Abbreviations

ERCC, External RNA Control Consortium
QRT-PCR, quantitative, real-time reverse transcriptase polymerase chain reaction
NIST, National Institute of Standards and Technology
polyA, polyadenylated
IVT, *in vitro* transcription
MW, molecular weight
DRC, dose response curve
Cy, Cyanine Dye
Ct, Cycle Threshold

Appendix

Appendix 1: Members of the external RNA controls consortium

Anne Bergstrom Lucas	Agilent Technologies, Inc.
Anne R. Kopf-Sill	NuGEN Technologies, Inc
Bin Chen	Centers for Disease Control and Prevention
Bud Bromley	ViaLogy Corp.
Carole Foy	LGC Ltd
Cecelia S. Hinkel	Centers for Medicare & Medicaid Services
Cecilie Boysen	ViaLogy Corp.
Chunmei Liu	Affymetrix Inc.
Daya Ranamukha-arachchi	FDA/CDRH/OSEL Division of Biology
Elizabeth Wagar	UCLA
Ernest S. Kawasaki	NCI/NIH
Federico M. Goodsaid	CDER/FDA
Friederike Wilmer	QIAGEN GmbH
Gavin Fischer	Stratagene
Gretchen L. Kiser	GE Healthcare
Helen C. Causton	Clinical Sciences Centre/Imperial College Microarray Centre
James C. Fuscoe	NCTR/FDA
James D. Brenton	University of Cambridge
Janet A. Warrington	Affymetrix, Inc.
Jesus Soriano	ATCC
John Coller	Stanford University
John D. Burrill	Applied Biosystems
Kate Rhodes	Cyntellect Incorporated
Kathleen F. Kerr	University of Washington
Kathryn C. Zoon	NIAID/NIH
Kathy Lee	Applied Biosystems
Laura H. Reid	Expression Analysis, Inc.
Leming Shi	NCTR/FDA
Marc Salit	NIST
Mary Satterfield	NIST

Matthew Marton	Rosetta Inpharmatics, LLC
Maureen Cronin	Genomic Health, Inc.
Michael P. Conley	Enzo Life Sciences, Inc.
Mickey Williams	Roche
Mike Fero	Stanford University
Mike Wilson	Ambion, Inc.
Natalia Novoradovskaya	Stratagene
Patrick Gilles	Invitrogen
Paul K. Wolber	Agilent Technologies, Inc.
Pranvera Ikonomi	American Type Culture Collection
Raj Puri	FDA/Center for Biologics Evaluation and Research
Richard P. Beyer	University of Washington
Richard Shippy	GE Healthcare
Robert Setterquist	Ambion, Inc.
Rosalie K. Elespuru	FDA/CDRH/OSEL Division of Biology
Shawn C. Baker	Illumina, Inc.
Stephen A. Chervitz	Affymetrix, Inc.
Steven R. Bauer	FDA/Center for Biologics Evaluation and Research
Steven Russell	University of Cambridge
Tamma Kaysser-Kranich	GE Healthcare
Theo K. Bammler	University of Washington
Thomas B. Ryder	Affymetrix, Inc.
Timothy J. Sendera	GE Healthcare
Uwe Scherf	CDRH/FDA
Xiaolian Gao	Atactic Technologies
Xiaoning Wu	Roche Molecular Systems, Inc.
Xu Guo	Affymetrix, Inc.
Z. Lewis Liu	USDA-ARS-NCAUR

Appendix 2: Summary of ERCC test plan workshop

The ERCC reviewed feedback on the Proposed Testing Plan at a NIST-hosted workshop on October 4-5, 2005. The meetings were attended by more than 50 participants, including ERCC members from Europe (Belgium, Germany, UK) and the US. During the first day of the meeting, summaries of the test plan were presented by representatives from each of the four subgroups. Suggested improvements and possible testing issues were discussed as listed below:

Reagent Production. Perhaps the ERCC set should include different types of controls (less than 100 bp or no polyA tail) that would support new labeling technologies or gene expression applications. Many members liked this possibility, and although it is outside the scope of the current ERCC initiative, it may develop into future ERCC activities. A request was made for early release of the submitted but not confirmed sequence of the external RNA controls. Another ERCC deliverable describing how to validate the quality of the external RNA controls was suggested.

Microarrays. The external RNA control concentrations used during phase 2 cross hybridization experiments as well as the exact definition of cross-reactivity were discussed. It was noted that identifying the specific transcript responsible for cross hybridization observed in a pool of 24 candidates may require many subsequent hybridizations with individual controls. Early definitions of the concentrations in the Latin square and dose response curve experiments are needed, with an emphasis on accommodating the ranges of both channels in two-color arrays. After a thorough discussion, it was decided to include pre-labeled external RNA controls in phase 2 of testing. Many of the microarray technology developers indicated that the test arrays might include multiple probes or probe sets in order to identify the best performing sequences. This format was acceptable, as long as participants agreed not to change probe sequence between phases.

QRT-PCR. The proposed test plan uses dilutions of equimolar pools of 96-140 the candidate RNA controls in the QRT-PCR assays. At the meeting, it was decided that plates of individual RNA transcripts would be more useful during the initial testing in phases 1-4. It was recognized that only 1-3 external RNA controls at a time would be used in QRT-PCR assays due to limitations in the number of fluorophors that can be detected. There was also some discussion that QRT-PCR assay optimization may require more than 2 primer sets described in the proposed test plan. A new application for the controls in the validation of thermal cyclers and other QRT-PCR instruments was discussed.

Informatics. The ERCC will need a plan for data storage and distribution. NCI, NCBI, and FDA representatives volunteered to provide a data repository for Test Plan data. A working group will be formed to identify needs of the ERCC and identify which of these data repositories would be a good fit. Participants were directed to the CLSI MM16-P document for a list of possible performance metrics. Preliminary review of these metrics could be initiated now using already available gene expression data on many of the candidate external RNA controls. There was some discussion on whether the external

RNA controls should be optimized and selected to have the best possible results (e.g. perfect dose response curve, no cross hybridization) or left imperfect to better reflect typical probes or probe sets on the microarrays.

After nine hours of review and discussion, consensus was achieved on the ERCC Test Plan by a show of hands with no major outstanding issues. On the second day of the meeting, preliminary resource requirements were identified and the scope of testing (sample number etc.) was better defined. Key features of this discussion are summarized below:

1. In an effort to efficiently utilize the RNA reagents, early stages of testing (Phases 1-4) will be performed only at developer sites and limited to one site per platform. We estimate that up to 10 microarray and 4 QRT-PCR commercial developers will participate. If resources permit, 3 or 4 noncommercial development sites (e.g. NCI or USDA) will be included to represent home-made, spotted microarrays.

