Summary of the MAQC Data Sets

The MicroArray Quality Control (MAQC) Consortium

Contact:
Leming Shi, National Center for Toxicological Research (NCTR), U.S. Food and Drug Administration (FDA)
3900 NCTR Road, Jefferson, Arkansas 72079, U.S.A. Tel: +1-870-543-7387, Fax: +1-870-543-7854
Leming.Shi@fda.hhs.gov; http://edkb.fda.gov/MAQC/; http://www.fda.gov/nctr/science/centers/toxicoinformatics/maqc/

October 2, 2006, 10:18 AM CDT

1. Results of the MAQC project has been described in a series of manuscripts published in *Nature Biotechnology*, September 8, 2006. PDF files of the manuscripts are freely available at http://www.nature.com/nbt/focus/maqc/.
2. The MAQC data set are publicly available starting through the following four mechanisms:
    - GEO (series accession number: **GSE5350**); All original files (e.g. CEL files) can be downloaded as a single tar file at ftp://ftp.ncbi.nih.gov/pub/geo/DATA/supplementary/series/GSE5350/ (3.99 GB).
    - ArrayExpress (accession number: **E-TABM-132**, pending);
    - ArrayTrack (http://www.fda.gov/nctr/science/centers/toxicoinformatics/ArrayTrack/); and
    - The MAQC website (http://edkb.fda.gov/MAQC/MainStudy/upload/) – a good place to download normalized data (filenames staring with "norm_") in a multi-column table format per platform (see Table 8 of this document). The filenames should be self-explanatory.
3. Image files (e.g., .DAT files from Affymetrix platform) are not publicly deposited, but will be made available upon request.
4. For updated information (e.g., corrections, detailed data set annotation, and feedback from users) about the MAQC data sets, please visit the MAQC website (http://edkb.fda.gov/MAQC/).
5. Corrections, questions, and comments should be addressed to Leming.Shi@fda.hhs.gov.

The total number of microarrays used in the MAQC project: **1,329**

Table 1. *"Official"* Platforms and Data Included in the MAQC Main Study: 573 microarrays*

(Eight microarray platforms and three alternative platforms)

Manufacturer	Code	Protocol	Platform	GEO GPL#	Array Express#	ArrayTrack Experiments	Number of Probes	Number of Test Sites	Number of Samples	Number of Replicates	Total Number of Microarrays
Applied Biosystems	ABI	One-Color Microarray	Human Genome Survey Microarray v2.0	GPL2986	**A-MEXP-503**	ABI_1 ABI_2 ABI_3	32,878	3	4	5	60
Affymetrix	AFX	One-Color Microarray	HG-U133 Plus 2.0 GeneChip®	GPL570	**A-AFFY-44**	AFX_1 AFX_2 AFX_3 AFX_4 AFX_5 AFX_6	54,675	6	4	5	120
Agilent	AGL	Two-Color Microarray	Whole Human Genome Oligo Microarray, G4112A	GPL1708	**A-AGIL-11**	AGL_1 AGL_2 AGL_3	43,931 (same printing; 41058 – GeneSpring[d])	3	2	10	60
	AG1	One-Color Microarray				AG1_1 AG1_2 AG1_3		3	4	5	60
Eppendorf	EPP	One-Color Microarray	DualChip® Microarray	GPL4096		EPP_1 EPP_2 EPP_3	294	3	4	5	60
GE Healthcare	GEH	One-Color Microarray	CodeLink™ Human Whole Genome	GPL2895	**A-GEHB-1**	GEH_1 GEH_2 GEH_3 GEH_2Fail	54,359	3[a]	4	5	80
Illumina	ILM	One-Color Microarray	Human-6 BeadChip, 48K v1.0	GPL2507	**A-MEXP-524**	ILM_1 ILM_2 ILM_3	47,293	3[b]	4	5	59
NCI_Operon	NCI	Two-Color Microarray	Operon Human Oligo Set v3	GPL4108	**A-MEXP-573**	NCI_1 NCI_2 NCI_2Fail NCI_3Fail	37,632	3[c]	4	5	74
Applied Biosystems	*TAQ*	*TaqMan® Assays*	*>200,000 assays available*	*GPL4097*		*TAQ_1*	*1,004*	*1*	*4*	*4*	*N/A (16)*
Panomics	*QGN*	*QuantiGene® Assays*	*~2,600 assays available*	*GPL4098*		*QGN_1*	*245*	*1*	*4*	*3*	*N/A (12)*
Gene Express	*GEX*	*StaRT-PCR™ Assays*	*~1,000 assays available*	*GPL4198*		*GEX_1*	*207*	*1*	*4*	*3*	*N/A (12)*
											573

