The MicroArray Quality Control (MAQC) Project:
An FDA-Led Effort Toward Personalized Medicine

# Summary of the 9th MAQC Project Meeting

## Development and Validation of Predictive Models Based on Microarray Data

September 18-19, 2008
US FDA, Silver Spring, Maryland


**Summary Author:** Leming Shi (leming.shi@fda.hhs.gov)
**Summary Date:** October 1, 2008
**Meeting Organizers:** Leming Shi and Federico Goodsaid (FDA)
**MAQC Contact:** Leming.Shi@fda.hhs.gov, Tel: +1-870-543-7387
**MAQC Website:** http://edkb.fda.gov/MAQC/


***MAQC-II Objective:***

***Reaching consensus on the "best practices" (Data Analysis Protocol, DAP) in developing and validating microarray-based predictive models (classifiers) for clinical and preclinical applications.***

The 9th face-to-face MAQC project meeting was held on September 18-19, 2008 at the US Food and Drug Administration's White Oak facilities in Silver Spring, Maryland. To make the discussions more effective, this meeting was by invitation only: one representative was invited from each data analysis team, each manuscript team, each data provider, and each platform provider. A total of 38 on-site participants and six WebEx participants representing 28 organizations attended the meeting. Presentations (PowerPoint or PDF files) are available by contacting Leming.Shi@fda.hhs.gov.

The main objectives of the meeting were: (1) Report on the selection of MAQC-II "candidate" models; (2) Analysis of prediction results on the validation sets; (3) Progress report on the preparation of manuscripts; and (4) Timeline for the project and manuscript preparation.

By September 17, 36 data analysis teams submitted prediction results from 18,202 models to the MAQC-II. The prediction performance (MCC, Accuracy, Sensitivity, Specificity, AUC, and RMSE) for these models were revealed during the meeting and resulted in many interesting discussions and debates. Participants agreed that this is a very productive and exciting meeting for everyone to learn and to share.

## September 18, 2008 (Day One)

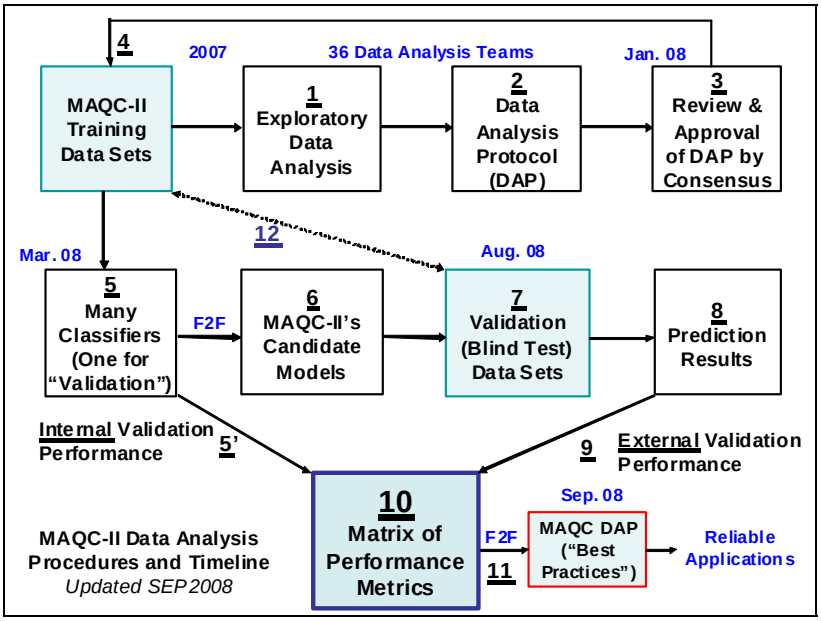
### Session I-A: MAQC-II Overview and Selection of "Candidate" Models

Chair: **Federico Goodsaid** (FDA/CDER)

Session I-A was aimed at updating MAQC-II participants of the progress that the MAQC-II project has made so far.

- **Leming Shi** (FDA/NCTR) provided an overview of the MAQC-II project and outlined the agenda for this 9th face-to-face meeting. Leming emphasized the main objective of the MAQC-II project: reaching consensus on the 'best practices' in developing and validating microarray-based predictive

models (classifiers) for clinical and preclinical applications. Such "best practices" will be incorporated in the MAQC's Data Analysis Protocol (DAP) that is expected to work reasonably well on many data sets beyond the six data sets being analyzed by the MAQC-II. Reliable and robust predictive models are essential to realize the great promises of personalized medicine. To accomplish this task, the MAQC-II has been creating a unique data set of predictive models, each of which is characterized by a set of modeling factors, internal cross-validation performance measures, and external performance measures (see summary of the 8th meeting for more details, http://www.fda.gov/nctr/science/centers/toxicoinformatics/maqc/docs/MAQC8_MeetingSummary_March24-26_2008.pdf). By the time of the 9th face-to-face meeting, prediction results of the validation sets from 18,202 models were received by the MAQC-II from 36 data analysis teams. We appreciate the 36 data analysis teams for their dedication to the MAQC-II project.