2. Phase 5 will be divided into two parts with separate goals. In Phase 5A, the last experiment of Phase 4 will be repeated at user sites to confirm that results can be reproduced using the same materials and protocols as at the developer sites. In Phase 5B, a set of the best performing external RNA controls will be distributed to user sites and incorporated into typical experiments with their materials and protocols. The intent is to establish performance of the external RNA controls in a broad variety of routine applications and protocols. One possible Phase 5B experiment would be to repeat a previous gene expression experiment to validate that similar differential expression results are observed with and without the inclusion of external RNA controls.

3. A tentative list of criteria for technology developers and user test sites was developed. It was decided that all sites should have the following attributes: 1) Demonstrated experience in gene expression platform; 2) Ability to contribute materials and labor; 3) Commitment to deposit all testing data in publicly available database; and 4) Agreement to meet ERCC timeline. In addition, developer sites should have demonstrated capacity to design and create the necessary reagents as well as a commitment to follow the test plan and collect data on all candidate clones. User test sites may be asked to contribute a novel application to the ERCC goals and to provide sample material for the complex RNA background. It is not necessary to have contributed sequences to be considered as a test site.

4. While the external RNA controls are likely to be useful in RNA samples from a variety of species, testing will emphasize human, mouse and rat RNA. A Stratagene representative offered to provide universal reference RNA samples from these three species to be used as the complex RNA background in Phases 1-4. Test microarrays will include probes or probe sets for both the candidate external RNA controls as well as genes included on their typical human, mouse and rat arrays (as determined by each developer). This design will enable global normalization methods and aid in cross hybridization experiments. The developers may elect to pre-screen their test arrays using RNA from a variety of species.

Action items and next steps will be further discussed in the monthly ERCC conference calls.

References

1. Cronin M, Ghosh K, Sistare F, Quackenbush J, Vilker V, O'Connell C: **Universal RNA reference materials for gene expression.** *Clin Chem* 2004, **50**:1464–1471.
2. **National Institute of Standards and Technology** [http://www.cstl.nist.gov/biotech/Cell&TissueMeasurements/GeneExpression/ERCC.htm]
3. **Dr. Janet Warrrington** [janet_warrington@affymetrix.com]
4. **Affymetrix GeneChip® Poly-A RNA Control Kit** [http://www.affymetrix.com/products/reagents/specific/poly_a.affx]
5. **GE Healthcare/Amersham Biosciences Codelink™ Whole Genome Controls** [http://www4.amershambiosciences.com/aptrix/upp00919.nsf/(FileDownload)?OpenAgent&docid=4F44E8ADAB46FEB3C1256EB400418062&file=63005456.pdf]
6. **Stratagene SpotReport™ Alien™ Array Validation System** [http://www.stratagene.com/products/showCategory.aspx?catId=17]
7. Choe SE, Boutros M, Michelson AM, Church GM, Halfon MS: **Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset.** *Genome Biology* 2005, **6:**R16 [http://genomebiology.com/2005/6/2/R16].
8. Kuhn K, Baker SC, Chudin E, Lieu M-H, Oeser S, Bennett H, Rigault P, Barker D, McDaniel TK, Chee MS: **A novel, high-performance random array platform for quantitative gene expression profiling.** *Genome Res* 2004, **14**:2347–2356.
9. Lockhart DJ, Dong H, Byrne MC, Follettie T, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL: **Expression monitoring by hybridization to high-density oligonucleotide arrays.** *Nature Biotechnology* 1996, **14**:1675-1680.
10. **Applied Biosystems ABI Prism 7770 Sequence Detection System User Bulletin #2.** [http://docs.appliedbiosystems.com/search-dodnum.taf?dodnum=4303859]
11. Innis MA, Gelfand DH, Sninsky JJ: *PCR Applications: Protocols for Functional Genomics.* Academic Press; 1999.
12. Vandecasteele SJ, Peetermans WE, Merckx R, and Van E: **Quantification of expression of *Staphylococcus epidermidis* housekeeping genes with Taqman quantitative PCR during in vitro growth and under different conditions.** *J Bacteriol* 2001, **183**:7094-7101.
13. Radonic A, Thulke S, Mackay IM, Landt O, Siegert W, Nitsche A: **Guideline to reference gene selection for quantitative real-time PCR.** *Biochem Biophys Res Commun* 2004, **313**:856-862.
14. Janssens N, Janicot M, Perera T, Bakker A: **Housekeeping genes as internal standards in cancer research.** *Mol Diagn* 2004, **8**:107-13.
15. Jeong YJ, Choi HW, Shin HS, Cui XS, Kim NH, Gerton GL, Jun JH: **Optimization of real time RT-PCR methods for the analysis of gene expression in mouse eggs and preimplantation embryos.** *Mol Reprod Dev* 2005, **71**:284-9.
16. Lee PD, Sladek R, Greenwood CMT, Hudson TJ: **Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies.** *Genome Research* 2002, **12**:292-297.
17. van de Peppel J, Kemmeren P, van Bakel H, Radonjic M, van Leenen D, Holstege FCP: **Monitoring global messenger RNA changes in externally controlled microarray experiments.** *EMBO Reports* 2003, **4**:387-393.

18. **MicroArray Quality Control Project**
[http://www.fda.gov/nctr/science/centers/toxicoinformatics/maqc/index.htm]
19. Box GEP, Hunter WG, Hunter JS: *Statistics for Experimenters*. New York: John Wiley & Sons; 1978.

Figure Legends

Figure 1: Illustrations of latin square and graeco-latin square designs
"A1" to "A4" number the 4 arrays used in the experiment, "G1" to "G4" number the 4 transcripts being studied and "L1" to "L4" denote 4 different concentrations for each transcript. The four pools of transcripts are labeled "W" to "Z". "g" and "r" note the gene concentrations or pools used in the green or red channel, respectively of a two-color experiment.

Figure 2: Illustration of chi square fit
Panel A. The distances from a straight-line fit (arrows) are calculated.
Panel B. The Chi square fit of the distances is then determined.

Figure 3: Illustration of spike performance

Figure 4: Illustration of model data (including modeled noise)
The values of m and b that were input into the model were m = 0.85 and b = 0.08. The noise model is realistic, in that it includes both constant (scanner) and proportional (chemical) noise.

