Table of Contents

Application of Toxicogenomics in Safety Evaluation and Risk Assessment

10. Toxicogenomics applied to ecosystem characterization - Dan Villeneuve and Gary Ankley (EPA)

11. Dose response considerations for *toxicogenomics* – Rusty Thomas (Hamner Institute)

12. Cross-species Extrapolations – Neal Cariello (GlaxoSmithKline)

13. Toxicogenomics and Animal Alternatives – J. Kleinjans (Masstricht University, The Netherlands)

14. Application of Toxicogenomics in chemical classification and prioritization – David Dix (EPA)

15. Toxicogenomics and the Regulatory Framework (U.S. EPA/U.S. FDA)

16. Toxicogenomics and the European Union Regulatory Framework

17. Integration of toxicogenomic data in regulatory decision making – A Japanese Perspective

18. Integration of Toxicogenomics in Safety evaluation studies – Lois Lehman-McKeeman (Bristol Myers Squibb)

19. Reframing the Risk Assessment Paradigm: Towards a Systems Biology Approach
20. Glossary

Application of Toxicogenomics in Safety Evaluation and Risk Assessment
Publisher- John Wiley and Sons Inc.
Editors: Darrell R Boverhof and B. Bhaskar Gollapudi

Summary and Scope

Toxicogenomics is an emerging discipline with the potential to change current toxicology practices by providing valuable enhancements to chemical safety assessment in the 21st Century. Although the independent contributions of this emerging discipline to risk assessment have thus far been limited, researchers are now re-evaluating the experience of the past decade in order to strategically apply toxicogenomics to enhance mode of action and weight-of-evidence approaches in the risk assessment process. This book provides a timely overview of the state of the science with special emphasis on the practical applications of this technology to the risk assessment process.

The book will begin by providing a general introduction to genomics and toxicology and the establishment of the field of toxicogenomics. Background will also be provided on the general principles of risk assessment including hazard identification, dose response, exposure assessment and risk characterization. The introductory chapters will conclude with an overview of why researchers are looking for new tools and technologies to inform and enhance the risk assessment process, approaches that have been applied to date, and a practical path forward for the application of toxicogenomics to risk assessment.

These introductory sections will be followed by a series of chapters highlighting practical and systematic applications of toxicogenomics in informing the risk assessment process including the areas of mutagenicity, carcinogenicity, endocrine toxicity, organ-specific toxicity, population monitoring, and ecotoxicology. The utility of this technology in dose response analysis, cross-species extrapolations, and in the establishment of animal alternatives is also presented. Chapters are also included which discuss the use of toxicogenomics in chemical classification and prioritization for toxicity testing as well as the implications of this technology in the global regulatory framework. The book will conclude with approaches for the integration of this technology in safety evaluation studies followed by an outlook on how toxicogenomics and complementary technologies can reframe the current risk assessment paradigm.

The novel aspect of this book is that the topics and authors have been chosen to provide an informative overview of the practical applications of toxicogenomics to the risk assessment process. This is a timely contribution to the field as many researchers are re-evaluating the place of this technology in risk assessment. The list of authors will include highly experienced and internationally recognized researchers representing academic, industrial and regulatory sectors. It is expected that this book will be a valuable resource for highlighting the realistic utility of toxicogenomics in chemical risk assessment.

Summary of the 7th MAQC Project Meeting

Development and Validation of Predictive Models

May 24-25, 2007
SAS Institute Inc., Cary, North Carolina

Summary Authors: Richard Shippy (richard_shippy@affymetrix.com)
Leming Shi (leming.shi@fda.hhs.gov)
Summary Date: June 3, 2007
Meeting Organizers: Russ Wolfinger (SAS), Weida Tong and Leming Shi (FDA/NCTR)
MAQC Contact: Leming.Shi@fda.hhs.gov, Tel: +1-870-543-7387
MAQC Website: http://edkb.fda.gov/MAQC/

"If a Statistical Analysis Plan (SAP) is hardwired to a specific data set, we will more likely end up fitting to noise." Kenneth Hess (M.D. Anderson Cancer Center)

The 7th face-to-face MAQC project meeting was held on May 24-25, 2007 (9:00 am – 6:00 pm EDT) at SAS Institute's main campus in Cary, North Carolina. A total of 116 on-site participants from 66 organizations attended the two-day meeting. Eighty two (82) of them came from outside of North Carolina and some from as far away as Belgium, China, Germany, Japan, Switzerland, and UK. In addition, ~20 people participated in part or all of the meeting via WebEx or phone. The main objectives of the meeting were: (1) to review the progress of the four Working Groups (Clinical WG, Toxicogenomics WG, Titration WG, and Regulatory Biostatistics WG); (2) to discuss the proposal of establishing a Genome-Wide Association WG (GWAWG) under MAQC-II; (3) to discuss the generation of new data for assessing reproducibility of microarray signatures across laboratories using the same tumor samples; (4) to review various data analysis strategies for the development and validation of predictive models and to streamline the review process of Statistical Analysis Plans; (5) to discuss the performance metrics to be used by MAQC-II for evaluating predictive models; and (6) to identify and confirm coordinators for specific data analysis tasks so that we can move to full-scale data analysis. The SAS Institute was an excellent venue for this event and MAQC attendees expressed sincere gratitude to Russ Wolfinger and the JMP team for hosting the MAQC meeting.

May 24, 2007 (Day One)

Session 1-A: MAQC-II Overview and Working Group Updates

Chair: **Russ Wolfinger** (SAS Institute Inc.)

This Session was aimed at updating MAQC-II participants of the general progress that each WG has made so far and reaffirming each WG's objectives.

- **John Sall**, co-founder of SAS, welcomed the MAQC group and stressed the importance of the MAQC-II effort. Mr. Sall noted SAS' "long commitment to life sciences analytics." "SAS Institute likes problems with lots of data," Sall said, "Little did we know when we got into problems involving microarrays that it would challenge our own limits and stretch us to support hundreds of thousands of columns and wide data formats."
- **Leming Shi** (FDA/NCTR) provided an overview of the MAQC project and outlined the agenda for this 7th face-to-face meeting. Leming explained the rationales behind the two phases of the MAQC

project. The MAQC-I demonstrated the technical performance of microarray platforms in the identification of differentially expressed genes. The objective of MAQC-II is toward development and "validation" of predictive models. Leming anticipates that a better understanding of the capabilities and limitations of microarray data analysis approaches in clinical and toxicogenomic applications could be reached and recommendations on the development and validation of classifiers may be put forward through MAQC-II. To accomplish this, we need to explore many different options in analyzing each of the data sets to generate the "matrix of performance metrics." Leming announced that there are currently 318 participants on the mailing lists of the four MAQC-II WGs; he also introduced a new WG for Genome-Wide Association studies. Leming discussed the bilateral CIDTA (Confidentiality Information Disclosure and Transfer Agreement), distributed on March 22nd 2007, which is being executed between Data Provider and Data Recipient. He also discussed the FDA IRB (Institutional Review Board) banking and withdrawal forms, which were distributed on March 26th 2007, and are required from Data Provider and Data Recipient for providing/accessing clinical data sets within MAQC-II. Leming emphasized that MAQC is research project; participation is completely voluntary and each participant is expected to cover her/his own costs. He expressed gratitude to the scientific community's enthusiastic participation in and support of the MAQC project. Leming then finished his talk with an overview of the meeting agenda and the MAQC-II timelines.

