



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: D3.13

Due Date: 18<sup>th</sup>-20<sup>th</sup> of May 2009

Delivery Date: 17<sup>th</sup> June 2009

Version 1

Partner responsible: UU/NTNU

## Minutes from EMERALD session at the 5<sup>th</sup> Annual Advances in Microarray Technology (AMT) conference in Stockholm 18<sup>th</sup> of May 2009

Our EMERALD session was held prior to the AMT meeting (19-20 May 2009) (<http://www.selectbiosciences.com/conferences/AMT2009/>) together with other pre conference sessions.

About 300 people worldwide attended the conference and about 25 attended the EMERALD workshop. The aim of the EMERALD workshop was to discuss and illustrate the importance of (microarray) data quality. Presentations highlighted the efforts on the implementation of quality metrics and standards to new high throughput array technologies.

The specific agenda for the EMERALD session looked like this:

Session chaired by: Joakim Lundeberg and Vidar Beisvåg

- 13:30 - 14:00 Quality measures for epigenetic measurements and protein expression profiling.  
*Jörg Hoheisel, Deutsches Krebsforschungszentrum, Germany*
- 14:00 - 14:30 Proximity ligation and other tools for standardized high-throughput analysis of biomolecules  
*Ulf Landegren, University of Uppsala, Sweden*
- 14:30 - 15:00 Suspension bead arrays and challenges towards multiplexed plasma profiling  
*Jochen Schwenk, KTH Royal Institute of Technology, Sweden*
- 15:00 - 15:30 Break
- 15:30 - 16:00 Minimum Information about a high-throughput SeQuencing Experiment – MINSEQE  
*Alvis Brazma, European Bioinformatics Institute, UK*
- 16:00 – 16.30 Quality Metrics and New Technologies  
*Audrey Kauffmann, European Bioinformatics Institute, UK*
- 16:30 - 17:00 NIST SRM 2374: A Certified Reference Material Designed to Support Confidence in Gene Expression Measurements  
*Marc Salit, NIST, US*

### Summary of the talks:

First talk was held by Jörg Hoheisel from the Deutsches Krebsforschungszentrum, Germany, and the title was Quality measures for epigenetic measurements and protein expression profiling.

Jörg started out by explaining the most recent advances in the different microarray technologies. This included regular gene expression microarrays and microRNA profiling. Jörg then went into details for epigenetic and protein expression profiling. Epigenetic profiling is currently available at several microarray platforms and which seem to work well. At the moment the main challenge with epigenetic profiling data seems to be the normalization. Today, epigenetic profiling has especially been used in cancer research. Further, Jörg talked about new development of protein expression arrays. This included the development they have done on transcription factor protein binding arrays based on self-complementary stem-loop arrangement and complex antibody arrays. For the transcription factor binding arrays, mismatch probes are included on the array, which is evaluated by a disassociation behavior due to temperature changes. For the antibody array, containing 825 different antibodies, Jörg went through the quality control step included in their protocol. This included position marker proteins at the array, protein labeling by Cypro Rybu and Goat anti mouse (or rabbit) IgG labeling. For this kind of analysis kinetics and mass transport may be a problem and if plasma is used as a protein source, depletion or not may be an issue.

Second talk was held by Ulf Landegren from the University of Uppsala, Sweden, and the title for his talk was EMERALD and New Diagnostics. Ulf started the talk by describing the need for a

database with updated protocols. He then presented MolMeth ([www.molmeth.org](http://www.molmeth.org)) an online database for protocols related to molecular biology, developed by his group. Ulf then switched to talk about multiplex molecular analysis (including standard microarray arrays, suspension microarrays, RealTime PCR microarrays and Next Generation sequencing). For all these methods improvement of probing methods are important. Ulf further talked about a new method they have developed for rewriting molecular information as a string of DNA. This method is based on proximity ligation assays, which shows a 100 fold greater sensitivity than regular Elisa assays. The advantages of paired tag arrays was then explained, and compared to regular arrays this technology showed very high signal to background levels compared to regular hybridization. Ulf also mentioned the advantages of situ PLA methods for visualizing of proteins, modifications and interactions.