Tables

Table 1: Summary of external RNA control clone library

Number	*Affiliation of Contributor*	*Genus species*	*Length of RNA*
1 - 28	Affymetrix	*B. subtilis*	700-2,000
29 - 40	Affymetrix	Artificial Sequences	500-1,900
41 - 43	USDA-ARS-NCAUR	*Bos taurus*	500
44 - 46	USDA-ARS-NCAUR	*Glycine max*	500
47	Ambion	Lamda phage	1,000
48 - 53	Ambion	Artificial Sequences	750-1,000
54 - 61	Ambion	*E. coli*	750-2,000
62 - 82	Stanford University	*Methanococcus*	500-750
83 - 85	Agilent Technologies	Artificial Sequences	500
86 - 90	GE Healthcare	*E. coli*	1,000
91 - 140	Affymetrix/Ambion/Atactic	Artificial Sequences	1,000

Table 2: Summary of testing phases

Testing Phase	*Specific Aim*	*Milestone*
1 – Design & Development	Generate Reagents	Distribution for prototype testing
2 – Prototype Testing	Validate Reagents	Initial data collected, acceptance criteria established
3 – Proof of Concept	Validate Assay	Candidate set of ERCC clones
4 – Functional Testing	Validate Product	Final set of ERCC clones
5 – Performance Review	Distribute Product	Symposium

Table 3: Description of pools and experiments in microarray testing

Pool	*External RNA Clones*	*Background RNA*	*No. of Arrays*
0	1 to 144	none	0
1	1 to 48 pre-labeled	none	3
2	49 to 96 pre-labeled	none	3
3	97 to 144 pre-labeled	none	3
4	1 to 144pre-labeled	none	0
5	1 to 72 (high conc.) 73 to 144 (low conc.)	human	3
6	1 to 72 (low conc.) 73 to 144 (high conc.)	human	3
7*	1 to 96 (diff. conc.)	human	12
8*	1 to 96	human	12
9*	1 to 96	human	12
10*	1 to 96	human	12
11	1 to 96	human	0
12	1 to 96	human	3
13	1 to 96	human	3
14	1 to 96	human	3
		Total	**75**

**Pools 7-10 may be further diluted to expand the concentration range tested.*
Array count is per one-color platform and does not include background RNA negative control samples.
Pools 0 and 11 will also be used in QRT-PCR assays.

Table 4: Concentration of controls in dilution 1 pools for modified latin square experiments

Pool	*Concentration Group A (Controls 1- 24)*	*Concentration Group B (Clones 25-48)*	*Concentration Group C (Clones 49-72)*	*Concentration Group D (Clones 73-96)*
Pool 7	125	1	5	25
Pool 8	25	125	1	5
Pool 9	5	25	125	1
Pool 10	1	5	25	125

Concentration is given as mass ratios, so that "125" represents 1:125,000 or 1ng of RNA transcript per 125 ng of background RNA where the spike amount is adjusted for its length.

Table 5: Modified latin square hybridization setup

Controls Group	*Pool 7*	*Pool 8*	*Pool 9*	*Pool 10*
A	Conc. 1	Conc. 2	Conc. 3	Conc. 4
B	Conc. 4	Conc. 1	Conc. 2	Conc. 3
C	Conc. 3	Conc. 4	Conc. 1	Conc. 2
D	Conc. 2	Conc. 3	Conc. 4	Conc. 1

Table 6: Concentration of controls in dilution pools for expanded range experiments

Pool	*Conc. A Dil. 2*	*Conc. B Dil. 2*	*Conc. C Dil. 2*	*Conc. D Dil. 2*	*Conc. A Dil. 3*	*Conc. B Dil. 3*	*Conc. C Dil. 3*	*Conc. D Dil. 3*	*Conc. A Dil. 4*	*Conc. B Dil. 4*	*Conc. C Dil. 4*	*Conc. D Dil. 4*
Pool 7	250	2	10	50	500	4	20	100	5,000	40	200	1,000
Pool 8	50	250	2	10	100	500	4	20	1,000	5,000	40	200
Pool 9	10	50	250	2	20	100	500	4	200	1,000	5,000	40
Pool 10	2	10	50	250	4	20	100	500	40	200	1,000	5,000

Concentration is given as mass ratios, so that "250" represents 1:250,000 or 1 ng of RNA transcript per 250 ng of background RNA, where the spike amount is adjusted for its length.
Concentrations of the initial stock, "Dilution 1" pools are shown in Table 4.

Table 7: Expected red:green ratios in two-color hybridizations

Array	*Pool in Green Channel*	*Pool in Red Channel*	*Group A Ratio*	*Group B Ratio*	*Group C Ratio*	*Group D Ratio*
1	P7-D1	P10-D1	0.008	5	5	5
2	P8-D1	P9-D1	0.2	0.2	125	0.2
3	P9-D1	P8-D1	5	5	0.008	5
4	P10-D1	P7-D1	125	0.2	0.2	0.2
5	P7-D2	P8-D2	0.2	125	0.2	0.2
6	P8-D2	P7-D2	5	0.008	5	5
7	P9-D2	P10-D2	0.2	0.2	0.2	125
8	P10-D2	P9-D2	5	5	5	0.008
9	P7-D3	P9-D3	0.04	25	25	0.04
10	P8-D3	P10-D3	0.04	0.04	25	25
11	P9-D3	P7-D3	25	0.04	0.04	25
12	P10-D3	P8-D3	25	25	0.04	0.04
13*	P7-D4	P7-D4	1	1	1	1
14*	P8-D4	P8-D4	1	1	1	1
15*	P9-D4	P9-D4	1	1	1	1
16*	P10-D4	P10-D4	1	1	1	1

**Self-to-self hybridizations*
Pools in each channel are labeled based on their pool and dilution numbers in Table 6.
Groups are sets of transcripts at the same concentration as defined in Table 4.

Table 8: Example single array DRC pool

Concentration	*No. of Targets*
1,000	2
2,000	2
4,000	2
5,000	0
10,000	4
20,000	4
25,000	0
40,000	10
50,000	0
100,000	12
125,000	0
200,000	12
250,000	12
500,000	12
1,000,000	12
5,000,000	12
Total	**96**

Concentration is given as mass ratios, so that "1,000" represents 1:1,000 or 1ng of RNA transcript per 1,000 ng of background RNA where the spike amount is adjusted for its length

Table 9: Illustration of a nomenclature system

Reagent	*Nomenclature*	*Legend*
RNA Transcript	ERCC-nnnnn-vv	nnnnn = unique 5-digit sequence number vv = 2 digit version number
PCR primer/probe microarray probe	ERCC-nnnnn-vv-pppp-lll-aaa	pppp = 4-digit positional location relative to the 0^{th} base at the 5' end of the transcript sequence lll = primer/probe length aaa = nucleic acid sequence of the initial triplet of the primer/probe (complement of the RNA sequence at pppp, pppp+1, pppp+2)