- **Greg Campbell** (FDA/CDRH) described the rationale and process for reviewing the data analysis protocols (DAPs) and selecting a candidate model for each of the 13 endpoints. This important exercise was coordinated by Greg Campbell, Lakshmi Vishnuvajjala and Tim Davison. The process of selecting candidate models emulates good classifier development practice and possibly helps reduce the bias and overfitting problems. Using the same data to build the model and then report the performance of the model is well-known to introduce a bias. The performance always deteriorates with independent confirmatory data even if it comes from exactly the same data source. In addition, bias may rise due to multiplicity, i.e. building a number of models and picking the best based on some performance measure. The review and selection process is very democratic: volunteers were solicited from RBWG and from the data analysis teams, and all were welcome and had an equal vote. Each volunteer was asked to rate each DAP from 1 to 10 where 10 is great (and 1 is not) and to indicate a passing level for the scale each used. Everyone who reports a threshold used 7 or more as "passing". Seven volunteers reviewed all or almost all the DAPs (a humongous feat!) and four others a significant proportion. Volunteers who were involved in data analysis teams were asked to not rate their own DAPs and whether they did or not their score was replaced with the median of the values from the reviewers who did all the DAPs. The second step of the process was the selection of a candidate model for each endpoint. Reviewers were informed of the results of the DAP rankings and then asked to select a list of data analysis teams from which to nominate a candidate model for each endpoint. The results were compiled and then discussed via teleconferences. Both the DAP rankings and the short list of models for each endpoint were transmitted to the MAQC-II Steering Committee.

- **Wendell Jones** (Expression Analysis) described the process by which the Steering Committee made its final decision on the selection of one candidate model for each endpoint based on the recommendations of the RBWG. A small group of Steering Committee members met to create a straw-man list of initial recommendations based upon RBWG committee review of individual DAPs (which were ranked overall by consensus). Many factors were considered in making this "straw-man" list, such as evidence of appropriate and unbiased measures of expected performance; normalization and feature selection was embedded within CV; clarity and completeness of performance measures provided; number of endpoints examined in total by the team; diversity of analytic approaches and organizations, and performance estimates should not appear to be outliers. After the initial straw-man selection of candidates and back-ups, further questions were asked to the respective data analysis teams regarding remaining ambiguities in the initial selection. Questions were resolved and candidates were partially reshuffled based on answers. Candidates were reviewed by Greg Campbell and Leming Shi and then by the entire Steering Committee. Final selections for candidates and back-up candidates were completed. Some caveats: Some backup models appear to have better estimated performance based on training than the candidate model (this demonstrates that estimated performance was not the sole selection criteria); some OS-related models appear to have better estimated performance based on training than the EFS-related models (and OS is apparently harder to predict), further demonstrating that estimated performance was not the sole criteria. If your model was not selected as MAQC's "candidate" model, this further demonstrates that estimated performance was not the sole criteria, but it is nothing personal about you (the Steering Committee thinks you all are wonderful!!).

A lot of time was reserved for representatives from the data analysis teams to comment on the process of DAP ranking and candidate model selection. Meeting participants were grateful for the energy that was put into this process by the RBWG and Steering Committee. Leming reiterated that this process is more of an evaluation of the RBWG reviewers and the Steering Committee rather than an evaluation of the data analysis teams since the prediction performance of the candidate models selected by the Steering Committee will be revealed and if the performance of these candidate models turns out to be worse than the average of the 36 data analysis teams, then we need to rethink about the DAP ranking and model selection process, and the judgment of the reviewers and the Steering Committee.

## Session I-B: Prediction Results from the Validation Data Sets

Chair: **Greg Campbell** (FDA/CDRH)

In Session I-B, representatives from data analysis teams were invited to share their experiences and observations in making predictions on the six data sets (13 endpoints) before Wendell Jones described the "rules" for calculating performance metrics followed by the revelation of the prediction performance of the 18,202 models by Leming Shi.

- **Andreas Scherer** (Spheromics, Finland) presented team Spheromics' observations on the five data sets (Spheromics did not analyze the MM data set). For each validation set, a PCA-type analysis was conduced to determine whether the validation set samples are considerably different from the training samples. If so, the validation samples are not predicted. Team Spheromics expected that good prediction performance would be achieved for endpoints C (NIEHS), L (NB: NEP_S), J (NB: OS_MO), and K (NB: EFS_MO), and reasonable prediction performance would be achieved for endpoints B (Iconix), D (BR:pCR), and E (BR:erpos). However, the prediction performance for endpoints A (Hamner) and M (NB:NEP_R) would be poor.
- **Matt McCall** (Johns Hopkins University) briefly described the Barcode approach for training and prediction. Using only "good" (NUSE < 1.02) arrays to create the barcode improved cross-validation prediction in the training data. However, performance in predicting on good and bad arrays were comparable; bad arrays cause a false signal to be picked up by the barcode, but are unlikely to affect

the few features that are used to make a prediction. Matt noted a considerable and persistent date effect in the MDACC training data, and excluded genes that strongly predicted date from the list of possible features. For endpoint I (MM_UAMS PCR1), Matt noted that the Johns Hopkins team could not make reliable predictions of the outcome based on microarray data and wondered whether this endpoint was intentionally designed.

