Summary of the 6th MAQC Project Meeting, November 28-29, 2006

Washington, DC and Silver Spring, MD

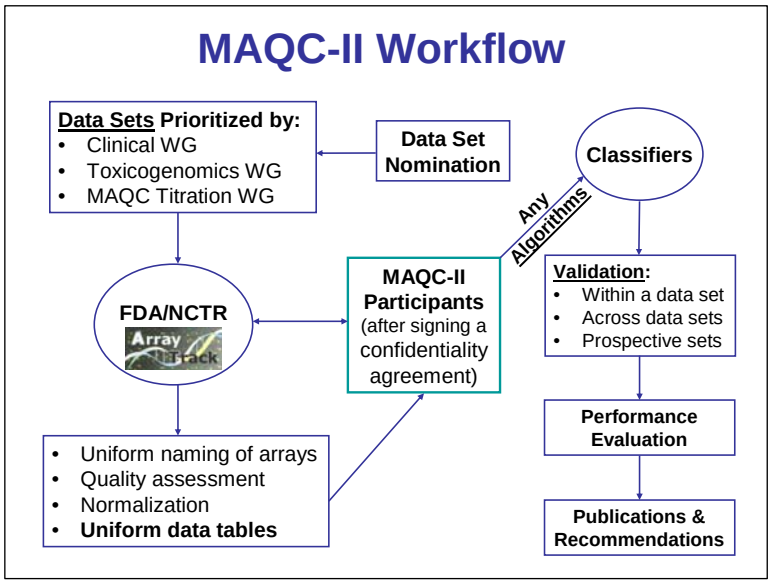
Leming Shi, December 12, 2006
Leming.Shi@fda.hhs.gov; http://edkb.fda.gov/MAQC/

The 6th face-to-face MAQC project meeting was held at the Washington Marriott Hotel on November 28, 2006 (3:00 pm – 5:00 pm) and the Central Shared Use Building of the FDA's White Oak facility at Silver Spring, MD on November 29, 2006 (9:00 am – 4:00 pm). A total of 93 on-site participants attended in addition to 12 people who participated via WebEx. The main objectives of the meeting were: (1) to review the progress of the three working groups (Clinical WG, Toxicogenomics WG, and Titration WG); (2) to discuss the data sets nominated for the Clinical WG and TGx WG; (3) to better define the goals of MAQC Phase II (MAQC-II); and (4) to discuss the criteria for evaluating classifier performance. Meeting participants were impressed by the nominated data sets and expressed strong interests in contributing to MAQC-II. Detailed meeting agenda and presentations (PowerPoint and PDF files) can be found at the MAQC web site http://www.fda.gov/nctr/science/centers/toxicoinformatics/maqc/.

MAQC-II Progress Report

Chair: Yvonne Dragan (FDA/NCTR)

Following an introductory remark by Yvonne Dragan on the importance of MAQC-II, Leming Shi (FDA/NCTR) gave an overview presentation, "MAQC: From Phase I to Phase II", summarizing main findings from MAQC-I and outlining the workflow for MAQC-II. It was emphasized that the main goal of MAQC-II was to reach consensus on the procedures for performance evaluation of different classifiers via three stages: initial discovery (internal validation within a data set); independent validation (cross-study prediction with multiple data sets), and clinical utility and validation (with prospective data sets). Although the prediction accuracy (sensitivity and specificity) is the main criterion for evaluating the performance of a classifier, the robustness and mechanistic relevance of the classifier are also important additional considerations. It is anticipated that a better understanding of the capabilities and limitations of microarrays in clinical and toxicogenomic applications could be reached and recommendations on the development and validation of classifiers (predictive signatures) may be put forward through MAQC-II. Leming announced the establishment of the fourth working group, Regulatory Biostatistics Working Group (RBWG) to be coordinated by Dr. Gregory Campbell, Director of the Division of Biostatistics of FDA/CDRH. Understanding the challenges ahead of MAQC-II, Leming encouraged the MAQC group by concluding his presentation with a quote from Niels Bohr: "It is very difficult to make an accurate prediction, especially about the future."