- **Wendell Jones** (Expression Analysis Inc.) gave a overview of the Clinical WG with the team's goals to (1) Understand the behavior of various prediction rules and gene selection methods that may be applied to microarray data sets to generate predictors of clinical outcomes; and (2) Identify and characterize sources of variability in multi-gene prediction results including: a) The impact of tissue acquisition and sample preparation, b) Inter- and intra-laboratory variation in prediction results, and c) Cross-platform performance of prediction results. Wendell outlined a work plan involving the solicitation of data from parties who possess large clinically annotated gene expression data sets that are relevant for the goals of the project. The MAQC members will collectively analyze the data, compare results, and make recommendations for suitable and/or best practices. New experiments were also suggested to generate data for independent prospective validation and assessment of reproducibility of prediction outcomes. Wendell discussed the administration of the Clinical WG that is also coordinated by Lajos Pusztai (MD Anderson) and Uwe Scherf (FDA/CDRH) and contains more than 200 participating individuals. Regular teleconferences take place on Tuesdays at 2 pm EDT. The three disease areas being actively analyzed by the Clinical WG are breast cancer, neuroblastoma, and multiple myeloma. Wendell discussed the formation of a QC subgroup that is assessing the impact of the quality of individual arrays on prediction performance.
- **Federico Goodsaid** (FDA/CDER) gave an overview of the Toxicogenomics WG that is currently comprised of 170 members from 'everywhere' (i.e. across government, academia, and industry). Federico stressed why we need to do something about toxicogenomics by referencing the IND submission requirement that states "Pharmacology/toxicology data on the basis of which the sponsor has concluded that it is reasonably safe to conduct the proposed clinical investigations" and the NDA submission requirements for non-clinical studies "Submit studies that are pertinent to possible adverse effects and clinical studies" and clinical studies "Submit data or information relevant to an evaluation of the safety and effectiveness of the drug product." Federico stated that the important goals toxicogenomics are to (1) identify clinical starting dose, (2) identify organ toxicity and reversibility, and (3) guide clinical dosing regimen and escalation strategy. Federico then discussed published literature illustrating how we know toxicogenomics works. The final slides of Federico's talk provided an overview of last 3.5 months activities of the Toxicogenomics WG. The data set being analyzed is from Rusty Thomas of The Hamner Institutes for Health Sciences, which comprises 70 Affymetrix GeneChip arrays from 2005 and 2006 experiments. Federico stated that "validation" and "exploration", two important objectives of data analysis, should proceed concurrently so that we will have a better understanding why some classifiers do not work. The Toxicogenomics WG is coordinated by Federico Goodsaid and David Dix (EPA/NCCT).

- **Richard Shippy** (Affymetrix Inc.) gave an overview of the Titration WG that is currently comprised of 40 members and holds weekly teleconferences on Fridays at 1 pm EDT. Richard discussed what we learned from the analysis of titration samples in MAQC-I and then outlined new goals for MAQC-II. In MAQC-I, we found utility in the titration samples for assessing relative accuracy and for determining the ability of each platform to detect predictable titration signatures across the A, B, C, and D samples. The titration samples give us a controlled experiment to measure subtle patterns in expression that is expanded in the MAQC Pilot Study to 13 (for Affymetrix, GE Healthcare, and Illumina platforms) or 19 (for Agilent one-color platform) titration points. In this study, the titration mixtures are as subtle as 99.5%A:0.05%B *vs.* 99%A:1%B, which provides benchmarks to evaluate different classifiers in terms of accuracy of prediction. The goals of the Titration WG are to (1) write a concept paper describing why you would do a titration study and what might be learned; (2) see if titrations can be used as part of a "method validation" process for a microarray-based clinical diagnostics; (3) determine if we can use titration experiments to compare the performance of classifiers; (4) estimate the effect of cell fractions on classifier and predictive model performance; and (5) determine if we can derive some generic titration "Performance Figures-of-Merit". Since this is a well controlled data set with determinable patterns, we can use it to learn more about classification strategies that may help reaching the goals of other MAQC-II WGs. The Titration WG is coordinated by Rich Shippy (Affymetrix), Russ Wolfinger (SAS Institute), and Rick Jensen (Virginia Bioinformatics Institute).
- **Greg Campbell** (FDA/CDRH) gave an overview of the Regulatory Biostatistics WG (RBWG) whose role is to generate a specific regulatory focus for data set and classifier algorithm selections, data analysis, procedures to validate the classifiers, prospective study designs, scientific conclusions, and potential impact in regulatory review of the work within MAQC-II. The focus of the RBWG is to identify study designs and performance measures for the evaluation of microarray technology and processes for establishing choice of classifier algorithm, validation strategy, normalization method and handling of missing data. Greg stated that there are 115 members in the RBWG that is coordinated by Greg Campbell, Lakshmi Vishnuvajjala (FDA/CDRH), and Tim Davison (Asuragen Inc.). Regular teleconferences take place on Thursdays at 12:30 pm EDT. Greg discussed the Standard Operating Procedure (SOP) that has been generated to provide practical advice to the other WGs on how to carry out the prediction model building and validation. The SOP document (1) identifies the method of data normalization for the training set and for the future test set; (2) indicates if any feature selection has been done; (3) defines the selected algorithm; (4) identifies the method of performance evaluation; and (5) recommends well-executed internal cross-validation. Greg also discussed the key SOP internal and external validation criteria and touched upon the problem of multiplicity. The SOP was distributed on March 22, 2007 as part of the "MAQC-II Research Plan" (contact Leming.Shi@fda.hhs.gov if you do not have it). At the end of Greg's talk is a list of progress items for the RBWG. The RBWG spent a lot of time in reviewing the SAPs for the Hamner TGx data set. Experiences gained through this review process should be helpful to other data sets.
- **Lakshmi Vishnuvajjala** (FDA/CDRH) focused her talk on the performance metrics to be used by MAQC-II for evaluating predictive classifiers/models. Lakshmi provided a list of statistical inference performance measures and explained criteria for a classifier to be considered informative. Examples were provided for a quite accurate but non-informative classifier as well as an informative classifier that may not be very accurate. Internal validation considerations were explained, including split sample and cross validation. The external validation data set is to be used once and only once, and after validation, the classifier may not be modified. Otherwise, a new validation data set will be required for validating the modified model. Testing multiple classifiers requires a multiplicity adjustment that is not always simple. An important question to ask is whether a microarray-based predictive model add anything new to the best clinical predictor. Lakshmi talked about the STARD (The Standards for Reporting of Diagnostic Accuracy) Initiative that aims at improving the quality of the reporting of diagnostic studies. She also mentioned the availability of an FDA/CDRH guidance

document, "*Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests*" (http://www.fda.gov/cdrh/osb/guidance/1620.pdf), which was issued on March 13, 2007.

Session 1-B: Analysis Results from the Titration and Clinical Working Groups
Chair: **Jim Fuscoe** (FDA/NCTR)

In this Session, the Titration WG and the Clinical WG presented in detail some of their data analysis results and plans for future data analysis.

- **Marc Salit** (NIST) discussed analysis results from the Titration WG. Marc stated that the MAQC-II Titration WG is developing objective and quantitative microarray performance measures from RNA titrations. Quantitative and objective performance metrics from titrations could lead to establishing a *Method Validation* approach for microarray gene expression. Marc then provided a description of Method Validation stating that, in general, methods for regulatory submission must include studies on specificity, linearity, accuracy, precision, range, detection limit, quantitation limit, and robustness. It is possible that titration samples can give us an estimate of these performance metrics. With a quantitative performance measure, we can objectively evaluate data processing (e.g. normalization), sample processing (e.g. reagent kits), and the ability to detect differentially expressed genes. Marc stated that titration experiments can possibly be used to model cell population fractions from clinical biopsies as well as robustness for diagnostic performance. Marc gave a detailed look at the observations and progress of the data analysis efforts within the Titration WG with illustrations of intensity distributions, PCA, Volcano plots, single-gene titration curves, multi-gene titration 'maps', and a multi-gene statistical model. The next steps are to enhance the statistical model (e.g. fitting to Langmuir isotherm for intensity concentration relationship), evaluate data pre-processing and normalization approaches, develop a manuscript on characterizing titrations and titration performance, and finally to integrate with classifier and predictive model teams.
- **Wendell Jones** (Expression Analysis Inc.) reported the quality review of MD Anderson Cancer Center's breast cancer data set and the University of Cologne's neuroblastoma data set. In addition, Wendell gave a preview of various QC subgroup issues, data analysis results, and analysis plans. The UAMS hold out set was also discussed. Among the 178 Affymetrix U133A CEL files from the MDACC breast cancer data set distributed for array quality review, 15 arrays had already been identified by the MDACC and excluded from the original publication. Fourteen institutions (MDACC, FDA/NCTR, Roche, Expression Analysis, University of Alabama at Birmingham, Genedata, Swiss Institute of Bioinformatics, SAS Institute, Systems Analytics, Chinese Academy of Sciences, University of Southern Mississippi, Ligand Pharmaceuticals, UCLA/CSHS, and Anhui Institute of Education) then reviewed all arrays using their own internal QC procedures. Additional QC assessments were submitted by Stanford University and DNAVision. QC assessment results were collated and quality methods captured by Leming Shi and Weida Tong (FDA/NCTR) and were then distributed to MAQC participants who generated arrays QC assessment results. Many participants were surprised by the disparity of the quality assessments. This provides the impetus for trying to formulate a consensus on quality control process. Nineteen of the 178 arrays were considered as questionable in quality through this "consensus voting" process. Clinical information for 130 cases was distributed to data analysis teams for developing classification models. Clinical information for the 19 "consensus" outliers and the 29 replicate arrays, along with a completely new data set being generated by MDACC, will be used for evaluating the prediction performance of the classifiers. In addition, 502 custom Agilent two-color microarrays on 251 cases with dye-flip hybridizations from the University of Cologne were distributed for quality review. Similarly disparate QC assessment results were found when ten institutions (Cologne, NCTR, SIB, SAS, CAS, SAI, EA, Agilent, UAB, and Stanford) reviewed all arrays using their own QC procedures. In both cases, MDACC and Cologne, some 50% or more of the microarrays were deemed aberrant by at least one institution.

