SANDIA REPORT

SAND2005-7536
Unlimited Release
Printed November 2005

Examining Microarray Slide Quality for the EPA using SNL's Hyperspectral Microarray Scanner

Jerilyn A. Timlin, Rachel M. Rohde

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation,
a Lockheed Martin Company, for the United States Department of Energy's
National Nuclear Security Administration under Contract DE-AC04-94AL85000.

Approved for public release; further dissemination unlimited.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865)576-8401
Facsimile: (865)576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800)553-6847
Facsimile: (703)605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND 2005-7536
Unlimited Release
Printed November 2005

Examining Microarray Slide Quality for the EPA using SNL's Hyperspectral Microarray Scanner

Jerilyn A. Timlin and Rachel M. Rohde

Biomolecular Imaging and Analysis
Sandia National Laboratories
PO Box 5800
Albuquerque, NM 87185-0895

Abstract

This report summarizes research performed at Sandia National Laboratories (SNL) in collaboration with the Environmental Protection Agency (EPA) to assess microarray quality on arrays from two platforms of interest to the EPA. Custom microarrays from two novel, commercially produced array platforms were imaged with SNL's unique hyperspectral imaging technology and multivariate data analysis was performed to investigate sources of emission on the arrays. No extraneous sources of emission were evident in any of the array areas scanned. This led to the conclusions that either of these array platforms could produce high quality, reliable microarray data for the EPA toxicology programs. Hyperspectral imaging results are presented and recommendations for microarray analyses using these platforms are detailed within the report.

Acknowledgements

Funding for this work was provided from the EPA through WFO # 018040121 in collaboration with Greg Toth at the EPA. The authors with to thank Ann Miracle for supplying the Amersham microarrays used in these studies, Patrick Larkin and Barbara Carter of EcoArray, Inc. for supplying the Agilent oligomer microarrays used in these studies, and Rong-Lin Wang at the EPA for his assistance in preparing this report. In addition, the authors thank Howland Jones, David Melgaard, Chris Stork, Mark Van Benthem, David Haaland, and Michael Keenan at Sandia National Laboratories for development of the multivariate analysis software tools which were critical to facilitate the image analysis required for this work.

Table of Contents

Figures

Tables

Nomenclature

CLS	Classical least squares
Cy3	Cyanine 3
Cy5	Cyanine 5
DNA	Deoxyribonucleic acid
EPA	Environmental Protection Agency
HSS	Hyperspectral microarray scanner
MCR	Multivariate curve resolution
PCA	Principal component analysis
R/G	Red/Green ratio (microarray statistic)
SNL	Sandia National Laboratories

Introduction

The study presented in this report is part of a larger collaboration between the Environmental Protection Agency (EPA) and Sandia National Laboratories (SNL) to develop and apply improved methods for microarray science for the purpose of increasing the reliability of microarrays used for toxicology studies at the EPA. In the current investigations arrays from two different commercial platform technologies were investigated with SNL's hyperspectral microarray scanner for extraneous and/or contaminant emissions that could adversely affect the reliability of the data they generate. This report is a formal communication of the findings of those investigations.

Previous research efforts in our group have led to the development of a novel hyperspectral scanner (HSS) capable of collecting an entire fluorescence emission spectrum (>500 spectral channels) at each pixel of a microarray sample. (Sinclair et al., 2004) Analysis of the hyperspectral data with advanced multivariate methods permits identification and quantitation of each emitting species on a microarray, including those that are unexpected and those that possess highly overlapped emission spectra. This capability differs from commercial microarray scanning technology because commercial scanners operate as univariate (or one channel at-a-time) instruments with a total of two spectral channels. Commercial array scanners typically use filters to pass all the photons from a particular wavelength range to a detector simultaneously and thus can confound extraneous emissions with signal from the analyte of interest if their spectra overlap in this wavelength range. This reduces the accuracy and reliability of data generated from commercial instruments if extraneous emissions are present. Using our HSS we previously identified extraneous emissions on glass substrate microarrays, most notably a persistent, but variable green channel contaminant that can significantly skew the resulting red/green channel ratios (R/G). These findings and our analysis methods are detailed elsewhere. (Timlin et al., 2005) Extraneous emissions on microarrays can arise from several sources such as the background substrate, dust, and additional fluorescent species that contaminate the slide during the array manufacturing, such as buffer solutions. The unique view offered by the HSS of all the emitting species motivated the EPA to employ our technology to perform quality control on two microarray platforms of interest to their research. These platforms were already under evaluation by the EPA using other methods to determine their fitness for large scale use in the EPA toxicology programs.

