Summary of the MAQC Data Sets

The MicroArray Quality Control (MAQC) Consortium

Leming.Shi@fda.hhs.gov
http://edkb.fda.gov/MAQC/
http://www.fda.gov/nctr/science/centers/toxicoinformatics/maqc/

September 8, 2006, 1:22 PM 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 from *Nature Biotechnology*'s website (http://www.nature.com/nbt/).
2. The MAQC data set will be publicly available starting from September 8, 2006 through the following four mechanisms:
    - GEO (series accession number: **GSE5350**);
    - ArrayExpress (accession number: **E-TABM-132**);
    - ArrayTrack (http://www.fda.gov/nctr/science/centers/toxicoinformatics/ArrayTrack/); and
    - The MAQC website (http://www.fda.gov/nctr/science/centers/toxicoinformatics/maqc/).
3. Image files (e.g., .DAT files from Affymetrix platform) are not deposited in GEO or ArrayExpress, 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 listed above (available on September 8, 2006).
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		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		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		OPN_1	37,584	OPN_1: Operon	2	5	10
NMC_ Operon	NMC	Two-Color Microarray	GPL4186		NMC_1	36,288	NMC_1: Norwegian Microarray Consortium	2	5	10
TeleChem	H25K	Two-Color Microarray	GPL4219		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		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

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

Official:	573
Additional:	130
Tumor:	20
Rat Toxicogenomics:	180
Pilots (I and II):	*426*

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) will also need to be 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.

Pilot-I and Pilot-II data (426 microarrays) will not be deposited in public repositories until further notice.

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/