BioFeedback

[Letter to the editor]

Unexpected presence of mycoplasma probes on human microarrays

The contamination of cell cultures by mycoplasmas poses a problem in that it can adversely affect cellular behavior and physiology; this is exacerbated by the fact that colonization is not easy to detect (1). In a previous study investigating the effects on microarray data, it was shown that mycoplasma contamination can alter patterns of human gene expression by upsetting host cell physiology (2). Miller et al. used Affymetrix Human Genome U133A GeneChip microarrays (Santa Clara, CA, USA) containing mostly well-characterized human genes, and demonstrated that mycoplasma infection compromises the validity of any data generated from such samples (2).

We have discovered the probeset 1570561_at—on the microarray that succeeded the HG-U133A chip (HG-U133 Plus 2.0)—that maps to the 16S-23S rRNA intergenic transcribed spacer (ITS) sequences from multiple species of mycoplasma. Interestingly, this sequence is already used in molecular detection and screening protocols for mycoplasma infections and genotyping, using PCR and microarrays (3,4). In contrast to the HG-U133A array, HG-U133 Plus 2.0 arrays contain more probesets for less–well-characterized sequences. The 176-nucleotide Affymetrix target sequence used to select probes for this probeset was designed to a single human expressed sequence tag (EST)–like sequence (GenBank accession no. AF241217); according to the company, this entry does not have any similarities to any known human transcripts or genomic sequences.

However, we demonstrate that the target sequence bears overwhelming similarities to 16S-23S rRNA ITS sequences from various different species of mycoplasma (Table 1). The table shows the top 10 most similar alignments between the 176-nucleotide target sequence and entries in the GenBank multi-species 'nr' database; 9 of these 10 are mycoplasma rRNA gene sequences, while the remaining match is for the AF241217 entry itself (5). Our conclusion from this finding is that a probeset representing mycoplasma 16S-23S rRNA ITS has been included on a human microarray.

The GenBank entry AF241217 is a 249-nucleotide sequence annotated as "*Homo sapiens* unknown sequence" and was submitted to the high-throughput cDNA (HTC) division in 2000. We propose that the cDNA sequenced as part of this particular HTC project may have been derived from a mycoplasma-contaminated human cell line. This means that some of the cDNA clones in the library would contain mycoplasma gene fragments leading to inclusion of mycoplasma sequences in the human database. Then, while designing the HG-U133 Plus 2.0 array, Affymetrix included the AF241217 GenBank entry, as it appeared human in origin. The HG-U133A array used in the previous mycoplasma study (2) does not contain any probesets designed to this particular sequence.

To find out if any other mycoplasma genes were represented on HG-U133 Plus 2.0, we aligned all 54,675 of its target sequences against an example mycoplasma genome (*Mycoplasma arthritidis* 158L3–1; complete genome sequence; GenBank accession no. CP001047; 820,453 bp) using the FASTA program (6). The results showed no significant alignments other than 1570561_at as judged by Smith-Waterman scores and long overlaps as measures of similarity (data not shown). We therefore conclude that 1570561_at is probably the only probeset that aligns with mycoplasma sequences on this particular array.

In order to assess whether high relative signals for this probeset could indicate mycoplasma contamination, we downloaded data from the Gene Expression Omnibus (GEO) database at NCBI (www.ncbi.nlm.nih.gov/geo), which stores data from microarray studies and makes it publicly available and searchable. We intended to find out whether (*i*) there were any samples showing high expression levels for this probeset, and (*ii*) the majority of samples with high expression for 1570561_at were from cultured cells