The two platforms investigated in this study were the Amersham Biosciences CodeLink Array platform and the Agilent in situ oligomer array technology. Details for each platform can be found at the following websites, respectively and will not be presented in depth here. (See References - URL 1, URL 2) Until recently DNA microarrays were typically printed using a robotic printer to place cDNA onto a glass slide substrate. Recent developments have produced several non-traditional platforms like the two referenced above for microarrays. These platforms have advantages over printed glass arrays, such as improved hybridization and stability and improved printing characteristics.

Sandia received microarrays from each of these platforms that had been hybridized for an experiment at the EPA to determine overall array quality and suitability using traditional microarray analysis methods. Before arriving at Sandia the commercial

microarray image data were assessed by the person performing the hybridization and each array was given a designation as "good or "bad" based on slide-to-slide variability and visual quality.

Experimental

Microarray slide details:

Amersham Biosciences CodeLink Arrays: Eleven slides containing ten arrays representing good and bad processing and one blank were sent to SNL in December 2004 by Ann Miracle (EPA, Cincinnati). These were test CodeLink arrays containing 125 30-mers representing fat head minnow genes printed in true 3x replicates. The arrays had been hybridized with DNA attached to Cy5 in a single-color experiment as directed by the manufacturer. The arrays were scanned at the EPA before being sent to SNL using their commercial microarray scanner and those supporting tiff image files were sent with the arrays. Once at SNL all arrays were entered into our sample database and stored in their original containers in a light free environment. Table 1 contains the details about the Amersham arrays imaged with the HSS.

Agilent Oligomer Arrays: Three slides containing 6 arrays representing good and bad processing were sent to SNL in late February 2005 by Barbara Carter (EcoArray, Inc. Florida). These were custom arrays containing 1000 sequences of 60-mers representing fat head minnow genes. There were two representations of each sequence on each array, but these were not true replicate sequences. The arrays had been hybridized in the two-color format (Cy3 & Cy5 DNA) as directed by the manufacturer and scanned prior to being sent to SNL at EcoArray, Inc. using their commercial microarray scanner. No supporting tiff images files were sent with the arrays, only a PowerPoint presentation. Once at SNL all arrays were entered into our sample database and stored in their original containers in a light free environment. Table 2 contains the details about the Agilent arrays imaged with the HSS.

Table 1: Amersham CodeLink Array Details

Array Identifier	EPA Quality Assessment
T00226259	Good
T00226147	Good
T00226240	Good
T00226142	Good
T00226138	Poor
T00226146	Poor
T00226258	Poor
T00226237	Poor
T00226139	Poor
T00226236	Poor
T00226233	Blank

Table 2: Agilent Oligomer Array Details

Array Identifier	Slide	EcoArray Quality Assessment
A	6	Poor
B	6	Good
E	8	Good
F	8	Good
G	11	Good
H	11	Poor

Hyperspectral scanning:

Amersham Biosciences CodeLink Arrays: The HSS configuration differed from that described in Sinclair et. al.(Sinclair et al., 2004) in that it was operating in its red excitation configuration using a 638 nm laser (CrystaLaser). In this configuration the power density at the sample is ~ 0.011 W/cm^2. A minimum of 2 areas on each array were scanned with the HSS at 10 μm (high) resolution, while the entire array was scanned at 30 μm (low) spatial resolution. Scanning was performed at speeds of 0.04 mm/sec and 0.12 mm/sec for the 10 μm and 30 μm resolutions, respectively and a 10x objective was used for both resolutions. A total of 23 hyperspectral images were collected amounting to 29 Gigabytes of spectral image data.

