A comprehensive 16 loci-based DNA fingerprinting dataset of a broad tumor cell line panel for cancer research

A. Somaschini, E. Scacheri, A. Nuzzo, N. Amboldi, D. Ballinari, G. Ukmar, A. Isacchi, R. Bosotti ✉
Business Unit Oncology, Nerviano Medical Sciences, Milan, Italy

Motivations

Tumor cell lines are widely used as in vitro models and as screening tools in cancer research. A significant portion of cell lines have been reported to be misidentified, potentially generating misleading data interpretation. Correct assessment of cell line genetic identity is therefore critical to cancer biology studies as well as to drug discovery. Several methods may be applied to authenticate cell lines, but DNA fingerprinting based on the detection of Short Tandem Repeats (STR) has emerged as the standard approach. STR profiles for several commercial cell lines can be retrieved from sparse literature reports or from vendor databases, however these are mainly based on the analysis of 8 loci, except for the NCI60 panel, on which 16 loci were profiled. Inconsistencies can be found in allele designation among the different authors, mainly when more than two alleles per locus are present, as is frequently found in tumor cell lines. A resource to facilitate literature interrogation is represented by the Cell Line Integrated Molecular Authentication database (CLIMA), but to date there are no reported comprehensive databases containing 16 loci profile sets generated in parallel. Here we disclose the in-house generation of a homogeneous 16 loci dataset, reporting the STR profiles for a panel of about 300 tumor cell lines, representative of diverse solid and circulating tumor types. Our intent was to facilitate the management of the internal cell bank, by assessing the identity and stability over time of the tumor cell lines used for research, with the highest accuracy in discrimination and with an optimized score for profile similarity checking. The STR database can be linked to Nerviano Medical Sciences (NMS) internal database, which contains information on cell line origin, morphology, growth conditions, as well as known somatic mutations (from the Wellcome Trust Sanger Institute Cosmic database), thus making it a comprehensive integrated platform for cell line utilization in drug discovery. Our plan is to make the STR database available to the scientific community.

Methods

DNA fingerprinting is based on the simultaneous amplification of highly polymorphic STR sites, which are short DNA sequences with a varying number of repeats in each cell line. Genetic abnormalities in tumor cell lines can result in more than two alleles at each locus, which complicates the analysis. Most STR profiles reported so far are based on the analysis of a limited number of STR loci, usually 8. The probability for two individuals to share the same STR profile is estimated to range from ~10-8 for 8 loci to ~10-17 for 16 loci. A 16 loci profile can be beneficial in cancer research, due to the intrinsic genetic instability of tumor cells that may result in allele acquisition or loss. In the current study we performed a 16-loci STR fingerprinting analysis on a panel of about 300 commercially available cancer cell lines. DNA was prepared directly from frozen cells, using NucleoSpin Tissue (Macherey-Nagel), and then analysed using AmpFlSTR Identifiler Plus PCR Amplification kit (Applied Biosystems), that amplifies simultaneously 15 tetranucleotide repeat loci and the amelogenin gender marker. For a more accurate comparison of STR profiles in different cell lines, we have introduced a modification to the standard calculation of similarity scores to account for partial allele identity at each locus, allowing a more precise definition of the level of similarity, which is particularly relevant when multiple alleles are present. This is obtained by applying the following formula that calculates the number of identical alleles at each locus, divided by the total number of alleles: given a cell line A and a cell line B, the score is defined as the identified sum of similarities at each i-th locus s(i) normalized by the number of not empty loci, where $s(i) = 2 \times [\text{alleles}(A) = \text{alleles}(B)] / [\#\text{alleles}(A) + \#\text{alleles}(B)]$. Based on a preliminary sensitivity analysis, the similarity score threshold to classify two cell lines as identical was set at 80%.

Results

A DNA fingerprinting analysis based on 16 loci was performed on a panel of about 300 widely

used tumor cell lines, purchased at NMS over several years from diverse providers. Using an established cut off of 80%, we found that all cell lines were correctly classified, suggesting that the cell line panel is genetically stable over time. As expected, an overall pairwise analysis of all the different cell line profiles revealed a generally low similarity degree (less than 50%), with no evidence for any similarity bias due to tumor type. The main outcome of this study is the availability of a large 16 loci STR dataset of tumor cell lines, which was integrated as a fundamental part of the NMS Cell Bank database. The STR dataset can be easily interrogated and distributed for use as a reliable reference for cancer research.