



Sixth Framework Programme for Quality of Life and Management of
Living Resources

Project no. LSHG-CT-2006-037686

EMERALD

Empowering the Microarray-Based European
Research Area to Take a Lead in Development
and Exploitation

EU Deliverable: 3.8 and 3.9

Due Date: September 2008
Delivery Date: November 2008

Partner responsible: EBI/NTNU

Protocol of the
EMERALD Workshop
at the MGED 11 meeting
4 September 2008
by Wolfgang Huber

Microarray and Gene Expression Data society (MGED) is the main world wide organization working with development of microarray technology, standardization, quality control and data analysis of microarray data. Annually MGED organizes an international meeting gathering the most experienced researchers from the whole world. This year MGED11 took place in Riva Del Garda, Trenton, Italy (1-4 September 2008). Approximately 200 people attended the conference.

According to our project plan, and also taking into account the advise of the proposal reviewers EMERALD should team up with MGED, schedule meetings and promote the results of our work to the European microarray society. By participating on this meeting we were able to get in close contact with MGED by informing the organization about our project effort and establishing an official connection.

The meeting covered seven different topics with keynote, plenary lectures, shorter talks selected from submitted abstracts and six workshops, where one of the workshops was organized by EMERALD (Wolfgang Huber). The title of the workshop was: EMERALD Workshop on Array Quality Assessment Methods. A summary of the workshop can be found below.

In addition to the workshop presentations we presented the project and disseminated results by a poster (see attachment 1) where we specifically presented some results from WP1 focusing on quality metrics and the development of additional MGED ontology (people responsible: Wolfgang Huber, Audrey Kauffmann, Helen Parkinson and James Malone all EBI). We also distributed a leaflet describing the EMERALD project, including all contact information for EMERALD (see attachment 2).

Array data quality assessment - where are we, what next?

The workshop discussions were organised around four presentations (summarized below) whose content was discussed among the ca. 30 participants during and after each presentation, as well as in a summary session in the end.

The motivating questions for the workshop were:

- What do people want to do with quality metrics?
- Different use cases imply different approaches to implementing and assessing quality metrics. How can we categorize them, and how can we reach consensus?
- Are the current experimental practices good enough for that or how can they be improved?
- Are the current data analytical approaches good enough?
- What software works, what doesn't? What more do we need?
- What is the role of large data repositories such as GEO or ArrayExpress?

Presentation 1: Marc Salit (NIST)

ERCC: You've got 100 spike in controls - what now?

The External RNA Control Consortium (ERCC) has been developing a set of standard controls to be used in gene expression assays. These poly-adenylated controls are designed to mimic eukaryotic

mRNAs, and are intended to be 'spiked-in' to a total RNA sample and carried through an assay. The ERCC has gone through 4 rounds of microarray testing on 5 different platforms to select a set of controls that perform reasonably well in most conditions. The presentation covered the ideas behind the testing protocols, the development of the library of controls as a formal reference material, and a model use scenario from the NIST's ultimate round of testing. This use scenario provided performance information on sensitivity, linearity, dynamic range, probe effect, and the ability to detect differential expression.

Presentation 2: Matthew McCall (Johns Hopkins School of Public Health), joint work with R. Irizarry

From a single CEL - quality metrics computed on individual arrays?

Some of the most widely used preprocessing algorithms for Affymetrix gene chip depend on the joint behaviour of the data from multiple chips, typically a complete dataset. Methods that proceed chip by chip are typically inferior in terms of accuracy and precision, because they are unable to estimate (sequence-specific) probe effects. Similarly, one of the most useful array-level quality metrics for Affymetrix data, the NUSE, is computed from the joint behaviour of the data from multiple chips. However, for some applications it would be beneficial to be able to perform preprocessing and quality assessment on a chip-by-chip basis, without reference to a complete dataset.

The presentation introduced the proposal of using a large database of chips to estimate probe effects beforehand, which allows performing single chip preprocessing and quality control. The authors have created such a database for the Affymetrix HGU133a platform and are currently using this to compare NUSE with their new single chip method GNUSE. Also, they applied different quality metrics to their database and observed a very strong lab effect.