Agilent Oligomer Arrays: The HSS was operated in both the green excitation mode as described in Sinclair et. al.(Sinclair et al., 2004) as well as the red excitation mode described above for the Amersham arrays. A minimum of 2 areas on each array were scanned with the HSS at 10 μm (high) and 30 μm (low) spatial resolution for each of the excitation modes. Scanning was performed at speeds of 0.04 mm/sec and 0.12 mm/sec for the 10 μm and 30 μm resolutions, respectively and a 10x objective was used for both resolutions. A total of 26 red excitation and 20 green excitation hyperspectral images were collected amounting to 48 Gigabytes of spectral image data.

Instrument alignment was verified at the start of each day of data acquisition visually as well as by imaging a test sample of fluorescent beads. Instrument calibration files were collected at least once every day.

Data pretreatment and analysis:

Data pretreatment is critical to the success of the multivariate analysis routines. Prior to analysis all image data sets were preprocessed to remove cosmic events using a modified subregion median filter and to correct for image curvature (an optical aberration common to line-imaging systems) using in-house written algorithms. Image curvature correction was performed using calibration data taken on the same day as the images from neon and krypton gas discharge lamps. This same calibration data served to generate the wavelength calibration for the emission spectra. In addition, images generated with green laser excitation at 10 μm resolution were also preprocessed before multivariate analysis to correct for slight keystone distortion occurring within our spectrometer. The images acquired with the red laser did not require keystone correction due to the reduced spectral range covered.

To help illustrate the performance of our multivariate analysis methods, a simulated hyperspectral image data set from a simple 2 component microarray is shown

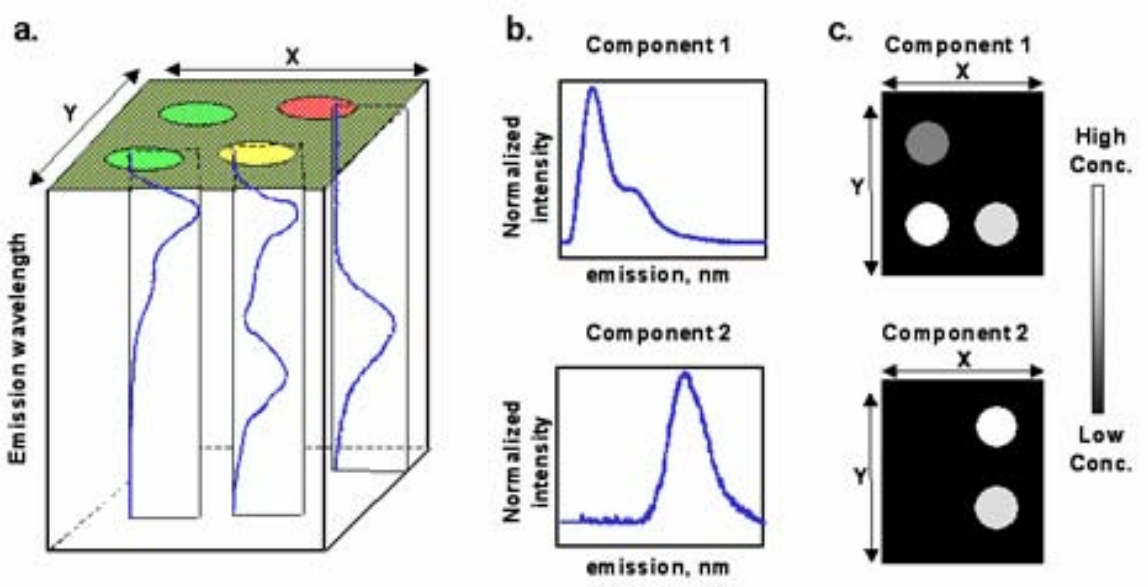

Figure 1: Illustration of hyperspectral data cube and multivariate analysis results. (a) Three-dimensional hyperspectral data cube for an idealized two-component microarray containing emission only from labeled DNA within the printed spots. Each pixel in the x-y image plane contains an entire fluorescence emission spectrum from 550–900 nm. (b-c) Results of multivariate curve resolution on two-component sample hyperspectral data cube shown in A. (b) Pure component spectra identify "what" species are fluorescing in an image. (c) Corresponding component concentration maps show "where" and "how much" of each component in Fig. 1b is present.

in Figure 1. The three-dimensional hyperspectral data contains an x-y image plane and a spectral, or wavelength, dimension. A PCA analysis is used to take an initial look at the hyperspectral data and determine the number of pure components present in the data. (not shown in Figure 1). The data representation is simplified using multivariate curve resolution (MCR), a factor analysis technique which uses correlations between variables to separate pure-component spectra and their corresponding concentration maps. (Figure 1b and 1c, respectively.) An advantage of the multivariate approach is that it has the powerful ability to separate highly overlapping spectral species as in Figure 1.