Uwe Scherf (FDA/CDRH) summarized the Clinical Working Group's (CWG's) progress since the September 21, 2006 meeting. Weekly conference calls were arranged. From the ~20 clinical data sets nominated, the CWG decided to examine data sets from four types of diseases: breast cancer, multiple myeloma, leukemia, and neuroblastoma. Criteria for data set selection were presented, including microarray data quality and sample quality, clinical information, and experimental design. Data sets were selected based on the potential for independent cross-study and prospective validation. It was emphasized that a classifier is useful only for the intended use population. Federico Goodsaid (FDA/CDER) gave an overview of the Toxicogenomics Working Group (TGx). Three teams within the TGx have been formed: data sets team (David Dix et al.), classifiers team (Tim Davison et al.), and validation team (Don Halbert et al.). The similarity in workflow between the CWG and TGx was emphasized and coordination is essential. Rich Shippy (GE Healthcare) described the Titration Working Group's (Titration) goal as providing a "positive control" for the evaluation of the performance of classifiers by using the titration data sets generated by the MAQC-I main study (A, B, C, and D samples) and the MAQC-I Pilot II titration (13 titration mixtures with A and B). Rick Jensen (UMass Boston) later on talked about some initial results from the Titration WG (Russ Wolfinger's group at SAS). If needed, additional titration samples may be created and profiled.

Fraser Symmans (MD Anderson Cancer Center) gave an excellent presentation on the state-of-the-art of gene expression profiling in breast cancer. Accurate prognosis is important, but selection of an effective treatment is even more important; and selection of the best among available treatment regimens could be the most important. Weaknesses in genomic assay performance, sample processing, and/or data analysis could introduce serious flaws, but weaknesses in clinical study design introduce the greatest flaw. The MAQC should pay special attention to clinical study design while analyzing the nominated data sets and interpreting the results.

Kurt Jarnagin (Iconix) shared with the MAQC group of Iconix's extensive experience in the development and validation of drug and toxicology signatures. The derivation of signatures requires identification of carefully considered phenotypes. The "forward validation" of signatures is the key. Criteria for validation should be determined based on fitness-for-use, and should show improvements over current gold-standards, not arbitrary thresholds or benchmarks. Multiple signatures with no overlapping genes for a given phenotype are typical.

Clinical Data Sets

Co-chairs: Wendell Jones (Expression Analysis) and Weida Tong (FDA/NCTR)

Wendell Jones (Expression Analysis) reiterated the criteria for selecting studies and associated data sets for the MAQC-II CWG. Such criteria include the quality and nature of clinical information (e.g., the goal of the study: prognostic, therapeutic, or diagnostic), the quality of the microarray data, the quality of biological samples, and the availability of additional samples for prospective analysis. Wendell emphasized that we may choose (as a group) to consider multiple data sets. Smaller working groups within CWG may be formed to focus on different nominated clinical studies, and MAQC CWG members may participate in any or all. Other data sets related to currently considered ones may appear in the near future, and we may decide (as a group) to include them if they provide additional value in the development and validation of classifiers. Ideally, at least one of the chosen studies will have prospective samples that we can use, especially with multiple independent sites processing the same set of samples.

Fraser Symmans (MD Anderson), John Shaughnessy (UAMS), Shujian Wu (Bristol-Myers Squibb), and André Oberthür (University of Cologne, Germany, via WebEx) presented data sets on four disease categories, breast cancer, multiple myeloma, acute lymphoblastic leukemia (ALL), and neuroblastoma, respectively. The data sets are summarized in Table 1. Clinical information was described for all data sets and quality information was available for some. As we are still in the early phases of data collection and organization, more information will soon be forthcoming. The quality of the microarray data presented to this point was very good. Additional data sets for each disease have been or will be identified for the purpose of cross-study or "prospective" validation.

Toxicogenomics Data Sets

Co-chairs: Federico Goodsaid (FDA/CDER) and David Dix (EPA)

David Dix (EPA) reiterated the goal of the TGx WG: Develop and compare methods for deriving genomic signatures from gene expression data that diagnose or predict toxicity of compounds in animal models. The individual entities that need to be predicted are compounds, not individual animals. David gave a brief summary of the nominated TGx data sets: 9 data sets on mice, 25 data sets on rats, and 2 data sets on human hepatocytes. Except for a few data sets, the nominated data sets were determined to be unsuitable for developing predictive classifiers due to the limited number of compounds involved in each data set. However, some of these small data sets might be useful during the validation process. Donald Halbert (Iconix), Rusty Thomas (CIIT), and David Dix (EPA) introduced their data sets. The data set nominated from NIEHS was described by David. David indicated that new data (and/or samples) from EPA's on-going ToxCast program and CIIT's mouse lung cancer study could serve as "prospective" validation. See Table 2 for a summary of the TGx data sets.