Presentation 3: Alvis Brazma (EMBL-EBI)

Quality metrics and ArrayExpress

How should microarray quality metrics be applied to data in public repositories? Dr. Brazma discussed how ArrayExpress deals (or rather not deals) with data quality issues in ArrayExpress now, and what their plans are for the future. He discussed what kind of quality metrics they need, how to apply them and what decisions to make. Finally, he discussed how will this change with the next generation sequencing based functional genomics data.

Presentation 4: Wolfgang Huber (EMBL-EBI)

Which way is up?

The need for quality metrics is uncontroversial. However, their choice, interpretation and application raises questions. Dr. Huber reviewed what manufacturing and industry engineers understand under the term "quality" and how this applies to the fields of microarrays and microarray datasets. He gave use-cases for quality metrics, categorized existing quality metrics approaches according to the type of information that they provide, and discussed their applications and limitations. In particular, he discussed the visualisation plots and quantities computed by the Bioconductor package arrayQualityMetrics (whose development is funded by Emerald), and reported on progress in the automation of the detection of 'bad' (i) arrays and (ii) experiments. He presented an example from a large, systematic survey of the datasets in ArrayExpress, where the omission of an array that was detected to be of doubtful quality by this approach indeed led to improved power for the detection of differentially expressed genes from the dataset.



EMERALD



Enhancing microarray data quality

The EMERALD consortium*

Project objectives

The European Union FP6 Coordination Action (CA) EMERALD, aims to establish and disseminate quality metrics (QM), microarray standards and best laboratory practices throughout the European microarray community. This will allow microarray data production to take full advantage of QA/QC, thereby significantly enhancing the quality of microarray data and setting a precedent for other array-based technologies. Data quality and meta data (documentation) are key to all microarray data generation and analysis, to ensure that maximum information can be extracted from the data. The need to reanalyse and reproduce data spawned a 'grassroots movement', now the MGED Society that established guidelines for experiment description (MIAME) and a structured data exchange model (MAGE-ML). MGED initiatives have predominantly been focused on data context, and its scope has only recently been extended to included data content. Quality and integrity of microarray data compendia (e.g. in ArrayExpress) are major determinants for information and extraction model building. High quality data will constitute one of the pillars of systems biology. This CA is designed to structure and amalgamate ongoing efforts across the Europe community, in close association with MGED and the ERCC.

Coordination and dissemination activities

Coordination activities are defined in six main areas relevant for microarray analysis: Development of quality metrics, ontology for data description, implementation of standards and best practices, selection of standards that are candidates for European Reference Materials, impact of QA/QC on data information content, and dissemination of QA/QC principles to novel experimental high-throughput techniques for the different -omics domains. These activities are made up of six work packages (WP).

WP1: Quality Metrics and Ontologies (EBI). The objective of this WP is to develop and disseminate quality metrics and tools for determining data quality and communicating data transformations. As part of the MGED ontology, a normalisation and transformation ontology (NTO) is being developed to describe data transformations (Figure 1). Recent information about the ontology work can be found at our web page: http://www.microarray-quality.org/ontology_work.html. We are also developing a new Bioconductor package, named arrayQualityMetrics, that provides a HTML report with diagnostic plots for one or dual color microarray data (Figure 2-4). The quality report contains the evaluation of different categories of quality metrics. The individual array quality is controlled by M versus A plots. The existence of spatial effects is checked by image representations of the arrays. Scatter plots are used to assess the reproducibility of the experiments. Boxplots and density plots allows the control of the homogeneity between the experiments. The report also contains a study of the GC content effects and the mapping of the reporters to test the array platform quality. A heatmap representing the distance between the samples allows the evaluation of the biological signal to noise ratio. In the case of Affymetrix experiments, some quality controls usually used for this platform are added to the report, such as Relative Log Expression (RLE) or Normalized Unscaled Standard Error (NUSE) plots for instance. This report can be used as a first step of the microarray analysis or to compare the efficiency of different methods of normalization. The quality metrics report will also be useful to assess the quality of public data in the context of meta-analysis for instance. Recent information can be found at our web page: http://www.microarray-quality.org/quality_metrics.html. **People responsible: Wolfgang Huber, Audrey Kauffmann, Helen Parkinson and James Malone (EBI).**

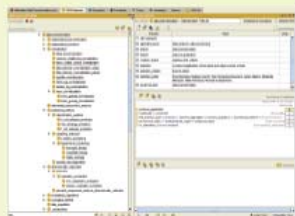


Figure 1. A Normalisation and Transformation Ontology (NTO).