The multivariate image analysis strategy utilized with the EPA hyperspectral microarray data was to perform principal component analysis (PCA) on a data matrix weighted for Poisson noise characteristics of our instrument to determine the number of spectral components present followed by multivariate curve resolution (MCR) to identify the underlying pure component spectra and corresponding concentration maps. PCA and MCR analysis were performed on a single, representative region of interest (ROI) image from each array taken at high spatial resolution (10 μm). SNL's proprietary MCR algorithms which are based on a constrained iterative least squares analysis, were employed for the MCR analysis. Additional details of the MCR analysis is given elsewhere.(Haaland et al., 2003; Kotula et al., 2003; Timlin et al., 2005). In all MCR analyses a baseline offset was fit to the data using an equality constraint and non-negativity constraints were employed for all concentrations and spectra to reduce solution abiguity. The pure component spectra of the emitting species identified by MCR on these ROI's represent a model for the species present on that array. The model for each array was tested on all of the remaining pixels from that array using a weighted classical

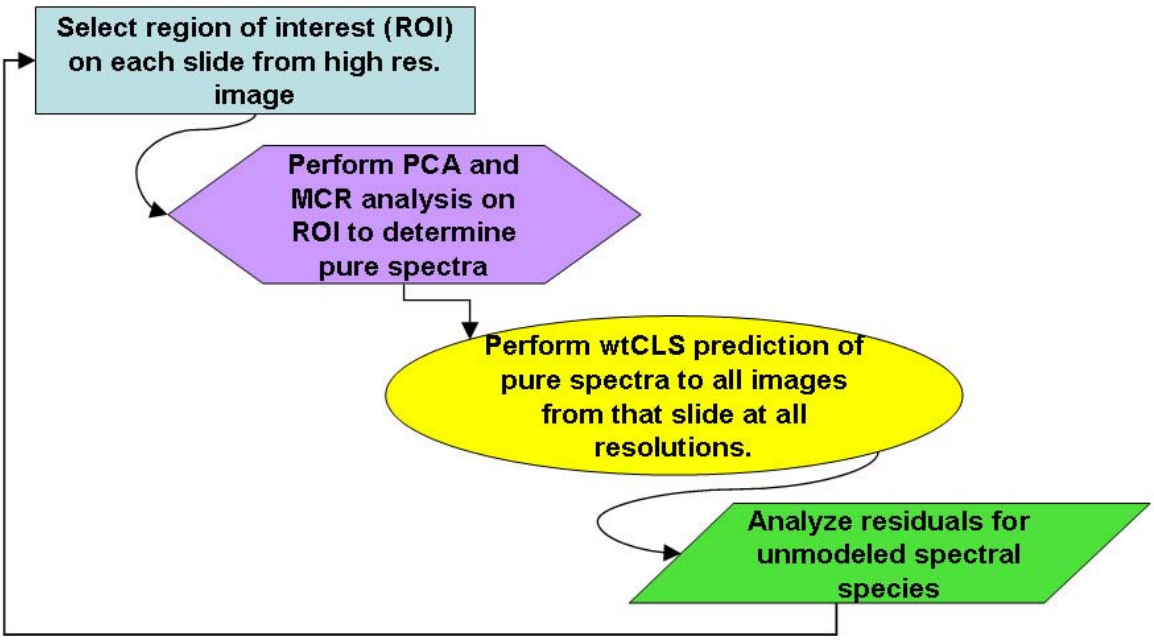

Figure 2: Flowchart of multivariate analysis procedure.

least squares (wtCLS) prediction algorithm and PCA was performed on the residuals to highlight any unmodeled spectral species. If an unmodeled spectral species was found, MCR analysis was performed on that image to identify the additional component. Resulting concentration images were reconstructed in a montage fashion from multiple adjacent images based on the instrument scan parameters. A flow chart of the analysis procedure is shown in Figure 2.