Figure 1

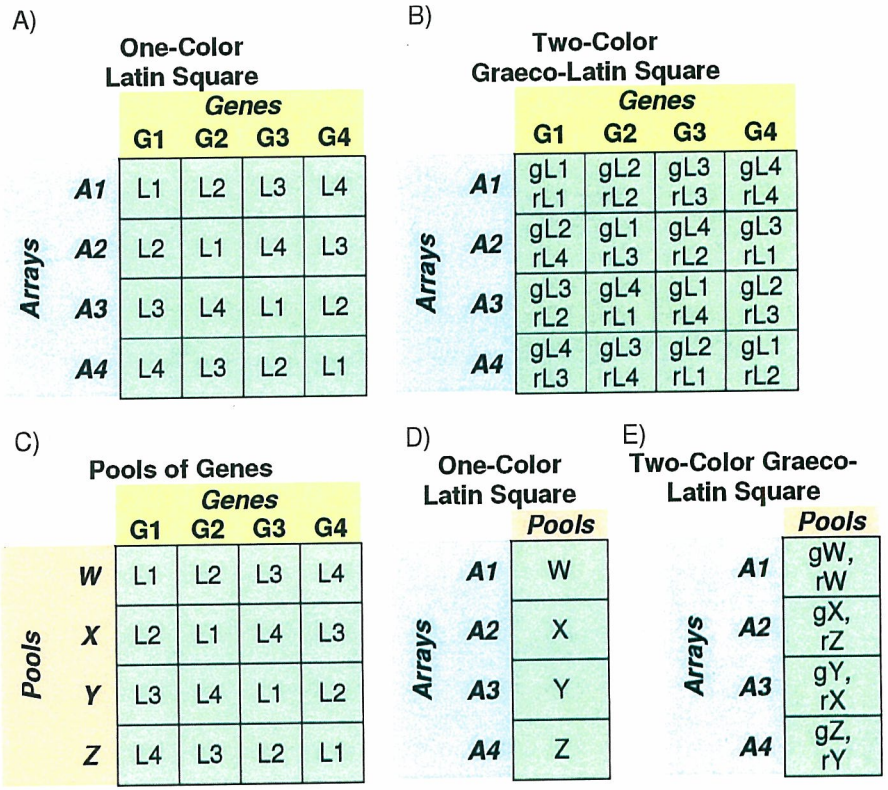
A)
One-Color
Latin Square
Genes
G1 G2 G3 G4
Arrays
A1 L1 L2 L3 L4
A2 L2 L1 L4 L3
A3 L3 L4 L1 L2
A4 L4 L3 L2 L1
B)
Two-Color
Graeco-Latin Square
Genes
G1 G2 G3 G4
Arrays
A1 gL1 rL1 gL2 rL2 gL3 rL3 gL4 rL4
A2 gL2 rL4 gL1 rL3 gL4 rL2 gL3 rL1
A3 gL3 rL2 gL4 rL1 gL1 rL4 gL2 rL3
A4 gL4 rL3 gL3 rL4 gL2 rL1 gL1 rL2
C)
Pools of Genes
Genes
G1 G2 G3 G4
Pools
W L1 L2 L3 L4
X L2 L1 L4 L3
Y L3 L4 L1 L2
Z L4 L3 L2 L1
D)
One-Color
Latin Square
Pools
Arrays
A1 W
A2 X
A3 Y
A4 Z
E)
Two-Color Graeco-
Latin Square
Pools
Arrays
A1 gW, rW
A2 gX, rZ
A3 gY, rX
A4 gZ, rY

Figure 2

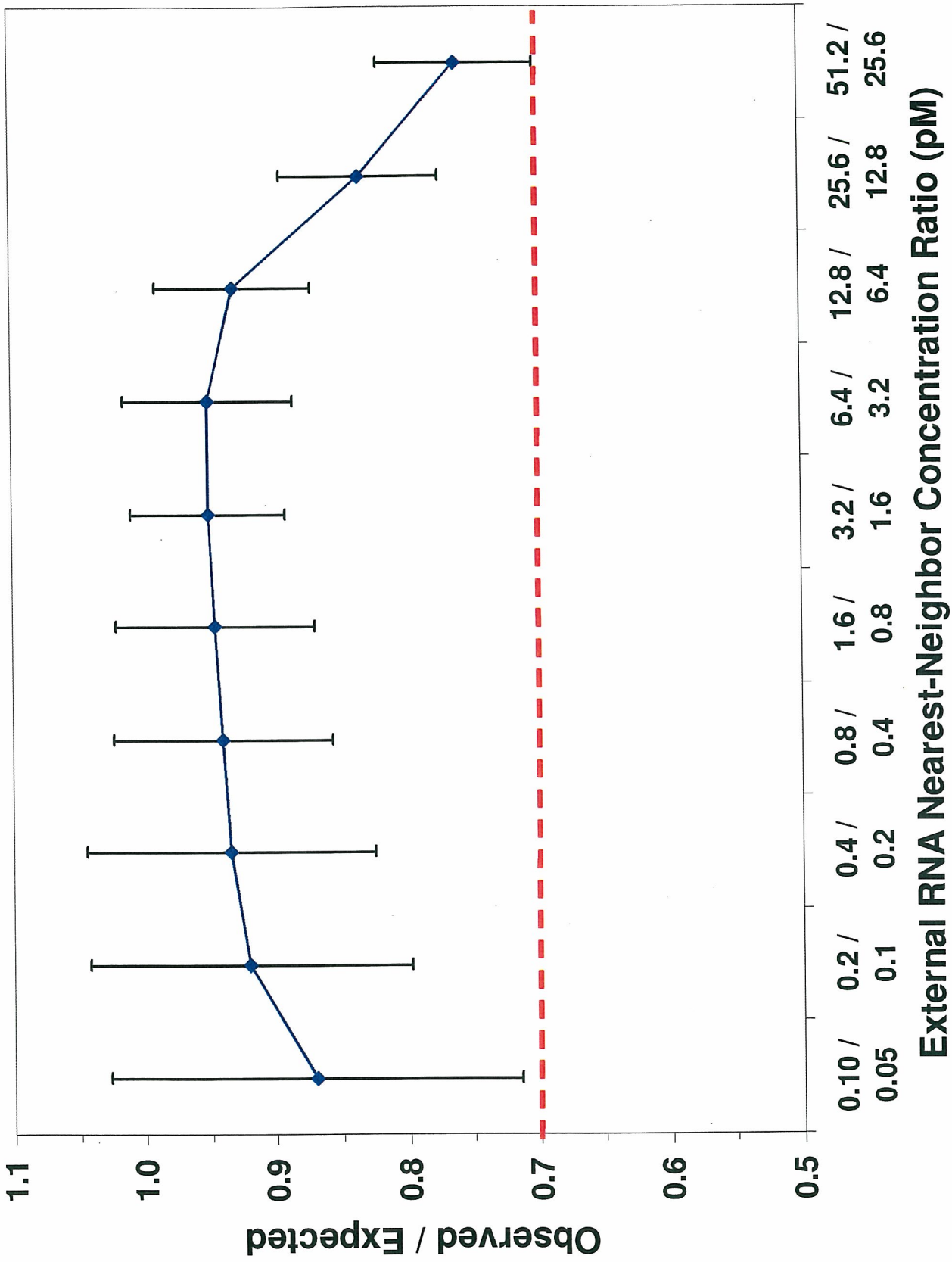
Observed / Expected
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.10 / 0.05
0.2 / 0.1
0.4 / 0.2
0.8 / 0.4
1.6 / 0.8
3.2 / 1.6
6.4 / 3.2
12.8 / 6.4
25.6 / 12.8
51.2 / 25.6
External RNA Nearest-Neighbor Concentration Ratio (pM)

