SNP Panel Identification Assay (SPIA): a genetic-based assay for the identification of cell lines

ID - 249

Demichelis Francesca[1,2,3], Heidi Greulich[2,4,5], Jill A. Macoska[6], Rameen Beroukhim[2,4,5], William R. Sellers[2,4,5], Levi Garraway[2,4,5], Mark A. Rubin[1,2,4,5]

[1]Department of Pathology, Brigham and Women's Hospital, Boston, MA, USA

[2]Harvard Medical School, Boston, MA, USA

[3]SRA, ITC-irst, Trento, Italy

[4]Dana Farber Cancer Institute, Boston, MA, USA

[5]Broad Institute of Harvard and the Massachusetts Institute of Technology, Cambridge, MA, USA

[6]Department of Urology and Comprehensive Cancer Center, University of Michigan, Ann Arbor, Michigan, USA

Motivation

Translational research hinges on the ability to make observations in model systems and to implement those findings into clinical applications, such as the development of diagnostic tools or targeted therapeutics. Tumor cell lines are commonly used to examine mechanisms of carcinogenesis. The same tumor cell line can be simultaneous studied in multiple research laboratories throughout the world, theoretically generating results that are directly comparable. One important assumption in this paradigm is that researchers are working with the same cells. However, recent work by our group and others suggest that experiments reported in the scientific literature may contain pre-analytic errors due to inaccurate identities of the cell lines employed. The MDA-MB435 cell line, which was thought to be a breast cancer cell line for many years, was determined to be genotypically identical to the M14 melanoma cell line (Garraway et al, 2005). Instances of inadvertent cross-contamination or other errors in sample processing that remain largely unrecognized by investigators may have profound adverse effects on experimental results and their interpretation. To address this problem, we developed a simple approach that enables an accurate determination of the identity of known cell lines by genotyping 34 single nucleotide polymorphisms (SNPs). Here, we describe the empirical development of a SNP panel identification assay (SPIA) compatible with routine use in the laboratory setting to ensure the identity of tumor cell lines and human tumor samples throughout the course of long term research use.

Methods

We know that extensive genotype profiles of DNA samples can work as a unique identifier of samples. We hypothesized that by using a small number of SNPs we would still be able to accurately distinguish samples, providing researchers with a convenient way to check the identity of their samples during the course of their use in the laboratory. To define the optimal SNP panel, we reasoned that the ideal SNPs should collectively maximize the probability of obtaining distinct genotype calls on different samples, i.e. exhibiting the greatest heterogeneity across different samples. The number of independent loci needed depends on the level of confidence one needs to make a definitive identification. Our dataset included genotype data of 155 cell lines derived from different organs, obtained through 50K Xba SNP arrays. The computational approach for the identification and the validation of the best SNP panel included: SNP filtering, a computational algorithm based on Hardy Weinberg equilibrium search, and a double probabilistic test based on Binomial distribution in order to discern when two cell lines are close enough to be called similar and when they are not. The test score depends on the number of single locus matches and on the total number of SNPs evaluated for the two cell lines. The SNP panel identification procedure was run 1000 times to ensure robust results. This procedure lead us to a ranked list of SNPs. Experimental validation of the identified SNP set was performed on a different platform (Sequenom mass spectrometric genotyping technology) on a set of ~ 90 cell lines.

Results

To measure the effect of the SNP selection process on the ability to distinguish different cell lines and to determine the minimum number of SNPs required to identify genetically similar cell lines, we evaluated the pair-wise distances using several sets of SNPs, by randomly sampling 80, 60, 40 and 20 SNPs from the top ranked SNPs. When comparing the distances obtained for the smaller sets of SNPs with the 5.3K SNP set (obtained after filtering), we see a significant increment of percentage differences. This result holds on the independent validation set of cell lines, confirming the ability of the selected SNPs of enforcing the differences between different samples. When comparing the results obtained with 80, 60, 40 and 20 SNPs to each other, we observe that the mean pair-wise distances do not change significantly. However, the standard deviations tend to increase when going from 80 to 20 SNPs. These results suggest that any set of 40 SNPs selected from the top 100 SNPs, provides researchers with a good SNP panel for DNA sample identification. However, the more SNPs in the panel, the more confident one can be in the final call. The

statistical test applied on 93 cell lines genotyped by Sequenom on 34 SNPs showed 100% accuracy. This assay can correctly identify a given DNA sample by comparing the genotype call set (bar code) within a reference database that contains bar codes of the most commonly used cell lines. Widespread application of this approach may reduce erroneous experimentation and data interpretation associated with inaccurate tumor sample identity, thereby providing a significant benefit to cancer scientists.
Email: fdemichelis@partners.org