March 13, 2007
External Review Draft

U.S. Environmental Protection Agency
DRAFT

Interim Guidance for Microarray-Based Assays: Data Submission, Quality, Analysis, Management, and Training Considerations

Prepared for the U.S. Environmental Protection Agency by Members of the Genomics Workgroup, a Group Tasked by EPA's Science Policy Council

Science Policy Council
U.S. Environmental Protection Agency
Washington, DC 20460

DISCLAIMER

This draft interim guidance, when finalized, will represent EPA's current thinking on this topic. It does not create or confer any legal rights for or on any person or operate to bind the public. The use of any mandatory language in this document is intended to describe laws of nature, scientific principles, or technical requirements and is not intended to impose any legally enforceable rights or obligations. Alternative approaches may be used if the approach satisfies the requirements of the applicable statutes and regulations. If you would like to discuss an alternative approach (you are not required to do so), you may contact the EPA staff responsible for implementing this guidance. Mention of trade names or commercial products does not constitute endorsement of recommendation for use.

Note: This is an external review draft, and is not approved for final publication.

Genomics Microarray Workgroup
Co-Chairs

William H. Benson
Office of Research and Development

Kathryn Gallagher
Office of the Science Advisor

J. Thomas McClintock
Office of Pollution Prevention and Toxics

Kerry Dearfield
Office of the Science Advisor
(2004 - June 2005)

Science Policy Council Staff

Jeremy Johnson
(2004)

Subgroup Co-Chairs

Performance Approach to Quality Assessment
David Lattier, ORD
Susan Lundquist, OEI

Data Management
Susan Hester, ORD
Joseph Retzer, OEI

Data Submission
Greg Miller, OPEI
Doug Wolf, ORD

Training
Bobbye Smith, Region 9 RSL
Julian Preston, ORD

Data Analysis
David Dix, ORD
Brenda Groskinsky, Region 7 RSL

Microbial Source Tracking
Jorge Santo Domingo, ORD
Ron Landy Region 3 RSL

Additional Coordinating Committee Members

Wafa Harrouk, US FDA
Lee Hofmann, OSWER
Robert Kavlock, ORD
Rita Schoeny, OW

Genomics Workgroup Lead for the Science Policy Council

Larry Reiter
Office of Research and Development

Genomics Microarray Workgroup Members

Gregory Akerman, OPPTS
Wenjun Bao, ORD
David Bencic, ORD
Lynn Bradley, OEI
Kevin Cavanaugh, ORD
Barbara Collins, ORD
Brion Cook, OPPTS
Don Delker, ORD
Michelle Embry, OPPTS
Robin Gonzalez, OEI
Susan Griffin, Region 8
Stephanie Harris, Region 10
Belinda Hawkins, ORD
Kenneth Haymes, OPPTS
Michael Hemmer, ORD
Todd Holdermann, OPPTS
Gene Hsu, ORD
Margo Hunt, OEI
Sid Hunter, ORD
Channa Keshava, ORD
Steven Kueberuwa, OW
Mitch Kostich, ORD
Richard Leukroth, OPPTS
Nancy McCarroll, OPPTS
Jesse Meiller, OPPTS
Elizabeth Mendez, OPPTS
Ann Miracle, ORD
Ines Pagan, ORD
Santhini Ramasamy, OPPTS
Ann Richard, ORD
Mitch Rosen, ORD
Phil Sayre, OPPTS
Judy Schmid, ORD
John Sykes, ORD
Freshteh Toghrol, OPPTS
Mark Townsend, OPPTS
Nancy Wentworth, OEI
Lori White, ORD
Witold Winnik, ORD
Steve Young, OEI

Additional Genomics Resources

Genomics Training Workgroup Members

Barbara Abbott, ORD
Gilberto Alvarez, Region 5
Michele Burgess, OSWER
Michelle Embry, OPPTS
Audrey Galizia, ORD
Karen Hamernik, OPPTS
Steven Kueberuwa, OW
David Lattier, ORD
David Lee, ORD
Roseanne Lorenzana, Region 10
Marian Olsen, Region 2
Jennifer Seed, OPPTS

Microbial Source Tracking Workgroup Members

Bobbye Smith, Region 9
Jafrul Hasan, OW
James Goodrich, ORD
Rita Schoeny, OW
Robin Oshiro, OW
Roland Hemmett, Region 2
Sally Gutierrez, ORD

Table of Contents

ACRONYMS

CBI	Confidential Business Information
cDNA	Complementary Deoxyribonucleic Acid
CEBS	Chemical Effects in Biological Systems knowledgebase
cRNA	Complementary Ribonucleic Acid
CWA	Clean Water Act
DER	Data Evaluation Record
DNA	Deoxyribonucleic Acid
DQO	Data Quality Objective
EPA	Environmental Protection Agency
FACS	Fluorescence Activated Cell Sorter
FDA	Food and Drug Administration
FNR	False Negative Rate
FPR	False Positive Rate
gDER	Genomics Data Evaluation Record
HPV	High Production Volume
IRB	Institutional Review Board
IVT	*In Vitro* Transcription
JPEG	Joint Photographic Experts Group
MAGE	Microarray And Gene Expression
MAGE-OM	Microarray And Gene Expression - Object Model
MGED	Microarray Gene Expression Data
MIAME	Minimal Information About Microarray Experiments
MOA	Mode of Action
MOPS-EDTA	[MOPS] 3-(N-Morpholino) propanesulfonic acid], [EDTA] ethylenediaminetetraacetic acid
MPSS	Massively Parallel Signature Sequencing
mRNA	Messenger RNA
MQO	Measurement Quality Objective
MST	Microbial Source Tracking
NHEERL	National Health and Environmental Effects Research Laboratory
NIEHS	National Institute of Environmental Health Sciences
NPDES	National Pollutant Discharge Elimination System
OEI	Office of Environmental Information
OPPTS	Office of Prevention, Pesticides and Toxic Substances
ORD	Office of Research and Development
OSWER	Office of Solid Waste and Emergency Response
OW	Office of Water
PCR	Polymerase Chain Reaction
PMN	Pre-Manufacture Notification
PMT	Photomultiplier Tube
QA	Quality Assurance
QAARWP	Quality Assurance Annual Report and Work Plan
QC	Quality Control

QMP	Quality Management Plan
qPCR	Quantitative Polymerase Chain Reaction
qRT-PCR	Quantitative Reverse Transcriptase PCR
RFU	Relative Fluorescent Unit
RNA	Ribonucleic Acid
RNase	Ribonuclease
RTP	Research Triangle Park
RT-PCR	Reverse-Transcription Polymerase Chain Reaction
SAGE	Serial Analysis of Gene Expression
SNP	Single Nucleotide Polymorphism
SOPs	Standard Operating Procedures
SPC	Science Policy Council
TIFF	Tagged Image File Format
TMDL	Total Maximum Daily Load
U.S. EPA	U.S. Environmental Protection Agency

EXECUTIVE SUMMARY

The mapping of diverse animal, plant, and microbial species genomes using molecular technologies has significantly affected research across all areas of the life sciences. The current understanding of biological systems is rapidly changing in ways previously unimagined and novel applications of this technology have already been commercialized. These advances in genomics will have significant implications for risk assessment policies and regulatory decision making. In 2002, the U.S. Environmental Protection Agency (EPA or "the Agency") issued its Interim Policy on Genomics (U.S. EPA, 2002a) that communicated the Agency's initial approach to using genomics information in risk assessment and decision making. The Interim Policy described genomics as the study of all the genes of a cell or tissue, at the DNA (genotype), mRNA (transcriptome), or protein (proteome) level. While noting that the understanding of genomics is far from established, the Agency stated that such data may be considered in the decision making process, but that these data alone are insufficient as a basis for decisions.

Following the release of the Interim Policy, the Science Policy Council (SPC) created a cross-EPA Genomics Task Force and charged it with examining the broader implications genomics is likely to have on Agency programs and policies. The Genomics Task Force developed a Genomics White Paper entitled "Potential Implications of Genomics for Regulatory and Risk Assessment Applications at EPA" (U.S. EPA, 2004). That document identified four areas likely to be influenced by the generation of genomics information within EPA and the submission of such information to EPA: 1) prioritization of contaminants and contaminated sites, 2) monitoring, 3) reporting provisions; and 4) risk assessment. One critical need in the area of technical development was identified: the need to establish a framework for analysis and acceptance criteria for genomics information for scientific and regulatory purposes. The Task Force recommended that the Agency charge a workgroup to establish such a framework and in doing so consider the performance of assays across genomic platforms (*e.g.*, reproducibility, sensitivity, pathway analysis tools) and the criteria for accepting genomics data for use in a risk assessment *(e.g.*, assay validity, biologically meaningful response).

In 2004, the Genomics Technical Framework and Training Workgroup was formed with the responsibility to ensure that the technical framework and training activities build upon the Agency's Interim Policy on Genomics while continuing to engage other interested parties. Information developed by these workgroups will be used by EPA program offices and regions to determine the applicability of specific genomics information to the evaluation of risks under various statutes.

To this end, the Genomics Technical Workgroup considered all of the "omics" technologies and applications and decided that an interim guidance document on the use of data generated by DNA microarray technology would be most beneficial to the Agency and regulated community at this time. Consequently, this document provides recommendations regarding: 1) data that should be considered for submission to the Agency for microarray studies, 2) the use of a performance approach to microarray quality assessment parameters, 3) data analysis approaches for microarrays, and 4) data management and storage issues for microarray data submitted to or used by the Agency. The guidance applies to both human health and ecological DNA microarray data.

With respect to experimental performance considerations, the Genomics Workgroup concluded that quality issues are critical considerations in the application of new technologies such as genomics. The Genomics Workgroup recommends that the Agency not prescribe specific methods to be used in microarray experiments at this time, but instead provide general guidance on the recommended performance of microarray experiments in order to obtain data of the quality required for a specific use; this guidance is provided herein. Investigators submitting data to the Agency in support of regulatory decision making, methods development, and technical transfer, may want to consider, in addition to compliance with MIAME (Minimal Information About Microarray Experiments) Workgroup standards (http://www.mged.org/Workgroups/MIAME/miame.html), the performance-related experimental and system factors outlined in this document (Appendix A). Further activities on the part of investigators to address experimental performance issues will serve to strengthen scientific arguments and experimental claims.

This document also provides information regarding submission of microarray data to EPA to ensure appropriate review and consistent evaluation of data from multiple sources. In accordance with accepted practice, it is recommended that submissions include sufficient information to allow an independent reviewer to reconstruct how the data were collected and analyzed. This approach allows reviewers to judge the quality of the data and the strength of any conclusions. Many scientific journal editors grappling with these issues have adopted the MIAME guidelines as a standard for submission of microarray data as part of a submitted publication. A slightly modified version of MIAME is proposed as the microarray data submission template for EPA; this submission template will be subject to change as the technology evolves.

With regard to data analysis, the Genomics Workgroup concluded that a systematic approach for genomics data evaluation is necessary for the further use of such data in risk assessments. A genomics Data Evaluation Record template is provided herein as a way to present and organize data from genomics studies in order to derive information necessary for a regulatory application (see Appendix C for the Genomics Data Evaluation Record [DER]). A completed sample DER is also provided in Appendix D to facilitate the use of the template. An overview of issues to be considered in analyzing microarray data is also provided. The transfer of these evaluations, and the underlying genomics data, into searchable, electronic databases will be essential to making the data useful in risk assessments. Furthermore, development of databases containing gene expression profiles for a wide variety of chemicals should facilitate creation of statistical/computational methods that will help predict the toxic potential of a chemical.

Due to potentially large volumes of genomic and associated toxicological data, it is essential that the Agency consider the development of a complete data management solution. The functional needs of a solution of this magnitude would minimally include items listed in the section on data management. In addition, this Agency data management solution should address needs unique to scientifically-based risk assessments, confidential and proprietary data security, public access, and other aspects of regulatory application. It should be noted that consistency, scientific and operational robustness, common access, and availability in a scalable environment

are data management needs for an Agency data management solution. While the Agency has begun to utilize bioinformatics research approaches, both intramurally (e.g., the National Center for Computational Toxicology in EPA's Office of Research and Development [ORD]) and extramurally (Environmental Bioinformatics Centers in North Carolina and New Jersey funded by EPA's Science to Achieve Results (STAR) Program), an Agency-wide data management solution integrating genomics, toxicological, and other key data required for regulatory applications is now necessary.

The document concludes with the Genomics Workgroup's recommendations to the Agency for follow-up activities to this interim guidance including: 1) further development of the outlined training materials and modules, to be offered throughout the Agency to risk assessors and decision makers who will be faced with the challenge of interpreting and applying genomics information, 2) continued collaboration of EPA personnel with staff from other federal agencies and stakeholders in the development of tools for the analysis of genomics data, 3) application of this guidance to a series of case studies to evaluate its utility in risk assessment and regulatory applications; and 4) the updating of this guidance as needed as the technology evolves.

This document is intended to provide information to the regulated community and other interested parties regarding submitting microarray data to the Agency and to provide guidance for EPA reviewers in evaluating such data and/or information. This interim guidance can be used by EPA program offices to determine the applicability of specific genomics information to the evaluation of chemical risks.

1.0 Introduction

1.1 Background

The mapping of diverse animal, plant, and microbial species genomes using molecular technologies has significantly affected research across all areas of the life sciences. The current understanding of biological systems is rapidly changing in ways previously unimagined and novel applications of this technology have already been commercialized. These scientific and technological advances have spurred many federal agencies to consider the far-reaching implications for policy, regulation, and society as a whole.

In 2002, EPA released the Interim Policy on Genomics (U.S. EPA, 2002a) communicating its initial approach to using genomics information in risk assessment and decision making (http://www.epa.gov/osa/spc/genomics.htm). This policy describes genomics as the study of all the genes of a cell or tissue, at the DNA (genotype), mRNA (transcriptome), or protein (proteome) level. The Interim Policy notes that while genomics offers the opportunity to understand how an organism responds at the gene expression level to stressors in the environment, understanding such molecular events with respect to adverse ecological and/or human health outcomes is far from established. This policy states that while genomics data may be considered in the decision making process at this time, these data alone are insufficient as a basis for decisions. Consequently, currently EPA will only consider genomics information for assessment purposes on a case-by-case basis.

Following the release of the Interim Policy, the Science Policy Council (SPC) created a cross-EPA Genomics Task Force and charged it with examining the broader implications genomics is likely to have on Agency programs and policies. To that end, the Genomics Task Force developed a Genomics White Paper entitled "Potential Implications of Genomics for Regulatory and Risk Assessment Applications at EPA" (USEPA, 2004, www.epa.gov/osa/genomics.htm). The Task Force identified scenarios to describe various circumstances under which EPA might receive these data. Four areas were identified as those

likely to be influenced by the generation of genomics information within EPA and the submission of such information to EPA: 1) prioritization of contaminants and contaminated sites, 2) monitoring, 3) reporting provisions; and 4) risk assessment. The Task Force also identified several challenges and/or critical needs that included research, technical development, and capacity (*i.e.*, strategic hiring practices and training).

The Genomics Task Force recommended that the Agency charge a workgroup with developing a technical framework for analysis and acceptance criteria for genomics information for scientific and regulatory purposes. The Genomics White Paper identified issues that need to be considered in developing such a framework including the performance of assays across genomic platforms (*e.g.*, reproducibility, sensitivity, pathway analysis tools) and the criteria for accepting genomics data for use in a risk assessment *(e.g.*, assay validity, biologically meaningful response).

In June, 2004, the Genomics Technical Framework and Training Workgroup was established with representatives from ORD, numerous program offices (OPPTS, OSWER, OW, OEI, OPEI) and regional offices (2, 3, 5, 7, 8, and 9). The Genomics Workgroup was comprised of a Coordinating Committee, several technical genomics guidance workgroups (Performance Approach Quality Assurance Workgroup, Data Submission Workgroup, Data Analysis Workgroup, and a Data Management and Storage Workgroup), a Training Workgroup, and a Microbial Source Tracking Workgroup. The Genomics Workgroup's responsibility was to ensure that the technical framework and training activities build upon the Agency's Interim Policy on Genomics while continuing to engage other interested parties. This document will be used by EPA program offices and regions to determine the applicability of specific genomics information to the evaluation of risks under various statutes.

1.2 Overview of Genomic Science

As a means of introduction to genomics and its potential impact on regulatory decision making, it is important to understand the basic principles behind genomic technology. Only about 1-2% of the human DNA actually codes for RNA that can be translated into proteins. This

1-2% is considered to be the theoretical functional genome. Any particular cell type (*i.e.*, from various organs or species) will have its own practical functional genome, which is a subset of the entire functional genome that encodes for functional proteins in that cell. The functional genome for any cell type can be assessed by determining the messenger RNA (mRNA) profile of the cell, tissue, or organ. The mRNA copies the necessary portion of the cell's DNA code and transports this information to the ribosomes where protein synthesis occurs. Thus, the assessment of mRNA profiles is called functional genomics. Such profiles are constructed using microarrays that contain all (or a sampling) of a cell's functional genome. Hybridization of a DNA copy (cDNA) of the mRNA that is being actively produced by the cell to these microarrays demonstrates which genes are currently active in that cell. Within the 98-99% of DNA not coding for RNA message is information that affects the activity of the functional genome by influencing where and when genes are active in an organism. Thus both coding and noncoding DNA are important in organismal function and response to perturbations.

The study of a cell's protein composition is called proteomics. Currently, it is possible to analyze only a fraction of a cell's proteins, but rapid advances in this field will allow more complete profiling in the near future. Another discipline of biology analyzes biofluids and tissues to determine the profiles of endogenous metabolites present under normal conditions or when the organism has been affected by factors such as exposure to environmental chemicals. This type of whole cell analysis is called metabolomics (or metabolic profiling). In order to understand how a cell functions under normal or stressed circumstances, it is necessary to characterize the proteins that are manufactured by the cell, as well as endogenous metabolites. This facilitates an understanding of global metabolism and how proteins interact along biochemical pathways. This approach describes the area of systems biology, in which the cell, tissue, or organism is considered as a complete, albeit complex, system.

Broadly defined, genomics tools provide the means to examine changes in gene expression, protein, and metabolite profiles within the cells and tissues, in contrast to current risk assessment methods which are restricted to whole organism effects or changes in single biochemical pathways. Genomics tools have the potential to provide detailed data about the underlying biochemical mechanisms of disease or toxicity (*i.e.*, disease etiology, biochemical

pathways), sensitive measures of exposures to chemicals, new approaches to detecting effects of such exposures, and methods for predicting genetic predispositions that may possibly lead to disease or higher sensitivity to particular stressors in the environment.