**There are 40 "virtual" microarrays corresponding to the three alternative platforms (TAQ, QGN, and GEX).*

Test sites and sample types are referenced using the following nomenclature: "platform code_test site_ sample ID". Sample A = 100% UHRR; Sample B = 100% HBRR; Sample C = 75% UHRR:25% HBRR; and Sample D = 25% UHRR:75% HBRR.
[a]Test site GEH_2 repeated an initial, failed experiment (GEH_2Fail, due to protocol issues).
[b]Test site ILM_1 only had 19 microarrays.
[c]Test site NCI_2 partially repeated an initial, failed experiment (NCI_2Fail, due to protocol issues) with 14 microarrays. Test site NCI_3 failed an initial experiment (NCI_3Fail, due to protocol issues) and partially repeated the study with another batch of printed slides (site NCI_3r in Table 2).
[d]Data from replicating spots were averaged within GeneSpring software to generate a single value for each unique probe.

Table 2. "*Additional*" Platforms and Data Included in the MAQC Main Study: 130 microarrays

(Seven microarray platforms)

Platform	Code	Protocol	GEO GPL#	Array Express#	ArrayTrack Experiments	Number of Probes	Test Sites	Number of Samples	Number of Replicates	Number of Microarrays Per Test Site
NCI_Operon (2nd printing)	NCI	Two-Color Microarray	GPL4185		NCI_4 NCI_3r	36,288	NCI_4: Harvard University	2	5	10
							NCI_3r: FDA/CBER (repeat)	2	5	10
CapitalBio_ Operon	BIO	Two-Color Microarray	GPL4187	A-MEXP-576	BIO_1	23,232 (same printing)	BIO_1: CapitalBio	2	10	20
	BIO1	One-Color Microarray			BIO1_1		BIO1_1: CapitalBio	2	5	10
Operon_ Operon	OPN	Two-Color Microarray	GPL4188	A-MEXP-574	OPN_1	37,584	OPN_1: Operon	2	5	10
NMC_ Operon	NMC	Two-Color Microarray	GPL4186	A-MEXP-552	NMC_1	36,288	NMC_1: Norwegian Microarray Consortium	2	5	10
TeleChem	H25K	Two-Color Microarray	GPL4219	A-MEXP-575	H25K_2	27,648 (same printing)	H25K_2: Yale University	2	15	30
	H25K1	One-Color Microarray			H25K1_1 H25K1_2 H25K1_3		H25K1_1: TeleChem	2	5	10
							H25K1_2: Yale University	2	5	10
							H25K1_3: Wake Forest University	2	5	10
										130

Table 3. "*Tumor*" Data Included in the MAQC Study: 20 microarrays[a]

(One microarray platform at two laboratories in Stanford University)

Manufacturer	Code	Protocol	Platform	GEO GPL#	Array Express#	ArrayTrack Experiments*	Number of Probes	Number of Test Sites	Number of Samples	Number of Replicates	Total Number of Microarrays
Affymetrix	AFX	One-Color Microarray	HG-U133 Plus 2.0 GeneChip®	GPL570	A-AFFY-44	AFX_1 AFX_2	54,675	2	2	5	**20**

[a]Tumor_Stanford_Lab1 and Tumor_Stanford_Lab2. T = Tumor (colon adenocarcinoma), N = Normal (normal colon tissue, patient matched). The tumor data set was analyzed in Lin, G., He, X., Ji H., Shi, L., Davis, R.W. and Zhong, S. *Nature Biotechnology*, **24**(10), 2006.

Table 4. "Rat *Toxicogenomics*" Validation Data Included in the MAQC Study: 180 microarrays

(Four microarray platforms in five laboratories)

Manufacturer	Code	Protocol	Platform	GEO GPL#	Array Express#	ArrayTrack Experiments	Number of Probes	Number of Test Sites	Number of Samples	Number of Replicates	Total Number of Microarrays
Applied Biosystems	ABI	One-Color Microarray	Rat Genome Survey Microarray	GPL2996	**A-MEXP-565**	ABI	26,857	1	6	6	36
Affymetrix	AFX	One-Color Microarray	Rat Genome 230 2.0 GeneChip®	GPL1355	**A-AFFY-43**	AFX AFX2	31,099	2	6	6	72
Agilent	AG1	One-Color Microarray	Whole Rat Genome Oligo Microarray, G4131A	GPL2877	**A-AGIL-19**	AG1	43,628 (41,070 – GeneSpring[a])	1	6	6	36
GE Healthcare	GEH	One-Color Microarray	Rat Whole Genome Bioarray, 300031	GPL2896	**A-GEHB-4**	GEH	35,129	1	6	6	36
											180

[a]Data from replicating spots were averaged within GeneSpring software to generate a single value for each unique probe. Therefore, the number of rows in the normalized data is fewer than that in the original data files from Agilent's Feature Extraction software. The rat toxicogenomics data set was analyzed in Guo, L. *et al*. *Nature Biotechnology*, **24**(9), 2006.