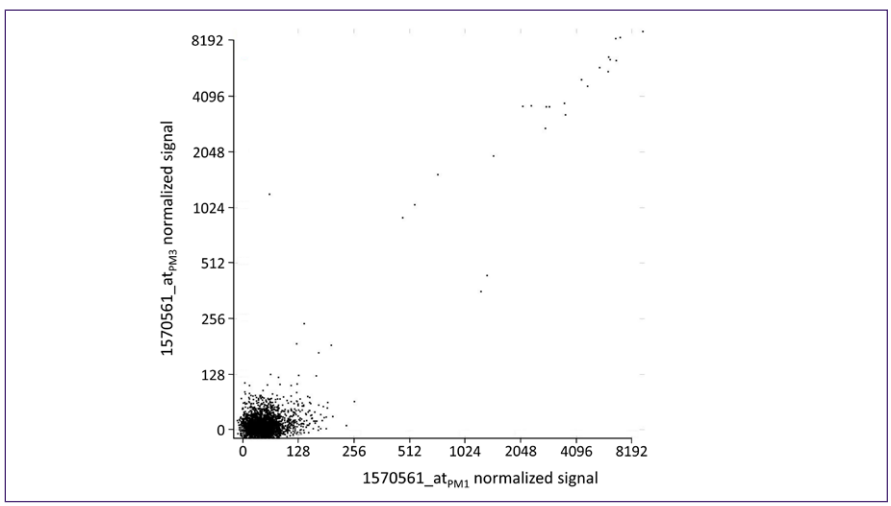

Figure 1. Scatterplot showing distribution of expression levels of 1570561_at, for 2757 samples in the GEO database hybridized to HG-U133 Plus 2.0 arrays. Each point on the graph represents the relative normalized signals for two of the PM probes (PM_1 and PM_3) from 1570561_at, for each of 2757 HG-U133 Plus 2.0 samples stored in the GEO database. Most of the samples are clustered together in the bottom left corner, with low relative expression levels, but there is a subset showing high signals for these probes. This plot was generated using the scatterplot function of RNAnet, hosted by the University of Essex, UK.

rather than non-cultured cells or tissues, which are less likely to be contaminated. GEO contained expression data from 2757 samples from HG-U133 Plus 2.0 hybridizations at the time of download (February 2007).

We analyzed a randomly-selected subset of these samples to find out the background frequency of cultured samples. Using this subset (1801 samples), we interpreted the sample descriptions and assigned each one into either a 'cultured' or 'non-cultured' group, depending on whether the cells in question had been subjected to standard culturing conditions. We found that the cultured samples made up 34% of the total (Supplementary Table S1).

Next, we analyzed the downloaded GEO HG-U133 Plus 2.0 data (2757 samples) and quantile-normalized the probe intensities using the RNAnet facility from the University of Essex (http://bioinformatics.essex.ac.uk/users/wlangdon/rnanet), which allows detailed querying of GEO data sets at the CEL file level (7). The scatterplot in Figure 1 shows that there is a cluster of samples that have high relative expression for 1570561_at; signals from two of the 11 perfect match (PM) probes (the first and the third, PM_1 and PM_3) were chosen as representative of the probeset and plotted against each other.

Using mean PM intensity as an expression measure, we created a list of samples ranked on 1570561_at relative expression level; the 33 samples with highest overall signals for 1570561_at contained 31 cultured and two non-cultured samples (Supplementary Table S2). This cultured fraction (94%) of high-expressing samples is significantly higher than the observed background frequency of 34% (chi-squared test; $X^2 = 0.0$). We suggest that high expression of this probeset appears to correlate with the act of physically culturing a sample. Coupled with the fact that the probeset has a high degree of similarity with many species of mycoplasma rRNA, we propose that some of the samples stored in GEO may have been derived from mycoplasma-contaminated cell cultures.

In conclusion, we suggest that: (*i*) the 1570561_at probeset has its origins not in the human genome, but from mycoplasma cDNA, (*ii*) it may detect the presence of mycoplasma RNA in a human microarray sample, and (*iii*) high expression levels for this probeset may be used as a post-hybridization biomarker of mycoplasma infection in samples processed on HG-U133 Plus 2.0 arrays. Furthermore, it is possible that a subset (perhaps ~1%) of all the data stored in the GEO database has been compromised, as detailed in Miller et al. (2), by mycoplasma infection. Further experiments including hybridizing different species of mycoplasma at known levels

Table1. Similarities between Affymetrix target sequence 1570561_at (176 nucleotides) and mycoplasma rRNA sequences

GenBank accession no.	Description	Max. score	Total score	Query coverage	E value	Max. ident.
FJ876260.1	*Mycoplasma orale* strain MT-4 16S ribosomal RNA gene, partial sequence; 16S-23S ribosomal RNA intergenic spacer, complete sequence; and 23S ribosomal RNA gene, partial sequence	326	326	100%	3.00E-86	100%
AF241217.1	*Homo sapiens* unknown sequence	326	326	100%	3.00E-86	100%
AF294995.1	*Mycoplasma orale* 16S-23S ribosomal RNA intergenic spacer, complete sequence; and 23S ribosomal RNA gene, partial sequence	326	326	100%	3.00E-86	100%
AY737010.1	*Mycoplasma orale* 16S ribosomal RNA gene, partial sequence; 16S-23S ribosomal RNA intergenic spacer, complete sequence; and 23S ribosomal RNA gene, partial sequence	320	320	100%	1.00E-84	99%
AY762640.1	*Mycoplasma indiense* 16S ribosomal RNA gene, partial sequence; 16S-23S ribosomal RNA intergenic spacer, complete sequence; and 23S ribosomal RNA gene, partial sequence	315	315	100%	6.00E-83	98%
EU859980.1	*Mycoplasma arthritidis* strain 91021 16S ribosomal RNA gene, partial sequence; 16S-23S ribosomal RNA intergenic spacer, complete sequence; and 23S ribosomal RNA gene, partial sequence	241	241	100%	1.00E-60	91%
CP001047.1	*Mycoplasma arthritidis* 158L3–1, complete genome	241	241	100%	1.00E-60	91%
AY973560.1	*Mycoplasma arthritidis* strain ATCC 19611 16S ribosomal RNA gene, partial sequence; 16S-23S ribosomal RNA intergenic spacer, complete sequence; and 23S ribosomal RNA gene, partial sequence	241	241	100%	1.00E-60	91%
EU859975.1	*Mycoplasma salivarium* strain MP1166 16S ribosomal RNA gene, partial sequence; 16S-23S ribosomal RNA intergenic spacer, complete sequence; and 23S ribosomal RNA gene, partial sequence	230	230	100%	2.00E-57	90%
AY973563.1	*Mycoplasma hyosynoviae* strain ATCC 25591 16S ribosomal RNA gene, partial sequence; 16S-23S ribosomal RNA intergenic spacer, complete sequence; and 23S ribosomal RNA gene, partial sequence	230	230	100%	2.00E-57	90%