- **Wendell Jones** (Expression Analysis) described the rules for calculating the prediction performance of models. If the predicted outcome value is provided on a continuum, a predicted outcome of less than 0.5 was treated as a negative (N or 0), whereas values greater than or equal to 0.5 were treated as a positive (P or 1). There is one exception: for team Cornell, the cutoff was 0, instead of 0.5. For some endpoints, there are missing outcome values for some subjects due to e.g. insufficient follow-up time; these subjects were omitted from the validation study and performance calculation. For the calculation of MCC, when all samples are predicted to be in the same class, the MCC cannot be calculated (denominator = zero); we assigned an MCC value of missing for this model. For some data analysis teams, there may be missing predicted outcome values (cells without a numerical value). There were a lot of discussions regarding how to treat these missing predictions. Wendell proposed several scenarios of different levels of penalty, but the approach taken by the FDA/NCTR (Dr. Zhenqiang Su) for performance calculation was the most stringent: if a sample was not predicted by a model, then it was considered to have been predicted wrong. This implementation seemed to be consistent with Greg's preference of more conservative estimation of prediction performance and in line with the "intention to diagnose" in clinical trials.
- **Leming Shi** (FDA/NCTR) presented the prediction performance of the 13 MAQC-II candidate models selected by the Steering Committee; this turned out to be one of the most anticipated and exciting moments of the face-to-face meeting. First, Leming showed how the external validation prediction performance of these 13 models compares with the cross-validation performance estimates. Not surprisingly, the cross-validation performance had a tendency of overestimating the models' external prediction performance; however, the extent of overestimation appeared to be acceptable to most meeting participants. Second, Leming asked the question of whether the external prediction performance of the 13 candidate models selected by the Steering Committee would be better than the average performance of the models nominated by the 36 data analysis teams for each endpoint. Greg and other meeting participants were delighted to see that the MAQC-II candidate models indeed performed better, demonstrating that the RBWG and Steering Committee did a reasonably good job in selecting the candidate models. Leming then released the big matrix of performance metrics consisting of 19,696 models, for which prediction performance was available for 18,202 models, to meeting participants and to all data analysis teams via e-mail. Many meeting participants skipped lunch so that they could explore this data set of predictive models! Leming noted that the information distributed should be considered preliminary and was subject to change as data analysis teams make final corrections of clerical mistakes before Thursday, September 25, 2008. In addition, the exact formulae for calculating AUC is still under discussion. The distribution of sample class labels will be depending on the correction of clerical errors and each data analysis team's confirmation that its submitted prediction results are correct and final.

### Sessions I-C: Manuscripts (1) – Modeling Factors and Microarray Reality Check

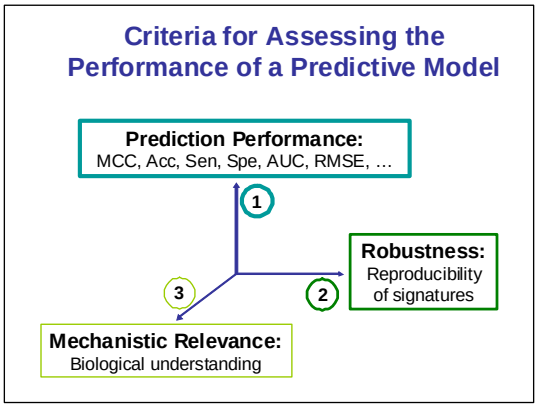
Chair: **Lakshmi Vishnuvajjala** (FDA/CDRH)

Sessions I-C to II-A were organized to allow each manuscript team to present its progress in preparing the proposed manuscript. Before the face-to-face meeting, 22 of the 35 manuscript teams confirmed their commitment to developing their original manuscript proposals into a manuscript, following the MAQC-II timeline. Each team was given an opportunity to present its progress at the meeting. Leming warned each manuscript team that it is a huge task to develop a high-quality manuscript under the tight timeline in

a collaborative environment and each team should be realistic regarding the amount of effort that it can devote to the MAQC-II project. If infeasible to develop a separate manuscript within the timeline, the original proposal team is encouraged to merge its findings to other manuscripts including the main paper.

Manuscripts presented in Session I-C aimed at exploring the impact of different modeling factors on the prediction performance. Cumulatively, this set of manuscripts would provide a reality check regarding the reliability microarray-based predictive models.

1. **Leming Shi** (FDA/NCTR) expected that the MAQC-II "main paper" will incorporate the most important findings of the MAQC-II participants with the main objective of reaching consensus on the "best practices" of developing and validating microarray-based predictive models. The "best practices", to be implemented as MAQC-II's Data Analysis Protocol (DAP), will be applicable to future data sets outside of the MAQC-II project. Realizing the great potentials of personalized medicine requires robust microarray-based predictive models. Leming reiterated the three categories of criteria for assessing the performance of predictive model: prediction performance, robustness of the signature, and mechanistic relevance of the signature, in decreasing order of importance for classification purposes. Each MAQC-II manuscript should use one or more of the tree types of criteria to objectively judge the degree of success of our work.




2. **John Zhang** (Systems Analytics) presented progress with the manuscript on "Minimizing the impact of batch effects in microarray data on the performance of predictive models". It has become more and more obvious from the prediction results of the MAQC-II participants that the appropriate handling of batch effects will hold the key to the successful extrapolation of a predictive model to external data sets where batch effects are almost certain of concern. The hypothesis: Batch-effect removal methods effectively improve the prediction performance of microarray-based predictive models.
3. **Ken Hess** (MD Anderson Cancer Center) was unable to attend the meeting but is still committed to developing a manuscript on the potential impact of normalization methods on prediction performance.
4. **Weida Tong** (FDA/NCTR) presented a lot of results from of the manuscript team that tries to evaluate the cross-platform consistency and transferability of microarray-based molecular signatures. The team's results demonstrate that (1) molecular signature genes identified from one microarray platform can be directly transferred to another platform to develop a predictive model; (2) a predictive model developed from one microarray platform can accurately predict the outcomes of samples profiled by another platform; and (3) signature genes independently identified from different microarray platforms can be highly concordant.
5. **Pierre Bushel** (NIH/NIEHS) presented via WebEx work conducted by the team on cross-tissue predictability of microarray genomic markers. The team demonstrated that (1) genomic classifiers from the blood can accurately predict liver necrosis manifested from exposure to a compendium of hepatotoxicants; (2) pathways and biologic mechanisms related to a severe inflammatory (immune) response, apoptosis, mitochondrial damage and angiogenesis are overrepresented by genes which confer prediction from the blood to the liver; and (3) genes representative of the toll-like receptor (TLR) signaling pathway leading to a cell proinflammatory response may play a key role in the hepatotoxicant-induced liver injury.