As part of the MGED ontology, a normalisation and transformation ontology is being developed to describe data transformations. The ontology will cover aspects of microarray data such as normalisation techniques, quality metrics and quality control and data transformation. The development of this ontology will employ several strategies that will be the subject of workshop group discussion, and it will include analysis of current vocabularies and text mining of relevant literature.

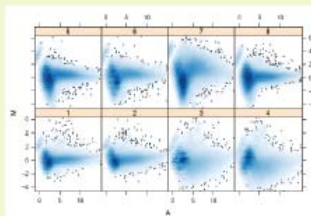


Figure 2. Represents MA plot for each array.

MA-plots are useful for pairwise comparisons between arrays. Rather than comparing each array to every other array, here we compare each array to a single median 'pseudo'-array. Typically, we expect the mass of the distribution in an MA-plot to be concentrated along the $M = 0$ axis, and there should be no trend in the mean of M as a function of A . Note that a bigger width of the plot of the M -distribution at the lower end of the A scale does not necessarily imply that the variance of the M -distribution is larger at the lower end of the A scale: the visual impression might simply be caused by the fact that there is more data at the lower end of the A scale. To visualize whether there is a trend in the variance of M as a function of A , consider plotting M versus $\text{rank}(A)$.

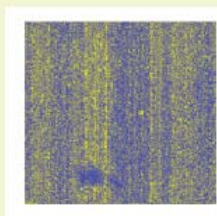


Figure 3. Intensity representation on the array (spatial plots).

False color representations of the spatial intensity distributions of each arrays. These graphical representation permit to show problems during the experimentation such as fingerprints, artifactual gradient or dye specific failure for instance.

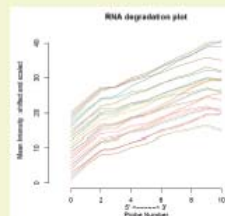


Figure 4. RNA digestion plot.

In this plot each array is represented by a single line. It is important to identify any arrays that has a slope which is very different from the others. The indication is that the RNA used for that array has potentially been handled quite differently from the other arrays.

WP2: Standards (LGC). The objective of this work package is to plan and advocate the use of standards by the microarray community. This will involve the identification of suitable reference materials (spikes, reference RNAs), the assessment of analytical "best practice" guidelines and standardised approaches to experimental design and execution.

WP3: Organisation and dissemination (NTNU). The purpose of WP3 is to organise and structure the community 'pull'. First, we will identify and bring together the key players in the field of transcriptome microarray use and further development. We will disseminate the results of WP1 and WP2 to the community through a series of workshops. Updated information will be available through our web page: www.microarray-quality.org.

WP4: Data Quality and Systems Biology (VIB). WP4 will assess the impact of QM-based filtering and general QA/QC implementation on the performance of various mining and modelling approaches of such data compendia.

WP5: Standards and European legislation (IRMM). The purpose of WP5 is to take the QA/QC criteria analysed, developed and discussed in the previous 4 work packages and translate these into commutability criteria for microarray-relevant reference materials. These criteria will form the basis for independent projects, aimed at developing and distributing European reference materials.

WP6: New Technologies (UU). A survey of new applications and development efforts in microarray technologies will be performed, in order to identify key academic and commercial players (research groups, users, product and service providers).

*Project partners

Martin Kuiper - VIB, Belgium and NTNU, Norway.
Arne K. Sandvik - NTNU, Norway.
Alvis Brazma - EBI, United Kingdom.
Carole Foy - LGC, United Kingdom.

Joaquin Dopazo - CIPF, Spain.
Laszlo Puskas - HAS, Hungary.
Heinz Schimmel - IRMM, Belgium.
Ulf Landegren - UU, Sweden.