Another type of application is chemical identification. By utilizing genomic expression profiles it is possible to identify and classify environmental contaminants. For example, Hamadeh *et al.* (2002a,b) found chemical-specific gene expression profiles in liver tissue of exposed rats. The authors demonstrated that 24-hour exposure to compounds from the same chemical class (peroxisome proliferators) resulted in gene expression profiles that were unique but more similar to each other than to patterns corresponding to exposure to a chemical of a different class (enzyme inducers). These gene expression profiles were associated with differences in histopathology between the different chemical classes following longer durations These and other published works indicate the utility of genomic approaches in chemical identification and in investigations of mode-of-action of chemical hazards.

1.3 Emerging Impacts of Genomics Technologies

Toxicology has been moving from observation of changes in tissue histology, physiology, and chemistry to a mechanistic understanding through assessment of large scale changes of gene activity within those tissues. Identification of changes in gene expression using microarrays is becoming an important tool for informing our understanding of toxicological processes as well as informing the hazard identification process and mode of action analysis as part of safety and risk assessment. As the price of conducting microarray experiments declines and an appreciation of their value increases, their use for basic research and as part of the environmental regulatory process is likely to increase.

The use of data generated by microarray technology in peer reviewed scientific publications has grown exponentially over the last few years. Microarray technology allows monitoring of changes in gene expression across thousands of genes, or even entire genomes or proteomes in response to experimentally manipulated or natural conditions. We are now

beginning to understand several important toxicological processes in terms of changes in the activity of single genes or ensembles of genes acting in concert. The identification of these changes is increasingly the product of the use of microarray technology. As a result of these research trends, EPA anticipates receiving increasing volumes of microarray data from environmental researchers, and as a part of the regulatory process. In order to ensure optimal utilization of these data, EPA has developed this guidance to address the quality, submission, analysis, and storage of microarray data.

While many new genomic technologies do exist, most are not as yet ready for application in risk/safety assessment and decision making. Therefore, it is important for the Agency to consider how these genomic technologies might be incorporated into existing programs. It should be noted that genomics will not fundamentally alter the risk assessment process, but is expected to serve as a powerful tool for evaluating the exposure to and effects of environmental stressors and will offer a means to simultaneously examine a number of response pathways. EPA and other regulatory agencies are beginning to address the use of genomics data for various risk assessment applications, including the need to establish a link between genomic alterations and adverse outcomes of regulatory concern. Given the rapidly evolving nature of genomics technologies, care should be taken to develop an acceptable scheme to simplify and refine the risk-related information and to distinguish it from the large amount of complex scientific and statistical data available. This strategy should remain dynamic and fluid in anticipation of continuing technical evolution at the molecular levels (*e.g.*, DNA, RNA, and protein levels). Furthermore, bioinformatic approaches for data acquisition and analysis, including technologies designed to store and analyze the profusion of data generated from microarray analyses, should be considered in parallel with the data generating methods. Finally, many scientific, policy, ethical, and legal concerns developing along with the emergence of this science will need to be addressed.

The Interim Policy on Genomics provides guidance concerning how and when genomics information should be used to assess the risks of environmental contaminants under the various regulatory programs implemented by the Agency at the present time. The standardization of experimental design, the selection of informative biomarkers, and data analysis for genomics is

important for the utility of genomics information in future risk assessment and regulatory decisions. Such standardization will enhance the reproducibility of results obtained and the reliability of conclusions drawn from microarray data. Furthermore, EPA is considering the development of data quality standards based on performance of microarrays, as well as other genomics technologies (*e.g.,* functional genomics). This in turn will help to ensure the integrity of EPA's approach to assessing the genomics information submitted to the Agency.

Genomics issues have already arisen in environmental decision-making. For example, a pesticide registrant has cited a published genomic article (Genter et al., 2002) as part of the data package submission for product registration to EPA's Office of Pesticide Programs. The data were submitted in support of an alternative mode of action that would affect human health assessment conclusions. Similar submissions are quite likely to be made by other pesticide registrants.

Although this document focuses on the use of microarrays for toxicological studies as they pertain to macroorganisms, it should be noted that the impact of microarray technologies goes beyond the exploration of toxicological effects in eukaryotic systems. For example, the use of microarray techniques in environmental and clinical microbiology has increased significantly in the last few years. Microarrays can also be used to screen for host specific markers that can be used in microbial source tracking (MST). As an example of the application of genomics to MST, a research consortium including State of California regulatory agencies, public utilities, and EPA recently participated in a study comparing the performance of various genomics-based methods designed to identify the source of fecal material in ambient waters in an MST approach (Griffith *et al*., 2003). Moreover, genomics methods are being evaluated to assist dischargers in complying with Clean Water Act (CWA) requirements to develop Total Maximum Daily Loads (TMDLs) for water bodies that are listed as impaired due to the presence of fecal coliforms. This MST work will also address the issue of beach closures; current microbial methods require several days to complete and do not distinguish between bacteria from humans and other sources such as sea gulls or seals. Further details on these MST efforts are described in *Microbial Source Tracking Guide Document* (available at: http://www.epa.gov/ORD/NRMRL/pubs/600r05064/600r05064.htm ; U.S. EPA, 2005) .

These examples indicate the need to make proactive policy decisions and to develop processes to address how genomics data will be used in Agency decision-making.

1.4 Purpose and Intent of this Document

As a result of research trends, EPA anticipates receiving increasing volumes of microarray data from environmental researchers, and as a part of the regulatory process. The Genomics Technical Workgroup considered all of the "omics" technologies and applications and decided that a guidance document on the use of data generated by DNA microarray analysis would be most beneficial to the Agency and regulated community at this time. This guidance applies to microarray data relevant to human health and ecological risk assessment and decision making. This guidance is provided in order to facilitate appropriate submission, consistent review, and optimal utilization of these data. Consequently, this document provides recommendations regarding: 1) data that should be considered for submission to the Agency for microarray studies, 2) the use of a performance approach to microarray quality assessment parameters, 3) data analysis approaches for microarrays, and 4) data management and storage issues for microarray data submitted to or used by the Agency.

The purpose of this document is to provide information to the regulated community and other interested parties regarding submitting microarray data to the Agency and to provide guidance for reviewers in evaluating and utilizing such data and/or information. This interim guidance can be used by EPA program offices to determine the applicability of specific genomics information to the evaluation of chemical risks. It is important to note that microarray technology is rapidly changing, such that methodologies for generating such data and ensuring its quality will likely change; however the need to ensure consistency and quality in generating, analyzing and using the data will not. As the state of the science develops, EPA plans to revisit the guidance as necessary.

2.0 The Performance Approach to Quality Assurance for Microarrays

Quality issues are critical considerations in the application of new technologies or approaches, such as genomics. The Workgroup recommends that the Agency not prescribe specific methods to be used in microarray experiments at this time. This section instead provides general guidance on the recommended performance of microarray experiments in order to obtain data of the quality needed for a specific use.

The Agency acknowledges that continued advancement of tools and platforms for describing biological phenomena will be pivotal in supporting claims for regulatory decision making. It is also noted that at this time there exist numerous approaches, investigator fabricated and commercially available platforms, hardware and other peripheral equipment by which to measure biologic trends and changes at the level of tissues and cells. The following technical statements relate primarily to "*expression*" measurements (up- and down-regulation of macromolecules) and certain other multiplex technologies used to generate and collect quantitative and qualitative data about changing biologic conditions. This guidance is also relevant to the evolving nature of "*expression*" measures, particularly as recommendations for standardization in experimental performance put forth by the combined efforts of academic, industry and government scientists, become universally accepted and applied.

Although there are currently numerous means by which to observe and acquire biological expression measurements, such as *Massively Parallel Signature Sequencing* (MPSS) and *Serial Analysis of Gene Expression* (SAGE), the most frequently used experimental approach to collecting expression data is microarray-based studies. This technology, which has expanded well beyond the sphere of human health, is exploited to describe changing transcriptional profiles in genes of countless species that are important to numerous areas of biological sciences. Unfortunately, many of these investigations are undertaken without the benefit of explicit consensus for quality assurance and quality control and there has yet to be firmly established criteria for intra-experimental and cross-platform performance evaluation.

Investigators submitting data to the Agency in support of regulatory decision-making, methods development, and technical transfer, should also consider at a minimum the performance-related experimental and system factors outlined in Appendix A, in addition to compliance with MIAME (Minimal Information About Microarray Experiments) Workgroup standards (http://www.mged.org/Workgroups/MIAME/miame.html) discussed in Section 3 below. Further activities on the part of investigators to address experimental performance issues will serve to strengthen scientific arguments and experimental claims.

Each EPA program, regional, or research and development office's Quality System should be defined and documented in their Quality Management Plan (QMP). A summary of their individual office's Quality System activities is detailed in a Quality Assurance Annual Report and Work Plan (QAARWP), which also includes information on their annual internal assessment of their Quality System.

Additional detailed discussion of the EPA Quality System and the performance approach to quality assurance for microarrays is provided in Appendix A.

3.0 Data Submission Guidance

3.1 Introduction

EPA developed the following information regarding submission of microarray data to facilitate appropriate review and consistent evaluation of data from multiple sources. The text that follows was written as a preliminary template guiding the submission of microarray data to the EPA. As the state of the science develops, EPA plans to revisit this submission format as necessary. In accordance with accepted practice, it is useful if submissions include sufficient information to allow an independent reviewer to reconstruct how the data were collected and analyzed. This approach allows reviewers to judge the quality of the data and the strength of any conclusions. It is also useful if the submission includes enough information in a format that facilitates comparison or integration with similar data from other experiments.

Microarray technology is rapidly evolving with many competing platforms, native data formats, and analysis tools. As a result, a data submission standard should not be so specific as to stifle flexibility or innovation. Similarly, standards should not be burdensome, discouraging submission or slowing scientific progress. Many scientific journal editors grappling with these issues have adopted the Minimal Information About Microarray Experiments (MIAME) guidelines as a standard for submission of microarray data as part of a submitted publication (http://www.mged.org/Workgroups/MIAME/miame.html). A slightly modified version of MIAME, described below in Sections 3.2 through 3.7 and Appendix B, is proposed as the recommended microarray data submission template for EPA, which will be subject to change as the technology evolves. As genomics science and the associated technologies evolve, it can be expected that the MIAME guidance will concomitantly evolve. If the MIAME guidance in this document conflicts with the most recent changes to the MIAME guidance, the reader is directed to consider the MIAME guidance as the most recent, correct version.

3.2 Abstract

An abstract or executive summary of the source and type of data as well as the type of data evaluation and its final interpretation would provide a useful introduction to the data submission. Such a summary would not need to be exhaustive but would optimally provide the key highlights so that the reader will know the source of the data and how it was interpreted. The abstract might be written in a similar manner as for the submission to a scientific meeting or a journal article. It is advantageous if the reader is able to extract the important features of the submission and its interpretation from the abstract, although it is understood that a thorough evaluation of the substance of the data will involve a review of all the submitted material.

3.3 Experimental Design

It would be beneficial if voluntary submissions of genomics data to EPA included a sufficient description of the experimental design necessary to understand the source and nature of the data as well as the materials used to conduct the research. The following discussion is not an exhaustive listing or meant to be complete but indicates the spectrum of information on the experimental design that might be submitted for review. The submitter should consider providing the standard information one would include in the materials and methods section of any scientific article including a list of all the endpoints examined in the study. Such information would include information about the biological model system, treatment methods and doses, husbandry of animals, and cell culture information for *in vitro* systems. If whole animal models were employed, then submission of information regarding the exposure system, exposure doses, time points, details on euthanasia, length of time between harvesting of tissues and freezing or other processing, numbers of samples utilized for DNA array analysis, methods of RNA processing, and RNA quantification should be considered. The submitter should consider providing information on the methods employed for hybridization and incorporation of label and the numbers of hybridizations. When relevant, the submission of additional information necessary for interpretation of the data should be considered. Such information might include reference sample information, sample amplification, or any additional information unique to the study. The submitter should also consider providing information regarding any problems that arose during the study that could have an impact on interpretation.

3.4 Array Design

The inclusion of a complete description the platform used for transcriptional expression analysis such that the reviewer can assess the appropriateness of the analysis should be considered. The platform might be a commercially available platform *(e.g.*, Affymetrix, Agilent, Clontech) such that reference may be made to the specific type of chip used and the locations (weblink) of the source of the proprietary information so that the reviewer may access this information to aid in the review of the data analysis. If the transcriptional expression analysis was derived from a custom array designed for or by the submitter, then a inclusion of complete description of the production of the array would be useful. This information would likely include but certainly not be limited to the source of the nucleotide sequences used on the array, how the arrays were prepared, equipment used to prepare the arrays, description of the slides or membranes on which the arrays were spotted, gene lists, and any supportive data which confirms the specificity of the sequences used. A more complete listing of the types of data that would be useful in supporting the submission of custom arrays can be found in Appendix B.

3.5 Biomaterials

It is advantageous if the submitted data package presents the physical characteristics of the studied biomaterials as these will likely vary between experiments. Such characteristics might include age, sex, cell type/line, and/or genetic variation. When applicable, this information would address the biological material from which nucleic acids (or proteins) have been extracted for subsequent labelling and hybridization. It is also recommended that submitted information on biomaterials detail the source properties, treatment, extract preparation, and labelling of the sample. Any pertinent information about sample controls would also be useful in analyzing submitted data.

The exposure conditions applied to each test organism or tissue are important parameters influencing the experimental response. As a result, it is useful to document the incubation and treatment conditions applied to the studied biomaterial. Other key submission information might

include the method of chemical or physical exposure using the appropriate dosing units. Furthermore, any processing of samples taking place after exposure would be of interest.

Information on the hybridization extract preparation protocol might include such details as the nucleic acid type and amplification method used. It would also be useful to record and submit the labeling materials and technique used in the experiment. Finally, the data submitter should consider outlining the type and position on the array of any external controls that may have been added to the hybridization extract(s). Please see Table B.4 in Appendix B for further information.

3.6 Hybridization

It would be usedul to submit a concise description of the procedures adopted for each hybridization. If a commercially available platform is utilized, reference may be made to the specific type of hybridization procedures and parameters adopted in the experiment. Web or literature citations describing the source of the hybridization protocol and materials are useful. Furthermore, information regarding the relationship between the labelled sample extracts and their corresponding arrays (design, batch and serial number) would be useful for understanding the experiment. Documentation of the steps taken in the hybridization including information regarding the solution, blocking agent and concentration used, wash procedure, quantity of labelled target used, time, concentration, volume, temperature, and a description of the hybridization instruments is encouraged.

3.7 Measurements

The submitter should consider completely describing the methods used to acquire the image of the array, the nature of the image (*e.g.*, TIFF), the nature of the extraction of image data into quantified image data, and the nature of the spreadsheets used to house the quantified data. Submission of the original TIFF images is encouraged as is the submission of the initial quantization matrix. The description of the spreadsheet normalization of the TIFF data and any subsequent data analysis is also of value in a submission. In addition, features of the data used

for analysis such as background correction, normalization methods, methods used to test usability of the raw data, and types of analytical approaches would be useful information for the reviewer. Analytical approaches might include statistical models, graphical models, image based displays of data, and various analytical software packages. Information about the software may include weblink, proprietary information from instruction manual, or specific description of custom analytic methods. More complete description of information that should be considered for a submission for review may be found in Appendix B.

4.0 Data Analysis Guidance

This section provides information that will assist in regulatory and risk assessment efforts when considering the use of genomics data. Genomics data can be used to aid in reducing the level of uncertainty in the decision making process and provide a means to further evaluate exposure and effects. This guidance effort is also an attempt to highlight the need for developing genomics data analysis tool criteria, and the standardization of methods for the use of these tools.

4.1 Introduction

Evaluation of qualified genomics data, which have been properly analyzed and submitted (see Sections 2.0 and 3.0), has the potential to dramatically improve the mechanistic understanding of toxicities and their relevance to human health and ecological hazard identification and risk assessments. For example, DNA microarrays may be used to identify gene expression profiles associated with exposure to particular compounds, or characteristic of certain modes of action or mechanisms of toxicity. When a correlation has been established between a gene expression profile and a toxic mechanism, then these genomic data provide supportive evidence for that mechanism. Even when the mechanism for a particular compound is unknown, genomic data can help identify plausible toxicity pathways that may be involved in the biological process under study (Crosby *et al*., 2000) for the purposes of prioritization or screening.

Genomic technologies generate vast amounts of data (gigabytes) quickly (during a single analytical session), especially when using DNA microarrays for gene expression profiling. This wealth of data increases the importance of careful documentation of experimental and analytical methods while working towards data interpretation and evaluation. The Minimal Information for the Analysis of Microarray Experiments (MIAME) guidelines have helped to standardize DNA microarray experiment documentation. Extension of the MIAME guidelines into toxicogenomics has provided even more applicable prerequisites for analysis

(http://www.mged.org/MIAME1.1-DenverDraft.DOC; Fostel at al., 2005). Also critical to analysis of genomics, and particularly microarray data, is access to the raw data from published or submitted experiments, and accompanying documentation of experimental and analysis details. Establishment of public genomic databases such as the Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) provides limited access to microarray data, but these are not compatible with all monitoring or regulatory applications.

In addition to data submission and management activities, computational tools for genomics data analysis are another critical need for routine application of genomics data. Although evaluation of many of the currently available computational tools for genomics data analysis is underway through multiple internal and external Agency research efforts, these tools have not been examined by the Agency in sufficient detail that would allow for specific final recommendations to be made. Furthermore, while the variability and complexity of microarray experiments make prescribing a common, all-encompassing protocol functionally problematic, general components for the successful analysis and interpretation of all microarray approaches are discussed. The Agency is currently participating in several projects designed to develop appropriate protocols and methods for microarray data analysis. These include collaborative efforts with Food and Drug Administration (FDA) on the Microarray Quality Control project (http://www.fda.gov/nctr/science/centers/toxicoinformatics/maqc/) and National Institute of Environmental Health Sciences (NIEHS) on the Chemical Effects in Biological Systems knowledgebase (http://cebs.niehs.nih.gov/). As an interim solution a genomics Data Evaluation Record (DER) template (Appendix C) is proposed as a means to outline a framework for genomics data analysis and documentation.

4.2 Data Analysis

A few general features of genomic data analysis areas are described below with the intent to provide a basic but broad overview.

4.2.1 Data Processing and Filtering

Data processing covers the steps from scanning the array, to obtaining reliable estimates for the relative abundance of each gene transcript in all of the samples. Generally, these steps are classified as image analysis, quality control filtering, background correction, transformation and normalization. Each hybridized array has an associated and unique image file from which individual values (pixel intensities) can be collected. Data can be filtered to exclude signals that fail quality criteria. The specifics of data filtering and the threshold levels chosen are dependent upon the details and goals of the experiment. Standardization of processing and filtering criteria will be a critical step toward intra- and inter-laboratory agreement. The final output of the initial processing will be data that can be analyzed further to identify differentially-expressed genes.