Figure 3

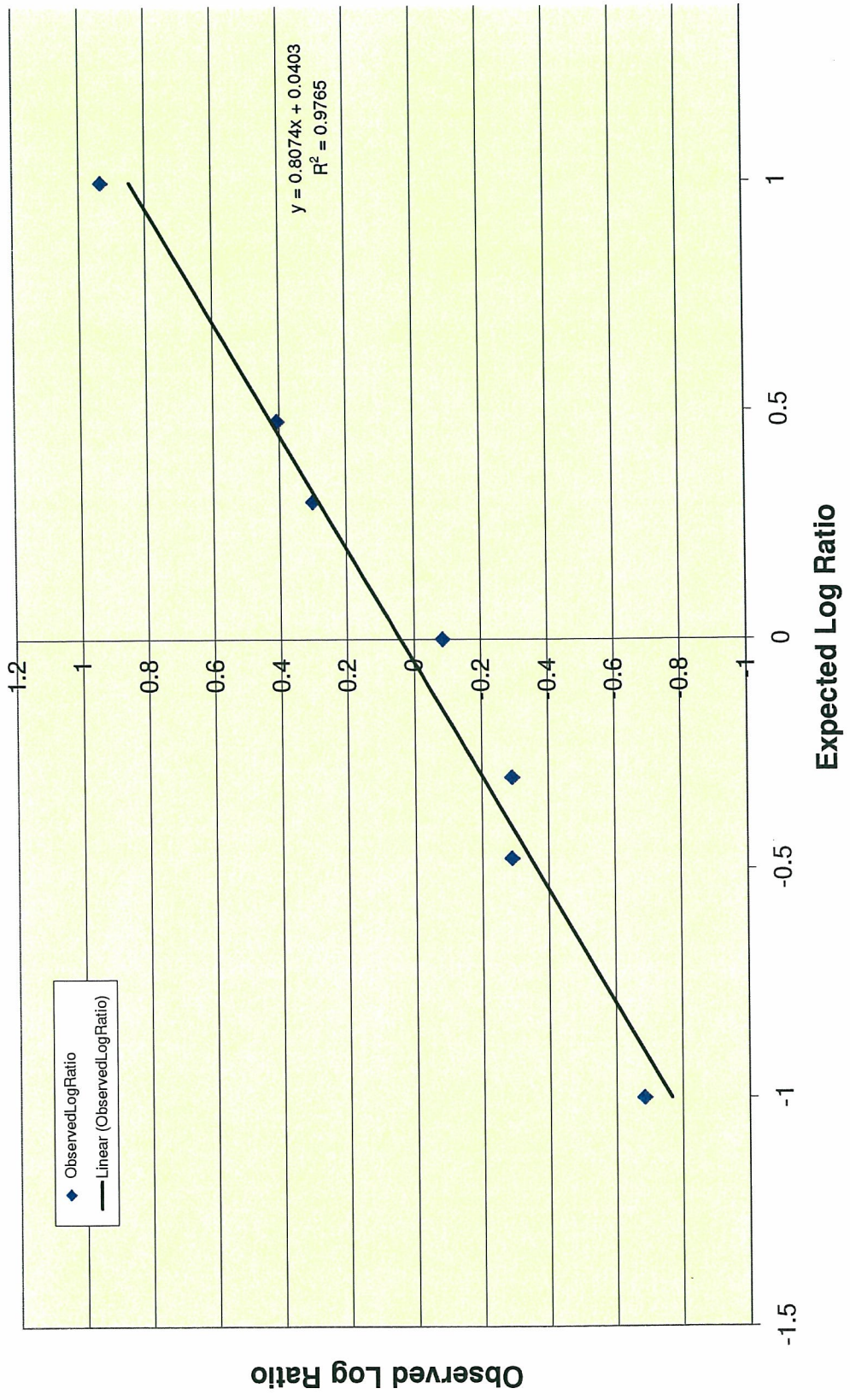
Observed Log Ratio
Expected Log Ratio
y = 0.8074x + 0.0403
R2 = 0.9765
ObservedLogRatio
Linear (ObservedLogRatio)
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1.5
-0.5
0
0.5
1

Figure 4

Comparative and Functional Genomics
Comp Funct Genom 2004; **5**: 633–641.
Published online in Wiley InterScience (www.interscience.wiley.com). **DOI:** 10.1002/cfg.447

Conference Paper

Standardization initiatives in the (eco)toxicogenomics domain: a review

Susanna Assunta Sansone[1]*, Norman Morrison[2], Philippe Rocca-Serra[1] and Jennifer Fostel[3]
[1]*EMBL-EBI, The European Bioinformatics Institute, Cambridge CB10 1SD, UK*
[2]*The University of Manchester, Kilburn Building, School of Computer Science, Oxford Road, Manchester M13 9PL, UK*
[3]*National Institute of Environmental Health Sciences, National Center for Toxicogenomics. Research Triangle Park, NC 27709, USA*

**Correspondence to:*
Susanna Assunta Sansone, EMBL-EBI, The European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK.
E-mail: sansone@ebi.ac.uk

Revised: 24 November 2004
Accepted: 26 November 2004

Abstract

The purpose of this document is to provide readers with a resource of different ongoing standardization efforts within the 'omics' (genomic, proteomics, metabolomics) and related communities, with particular focus on toxicological and environmental applications. The review includes initiatives within the research community as well as in the regulatory arena. It addresses data management issues (format and reporting structures for the exchange of information) and database interoperability, highlighting key objectives, target audience and participants. A considerable amount of work still needs to be done and, ideally, collaboration should be optimized and duplication and incompatibility should be avoided where possible. The consequence of failing to deliver data standards is an escalation in the burden and cost of data management tasks. Copyright © 2005 John Wiley & Sons, Ltd.

Keywords: toxicogenomics; ecotoxicogenomics; toxicology; environment; functional genomics; standards; database

Introduction

Molecular-based approaches, such as transcriptomics, proteomics, metabolomics and metabonomics, are being used to study the impact of chemicals on human and wildlife populations. These high-throughput (eco)toxicogenomics investigations are information-intensive and, by producing massive amounts of data, have placed the informatics challenge under the spotlight. The need to provide easy access to integrated data in a structured standard format is clearly significant. Several efforts are already under way to promote standardization, tackle data management issues and develop databases to facilitate data exchange. We have seen the value of these collaborative efforts already. The Microarray Gene Expression Data (MGED; **http://www.mged.org**) Society has been successful in developing the MIAME standard and related ontology and object models for microarray data (reviewed in Quackenbush 2004). The Reporting Structure for Biological Investigations (RSBI; **http://www.mged.org/Workgroups/rsbi**) is a new working group formed under the MGED Society umbrella, planning to act as a 'single point of focus' for Toxicogenomics, Environmental Genomics and Nutrigenomics communities working towards an international and compatible informatics platform for data exchange. Discipline-specific initiatives are regarded as important because they target 'real world' data capture requirements for the particular omics technologies being used. A consequence of this, however, is that, by remaining within each given discipline, the standardization effort fragments, resulting in duplication and the development of different terminology and data models, thereby limiting the potential for data exchange. One of the objectives of the RSBI working group is to ensure that these initiatives are coordinated, so that synergy and cross-discipline communication can be maximized, and duplicated effort can be minimized.