6. **Yiming Zhou** (University of Arkansas for Medical Sciences) was unable to attend the meeting but is still committed to developing a manuscript that compares the prediction performance of models developed on the same set of patients but with different generations of Affymetrix microarrays. Classification models developed from an older generation of the widely Affymetrix microarray gene expression platform (U95Av2 or U133A) can be applied to accurately predict clinical outcomes of samples profiled on a newer generation of the same platform (U133Plus2.0). Classification models based on the combined data set from three generations of the Affymetrix platform offers more reliable prediction results on external validation data sets, suggesting the utility of legacy gene expression data for model development. Transforming absolute gene expression intensity values into a relative scale, explicitly or implicitly, is essential to achieve across-generation predictability.
7. **Benedikt Brors** (DKFZ) presented via WebEx a manuscript proposal that compares one-color and two-color microarray platforms for the classification of neuroblastoma based on gene expression profiles. About 400 NB tissue samples are being profiled at the University of Cologne using the Agilent one-color platform, and the resulting data will be compared with the two-color data being analyzed by the MAQC-II. One important hypothesis to be tested is that the use of reference RNA sample helps the microarray system resist systematic bias or batch effect. The influence of dye bias on predictive performance for two-color data will also be investigated.
8. **Wendell Jones** (Expression Analysis) described results from the "QC team" that investigated the microarray data quality and its impact on classifier performance. It included a simulation of the impact of common technical defects in microarray data on classification and prediction results.
9. **Huixiao Hong** (FDA/NCTR) evaluated the robustness of genotyping technologies and genotype calling algorithms. It was found that genotypes from different SNP arrays called by the same calling algorithm for the same sample and the same SNPs were found to be variable. Variations in genotypes called by different calling algorithms on the same raw data set were also observed. Variations from different SNP arrays and calling algorithms were observed to propagate to the genetic markers identified in the downstream association analysis. The results demonstrate that false positive associations do exist and true positive associations might be missed in current genome-wide association studies. An ongoing experiment that genotypes three HapMap samples (one trio) and three additional DNA samples with known copy number variation is in progress. Each of the six samples is being processed in quadruplicates in each laboratory. Each platform is being tested at multiple sites.

**Session I-D: Manuscripts (2) – Functional Analysis, SOP, and Multiplicity**
Chair: **Jim Fuscoe** (FDA/NCTR)

Manuscripts presented in Session I-D addressed issues on functional analysis, standard operating procedures (SOP), and multiplicity of prediction.

10. **Tieliu Shi** (Chinese Academy of Sciences) presented work for two manuscripts: (1) "The tumor gene signature selection using dynamic biological networks at pathway level" and (2) "Breast cancer gene signature selection based on multi-level similarity analysis". A subset of genes from the gene signatures developed by different data analysis teams provides better prediction performance. Gene signatures of different sizes can perform equally well in prediction; minimized feature size is preferred. Gene signatures can be organized into a simple network related to breast cancer.
12. **Youping Deng** (University of Southern Mississippi) proposed a manuscript on "Meta-analysis of gene features to compare predictive models" but did not report progress.
13. **Yuri Nikolsky** (GeneGo) presented results on the comprehensive functional analysis of data sets and gene signatures used in the MACQ-II project. It was demonstrated that gene signatures from

different models for the same endpoint significantly correlate with biological functions at the levels of gene feature intersections, protein function composition, enrichment in ontologies and network topology. Similarity in gene signatures correlates with the models' prediction performance, in particular when similarity is measured at the levels of pathways, disease biomarkers, cellular processes, rather than at the level of gene feature intersections. Gene signatures are highly interconnected between each other and with the rest of human proteins (whole proteome). For each endpoint, genes in multiple signatures are co-regulated by strikingly few transcription factors. Likewise, genes in multiple signatures co-regulate very few downstream targets.

14. **Greg Campbell** (FDA/CDRH), representing RBWG, briefly presented two manuscript proposals, (1) Principles of classifier development (based on the RBWG SOP) and (2) Multiplicity and selection of candidate models. The problem of multiplicity is of great concern: the performance is overestimated for a reason that is related to the Regression-to-the-Mean effect or the Rookie effect in baseball and the difficulty is that it is unclear how to adjust for this bias. The variance is underestimated but there are multiplicity methods to adjust for this bias. Greg indicated that some of the statistics-oriented manuscript topics might be merged.

16. **Gene Pennello** (FDA/CDRH) presented a manuscript proposal on "Analysis of external validation results with adjustment for multiplicity in the MAQC-II". For each of the 13 endpoints on six data sets, the candidate model was compared with 35 other nominated models for classification performance on the validation data set. The comparisons were adjusted for multiplicity and for correlation in model results on the same data set. For each endpoint, each of the 36 nominated models was given the Bayesian posterior probabilities that the model is the best performer while adjusting for multiplicity and correlation of model results. Multiplicity adjusted p values were also provided. In general, the candidate model performed favorably relative to the nominated models, as it could not be rejected as the model with the best performance. Gene pointed out potential overlap with two other manuscripts: (1) Multiplicity and selection of candidate models (Greg Campbell) and (2) Significance tests for comparing multiple results in the MAQC-II (Xuegong Zhang).