4.2.2 Statistics

A standard, or common, statistical approach, that would be appropriate for all microarray experiments, cannot be specified because of unique experimental variables such as differences in microarray platforms, experimental design (reference versus matched), levels of replication (technical versus biological), as well as within experiment sources of variation (spot to spot, slide to slide, etc.). Therefore, the types of methods and tools used for statistical analyses of microarray results often differ not only from more traditional experimental approaches, but also from one microarray experiment to another. Sample size strongly affects the statistical method chosen for analysis. For example, while a relative balance may exist between the number of samples and data points measured in a standard non-genomic experiment, microarrays, as well as proteomic and metabonomic technologies, generate hundreds and often thousands of data points from each sample. Furthermore, a variety of formulae exist to calculate appropriate microarray sample sizes, depending on experimental design. Nevertheless, the cost of conducting such experiments prohibits large scale studies with multiple sample sizes. Another constraint is sample pooling, at times a necessity due to the complex nature and paucity of biological material (*i.e.*, tissues and/or RNA quantities). It is, nonetheless, important to recognize that sample pooling may impact microarray experiments at multiple levels, including experimental design and subsequent analyses. Finally, data replication should be considered. It is important to distinguish the two types of replication that exist in biological experiments, including microarrays: technical (repeats of the same sample) and biological (starting material from unique

sources, such as different animals in a test group). For scientifically sound reasons, the latter assumes greater significance in most biological assays including microarray experiments.

4.2.3 Interpretation

Numerous approaches can be used as a secondary level of analysis to interpret differentially expressed genes detected using microarray experiments. For example, genes can be sorted by ontology (gene ontology, GO) and subsequent cluster analyses (principal component analysis, hierarchical clustering, and κ-means clustering) can be used to better organize the data and help identify patterns of gene expression.

Various bioinformatics (mathematical and statistical) algorithms can be used to integrate these patterns of expression with common biological pathways and networks of co-regulated genes. Linking these functional and pathway analyses to concurrent and previously identified phenotypic characteristics will significantly advance the understanding of the biological processes involved along the source-to-outcome continuum.

4.2.4 Inference

Integration of these various data analyses and interpretation tools can be used to infer cause and effect relationships from these genomic data (Freeman, 2005). Biological inference may lead to biomarker development as well as descriptions of dose-response relationships, mechanisms of action, and predictive toxicity. Biomarkers are recognized as providing data linking exposure to internal dose and effect. The application of biomarkers to the risk assessment process that is linked to toxic processes or mechanisms may provide additional information for risk assessors. Additionally, data generated from microarray studies on model test organisms could be 1) applied to the identification of susceptible subpopulations, 2) used to develop surrogate species for toxicity testing, and 3) extrapolated to additional species, once the biomarkers and mechanism(s) of action are identified.

4.3 Data Evaluation

The goals of the evaluation of genomics data are directed toward risk assessment for regulatory applications. Currently, however, decisions cannot be made based solely upon gene expression pattern recognition, according to EPA's Interim Genomics Policy; this technology has not yet come to set precedence on its own. Currently, confirmatory studies are useful for potential risk assessment and regulatory use. If the data generated from microarray assays are confirmed using other techniques (*i.e.*, real-time quantitative PCR, functional enzyme assays, protein and metabolite profiles and/or linked to bioassay results), these data will help support links between gene expression, exposure and the resulting adverse effects in organisms. Furthermore, interpretation of microarray data with respect to existing toxicity profiles and endpoints of other perhaps higher level tests (clinical chemistry, immunochemistry, histopathology, and reproductive endpoints) should significantly increase the diagnostic and predictive applications of these technologies in the future.

A genomics Data Evaluation Record is used here as a way to present and organize data from genomics studies in order to derive information necessary for a regulatory application (see Appendix C for the Genomics Data Evaluation Record (DER) Template). For monitoring applications such information and standardization is recommended. The sections of the DER include the general information about a study and a brief executive summary as well as the materials and methods used. The test performance section includes: treatment and sampling times, tissues and cells examined, details of tissue harvest and storage, sample preparation, data analysis, evaluation criteria and statistical analysis. The results, discussions and conclusions are also components of the DER. Sections of the DER are included to provide example information to the risk assessor as a means to document the incorporation of genomics information in the risk assessment process. Genomic data used to support the more conventional data (*e.g.,* limited clastogenesis *in vitro* associated with cytotoxicity, DNA strand breaks, lipid peroxidation) are presented in an example DER for rats exposed to alachlor (see Appendix D: Draft Genomics Data Evaluation Record for Alachlor)

4.4 Data Analysis Conclusions

The above considerations demonstrate that a systematic approach for genomics data evaluation is necessary for further use of genomic data in risk assessment efforts. Documentation methods, like those in the proposed genomics DER (Appendix C) can help capture some requisite information, but the transfer of these evaluations, and the underlying genomics data, into searchable, electronic databases will be essential to making the data useful in risk assessments. Furthermore, development of databases containing gene expression profiles for a wide variety of chemicals should facilitate creation of statistical/computational methods that predict the toxic potential of a chemical.

5.0 Data Management

The goal of this section is to outline recommendations to EPA for an approach to managing genomic data submitted to the Agency or developed internally by EPA scientists. This includes the need to consider an Agency-wide warehouse for storage, retrieval and analysis of information submitted for regulatory or risk assessment purposes.

There are several major types of needs to consider in addressing the issue of an EPA-wide database: broad scientific needs for risk assessment purposes, program-specific regulatory needs, Agency Information Technology (IT) security needs, and public access needs. Although there is an overlap of issues for each of these purposes, it is useful to think of each additional purpose adding another layer of needs.

For scientific risk assessment purposes the key needs include the following items:

1) Standardization of data inputs as identified by the Data Submission Workgroup. This includes both microarray data and experiment parameters associated with the toxicogenomics study. It also provides for electronic submission of data.
2) Provision of connectivity to external public biological databases such as Affymetrix, Agilent, and GenBank
3) A quality control mechanism to ensure the fidelity of entered data
4) Capability for importing and exporting data by means of automatic routines
5) Inclusion of a wide range of data analysis and visualization tools such as filtering, clustering, and statistical analysis
6) Sufficient scalability to address large data submissions, many users, and later addition of metabonomics and proteomics data at times in the future
7) Audit trail capability. This would provide a time line and information on who added, changed or deleted specific data. It would also provide versions prior to deletions and changes.
8) Automatic data back up and recovery system

For security and management purposes, additional key needs include:

1) Database hosting, administration and management. This includes managing data submission, database access and privileges, software and hardware updates, back-up and storage.
2) Physical and electronic security, including user authentication, firewalls, and virus protection.
3) Governance structure to provide policies and procedures for submissions, access, security, cost sharing, and priority for development of new features.

For regulatory purposes, additional considerations may be necessary:

1) Electronic signature or other formal identity management capability. If the data are submitted electronically as part of a regulatory submission, the system needs to ensure that the submission is linked to the submitter.
2) Capability of partitioning the database to secure Confidential Business Information (CBI) or other non-public information, if this is part of a regulatory submission.
3) Workflow enabled, so that reviewer can address data in systematic steps needed for response to submission.

For public access purposes, key needs are:

1) Database is Web enabled, with easy routine for export of data.
2) Clear policies governing the management of the public database as opposed to an internal or staging database.

There may also be staging considerations in building or adopting an Agency-wide genomic database. The first phase might include genomic data only, and have limited analytic capability. Eventually the database should provide quality assessment tools, extensive analytical capability, gene-centric queries, and encompass proteomic, metabonomic, and conventional

toxicology assay results. Integrating these diverse types of experimental data will support data mining as well as the development of predictive toxicology systems.

Currently, there is no single database at EPA for managing genomics data each program or lab is developing its own approach. As the needs are currently identified above, there are several advantages to creating and maintaining an EPA-wide genomics database:

1) **Cost.** All of the scientific and management/security needs identified above should be addressed by any genomic database used at EPA. Addressing these items once in a uniform way would avoid duplication of these costs.
2) **Data Access.** All users in the Agency would have access to all Agency genomic data (except CBI data), greatly enhancing our risk assessment capabilities.
3) **Quality Control and Consistency.** A quality control mechanism would ensure that all Agency data passes a consistency test.
4) **Availability of a Common Set of Analysis Tools**. As new tools are developed, they would become available to all users.
5) **Scaleable**. While lab or program specific databases may focus on a narrow range of data or analysis, an EPA database would be built to include a wider range of "omics" data and a full portfolio of analytical tools enabling Agency scientists to pursue a wider range of data mining and biological systems-oriented studies.

6.0 Additional Recommendations

The Genomics Workgroup recommends that the Agency undertake a number of follow-up activities to this interim guidance including: 1) further development of the training materials and modules outlined below, to be offered throughout the Agency to risk assessors and decision makers who will be faced with the challenge of interpreting and applying genomics information, 2) continued collaboration of EPA personnel with staff from other federal agencies and stakeholders in the development of tools for the analysis of genomics data, 3) application of this guidance to a series of case studies to evaluate its utility in risk assessment and regulatory applications; and 4) the updating of this guidance as needed as the technology evolves.

6.1 Training Needs and Recommendations

The charge to the Genomics Task Force Training Workgroup was to develop an approach and appropriate delivery mechanisms for training Agency risk assessors and managers to understand and interpret genomics data in the context of risk assessment. The need for a better understanding of molecular biology concepts, and ultimately how genomics, proteomics, and other "omics" data may be used to support decision making, is the primary driver for the development of such training for staff and managers.

In designing training genomics, the Training Workgroup considered several issues: 1) the need to develop a modular approach that could build on basic information and change as new information becomes available, 2) the need to vary the level of complexity based on the needs of a particular audience, 3) the importance of considering the target audience, based on the recognition that different staff and managers will have different needs, 4) the need to develop a schedule for production of training materials, recognizing that, by taking advantage of existing public sector resources to build the initial version of the Genomics Training, time and resources may be saved; and 5) identifying internal capacity to provide training, such as ORD scientists and risk assessors to save time and resources.

Presented below is a draft outline that describes a modular training course in molecular techniques, in general, and genomics data interpretation, in particular. The Genomics Training would consist of three levels of training targeted to specific audiences, each consisting of a series of modules devoted to a particular group of concepts and/or techniques. Each training level is outlined below in Table 1, with descriptions of the content and instructional goals. More detail on the proposed training is provided in Appendix G.

Table 1. Overview of Genomics Training Plan (see Appendix G for more detail)

Training Level	Number of Module	Target Audiences	Content	Goal
Level I: Introductory Modules-	8	Non-scientists and/or technical staff without training in biological sciences.	Molecular Biology concepts: cell structure and function, DNA, RNA, proteins, gene arrays, risk assessment concepts, regulatory and risk assessment communication, EPA's current genomics policy.	Provide basic information necessary for understanding assessments of cellular functions at the molecular level and how genomics data may affect risk assessments.
Level II: Intermediate Modules	3	Scientists and/or those likely to use genomics: Intended for staff who need more in-depth understanding of genomics data generation, but do not necessarily generate data.	Background on molecular techniques such as microarrays, DNA amplification techniques, DNA fingerprinting, protein analysis, etc. Modules to be targeted for specific applications. *(e.g.*, microbial source tracking, homeland security, field inspectors, etc.)	Provide a general understanding of various applications that may be currently considered by programs throughout EPA. Intended to support human health and ecological risk assessors.
Level III Advanced Modules	Dependent on specific technical needs.	Scientists and those likely to use genomics data to generate risk assessments.	Modules would include statistical, computational and bioinformatics approaches to analyze genomic data, the use of molecular biology in mode-of-action determinations, and using genomics data in hazard/risk assessments. Flexible to account for changes in the field and to meet needs of the different EPA programs. As new technologies/ applications appear, additional modules developed, enhanced and/or revised.	Provide advanced-level knowledge on specific technical needs that scientists performing research or developing hazard/risk assessments associated with chemical registrations and other regulatory activities may face.

6.2 Collaborative Development of Genomic Tools for Data Analysis and Data Management

The Agency, in concert with other federal agencies, has begun to investigate and evaluate the currently available computational tools for genomic data analysis. EPA has been testing the toxicogenomic data management and analysis features of the NIEHS Chemical Effects in Biological Systems (CEBS) knowledgebase and FDA National Center for Toxicological Research's ArrayTrack database. EPA has also been collaborating with FDA, National Institutes of Health (NIH), National Institute of Standards and Technology (NIST), and other stakeholders on the microarray quality control (MAQC) project to establish protocols for genomic data analysis. Further, EPA has participated in National Academy of Sciences (NAS) workshops and International Life Sciences Institute (ILSI) projects on the application of genomics to toxicology and risk assessment. Building on these prior efforts, recommendations on the use of genomics tools should be identified recognizing that the goal is the appropriate application of genomic data in risk assessments and regulatory decision making. The Agency should also consider and identify limitations of the currently available tools. Ultimately, the Agency is looking for quantitative and predictive modeling tools, which will likely call for the development of new algorithms and models. These tools will need to provide reliable and repeatable data analyses, and the consistent and necessary information for EPA decision making processes. The scientific, mathematical, and statistical methods that are used for these models and analyses will need to be validated and standardized.

Due to the potentially large volumes of genomic and associated toxicological data, it is essential that the Agency consider the development of a complete data management solution. The functional needs of a solution of this magnitude should minimally include items listed in Section 5.0 Data Management. In addition, this data management solution should address needs unique to scientifically-based risk assessments, confidential and proprietary data security, public access, and other aspects of regulatory application. It should be noted that consistency, scientific and operational robustness, common access, and availability in a scalable environment are important data management needs. While the Agency has begun to develop bioinformatics

research efforts, both intramurally (e.g., ORD's National Center for Computational Toxicology) and extramurally (the STAR funded Environmental Bioinformatics Center in NC and NJ), an Agency-wide data management solution integrating genomics, toxicological, and other key data for regulatory applications is now needed.

6.3 Applying this Interim Guidance for Microarray-Based Assays to Case Studies to Verify its Utility in Risk Assessment and Regulatory Applications

The EPA's Risk Assessment Forum and other appropriate groups should apply this interim guidance to several case studies to evaluate its utility in risk assessment and regulatory applications and to identify potential areas for improvement.

6.4 Updating Genomics Guidance as Needed

This interim guidance should be revised and updated as indicated through its application to case studies (see section 6.3 above), and as genomics technologies evolve. Additional genomics guidances (*e.g.*, proteomics, metabonomics) should be developed as needed to ensure the Agency is prepared to receive and apply such data as the need develops.

References

American Public Health Association, American Water Works Association, & Water Environment Federation. Standard Methods for the Examination of Water and Wastewater. Revision in process.

Brooks, A.N., Pennie, W.D. 2001. Transcript profiling of the response to environmental hormone mimics. Comments Toxicol 7:303-315.

Burczynski, M.E., McMillian, M., Ciervo, J., Li, L., Parker, J.B., Dunn, R,T., Hicken, S., *et al*. 2000. Toxicogenomics-based discrimination of toxic mechanism in HepG2 human hepatoma cells. Toxicol Sci 58:399-415.

Crosby, L.M., Hyder, K.S., DeAngelo, A.R., Kepler, T.B., Gaskill, R., Benavides, G.R., *et al*. 2000. Morphologic analysis correlates with gene expression changes in cultured F344 rat mesothelial cells. Toxicol Appl Pharmacol 189:205-222.

Fostel, J., Choi, D., Zwick, C., Morrison, N., Rashid, A., Hasan, A., Bao, W., *et al.* 2005. Chemical Effects in Biological Systems—Data Dictionary (CEBS-DD): A Compendium of Terms for the Capture and Integration of Biological Study Design Description, Conventional Phenotypes, and 'Omics Data. Toxicol Sci 88:585–601.

Freeman, M.R., Cinar, B., and Lu, M.L. 2005. Membrane rafts as potential sites of nongenomic hormonal signaling in prostate cancer. Trends Endocrinol Metab 16(6):273-9
(Freeman, 2005

Genter, M.B., Burman, D.M., Soundarapandian, V., Ebert, C.L., Aronow, B.J. 2002. Genomic analysis of alachlor-induced oncogenesis in rat olfactory mucosa. Physiol. Genomics 12:35-45.

Hamadeh, H.K, Bushel, P.R., Jayadev, S., Martin K., DiSorbo 0., Sieber S., *et. al*. 2002. Gene expression analysis reveals chemical-specific profiles. Toxicol Sci 67:219-231.

Moreau, Y., Aerts, S., De Moor, B., De Strooper, B., Dabrowski, M. 2003. Comparison and meta-analysis of microarray data: from the bench to the computer desk. Trends Genetics 19:570-577.

U.S. Environmental Protection Agency. 2005. Microbial Source Tracking Guide Document. Office of Research and Development, Washington, DC EPA-600/R-05/064. 131 pp. http://www.epa.gov/ORD/NRMRL/pubs/600r05064/600r05064.htm

U.S. Environmental Protection Agency, Science Policy Council. 2004. Potential Implications of Genomics for Regulatory and Risk Assessment Applications at EPA. EPA 100/B-04/002. available at: www.epa.gov/osa/genomics.htm

U.S. Environmental Protection Agency, Science Policy Council. 2002a. Interim Policy on Genomics

U.S. Environmental Protection Agency. 2002b. NELAC Constitution, Bylaws and Standards EPA/600/R-03/049.

U.S. Environmental Protection Agency. 1997. "Performance Based Measurement System," 62 Federal Register 52098 – 52100, October 6, 1997.

Waring, J.F., Ciurlionis, R., Jolly, R.A., Heindel, M., Ulrich, R.G. 2001. Microarray analysis of hepatotoxins in vitro reveals a correlation between gone expression profiles and mechanisms of toxicity. Toxicol Lett 120:359-368.

Appendix A: EPA Quality System and the Performance Approach to Quality Measurement Systems

The best quality data may not be technically available, affordable, or even applicable to the exact problem at hand. To address a variety of circumstances, EPA has developed a Quality System by which reasonable quality assurance (QA) guidelines or policies are offered for assuring, documenting, and assessing data quality. EPA's Quality System is defined in *EPA Order 5360.1 A2, Policy and Program Requirements for the Mandatory Agency-Wide Quality System*, the *EPA Quality Manual for Environmental Programs, EPA Manual 5360 A1*, the Contracts Management Manual, and the Agency's Website (www.epa.gov/quality). The requirements for EPA-funded organizations and organizations submitting data to EPA under applicable statutes and regulations are also found in the Code of Federal Regulations (48 CFR Part 46), also available through www.epa.gov/quality. Parties submitting data under applicable statutes and regulations are expected to document the quality of the data submitted as well as how it was achieved. Quality System parameters apply to environmental data operations and measurements or information that describe: (1) environmental processes, (2) location or conditions, (3) ecological or health effects and consequences; and (4) performance of environmental technology

What is a Quality System?