To capitalize on these efforts, representatives of the RSBI working group are also directly participating in certain initiatives and, by fostering interactions, are laying the ground for further collaborations. One forum for such interaction is the Standards and Ontologies for Functional Genomics (SOFG; **http://www.sofg.org**) Conference. We invite comments on the work of the RSBI at **mged-rsbi@lists.sourceforge.net**

Standardization initiatives

Data standardization is now considered beyond the research application of high-throughput technologies (reviewed in Quackenbush, 2004) and regulatory bodies, such as the US Food and Drug Administration (FDA) and Environmental Protection Agency (EPA), are developing their policy or guidance on genomics data submissions (**http://www.fda.gov/cder/guidance/5900dft.doc; http://www.epa.gov/osa/genomics.htm**). Several organizations and committees are tackling data standardization; however, there is a fundamental difference in both the design and objectives of the efforts around regulatory submission of data vs. the needs of the research community, who need databases and tools for discovery. The former aims to accelerate the review process, facilitate proprietary data submission and optimize data visualization in a way that does not impact the vocabulary used by the individual submitter. The research community needs to ease deposition in public databases and facilitate data mining by the use of common annotation standards and ontologies. There is some overlap between the needs of these communities and some level of interaction. Thus, there is value in assessing the commonality between regulatory, research community and database designers' objectives in the design of data standards. Specifically, a unified approach to describing and reporting the experimental biological metadata that is common to the different 'omics' technologies (transcriptomics, proteomics and metabonomics/metabolomics) or disciplines (e.g. pharmacogenomics, toxicogenomics, environmental genomics) is a goal of the RSBI. Undoubtedly specialized information is needed by certain applications, but a high-level unified model for description of metadata would be able to encompass these applications. Here, metadata, refers to biological information relating to samples and the information about experimental design. Data refers to measured values relating to samples (e.g. toxicological endpoints and gene expression) under given experimental conditions.

This paper is not an exhaustive list of all activity but provides a summary of standardization efforts for toxicological and environmental applications, which address reporting standards (e.g. what should be reported), and management issues (e.g. how reported information should be stored and exchanged, and which ontologies should be used to annotate data and metadata). The various initiatives fall into six broad categories, summarized in Table 1 and explored in detail below.

'Omics' technology communities

These are academic grass roots communities that have joined forces with commercial vendors to address content standards and reporting needs for a single high-throughput technology.

MGED Society

The MGED Society has established standards for microarray data annotation (MIAME; Brazma *et al.*, 2001; Ball *et al.*, 2002) and exchange (MAGE-ML; Spellman *et al.*, 2002) that have facilitated the creation of microarray databases and related supporting software (MAGE-OM; Spellman *et al.*, 2002). The response from the scientific community to these community standards has been extremely positive (Editorial, 2002). Most of the major scientific journals and some funding agencies require publications describing microarray experiments to comply with MIAME, for the data to be submitted to public repositories, such as ArrayExpress (Brazma *et al.*, 2003), GEO (Edgar *et al.*, 2002) and CIBEX (Ikeo *et al.*, 2003). Consequently, the MIAME model has been adopted by other communities (Quackenbush, 2004). MGED is now working with other initiatives, such as HUPO-PSI in the proteomics field and SMRS (see below). There have been several extensions to MIAME: MIAME-Tox, an array-based toxicogenomics standard developed by the ILSI Health and Environmental Sciences Institute (HESI) (**http://hesi.ilsi.org/index.cfm?pubentityid=120**); the National Institute of Environmental

Table 1. The initiatives divided according to six broad categories

Category	Description	Acronym	Domain	URL
Omics technology communities	Academic grass roots communities that have joined forces with commercial vendors to create technology-driven standards	MGED	Microarray	**http://www.mged.org**
		PSI	Proteomics	**http://psidev.sourceforge.net**
		SMRS	Metabolomics and metabonomics	**http://www.smrsgroup.org**
Measurement and methods validations	Efforts focusing on validation programs and production of standard materials and methods	ECVAM	Array-based toxicogenomics	**http://ecvam.jrc.cec.eu.int**
		ERCC	Microarrays and quantitative RT-PCR	**http://www.cstl.nist.gov/biotech/workshops/ERCC2004**
		MARG	Microarray	**http://www.abrf.org/index.cfm/group.s how/Microarray.30.htm**
		ABRF		
		MFB		**http://www.mfbprog.org.uk**
Regulatory driven discussion fora	Efforts aiming for a broader understanding and use of omics data, defining data models for data submission to regulators. That preserve the terms and observations used by the submitter	CDISC	Clinical data	**http://www.cdisc.org**
		PGx	Pharmacogenomics data	To be announced
		SEND	Animal toxicity data	**http://www.cdisc.org/models/send/v1.5**
Domain-driven discussion fora	Efforts aiming to a broader exchange and integration of toxicity and ecological data	DSSTox	Chemical toxicity data	**http://www.epa.gov/nheerl/dsstox/**
		SEEK	Ecological data	**http://seek.ecoinformatics.org**
World-wide organizations	Efforts producing internationally agreed instruments, decisions and recommendations or acting as facilitator	IPCS	Toxicogenomics	**http://www.who.int/ipcs/en/**
		NAS	(Eco)toxicogenomics	**http://dels.nas.edu/emerging-issues**
		OECD	Ecotoxicogenomics	**http://www.oecd.org**
		BSC IEEE	Bioscience	**http://www.csbcon.org**
Infrastructure	Standards-compliant infrastructure, assisting in development of useful and usable standards	ArrayExpress and Tox-MIAMExpress	Array-based data and toxicology endpoints values	**http://www.ebi.ac.uk/arrayexpress** **http://www.ebi.ac.uk/tox-miamexpress**
		CEBS	Toxicogenomics	**http://cebs.niehs.nih.gov**
		CTD	Genes and proteins	**http://ctd.mdibl.org**
		maxd	Array-based data and environmental metadata	**http://bioinf.man.ac.uk/microarray/maxd**
		TIS (ArrayTrack)	Toxicogenomics	**http://www.fda.gov/nctr/science/centers/toxicoinformatics/ArrayTrack**

Health Sciences (NIEHS); the National Center for Toxicogenomics (NCT; **http://www.niehs.nih.gov/nct**); the FDA National Center for Toxicological Research (NCTR; **http://www.fda.gov/nctr**); and the European Bioinformatics Institute (EBI; **http://www.ebi.ac.uk**). MIAME/Env has been developed by the NERC Environmental Genomics Thematic Programme Data Centre (EGTDC; **http://envgen.nox.ac.uk**) to fulfil the diverse needs of those working in the functional genomic of ecosystems, invertebrates and vertebrates which are not covered by the model organism community. However, extending MIAME to meet domain-specific requirements is only a partial solution. As multi-technology investigations become commonplace, these checklists will soon be insufficient. Currently, the above communities are working together with the RSBI group to develop a reporting structure for describing multi-platform technologies investigations. The proposed RSBI Tiered Checklist (RSBI TC; **http://www.mged.org/Workgroups/rsbi**) will be a modular context-dependent structure.