Before the close of day one, **Russ Wolfinger** (SAS Institute) gave an interactive presentation on the variance components analysis of the difference between external prediction performance and cross-validation performance (Diff_MCC or Diff_AUC). Russ' initial analysis used all 18K models and revealed some modeling factors that explained some of the variance in Diff_MCC or Diff_AUC. On the next morning, Russ re-ran the analysis by focusing on the 320 models nominated by the 36 data analysis teams. EndpointCode, InternalValidation, and OrganizationCode were the only nonzero components. Upon request of several meeting participants, Russ made his JMP Genomics script available to MAQC-II participants (see instructions in "MAQC9_ID_Wolfinger_SAS_VarianceComponentsResults.ppt").

## September 19, 2008 (Day Two)

With deep sadness, Leming Shi informed the MAQC-II participants of the tragic loss of a visionary and inspiring colleague, Dr. Robert F. Wagner of FDA/CDRH. Bob passed away from prion disease on June 30, 2008 (http://sites.google.com/site/robertfwagnermemorial/). Bob enthusiastically participated in and contributed to the MAQC-II project and attended the 6th, 7th, and 8th face-to-face meetings. Many of us still vividly remember the thought-provoking keynote speech that Bob gave during the 7th MAQC face-to-face meeting at SAS Institute, May 25, 2007 regarding uncertainties of classifier performance from finite training and finite testing. We miss Bob greatly! It is gratifying to know that Weijie Chen, Brandon Gallas and colleagues are continuing the important work that Bob initiated on the estimation of uncertainties of predictive models.

**Session II-A: Manuscripts (3) – Uncertainty, Clinical Utility, and Consensus**
Chair: **Weida Tong** (FDA/NCTR)

17. **Weijie Chen** (FDA/CDRH) presented progress of the manuscript on "Uncertainty estimation in prediction models". Weijie described the concept of performance uncertainty and classifier stability in the high dimension, low sample size setting. He then proposed a methodology for uncertainty estimation with a finite data set and validation study designs for the assessment of predictive classification models. Finally, he demonstrated the utility of the proposed approach with MAQC-II clinical data sets and models.
18. **Xuegong Zhang** (Tinghua University) was unable to attend the meeting but is still committed to developing a manuscript on "Variability of the estimated classification errors due to sampling". A given data set is only a sample drawn from the population of the studied disease, and it is therefore necessary to investigate the confidence interval of estimated classification errors obtained with a given data set. Theoretical study shows that error rate confidence interval could be estimated unbiasedly using iterated cross validation under mild conditions. A strategy to estimate error rate confidence interval is proposed and experimented on both simulated data and the MAQC-II results.
19. **Samir Lababidi** (FDA/CDRH) presented progress of the manuscript on "The clinical benefit of a microarray-based classifier". Given the classification prediction performance of a genomic-based classifier, what is the chance that this classifier would have an added clinical benefit over and beyond the clinical covariates alone? Could a genomic-based classifier provide an acceptable clinical benefit even if it is not picked by MAQC-II as the "candidate/best" model? Samir plans to explore the relationship between the clinical benefit of a classifier and the prediction performance for all the candidate models by all data analysis teams.
20. **Lajos Pusztai** (MD Anderson Cancer Center) was unable to attend the meeting but is still committed to developing a manuscript on "Predicting treatment outcomes of breast cancer patients with microarray gene expression profiles". This manuscript team has already held several teleconferences.
21. **Guy Tillinghast** (Riverside Cancer Care Center) presented on "Guideline for good clinical practice in conducting microarray studies". Guy listed some barriers to the clinical use of microarrays: Quality concerns (cross-platform reproducibility, continual updating of microarray platforms, and lack of uniform quality control standards), changing practices (requirement for freezers and educating pathology technicians to handle frozen samples), funding (pharmaceutical unwillingness to limit product label and insurance unwillingness to fund research), and knowledge gap (readable guidelines for conducting microarray clinical trials). The proposed guideline for good clinical practice (GCP) should help move the microarray technology to the clinic.
22. **Federico Goodsaid** (FDA/CDER) pointed out that the MAQC-II effort completes the work initiated with MAQC-I together with the experience at the FDA with Voluntary eXploratory Data Submissions (VXDS) to develop recommendations for the generation, analysis, interpretation and submission of gene expression data from microarrays. These recommendations, which capture a consensus in this area, serve as major references for the Companion Guidance for the Pharmacogenomic Guidance of the FDA. A coherent consensus on how to develop classifiers on the basis of microarray data will be of great value for an expanded use of microarray data.
23. At the end of the two-day meeting, several participants suggested that a more thorough investigation about the impact of different feature selection methods on prediction performance. This manuscript topic was decided at the March 2008 face-to-face meeting, but unfortunately Simon Lin who originally agreed to work on it could not commit the time to develop the manuscript due to other commitments. **Jie Cheng** of GSK (jie.j.cheng@gsk.com) has agreed to take a leading role on this manuscript "Feature selection methods play a critical role in determining the performance of

microarray-based predictive models". A small team has been formed and a solid data analysis plan has been agreed upon with encouraging preliminary results.

### Session II-B: Manuscripts (4) – Timeline, Target Journal, and Discussion

Chair: **Leming Shi** (FDA/NCTR)

Leming Shi proposed an updated timeline for manuscript preparation and noted that it reflected a three months delay from what was decided at the March 2008 meeting, mainly due to the much longer than expected time required for reviewing/ranking DAPs and selecting candidate models. While the new timeline is aggressive, most participants who were intimately involved in data analysis and manuscript preparation agreed that it is doable. Working with this timeline was considered essential to keep the momentum of the MAQC-II project. Manuscripts that cannot keep up with the timeline will be automatically dropped out and will not be considered as one of the MAQC-II manuscripts to be proposed for publication as a group.

