profiles based on the degree of similarity to the input profile of the cell line of interest. The tool is built on a dataset which includes about 300 commercially available human tumor cell lines, profiled in house. The panel is representative of diverse tumor tissue types. Users can insert data relative to a cell line of interest specifying either the cell line name or its STR profile. Then a query is run against the dataset in order to compute a similarity score versus each cell line and to generate the final summary table. The similarity is computed using a score that we implemented in order to account for multiallelic values at individual loci, which may be easily found in aberrant tumor cell line genomes. The similarity threshold by which two cell lines are considered identical is set at 80%. This cut-off value has been defined through a sensitivity analysis of available profiles. The tool has been developed using the Java programming language, which makes it easily portable on different platforms. The Graphical User Interface is composed of four main sections: i) the "Cell line" identification section by which the user may enter a specific identifier name or choose a cell line from the underlying dataset; ii) the "Loci" input pane, where the user can enter allelic values for each locus, or automatically retrieve loci values for an available cell line; iii) the "Command" pane containing the search task launcher button; iv) the "Result table" section, which reports all cell line matches found with a similarity score of at least 80%.

Results

The tool has been implemented and used to perform a complete characterization of the STR profiles of a 300 tumor cell line panel. Query automation allowed to calibrate optimal settings for the parameters involved in cell identification. In particular, the optimal threshold cut-off value has been identified through a sensitivity analysis using the available profiles. Moreover, the similarity value that we computed overcomes critical aspects of other adopted scores, which usually either do not take into account or provide inconsistent values for multiallelic loci. The availability of a portable tool will allow bench scientists to have an immediate authentication of the cell line of interest, which is nowadays a mandatory requirement for paper submission to most scientific journals.