Proteomics Standardization Initiative (PSI)

The HUPO (Human Proteome Organization; **(http://www.hupo.org)** PSI **(http://psidev.sourceforge.net)** includes the major protein databases, government and industry and is defining standards for data representation in proteomics to facilitate data comparison, exchange and verification. Current focus is on mass spectrometry and protein–protein interaction data. A set of open source standards are being developed along MIAME lines, including a content standard, the Minimum Information About Proteomics Experiments (MIAPE), an XML standard data exchange format (Hermjakob *et al.*, 2004) and an ontology of clearly defined general proteomics terms.

Standard Metabolic Reporting Structure (SMRS)

SMRS **(http://www.smrsgroup.org)** comprises industry, software developers, governmental representatives and academia, who are investigating the reporting and design of metabonomics and metabolomics studies in plants, microbial systems, environment, *in vivo* and *in vitro* applications, as well as human studies. A set of draft recommendations has been produced as a discussion document. It considers the factors in a metabolic study that could be recorded and standardized, including the origin of a biological sample, the technologies and methods for analysis and the chemometric and statistical approaches. The recommendations also touch on the granularity of information required for different reporting needs, including journal submissions, public databases and regulatory submissions.

Measurement and methods validations

As high-throughput technologies are used in industry and are considered by regulatory agencies, the methodology itself comes under scrutiny. Agreement on data formats will do little good if experimental protocols are inconsistent. Currently, standardization of microarray experiment procedures is key to the broad acceptance and use of these data. The very variability of microarray data generation, analysis, future validation of the technology and production of standard materials is now the focus of many initiatives.

MfB (Measurements for Biotechnology) program

MfB **(http://www.mfbprog.org.uk)** is a UK programme that addresses bio-measurements of importance for industry. The 'Comparability of Gene Expression Measurements on Microarrays' is an industry-based consortium led by LGC **(http://www.lgc.co.uk)**. The project is designed to determine the accuracy and comparability of gene expression measurements made on different array platforms and also evaluates data analysis methods. A second phase is now looking at the standardization of array-based toxicogenomics and will build up on the analysis framework to develop a panel of quality metrics for validating and standardizing array-based toxicogenomics measurements.

The Microarray Research Group (MARG) of the Association of Biomolecular Resource Facilities (ABRF)

The MARG **(http://www.abrf.org/index.cfm/group.show/Microarray.30.htm)** is a research-focused consortium of academic laboratories promoting communication and cooperation among core academic and industrial microarray and data analysis services providers. The resulting data is used to help laboratories evaluate their performance and achieve the highest quality results possible from the use of microarray technologies.

The European Centre for the Validation of Alternative Methods (ECVAM)

The ECVAM **(http://ecvam.jrc.cec.eu.int)** coordinates and funds validation studies of alternative methods that could reduce, refine or replace the use of laboratory animals in regulatory toxicology. Both the new EU Chemical Policy (REACH) (Editorials, 2003a, 2003b) that proposes the re-evaluation of about 30 000 chemicals, and the 7th Amendment to the Cosmetics Directive, which foresees the complete replacement of animal experiments by 2013, call for the development and implementation of alternative methods. ECVAM is working with the US Interagency Coordinating Committee on the Validation of Alternative Methods (ICCVAM; **http://iccvam.niehs.nih.gov/home.htm**) and National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods (NICEATM; **http://iccvam.ni-**

ehs.nih.gov/home.htm) to investigate the specific considerations necessary for adequate validation of array-based toxicogenomics-based test methods. At present, recommendations are being prepared which will cover topics such as description of the biological systems, methodological/technical issues, data analysis, and data format and storage.

External RNA Controls Consortium (ERCC)

ERCC (**http://www.cstl.nist.gov/biotech/workshops/ERCC2004**) originated at a US National Institute of Standards and Technology (NIST; **http://www.nist.gov**) meeting and is composed of representatives from the public, private and academic sectors, addressing experimental control and performance evaluation for gene expression analysis. ERCC is considering the utility of universal (platform-independent) spike-in controls, protocols, and informatics tools intended for use across one- and two-channel microarray and quantitative RT-PCR (QRT-PCR). Outcomes of this work will be published and resulting data submitted to a public database.

Regulatory-driven fora

To streamline regulatory electronic submissions a number of technical issues need to be addressed. These efforts intend to identify the kind of data that should be included in submissions to regulatory bodies and automate the largely paper-based clinical trials and non-clinical research processes.

Clinical Data Interchange Standards Consortium (CDISC)

CDISC (**http://www.cdisc.org**) is an open, multidisciplinary, non-profit organization committed to the development of worldwide pharmaceutical industry standards, vendor-neutral, platform-independent data models to support the electronic acquisition, exchange, and the submission and archiving of clinical trials data and metadata.

Standard for Exchange of Non-clinical Data (SEND)

SEND (**http://www.cdisc.org/models/send/v1.5**) is a consortium formed among the pharmaceutical industry, contract laboratories, software developers and the FDA. The goal of SEND is to develop a common format for the electronic submission of animal toxicity data and study description to a regulatory agency. Once the SEND standard is finalized, it will be merged with CDISC's model to form the Study Data Tabulation Model (SDTM).

Pharmacogenomics (PGx) Standards Group

The Pharmacogenomics (PGx) Standards Group was formed in November 2003 at a workshop organized by the Drug Information Association (DIA), FDA, Pharmacogenetics Working Group (PWG), Pharmaceutical Research and Manufacturers of America (PhRMA) and Biotechnology Industry. Organization (BIO) to review the FDA draft, 'Guidance for Industry — Pharmacogenomic Data Submissions'. The PGx Standards Group encompasses regulatory bodies, pharma, and industry organizations. The goal of this joint project is to help define the requirements for pharmacogenomics submission to the FDA and define data formats and standards. This project focuses on the use of pharmacogenomics and toxicogenomics data to support pharmacological and toxicological conclusions. There is a consensus within this group to use existing standards (e.g. MIAME, MAGE, SEND, CDISC) if available, and to extend them if needed.

Domain-driven fora

These toxicoinfomatics and ecoinformatics specific initiatives are an example of international coordination for the development and adoption of controlled vocabularies and format for exchanging chemical toxicity, and ecological and environmental data.