However, at least with the MDACC data, the really poor arrays (i.e., those originally identifies by MDACC) were identified by the great majority. Leming Shi was informed by André Oberthür that a relatively small portion of arrays (3~5%) failed the University of Cologne's internal QC thresholds and were not included in the neuroblastoma data set of 502 arrays. This may explain why none of the 502 arrays received outlier status from >60% of the institutions participated in the QC assessments. There is a wide variety of techniques and approaches in QC assessment. The Quality Control Subgroup within CWG was created to assess which methods are most relevant to classifier prediction. This subgroup will try to determine thresholds for appropriate quality assessments while simulating bias and variance with examples of induced 'bubbles' and 'black holes' provided. Wendell concluded his talk with a discussion of the UAMS hold-out data set. The UAMS multiple myeloma data set consists of more than 500 Affymetrix HG-U133Plus2.0 arrays. The training set was divided in time (processing date) with the training arrays mostly related to TT2 clinical trial. The forward validation set is comprised of both TT2 and TT3 subjects. A preliminary view of quality appears to be comparable to the first two studies (MDACC and Cologne), and the MAQC will soon distribute the training arrays for QC assessment, followed by clinical information for developing predictive models.

- **Roger Perkins** (FDA/NCTR), **Russ Wolfinger** (SAS Institute Inc.), **Weiwei Shi** (GeneGo Inc.), and **John Zhang** (Systems Analytics Inc.) provided an overview of their individual analysis results for the Clinical WG on the MDACC breast cancer data set. Roger provided a flow chart of the NCTR modeling logic that started with removal of the half of the lowest average intensity genes. Roger concluded his talk with an initial call for analysis teams for the UAMS multiple myeloma data set for which he has been suggested to coordinate the analysis efforts. Russ showed a valuable graph of AUC *vs.* RMSE for different predictive methods as a basis for recommending RMSE (Root Mean Square Error) as an additional performance metric for assessing predictive methods. Weiwei concluded that classifiers derived from GeneGo processes gave prediction performance similar to the MDACC 31-probe set signature but with better biological interpretation and robustness. John illustrated that batch effect may be an important issue in model development. John provided analysis results using MAS5 and RMA on the MDACC data set. John noted that for predictive model development, RMA on all samples is not appropriate since it is across-sample normalization. The removal of low intensity probes (i.e. noise) lessened the batch effect that may also be confounded with other experimental factors from this data set across different years.

Session 1-C: **Analysis Results from the Toxicogenomics WG**
Chair: **Greg Campbell** (FDA/CDRH)

In this Session, the Toxicogenomics WG presented in great detail some of their data analysis results on the lung tumor data set from The Hamner Institutes (Rusty Thomas). There were also discussions on the RBWG review of the Statistical Analysis Plans (SAPs) developed for the Hamner data set.

- **Weida Tong** (FDA/NCTR) has been coordinating the data analysis effort on the Hamner data set over the past 3.5 months. Weida provided an overview of the coordination of analysis efforts. This is a lung tumor data set provided by Rusty Thomas at the Hamner Institutes for Health Sciences. The experiments for this data set were conducted in three different years ('05, '06, and '07). The model development (training) data set is from '05 and '06 (70 arrays), while the confirmatory (external validation) data set is from '07 (40 arrays, not distributed yet). PCA plots showed a pronounced batch effect between the '06 data and the '05 data, which Rusty pointed out could also be due to biological differences between these two time points. The analysis approaches were to develop a classifier using the '06 data alone as well as to develop a classifier by using a combination of the '05 and '06 data sets. Weida listed some specific questions for the Toxicogenomics WG such as (1) how to conduct the cross-validation? and (2) how to calculate the performance metrics? The preprocessing methods outlined are raw data, MAS5, RMA, dChip, and PLIER. Weida provided the relationship to the

Regulatory Biostatistics WG that serves as a review source and independent evaluation of the analysis efforts. He then discussed the Statistical Analysis Plan (SAP) that provides an overview of each group's experimental design, description of data set, summary of QC process, and the general analysis approach. Weida finished the discussion by proposing a list of decisions that need to be made as soon as possible: (1) when to submit the SAPs: before seeing the data or after analysis is done? (2) how to interact with RBWG: Anonymous or close and direct interactions? (3) what to be submitted: single classifier or multiple classifiers?

- **Pierre Bushel** (NIH/NIEHS), **Mauro Delorenzi** (Swiss Institute of Bioinformatics), **Venkata Thodima** (University of Southern Mississippi), **Roger Perkins** (FDA/NCTR), **Tieliu Shi** (Chinese Academy of Sciences), **Russ Wolfinger** (SAS Institute Inc.), and **John Zhang** (Systems Analytics Inc.) each provided a summary of their classifiers using the Hamner data set. Despite the large number of different individuals independently analyzing this data set, they all reached similar conclusions on the predictive nature of the data set. These analysis teams were commended for their hard and collaborative work over the past several months under Weida's coordination. The disparity of the estimated performance measures of the models appeared to be a reflection of some biases that might have been introduced in some of the analysis processes. Future development and RBWG review of Statistical Analysis Plans (SAPs) should focus on eliminating such biases; close interactions between data analysis teams and the RBWG review teams are essential. **Tim Davison** (Asuragen) ended this session with a summary of the reviews conducted by the Regulatory Biostatistics WG on the submitted SAPs.

Session 1-D: **Availability of Tissue/RNA Samples and Generation of New Data**
Co-Chairs: **Fraser Symmans** (MD Anderson Cancer Center) and **Federico Goodsaid** (FDA/CDER)

Information about new data sets and tissue/RNA samples available to the MAQC-II was presented. Platform providers confirmed their support in generating additional data for addressing specific questions determined by the MAQC-II.

- **Fraser Symmans** (MD Anderson Cancer Center) presented a compelling case for conducting an MAQC-II breast cancer signature reproducibility study. The objective of this study will be to establish the reproducibility of microarray-based tests for human breast cancer between different laboratories evaluating the same tumor sample. Fraser proposes that RNA from operative breast cancer tissue samples be profiled in three independent reference laboratories. Investigators would blindly report gene expression signature results and predictions. An independent data center (e.g. the FDA/NCTR) will evaluate the concordance of those results. In all, Fraser called for two microarray technologies (Affymetrix and Agilent) to be studied since most of the published breast cancer classifiers were based on these two platforms. The second objective of this study would be to establish the reproducibility of microarray-based tests for human breast cancer between cytologic and tissue samples from the same tumor. A fourth reference laboratory would profile a cytologic sample from every breast cancer with matching tissue sample that was profiled in triplicate. Fraser provided details on obtaining the samples, the collaborators doing the collecting, as well as the logistics for RNA purification/QC and processing laboratories. The study proposal received a very positive response from the MAQC-II attendees. Fraser acknowledged the statistical input from Christs Hatzis (Nuvera Biosciences Inc.) and Sue-Jane Wang (FDA/CDER) on sample size considerations.
- **Christine Desmedt** (Jules Bordet Instituut, Belgium) discussed the Amsterdam 70-gene and the Rotterdam 76-gene signatures that have been independently validated using the same TRANSBIG validation series. The Amsterdam 70-gene test is robust (laboratory reproducibility) and available for patient diagnostic testing. Christine then discussed the TOP Trial, which is a prospective evaluation of topoisomerase II alpha gene amplification and protein overexpression as markers for patient outcome prediction of epirubicin primary treatment of breast cancer. One hundred and five (105)

patients were included over eight centers. Eighty (80) cases were profiled using Affymetrix HG133Plus2.0 arrays with the preliminary analysis completed in collaboration with MD Anderson and IGR. The prospective study is ongoing.