Results

The MCR extracted spectral components are shown in Figure 3 resulting from a representative ROI from a high resolution hyperspectral image of a "poor array" from each of the platforms. These extracted curves match well with the literature values for the dyes and our expectations based on previous experience with similar labels and glass substrates. There were no significant spectral residuals evident from the analysis of these or other high resolution images of poor or good classification that would indicate extraneous sources of emission not accounted for by the spectral model comprised of dye(s), glass, and a linear offset. A visual comparison of the extracted spectral species from different hyperspectral images indicates good correlation within data sets from the same platform regardless of array designation as poor or good. It would be redundant to include all images taken in this report. Instead sample HSS images are shown in Figures 4 and 5 for an array from each platform designated as poor based on observations made prior to hyperspectral imaging.

Figure 4 shows the resulting commercial scanner and HSS images of Array T00226139. The HSS images are generated from the MCR predicted concentration map corresponding to the Cy5 spectral component for two sample images at high spatial resolution and the wtCLS predicted concentration map corresponding to the Cy5 spectral component in an image acquired with low spatial resolution.

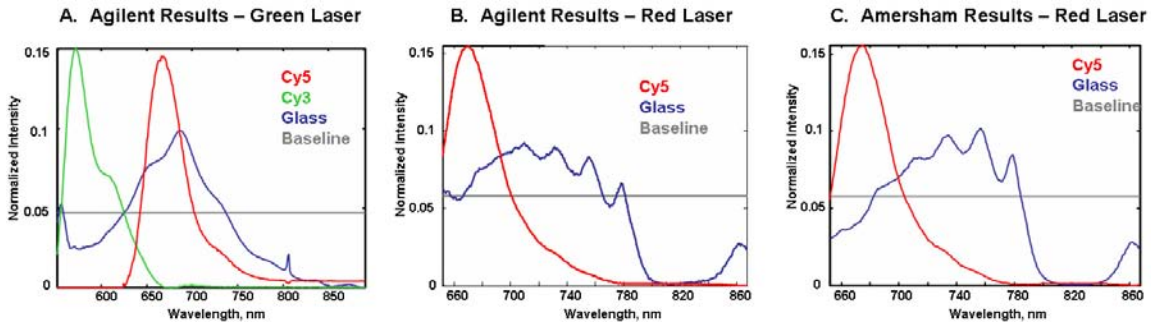

Figure 3: MCR extracted spectral components from representative high resolution hyperspectral images of poor quality arrays. A. Results from an Agilent oligomer microarray scanned with green laser excitation. B. Results from same Agilent oligomer microarray as in A. scanned with red laser excitation. C. Results from Amersham CodeLink microarray scanned with red laser excitation.

Figure 5 shows the resulting commercial scanner and HSS images of Array H. The HSS images are generated from either the MCR predicted concentration maps or the wtCLS predicted concentration maps corresponding to the Cy3 spectral component of green excitation images at high spatial resolution.

The hyperspectral images, both shown and not shown, match well (based on a qualitative assessment of intensity patterns of spots) with the commercial scanner images provided, indicating no significant degradation or other problems with the slides before scanning with the HSS.

Discussion

One of the benefits of using the HSS is that as a hyperspectral imaging system it is capable of using one color laser to excite all the dyes on the sample, thus negating the need for scanning with both colors as commercial array scanners do. When we do this we do not optimally excite the red fluorophores in the sample; however there is enough signal to quantify the Cy5 label often used for microarrays. The HSS was operated using red laser excitation for the Amersham CodeLink arrays because they contained only a red fluorophore (manufacturer's recommendation) and the red laser would provide maximal signal to noise for these arrays and excite the same set of constituents as the commercial scanner would. With the Agilent two-color arrays enough signal can be generated with just the green laser excitation to quantitate the Cy5 label; however it is possible that a red contaminant that would be excited with the red laser could be weak enough to be missed with our HSS using green excitation images were taken with both the red and the green excitation to provide a thorough look at the sample components and a direct comparison with the commercial scanner excitation.