**New timeline for MAQC-II manuscript preparation:**

VO: April 28 (Detailed manuscript outlines, distributed)
V1: October 6 (Full manuscript draft)
V2: Nov. 3 (Revised)
V3: Nov. 17 (Revised, ready for institutional clearance)
V4: Dec. 1 (Revised, almost ready for peer review)
***VS: Dec. 8, 2008 (Submission for peer review)***
Publication Date: April-June, 2009

A lot of discussions were about where to publish the MAQC-II results. Many participants strongly felt that the impact of the MAQC-II project will be more significant than the MAQC-I since the new outcomes will be more directly used to affect and improve clinical practices. It was also pointed out that publishing the important results in a prestigious journal is critical for the huge MAQC-II effort to generate its intended impact to the community. Several options were discussed and the participants agreed that *Nature Biotechnology*, which published the MAQC-I results, should be the first journal to be considered. It was agreed that *Nature Biotechnology* is a good place to disseminate the MAQC-II findings. Leming Shi has been discussing with *Nature Biotechnology* about the MAQC-II publication proposal and will keep MAQC-II participants informed as more information becomes available.

### Session II-C: Open Discussions and Presentations

Session II-C provided an opportunity to meeting participants to freely present their ideas and comments on the MAQC project and related topics. Each presentation was followed with lively discussions.

- **Christophe Lambert** (Golden Helix) presented on "Methods and discoveries drawn from twenty whole genome copy number variation studies". Christophe's analysis was done with Golden Helix SNP & Variation Suite.
- **Federico Goodsaid** (FDA/CDER) briefly described the efforts of MAQC-II GWAWG (Genome-Wide Association Working Group) and in particular, the analysis of the Wellcome Trust Case Control Consortium (WTCCC) data sets.
- **Leming Shi** (FDA/NCTR) presented a proposal that he made in August 2008 to the RBWG review group and the MAQC-II Steering Committee on the selection of candidate models. The idea was to select a good DAP from a data analysis team and the 13 candidate models nominated by that team

would automatically become MAQC-II's candidate models. The proposal was not adopted by the Steering Committee. However, prediction performance of the models from the candidate DAPs appeared to be very good, indicating that it is possible to develop the "MAQC DAP" that will be applicable to future data sets beyond the MAQC project.

- **Jie Cheng** (GSK) presented his ranking of the data analysis teams based on the prediction performance (AUC and MCC) and the consistency between external prediction performance and the cross-validation estimated performance. While the actual ranking changes as different criteria were used, a group of data analysis teams surfaced on the top. Interestingly, this group of "good performers" largely overlapped with the candidate DAPs suggested in Leming Shi's August 2008 proposal. Leming emphasized that the ultimate goal of the MAQC-II project is not to pick a "winning team" or "winning model"; instead, by examining the practices of the teams whose models performed well in validation, we will be able to develop "best practices" for future data sets.
- **Simon Lin** (Northwestern University) presented his findings that, not surprisingly, the overlapping between features from different models with similar performance is limited. Features for classification purposes are redundant. Thus, better biological interpretation becomes very important besides classification performance; well characterized genes are preferred than unknown transcripts as selected features for classification purposes.
- **Yuri Nikolsky** (GeneGo) presented additional results on the comprehensive functional analysis of the MAQC-II results.
- **Fabien Campagne** (Cornell University) presented an approach for characterizing the average bias of biomarker development protocols and implementations. Fabien illustrated his approach using a large data set with over 500 patients (Setlur SR *et al*., JNCI 2008). The distribution plots can be used as an objective sanity/diagnostic test: (1) Run a DAP on a reference data set; (2) Compare the distributions of deltaMCC/deltaAUC with a reference DAP; (3) Significant differences should indicate problems with modeling protocol or implementation tested.
- **May Wang** (GeorgiaTech and Emory University) presented a case study on how microarray data quality control was successfully implemented in a translational research setting.
- **Dalila Megherbi** (University of Mass. Lowell) presented a "proprietary" metric that was said to be better than t-test, fold-change, or SAM for microarray data analysis. Details about this method were not revealed.
- The following is the straw-man of the simple MAQC DAP for developing and validating predictive models (yes, we agreed that microarray data analysis can be as simple as this!). Several data analysis teams have agreed to implement it and apply it to the 13 endpoints.

**The MAQC DAP (Tentative)**

| | |
|---|---|
| ▪ SummaryNormalization: | MAS5-like |
| ▪ FeatureSelectionMethod: | FC ($P$<0.05) |
| ▪ NumberOfFeatureUsed: | <36 (< one feature per team ☺) |
| ▪ ClassificationAlgorithm: | DLDA |
| ▪ BatchEffectRemovalMethod: | TBA (YES) |
| ▪ PerformanceMetrics: | MCC, AUC |
| ▪ InternalValidation: | 5-CV |
| ▪ ValidationIterations: | 10 |

- The meeting was adjourned at 5 pm CDT, Friday, September 19, 2008.