- **Vincent Bertholet** (Eppendorf, Belgium) described the availability of a new data set and tissue/RNA samples for breast caner from the BreastMed Consortium. The BreastMed Consortium is an international project funded by the European Union, representing genetic profile comparisons of breast cancers in Mediterranean countries with implications in preventative and predictive medicine. The BreastMed RNA samples are from the following three categories of breast cancer: hereditary, consanguine, and sporadic. There are 95 samples that have greater than 10 ug of total RNA remaining. In addition, detailed clinical information is available for each sample.
- **George Mulligan** (Millennium Pharmaceuticals Inc.) presented details on a multiple myeloma data set with an overview of the disease. Analysis results were provided for 124 different myeloma biopsies processed in triplicate on Affymetrix arrays in two different sets. In all, 84 patient samples were analyzed across six replicates to assess the variance component and estimate the variance component model for each probeset. George then gave an overview of the complete MPI myeloma data set consisting of 264 cases and future data analysis objectives.
- **Wendell Jones** (Expression Analysis Inc.) presented additional information on the multiple myeloma data set from the University of Arkansas for Medical Sciences (John Shaughnessy). This data set is comprised of Affymetrix U133Plus2.0 CEL files on purified myeloma plasma cells from 351 consecutive cases enrolled on Total Therapy 2 (TT2) and 181 enrolled on TT3. An additional ~100 cases, enrolled on TT3, have been put on the U133Plus2.0 arrays and that data can be made available to MAQC-II as well. There are also U133A/B data on 146 of the 351 TT2 cases for which U133Plus2.0 data are available. In addition, there are ~500 cases on U133A/B and/or U133Plus2.0 that have been treated off front line protocols. The UAMS data set and the MPI data set could be used as a mutual external validation set.
- **André Oberthür** (University of Cologne, Germany) gave a presentation on new data and samples about neuroblastoma, a tumor of the sympathetic nervous system. A custom neuroblastoma microarray manufactured by Agilent was used in this study. The array is comprised of 10,263 pre-selected genes with biological or clinical relevance. Two hundred and fifty one (251) neuroblastoma samples were processed in the two-color dye-flip experiments, resulting in 502 arrays. Material of about 100 tumor RNA/tissue samples is available for further analyses on other microarray platforms/laboratories. A second data set is upcoming with 250-300 additional neuroblastoma samples, and completion of expression analyses is expected by the end of July, 2007. Tumor RNA/tissue is available for at least 150 of these second-set samples. These samples could be made available to MAQC-II participants under certain conditions.
- **Rusty Thomas** (The Hamner Institutes for Health Sciences) described the generation of new data at The Hamner Institutes for predicting chemically-induced lung and liver tumorigenicity. The submitted data set is comprised of 13 chemical treatments plus vehicle controls. The prospective data set (at FDA/NCTR) is comprised of several chemical treatments plus vehicle controls. The scheduled new studies (beginning in summer of '07) will consist of 12 chemical treatments plus vehicle controls and in addition one chemical treatment in 5-point dose response.
- **Christos Hatzis** (Nuvera Biosciences Inc.) expanded upon Fraser's proposal with a discussion on the power analysis he performed to determine the appropriate number of samples (arrays) needed in the reproducibility study proposed by Fraser. Christos reiterated this study's objective, that is, to show that gene expression-based indices are consistent or equivalent when measured by three different laboratories following the same protocol. His power analysis showed that 125 arrays are needed at each of the three sites to achieve a power level of 98.2%. Therefore, the proposal Fraser and Christos together outlined would require 125 arrays being processed at each of the three laboratories.

Confirmation of Support from Representatives of Platform Providers

- **Paul Haje** (TeleChem ArrayIt) played a video clip to illustrate why TeleChem is committed to the MAQC-II project. The film clip is Chapter 13 from the PBS program "Cracking the Code of Life". This 5-minute video presents film clips from the movie *GATTACA* and explains how DNA chips would allow doctors to screen babies for numerous diseases and discusses some of the implications of this type of information. The video clip can be viewed from the following link (select Chapter 13): http://www.pbs.org/wgbh/nova/genome/program_t.html.
- Paul Haje asked each platform provider to stand up and announce whether they support the MAQC-II project. The platform providers in attendance at this meeting announced support for the efforts outlined by the MAQC-II group. Specific details will need to be provided in written proposals so array manufacturers can determine their individual level of support. Affymetrix (**Xu Guo**), Agilent (**Anne Bergstrom Lucas**), SuperArray (**Paul Nisson**), Gene Express (**Nick Lazaridis**), TeleChem ArrayIt (**Paul Haje**), Eppendorf (**Katrin Welzel**), NimbleGen (**Tsetska Takova**), and PhalanxBiotech (**Charles Ma**) had representatives in attendance who expressed interest in supporting the MAQC-II project. Illumina, Panomics, and ABI, previously pledged support, were not represented at this meeting so Leming Shi will follow up with these platform providers to discuss next steps and their anticipated level of involvement.
- **Leming Shi** (FDA/NCTR) summarized what have been discussed on day one and reviewed the agenda for day two. All meeting participants were invited to diner, generously sponsored by JMP. It has been a great opportunity for MAQC-II participants to exchange ideas and to know each other better.

May 25, 2007 (Day Two)

Session 2-A: FDA Regulatory Perspectives on Genomics
Chair: **Uwe Scherf** (FDA/CDRH)

FDA representatives discussed the regulatory and scientific perspectives on the use of genomic information in the development of medical products such as devices and drugs.

- **Reena Philip** (FDA/CDRH), chief reviewer of Agendia's MammaPrint® application, gave a presentation on "MammaPrint®: FDA Review and Approval Process". MammaPrint® is a microarray-based IVD breast cancer prognostic test. Service is performed solely in a central CLIA and CAP accredited and ISO certified laboratory at Agendia Amsterdam. MammaPrint® microarrays are manufactured by Agilent Technologies as a custom array design. The MammaPrint® microarrays contain eight 1,900-feature subarrays per 1x3" slide. There are 232 reporter genes per array in triplicate including a 70 gene MammaPrint® prognostic profile. In addition, there are 915 normalization genes and 290 spots for hybridization and printing QC on the same array. The sample preparation involves cutting frozen tumor sections, followed by staining to determine the tumor cell percentage, RNA isolation, Bioanalyzer QC, RNA amplification and labeling (two-color dye-flip with Cy3 and Cy5), and then overnight hybridization. Reena discussed the regulatory route for clearance. Agendia submitted *de novo* 510 (k) in September 2006 and the test was cleared on February 6, 2007 (decision summary is available at http://www.fda.gov/cdrh/reviews/K062694.pdf). Special Controls guidance was issued on May 9, 2007 (http://www.fda.gov/cdrh/oivd/guidance/1627.pdf). During the "Intended Use" section of the talk, Reena described the definitions of Prognostic *vs.* Predictive. With regard to intended use, the MammaPrint® result is indicated for use by physicians as a prognostic

marker only, along with other clinicopathological factors. Reena described the analytical performance studies and specimen requirements.