As illustrated in Figure 1, a Quality System is viewed as a tiered organizational approach for its work processes because it defines how the work is conducted, and provides a scientific and technical basis for EPA's decision making process. The Quality System is a documented management structure to ensure the quality of an organization's work processes, products and services. Adhering to the Quality System helps to ensure that all operations, no matter where they are performed, occur in a consistent manner and that the processes and outputs in the system are effective, stable, and consistently followed. Key components in a Quality System are: (1) Quality management, (2) Quality assurance (QA), and (3) Quality control (QC).

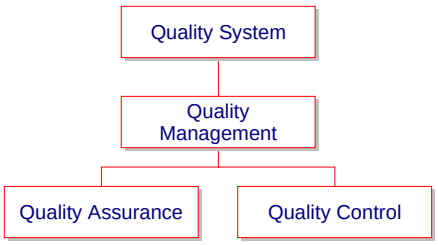

Figure 1. A Generic Quality System

What Documentation is Needed for Organizations Submitting "Genomics Data??

An organization documents its Quality System in a Quality Management Plan while a laboratory may document its implementation of specific quality policies and practices in a document entitled a Quality Manual or Quality Assurance Plan. However named, the document details the efforts to produce data that are adequate for their intended use and for assuring conformity with regulations and customer requirements for data quality. Examples of a Quality Management Plan are available at www.epa.gov/quality/qmps.html.

What is a Performance Approach?

A Performance Approach conveys "what" needs to be accomplished, but not prescriptively "how" to do it. EPA defines the performance approach as a set of processes wherein the data needs, mandates, or limitations of a program or project are specified, and serve as criteria for selecting appropriate methods to meet those needs in a cost-effective manner. The criteria may be published in regulations, technical guidance documents, permits, work plans, or enforcement orders. Under a performance approach, EPA would specify the questions to be answered, the decisions to be supported by the data, the level of uncertainty acceptable for making decisions, and the documentation to be generated to support this approach (see

http://www.epa.gov/fedrgstr/EPA-WASTE/1997/October/Day-06/f26443.htm, or 62 FR 52098 for more details about Agency policy regarding the performance approach)

Performance approaches can be defined as either: (1) measurement data that are of specified quality when demonstrating compliance (measurement quality objective (MQO) approach), or (2) a demonstration of compliance that achieve specified statistical confidence (the data quality objective (DQO) approach). Any appropriate measurement technology and sampling frequency/thoroughness may be used as long as MQO or DQO is documented and met.

Key components that need to be considered in a performance approach are:

a) Sampling procedures and sample acceptance criteria, describing procedures for collecting, handling (*e.g.*, time and temperature), accepting, and tracking submitted samples, and procedures for chain-of-custody.

b) Analytical methods, listing the laboratory's scope for testing and denoting accreditation/certification status for individual methods, for non-standard methods or new methods, the laboratory's validation procedures.

c) Analytical quality control measures, stating the laboratory's requirements for measurement assurance, e.g., method verification and documentation, error prevention, and analytical checks such as duplicate analyses, blanks, positive and negative culture controls, sterility checks, and verification tests.

d) Documentation control and record keeping specifications, identifying recordkeeping procedures to ensure data review, acquisition, traceability; accountability noting procedures to ensure customer confidentiality; and other parameters such as control, security, storage, retention, and disposal of laboratory records.

e) Assessments, describing the laboratory's processes to monitor the effectiveness of its QA program.

1) Internal audits of laboratory operations, performed on a routine basis, minimally annually, by the QA officer and supervisor. For a small laboratory, an outside expert may be needed.

2) On-site evaluations by outside experts to ensure that the laboratory and its personnel are following an acceptable QA program.

3) Proficiency test studies, in which the laboratory participates. These collaborative studies confirm the abilities of a laboratory to generate acceptable data comparable to those of other laboratories and to identify potential problems.

f) Correction and preventative activities, identifying procedures used to determine the causes of identified problems and to record, correct, and prevent their re-occurrence.

Systematic Project or Experimental Planning

In general, systematic project planning is essential before any activity begins, whether it's sampling or analysis. For any project, the scientist needs to develop the experimental study design by first identifying and documenting what the problem is, why the new information is needed, and the objectives for the experiment or series of experiments.

Once the study objectives are defined, the hypothesis is then developed. In systematically planning a project, the team or researcher then needs to determine the study parameters or test variables, both critical and the secondary (if any). The data quality objectives or performance criteria (*i.e.*, how good the data should be for the intended purpose) should be defined before the experiment starts along with all the appropriate quality control activities. For example, how types and numbers of replicates will be followed in the experiment, how is the specificity/selectivity of the analytical method to the target determined, how will the precision be determined in terms of repeatability and reproducibility. In the process of determining all these quality control activities, the experimental design can be optimized and documented.

Parameters of Microarray Platform Performance

Most microarray gene expression experiments fall into three broad classes, depending on outcome, that should have distinct QC reporting needs:

1. The first class of microarray experiments is that for which the investigator concludes that a treatment/exposure causes a biological effect. This is the most common conclusion from published microarray experiments, and the simplest from a QC point of view.
2. The next category of experiments is one wherein the investigator concludes that the treatment has no observable effect. These results are rarely reported in the literature, but might be common in regulatory submissions. These are somewhat more complicated from a QC perspective.
3. The third group offers claims about the magnitude of changes in transcription.

Examples of this last class of experiments are rare, and are the most difficult and expensive on which to perform adequate QC. Currently, a cost effective microarray platform on which to perform this last class of experiments is not available. While the minimum QC of the experiment may be unchanged, the extent of documentation needed to verify that an ensuing experimental report is acceptable, may vary based on accompanying results.

Although *negative* and *positive* controls are part of experimental designs and investigative approaches, the requisite controls for "expression" studies – particularly microarray experiments - are not always obvious. Investigators are encouraged to consider not only the biological system under scrutiny, but also the nature of the assertions about the system. The need to have adequate controls should be considered in an experimental scheme, in order to demonstrate that measurements are accurate enough to support scientific claims and assertions. Needs for several straightforward situations are listed below, and can be applied as a guide to more complex scenarios. It is useful if control samples are constructed in a way that ensures the control and experimental samples are as similar to one another as possible (*e.g.*, with regard to

biologic composition and complexity of RNA) except in such cases where control sample characteristics are unambiguously presumed to differ.

In cases where the investigator proposes that a biologic effect is present (the first case noted above), the primary QC issues are precision and specificity, and the use of a negative control is encouraged demonstrate that the measurement system is not likely to produce false positives. Accuracy is rarely a concern, since claims are not being made regarding the magnitude of differential expression between experimental and control groups, but only whether a difference exists. Sensitivity is also not relevant, since no difference or effect would be observed if sensitivity were too low. It is useful if the negative control and experimental groups include sufficient replicates, relative to the magnitude of effect and experimental variability, in order to show that the claimed effect, and no effect cases, can be statistically distinguished with desired confidence. RNA from untreated samples is usually as adequate and readily available as a control in this case. Precision is accounted for by statistical procedures (*e.g.*, t-test, chi squared test, and respective non-parametric analogs) routinely used to determine whether the experimental and control groups differ significantly. Additional consideration should be given to demonstrating the specificity of measurement(s) for the effect of interest. Demonstrating that a variety of probes exhibit binding affinity for discrete regions of a given transcript, can provide congruent results and is often a sound way to address the issue of platform specificity. The ability of probes to distinguish similar transcripts, including splice variants, is also useful to address. The issue of specificity is best addressed by using complementary "expression" measurement technologies (*e.g.*, quantitative real-time PCR, Northern blot analysis, RNase protection assays, S1 nuclease protection) to confirm microarray results. This will control for technique-specific effects, and by using distinct set(s) of amplification primers, help control for non-specific or unintended hybridization to microarray probes. Alternatively, a different microarray platform could be used to confirm specificity if the second microarray platform uses distinct probe sequences for detecting the transcripts of interest. A useful way to control systematic error is to ensure randomization of both processing order and acquisition of measurements for control and experimental samples. Using blind samples can be a useful approach to avoid operator bias.

In cases where the investigator maintains that there is no biologic effect (class two, above), a positive control is useful to show that the measurement system is capable of detecting the smallest effect sizes for which the claim is being made. In scenarios such as this, the additional QC factor of sensitivity comes into play, while specificity becomes less important. It is advantageous if the positive control and experimental groups both contain a sufficient number of replicates to show that the two groups can be statistically distinguished with the desired confidence level. It is useful to avoid absolute claims to the effect of '*there is no effect whatsoever,*' since only effects equal to or larger than those readily observable in the positive control can realistically be ruled out. Instead, conclusions might take the form of '*there is no effect larger than X*', where '*X*' represents the smallest magnitude of effect readily detectable in positive controls. Some validated positive controls, such as samples subjected to a treatment widely recognized to produce the desired effect, are considered the preferred source for positive controls. However, there are cases for which no adequate model exists for the effect being studied. Then, it is useful to construct a positive control using methods such as spiking complex RNA samples with purified and quantified RNA of interest. Alternatively, investigators might use mixtures of complex RNA samples in which the RNA of interest is present in varying known concentrations (see also section on **System Linearity and Calibration**). These controls are useful for demonstrating that the measurement system can readily detect the effect sizes for which the negative claim is being asserted.

As always, it is beneficial to randomize the order of processing and measurement relative to the sample group, and using blind samples should be considered. The same statistics, as applied to the '*there is an effect*' case, are generally used to control for inconsistencies in precision, but in this case acceptable performance means that the positive and negative controls may be reliably distinguished from one another, while the experimental sample is statistically indistinguishable from (appears to come from the same population as) the negative control.

When an investigator submits a claim regarding the magnitude of an effect, and not only the presence or absence of effects, a more complex system of control (i.e., calibration curve) should be considered (see also section on **System Linearity and Calibration**). In cases such as this, where quantification of differential expression is critical (e.g., when stating that

transcription of *gene X* increases 1.8 fold after exposure/treatment), accuracy becomes the foremost QC factor, and more complex positive controls and statistics should be considered. A calibration curve typically demonstrates the accuracy of the measurement system across the range of concentrations being considered. Researchers should consider assembling appropriate materials for constructing calibration curves in cases where standard reference RNA is not available. In many cases, investigators may consider methods such as spiking a complex RNA sample with a known series of concentrations of the RNA species of interest, or using a mixture series of two complex samples where the concentration of the RNA species of interest differs by a known amount. In the latter case, combining the two RNA 'targets' at different ratios produces a series of known concentrations of the RNA species of interest. It is useful to adjust the range and spacing of concentrations on the calibration curve (*e.g.*, log linear scale) and the number of replicates per concentration based on the level of precision desired and amount of experimental variability observed. Specificity of signals of interest might be confirmed by showing congruence with signals produced by probes that hybridize to a different portion of the same transcript. It is beneficial if conclusions on the magnitude of an effect include confidence intervals that reflect the performance of the measurement system during calibration curve construction, as well as variability seen in the experimental samples.

Overview of Array Technology – The Physical Platform

The current method for fabricating DNA microarrays (DNA chips) is to use either cDNA or oligonucleotides as probes that represent specific genes in the organism of interest, attached to a suitable solid substrate such as a glass microscope slide. It is acknowledged that the specificity of these probes is limited by the current understanding of gene sequence, among other things. It is useful if all sequences are periodically reevaluated based on the newest gene sequence information to ensure valid assessments.

Microarrays populated with cDNA probes are created by 'spotting' amplified cDNA fragments in a desired density pattern onto a solid medium such as a glass slide. Arrays using oligonucleotide probes are either mechanically 'spotted' or assembled by chemically synthesizing short, unique oligonucleotide probes directly onto a glass or silicon surface using

covalent chemistry or photolithographic technologies. It has been well established that numerous possibilities exist for errors to become 'fixed' during the manufacture of the arrays; therefore, the fidelity of the DNA fragments immobilized to microarray surfaces may be compromised by several different kinds of experimental and manufacturing inconsistencies. Given the QA/QC challenges in manufacturing gene arrays, a trend has emerged in recent years towards the use of gene arrays from several large vendors rather than arrays from smaller scale manufacturers or those prepared "in-house." While this may limit choice, it may also offer an advantage to the array community when addressing issues of cross-platform compatibility. There are a number of sources of technical error which can adversely impact data quality of a gene array experiment. These include, but are not limited to, poorly functioning probes or probe sets, cross hybridization of related genomic sequences, scanner settings and function, and atmospheric ozone. Unfortunately, a set of performance standards by which individual laboratories may be evaluated are not currently in place, although it is anticipated that such standards may be developed in the near future.

In many array-based studies, the investigators report microarray data for which there is no corroborating validation for the observed transcriptional measures. For profile data observed on array platforms regarding novel findings that are not readily supported in the peer-reviewed scientific literature, it is useful to include supporting data generated by traditional methods of evaluating gene expression, such as PCR, Northern blot hybridization analysis and RNase protection assays. In addition, the quality of probe sequences selected for particular transcribed regions incorporated onto the array is also a critically important consideration. For example, if probes are selected primarily from the 3' end of given genes, splice variants of those genes can evade identification, if the alternative splicing events occur 5' of a probe region. Additionally, by microarray analysis, it is very difficult to distinguish between two expressed genes that share a high degree of sequence homology. Variation in probe specificity is also a commonly encountered problem in oligonucleotide arrays. This problem frequently arises in instances where nucleic acid sequences are practically identical between two coding regions and the oligonucleotide probes are synthesized from 3'ends of the genes.

Isolation of Nucleic Acid 'Targets"

Since this biological analyte comprises the molecular species that will be measured, it is beneficial to ensure efficient isolation as well as post-isolation stability and structural integrity. Total RNA is generally used for gene expression analysis, although, mRNA is also used. RNA isolation techniques often involve homogenization of either fresh or frozen samples at high concentrations of guanidine isothiocyanate followed by phenol extraction and alcohol precipitation, although other methods can produce RNA of high quality.

Methods for determining purity (*i.e*., absence of contaminating reagents) include nucleic acid analysis by spectrophotometry at absorbance ratios of 260/280 nm, with expected values between 1.90 and 2.10 at pH 7.5. Another conventional method for determining structural integrity is through the use of MOPS-EDTA formaldehyde (or glyoxal) agarose gel electrophoresis, during which either the integrity of ribosomal RNA or the relative size distribution of mRNA can be evaluated. Recent advances in microfluidics and analytical equipment, (*e.g.,* 2100 Bioanalyzer , Agilent, Inc.), allow investigators to evaluate the integrity of nucleic acids with greater speed and accuracy than possible with agarose gel electrophoresis. It is anticipated that this technology will soon replace the more frequently used methods.

Experimental Design

The importance of pre-planning the experimental design cannot be overemphasized. Since the critical outputs from biological "expression" analyses are largely dependent upon experimental design investigators should consider devoting extensive attention to performing experiments with the appropriate design parameters. It is advantageous if the chosen experimental design provides sufficient statistical power to unambiguously test the biologic argument. The level of analytical power needed to allow for the detection of differentially expressed transcripts at a ratio greater or equal to '*X*-fold' should be considered. In addition, it is useful if such analyses take into account the percentage of false positives that the researcher is willing to accept. The false positive rate (FPR) and the false negative rate (FNR) are necessarily dependent upon each other *i.e*., a decrease in one results in an increase in the other.

When designing an experiment, adequate consideration should be given to sample sizes, the use of controls, the use of sample randomization, and blind sample procedures. Specific needs will depend on a number of factors, including the nature of the conclusion being presented, the manner by which samples are compared with one another, the range of measured effects, and experimental precision. To ensure adequate statistical power will be realized to support scientific arguments and conclusions, it is advantageous to consult a statistician during the experimental design phase. In order to estimate the projected magnitude of effect and experimental precision conducting a small scale pilot experiment in advance of the definitive experiment might be considered. Alternatively, one technology (*e.g.*, DNA microarrays) might be used as an exploratory tool for hypothesis generation, followed by the use of a secondary technology (*e.g.*, quantitative RT-PCR, qPCR) to generate adequate numbers of experimental and control replicates to fulfill hypothesis testing. Some general issues that should be considered are listed below, but their specific applicability will vary among experiments. Careful consideration of these issues should provide sufficient information for a reliable estimate of overall experimental performance, and the statistical strength of conclusions put forth.

It is expected that technical variation will be introduced at each critical laboratory step during expression analysis. In addition, unique sources of variation are likely to be associated with individual laboratories and/or technicians. It is important, therefore, that this variation be considered during study design and statistical analysis in order to avoid confounding of these sources of variation with treatment effects. For those experiments in which data are collected from array-based studies, there are three design schemes typically used; these are briefly described below. Although these are certainly not all inclusive, identification of the acceptable system is left to individual research teams. The three fundamental design alternatives typically used are 1) the flexible ***universal reference design***, which is used for analysis of many experimental factors of equal importance, or those that will be integral to future meta-analysis, 2) the efficient ***balanced block design***, for use in looking for genes that are upregulated or downregulated between two samples, and 3) the more integral ***loop design***, which when comparing samples of equal interest and high quality results in half the variance per estimate, because each sample is included two times, rather than once, at the minimal expense of one

additional chip. There is, however, a rather large experimental cost of this latter design, because it relies on not even one chip failing to reach the highest quality level.

The use of universal reference RNA has appeal when conducting experiments using gene arrays in a two-color hybridization approach. In such experiments, both the control and treated samples are labeled separately with a single sulfonated indocyanine fluorescent dye (Cy ™; e.g., Cy3 or Cy5) and are compared to a reference RNA sample which is labeled with the other of the two Cy ™ dyes. Not only does this approach help minimize the potential for dye bias, which is a significant concern when using the two-dye hybridization approach, but this also allows for comparison of data across studies that use the same reference RNA. One practical approach may be to take advantage of commercially available universal reference RNAs for gene expression profiling which, at this time, are offered for use with arrays representing a limited number of organisms (human, mouse, rat). Another experimental design often used to address dye bias is the 'dye swap' or 'dye flip'. In this method a second experiment is conducted by exchanging labeling reactions such that the treated and control samples are conversely labeled with the respective Cy ™ dyes. The approach entails the use of additional arrays; however, because dye bias has been observed by numerous investigators and noted in the literature, such a scheme should be considered when designing experiments using two-color array systems.