Table 5. "*Pilots*" Data (Pilot-I and Pilot-II) Included in the MAQC Study: 426 microarrays

1. MAQC Pilot-I (RNA Sample Selection): **160** microarrays (four human platforms; four RNA samples).
   - AFX (2 sites): 40
   - AGL (2 sites): 60
   - GEH (2 sites): 40
   - ILM (1 site): 20
2. MAQC Pilot-II (RNA Sample Titration): **266** microarrays (six human platforms; 13 titration points).
   - AFX (1 site): 45
   - AGL (1 site) 45
   - AG1 (1 site): 45
   - GEH (1 site): 46
   - ILM (1 site): 51
   - N/A (1 site): 34

Table 6. The total number of microarrays used in the MAQC project: 1,329

Category	Number of Microarrays
Official	573
Additional	130
Tumor	20
Rat Toxicogenomics	180
Pilots (I and II)	426
Total	***1,329***

Data from **903** microarrays (Official-573 + Additional-130 + Tumor-20 + Rat Toxicogenomics-180) based on 19 platforms need to be deposited into GEO and ArrayExpress. In addition, expression data from three alternative platforms (TAQ, QGN, and GEX) are also deposited, corresponding to **40** "virtual" microarrays in public database records (e.g., GEO GSM numbers). In total, there will be **943** GSM records in GEO.
Note: Pilot-I and Pilot-II data (426 microarrays) will not be deposited in public repositories until further notice.

Table 7. Original MAQC Data Files Distributed under http://edkb.fda.gov/MAQC/MainStudy/upload/

No.	Filename	Description of the Original Data
1	MAQC_ABI_123_60TXTs.zip	ABI non-normalized data.
2	MAQC_AFX_123456_120CELs.zip	Affymetrix CEL files.
3	MAQC_AG1_123_60TXTs.zip	Agilent FE8.5 output files.
4	MAQC_AGL_123_60TXTs.zip	Agilent FE8.5 output files.
5	MAQC_BIO1_1_10LSRs.zip	Output from SpotData software (CapitalBio Corp.).
6	MAQC_BIO_1_20LSRs.zip	Output from SpotData software (CapitalBio Corp.).
7	MAQC_EPP_123_180TXTGPRs.zip	Each array was scanned at three different PMT gain settings.
8	MAQC_GEH_1232Fail_80TXTs.zip	GenePix 4000B output files.
9	GEX	N/A
10	MAQC_H25K1_123_30GPRs.zip	GenePix output files.
11	MAQC_H25K_2_30GPRs.zip	GenePix output files.
12	MAQC_ILM_123_59TXTs.zip	Illumina non-normalized data.
13	MAQC_NCI2nd_3r4_20GPRs.zip	GenePix output files.
14	MAQC_NCI_122Fail3Fail_74GPRs.zip	GenePix output files.
15	MAQC_NMC_1_10GPRs.zip	GenePix output files.
16	MAQC_OPN_1_10GPRs.zip	GenePix output files.
17	QGN	N/A
18	MAQC_Rat_180Files.zip	Original data from ABI, Affymetrix, Agilent, and GE Healthcare platforms.
19	TAQ	N/A
20	MAQC_Tumor_20CELs.zip	Affymetrix CEL files.

*You can download ALL original data in one single file from GEO at ftp://ftp.ncbi.nih.gov/pub/geo/DATA/supplementary/series/GSE5350/ (3.99 GB!).