Criteria for Performance Evaluation of Classifiers

Co-chairs: Richard Simon (NIH/NCI) and Gregory Campbell (FDA/CDRH)

Richard Simon (NIH/NCI) highlighted some "guiding principles" on the development and evaluation of predictive classifiers. The validation of predictive classifiers should NOT involve (1) measuring overlap of gene sets used in classifier developed from independent data; (2) statistical significance of individual gene expression levels or summary signatures in multivariate analysis; (3) confirmation of gene expression measurements on other platforms; and (4) demonstrating that the classifier or any of its components are "validated biomarkers of disease status". Valid metrics for the validation of predictive classifiers include (1) predictive accuracy; (2) reproducibility of classification for individual patients; and (3) medical utility.

Gregory Campbell introduced the newly formed Regulatory Biostatistics Working Group (RBWG). The goal for RBWG is "to generate a specific regulatory focus for data set and classifier algorithm selections, data analysis, procedures to validate the classifiers, prospective study design, scientific conclusions, and potential impact in regulatory review of the work within MAQC-II". If you are interested in biostatistics or in the regulatory issues about microarrays, you are welcome to join RBWG (e-mail greg.campbell@fda.hhs.gov and cc leming.shi@fda.hhs.gov).

Gene Pennello (FDA/CDRH) presented the possible statistical goals for MAQC-II from an FDA biostatistics viewpoint. For the FDA, the value of MAQC-II is not in evaluating whether particular prediction rules are better than others, *per se*, but in evaluating if strategies for validating a prediction rule are better than others. Validation strategies that work can be used to support approval of genomic signatures, and validation strategies that are least burdensome can shorten time to market. Strategies for evaluating classifiers should include performance validation, algorithm stability, and reproducibility. The evaluation of strategies for developing classifiers is useful to the FDA because (1) the dissemination of good principles for classifier development can lead to the decreased likelihood of an approvable, but flawed classifier; and (2) the proper assessment of error rates is needed to properly determine the sample size for a Phase III or pivotal trial. Gene pointed out that normalization across multiple arrays (e.g., dCHIP, RMA, or quantile) is problematic because the normalization is used on a specific set of arrays during the training phase and is potentially incongruous with practice when microarrays are used on asynchronous samples, unless one employs static reference distributions.

Timothy Davison (Asuragen) discussed the need for a common glossary of terms for classification and methods for performance evaluation among the different working groups; interactions and coordination among the working groups are essential. Mark Porter (Gene Logic) listed six potential predictive modeling projects that the MAQC-II may consider working on.

Open Discussions and Prospective Studies

Chair: Wendell Jones (Expression Analysis)

There were lengthy discussions on the nominated data sets and the potential for "prospective" studies with additional data sets and/or biological samples. All manufactures present, including Affymetrix, Agilent, Eppendorf, GE Healthcare, PhalanxBiotech, and Telechem agreed in principle to supply substantial numbers of microarrays for the MAQC-II validation and "prospective" efforts. Illumina, while not present, also has agreed to support the "prospective" efforts. Separately, ABI, Gene Express, Panomics, and SuperArray have also pledged support. One concern expressed is the availability of RNA in the test/training sets for multiple platforms. For example, John Shaugnessy's (UAMS) multiple myeloma study requires 10 ug of total RNA per subject, which typically doesn't leave enough for other parallel assays. Some TGx studies have more potential for sharing RNA samples cross multiple platforms. This needs to be investigated. Rich Shippy (GE Healthcare) asked about the availability of cRNA (labeled) from existing studies, regardless of the platform finally employed, and this needs to be investigated on a case-by-case basis.

It was generally supported and understood that we may not be able to run samples and perform testing/validation on multiple platforms for an individual study that was initially trained on one platform, especially for the clinical studies. However, attendees agreed that if there is success of having accurate, reproducible (i.e., more than one lab), prospective or forwardly validated classifiers on one particular microarray platform with one study, and possibly in parallel for a different study with a separate platform, then this was still in harmony with the overall MAQC effort to be multi-platform and would be considered acceptable and novel.