If you are interested to participate, or have information relevant to this project, please contact:

Project coordinator:
Martin Kuiper (kuiper@nt.ntnu.no)

Project fellows:
Vidar Beisvåg (vidar.beisvag@ntnu.no)
Ewa Sugajska (ewsug@psb.ugent.be)

Funding

EMERALD is funded by the Sixth Framework Programme of the Quality of Life and Management of Living Resources. Project no. LSHG-CT-2006-037689. Scientific officer: Christina Kyriakopoulou (ec.europa.eu)

www.microarray-quality.org

Attachment 2. EMERALD leaflet distributed at MGED11.

Project management

The project is managed by a project board which has representatives of the eight partners:

Martia Kulper	Flanders Institute for Biotechnology, VIB, Gent, Belgium and Norwegian University of Science and Technology, NTNU, Norway.
Aree K. Sandvik	Norwegian University of Science and Technology, NTNU, Norway.
Alvis Brazma	European Bioinformatics Institute, EBI, United Kingdom.
Carole Fay	LGC, United Kingdom.
Joaquín Dopazo	Centro de Investigación Príncipe Felipe, Spain.
Laszlo Puskas	Biological Research Center of the Hungarian Academy of Sciences, Hungary.
Helax Schinmel	Institute for Reference Materials and Measurements, Belgium.
Ulf Landegren	Uppsala University, Sweden.

The project management is assisted by a scientific advisory board:

Frank Holsteg	Utrecht University, Netherlands.
Helen Causton	Imperial College London, United Kingdom.
Rafael Irizarry	Johns Hopkins University, United States.
Joerg Hohenseil	German Cancer Research Center, DKFZ, Germany.
Astrid Lægreid	Norwegian University of Science and Technology, Norway.
Marc Salt	National Institute of Standards and Technology, NIST, United States.



www.microarray-quality.org

EMERALD

A European Project funded by the Sixth Framework Programme for the Quality of Life and Management of Living Resources
Project no. LSHG-CT-2006-037689
Scientific officer: Christina Kyriakopoulou (@ec.europa.eu)




Contact:


Project coordinator:
Martin Kulper
Norwegian University of Science and Technology (NTNU),
Department of Biology, RealFagbygget, NTNU, 7491 Trondheim, Norway.
Email: kulper@nt.ntnu.no
Phone +47 73550348
Fax +47 73596100

Project fellows:
Vidar Betsvåg
Dept. of Cancer Research and Molecular Medicine,
Norwegian University of Science and Technology,
Medisinsk Teknisk Forskningscenter
Olav Kyrres gt. 9,
7489 Trondheim, Norway
Email: vidar.betsvag@ntnu.no
Phone +47 73598615
Fax +47 72576400
and
Ewa Sugajska
VIB Department of Plant Systems Biology,
UGent-VIB Research Building F5VM,
Technologiepark 927,
BE-9052 GENT, Belgium
Email: ewusug@psb.ugent.be
Phone +32 93313823
Fax +32 93313809

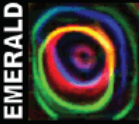
www.microarray-quality.org



A European Project funded by the Sixth Framework Programme for the Quality of Life and Management of Living Resources



EMERALD



European Project on Standards and Standardisation of Microarray Technology and Data Analysis

www.microarray-quality.org

Project objectives

This European Union Framework Program 6 Coordination Action (CA) will serve to establish and disseminate quality metrics (QM), microarray standards and best laboratory practices throughout the European microarray community. This will allow microarray data production governed by QA/QC, significantly enhancing the quality of microarray data and setting a precedent for other array-based technologies. Over the last 15 years microarray technology has proved the method of choice for capturing molecular biological data in a massively parallel fashion. Data quality and meta-data (documentation) are key to all microarray data generation and analysis, to ensure that maximum information can be extracted from the data. Very early in the development of microarray-based transcript profiling the microarray community has realised the importance of structured documentation accompanying microarray

www.microarray-quality.org

data. The need to reanalyse and reproduce data spawned a 'grassroots movement', now the MGED Society that established guidelines for experiment description (MIAME) and a structured data exchange model (MAGE-ML). MGED initiatives have predominantly been focused on data context, and has only recently been extended to included data content. Quality and integrity of microarray data compendia (e.g. in ArrayExpress) are major determinants for information extraction model building and high quality data will be one of the pillars of systems biology. This CA is designed to structure and amalgamate ongoing efforts across Europe, in close association with MGED and the ERCC.