Table 8. Normalized MAQC Data Files Distributed under http://edkb.fda.gov/MAQC/MainStudy/upload/

No.	Filename	Description of the Normalized Data
1	norm_MAQC_ABI_123_qNorm.zip	The signal value is the fully corrected, background subtracted measurement of chemiluminescent signal for gene expression values. Data were quantile-normalized per test site (20 arrays).
2	norm_MAQC_AFX_123456_qPLIER16.zip	The CEL file data were analyzed with BioConductor to generate probeset level data using the justPlier() function. Probe-level data were quantile normalized before PLIER summarization per test site (20 arrays). An offset value of 16 was then added to each probeset-level data point.
3	norm_MAQC_AG1_123_Median1GeneSpring.zip	Data were transformed by setting all measurements less than 5.0 to 5.0. Data points that did not have detectable signal and those that represent microarray controls were labeled as Absent, those representing either non-uniform or saturated features were labeled as Marginal, and all remaining data points were labeled as Present. All data points were median-scaled to 1 using the median signal intensity value for data points labeled as Present.
4	norm_MAQC_AGL_123_FE8.5.zip	Ratio defined as CH1/CH2 (Cy3/Cy5), Cy3 and Cy5 channel intensities (background-subtracted). Data were extracted using Agilent's Feature Extraction software, version 8.5 (P/N G2567AA). AGL sample-pair convention: A (A/A, Cy3/Cy5), B (B/B, Cy3/Cy5), C (A/B, Cy3/Cy5), and D (B/A, Cy3/Cy5). Only MAQC samples A and B were used in AGL two-color experiments.
5	norm_MAQC_BIO1_1_Median1000.zip	Slides were scanned with a confocal LuxScan scanner (CapitalBio Corp.). The data from the obtained images were extracted with SpotData software (CapitalBio Corp.). Median-scaling of background-subtracted intensity was applied to each array to make median equal to 1000.
6	norm_MAQC_BIO_1_Median1000Lowess.zip	Ratio defined as CH1/CH2 (Cy3/Cy5), Cy3 and Cy5 channel intensities (background-subtracted). The data from the obtained images were extracted with SpotData software (CapitalBio Corp.). The raw data were submitted to MAQC and further normalized by using linear-scaling (to median of 1000) and LOWESS. BIO sample-pair convention: A (A/B, Cy3/Cy5), B (B/A, Cy3/Cy5), sA (A/A, Cy3/Cy5), sB (B/B, Cy3/Cy5). The human genome-wide long oligonucleotide microarray was constructed inhouse at CapitalBio Corporation, Beijing, China.
7	norm_MAQC_EPP_123_Ratio.zip	Each array was scanned at three different PMT gain settings. Eppendorf has developed specialized data analysis scheme treating the scanning, the background correction and normalization in a suited framework for low density arrays generating array to array gene expression ratio. The data sets submitted are divided in two groups. The first data set (norm) provides gene expression ratio comparing two samples and constitutes the standard result for the Eppendorf platform. Provided here are ratio data (compared to sample A1 for each test site).
8	norm_MAQC_GEH_1232Fail_Median1.zip	Local background subtracted intensity data after median-scaling each array to a median of 1. GEH's original flag information ('Quality_flag') is provided: G = 'Good', L = 'Absent', S = 'Saturated', M = 'Manufacturing Defect', I = 'Irregular Shape', C = 'Contaminated Spot'.
9	norm_MAQC_GEX_1_ACTB.zip	The data values are the number of molecules normalized to 1,000,000 ACTB molecules. No threshold was set; every gene was identified as Present ("0").
10	norm_MAQC_H25K1_123_Median1000.zip	Local background subtracted intensity data after median-scaling each array to a median of 1,000. GenePix's original flagging information - 0: Good; -50: Not found; -75: Absent; -100: Bad
11	norm_MAQC_H25K_2_Median1000Lowess.zip	Ratio defined as CH1/CH2 (Cy3/Cy5), Cy3 and Cy5 channel intensities (background-subtracted). The raw data were submitted to MAQC and further normalized by using linear-scaling (to median of 1000) and LOWESS. H25K two-color sample-pair convention: A (A/B, Cy3/Cy5), B (B/A, Cy3/Cy5), sA (A/A, Cy3/Cy5), and sB (B/B, Cy3/Cy5). Only test site 2 submitted two-color H25K data.
12	norm_MAQC_ILM_123_qNorm16.zip	Data was processed using Illumina BeadStudio version 1.5.0.34 software using background subtraction and cubic spline normalization using quantile (20 arrays per test site). Normalized hybridization intensity values were adjusted by adding a constant such that the lowest intensity value for any sample equaled 16. Flag_Detection = Present: >= 0.99, Absent: <0.99.