The results of the hyperspectral image analysis (Figure 3-5 and additional data not shown) did not indicate the presence of any spectrally overlapping emission source such as a contaminant for any of the image areas investigated. In all images, the data could be adequately modeled with the dye or dyes, glass, and a baseline offset. The only spectral residuals observed during the analysis of the images were due to random noise, known optical effects of our imager (residual keystone distortion), and the occasional dust particle. It is possible that since the entire array was not scanned on every Agilent slide with the HSS there may be some additional emission species present on these slides, but

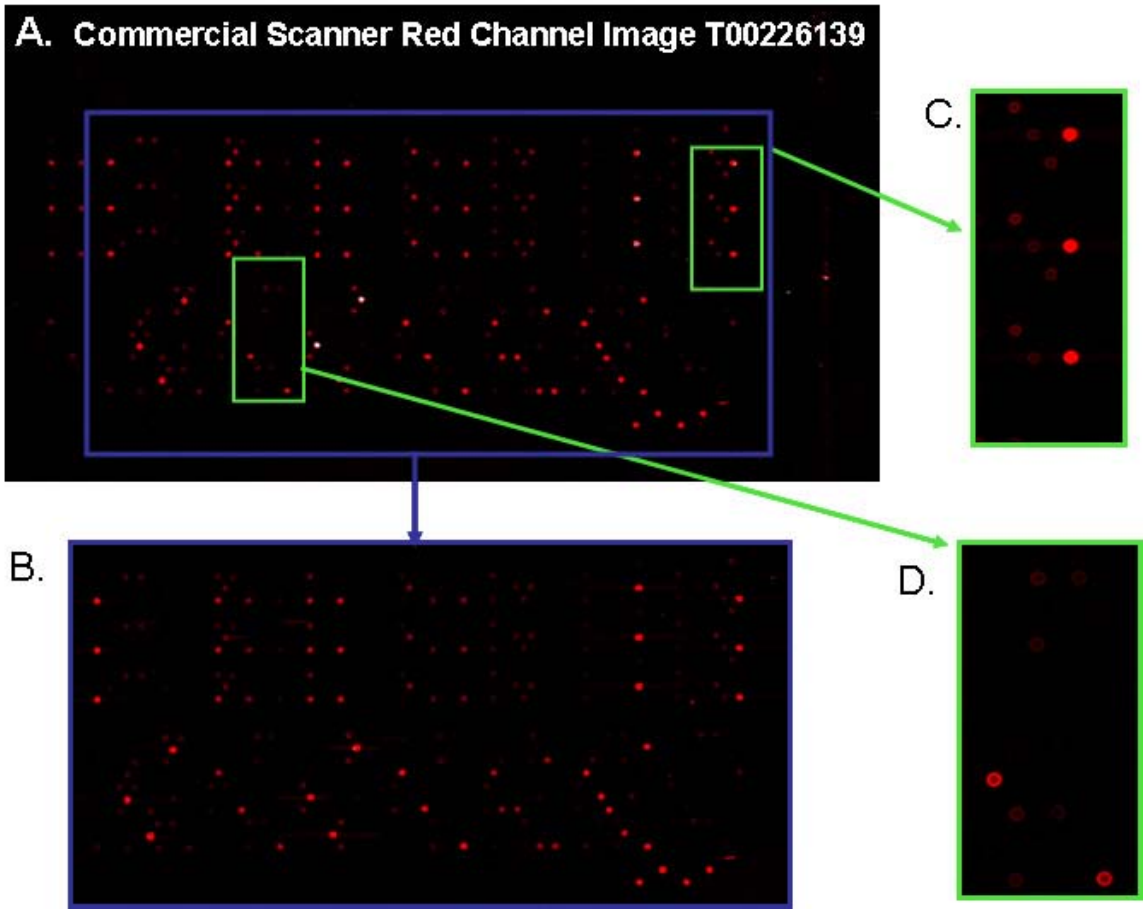

Figure 4: Images of Amersham CodeLink array #T00226139. A. Commercial scanner red channel image (provided by the EPA). Areas marked with colored rectangles represent regions scanned with HSS. B. Low resolution (30 μm) hyperspectral image generated from CLS predicted concentration map corresponding to Cy5 spectral component. C. & D. High resolution (10 μm) hyperspectral images generated from MCR extracted concentration maps corresponding to Cy5 spectral component.

this is highly unlikely. The areas scanned and analyzed were fairly representative and the possible contaminants are typically introduced during the hybridization process so they would be expected to be found throughout the slides, rather than in small, isolated areas.