Experimental Replication

It is not possible to analyze expression data without an estimate of variance. Since experimental variance has both technical and biological components, replication could be incorporated at several levels. In the case of a gene array experiment, technical replication could be in the form of multiple spots per gene on the same array or, perhaps, multiple arrays for a given sample. While including technical replicates will improve data analysis it is not an absolute necessity. On the other hand, biological replication is an important consideration. While it is generally accepted that in a gene array experiment an absolute minimum of three biological replicates is needed, additional replication is often needed to detect a treatment effect when less than robust changes in gene expression are observed. Pilot studies could be conducted to estimate variance and give insight as to what level of replication may be useful. Although not

comprehensive, additional considerations for determination of replicate numbers are the relative quality and integrity of samples, the range of expected effects and the method of raw data analysis. The optimal replicate number is affected other factors such as the type of array technology and platform (single or dual channel RFU capture), array platform linearity (precision), feature density (number of representative gene probes), and the selected percentage value of FPR. Since replication is an asymptotic process, even a small number of replicates will strengthen any conclusions that can be drawn from the data, irrespective of the technological approach used to collect these data.

Pooling of Samples

From a theoretical perspective, most biological material used in expression studies arises from 'pooled' sources because most tissues used in such investigations contain many distinct cell types. Pooling of samples is primarily encouraged in those cases where the quantity of nucleic acid 'target' (total RNA) is limiting to the point that this represents the only means by which to obtain the requisite mass. It is recognized that, in certain studies, pooling of samples across individuals is a logical approach in order to limit study size. In fact, pooling of samples can help to minimize biological variation. However, it should be recognized that pooling will not be as effective in controlling biological variation as increasing the number of biological replicates in a study. Theoretically, most total RNA samples are pooled, since they are isolated from cells of related or different types after having been amplified from the original source to produce the test product. Combining samples does have the advantage of decreasing noise in the system. If biological variability is not a major concern, a pooled sample could be considered the same as a single individual when applying an experimental design. If biological variability is important to the interpretation of the data, and RNA from pooled sources is used in determination of expression measurements, it is useful to include more than one independent pool of samples for the purpose of estimating biological variability. Biological replicates are generally regarded as more critical than are technical replicates to measures of expression in biologic systems unless otherwise indicated.

Specificity and Sensitivity

Specificity and **sensitivity** of assays are affected by sequence-dependent (length and inclusive base composition) and sequence-independent (relative concentrations of probes and targets, hybridization time, temperature, etc.) factors. The specificity and sensitivity of assays have been the subject of numerous cross platform comparison studies recently cited in scientific literature (Venkatasubbarao, 2004; Enders, 2004; de Longueville *et al*., 2004). The term specificity refers to the ability of an expression platform to discriminate or select between distinct members of the same gene family, whereas sensitivity is the potential to discriminate transcripts expressed at low level in a complex background. In recent years there has been a trend in microarray design towards oligonucleotide probe sets to improve the specificity of gene targeting. Oligonucleotide microarrays (25 to 70 bp) have some advantages over arrays on which cDNA probes have been affixed. Oligonucleotide probes are designed to be identical with respect to the number of bases (length) and concentration, with comparable annealing temperatures of hybridization. These considerations account for enhanced uniformity over the entire platform. Oligonucleotides are also designed to reduce inadvertent target cross-hybridization, thereby increasing specificity during hybridization reactions. These combined properties increase the stability and reproducibility of hybridization signal on each feature on the array.

In addition to the quality of the probe sequences, the specific region of a gene that is selected as a probe to be incorporated onto the array is also critically important. For example, if probes are selected primarily from the 3' end of given genes, as is often the case, there is a distinct possibility that splice variants of those genes will evade identification if the alternative splicing events occur 5' of a probe region. Additionally, it is very difficult to distinguish between two expressed genes that share a high degree of sequence homology by microarray analysis. Irregularity in probe specificity is also a frequently encountered problem in oligonucleotide arrays. This problem frequently arises in instances where nucleic acid sequences are practically identical between two coding regions and the oligonucleotide probes are synthesized from 3' ends of the genes.

Decrease in specificity on microarray platforms generally results from the technical limitations inherent in enzymatic labeling of the RNA target. One of the most widely used

methods for enzymatic modification of total RNA, for microarray analysis of gene expression, uses T7 viral RNA polymerase *in vitro* transcription (IVT) to produce complementary RNA (cRNA) that can be hybridized to gene-specific probes affixed to arrays. Multiple rounds of amplification are used to label a limited mass of RNA by this IVT method, which has been shown to inadvertently introduce errors. Because cRNA-DNA sequence mismatches are more thermo stable than comparable cDNA-DNA mismatches, intensity artifacts have been observed due to increased non-specific hybridization.

System Linearity and Calibration

Linearity of signal responses and other measurable output are perhaps among the most significant aspects of obtaining reliable gene expression data. Regardless of technological modes (*e.g.*, microarray-based studies, semi-quantitative gel based PCR, 'real-time' PCR, or densitometric scanning of pixel density), usable data collected within the linear region of the output curve for any chosen system is essential. Given the increased number and overall density of gene-specific probes present on microarrays, it is particularly useful to demonstrate linearity of relative fluorescence units (RFUs) for the greatest number of discrete features represented on the chip. Recent observations from microarray workgroups suggest that specific reference RNA is the most efficient means by which to accomplish this. In an attempt to measure precision (B. Aronow, personal communication), it was determined that the greatest coverage of features was attained by hybridizing 4-5 different ratio mixtures, on as many chips, of species-specific RNA obtained from different sources. For instance, the study in question mixed RNA prepared from the colons of 8-week old C57BL/6 8 mice and post partum day one C57BL/6 whole animals in different proportions. The relative fluorescent unit (RFU) value changes for every gene probe that yielded a response to the mixture of mouse RNA, were statistically analyzed using least-square linear regression. This suggested approach permits investigators to ascertain a global perspective regarding the degree of linear response in a chosen system.

Randomization of Samples

Technical variations or differences in "expression" measurements can be introduced at several junctures in the experimental process including, but not limited to, methods of RNA labeling, the choice of microarray platform, capture methods for RFU intensity and signal quantitation, ozone-mediated fluorescent signal degradation, humidity and temperature, and moreover, those individuals charged with performing the experiments.

Many experimental designs suggest that blind randomization of samples is integral to the analyses. This approach offers the promise of 'flattening' both internal and external experimental sources of variation. Sample randomization should be considered wherever practical. Numerous confounding variables have been identified that can distort microarray results. Some of these sources of variation are well known (*e.g.*, RNA degradation during tissue extraction), and others have been more recently identified (*e.g.,* ozone-mediated bleaching of some florescent dyes), and some causative factors have yet to be characterized. If adequate care is taken to randomize the order in which samples are processed, and operators are unaware of the nature of each sample, known and unanticipated sources of variability are not likely to bias the outcome of the experiments. However, such sources of variability can nevertheless exert influence on the observed precision of the system. For instance, if all the experimental group samples are run on a given day, and all the negative control samples are run on the following day, it is possible that experimental features can differ on the two days (*e.g.*, operator identity, photomultiplier drift, and/or ozone concentration). Such differences could systematically bias results for the experimental samples relative to the control samples, creating the false impression of a real difference between the two groups. On the other hand, if samples for the two groups are randomized, with half the samples run on day one, and the other half on day two, factors that differ between the two days will decrease the precision observed in both groups (a readily detected and addressed occurrence), without creating the false impression of systematic differences between the groups in question.

APPENDIX B: MIAME-Based Data Submission Tables

Table B.1 Abstract				
MIAME	**Description**	**When applicable**	**Notes**	**Values**
B.1 Abstract	Brief summary of the purpose and findings of the experiment.	Always		

Table B.2 Experimental Design				
MIAME	**Description**	**When applicable**	**Notes**	**Values**
B.2 Experiment design	Design and purpose common to all hybridizations	Always	Related hybridizations interpreted as a single experiment.	
Author, laboratory, and contact	Person(s), organization(s), names and contacts (address, phone, FAX, email, URL).	Always		Contact details
Experiment type(s)	A controlled vocabulary that classifies an experiment.	Always	Experimental Factor(s).	Time course, dose response, comparison (disease vs normal, treated vs untreated), temperature shock, gene knock out, gene knock in (transgenic), etc.
Experiment Description	Description of the experiment and relevant electronic peer-reviewed journal publication(s)	When additional information is available and an electronic publication exists.	Consistent with experimental design.	Text description, citation, URL. Database entry

Table B.2 Experimental Design

MIAME	Description	When applicable	Notes	Values
Experimental factor(s)	Parameter(s) or condition(s) tested in the experiment.	Always	Experimental factor(s) consistent with Experiment Type(s)	Time, dose, compound, temperature, extraction, hybridization, labelling, scanning
Number of hybridization replicates	Number of hybridization replicates	Always	Consistent with Experiment Type(s)	Single, multiple
Common reference	A hybridization to which all the other hybridizations have been compared.	Always		Yes, no
Quality control steps	Measures to ensure quality: replicates (number and description), dye swap (for two channel platforms) or other	When appropriate		Text description. biological, technical
Qualifier, value, source (may use more than once)	Any further information about the experiment .	Additional useful information		Qualifier= name Value= value Source= database entry or ontology entry

Table B.3 Array Design

MIAME	Description	When applicable	Notes	Values
B.3 Array design	Array layout. Description of the common features of the array and each array element.	When an array design is novel and cannot refer to manufacturer	Array design should be provided by the array manufacturer.	
B.3.1. Array related information	Overall description of the array.			
Array design name	Unique name, that identifies a specific design	Array is novel and cannot refer to manufacturer	Consistent with the design name given for the array.	Design name, number of features, version (e.g.: EMBL yeast 12K ver1.1)
Platform type	Technology to place biological sequence on array.	Array is novel and cannot refer to manufacturer		in situ synthesized, spotted cDNA, etc.
Surface and coating specification	Surface coating type and name	Array is novel and cannot refer to manufacturer	Consistent with Platform Type	SurfaceType: glass, membrane, coating type
Array dimensions	Dimensions of the array support slide.	Array is novel and cannot refer to manufacturer		width, length
Number of features on the array	The number of features on the array.	Array is novel and cannot refer to manufacturer		number of features
Production protocol	A description of how the array was manufactured.	Array is novel and cannot refer to manufacturer		Protocol description, printing hardware, printing software
Provider	The primary contact (manufacturer) for the information on the array design.	Always		Contact details of manufacturer

Table B.3 Array Design

MIAME	Description	When applicable	Notes	Values
B.3.2 Reporter related information	Information on the nucleotide sequence present in a particular location on the array.			
B.3.2.1 For each reporter type				
Reporter type	Physical nature of the reporter (e.g. PCR product, synthesized oligonucleotide).	Array is novel and cannot refer to manufacturer	Consistent with Platform Type	Types: empty, PCR, synthesized oligonucleotide, plasmid, colony, etc.
Single or double stranded	Reporter sequences are single or double stranded.	Array is novel and cannot refer to manufacturer	Consistent with Platform Type	Single, double
B.3.2.2 For each reporter				
Reporter sequence information	Nucleotide sequence for each reporter: accession number (from DDBJ/EMBL/GenBank), the sequence itself or reference sequences and primers pair information	Array is novel and cannot refer to manufacturer	Consistent with Platform Type and clone	Sequence annotation, accession number, PCR primer pair
Reporter approximate length	The approximate length of the reporter sequence.	When the exact reporter sequence is NOT known		Number of bases
Clone information	For each reporter, identity of the clone, clone provider, date obtained, and availability.	When elements are from clones When an array design is novel and cannot refer to manufacturer	Consistent with Platform Type	Clone ID, provider, date obtained, availability
Reporter generation protocol	A description of how the reporters were generated.	Array is novel and cannot refer to manufacturer		Protocol

Table B.3 Array Design				
MIAME	**Description**	**When applicable**	**Notes**	**Values**
B.3.3 Features related information	Information on the location of the reporters on the array			
B.3.3.1 For each feature type				
Feature dimensions	Dimensions of each feature.	Array is novel and cannot refer to manufacturer	Consistent with array dimensions and number of features	Width, length, height, diameter
Attachment	How the elements (reporters) are physically attached to the array.	Array is novel and cannot refer to manufacturer	Consistent with element generation protocol	Covalent, ionic, hydrophobic, etc.
B.3.3.2 For each feature				
Reporter and location	Arrangement and system used to specify location of each feature	Array is novel and cannot refer to manufacturer	Consistent with array dimensions and number of features	Row, column, x microns, y microns, zone
B.3.4 Composite sequence related information	Information on the set of reporters used collectively to measure an expression of a particular gene.			
B.3.4.1 For each composite sequence				
Composite sequence information	The set of reporters contained in the composite sequence.	When elements are composite array is novel cannot refer to manufacturer	Consistent with element type	Oligonucleotide sequences, number of oligonucleotides, reference sequence
Gene name	The gene represented at each composite sequence	Array is novel and cannot refer to manufacturer	Consistent with clone and composite sequence information	Gene name, accession number, annotation
Qualifier, value, source (may use more than once)	Describe any further information about the array in a structured manner.	When additional information is available that would be useful to base queries on		Qualifier= name Value= value Source= database entry or ontology entry

Table B.3 Array Design

MIAME	Description	When applicable	Notes	Values
B.3.5 Control elements related information	Array elements that have an expected value and/or are used for normalization.			
Control element position	The position of the control features on the array.	When any elements on the array were used as controls	Consistent with Quality Control Description	Row, column, x microns, y microns, zone
Control type	The type of control used for the normalization and their qualifier.	When any elements on the array were used as controls	Consistent with Quality Control Description	Control type (spiking, negative, positive), control qualifier (endogenous, exogenous)

Table B.4 Biomaterials

MIAME	Description	When applicable	Notes	Values
B.4 Biomaterials	The biological material from which the nucleic acids have been extracted for subsequent labelling and hybridization.	Always		
B.4.1 Biosource properties	Information on the source of the sample.			
Organism	The genus and species (and subspecies) of the organism from which the biomaterial is derived.	Always		Genus, species, subspecies from NCBI taxonomy

Table B.4 Biomaterials

MIAME	Description	When applicable	Notes	Values
Sample Contact details	The resource used to obtain the biomaterial	When biomaterial was prepared or grown outside of the laboratory listed for the author		Biosource provider Type of specimen (tumor biopsy, paraffin section, stool sample}
Cell type	Cell type(s) or organs used in the experiment.	Always	Consistent with organism and targeted cell type	Name of organ tissue cell type (ATCC #) and source
Sex	Term applied to any organism able to undergo sexual reproduction in order to differentiate the individuals or type involved.	When applicable	Consistent with organism	Mating type alpha, F^+, F^-, Hfr, Mating type a, Mixed sex, Unknown sex
Age	The time period elapsed since an identifiable point in the life cycle of an organism.	When applicable	Consistent with organism	Age = combination of real number (measurement) and initial time point e.g.: coitus, birth, planting, beginning of stage
Developmental stage	The developmental stage of the organism's life cycle during which the biomaterial was extracted.	For multicellular species	Consistent with organism	Developmental stage (i.e., embryo, fetus, adult)
Organism part	The part or tissue of the organism's anatomy from which the biomaterial was derived.	For multicellular species	Consistent with organism	Organism part term)
Strain or line	Animals or plants that have an ancestral breeding.	When known	Consistent with organism	Strain or line (e.g.: Jax mouse strains, Cultivar,NCBI taxonomy)

Table B.4 Biomaterials

MIAME	Description	When applicable	Notes	Values
Genetic variation	The genetic modification introduced into the organism from which the biomaterial was derived.	When the source organism is genetically modified	Consistent with organism	Examples of genetic variation include specification of a transgene or the gene knocked-out.
Individual number	Identifier or number of the individual organism from which the biomaterial was derived.	When the organism can be distinguished on an individual basis with a unique ID	Consistent with organism	Individual ID. For patients, the identifier should be approved by Institutional Review Boards (IRB, review and monitor biomedical research involving human subjects) or appropriate body.
Individual genetic characteristics	The genotype of the individual organism from which the biomaterial was derived.	When applicable	Consistent with organism	Allele, genotype, haplotype, polymorphisms.
Disease state	The name of the pathology diagnosed in the organism from which the biomaterial was derived.	When applicable	Consistent with organism	"Normal" or . disease state description
Targeted cell type	Cell of primary interest.	Biomaterial is a mixed population of cells	Consistent with organism and cell type Biomaterial may be derived from a mixed population of cells although only one cell type is of interest.	Targeted cell type= term, (Mouse Anatomical Dictionary, FlyBase, CBIL vocabulary)
Cell line	Identifier for the cell line	Biomaterial is derived from an immortalized cell line	Consistent with organism and cell type	Cell line term, source of term (ATCC #),e.g.,Hela, Caco-2

Table B.4 Biomaterials

MIAME	Description	When applicable	Notes	Values
B.4.2 Biomaterial manipulation	Information on the treatment applied to the biomaterial			
Growth conditions	Description of environment used to grow organisms			Culture condition details
In vivo treatment	Manipulation to generate variable(s) under study .	When sample has been treated or manipulated for the study	Consistent with Experiment Type and Experimental Factors	Documentation of the set of steps taken in the treatment
In vitro treatment	Manipulation of cell culture condition for generating variables under study.	When the sample has been treated or manipulated in vitro for the study purpose	Should be consistent (where appropriate) with Experiment Type, Experimental Factors	Documentation of the set of steps taken in the treatment
Treatment type	Manipulation for generating variables under study.	When sample has been treated or manipulated for the study	Consistent with experiment type, experimental factors and treatment	Description of treatment (behavioral stimulus, compound based treatment, infection, modification (genetic, somatic)),
Compound	Drug, solvent, chemical, etc., that can be measured.	When sample has been treated or manipulated with a compound	Consistent with treatment type	Description of compound's physical and chemical characteristics
Separation technique	Technique to separate tissues or cells.	When the cells or tissue are separated from a heterogenous sample		Protocol

Table B.4 Biomaterials				
MIAME	**Description**	**When applicable**	**Notes**	**Values**
B.4.3 Hybridization extract preparation	Information on the extract preparation for each extract prepared from the sample			
Extraction method	The protocol used to extract nucleic acids from the sample.	Always		Protocol
Nucleic acid type	The type of nucleic acid extracted (e.g. total RNA, mRNA).	Always		Polymer type (total RNA, mRNA, DNA)
Amplification method	The method used to amplify the nucleic acid extracted.	When applicable		Protocol
B.4.4 Sample labelling	Information on the labelling preparation for each labelled extract.			
Amount of nucleic acid labelled	Amount of nucleic acid labelled.			Protocol
Label used	Label used.	Always		Label (Cy3, Cy5, etc.)
Label incorporation method	Label incorporation method	Always		Protocol
B.4.5 Spiking control	External controls added to the hybridization extract(s).			
Spiking control feature	Position of the feature(s) on the array expected to hybridize to the spiking control.	When applicable	Consistent with quality control description	row, column, x microns, y microns, zone

Table B.4 Biomaterials

MIAME	Description	When applicable	Notes	Values
Spike type and qualifier	Type of spike used and its qualifier	When applicable	Consistent with quality control description	Oligonucleotide, plasmid DNA, transcript,. concentration, expected ratio, labelling methods
Qualifier, value, source (may use more than once)	Describe any further information about the sample in a structured manner.	When additional information is available that would be useful to base queries on		

Table B.5 Hybridization

MIAME	Description	When applicable	Notes	Values	Included in DER?
B.5 Hybridization	Procedures and parameters for each hybridization.	Always			
Relationship between samples and arrays	Relationship between the labelled extract	Always	Consistent with technology quality control	Which sample, which extract "array design, batch and serial number, during which hybridization	Yes
Hybridization protocol	Set of steps taken in the hybridization: (solution blocking agent, concentration, wash procedure); quantity of labelled target used; time; concentration; volume, temperature.	Always		Description of the hybridization instruments	Yes

Table B.5 Hybridization					
MIAME	**Description**	**When applicable**	**Notes**	**Values**	**Included in DER?**
Qualifier, value, source (may use more than once)	Describe any further information about the hybridization in a structured manner.	When additional information is available that would be useful to base queries on			Non-specific

Table B.6 Measurements (MIAME distinguishes between three levels of data processing: image (raw data), image analysis and quantitation, gene expression data matrix (normalized and summarized data).				
MIAME	**Description**	**When applicable**	**Notes**	**Values**
B.6.1 Raw data	Each hybridization has at least one image.			
Scanner image file	The image file including header	Always		TIFF, JPEG
Scanning protocol	Steps taken for scanning array and generating an image	Always		Description of the scanning instruments and the parameter settings.
B.6.2 Image analysis and quantitation	Each image has a corresponding image quantitation table, where a row represents an array design element and a column represents different quantitation types			Mean or median pixel intensity.
Image analysis output	The complete image analysis output for each image.	Always.		Spreadsheet or tab-delimited file

Table B.6 Measurements
(MIAME distinguishes between three levels of data processing: image (raw data), image analysis and quantitation, gene expression data matrix (normalized and summarized data).