A suggestion was voiced from Fraser Symmans, and initially mentioned in Richard Simon's presentation, that we could examine the reproducibility of microarrays with respect to a particular or even multiple classifiers independent of an actual clinical trial. In fact, this step was seen by many as a necessary step that may be accomplished prior to or in parallel with novel analysis of data from existing or soon-to-be released studies. This could be termed as a MAQC-II Pilot study. The scenario is the following: Take one or more existing classifiers for a particular platform (such as Affymetrix), and also collect anonymous human samples from non-treated subjects related to the condition (such as breast cancer) for which the classifier(s) was(were) originally built. Take the total RNA per subject from the samples and allocate them for two or more independent reference labs to run on a platform/chip identical or similar enough to the original study (e.g., HG-U133Av2 vs HG-U133A). Then, after processing, run the classifier(s) on each chip from each reference lab and quantify the per subject reproducibility of each classifier outcome across labs, whether continuous or discrete (or binary). If the results from the reference labs are found to be reproducible, then we would feel much more confident going forward with a larger, more formal prospective validation. However, if the classifiers are found to be lacking in reproducibility, then we need to retool and investigate why and delay the forward or prospective validation until we have achieved success in this area. In addition, we may be able to do this reproducibility test multiplatform, and thus it would be appealing to continue to do studies that are inherently multiplatform. Fraser felt that a positive result on reproducibility would remove many of the concerns of clinicians and be deemed a success by clinicians and doctors prescribing treatment, and volunteered to write the IRB protocol and to acquire the resources required from pathology to carry out this "Pilot" study for classifier outcome reproducibility. One desirable attribute of this smaller study is the absence of the requirement of knowledge of the eventual clinical outcome of interest: we are only testing whether the classifier would provide a similar predicted outcome across processing labs given the same biological sample.

There were questions related to both this reproducibility study and to the validation study as to whether we would use specialized arrays or whole-genome arrays. In addition, we discussed the impact on the diagnosis using specialized arrays if the diagnosis had been built/trained using whole-genome arrays. For example, how would normalization be performed on specialized arrays? What about handling multiple spots/probes for the same transcript on the array?

Leming Shi (and others) pointed out the difficulties in reproducing absolute intensity values in cross-laboratory and cross-platform comparisons. It may be unrealistic to expect that classifiers built on absolute intensity values from one lab would be consistently predictive of phenotype for samples processed in another lab or platform. What about classifiers that use relative intensity vs. absolute intensity, two-color vs. one-channel arrays? Should we attempt to validate classifiers that use absolute intensity at this point? Perhaps the Pilot study will provide answers. Consensus was reached that for validation or other studies, we will no longer restrict the features that a platform has available as was done in the original MAQC-I study.

It was agreed that we should consider all Pilot and validation studies together and use this to create a solid proposal for the array and alternative platform manufacturers so that they can adequately plan and solicit funds for resources related to the prospective effort.

Action Items

1. Biweekly conference call to be set up for each WG
2. Monthly conference call to be set up for the entire MAQC
3. Raw data to be submitted to Leming Shi at FDA/NCTR by 12/15/06
4. A small team of volunteers to assess the quality of the microarray data (without access to phenotypic information); volunteers are welcome
5. Each participant is expected to sign a Confidential Disclosure Agreement (CDA) before gaining access to the data sets
6. Data sets to be distributed to participants by 1/31/07
7. Initial results/issues/problems to be discussed face-to-face at (around) the FDA Science Forum in Washington, DC, April'07 (tentative).

Remaining Issue: Intellectual Properties

We did not have time to discuss the IP issue during the 6th MAQC meeting. There have been follow-up discussions between potential data (and sample) providers and Leming Shi since the meeting. Questions and concerns were brought up about the handling of the IP issue. Suggestions on the proper handling of potential new intellectual properties out of MAQC-II are welcome. We will discuss this during our forthcoming conference calls. Data sets will not be distributed to participants until the IP issue is fully addressed and mutually agreed upon.

Table 1. Clinical Data Sets Nominated and Discussed for MAQC-II

Disease	Data Source	Clinical Phenotype	Number of Samples	Additional Data
Breast Cancer	UNC;NKI/Rosetta; Brussels JBI; MD Anderson Cancer Center	Subtype classification; Prognosis; Chemotherapy	133; 97; 189; 133	MD Anderson
Multiple Myeloma	University of Arkansas for Medical Sciences (UAMS)	Subtype classification; Prognosis; Chemotherapy outcome	~700	Millennium (250); Univ. Heidelberg (???)
Acute Lymphoblastic Leukemia (ALL)	St. Jude Children's Research Hospital; Erasmus University Medical Center, The Netherlands	Subtype classification; Prognosis; Chemotherapy outcome	98; 173	St. Jude
Neuroblastoma	University of Cologne, Germany	Subtype classification; Prognosis; Chemotherapy outcome	77; 174	Intl. Collaborators (200-250 samples)