- **Federico Goodsaid** (FDA/CDER) gave a talk entitled "Companion Guidance to the FDA Pharmacogenomics Guidance," which discussed the areas that need guidance in the microarray field. He stated that guidance documents follow the science and not the other way around. In addition, guidance documents should give new information not otherwise known. The draft Companion Guidance was developed as a result of experiences gained through the VGDS and MAQC-I. Federico cited specific areas for guidance with expression microarrays such as RNA isolation, handling, characterization, labeling, and biological interpretation. Genotyping guidance was also discussed such as the genotyping method, DNA handling, and genotyping report. Also touched upon were proficiency testing, genomic data in clinical study reports, genomic data from non-clinical toxicology studies, and data submission format. Possible areas for future guidance are classifiers, quantitative RT-PCR, proteomics, and metabolomics.
- **Bob Wagner** (FDA/CDRH) gave a thought-provoking presentation on uncertainties in bioinformatics, machine learning, and the multiple-biomarker classifier problem. He provided an overview covering some fundamental concepts from the field of Statistical Learning Machines and Statistical Pattern Recognition. Uncertainties of classifier performance from finite training, a statistical/historical quirk-embedded in analysis of training, and two random effects being finite training plus finite testing leading to variance. A recently proposed solution is an approach to estimate that total variance (or error bars) and a suggested three-stage approach to classifier assessment. Bob concluded his talk with a summary of an ideal world having the following characteristics: (1) carry out pre-clinical research and then data mining for feature and architecture selection; (2) obtain new samples for pilot study and analyze via Paradigm I; (3) estimate mean performance and uncertainties (e.g., via hold-out bootstrap and influence function); (4) at the end train again with entire available set (this defines the interim classifier); (5) use uncertainties to design a pivotal study in the form of Paradigm II. Please refer to Bob's presentation for the fine details that are difficult to summarize without seeing some of the equations that he used.

Session 2-B: Genome-Wide Association Working Group (GWAWG)
Co-Chairs: **Federico Goodsaid** (FDA/CDER) and **Leming Shi** (FDA/NCTR)

After an excellent keynote presentation by Teri Manolio (NIH/NHGRI) on genome-wide association studies, a proposal for establishing the Genome-Wide Association WG (GWAWG) under MAQC-II was presented by Federico Goodsaid (FDA/CDER) and Nianqing (Nick) Xiao (NCI/SAIC), followed by discussion. Leming Shi (FDA/NCTR) announced that the proposal was approved by meeting participants, and named Federico and Nick as coordinators of the GWAWG, the 5th WG under MAQC-II.

- **Teri Manolio** (NIH/NHGRI) gave an excellent keynote presentation entitled "Genome-Wide Association Studies: New Paradigm or the Same Old Genes?" which started with a discussion on what is a Genome Wide Association (GWA) Study. Teri described a GWA study as a method capable of interrogating all 10 million variable points across the human genome. Variation is inherited in groups, or blocks, so not all 10 million points have to be tested. Blocks are shorter (so that more points are needed to be tested) the less closely people are related. She pointed out that technologies now allow studies in unrelated persons, assuming ~10,000 base pair lengths in common (300,000 – 500,000 markers). Teri then provided an overview on how relationships among SNPs are mapped. One slide illustrates the extraordinary progress in genotyping technology with cost per genotype *vs.* SNP number and year. Another slide illustrates cost per person *vs.* year between the Affymetrix 500K and Illumina microarray technologies. Teri also provided 2007 cost numbers for a GWA study in 2,000 people which is approximated at a total cost of $1 million dollars, or 0.1 cent per SNP. Published results were then provided for GWA scans for age-related macular degeneration, prostate

cancer, and type 2 diabetes. Teri stressed the need for consensus on what constitutes replication in association studies, especially considering the "avalanche" of GWA and candidate gene studies now and in the near future. The NCI-NHGRI has a working group on replication in association studies and it is important to have sufficient sample size to convincingly distinguish proposed effect from no effect. Teri provided a number of different cited publications illustrating the scientific value of GWA studies that are detecting robust associations never previously suspected. She also emphasized the importance of data sharing in GWA studies and pointed to NCBI resources such as dbGaP. (*Note from Leming Shi*: Teri and her colleagues recently developed an excellent document on "What Constitutes Replication of a Genotype-Phenotype Association? Summary of an NCI-NHGRI Working Group", S Chanock, T Manolio, *et al.*, *Nature* 2007, in press. This document should serve as an excellent starting point for the MAQC GWAWG effort. MAQC participants are encouraged to read this paper when it is published within a few weeks. Leming thanks Stephen Chanock for sharing a draft of the manuscript.)

- **Nianqing (Nick) Xiao** (Core Genotyping Facility at NCI/SAIC) gave a presentation entitled "Genome-Wide Association Working Group (GWAWG): Can what we learned from MAQC help?", which began with the major applications for GWA technologies. These major applications are (1) Comparison: Identify disease-associated markers/genes for better understanding of disease etiology and identify therapeutic targets (CGEMS, GEI, GAIN, etc.) and (2) Prediction: Predict efficacy and adverse event based on genotypes – personalized medicine. Nick illustrated the diversity within the technology from Whole Genome Scan (Illumina Infinium and Affymetrix 500K/5.0), Large Multiplex (Illumina GoldenGate and Affymetrix ParAllele), Medium Multiplex (Sequenom iPLEX and SNPlex), and Single-plex (TaqMan and MGB Eclipse). Nick discussed the genome-wide scan *vs.* candidate gene approach, illustrating the value in WGA studies. Nick gave a bullet point list for data quality assurance: (1) Completion rate: by sample and by SNP; (2) Concordance between duplicates; (3) Concordance with reference data (HapMap); (4) Hardy–Weinberg Equilibrium (HWE); (5) Mendelian check; (6) Cryptic relatedness; and (7) Population structure and admixture. An overview of data analysis metrics was provided, followed with a discussion on how the MAQC experiences can help with future direction for GWA technologies. The concluding slide of Nick's talk focused on what needs to be determined next: the scope for platforms, planning for multiple phases (comparison, prediction, combined expression and genotype analysis), data sources, identifying participants and expertise in this area.
- **Federico Goodsaid** (FDA/CDER) presented on "Genome-Wide Association Studies and FDA's Voluntary eXploratory Data Submission (VXDS)". The VXDS goals are (1) Training reviewers in the analysis and interpretation of microarray data and integration of this training within reviewer training in pharmacogenomics; (2) Training sponsors in the capabilities of our reviewers for the analysis and interpretation of microarray data; (3) Development of a review strategy for microarray data including both data analysis protocols applied by the sponsor as well as modifications to these analysis protocols; (4) Development of a review strategy for discriminating between the reproducibility of biomarker definitions and the reproducibility of the biological interpretation of microarray data; (5) Assessment of strengths and weaknesses in current microarray data analysis protocols; (6) Identification of technical issues in the analysis protocols included in the Draft Companion Guidance for the Pharmacogenomic Guidance covering the generation and submission of genomic data; (7) An accessible database for microarray data in VXDS; and (8) A basic framework for infrastructure required in electronic submission of microarray data. Federico provided detailed diagrams of the Whole Genome Scanning assays of Affymetrix and Illumina. He also touched upon classifier development and noted that the genotyping assays have similar classification considerations as in MAQC-II for the gene expression assays. In particular, there are multiple combnations with considerations of QC, normalization and appropriate algorithms for analysis. Federico provided a draft work plan that calls for raw data files from Whole Genome Scanning Affymetrix or Illumina arrays. The last slide of Federico's presentation provided a list of what is needed: training data sets,

test data sets, multiple classifiers per training data set, and confirmation in test data set for each classifier.

- After the discussion, **Leming Shi** (FDA/NCTR) announced the decision to establish the Genome-Wide Association Working Group (GWAWG), the fifth WG under MAQC-II, and named **Federico Goodsaid** (FDA/CDER) and **Nianqing (Nick) Xiao** (NCI/SAIC) as coordinators of the GWAWG. The first tasks for the group will be (1) to identify experts in academia, industry, and government who will be willing to assist with the project, and (2) to identify appropriate data sets for developing genotype-phenotype association and prediction models. Leming noted that GWA studies present several new challenges for the microarray community, and there is a disconnection between people working in gene expression and those in genotyping (GWA). However, GWA studies share a lot of similarity with what we have been seeing in gene expression. Recently published studies focus on the identification of SNPs that are associated with phenotypes such as disease susceptibility, which is similar to the identification of differentially expressed genes (investigated under MAQC-I). The ultimate utility of SNP data, and the goal of the GWAWG, will be dependent on the development of SNP-based classification models that are clinically predictive of patient outcomes (diagnostics, prognostics, efficacy, or toxicity), which is similar to the existing focus of MAQC-II (development and validation of predictive models). For personalized medicine to be realized, we have to be able to make an accurate prediction for each patient based on information from gene expression, genotyping, or both. It is expected that the MAQC project will provide a unique opportunity for researchers in gene expression and genotyping to collaborate with each other, share experiences, and understand each other's "language". Interested participants should contact Federico.Goodsaid@fda.hhs.gov and xiaon@mail.nih.gov (and cc Leming.Shi@fda.hhs.gov) to join the GWAWG.