Table 1. Participants of the 9th MAQC Project Meeting, September 18-19, 2008, Silver Spring, Maryland, USA

| No. | Name | Organization | No. | Name | Organization |
|---|---|---|---|---|---|
| 1 | Fabien Campagne | Cornell University | 26 | Andreas Scherer | Spheromics |
| 2 | Gregory Campbell | FDA/CDRH | 27 | Uwe Scherf | FDA/CDRH |
| 3 | Weijie Chen | FDA/CDRH | 28 | Leming Shi | FDA/NCTR |
| 4 | Jie Cheng | GlaxoSmithKline | 29 | Tieliu Shi | Chinese Academy of Sciences |
| 5 | Viswanath Devanarayan | Abbott | 30 | Zivana Tezak | FDA/CDRH |
| 6 | Joaquin Dopazo | CIPF | 31 | Russell Thomas | The Hamner Institutes |
| 7 | James Fuscoe | FDA/NCTR | 32 | Guy Tillinghast | Riverside Cancer Care Center |
| 8 | Brandon D Gallas | FDA/CDRH | 33 | Weida Tong | FDA/NCTR |
| 9 | Federico Goodsaid | FDA/CDER | 34 | Lakshmi Vishnuvajjala | FDA/CDRH |
| 10 | Huixiao Hong | FDA/NCTR | 35 | May Wang | GeorgiaTech and Emory Univ. |
| 11 | Roderick Jensen | VirginiaTech | 36 | Russell Wolfinger | SAS Institute |
| 12 | Wendell Jones | Expression Analysis | 37 | John Zhang | Systems Analytics |
| 13 | Giuseppe Jurman | Fondazione Bruno Kessler | 38 | Ying Zhang | SABiosciences Corp. |
| 14 | Samir Lababidi | FDA/CDRH | | | |
| 15 | Christophe Lambert | Golden Helix | | | |
| 16 | Lili Li | SAS Institute | | | |
| 17 | Simon Lin | Northwestern University | | | |
| 18 | Guozhen Liu | SABiosciences Corp. | | | |
| 19 | Francisco Martinez-Murillo | FDA/CDRH | | WebEx Participants: | |
| 20 | Matt McCall | Johns Hopkins University | 1 | Benedikt Brors | German Cancer Research Center |
| 21 | Dalila Megherbi | University of Mass. Lowell | 2 | Pierre Bushel | NIH/NIEHS |
| 22 | Yuri Nikolsky | GeneGo Inc. | 3 | Timothy Davison | Asuragen |
| 23 | Richard S. Paules | NIH/NIEHS | 4 | Youping Deng | Univ. of Southern Mississippi |
| 24 | Gene Pennello | FDA/CDRH | 5 | Juraeva Dilafruz | German Cancer Research Center |
| 25 | John Phan | GeorgiaTech | 6 | Venkata Thodima | Univ. of Southern Mississippi |

**AGENDA**

The MicroArray Quality Control (MAQC) Project:

An FDA-Led Effort Toward Predictive and Personalized Medicine

# The 9th MAQC Project Meeting

## Best Practices for Developing and Validating Microarray-based Predictive Models

Thursday-Friday
September 18–19, 2008
9:00 am – 6:00 pm Eastern Daylight Time

at
US Food and Drug Administration
Building 51, Room 6200
10903 New Hampshire Avenue
Silver Spring, MD 20903, USA
http://www.fda.gov/oc/whiteoak/

**Meeting Objectives:**

1. Report on the selection of MAQC-II "candidate" models
2. Analysis of prediction results on the validation sets
3. Progress report on the preparation of manuscripts
4. Timeline

Leming.Shi@fda.hhs.gov

Tel: +1-870-543-7387

http://edkb.fda.gov/MAQC/

http://www.nature.com/nbt/focus/maqc/

*Participants should consider information exchanged during the MAQC meeting as confidential.*

## Thursday, September 18, 2008 (Day One)

| | | |
|---|---|---|
| 8:00 am | Registration (participants should arrive at the FDA campus no later than 8:30 am in order to be cleared at the security checkpoint in time) | |
| **Session I-A**: MAQC-II Overview and Selection of "Candidate" Models<br>Chair: **Federico Goodsaid** (CDER/FDA) | | |
| 9:00 am | Welcoming remarks | Federico Goodsaid |
| 9:10 am | Overview of submissions of models and prediction results; Review of meeting agenda | Leming Shi (NCTR) |
| 9:30 am | Review of Data Analysis Protocols (DAPs) and ranking of candidate models by the RBWG | Greg Campbell (CDRH) |
| 9:50 am | Selection of MAQC-II candidate models by the Steering Committee | Wendell Jones (EA)<br>Russ Wolfinger (SAS) |
| 10:10 am | Comments from Data Analysis Teams (DATs) | DAT Leaders |
| 10:40 am | Coffee Break | |
| **Session I-B**: Prediction Results from the Validation Data Sets<br>Chair: **Greg Campbell** (CDRH/FDA) | | |
| 11:00 am | Important observations from Data Analysis Teams regarding the prediction of validation sets | Volunteering DATs are welcome. |
| 12:00 pm | Rules on the calculation of prediction performance metrics for all models | Wendell Jones (Expression Analysis) |
| 12:15 pm | Prediction performance of MAQC-II candidate models; Release of prediction performance metrics and individual sample prediction results for all models | Leming Shi (NCTR) |
| 12:30 pm | Lunch (on your own) | |
| **Session I-C**: Manuscripts (1) – Modeling Factors and Microarray Reality Check<br>Chair: **Lakshmi Vishnuvajjala** (CDRH/FDA) | | |
| 2:00 pm | 1. MAQC-II "main paper": Reaching consensus on the "best practices" in developing and validating microarray-based predictive models for personalized medicine | Leming Shi (NCTR) |
| 2:15 pm | 2. Minimizing the impact of batch effects in microarray data on the performance of predictive models | John Zhang (SAI) |