Table 2. Toxicogenomics Data Sets Nominated and Discussed for MAQC-II

Data Set	Data Sources	Phenotype Information	Number of Compounds, Doses, Time Points, and Replicates	Platform
Iconix Rat Liver Toxicity	Iconix	Liver toxicity. Clinical chemistry and histopathology data available	22 cmpds; 2-4 does per cmpd; 4-5 time points; 3 rats per group	CodeLink RU1
Iconix Rat Liver Cancer	Iconix	Carcinogenicity. ALT, necrosis, hypertrophy, relative liver weight data available	146 cmpds; Single high dose; 1-3 time points; 3 rats per group	CodeLink RU1
Iconix Rat Kidney Toxicity	Iconix	Kidney toxicity. Albumin, BUN, CRE, and cholesterol data available	75 cmpds; Single dose (MTD); 1-3 time points; 3 rats per group	CodeLink RU1
CIIT Mice Lung Tumor	CIIT	Lung tumor formation in 2 year rodent cancer bioassay. Histopathology, serum NMR (subset of samples), liver gene expression (subset of samples) available	13 (7 carcinogens and 6 non-carcinogens) plus controls. Single dose (MTD); 90 days; 3-4 mice per group (addl. mice available for most groups). Archived tissues available for most treatments and samples; Additional data from on-going studies	Affymetrix 430 2.0
NIEHS Rat Liver Toxicity	NIEHS	Liver toxicity (various phenotypes). Clinical chemistry and liver histopathology data available	8 cmpds (7 acute liver toxicants and 1 non-toxic control). 4 doses; 3 time points; 4 rats per group.	Affymetrix and Agilent for liver, and Agilent for blood
EPA/Iconix Rat Liver Toxicity	EPA	Liver toxicity. Clinical chemistry and histopathology data available	5 cmpds; Single dose (MTD); 3 time points (1, 3, 5-day); 3 rats per group	CodeLink RU1 (3 time points); Affymetrix 230_2.0 and AB (day 3 only)
EPA/Gene Logic Rat Liver Toxicity	EPA	Liver toxicity. Clinical chemistry and histopathology data available	2 cmpds; Single dose (MTD); 3 time points (6 hrs,1 and 14 day); 5 rats per group	Affymetrix 230_2.0
EPA Rat/Human Hepatocytes	EPA	Liver toxicity. Cytotoxicity data available	12 cmpds; 3 doses; 1 time point (24 hrs); 3 rats per group	Affymetrix 230_2.0 or U133Plus2

Table 3. Coordinators of the MAQC-II Working Groups

Working Group	Coordinator	E-mail
Clinical WG	Uwe Scherf Wendell Jones Lajos Pusztai	uwe.scherf@fda.hhs.gov wjones@expressionanalysis.com lpusztai@mdanderson.org
Toxicogenomics WG	Federico Goodsaid David Dix	federico.goodsaid@fda.hhs.gov dix.david@epa.gov
MAQC Titrations WG	Richard Shippy Rick Jensen Russ Wolfinger	richard.shippy@ge.com roderick.jensen@umb.edu russ.wolfinger@sas.com
Regulatory Biostatistics WG	Greg Campbell	greg.campbell@fda.hhs.gov

Everyone is welcome to join the MAQC project. If you are interested in contributing to a particular WG, please contact the coordinators of the corresponding WG (cc leming.shi@fda.hhs.gov to ensure that your e-mail will be listed on the MAQC distribution).