Session 2-C: Data Analysis Strategies
Chair: **Tim Davison** (Asuragen Inc.)

Several speakers presented their views on the "best practices" for the development and validation of predictive models using microarray data. **Weida Tong** (FDA/NCTR) shared his experience with the process of RBWG review of Statistical Analysis Plans (SAPs) for the Hamner TGx data set. During the discussion, Uwe Scherf (FDA/CDRH) emphasized the importance of direct interactions between RBWG review teams and data analysis teams. The existing "anonymous" RBWG review practice, which may be preferred by a few reviewers, prevents such direct interactions and does not appear to be consistent with MAQC's open collaboration spirit. **Gene Pennello**, **Lakshmi Vishnuvajjala** (FDA/CDRH) and other biostatisticians raised concerns on multiplicity when multiple models are proposed for "validation".

- **Kenneth Hess** (MD Anderson Cancer Center) gave a presentation on the development and assessment of microarray predictive models. Kenneth stressed that identified biomarkers need to be tested in good, large, prospective studies. Multiplex predictive markers were described to select and combine markers from thousands of candidates using only dozens of samples. The solution is supervised learning models since the rules for marker selection and integration are learned from the data. Kenneth pointed out that supervised learning is not new and we need to build on decades of research in computer science (machine learning), engineering (pattern recognition), and statistics (data mining). It is important to take into serious consideration the bias-variance trade-off that operates differently for prediction. Using too few training samples leads to under-fitting bias. Including all model selection steps in cross-validation reduces bias. Using too few testing samples leads to excessive variance. Cross-validation reduces bias, but increases variance. Kenneth summarized the key issues: (1) Minimum number of training samples? (2) Which type of normalization? (3) Which genes? (4) How many genes? (5) Which prediction algorithm? (6) Which performance measure? (7) Which K in K-fold CV? (8) Minimum acceptable accuracy? (9) How to choose classifier parameters? (10) Minimum number of testing samples? These are difficult questions

to answer and a good strategy is needed to use many data sets to assess and refine strategies. MDACC used 12 public Affymetrix data sets, each having at least 40 clinical samples. Gene filtering is an important consideration since noisy genes can reduce prediction accuracy. The approaches for gene selection are Univariate Ranking and Multivariate Searching, with the end result being that "simpler may be better". Kenneth stressed the importance of independent validation. Calculations indicate that 100 to 200 samples are needed for independent validation.

- **Russ Wolfinger** (SAS Institute Inc.) described a strategy to explore factors involved in predictive modeling analysis. He explained the process as wandering through a fairly complex space of many dimensions. Russ described initial data preprocessing consisting of non-background subtraction, background subtraction and other instrument-specific methods. In addition, he described different data transformation, summarization, and normalization methods that could be applied to the data. Different gene (predictor) filtering methods could be used. A large number of predictive modeling methods were outlined as well as cross-validation methods. The performance criteria included accuracy, specificity, sensitivity, AUC, pAUC, and RMSE. Unfortunately, there are millions of combinations so Russ provided some methods for covering these combinations. One possibility is an experimental design approach to derive a subset of combinations to cover the space. The talk ended with final thoughts about how model averaging could reduce variance and boost confidence. CV is highly parallelizable; linking idle computers can help with the analysis load. Russ' talk gave us a lot to think about, while providing new insights.

- **Jun Kanno** (National Institute of Health Sciences, Japan) gave a talk on Percellome: "Per Cell" Normalization Method for mRNA Measurements by Quantitative PCR and Microarrays. Details on this method can be found in Kanno *et al.*, *BMC Genomics* 2006 that describes the "Per cell" normalization method. The goal is to obtain transcriptome data as copy number of mRNA per cell (average) for the direct comparison of mRNA expression data. Essentially, a grade-dosed spike cocktail (mRNAs) are used that do not cross-hybridize with sample mRNA and are in proportion to the sample cell number. This cocktail contains five mRNAs of known copy numbers that serve as quantitative standards. Refer to Kanno's presentation for many informative data analysis graphs.

- **Weida Tong** (FDA/NCTR) opened the floor for discussion on the RBWG review and SAPs. Questions were raised with respect to submitting multiple classifiers versus a single classifier for "validation". In particular, do we want to generate two groups of classifiers: reviewed and not-reviewed? How critical is the multiplicity issue related to submitting multiple classifiers? Do two classes of classifiers add additional/unique flavor to MAQC-II? Also, questions were proposed on when to release the confirmatory data. These questions can be resolved in future discussions.

- During open discussion at the end of this Session, **Kenneth Hess** (MD Anderson) made a very important comment regarding the "best practice" in formulating a data analysis strategy for microarrays: "*If a Statistical Analysis Plan (SAP) is hardwired to a specific data set, we will more likely end up fitting to noise.*" We should keep this advice in mind as we enter into full-scale model development/validation phase so that we may minimize the chance of over-fitting. **Uwe Scherf** (FDA/CDRH) emphasized the importance of direct interactions between RBWG review teams and data analysis teams and pointed out that the current "anonymous" RBWG review practice, while preferred by a few reviewers, prevented such effective direct interactions and did not appear to be consistent with MAQC's open collaboration spirit. **Gene Pennello**, **Lakshmi Vishnuvajjala** (FDA/CDRH) and a few other biostatisticians commented on multiplicity issues when multiple models are proposed for "validation". During follow-up discussions with key members of the RBWG, **Leming Shi** urged the review teams to directly communicate with data analysis teams so that RBWG's review comments may have a better chance to help improve the data analysis process. Leming also stated his plan of distributing confirmatory data sets only until after each data analysis team gets a chance to explore all the MAQC-II training data sets, learn from each data set, and finalize the SAP that will be applied to all data sets.

Session 2-D: Data Analysis Teams and Coordinators
Chair: **Wendell Jones** (Expression Analysis Inc.)

On behalf of Leming Shi (FDA/NCTR), **Wendell Jones** announced the formation of multiple data analysis teams and the appointment of their corresponding coordinators. This list of data analysis tasks and the coordinators (Table 1) was proposed by **Leming Shi** based on his extensive interactions with many data analysis groups over the past several months. Hopefully, a manuscript will be developed from each data analysis task through the collaborative efforts within each task team. Leming used the following criteria during the selection of data analysis coordinators: (1) commitment of the proposed coordinator or his group in conducting hands-on data analysis with enthusiasm; (2) the coordinator is not the data provider; (3) the coordinator is preferred to physically reside in the US for logistic reasons (e.g. time differences); and (4) the degree of the coordinator's involvement in the MAQC-II over the past several months. MAQC-II participants who would like to be part of the data analysis efforts and to make intellectual contributions to the data analysis tasks should contact the coordinators directly (and cc Leming.Shi@fda.hhs.gov). The data analysis coordinators have been instructed by Leming Shi to be as inclusive as possible in terms of team members. All data analysis coordinators at the meeting restated their commitments to coordinating the data analysis tasks, and four of them, **Xutao Deng** (UCLA/CSHS), **Richard Judson** (EPA/NCCT), **John Zhang** (Systems Analytics Inc.), and **Gene Pennello** (FDA/CDRH), made brief presentations on the tasks assigned to them and extended their invitations to meeting attendees for participation. The last hour of the meeting was reserved for meeting participants to freely present their results and/or to make comments about the future directions of the MAQC project.