MIAME	Description	When applicable	Notes	Values
Image analysis protocol	Documentation of the set of steps taken to quantify the image	Always.		Image analysis software, the algorithm and all the parameters used
B.6.3 Normalized and summarized data	Several quantitation tables are combined using data processing metrics to obtain the 'final' gene expression measurement table (gene expression data matrix) associated with the experiment.			
Data processing protocol	Documentation of the set of steps taken to process the data.	When normalization has been performed		Normalization strategy and the algorithm used to allow comparison of all data.
Final gene expression table (s)	Derived measurement value summarizing related elements and replicates, providing the type of reliability indicator used.	When a value used for a reliability indicator has been generated	Should be consistent with quality control description and replicate description	Replicates of the elements on the same or different arrays or hybridizations, as well as different elements related to the same entity (e.g., gene). Reliability indicator for each data point (e.g., standard deviation)
Qualifier, value, source (may use more than once)	Describe any further information about the measurements in a structured manner	When additional information is available that would be useful to base queries on		

Appendix C: Genomics Data Evaluation Record (gDER) Template

Genomics
DATA EVALUATION RECORD

STUDY TYPE:

PC CODE: **DP BARCODE:**
SUBMISSION NO.:

TEST MATERIAL (PURITY):

SYNONYMS:

CITATION:

SPONSOR:

EXECUTIVE SUMMARY:

COMPLIANCE:

I. MATERIALS AND METHODS

A. MATERIALS:

1. Test Material:
Description:
Lot/Batch #:
Purity:
CAS # of TGAI:
[Structure]

2. Control Materials:
Negative control (if not vehicle) :
Final Volume:
Route:
Vehicle:

3. Test animals:

Species:
Strain:
Age/weight at study initiation:
Source:
No. animals used per dose per duration:
Properly Maintained?

4. Compound administration:

a. Test material

Dose levels:
Route:
Method:

b. Vehicle control:

c. Positive control:

B. TEST PERFORMANCE

1. Treatment and Sampling Times:

Duration of dosing:
Frequency of dosing:
Total number of doses:
Timing and frequency of sampling:
Time elapsed between dosing and sampling:

2. Tissues and Cells Examined:

3. Details of tissue harvest:

4. Detail of tissue storage:

5. Sample Preparation:

a RNA Isolation, Labelling, Amplification:

b. Histology:

c. Immunochemistry:

d. Western blot analysis:

e. Array analysis:

Characteristics of the arrays:

Methods:

5. Data Analysis:

6. Evaluation Criteria/Statistical Analysis

II. REPORTED RESULTS

A. IMMUNOCHEMISTRY:

B. WESTERN BLOT ANALYSIS:

C. MICROARRAY ANALYSIS:

III. DISCUSSION and CONCLUSIONS

A. INVESTIGATORS' CONCLUSIONS:

B. REVIEWER COMMENTS: .

C. STUDY DEFICIENCIES:

REFERENCES

ATTACHMENTS and TABLES

Appendix D: Genomics Data Evaluation Record (gDER) for Alachlor (Sample)

Genomics
DATA EVALUATION RECORD

STUDY TYPE: Mode of Action *In vivo* Genomic Analysis in Rat Olfactory Mucosa

PC CODE: **DP BARCODE:**
SUBMISSION NO.:

TEST MATERIAL (PURITY): Alachlor (Purity not listed)

SYNONYMS:

CITATION: Genter, M.B., Burman, D.M., Vijayakumar, S., Ebert, C.L., Aronow, B.J. (2002). Genomic analysis of alachlor-induced oncogenesis in rat olfactory mucosa. Physiol. Genomics 12:35-45.

SPONSOR:

EXECUTIVE SUMMARY:

In an *in vivo* genomic analysis, groups of male Long-Evans rats (1-2 rats/ group) were administered dietary preparations of an established tumorgenic dose of Alachlor (126 mg/kg/day) or untreated feed 1 day to 18 months. Ethmoid turbinates were removed and frozen in liquid N_2. Animals were sacrificed at 3, 4 or 5 months and two separate olfactory mucosal RNA samples were isolated. Other RNA samples were harvested from single rats treated with alachlor for 1 or 4 days. After 18 months of treatment, single RNA samples derived from alachlor-induced tumors were also isolated. Total RNA was extracted from frozen tissue homogenates by precipitation with ethanol/sodium acetate, screened for quality, labeled with biotin and hybridized. Histological examinations were performed on additional rats dosed for the same treatment duration. For the determination of ebrenin (a gene related to the human tumor suppressor gene, DMBT1) or ∃-catenin (gene/product associated with the wnt signaling pathway), immunochemistry was also performed on sections prepared for histology using an anti-hensin antibody or a commercial antibody; intestinal sections served as the positive control for antibody staining. CYP2A3 levels were assessed by Western blot analysis. For the array analysis, total RNA was reverse transcribed followed by second-strand cDNA synthesis. The resulting cRNA was biotinylated and hybridized to the Affymetrix GeneChip Rat U34A.

Based on an independent review of qualitative data only (presented as graphs or photographic copies of tissue sections), it was concluded that alachlor induces olfactory nasal carcinomas through a nongenotoxic mode of action (*i.e.,* oxidative stress). Support for this conclusion comes from data showing upregulation (∃2-fold increase over untreated control) of genes correlated with the following steps in the carcinogenic process: oxidative stress and damage to DNA (8heme oxygenase, glutathione and metalothionein, GADD 45, apurinic/apyramidinic endonuclease);progression of adenomas to malignant adenocarcinomas (activation of the wnt signaling pathway), and transformation to adenocarcinomas (activation of nuclear ∃-catenin genes, also associated with the wnt signaling pathway).

This study is classified as acceptable (non-guideline) but does not satisfy current regulatory data requirements for pesticides. Although guidelines do not exist for genomic data, the results presented in this published article provided critical information that enhances the understanding of the nongenotoxic mode of action for olfactory mucosal tumors induced by alachlor.

COMPLIANCE: Not applicable; the publication, however, comes from a reputable, peer-reviewed scientific journal.

I. MATERIALS AND METHODS

A. MATERIALS:

1. Test Material: Alachlor

Description: Not provided
Lot/Batch #: Not reported
Purity: % a.i. Not reported
CAS # of TGAI:

[Structure]

2. Control Materials:

Negative control (if not vehicle) : **Final Volume:** NA **Route:** NA
Vehicle: Harlan powder diet

3. Test animals:

Species: Rat
Strain: Long-Evans
Age/weight at study initiation: Not specified
Source: Harlan, Indianapolis, IN
No. animals used per dose per duration 1-2 males; 0 females
Properly Maintained? Not specified

4. Compound administration:

a. Test material

Dose levels: 126 mg/kg/day Route Oral – Feeding

Preliminary: Not performed, referred to a citation (Genter et al., (2000).

Main Study: Dietary administration of 0 or 126 mg/kg/day (tumorigenic dose as per EPA 1985)

b. Vehicle control:

Untreated Harlan powdered feed

c. Positive control:

See Immunochemistry

B. TEST PERFORMANCE

1. Treatment and Sampling Times: Male rats were fed dietary preparations of 0 or 126 mg/kg/day for 1 day to 126 mg/kg.

Sampling (after last dose): 1 and 4 days, 3, 4 and 5 month. Tumors were harvested from rats treated for 3, 4, 5, 11 or 18 months.

2. Tissues and Cells Examined: Ethmoid turbinates and/or oflactory mucosal tumors

3. Details of tissue harvest: Rats were sacrificed by a pentobarbitol overdose and decapitated. Ethmoid turbinates were rapidly removed and frozen in liquid N_2 until use. RNA samples of olfactory mucosal tumors derived from two different rats were also harvested.

4. Detail of tissue storage: Stored in liquid N_2 until use.

5. Sample Preparation:

a **RNA Isolation:** Selected frozen tissues were homogenized and total RNA was precipitated with ethanol/sodium acetate, resuspended in DEPC-treated water and screened for RNA quality using an Agilent Bioanalyzer. Acceptable samples had cut-off ratios of 1.8 for the 28S:18S ribosomal subunits. Duplicate samples were prepared for 2 rats/group at each sampling time (3, 4, 5 months) and one sample was used for single rats at days 1 and 4.

b **Histology:** Progression of the olfactory mucosa tumors was followed by harvesting tissue at 3, 4, 5, 11, and 18 months from dosed rats doses. Tissue was prepared for histological examinations as previously described (Genter *et al*., 2000)[1] with the exception that decalcified after fixation in multiple changes of cold 0.3 M EDTA prior to embedding in paraffin.

[1]Genter, MB, Burman, DM, Dingeldein, MW, Clough, I , Bolon, B (2000). Characterization of cell proliferation and immunochemical markers of alachlor-induced olfactory mucosal tumors in Long-Evans rats. Toxicol Pathol 28:770-781.

c **Immunochemistry:** Sections (5-Φm) prepared for histology were stained for detection of ebnerin (a gene related to the putative human tumor suppressor gene, DMBT1) using immunochemistry techniques. Ebnerin was localized in the nasal cavity sections and in tumors with anti-hensin antibody. Localized ebnerin was reacted with anti-guinea pig horseradish peroxidase (HRP)-conjugated secondary antibody (1:100) with tyramide signal amplification. Since this antibody also detects the intestinal crp-ductin, intestinal sections were taken from the rats, stained and served as the positive control. ∃-catenin (associated with the activation of the wnt signaling pathway)was localized with antibody from BD-Transduction Laboratories, Lexington, KY and visualized using an HRP-conjugated anti-mouse secondary antibody and TSA amplification as described.

d **Western blot analysis:** Gene expression changes in the olfactory specific cytochrome P-450 enzyme (CYP2A3), were assessed by Western blot analysis of 5 Φg of olfactory mucosal microsomal protein per lane. Visualization was achieved with HRP-conjugated secondary antibody, enhanced chemiluminescence and exposure to X-ray film.

e **Array analysis:** Characteristics of the arrays: The total number of probe sets (genes or expressed sequence tags, ESTs) interrogated was not reported. However, the study authors provided the following information: ESTs represented on the U34A GeneChip were derived from the Rat Unigene Build No. 34 assembly. All clones represented on the chip were ESTs on gene lists of interest and were subjected to re-annotation by use of Unigene and execution of the National Center for Biotechnology Information (BLASTN) searches (http://www.ncbi.nlm.gov/BLAST) against non-redundant nucleotide databases during February-April, 2002. Gene category information was based on all publically available gene ontology information from the Gene and Ontology Consortium (http://www.geneontology.org) as harvested from SWISS-PROT, GeneCards, Compugen. LocusLinks, and GeneBank as well as exhaustive Medline literature searches.

Methods: Total DNA was reverse transcribed with an oligo-dT primer and second-strand cDNA was synthesized. The resulting T7 RNA polymerase-mediated cRNA was biotin-labeled and hybridized on the Affymetrix GeneChip Rat U343A using the recommended protocol provided by Affymetrix

7. Data Analysis:

The study authors provided the following information. "MicroArray Suite 5.0 software was used to scan and quantitate the GeneChip data using a default scan setting; intensity data were scaled to target intensity of 1,500, and results were analyzed using the MicroArray Suite 5.0 and GenSpring 4.1.5 software. Data values used for filtering and clustering were "signal", signal confidence", "absolute call" (absent or present), and "change" (increased, decreased, unchanged) as implemented in MicroArray Suite 5.0.

Data were normalized as follows: the 50th percentile of all measurements was used as the positive control for each array. Each measurement for each gene was divided by this synthetic positive control, assuming that this was at least 10. The bottom 10th percentile signal level was used as a test for correct background subtraction. The measurement of each gene in each sample was divided by the corresponding value in untreated samples, assuming that the value was at least 0.01. Genes regulated across the experimental study were identified by data filtering for those over-or under- expressed in at least two samples whose signal strength was greater than 500 in two samples, and were also called "present" in at least two samples. An additional approach combined those with genes that could predict length of alachlor exposure or histological responses using Kruskal-Wallis ANOVA at $p<0.001$ and a Benjamini-Hochberg multiple testing correction as implemented in GeneSpring. K-means analysis were similarly executed in GeneSpring to organize genes into clusters based on similar expression across the treatment time course."

8. Evaluation Criteria/Statistical Analysis

Initial Filter Criteria: The study authors indicated that 4,777 probe set elements (a pool of genes that fulfill a series of initial filter criteria) were called "present" by the Affymetrix algorithm on at least one of 26 chips, 998 probe set elements were overexpressed by 1.8X or more in at least 2 samples, whereas 584 were underexpressed 0.5 X in at least 2 samples. Additionally, significant gene regulation was detected using a Welch t-test with a cutoff of $p<0.001$ (without correction for false discovery rate error). Using this approach, alachlor-exposed samples could be distinguished from untreated controls based on differential expression of 644 genes.

Cluster analysis: 1392 probe sets elements were provided for cluster analysis by combining the under and over expressed genes along with the alachlor-regulated genes and then restricting these to only genes that met the 'present' criteria. Using the K-means algorithm, 16 set were determined to serve as excellent representations of the prominent patterns in the data set. Clusters with "highly chaotic" patterns were eliminated from further analysis. Accordingly, 1,265 genes whose variance was well represented by 16 K-means sets were found. These K-means sets were grouped into the following behavior patterns:

- Sets that were upregulated acutely
- Sets that were upregulated only in alachlor-induced tumors
- Sets that were downregulated following alachlor treatment
- Sets that were downregulated in alachlor-induced tumors
- Sets that were persistently upregulated across the treatment intervals.

II. REPORTED RESULTS

A. IMMUNOCHEMISTRY: The study authors stated that ebnerin was highly expressed after 4 months of treatment with 126 mg/kg/day alachlor in nasal respiratory mucosa. Tissue sections of olfactory mucosal tumors induced by alachlor and control nasal mucosa were provided to

support the study authors' claim that this gene product was detected in vehicle control nasal respiratory tissue but was absent in the control olfactory mucosa. Increased gene expression of this protein was also noted in the progression of alachlor-induced tumors. In the olfactory tumors, ebnerin was displayed on the surface and in the ductal lumens of the tumor. In addition, nuclear localization of ∃-catenin was also confirmed using immunochemical staining. It was stated that alachlor-induced polyps and early adenomas did not exhibit nuclear localization of ∃-catenin but more advanced adenocarcinomas displayed abundant cytoplasmic and nuclear ∃-catenin; a tissue section of olfactory mucosal tumors induced by alachlor was presented to support this claim.

B. <u>WESTERN BLOT ANALYSIS</u>: K-mean analysis indicated that after exposure of the rats to alachlor for 2 or 4 days or 1 month, 137 genes were downregulated; included among these genes was the olfactory predominant cytochrome P-450 enzyme, CYP2A3 and CYP2F1, an olfactory marker protein. Western blot analysis, depicted in graphs and accompanied by histological alterations, also indicated that these genes/products returned to background levels in the presence of foci of respiratory metaplasia (3- and 4- month samples), in the presence of more pronounced epithelial atypia and small neoplasms present in –25% of the animals (5-month samples). CYP2A3 and CYP2F1 were downregulated in the presence of numerous tumors, some of which were invasive. These results were supported by the composite graphs of the 16 K-means sets presented in the publication.

C. <u>MICROARRAY ANALYSIS</u>: One-hundred and forty-eight genes and ESTs that were upregulated (*i.e.*, ∃3-fold increase in the normalized intensity value of 1.0) with acute (1 day to 1 month) exposure to alachlor were identified. These include genes associated with the control of extracellular matrix such as matrix metalloproteinase-9 (MMP-9, upregulated 9-fold), carboxypeptidase Z (upregulated 7-fold) and tissue inhibitor of metalloproteinase-1 (upregulated 3-fold); immune system functions; cell proliferation/ cell cycle regulation, including apoptosis-related genes; calcium homeostasis/signaling; olfactory-related; nervous system-related; oncogene-related; transporters; and structural machinery. These genes, subgrouped according to the key functional categories listed above are presented in Table 1 of the article (see Attachment 1). Other genes mentioned by the study authors as being upregulated ∃2-fold following acute exposure included multiple genes which encode proteins associated with oxidative stress; these included heme oxygenase, glutathione synthase and metallothionein (MT)-1 and MT-2. Additionally, the GADD 45 gene (associated with mutagenesis possibly caused by oxidative damage to DNA) was listed as one of the most highly regulated genes by alachlor.

An additional, 417 genes and ESTs were identified, based on a ∃2-fold upregulated expression, in alachlor-induced tumors as compared to the untreated mucosa. These genes included several immune response genes (*i.e.,* neutrophil defensin, mast cell proteases, squamous cell carcinoma antigens and major histocompatible complex antigens and genes associated with cell proliferation (*e.g.,* nucleolin, the major nucleolar protein in exponentially-growing eukaryotic cells). Another set of highly expressed gene were axin2 and frizzled. The study authors claim that the increased expression of these genes is suggestive of activation of the wnt signaling pathway. This pattern is consistent with the results of immunochemical staining confirming nuclear localization of ∃-catenin late in the carcinogenesis process. Primary

normalized data, gene lists and K-means groups can be obtained from http://genet.chmcc.org in the U34A folder listed under Genter *et al.,* 2002.