As noted in both Table 1 and Table 2, the microarrays received had a designation of poor or good based on assessments of array quality made by the lab producing the microarrays prior to hyperspectral imaging. These designations were based on factors such as repeatability of hybridization or physical appearance and were estimated from a limited number of samples. In Figure 5A the large circular region of the image that appears to have reduced signal levels shows an example of poor quality based on physical appearance. The hyperspectral image results coupled with our previous experience with microarray quality control, would lead to the conclusion that the designation of poor on the slides investigated does not arise from the presence of a contaminant, but most likely other problems with microarray preparation such as hybridization mixing, etc. For example, previous investigations with other collaborators microarrays have shown the effect seen in Figure 5A to be from a large air void under the coverslip during

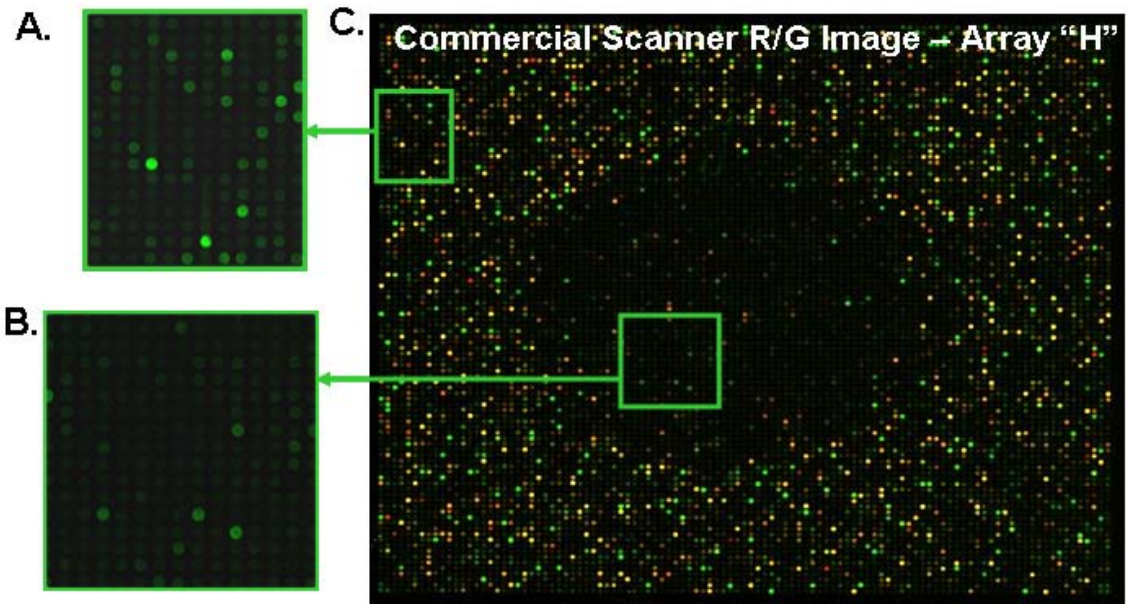

Figure 5: Images of Agilent oligomer array H (on slide #11). A. & B. High resolution (10 μm) hyperspectral images generated from MCR extracted or CLS predicted concentration maps corresponding to Cy3 spectral component. C. Commercial scanner red/green channel ratio image (provided by the EcoArray, Inc.). Areas marked with colored rectangles represent regions scanned with HSS.

hybridization. With appropriate investigation and subsequent optimization of the hybridization and microarray processing protocols these effects leading to poor designation should be able to be minimized. In addition to optimizing protocols, it is also important to have controls in place to discover slides where there are processing problems, and remove the affected outlier data from any statistical analysis.