- **Xutao Deng** (UCLA/CSHS) gave a presentation describing a preliminary analysis of the MDACC breast cancer data set. Xutao summarized the CEL data with PLIER, followed by adding an offset value of 16 (a procedure recommended by Affymetrix in MAQC-I) and log2 transformation. All probe sets were included. The classification algorithms are Naïve Bayesian (NB) and KNN. The NB classifier using only 6 or 7 genes can achieve accuracy 88%. Using KNN we see FC biomarkers still achieved the best performance in about 30 biomarkers, also about 88%. The other methods produce very similar results. Xutao finished the talk with some conclusions from the preliminary analysis. In particular, FC produces more reproducible biomarker set than does p-value or CR in this clinical data set. Higher reproducibility of biomarker set potentially leads to better prediction. Surprisingly, even random feature sets can achieve ~80% accuracy (Leming's notes: This might be due to the RD prevalence (~75%) of both the training and validation sets. In addition to accuracy, sensitivity and specificity should be examined in future analysis.) A modest and reproducible accuracy, e.g. 85%, is a reasonable goal for this data set. After all, classifiers are functions applied to biomarkers sets, if the biomarker goes wrong, the classifier become useless. The goal here is always to identify robust biomarkers rather than to find the universal classifier or models.
- **Richard Judson** (EPA/NCCT) gave a presentation on the Iconix Data Analysis Project. The data set characteristics were provided where the endpoint predicted is non-genotoxic hepatic tumorigenicity in rats. One hundred and forty seven (147) compounds (100 compounds in training set and 47 compounds in test set, in Iconix's original data analysis). Positive class: either known to cause liver tumors in two-year carcinogenicity assay in one strain or gender of rat based on literature annotation, or reasonably expected to be tumorigenic to the liver based on class effect (e.g. PPAR agonists). Single high dose (MTD) per compound (MTD= dose that causes 50% decrease in weight gain in five-day range-finding study, or pathology inducing). Three rats per dose-time combination, 20 control rats per dose-time combo with rat livers profiled on 10K Codelink arrays. In addition, five compounds tested by EPA under the same experimental protocols will be used in the MAQC-II validation set. David Dix (EPA/NCCT) and Hong Fang (FDA/NCTR) will split the data set into training and validation sets within the next two weeks. The array data of training set will be

distributed to MAQC data analysis teams for array QC assessment, followed by the distribution of endpoint information of the training set for model development.

- **John Zhang** (Systems Analytics Inc.) gave a presentation describing batch effect removal on model prediction performance. There are many sources for batch effect such as chip type/platform, RNA isolation batch, hybridization batch, wash and stain, reagent batches, chip lot, and operator. In the Hamner TGx data set, there may be other confounding factors such as biology since the animals may have been treated differently in the subsequent experiment. Differences are seen between '05 and '06 data sets but there are many confounding variables that could be causing this effect. RMA and MAS5 normalization procedures were assessed and noise removed. When the batch effect was removed by adjusting '05 data to the level of '06, the prediction accuracy of '05 data based on models developed on '06 data was improved. This procedure will be considered for other MAQC-II data sets.
- **Gene Pennello** (FDA/CDRH) presented on "Regulatory Biostatistics WG: Statistical Methodologies". Gene described the SAP review process, external validation process, performance measures, and good practices. For the Hamner toxicogenomics lung tumor data set, the SAPs were submitted, reviewed by the RBWG, revised, and new classifiers built. The external validation process will be to release the validation data set in full and without class labels. Gene outlined possible factors to explore such as feature selection, internal validation scheme, algorithm (SVM, DLDA, Naïve Bayes, etc.), normalization (RMA, MAS5, etc.), number of features, and number of items adhered to on SOP checklist. The Regulatory Biostatistics WG's effort is to provide scientific and statistical advice on good practices for building and validating predictive classifiers. For a model to be "validated", it should be completely fixed before the confirmatory data set is distributed. For the MAQC to search for factors that produce the best and worst models, multiple classifiers resulting from different analysis procedures will be proposed and tested with the confirmatory data set. Gene also pointed out the multiplicity issue due to the large number of models to be reviewed by RBWG, and eventually assessed by the confirmatory data set. He also suggested alternative ways of removing batch effects to make the process as less biased as possible.

Table 1: MAQC-II Data Analysis Tasks and Coordinators

No.	Data Analysis Task	Coordinator	E-mail
Date Set (Disease) Specific Task			
1	Hamner mouse lung tumor	Weida Tong	weida.tong@fda.hhs.gov
2	Iconix-EPA rat liver cancer	Richard Judson	judson.richard@epa.gov
3	NIEHS rat liver toxicity	Simon Lin	s-lin2@northwestern.edu
4	Breast cancer	Xutao Deng	xutao.deng@cshs.org
5	Neuroblastoma	Russ Wolfinger	russ.wolfinger@sas.com
6	Multiple myeloma	Roger Perkins	roger.perkins@fda.hhs.gov
7	MAQC-I titration pilot data	Marc Salit Russ Wolfinger	salit@nist.gov russ.wolfinger@sas.com
Problem-targeted Tasks			
9	Array QC and prediction performance	Wendell Jones	wjones@expressionanalysis.com
10	Prediction reproducibility across laboratories	Fraser Symmans	fsymmans@mdanderson.org
11	Batch effect and prediction performance	John Zhang	johnz@systemsanalytics.com
12	Statistical methodologies	Gene Pennello Lakshmi Vishnuvajjala	gene.pennello@fda.hhs.gov lakshmi.vishnuvajjala@fda.hhs.gov

To join the data analysis efforts, contact the coordinators and cc Leming.Shi@fda.hhs.gov.

Open Presentation and Discussion

- **Guozhen (Gordon) Liu** (SuperArray Bioscience Corp.) presented his classifiers developed from the MDACC breast cancer data set and compared the probesets used in his classifiers with those used in the MDACC study's original DLDA model.
- **Jean Thierry-Mieg** (NIH/NCBI) presented a talk on the quality control lessons obtained from the MAQC-I data set. There is great value in using all data and self consistent optimization. Jean and Danielle Thierry-Mieg thought that titration is more powerful than p value or fold change. Jean and Danielle averaged across labs the 15 values for each sample, removed outliers, optimized normalization to maximize the number of titrating probes. Probes should titrate in one of two different directions: Either A<C<D<B or A>C>D>B since A and B represent independent samples while C and D are mixtures of A and B. To optimize normalization, Jean and Danielle recommend ignoring the qualitative calls, using the unprocessed data, centering over non-differentially expressed genes, and removing outlier arrays. The biggest gain in titrating probes occurred when ignoring qualitative calls and using unprocessed data rather than processed data. Maximizing the number of titrating probes actually maximized the number of genes coherently seen as differentially expressed across all platforms. Approximately 14,000 genes titrate in at least three platforms. Probes from these genes show 70.71% are co-titrating, 3.44% are discordant, 25.86% are not titrating when they should. Jean provided evidence for improving probe design by looking at titration success versus probe Tm. Co-titration is highly dependent on Tm. Mapping probes to the complete transcriptome also matters. AceView is a non-redundant comprehensive manually supervised summary of all cDNA sequences in GenBank and dbEST into genes and alternative transcripts. There is evidence for at least 58,000 main human genes and 190,000 spliced transcripts. All platforms are remarkably coherent.
- **Jim Willey** (Gene Express Inc.) made a brief presentation on gene-by-gene array QC assessment (presentation not available for distribution).
- **Guy Tillinghast** (Riverside Cancer Care Center) commented at the meeting from a physician's perspective on the development and validation of predictive models based on microarray data. Modelers should think about how their microarray classifier would be used in the clinic. Rather than competing with clinical information, an ideal microarray predictor would complement what is already clinically known and used, especially when there are areas of clinical ambiguity. There is a wide gap in the public's understanding of gene-based diagnostic test capabilities and the reality. This gap will need to be addressed when it comes time for informed consent and when meeting IRB approval. The chief advantage of microarray (over competing technologies such as PCR) is the ability to contain multiple classifiers within one laboratory test. Guy noted that none of the available MAQC-II data sets address the issue of selecting the best drug by which to treat a patient. The NCI-60 cell line data, however, may be a start for addressing this issue. Microarray may provide a unique opportunity of determining whether a patient is too different from the training and validation sets from which the microarray classifier was trained and evaluated. The mandate given to the MAQC appears to be a good one!
- **Leming Shi** (FDA/NCTR) gave a brief summary of the meeting and thanked meeting participants for making the meeting a success. Leming appreciated session chairs and presenters for strictly following the tight meeting schedule. A major outcome of the meeting is the establishment of the 12 data analysis task teams and the confirmation of the coordinators who will be leading the extensive data analysis efforts in the following six months. The coordinators' effective leadership will be essential to keep the MAQC-II on schedule. Leming appreciated the coordinators for their commitments and asked MAQC-II participants to help them. The success of the MAQC project will depend on many individuals' voluntary efforts.