Table 4. Participants of the 6th MAQC Project Meeting, November 28-29, 2006

No.	Name	Organization	No.	Name	Organization
1	Shashi Amur	FDA/CDER	54	Mark Porter	Gene Logic
2	Angela Men	FDA/CDER	55	Vitali Proutski	Almac Diagnostics
3	Emi Arikawa	SuperArray	56	Laura H. Reid	Expression Analysis
4	Anne Bergstrom Lucas	Agilent	57	Robert J Rooney	Genome Explorations Inc.
5	Eric Bremer	Precision Biomarker Resources	58	Raymond R. Samaha	Applied Biosystems
6	Andrej Bugrim	GeneGo Inc.	59	Susanna-Assunta Sansone	EBI
7	Gregory Campbell	FDA/CDRH	60	Uwe Scherf	FDA/CDRH
8	Jennifer G. Catalano	FDA/CBER	61	Joe Shambaugh	Genedata (USA) Inc.
9	Kervin Chen	Phalanx Biotech Group	62	John D. Shaughnessy	Univ. Arkansas for Medical Sci.
10	Luke Chen	Phalanx Biotech Group	63	Leming Shi	FDA/NCTR
11	Tzu-Ming Chu	SAS Institute	64	Song Shi	BD Diagnostics
12	Timothy S. Davison	Asuragen	65	Toshi Shioda	Harvard Medical School
13	Mauro Delorenzi	Swiss Institute of Exp. Cancer Res.	66	Richard Shippy	GE Healthcare
14	David J. Dix	EPA	67	Richard Simon	National Cancer Institute
15	Yvonne P. Dragan	FDA/NCTR	68	Dave D. Smith	Luminex
16	Rosalie Elespuru	FDA/CDRH	69	Mat Soukup	FDA/CDER
17	Felix W. Frueh	FDA/CDER	70	Yongming Sun	Applied Biosystems
18	James C. Fuscoe	FDA/NCTR	71	W. Fraser Symmans	MD Anderson Cancer Center
19	Federico M. Goodsaid	FDA/CDER	72	Zivana Tezak	FDA/CDRH
20	Lei Guo	FDA/NCTR	73	Danielle Thierry-Mieg	NIH/NCBI
21	Paul K. Haje	TeleChem ArrayIt	74	Jean Thierry-Mieg	NIH/NCBI
22	Donald N Halbert	Iconix	75	Russell S. Thomas	CIIT Centers for Health Res.
23	Damir Herman	NIH/NCBI	76	Karol L. Thompson	FDA/CDER
24	Peter Herzer	Eppendorf Biochip Systems	77	Jawahar Tiwari	FDA/CBER
25	Kurt Jarnagin	Iconix	78	Weida Tong	FDA/NCTR
26	Roderick V. Jensen	Univ. of Mass. Boston	79	Jonathan D. Tugwood	AstraZeneca (early access)
27	Adam C Jerauld	Virginia Bioinformatics Inst.	80	Yaron Turpaz	Affymetrix
28	Hanlee Ji	Stanford University	81	Stephen J. Walker	Wake Forest University
29	Donald F. Jin	Gene Logic	82	Eric Wang	Systems Analytics
30	Jason Gang Jin	ShanghaiBio Corporation	83	Sue Jane Wang	FDA/CDER
31	Charles D. Johnson	Asuragen	84	Janet A. Warrington	Affymetrix
32	Wendell D. Jones	Expression Analysis	85	James C. Willey	Ohio Medical University
33	Ernest S. Kawasaki	NIH/NCI	86	Paul K. Wolber	Agilent
34	Samir Lababidi	FDA/CDRH	87	Russ Wolfinger	SAS Institute
35	D.J. Dave Li	FDA/CDRH	88	Bill Worzel	Genetics Squared
36	Wayne Liao	Phalanx Biotech Group	89	Shujian Wu	Bristol-Myers Squibb
37	Simon Lin	Northwestern University	90	Chunlin Xiao	Applied Biosystems
38	Jun Luo	Systems Analytics	91	Jingping Yang	SuperArray
39	Yuling Luo	Panomics	92	Xiao Zeng	SuperArray
40	Charles Ma	Phalanx Biotech Group	93	John Zhang	Systems Analytics
41	Scott R. Magnuson	GenUs BioSystems	**WebEx Participants**		
42	Diana Matkovich	Eppendorf North America	1	Meyling Cheok	St. Jude Children's Research Hospital
43	Donna L. Mendrick	Gene Logic	2	Hong Fang	FDA/NCTR (Z-Tech)
44	George J. Mulligan	Millenium Pharmaceuticals	3	Connie Kohne	Jaden BioScience
45	Padraic Neville	SAS Institute	4	David Kohne	Jaden BioScience
46	Michael S. Orr	FDA/CDER	5	Yong Mao	Zhejiang University
47	Jim Parina	Expression Analysis	6	André Oberthuer	University of Cologne
48	Kyunghee Park	Samsung	7	Roger Perkins	FDA/NCTR (Z-Tech)
49	Xuejun Peng	Takeda	8	Tieliu Shi	Chinese Academy of Sciences
50	Gene A. Pennello	FDA/CDRH	9	Soo Kyung Suh	Korean Food and Drug Adm.
51	Mette A. Peters	Rosetta Biosoftware	10	Charles Wang	UCLA/Cedars-Sinai
52	Ron Peterson	Novartis	11	Wenjian Yang	St. Jude Children's Research Hospital
53	P. Scott Pine	FDA/CDER	12	Liang Zhang	CapitalBio Corporation

Quote of the meeting:

"We should not waste good thoughts on bad data."

Anne Bergstrom Lucas (Agilent)