The table shows the results of a MegaBLAST search of the NCBI 'nr' database that contains all RefSeq and UniGene information for multiple species. MegaBLAST is a modified BLAST tool used to find highly similar sequences. The data in the table is ranked by the E value, which indicates the statistical probability that the sequence similarity occurred by chance; the lower the score, the more significant the match. The second entry (AF241217.1) in the table corresponds to the original human cDNA used by Affymetrix to prepare the 176-nucleotide target sequence. This entry is a 249-nucleotide clone from the high-throughput cDNA (HTC) division of GenBank. Ident., identity (NCBI defines "max. ident." as "the highest percent identity for a set of aligned segments to the same subject sequence").

of contamination to HG-U133 Plus 2.0 arrays, or subjecting archived array hybridization cocktails to molecular mycoplasma detection assays, such as NAT (3,4), would be required to prove these hypotheses. However, we do not advocate this resource as a replacement for standard screening and detection of mycoplasma in cell cultures.

Acknowledgments

The authors specifically wish to thank Tanya Barrett and Alexandra Soboleva from the NCBI GEO database team for their help in supplying custom-formatted expression data. We also acknowledge the scientific contribution of Adeel Riaz and Mahvash Tavassoli at King's College London in helping to identify the probeset.

Competing interests

The authors declare no competing interests.

References

1. **Rottem, S. and Y. Naot.** 1998. Subversion and exploitation of host cells by mycoplasmas. Trends Microbiol. *6*:436-440.
2. **Miller, C.J., H.S. Kassem, S.D. Pepper, Y. Hey, T.H. Ward, and G.P. Margison.** 2003. Mycoplasma infection significantly alters microarray gene expression profiles. BioTechniques *35*:812-814.
3. **Kong, H., D.V. Volokhov, J. George, P. Ikonomi, D. Chandler, C. Anderson, and V. Chizhikov.** 2007. Application of cell culture enrichment for improving the sensitivity of mycoplasma detection methods based on nucleic acid amplification technology (NAT). Appl. Microbiol. Biotechnol. *77*:223-232.
4. **Jang, H., H. Kim, B. Kang, C. Kim, and H. Park.** 2009. Oligonucleotide array-based detection and genotyping of mollicutes (*Acholeplasma, Mycoplasma*, and *Ureaplasma*). J. Microbiol. Biotechnol. *19*:265-270.
5. **Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. *215*:403-410.
6. **Pearson, W.R. and D.J. Lipman.** 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA *85*:2444-2448.
7. **Langdon, W.B., G.J.G. Upton, R. da Silva Camargo, and A.P. Harrison.** A survey of spatial defects in Homo sapiens Affymetrix GeneChips. IEEE/ACM Trans. Comput. Biol. Bioinf. (In press).

Received 13 May 2009; accepted 17 September 2009.

Estibaliz Aldecoa-Otalora[1], William B. Langdon[2], Phil Cunningham[3], and Matthew J. Arno[1]

[1]Genomics Centre, School of Biomedical and Health Sciences, King's College London, London, UK, [2]Department of Computer Science, King's College London, London, UK, and [3]Department of Biochemistry, School of Biomedical and Health Sciences, King's College London, London, UK

Address correspondence to Matthew Arno, Genomics Centre, School of Biomedical and Health Sciences, King's College London, Franklin-Wilkins Building, 150 Stamford Street, London, SE1 9NH, UK. email: matthew.arno@kcl.ac.uk

Supplementary material for this article is available at www.BioTechniques.com/article/113271.

BioTechniques 47:1013-1016 (December 2009) doi 10.2144/000113271