Coordination and dissemination activities

Coordination activities are defined in six main areas relevant for microarray analysis: Development of quality metrics; ontology for data description; implementation of standards and best practices; selection of standards that are candidates for European Reference Materials; impact on data information content; and dissemination of QA/QC principles to novel experimental high-throughput techniques for the different -omics domains. These activities are made up of six work packages (WP).

Work packages

WP1: Quality Metrics and Ontologies (EBI). The objective of this WP is to develop and disseminate quality metrics and tools for determining data quality and communicating data transformations. An Ontology for describing microarray experiments and Normalization and Transformation is now under development (http://www.microarray-quality.org/ontology_work.html). And recently a new Bioconductor package named `arrayQualityMetrics` (<http://bioconductor.org/packages/2.1/bioc/html/arrayQualityMetrics.html>) is released, that provides a HTML report with diagnostic plots for one or dual color microarray data. The quality report contains the evaluation of the individual array quality, the existence of spatial effects, the reproducibility of the experiments, the homogeneity between the experiments, the GC content effects, the mapping of the reporters, the evaluation of the biological signal to noise ratio. This report can be used as a first step of the microarray analysis or to compare the efficiency of different methods of normalization.

WP2: Standards (LGC). The objective of this work package is to plan and advocate the use of standards by the microarray community. This will involve the identification of suitable reference materials (spikes, reference RNAs), the assessment of analytical 'best practice' guidelines and standardised approaches to experimental design and execution.

WP3: Organisation and dissemination (NTNU). The purpose of WP3 is to organise and structure the community 'pull'. First, we will identify and bring together the key players in the field of transcriptome microarray use. We will disseminate the results of WP1 and WP2 to the community through a series of workshops.

WP4: Data Quality and Systems Biology (VIB). WP4 will assess the impact of QM-based filtering and general QA/QC implementation on the performance of various mining and modelling approaches of such data compendia.

WP5: Standards and European legislation (IRMM). The purpose of WP5 is to take the QA/QC criteria analysed, developed and discussed in the previous 4 work packages and translate these into community criteria for microarray-relevant reference materials. These criteria will form the basis for independent projects, aimed at developing and distributing European reference materials.

WP6: New Technologies (UIU). A survey of new applications and development efforts in microarray technologies will be performed, in order to identify key academic and commercial players (research groups, users, product and service providers).

EMERALD workshops

WS4: Launch of EMERALD arrayQualityMetrics system (in collaboration with MGED11, 1-5 September 2008, Riva Del Garda Trentino, Italy).

WS5: Microarrays and clinical applications (in collaboration with the 3rd ESF Conference on Functional Genomics and Disease, 1-4 October 2008, Innsbruck, Austria).

WS6: Data quality and Systems Biology (in collaboration with the 4th EMBO Conference: From Functional Genomics to Systems Biology, 15-18 November 2008, Heidelberg, Germany).

WS7: Data quality Control and Transformation workshop (in collaboration with the 8th International conference for the Critical Assessment of Microarray Data Analysis CAMDA, 4-6, December, Vienna, Austria, 2008).

WS8: Ontology workshop (November 2008, EBI, Hinxton UK).

WS9: Towards federal standards (planned Spring 2009).

WS10: Implications for new technologies (planned Spring 2009).

WS11: Dissemination of results to larger community (planned Autumn 2009).

Web pages relevant for the project

EMERALD (www.microarray-quality.org)
Microarray Gene Expression Data (MGED) Society (www.mged.org)
National Institute of Standards and Technology (NIST) (www.nist.gov/)
External RNA Control Consortium (ERCC) (www.cstl.nist.gov/biotech/Cell&TissueMeasurements/GeneExpression/ERCC.htm)
MicroArray Quality Control (MAQC) project (www.fda.gov/nct/science/centers/toxicoinformatics/maqc/)