| | | |
|---|---|---|
| | *3. Microarray normalization methods and prediction performance* | *(Ken Hess, MDACC, absent)* |
| 2:30 pm | 4. Evaluation of cross-platform consistency and transferability of microarray-based molecular signatures | Weida Tong (NCTR) |
| 2:45 pm | 5. Cross-tissue predictability of microarray genomic markers | Pierre Bushel (NIEHS, via WebEx) |
| | *6. Cross-generation consistency in the prediction of treatment outcomes of multiple myeloma patients using different generations of the Affymetrix GeneChip microarrays (U95Av2, U133A, and U133Plus2.0)* | *Yiming Zhou (UAMS, absent)* |
| 3:00 pm | 7. Comparison of one-color and two-color microarray platforms for the classification of neuroblastoma based on gene expression profiles | Russ Wolfinger (SAS)<br>Benedikt Brors (DKFZ) |
| 3:15 pm | 8. Microarray data quality and its impact on classifier performance: a simulation of the impact of common technical defects in microarray data on classification and prediction results | Wendell Jones (Expression Analysis) |
| 3:30 pm | 9. Evaluation of technical robustness of genotyping in genome-wide association studies | Huixiao Hong (NCTR) |
| 3:45 pm | Coffee Break | |
| **Session I-D**: Manuscripts (2) – Functional Analysis, SOP, and Multiplicity<br>Chair: **Jim Fuscoe** (NCTR/FDA) | | |
| 4:10 pm | 10. Biomarker discovery from dynamic biological networks<br>11. Biomarker discovery using meta-analysis | Tieliu Shi (CAS) |
| 4:30 pm | 12. Meta-analysis of gene features to compare predictive models | Youping Deng (USM, WebEx) |
| 4:45 pm | 13. Comprehensive functional analysis of data sets and gene signatures used in the MACQ-II project | Yuri Nikolsky (GeneGo) |
| 5:00 pm | 14. Principles of classifier development: SOP<br>15. Multiplicity and selection of candidate models | Greg Campbell (CDRH) |
| 5:20 pm | 16. Analysis of external validation results with adjustment for multiplicity in the MAQC-II | Gene Pennello (CDRH) |
| 5:35 pm | Discussion | All |
| 5:55 pm | Summary of Day One | Leming Shi (NCTR) |
| 6:00 pm | Adjourn Day One | |

## Friday, September 19, 2008 (Day Two)

| | | |
|---|---|---|
| **Session II-A**: Manuscripts (3) – Uncertainty, Clinical Utility, and Consensus<br>Chair: **Weida Tong** (NCTR/FDA) | | |
| 9:00 am | 17. Uncertainty estimation in the assessment of classification models with a finite data set | Weijie Chen (CDRH) |
| | *18. Significance tests for comparing multiple results in the MAQC-II* | *Xuegong Zhang (Tsinghua, absent)* |
| 9:15 am | 19. The clinical benefit of a microarray-based classifier | Samir Lababidi (CDRH) |
| | *20. Predicting treatment outcomes of breast cancer patients with microarray gene expression profiles* | *Lajos Pusztai (MDACC, absent)* |
| 9:30 am | 21. Good Clinical Practices (GCP) in using microarray gene expression data | Guy Tillinghast (Riverside) |
| 9:45 am | 22. MAQC, VXDS, and FDA guidance | Federico Goodsaid (CDER) |
| 10:15 am | Discussion | |
| 10:40 am | Coffee Break | |
| **Session II-B**: Manuscripts (4) – Timeline, Target Journal, and Discussion<br>Chair: **Leming Shi** (NCTR/FDA) | | |
| 11:00 am | Timeline<br>V1: October 6 (Full manuscript draft)<br>V2: Nov. 3 (Revised)<br>V3: Nov. 17 (Revised, ready for institutional clearance)<br>V4: Dec. 1 (Revised, almost ready for peer review)<br>VS: Dec. 8, 2008 (Submission for peer review)<br>Target journal | Leming Shi |
| 11:30 am | Discussion and action items | All |
| 12:30 pm | Lunch (on your own) | |
| **Session II-C**: Manuscripts (5) – Parallel Discussions<br>Co-Chairs: **Manuscript Team Leaders** | | |
| 1:30 pm | Analysis of prediction results | Russ Wolfinger (SAS) and more volunteers |
| 1:30 pm | Parallel discussions by individual manuscript teams | |
| 5:00 pm | Adjourn the Meeting | |

## Registration

This meeting is by invitation only. The following individuals are invited:

- Leaders of Data Analysis Teams (1 representative per team)
- Leaders of Manuscript Proposals (1 representative per proposal)
- Data Providers (1 representative per provider)
- Microarray manufacturers (1 representative per manufacturer)
- MAQC-II Steering Committee Members

If you plan to attend the meeting, please contact Leming Shi (Leming.Shi@fda.hhs.gov, +1-870-543-7387) as soon as possible so that a seat will be reserved for you.

## Meeting Venue

US Food and Drug Administration
Building 51, Room 6200
10903 New Hampshire Avenue
Silver Spring, MD 20903, USA
http://www.fda.gov/oc/whiteoak/

Contact: Federico Goodsaid
federico.goodsaid@fda.hhs.gov
301-796-1535 (O), 301-520-4063 (C)

## Airports

Ronald Reagan Washington National Airport (DCA)
Washington Dulles International (IAD)
Baltimore/Washington International Thurgood Marshall (BWI)

## Suggested Hotels

Crowne Plaza Hotel Washington DC-Silver Spring
8777 Georgia Avenue
Silver Spring, MD 20910
301-589-0800
http://www.cpdcsilverspring.com/?src=ppc_google_brand

Marriott Courtyard Silver Spring
8506 Fenton St.
Silver Spring, MD 20910
Telephone: 301-589-4899
http://www.silverspringdowntown.com/go/marriott-courtyard-silver-spring