Additionally it can be seen from Figure 3 that a slight spectral shift is present when comparing the Cy5 spectra recovered from the MCR analysis of images from the Agilent versus Amersham platforms. With the Agilent images the Cy5 emission maximum is ~670 nm regardless of red or green excitation, whereas the Cy5 emission maximum is ~678 nm for the data from the Amersham platform. This was consistent for all images investigated. The spectral properties of the Cy dyes are known to shift based on the surroundings of the dye and its attachments following a trend of the larger the oligomer, piece of genetic material, or molecule attached to the Cy dye, the more red the emission maxima. (unpublished data- Amersham BioSciences website) The HSS is very sensitive to these small changes in the spectral data.(Timlin et al., 2005) These two platforms incorporate the Cy5 into the DNA or RNA using different linkages thus creating the observed differences in emission maximum.

Conclusions and Recommendations

We conclude that there were no extraneous emissions present on either of the two microarray platforms investigated during this study. The impact of this is that either of these platforms should be capable of producing high quality repeatable and reliable microarray data with appropriate controls in place for optimizing and monitoring the microarray process. In our previous research the most commonly observed green

contaminant was dependant on microarray processing, including buffer solutions and exposure before printing time(Martinez et al., 2003). Because of this dependency, we recommend that if the microarray preparation and/or hybridization procedures are changed the arrays should be reevaluated for problems related to extraneous emissions. Changes that would warrant reevaluation would be a change in buffer solution or solution manufacturer, a change in label used, a change in laboratory preparing/hybridizing the arrays, or a significant change in protocols such as waiting a much longer time before or after a step, additional or reduced washing, etc.

Although there is no clear advantage of one platform over the other in terms of contaminant emissions, it should be noted that the Amersham CodeLink array platform is designed and marketed only as a single color array (red). During our investigation it became apparent that the substrate exhibit strong green fluorescence if excited with a green laser and thus this would prevent one from using it in a two-color format with a commercial scanner. If having the two-color format is important for the toxicology studies at the EPA we would not recommend the Amersham arrays unless one were to utilize the HSS and use two overlapping red labels.

References

Haaland, D. M., Timlin, J. A., Sinclair, M. B., Van Benthem, M. H., Martinez, M. J., Aragon, A. D., and Werner-Washburne, M. (2003). Multivariate curve resolution for hyperspectral image analysis: Applications to microarray technology. *In* "Spectral Imaging: Instrumentation, Applications, and Analysis", Vol. 4959, pp. 55-66. International Society for Optical Engineering, San Jose, CA.

Kotula, P. G., Keenan, M. R., and Michael, J. R. (2003). Automated analysis of SEM X-Ray spectral images: a powerful new microanalysis tool. *Microscopy & Microanalysis* 9, 1-17.

Martinez, M. J., Aragon, A. D., Rodriguez, A. L., Weber, J. M., Timlin, J. A., Sinclair, M. B., Haaland, D. M., and Werner-Washburne, M. (2003). Identification and removal of contaminating fluorescence from commercial and in-house printed DNA microarrays. *Nucleic Acids Research* 31, e18.

Sinclair, M. B., Timlin, J. A., Haaland, D. M., and Werner-Washburne, M. (2004). Design, construction, characterization, and application of a hyperspectral microarray scanner. *Applied Optics* 43, 2079-2089.

Timlin, J. A., Sinclair, M. B., Haaland, D. M., Aragon, A. D., Martinez, M. J., and Werner-Washburne, M. (2005). Hyperspectral microarray scanning: Impact on the accuracy and reliability of gene expression data. *BMC Genomics* 6.

URL 1
http://www4.amershambiosciences.com/aptrix/upp01077.nsf/Content/Products?OpenDocument&parentid=568694&moduleid=165695#content)

URL 2

http://www.chem.agilent.com/scripts/generic.asp?lPage=10408&indcol=N&prodcol=Y)

Distribution

External:
Rong-Lin Wang
US EPA
26 W. Martin Luther King Dr., MS 525
Cincinnati, OH 45268

Greg Toth
US EPA
26 W. Martin Luther King Dr., MS 525
Cincinnati, OH 45268

Internal:

6	MS 0895	Jerilyn Timlin
1	MS 0895	Anthony Martino
1	MS 0895	David Haaland
1	MS 0895	Howland Jones
1	MS 1411	Michael Sinclair
1	MS 1413	Linda Nieman
1	MS 1413	Grant Heffelfinger
1	MS 0886	Chris Stork
2	MS 9018	Central Technical Files, 8945-1
2	MS0899	Technical Library, 4536