The Distributed Structure-Searchable Toxicity (DSSTox)

DSSTox (**http://www.epa.gov/nheerl/dsstox**) is a network project by the US EPA, providing a community forum for publishing standard format, structure-annotated chemical toxicity data files for open public access. Although a primary focus of this effort is aimed towards inclusion of chemical structures and standardized chemical fields, DSSTox will also promote the use of a controlled vocabulary, i.e. common data field names

and entry formats for the same types of toxicity data across databases. It will link to such public toxicity data by incorporating DSSTox Standard Fields and Indices in the custom databases, making common queries possible using a standard DSSTox identifier. DSSTox is collaborating with, or using standards from, several other efforts, including the LeadScope In Silico Tox (LIST) Focus Group, the National Cancer Institute (NCI), NIEHS's National Center for Toxicogenomics and the National Toxicology Program, the National Library of Medicine (NLM) TOXNET, the International Union of Pure and Applied Chemistry (IUPAC), the National Institutes of Standards and Technology (NIST), the ILSI HESI SAR Toxicity Database Project and MGED's MIAME/Tox, as well as numerous vendors and consortia (**http://www.epa.gov/nheerl/dsstox/CoordinatingPublicEfforts.html**).

The Science Environment for Ecological Knowledge (SEEK)

SEEK (**http://seek.ecoinformatics.org**) is a multidisciplinary initiative designed to create cyberinfrastructure for ecological, environmental and biodiversity research and to educate the ecological community about eco-informatics. SEEK participants are building an integrated data grid (EcoGrid) for accessing a wide variety of ecological and biodiversity data and analytical tools (Kepler; **http://kepler-project.org**). Ecological Metadata Language (EML) is a metadata specification developed in association with SEEK and the Knowledge Network for Biocomplexity (KNB; **http://knb.ecoinformatics.org**) that can by used in a modular and extensible manner to document ecological data.

World-wide organizations

Global organizations have initiated a dialogue between technological experts, regulators and the principal validation bodies to draw road maps for development, validation and regulatory use of omics-based technologies in chemical assessment. Others are liaising with different life sciences disciplines, offering support, mediation and consultancy to speed up the standards development process.

Organization for Economic Co-operation and Development (OECD) and the International Program on Chemical Safety (IPCS)

IPCS (**http://www.who.int/ipcs/en/**) is a joint program of three cooperating organizations — the International Labour Organization, the United Nations Environment Network and the World Health Organization — implementing activities related to chemical safety. In collaboration with the Organization for Economic Cooperation and Development (OECD, **http://www.oecd.org**), the IPCS has organized a series of workshops to identify the possible application of methods based on (eco)toxicogenomics in regulatory hazard assessment, to determine the current limitations to the use of (eco)toxicogenomics in regulatory assessment and develop a plan to overcome such limitations, to identify the need for future activities with regard to the use of these methods in test guidelines, new and existing chemicals, pesticides and biocides programs. At present, recommendations are being prepared and will be published. In view of these recommendations, the development of a coordinated international research program on (eco)toxicogenomics will be initiated, aiming to optimize the integration of genomic techniques into (eco)toxicology and their use in ecological and human health risk assessment.

The National Academy of Sciences (NAS)

The NAS Committee on Emerging Issues and Data on Environmental Contaminants (**http://dels.nas.edu/emergingissues**) is a public forum for communication among government, industry, environmental groups and the academic community about emerging evidence and issues in toxicogenomics, environmental toxicology, risk assessment and exposure assessment. The Committee will develop a framework for how the emerging field of genomics will be incorporated into risk assessment.

Institute of Electrical and Electronics Engineers (IEEE) Computer Society

The Bioinformatics Standards Committee (BSC; **http://www.csbcon.org**) has a mission to act as a liaison between groups in the bioscience community, developing standards for biological objects

in the life sciences disciplines and the IEEE Standards Association. BSC will provide a neutral forum for the global bioinformatics community to work towards common agreements on standards in new areas and integration between established standards.

Standard(s)-compliant infrastructure

This section provides a short review of public infrastructure currently available for toxicogenomics and environmental genomics data. These efforts are in different stages of development, serving specific needs of their user community and relying on diverse types of funding support. Nevertheless, these are examples of institutions working together, sharing expertise and moving towards an internationally compatible informatics platform for data exchange, interacting closely with standardization initiatives listed here.

ArrayExpress and Tox-MIAMExpress

ArrayExpress (**http://www.ebi.ac.uk/arrayexpress**) (Brazma *et al.*, 2003) is a MGED standards-compliant, public infrastructure for microarray-based gene expression data at the EBI. The infrastructure has been extended to link biological endpoint values with gene expression data as result of a collaborative undertaking with the ILSI HESI Committee on the Application of Toxicogenomics Data to Mechanism-based Risk Assessment (**http://www.ebi.ac.uk/microarray/Projects/tox-nutri**). Their toxicogenomics datasets (Pennie *et al.*, 2004) have been submitted to ArrayExpress using Tox-MIAMExpress, the online MIAME/Tox-compliant data input tool (Mattes *et al.*, 2004) (**http://www.ebi.ac.uk/tox-miamexpress**). The ILSI HESI Committee research programme has provided the first large array-based toxicogenomics dataset in the public domain annotated according to the MGED standards.

Chemical Effects in Biological Systems (CEBS) Knowledgebase

CEBS (**http://cebs.niehs.nih.gov**) (Waters *et al.*, 2003) is a public toxicogenomics knowledgebase in year two of its 10 year development at the NIEHS's NCT. CEBS aims to integrate omics datasets in the context of toxicology to advance knowledge discovery about toxicity (Waters *et al.*, 2003; Waters and Fostel, 2004; Mattes *et al.*, 2004). CEBS implements standards developed by the MGED Society and the HUPO PSI in the CEBS SysBio object model (Xirasagar *et al.*, 2004). CEBS is designing an ontological representation of data and terms used by its collaborators, which includes descriptors for different study design types and metadata vocabularies.

maxd

maxd (**http://bioinf.man.ac.uk/microarray/maxd**) is an open-source data warehouse and visualization environment for genomic expression data employed by the NERC EGTDC. The maxd software suite includes two major components. The first, maxdLoad2, is a database schema and data loading and curation application designed to enable biologists to store expression data, annotate it to MIAME and MIAME/Env standards, and export it in MAGE-ML format to ArrayExpress. The second, maxdView, is a modular analysis and visualization environment for interactive exploration of transcriptomics data and associated metadata.

Toxicoinformatics Integrated System (TIS)

ArrayTrack (**http://www.fda.gov/nctr/science/centers/toxicoinformatics/ArrayTrack**; Tong *et al.* 2003) is an integrated software system for managing, mining and visualizing microarray gene expression data at NCTR-FDA. The system has three integrated components: a MIAME-compliant database storing array-based toxicogenomics data; a set of tools providing data visualization and analysis capability; and a library containing functional information about genes, proteins, pathways and toxicants. ArrayTrack is the first module of TIS, a system to integrate genomic, proteomic and metabonomic data with data from the public repositories, as well as conventional *in vitro* and *in vivo* toxicology data. TIS will serve as a general toxicogenomics repository for diverse data sources, supporting broad data mining and meta-analysis activities, as well as the development of robust and validated predictive toxicology systems.