Statistical Applications in Genetics and Molecular Biology

Volume 3, *Issue* 1 2004 *Article* 25

Concerns About Unreliable Data from Spotted cDNA Microarrays Due to Cross-Hybridization and Sequence Errors

Daniel Handley* Nicoleta Serban†

David G. Peters‡ Clark Glymour**

*University of Pittsburgh, dhandley@andrew.cmu.edu

†Carnegie Mellon University, nserban@stat.cmu.edu

‡University of Pittsburgh, dgp@imap.pitt.edu

**Carnegie Mellon University, cg09@andrew.cmu.edu

Concerns About Unreliable Data from Spotted cDNA Microarrays Due to Cross-Hybridization and Sequence Errors

Daniel Handley, Nicoleta Serban, David G. Peters, and Clark Glymour

Abstract

We discuss our concerns regarding the reliability of data generated by spotted cDNA microarrays. Two types of error we highlight are cross-hybridization artifact due to sequence homologies and sequence errors in the cDNA used for spotting on microarrays. We feel that statisticians who analyze microarray data should be aware of these sources of unreliability intrinsic to cDNA microarray design and use.

KEYWORDS: Cross-hybridization, Sequence Errors, cDNA microarray, Sources of Error

To the editor:

We would like to bring to your attention our concerns regarding the reliability of data generated by spotted cDNA microarrays. Spotted cDNA microarrays are both commercially available as well as often fabricated by individual investigators in their laboratories. This type of microarray is used extensively in nearly all areas of life sciences and biomedical research, and thus investigators should be made aware of any indications that there may be serious unreliability issues associated with it.

In a recent publication (Handley et. al.), we present evidence of significant cross-hybridization in expressed sequence tags (EST) IMAGE clones used on a commercially available single-dye cDNA microarray. A high proportion of these EST sequences were reported to contain 5'-end poly(dT) sequences that are remnants from the oligo (dT)-primed reverse transcription of polyadenylated mRNA templates used to generate EST cDNA for sequence clone libraries. Analysis of two expression data sets revealed that such sequences appear to be highly co-expressed, and that the degree of expression variability correlates with poly(dT) tract length. This prompted us to suspect the expression values stemmed from systematic cross-hybridization of these poly(dT) tracts rather than reflecting true mRNA expression levels.

Spotted cDNA microarray protocols commonly involve a prehybridization step in which poly(dA) oligonucleotide is used to block polyadenylated sites. However, the cDNA target deposited on these microarrays during manufacture is denatured double-stranded DNA. This means the target contains both poly(dT) tracts as well as their complementary poly(dA) sequences. The oligo(dA) added during this prehybridization step may bind to any exposed poly(dT) tracts, but not to the poly(dA). In mRNA expression experiments, the probe is typically produced by reverse transcribing oligo(dT)-primed mRNA to single-stranded labeled cDNA (e.g., using 33P-dCTP). Thus, the probe will have poly(dT) tracts that may hybridize to any exposed poly(dA) tracts on the microarray. However, only poly(dT) tracts on the microarray are blocked in the prehybridization step. We believe that this may be one potential source of the cross-hybridization signal we observed.

This cross-hybridization potential might be alleviated by annealing oligo(dA) to the probe, but since such annealing relies on probabilistic kinetics, it is far from certain that all the exposed poly(dA) or poly(dT) tracts will be completely suppressed. Thus, as long as poly(dA) and poly(dT) tracts remain on the target sequences, there should always be a concern for cross-hybridization artifact.

Also of concern is that in 2001, Halgren et. al. reported an exceedingly high error rate (approaching 38%) in the sequence of IMAGE mouse cDNA

clones versus the reported sequence. Much of the sequence error appeared to stem from contamination of the desired sequence with vector sequences.

A major focus of microarray data analysis appears to be centered on the statistical treatment of issues such as image analysis, normalization, and background subtraction. However, we feel that more attention should be give to artifact generated by target sequence errors and unintentional homologies. In this case in particular, we feel that investigators and statisticians should be especially cautious when interpreting data obtained from cDNA microarrays spotted with cloned sequences derived from eukaryotic mRNA.

References:

1. Handley D, Serban N, Peters D, O'Doherty R, Field M, Wasserman L, Scheines R, Spirtes P, Glymour C. Evidence of systematic expressed sequence tag IMAGE clone cross-hybridization on cDNA microarrays, *Genomics* 2004 83(6): 1169-75.

2. Halgren RG, Fielden MR, Fong CJ, Zacharewski TR, Assessment of clone identity and sequence fidelity for 1189 IMAGE cDNA clones, *Nucleic Acids Research*, 2001, Vol. 29, No. 2 582-588.