MAQC-II Tentative Timelines

- **9/21/06**: Kickoff meeting at FDA/NCTR
- **11/28,29/06**: Data set and algorithm evaluation at FDA/CDER (6th F2F)
- **12/15/06**: Data sets (raw data) submitted to FDA/NCTR
- **1/31/07**: Data sets distributed to participants
- **5/24,25/07**: Initial analysis results discussed at SAS Institute (7th F2F); Data analysis tasks and coordinators identified
- **Jun-Jul/07**: Manuscript-drafting teams assembled
- **Jun-Oct/07**: Exploratory analysis with all MAQC-II training data sets
- **Oct-Nov/07**: Detailed analysis results discussed at the 8th F2F meeting (TBA); Statistical Analysis Plans finalized; Confirmatory data sets distributed
- **3/31/08**: Submission of manuscripts
- **9/08**: Peer-reviewed publications
- **12/08**: MAQC recommendations on the development and validation of predictive models/classifiers

Disclaimer

- This meeting summary serves as a shortcut to the many presentations made at the MAQC meeting. Please refer to speakers' original presentations for more accurate information of their point of views.
- Participation in the MAQC project is completely voluntary. No fund whatsoever is available from the MAQC to any participant. Participants agree to cover all their own costs as a result of voluntary involvement in the MAQC project. The US Food and Drug Administration (FDA) has solicited DNA microarray gene expression data sets as well as proposals to analyze these data sets in order to evaluate the impact of different analysis protocols on the selection of genes and their associated predictive models for biomarker pattern development (*Federal Register*, 71(77), 20707-8, April 21, 2006; available at http://www.fda.gov/nctr/science/centers/toxicoinformatics/maqc/docs/FederalRegister_MAQC_FollowUp.pdf).
- The MAQC project is being coordinated by the FDA, but there are no regulatory rights conveyed to anyone by the participation of FDA personnel in this project. Although FDA personnel are involved in the MAQC project, the views expressed in the MAQC-II Research Plan, at face-to-face meeting or other circumstances are not FDA guidance and do not necessarily represent FDA policy.

Table 2. Participants of the 7th MAQC Project Meeting, May 24-25, 2007, Cary, NC

No.	Name	Organization	No.	Name	Organization
1	Wenjun Bao	SAS Institute	61	Padraic Neville	SAS Institute
2	William T. Barry	Duke University	62	Paul E. Nisson	SuperArray
3	Anne Bergstrom Lucas	Agilent	63	André Oberthuer	University of Cologne
4	Vincent Bertholet	Eppendorf Array Technology	64	Hiroshi Okada	Okada Consulting Inc.
5	Norman Birchfield	EPA	65	Jason A. Osborne	University of North Carolina
6	Steve Bischoff	North Carolina State University	66	Grier P. Page	University of Alabama
7	Guy Bowman	InforSense, LLC	67	Joel Parker	Constella Group, LLC
8	Anne Bullard	SAS Institute	68	Richard S. Paules	NIH/NIEHS
9	Pierre Bushel	NIH/NIEHS	69	Thomas Pedersen	SAS Institute
10	Gregory Campbell	FDA/CDRH	70	Gene A. Pennello	FDA/CDRH
11	Jennifer G. Catalano	FDA/CBER	71	Edward J Perkins	US Army Engineer R&D Center
12	Damien Chaussabel	Baylor Inst for Immunology Res	72	Roger Perkins	FDA/NCTR (Z-Tech)
13	Tzu-Ming Chu	SAS Institute	73	Ron Peterson	Novartis
14	Shannon Conners	SAS Institute	74	Reena Philip	FDA/CDRH
15	Wendy Czika	SAS Institute	75	Vitali Proutski	Almac Diagnostics
16	Timothy S. Davison	Asuragen	76	Laura H. Reid	Expression Analysis
17	Francoise de Longueville	Eppendorf Array Technology	77	Todd Richmond	NimbleGen Systems, Inc.
18	Mauro Delorenzi	Swiss Institute of Bioinformatics	78	Marc Salit	NIST
19	Xutao Deng	UCLA/Cedars-Sinai	79	John Sall	SAS Institute
20	Youping Deng	University of Southern Mississippi	80	Uwe Scherf	FDA/CDRH
21	Christine Desmedt	Institut Jules Bordet	81	Banalata Sen	EPA
22	Pan Du	Northwestern University	82	Imran Shah	EPA
23	Stephen W Edwards	EPA	83	Ruchir Shah	Constella Group, LLC
24	Meg G Ehm	GlaxoSmithKline	84	Joe Shambaugh	Genedata (USA) Inc.
25	Fathi Elloumi	EPA	85	Leming Shi	FDA/NCTR
26	Patton Fast	InforSense, LLC	86	Tieliu Shi	Chinese Academy of Sciences
27	Kazuhisa Fukushima	Yokogawa Electric Corp.	87	Weiwei Shi	GeneGo Inc.
28	James C. Fuscoe	FDA/NCTR	88	Richard Shippy	Affymetrix
29	Federico M. Goodsaid	FDA/CDER	89	Dave D. Smith	Luminex
30	Xu Guo	Affymetrix	90	W. Fraser Symmans	MD Anderson Cancer Center
31	Paul K. Haje	TeleChem ArrayIt	91	Tsetska Takova	NimbleGen Systems, Inc.
32	Christos Hatzis	Nuvera Biosciences, Inc.	92	Pei-Yi Tan	SAS Institute
33	Kenneth Hess	MD Anderson Cancer Center	93	Danielle Thierry-Mieg	NIH/NCBI
34	Susan Hester	EPA	94	Jean Thierry-Mieg	NIH/NCBI
35	Lingkang Huang	NIH/NIEHS	95	Venkata Thodima	University of Southern Mississippi
36	Robin Hughes	SAS Institute	96	Russell S. Thomas	The Hamner Inst for Health Sciences
37	Melody Humbles	Tecan	97	Guy Tillinghast	Riverside Cancer Care Center
38	Ansar Jawaid	AstraZeneca	98	Bernadette Toner	GenomeWeb
39	Brandon D. Jeffy	Iconix	99	Weida Tong	FDA/NCTR
40	Charles D. Johnson	Asuragen	100	Shengdar Tsai	North Carolina State University
41	Wendell D. Jones	Expression Analysis	101	Silvia Vega	Rosetta Biosoftware
42	Richard Judson	EPA	102	Lakshmi Vishnuvajjala	FDA/CDRH
43	Jun Kanno	National Inst of Health Sci., Japan	103	Juergen von Frese	Almac Diagnostics
44	Channa Keshava	EPA	104	Robert F Wagner	FDA/CDRH
45	Nagalakshmi Keshava	EPA	105	Stephen J. Walker	Wake Forest University
46	Samir Lababidi	FDA/CDRH	106	May Dongmei Wang	Georgia Tech and Emory University
47	Nick Lazaridis	Gene Express	107	William Ward	EPA
48	Jianying Li	NIH/NIEHS	108	Jeffrey F Waring	Abbott
49	Wayne Liao	Phalanx Biotech Group	109	Liling L Warren	GlaxoSmithKline
50	Simon Lin	Northwestern University	110	Katrin Welzel	Eppendorf Biochip Systems GmbH
51	Guozhen (Gordon) Liu	SuperArray	111	James C. Willey	University of Toledo
52	Edward K. Lobenhofer	Cogenics	112	Russell D Wolfinger	SAS Institute
53	Kevin Long	Consultant	113	Bill Worzel	Genetics Squared
54	Jun Luo	Systems Analytics	114	Nianqing Xiao	NCI (SAIC)
55	Charles Ma	Phalanx Biotech Group	115	George Yuan	Affymetrix
56	Sergei Makarov	attagene	116	John Zhang	Systems Analytics
57	Geoffrey Mann	SAS Institute	Among the 116 meeting participants from 66 organizations, 82 people came from outside of North Carolina. In addition, ~20 people participated in all or part of the two-day meeting via WebEx or phone.		
58	Teri Manolio	NIH/NHGRI			
59	Stan Martin	SAS Institute			
60	George J. Mulligan	Millennium Pharmaceuticals			