III. DISCUSSION and CONCLUSIONS

A. INVESTIGATORS' CONCLUSIONS: Based on these analyses, the study authors concluded that "initiation and progression of alachlor-induced olfactory mucosal tumors is associated with alterations in extracellular matrix components, induction of oxidative stress, upregulation of ebnerin, and final transformation to a malignant state by **wnt pathway** activation."

B. REVIEWER COMMENTS: Based on an independent analysis of the genomic data presented by the study authors, Agency reviewers conclude the following with respect to the proposed steps in the **alachlor**-mediated carcinogenesis model:

- **Initial progression from histologically normal olfactory mucosa to foci of abnormal mucosa**
  This step, which is regulated by genes in the acute phase of exposure, is accompanied by "upregulation" (∃2-fold increase) of genes consistent with a mutagenic response possibly as a result of oxidative damage to DNA **(8GADD 45, apurinic/apyramidinic endonuclease).** While the exact role of GADD (growth arrest and DNA-damage inducible) gene products is not known, this gene group is upregulated in response to stress to allow cells time to repair macromolecular damage or to lead cells into apoptosis so that a genetic defect is not propagated. Types of environmental stress that induce GADD genes include UV irradiation, alkylating agents and glucose starvation (Takahashi *et al*., 2001; Jackman *et al*., 1994). Stokes *et al*. (2002) also demonstrated that GADD 45 gene induction occurs in response to reactive oxygen species (ROS) and quinones and is abolished in the presence of the antioxidant, ascorbic acid. It is of note that quinones, which are operationally non-genotoxic (Clayson *et al*., 1994), are highly redox active molecules which can redox cycle with their semiquinone radicals, leading to formation of ROS, including superoxide, hydrogen peroxide, and ultimately the hydroxyl radical. Production of ROS can cause severe oxidative stress within cells through the formation of oxidized cellular macromolecules, including lipids, proteins and DNA (Bolton et al., 2000). Supporting the hypothesis of oxidative stress, Genter *et al*., also observed upregulation of other genes associated with oxidative stress,[***i.e.,*** **heme oxygenase** (Otterbein *et al*., 2000), **glutathione synthase and metallothionein** (Andrews 2000)].

- **Progression from histologically altered olfactory mucosa to the development of adenomas**
  The study authors stated that this step was accompanied by expression of genes indicating inhibition of apoptosis **[Bid3(AI102299)]** and enhancement of cell proliferation **(zyxin)**. However, no data were provided to support this claim. Nevertheless, it is of note that Sarafian and Bredesen (1994) state that ROS can serve as common mediators of apoptosis.

- **Progression to a malignant adenocarcinoma phenotype**
  This phase was indicated by induction of genes **(*i.e.*, axin2 and frizzled)** related to activation of the **wnt signaling pathway**, which are generally upregulated late in the carcinogenesis process.

- **Transformation to adenocarcinomas**
  In the late stages of tumor progression, the activation of **nuclear ß-catenin genes,** which is critical for tumor formation in other organs and is associated with mutations in the **wnt pathway**.

Several other studies support a role for oxidative stress in **Alachlor**-induced toxicity. Burman *et al*. (2003) show that dietary exposure of Long-Evans rats to 126 mg/kg/day for 1 day caused an –20% depletion of the olfactory mucosa antioxidant, GHS followed by a significantly ($p<0.001$) increased expression of genes associated with increased GHS production after 2 and 4 days of treatment. A return to control values was seen by 10 days of treatment. A pattern somewhat similar to GSH was observed for ascorbate in the olfactory tissue of 126-mg/kg/day male rats (*i.e.,* initially, a significant decrease 1 day post-treatment, followed by significant increases 2 and 4 days after dosing). In contrast to the GSH data, there was a reduction in ascorbate at 10 days. We noted, however, that the response with either antioxidant was not dose related. From these results, the investigators concluded that, "Despite the fact that GHS levels recovered, acute antioxidant perturbations may have been sufficient to trigger other steps in the carcinogenic process. Therefore, acute depletion of GSH and ascorbate may trigger more sustained events involved in both the initiation and promotion of the carcinogenic process."

There is also evidence of the ability of **alachlor** to induce oxidative stress in other tissues. Bagchi *et al*.(1995) evaluated the potential of **alachlor** to induce oxidative stress and oxidative tissue damage, as measured by production of lipid peroxidation and DNA-single strand breaks (SSB), in the liver and brain of Sprague-Dawley rats administered two equal oral doses (at 0 and 21 hours) of 300 mg/kg. As noted by Clayson *et al.* (1994), SSB are considered by to be a good indicator of oxygen damage to DNA. Results from the study of Bagchi *et al.* (2003) show that **alachlor** induced moderate lipid peroxidation in liver and brain tissues and SSB in brain but not liver DNA in samples harvested 24 hours after exposure to the first dose. The same authors also conducted *in vitro* studies of chemiluminescence on liver and brain homogenates, and found that 1nmol/mL **alachlor** induced 3-fold increases in chemiluminescence in both tissues further suggesting that **alachlor** induced ROS. Finally, the results from *in vitro* studies with cultured PC-12 neuroactive cells exposed to 100 nM **alachlor** illustrate the sequence of early events postulated for this MOA (generation of ROS ≡DNA damage ≡tissue damage) with a 2-fold increase in DNA-SSB and a 3-fold increase in LDH leakage. Although olfactory nasal tissue was not examined in this series of assays, the ability of **alachlor** to generate ROS with subsequent DNA damage and tissue damage both *in vivo* and *in vitro* has been established. Finally, Bagchi *et al.* cite the work of Akubue and Stohs (1991) showing that the oral administration of 800 mg/kg **alachlor** to rats caused the increased urinary excretion of the "oxidative lipid metabolites, malondialdehyde, formaldehyde, acetaldehyde and acetone".

Based on the above considerations, the postulated MOA (generation of ROS ≡DNA damage ≡tissue damage ≡ cell proliferation≡olfactory nasal tumors) in rats is plausible and coherent. An additional factor favoring this MOA is the evidence of weak and sporadic mutagenic effects, generally seen only at concentration near or at cytotoxic concentrations.

C. STUDY DEFICIENCIES:

The independent review of the data presented in this publication was limited to the analysis of qualitative results presented in graphs or photographs copies of tissue sections. Attempts to access the link for raw data provided in the article failed. Additionally, there were no data to support the study authors' claim of upregulation of genes associated with apoptosis or cell proliferation. These data would complete the sequence of key events in the carcinogenic process for alachlor. Access to the primary microarray data through a functioning, public website would have been preferable.

Based on an independent review of qualitative genomic data (presented as graphs or photographic copies of tissue sections) in conjunction with the conventional data, it was concluded that alachlor induces olfactory nasal carcinomas through a nongenotoxic mode of action (*i.e.,* cytotoxicity manifested through oxidative stress). Partial support for this conclusion comes from data showing upregulation (2-fold increase over untreated control) of genes correlated with the following steps in the carcinogenic process: oxidative stress and damage to DNA progression of adenomas to malignant adenocarcinomas, and transformation to adenocarcinomas. Although guidelines do not yet exist for genomic data, the results presented in this DER provided critical information that enhanced the understanding of the nongenotoxic mode of action for olfactory mucosal tumors induced by alachlor in the rat.

REFERENCES

Andrews, G.K., .2000. Regulation of metallathionein gene expression by oxidative stress and metal ions. Biochem. Pharm. 59: 95-104.

Bagchi, D., Bagchi, M., Hassoun, E.A., Stohs, S.J. 1995.. In vitro and in vivo generation of ROS, DNA damage and lactate dehydrogenase leakage by selected pesticides. Toxico. 104: 129-140.

Bolton, J.L., Trush, M.A., Penning, T.M., Dryhurst, G. Monks, TJ. 2000. Role of quinones in toxicology. Chem. Res. Toxicol. 13:135-160.

Burman, D.M., Shertzer, H.G., Senft, A.P., Dalton, T., Genter, M.B. 2003. Antioxidant perturbations in the olfactory mucosa of alachlor-treated rats. Biochem Pharm 66:1707-1715.

Clayson, D.B., Mehta, R., Iverson, F. 1994. Oxidative DNA damage - The effect of certain genotoxic and operationally non-genotoxic carcinogens. Mutat. Res.317: 25-42.

Kasai, H.1997. Analysis of a form of oxidative DNA damage, 8-hydroxy-2'-deoxyguanosine, as a marker of cellular oxidative stress during carcinogenesis. Mutat. Res. 387:147-163.

Jackman, J., Alamo I.Jr.,, Forance, A.J. Jr. 1994. Genotoxic stress confers preferential and coordinate messenger RNA stability on the five *gadd* genes. Cancer Res. 54:5656-5662.

Otterbein, L.E., Augustine, M.K.C. 2000. Heme oxygenase: colors of defense against cellular stress. Am. J. Physiol. Lung Cell. Mol. Physiol. 279: 1029-1037.

Sarafian, T.A. and Bredesen, D.E. 1994. Is apoptosis mediated by ROS? Free Rad. Res. 21:1-8.

Stokes, A.H., Freeman, W.M., Mitchell, S.G., Burnette, T.A., Hellman, G.M., Vrana, K.E. 2002. Induction of GADD 45 and GADD153 in Neuroblastoma Cells by Dopamine-Induced Toxicity. Neuro.Toxicol. 23:675-684.

Takahashi, S., Saito, S., Ohtani, N., Sakai, T. 2001. Involvement of the Oct-1 regulatory Element of the *gadd45* Promoter in the p53-independent Response to Ultraviloet Irradiation. Cancer Res. 61:1187-1195.

Appendix E: MIAME Glossary

For the most recent version of the MIAME glossary, please see:
http://www.mged.org/Workgroups/MIAME/miame_glossary.html

Age: The time period elapsed since an identifiable point in the life cycle of an organism. (If a developmental stage is specified, the identifiable point would be the beginning of that stage. Otherwise the identifiable point must be specified such as planting) [MGED Ontology Definition]

Amount of nucleic acid labeled: The amount of nucleic acid labeled

Amplification method: The method used to amplify the nucleic acid extracted

Array design: The layout or conceptual description of array that can be implemented as one or more physical arrays. The array design specification consists of the description of the common features of the array as the whole, and the description of each array design elements (*e.g.*, each spot). MIAME distinguishes between three levels of array design elements: feature (the location on the array), reporter (the nucleotide sequence present in a particular location on the array), and composite sequence (a set of reporters used collectively to measure an expression of a particular gene)

Array design name: Given name for the array design, that helps to identify a design between others (e.g., EMBL yeast 12K ver1.1)

Array dimensions: The physical dimension of the array support (e.g. of slide)

Array related information: Description of the array as the whole

Attachment: How the element (reporter) sequences are physically attached to the array (e.g. covalent, ionic)

Author, laboratory, and contact: Person(s) and organization (s) names and details (address, phone, FAX, email, URL)

Biomaterial manipulation: Information on the treatment applied to the biomaterial

Bio-source properties: Information on the source of the sample

Cell line: The identifier for the immortalized cell line if one was used to derive the BioMaterial [MGED Ontology Definition]

Cell type: Cell type used in the experiment if non mixed. If mixed the targeted cell type should be used [MGED Ontology Definition]

Clone information: For each reporter, the identity of the clone along with information on the clone provider, the date obtained, and availability

Common reference: A hybridization to which all the other hybridizations have been compared

Composite sequence information: The set of reporters contained in the composite sequence. The nucleotide sequence information for each composite element: number of oligonucleotides, oligonucleotide sequences (if given), and the reference sequence accession number (from relevant databases)

Composite sequence related information: Information on the set of reporters used collectively to measure an expression of a particular gene

Compound: A drug, solvent, chemical, etc., that can be measured [MGED Ontology Definition]

Contact details for sample: The resource (*e.g.*, company, hospital, geographical location) used to obtain or purchase the BioMaterial and the type of specimen [MGED Ontology Definition]

Control elements position: The position of the control features on the array

Control elements related information: Array elements that have an expected value and/or are used for normalization

Control type: The type of control used for the normalization and their qualifier

Data processing protocol: Documentation of the set of steps taken to process the data, including: the normalization strategy and the algorithm used to allow comparison of all data

Developmental stage: The developmental stage of the organism's life cycle during which the BioMaterial was extracted [MGED Ontology Definition]

Disease state: The name of the pathology diagnosed in the organism from which the BioMaterial was derived. The disease state is normal if no disease has been diagnosed [MGED Ontology Definition]

Element dimensions: The physical dimensions of each features

Experiment description: Free text description of the experiment and link to an electronic publication in a peer-reviewed journal

Experiment design: Experiment is a set of one or more hybridizations that are in some way related (*e.g.*, related to the same publication MIAME distinguishes between: the experiment design (the design, purpose common to all hybridizations performed in the experiment), the sample used (sample characteristics, the extract preparation and the labeling), the hybridization (procedures and parameters) and the data (measurements and specifications)

Experiment type (s): A controlled vocabulary that classify an experiment

Experimental design: Design and purpose common to all hybridizations performed in the experiment

Experimental factor (s): Parameter (s) or condition (s) tested in the experiment

Extraction method: The protocol used to extract nucleic acids from the sample

Features related information: Information on the location of the reporters on the array

Final gene expression table (s): Derived measurement value summarizing related elements and replicates, providing the type of reliability indicator used

Gene name: The gene represented at each composite sequence: name and links to appropriate databases (*e.g.* SWISS-PTOR or organism specific database)

Genetic variation: The genetic modification introduced into the organism from which the BioMaterial was derived. Examples of genetic variation include specification of a transgene or the gene knocked-out [MGED Ontology Definition]

Growth conditions: A description of the isolated environment used to grow organisms or parts of the organism [MGED Ontology Definition]

Hybridization protocol: Documentation of the set of steps taken in the hybridization, including: solution (*e.g*. concentration of solutes); blocking agent and concentration used; wash procedure; quantity of labelled target used; time; concentration; volume, temperature, and description of the hybridization instruments

Hybridization extract preparation: Information on the extract preparation for each extract prepared from the sample

Hybridizations: Procedures and parameters for each hybridization

Image analysis and quantitation: Each image has a corresponding image quantitation table, where a row represents an array design element and a column to a different quantitation types (e.g. mean or median pixel intensity)

Image analysis output: The complete image analysis output for each image

Image analysis protocol: Documentation of the set of steps taken to quantify the image including: the image analysis software, the algorithm and all the parameters used

In vitro treatment: The manipulation of the cell culture condition for the purposes of generating one of the variables under study and the documentation of the set of steps taken in the treatment

In vivo treatment: The manipulation of the organism for the purposes of generating one of the variables under study and the documentation of the set of steps taken in the treatment

Individual genetic characteristics: The genotype of the individual organism from which the BioMaterial was derived [MGED Ontology Definition]

Individual number: Identifier or number of the individual organism from which the BioMaterial was derived. For patients, the identifier must be approved by Institutional Review Boards (IRB, review and monitor biomedical research involving human subjects) or appropriate body [MGED Ontology Definition]

Label incorporation method: The label incorporation method used

Label used: The name of the label used

Measurements: MIAME distinguishes between three levels of data processing: image (raw data), image analysis and quantitation, gene expression data matrix (normalized and summarized data)

Normalized and summarized data: Several quantitation tables are combined using data processing metrics to obtain the 'final' gene expression measurement table (gene expression data matrix) associated with the experiment

Nucleic acid type: The type of nucleic acid extracted (*e.g*. total RNA, mRNA)

Number of elements on the array: The number of features on the array

Number of hybridizations: Number of hybridizations performed in the experiment

Organism: The genus and species (and subspecies) of the organism from which the BioMaterial is derived [MGED Ontology Definition]

Organism part: The part or tissue of the organism's anatomy from which the BioMaterial was derived MGED Ontology Definition]

Platform type: The technology type used to place the biological sequence on the array

Production protocol: A description of how the array was manufactured

Provider: The primary contact (manufacturer) for the information on the array design

Qualifier, value, source (may use more than once): Describe any further information about the array in a structured manner

Quality control steps: Measures taken to ensure or measure quality: replicates (number and description), dye swap (for two channel platforms) or others (unspecific binding, low complexity regions, polyA tails)

Raw data: Each hybridization has at least one image

Relationship between samples and arrays: Relationship between the labelled extract (related to which sample which extract) and arrays (design, batch and serial number) in the experiment

Reporter and location: The arrangement and the system used to specify the location of each features on the array (*e.g*. grid, row, column, zone)

Reporter approximate length: The approximate length of the reporter's sequence

Reporter generation protocol: A description of how the reporters were generated

Reporter related information: Information on the nucleotide sequence present in a particular location on the array

Reporter sequence information: The nucleotide sequence information for reporter: sequence accession number (from DDBJ/EMBL/GenBank), the sequence itself (if known) or a reference sequences *(e.g.* for oligonucleotides) and PCR primers pair information (if relevant)

Reporter type: Physical nature of the reporter (e.g. PCR product, synthesized oligonucleotide)

Sample: The biological material from which the nucleic acids have been extracted for subsequent labelling and hybridization. MIAME distinguishes between: source of the sample (bio-source), its treatment, the extract preparation, and its labeling

Sample labeling: Information on the labeling preparation for each labelled extract

Scanner image file: The TIFF file including header

Scanning protocol: Documentation of the set of steps taken for scanning the array and generating an image including: description of the scanning instruments and the parameter settings

Separation technique: Technique to separate tissues or cells from a heterogenous sample (e.g. trimming, microdissection, FACS)

Sex: Term applied to any organism able to undergo sexual reproduction in order to differentiate the individuals or type involved. Sexual reproduction is defined as the ability to exchange genetic material with the potential of recombinant progeny [MGED Ontology Definition]

Single or double stranded: Whether the reporter sequences are single or double stranded

Spike type and qualifier: The type of spike used (*e.g*. oligonucleotide, plasmid DNA, transcript) and its qualifier (e.g. concentration, expected ratio, labeling methods)

Spiking control: External controls added to the hybridization extract (s)

Spiking control feature: Position of the feature (s) on the array expected to hybridize to the spiking control

Strain or line: Animals or plants that have a single ancestral breeding pair or parent as a result of brother x sister or parent x offspring matings [MGED Ontology Definition]

Surface and coating specification: Type of surface and name for the type of coating used

Targeted cell type: The targeted cell type is the cell of primary interest. The BioMaterial may be derived from a mixed population of cells although only one cell type is of interest [MGED Ontology Definition]

Treatment type: The type of manipulation applied to the BioMaterial for the purposes of generating one of the variables under study [MGED Ontology Definition]

Appendix F: Additional Glossary from Genomics White Paper

Allele: An alternative form of a gene or any other segment of a chromosome

Bioinformatics: The analysis of biological information using computers and statistical techniques; the science of developing and utilizing computer databases and algorithms to accelerate and enhance biological research.