13	norm_MAQC_NCI2_3r4_Median1000Lowess.zip	Ratio defined as CH1/CH2 (Cy3/Cy5), Cy3 and Cy5 channel intensities (background-subtracted). The raw data were submitted to MAQC and further normalized by using linear-scaling (to median of 1000) and LOWESS. NCI2 sample-pair convention: A (A/B, Cy3/Cy5), B (B/A, Cy3/Cy5), sA (A/A, Cy3/Cy5). This is the second (and different) printing of the NCI. GenePix's original flagging information, - 0: Good; -50: Not found; -75: Absent; -100: Bad
14	norm_MAQC_NCI_122Fail3Fail_Median1000Lowess.zip	Ratio defined as CH1/CH2 (Cy3/Cy5), Cy3 and Cy5 channel intensities (background-subtracted). The raw data were submitted to MAQC and further normalized by using linear-scaling (to median of 1000) and LOWESS. NCI sample-pair convention: A (A/A, Cy3/Cy5), B (B/A, Cy3/Cy5), C (C/A, Cy3/Cy5), and D (D/A, Cy3/Cy5). MAQC sample A was used as the common reference sample (Cy5; channel 2) in NCI two-color experiments. Test sites NCI_2 and NCI_3 failed in the first round of experiments due to protocol issues. The failed hybridizations are named as NCI_2Fail and NCI_3Fail. NCI_2 repeated part of the experiment with 14 hybridizations. Site NCI_3 repeated part of the experiment wit hthe 2nd NCI printing (NCI_3r). GenePix's original flagging information, - 0: Good; -50: Not found; -75: Absent; -100: Bad
15	norm_MAQC_NMC_1_Median1000Lowess.zip	Ratio defined as CH1/CH2 (Cy3/Cy5), Cy3 and Cy5 channel intensities (background-subtracted). Images were processed and quantified using GenePix Pro 6.0.0.56. The data were filtered (or flagged) on the basis of signal levels and spot quality. The raw data were submitted to MAQC and further normalized by using linear median-scaling (to median of 1000) and LOWESS. NMC sample-pair convention: A (A/B, Cy3/Cy5), B (B/A, Cy3/Cy5). The slides were printed by the Norwegian Microarray Consortium. GenePix's original flagging information, - 0: Good; -50: Not found; -75: Absent; -100: Bad
16	norm_MAQC_OPN_1_Median1000Lowess.zip	Images were processed and quantified using GenePix Pro. The data were filtered (or flagged) on the basis of signal levels and spot quality. The raw data were submitted to MAQC and further normalized by using linear median-scaling (to median of 1000) and LOWESS. GenePix's original flagging information, - 0: Good; -50: Not found; -75: Absent; -100: Bad
17	norm_MAQC_QGN_1_Signal.zip	Background signals were determined in the absence of RNA samples and subtracted from signals obtained in the presence of RNA samples. Because QuantiGene assay measures RNA directly, no data normalization against a reference gene is required in the data analysis. Flag_Detection = "A": VALUE < LOD; "P": VALUE >= LOD. LOD = limit of detection of the assay, and equals to "Background + 3SD of Background".
18	norm_MAQC_RatTGx_4Platforms5Labs.zip	Each of the five labs ran 36 samples, resulting in 180 microarrays from four platforms. The data from each lab were analyzed according to the platform provider's preferred procedures as defined above for the human platforms.
19	norm_MAQC_TAQ_1_POLR2A.zip	POLR2A was chosen as the reference gene and each replicate CT was subtracted from the average POLR2A CT to give the log2 difference (delta CT), the normalized value. For delta CT calculations, a CT value of 35 was used for any replicate that had CT > 35. Provided is the normalized expression value, i.e., CT of POLR2A minus CT of the tested gene. Thus, a larger value indicates a higher gene expression level. Flag_Detection = "A": CT>35; "P": CT<=35.
20	norm_MAQC_Tumor_AFX_12_qPLIER16.zip	The CEL file data were analyzed with BioConductor to generate probeset level data using the justPlier() function. Probe-level data were quantile normalized before PLIER summarization per test site (10 arrays). An offset value of 16 was then added to each probeset-level data point. BioConductor-calculated PLIER Signal intensity with the addition of an offset value of 16. Flag information was generated using mas5calls() function in BioConductor: P = Present, A = Absent, and M = Marginal.

*Notes: 1. For quantile normalization, data from each test site were considered independently; 2. Detailed array annotation information can be found at GEO.

A list of the 943 GEO GSM records ("microarrays") and the corresponding original data files is available at http://edkb.fda.gov/MAQC/MainStudy/upload/GSE5350_943GSMs_MAQC_FileName_Mapping.txt.