Biomarker: A molecular indicator of a specific biological property; a biochemical feature or facet that can be used to measure the progress of disease or the effects of treatment.

Complementary DNA (cDNA): DNA made from a messenger RNA (mRNA) template. The single-stranded form of cDNA is often used as a probe in physical mapping.

Biotechnology: Set of biological techniques developed through basic research and now applied to research and product development. In particular, biotechnology refers to the use by industry of recombinant DNA, cell fusion, and new bioprocessing techniques.

Computational Toxicology (Comp Tox): Word used first in EPA's Interim Policy on Genomics - "Computational Toxicology is defined as the application of models from computational and mathematical biology and computational chemistry for prediction and understanding mechanisms" - Computational Toxicology Framework Document, ORD, April 2003.

DER: Data Evaluation Record

Deoxyribonucleic acid (DNA): Nucleic acid that constitute the genetic material of all cellular organisms and DNA viruses. The genetic information is used in the synthesis of ribonucleic acids (RNAs) from DNA templates (transcription), and in the synthesis of proteins from messenger RNA (mRNA) templates (translation).

DNA Microarray:. Microarray is a tool used to sift through and analyze the information contained within a genome. A microarray consists of different deoxyribonucleic acid (DNA) probes that are chemically attached to a substrate, which can be a microchip, a glass slide or a microsphere-sized bead.

Expressed sequence tag: A unique stretch of DNA within a coding region of a gene that is useful for identifying full-length genes and serves as a landmark for mapping.

FACS: Fluorescence Activated Cell Sorter

Gene: The fundamental physical and functional unit of heredity. A gene is an ordered sequence of nucleotides located in a particular position on a particular chromosome that encodes a specific functional product (*i.e.,* a protein or RNA molecule).

Gene chip technology: Development of cDNA microarrays from a large number of genes. Used to monitor and measure changes in gene expression for each gene represented on the chip.

Gene expression: Process by which a gene's coded information is converted into the structures present and operating in the cell. Expressed genes include those that are transcribed into mRNA and then translated into protein and those that are transcribed into RNA but not translated into protein (*e.g.*, transfer and ribosomal RNAs).

Genetics: Study of inheritance patterns of specific traits.

Genetic testing: Analyzing an individual's genetic material to determine predisposition to a particular health condition or to confirm a diagnosis of genetic disease.

Genomics: Comprehensive study of whole sets of genes, gene products and their interaction.

Genome: All the genetic material in the chromosomes of a particular organism; its size is generally given as its total number of base pairs.

Genotype: The genetic composition of an organism or a group of organisms; a group or class of organisms having the same genetic constitution.

Hazard Assessment: The process of determining whether exposure to an agent can cause an increase in the incidence of a particular adverse health effect (*e.g.,* cancer, birth defect) and whether the adverse health effect is likely to occur in humans.

Hazard Characterization: A description of the potential adverse health effects attributable to a specific environmental agent, the mechanisms by which agents exert their toxic effects, and the associated dose, route, duration, and timing of exposure.

Hazard identification: The process of determining whether it is scientifically correct to infer that toxic effects observed in one setting will occur in other settings (*e.g.*, whether substances found to be carcinogenic or teratogenic in experimental animals are likely to have the same results in humans).

In Vitro: A biological study is one which is performed in isolation from a living organism (in contrast to In Vivo studies).

In Vivo: A biological study is one which is performed within a living biological organism (as opposed to an In Vitro study).

Knockout: Inactivation of specific genes. Knockouts are often created in laboratory organisms such as yeast or mice so that scientists can study the knockout or null organism as a model for a particular disease.

MAGE: MicroArray and Gene Expression; the group aims to provide a standard for the representation of microarray expression data that would facilitate the exchange of microarray information between different data systems.

MAGE-OM: Microarray Gene Expression: Object Model

MGED: The Microarray Gene Expression Data (MGED) Society is an international organization of biologists, computer scientists, and data analysts that aims to facilitate the sharing of microarray data generated by functional genomics and proteomics experiments.

Mapping: Charting the location of genes on chromosomes.

Mass spectrometry: A method used to determine the masses of atoms or molecules in which an electrical charge is placed on the molecule and the resulting ions are separated by their mass to charge ratio.

Metabolome: Entire complement of all the small molecular weight metabolites inside a cell suspension (or other sample) of interest (Aberystwyth, University of Wales Web site-http://dbk.ch.umist.ac.uk/metabol.htm). This profile is a product of the genome of the organism, the expression of that genome, and the operation of the metabolism is a particular part of the organism, in a particular environment.

Metabolomics: Involves the systematic estimation of metabolomes from a range of organisms, followed by statistical analyses and other investigations of that large quantity of data.

Metabonomics: Study of the endogenous composition of biofluids and tissues of an organism in order to probe the metabolic state in homeostasis, and when under interventional stress. Hector Keun (Biological Chemistry and Biological Sciences, Imperial College, London); Metabolic Profiling: Application to Toxicology and Risk Reduction. International Conference, May 14-15, 2003, NIEHS, Research Triangle Park, North Carolina.

MIAME: Minimum Information About a Microarray Experiment that is needed to enable the interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment

Microarray: A tool used to sift through and analyze the information contained within a genome. A microarray consists of different nucleic acid probes that are chemically attached to a substrate, which can be a microchip, a glass slide or a microsphere-sized bead.

Mode of Action: Key events and processes, starting with the interaction of an agent with a cell, through functional and anatomical changes observed on the progression to toxicity

MOPS-EDTA: [MOPS] 3-(N-Morpholino) propanesulfonic acid], [EDTA] ethylenediaminetetraacetic acid

Northern blot: A technique used to separate and identify RNA.

Nucleotide: A subunit of DNA or RNA. To form a DNA or RNA molecule, thousands of nucleotides are joined in a long chain.

"**Omics":** Term including genomics, proteomics, metabonomics (some differentiate this term from metabolomics), transcriptomics, and associated bioinformatics (Environmental Health Perspectives, 110: 2002, 1047-1050; Meeting Report: Use of Genomics in Toxicology and Epidemiology: Findings and recommendations of a workshop). Carol J. Henry and Vanessa Vu, first and last authors, respectively.

Omics Technologies: A quote often cited describes this phrase"…are based on comprehensive biochemical and molecular characterizations of an organism, tissue or cell type" Sumner ***et. al***. 2003.

Phenotype: The observable physical or biochemical traits of an organism, as determined by genetics and the environment; the expression of a given trait based on phenotype; an individual or group of organisms with a particular phenotype.

PMT: Photomultiplier tube; used in the capture of raw data

Polymorphism: The quality or character of genes occurring in several different forms.

Proteome: All of the proteins produced by a given species, just as the genome is the totality of the genetic information possessed by that species.

Proteomics: Study of the function of all expressed proteins (Nature, 422: 2003, 193-197).

Quality policy statement: Describing the specific objectives and commitment of the laboratory and its management to quality and data integrity. An ethics statement may be included at this point.

RNA: Nucleic acid found in all living cells that plays a role in the transfer of information from DNA to the protein-forming system of the cell. The base sequence of an RNA is specified by the base sequence of a section of the DNA (a Gene) which is used as the template for RNA synthesis (transcription). (Dorland's Medical Dictionary)

Risk Assessment (in the context of human health): The evaluation of scientific information on the hazardous properties of environmental agents (hazard characterization), the dose-response relationship (dose-response assessment), and the extent of human exposure to those agents (exposure assessment). The product of the risk assessment is a statement regarding the probability that populations or individuals so exposed will be harmed and to what degree (risk characterization).

Signal transduction pathway: The course by which a signal from outside a cell is converted to a functional change within the cell.

Single nucleotide polymorphism (SNP): A change in which a single base in the DNA differs from the usual base at that position.

Standard operating procedures (SOPs): listing all routine laboratory operations documented and signed by management which are available to clients upon request and readily accessible to staff. Also known as laboratory operating procedures and protocols.

Susceptibility: Increased likelihood of an adverse effect, often discussed in terms of relationship to a factor that can be used to describe a human subpopulation (*e.g*. life stage, demographic feature, or genetic characteristic).

Susceptible Subgroups: May refer to life stages, for example, children or the elderly, or to other segments of the population, for example, asthmatics or the immune-compromised, but are likely to be somewhat chemical-specific and may not be consistently defined in all cases.

Systems Biology: A holistic approach to the study of biology with the objective of simultaneously monitor all biological processes operating as an integrated system. Sumner *et. al.,* 2003.

Systems Toxicology: "…involves the study of perturbation of organisms by chemicals and stressors, monitoring changes in molecular expression and conventional toxicological parameters, and iteratively integrating biological response data to describe the functioning organism".

Throughput: Output or production, as of a computer program, over a period of time.

Toxicity: Deleterious or adverse biological effects elicited by a chemical, physical, or biological agent.

Toxicology: The study of harmful interactions between chemical, physical, or biological agents and biological systems.

Toxicogenomics: The collection, interpretation, and storage of information about gene and protein activity in order to identify toxic substances in the environment, and to help treat people at the greatest risk of diseases caused by environmental pollutants or toxicants. Study of the roles that genes play in the biological responses to environmental toxicants and stressors (Environmental Health Perspective Toxicogenomics (NIEHS).

Transgenic: Having genetic material (DNA) from another species. This term can be applied to an organism that has genes from another organism.

Web-based Glossary Sources

1- National Center for Toxicogenomics (NCT, NIEHS) Glossary
<http://www.niehs.nih.gov/nct/glossary.htm>

2- Human Genome Project Information Web Glossary
<http://www.ornl.gov/sci/techresources/Human_Genome/glossary/>

3- Cambridge Healthtech Institute <http://www.genomicglossaries.com/CONTENT/omes.asp>

4- The Physical and Theoretical Chemistry Laboratory, Oxford University Chemical and Other Safety Information <http://ptcl.chem.ox.ac.uk/MSDS/>

5- NIH Glossary <http://www.accessexcellence.org/AE/AEPC/NIH/gene27.html>

6- Integrated Risk Information Systems (IRIS, EPA) Glossary
<http://www.epa.gov/iris/gloss8.htm>

Appendix G: Content and Instructional Goals for the Three Levels of EPA Genomics Technical Training:

Level I: Introductory Modules – Molecular Biology Concepts

Modules 1-8

Goal: Provide the basic information necessary for understanding the more intricate assessments of cellular functions at the molecular level. Introduce gene arrays and discuss how genomics data may affect risk assessments in the future – this module will tie into EPA's current Genomics Policy. Issues relating to how to communicate genomics information to risk managers and the public will be addressed.

Target Audience: Non-scientists and/or technical staff without training in biological sciences, such as:

Managers from Office of Research and Development, Regional and Program Offices
Regional Risk Managers (e.g., Remedial Project Managers, On Scene Coordinators)
Attorneys
Staff from Regional Office Programs (e.g., Air, Water, Waste, Pesticides, Community Involvement, Tribal Program)
Staff from States and Tribes

Components: Cell structure and function
DNA structure and replication
RNA – Types, functions, transcription (gene expression)
Proteins – General features, formation (translation)
Gene Arrays – General principles and types
Risk Assessment Concepts – Cancer and non-cancer risk, how genomics data may affect risk assessments in the future
Regulatory Framework and Risk communication (different regulatory applications)

Level II: Intermediate Level Modules – Techniques in Molecular Biology

Modules 9-12

Goal: Provide a general understanding of all of the various applications that may be currently considered by programs throughout EPA and is intended to support human health and ecological risk assessors. Specific modules for individual program applications are considered separately (see Level II Modules – Specific Applications for Molecular Tools)

Target Audience: Scientists and/or those likely to use genomics data generated by risk assessors are the audience. Modules are intended for staff who need a more in-depth understanding of how genomics data is generated, but do not necessarily need to generate that data to support decision-

making. Modules for specific applications will be developed (e.g., microbial source tracking, homeland security, field inspectors). Examples include:

- Laboratory Staff
  - Regional Laboratories
  - Office of Research and Development
  - Enforcement/Compliance Staff (e.g., Water programs, TMDLs, FIFRA)
- Risk Assessors - Human Health and Ecological
  - Regional Offices
  - Office of Research and Development
  - Program Offices

Components: Background on molecular techniques, such as:

- Microarrays
- DNA amplification using PCR and RT-PCR
- Isolation kits
- Restriction enzymes
- Electrophoresis
- DNA fingerprinting
- Protein Analysis

Laboratory exercises using various molecular techniques (see above)

Techniques for specific applications, such as:

- Microbial source tracking
- Homeland security
- Field inspection

Molecular Biology Approaches in Quantitative Risk Assessment

Level II: Intermediate Level Modules – Specific Applications for Molecular Tools

Module 13: Homeland Security
Module 14: Microbial and/or Bacterial Source Tracking
Module 15: Molecular Techniques to Assess Exposure in Environmental Media
Module 16: Molecular Techniques for Genetically Modified Crop Plant Inspectors

Goal: Reinforce information and techniques learned in the Level II Modules – Techniques in Molecular Biology, and to provide more in-depth knowledge and skills in the performance of molecular techniques. Each of these modules is focused on a separate and specific application of the molecular tools (introduced in modules 8-11) to support different programs and needs of the Agency and its staff. Each module is intended to provide technical training to staff to increase the breadth of scientific understanding that will assist in improving job competencies with respect to science in their particular program area.

Target Audience: Same as Level II Modules – Techniques in Molecular Biology.

Components: Technical training in particular program areas, focusing on research and tools currently under development by or through ORD. For example, Module 13: Microbial and/or Bacterial Source

Tracking will use a newly developed Guide on Tools for Microbial Source Tracking (Jorge Santodomingo, in preparation), which compares a number of molecular (RT-PCR, DNA finger printing) and non-molecular (antibiotic resistance) techniques for identifying pathogenic bacteria from in water. This information may be supplemented by laboratory exercises.

Level III: Advanced Modules

Module 17: Data Analysis (1) – Statistical Analysis
Module 18: Data Analysis (2) – Bioinformatics Approaches, Computational Toxicology
Module 19: Use of Molecular Biology in Mode-Of-Action Determinations
Module 20: Using Genomics Data in Chemical Hazard/Risk Assessment

Overall Goal: Provide advanced-level knowledge on specific technical needs that scientists performing research or developing hazard/risk assessments associated with chemical registrations and other regulatory activities may face. Due to the novel and continually evolving nature of the genomics field, the advanced training modules will be flexible to account for these potential dynamic changes. As new technologies and applications appear, additional or existing training modules will be developed, enhanced and/or revised. Modules will also be flexible to meet the needs of the different EPA programs.

Target Audience: Scientists and those likely to use genomics data to generate Risk Assessments, such as:

ORD Researchers
Program Office Risk Assessors

Modules 17 & 18 (Data Analysis 1 & 2)

Goal: Provide information to research scientists and program office risk assessors on computational toxicology, bioinformatics and statistics. The modules will focus on how to identify and interpret patterns within the large volumes of genomics data and assess data significance and accuracy, offering insight into the critical evaluation, including pros, cons and limitations of possible approaches.

Components – Module 17:
Statistical approaches to microarray data analysis including, but not limited to:

Bayesian statistics
Correlation
Clustering
Principle component analysis

Components – Module 18:
Computational toxicology and bioinformatic approaches and tools used to analyze genomics data.

Models and molecular biological applications used to predict effects and understand the cascade of events leading to an effect and how statistical analyses fits together with other information to form a bigger picture.
Bioinformatics tools (algorithms and statistics) that will be used to discriminate unique signature and families of signatures indicative of stressors and groups of stressors.
Data access (i.e. data mining) and management of data

Module 19: Use of Molecular Biology in Mode-Of-Action Determinations

Goal: Introduce the approaches for and limitations of data interpretation. This module will provide a link between molecular biology methods and information and the risk assessment process.

Components:

The module will present the general concept that an understanding of the key events associated with the production of adverse health outcomes at the molecular level could enhance our ability to predict these outcomes in a qualitative and quantitative sense. In addition, variability and other uncertainties (e.g., adaptive responses and homeostatic compensation) surrounding the analysis and interpretation of microarray data for making quantitative conclusions about effect/response levels will be discussed. The concept of mode-of-action (MOA including key events) will also be introduced. The different classes of MOA will be discussed: these will include genotoxicity, mutagenicity, receptor-mediated, cell killing regenerative cell proliferation, and mitogenic responses. Each of these will be discussed in terms of the current understanding at the molecular level. For example, what is cell signaling and how do changes affect cell function; what is apoptosis and how is it induced; what controls the cell cycle and how can it be abrogated; what is the mechanism for the induction of mutations and chromosome changes and the role of DNA repair and replication? These molecular underpinnings will allow for examples of key event pathways to be discussed and how chemicals might potentially impact the various pathways.

Module 20: Using Genomics Data in Chemical Hazard/Risk Assessment

Goal: Provide guidance on the incorporation of genomics (microarray) data in a weight-of-evidence approach for hazard assessment. Present principles and pitfalls using simple case studies. Case studies will be flexible to meet the needs of the programs and offices, for example, case studies may focus on homeland security and microbial source tracking applications.

Components:
Case studies such as:
Examples where microarray data quality is high
Examples that demonstrate data concerns which could lead to erroneous conclusions
Case Studies should be developed to support the need of the programs and Regional offices, e.g., homeland security, microbial source tracking, ambient water quality monitoring, etc. to support the use of microarray data or for other molecular-biology-based or “omics” approaches.
Examples include, but are not limited to:

- Demonstration of purported evidence that a particular chemical belongs to a particular class of hepatotoxins
- Demonstration of purported evidence that chemical has characteristics of a certain class of hormonally active substances

These Case Studies should include the following elements:

- Purpose
- Overall (microarray or other "omics" approach) study design
- Purported mode of action and details of how data support proposal, including purported rationale for utility of microarray data; arguments to support conclusions
- Conventional mechanistic support: histopathology, clinical chemistry, metabolic profile, time course to appearance of critical elements, dose-response information, special studies, etc.
- Microarray data: summary gene expression profile data presentation and necessary supporting raw data, proposed up and down regulated and constituent gene identification, rationale for platform and chip design, demonstration of reproducibility, analysis of variability, positive and negative controls, dose response/temporal elements analysis, statistical analysis; RNA stability
- Correlation and comparison: between conventional and microarray data to support argument; phenotypic anchoring
- Other Evidence: Structure-Activity Relationship, etc.
- Any perceived data gaps
- Potential relevance to humans
- Weight-Of-Evidence Conclusion