The third talk was held by Jochen Schwenk from the KTH Royal Institute of Technology, Sweden. The title of his talk was: Suspension bead arrays and challenges towards multiplexed plasma profiling.

Antibody microarrays offer a powerful tool to screen for target proteins in complex samples and Jochen described an approach for systematic analysis of serum, based on antibodies and using color-coded beads for the creation of antibody arrays in suspension. This method, adapted from planar antibody arrays, offers a fast, flexible, and multiplexed procedure to screen larger numbers of serum samples, and no purification steps are required to remove excess labeling substance. The assay system detected proteins down to lower picomolar levels with dynamic ranges over 3 orders of magnitude. The feasibility of this workflow was shown in a study with more than 200 clinical serum samples tested for 20 serum proteins.

The fourth talk was held by Alvis Brazma from European Bioinformatics Institute, UK and the title was Minimum Information about a high-throughput Sequencing Experiment – MINSEQE.

Alvis talked about how MIAME (Minimal Information of Microarray Experiments) was developed and the reason for its success. He continued to explain how their experience with establishing MIAME was used during the development of MINSEQE, which is an equivalent to MIAME but related to high throughput sequencing technology instead of microarray technology. The main factors involve recommendations for how description of the biomaterials, protocols and data should be reported in a structured order, related to the relatively new high throughput sequencing technologies available. Important issues are also how the data are possessed and what data that are stored in public repositories like Array Express. The first draft of MINSEQE was released in 2008 at the MGED web page (<http://www.mged.org/minsege/>) and these days a paper is in preparation, to inform the scientific community about this recommendations and this will hopefully be published early autumn 2009.

The fifth speaker was Audrey Kauffmann from the European Bioinformatics Institute, UK and the title of the talk was Quality Metrics and New Technologies.

Audrey presented the EMERALD Quality diagnostics program developed at EBI, that provides HTML reports with diagnostic plots for one and two color arrays. The report contains the evaluation of different categories of quality metrics to cover the identification of numerous types of problems, both per slides and between slides. The individual array quality, the existence of spatial effects, the reproducibility, the homogeneity between experiments and the biological signal to noise ratio are evaluated. A new feature for outlier detection was described. This useful function is now added to the newest version of the program. The program can be downloaded through this page: <http://bioconductor.org/packages/2.1/bioc/html/arrayQualityMetrics.html>. Audrey also mentioned that this software has been used on several “new” microarray platforms like exon arrays, miRNA, Chip-chip arrays, protein arrays, CGH arrays, SNP arrays and Mass spec data, in addition to the conventional expression arrays.

The sixth and last speaker was Marc Salit from NIST, US and the title of his talk was NIST SRM 2374: A Certified Reference Material Designed to Support Confidence in Gene Expression Measurements.

Marc presented the work done by his group at the National Institute of Standards and Technology (NIST) and The External RNA Control Consortium (ERCC) who have been developing a set of standard controls (96) 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. They have gone through 5 rounds of microarray testing on 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. The library of clones will be commercial available during summer 2009 and mixes of RNA may be available a bit later. These external standards are also hypothesized to be important in evaluation of new technologies like high throughput sequencing analysis.

### **Additional dissemination**

In addition to the workshop 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 the leaflet describing the EMERALD project, including all contact information for EMERALD (see attachment 2).

# Attachment 1: EMERALD poster presented at 5<sup>th</sup> AMT Conference, 2009.



## 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 (IMAGE-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).

### Normalisation and Transformation ontology (NTO)

The diversity in microarray experiment designs and applications requires that a large number of pre-processing approaches are available. In order to facilitate unambiguous and consistent descriptions of experimental data transformation the development of a 'normalisation and transformation ontology' (NTO) has been undertaken. This provides a means to conceptualize and classify the approaches used, describe relationships between these concepts and store these in a machine readable form. Such a representation can offer a useful checking mechanism to ensure that data is correctly modelled as well as a more powerful querying mechanism. The NTO has been developed as part of the Ontology for Biomedical Investigations (OBI), a large, multi-national, collaborative community development project. A beta version of the ontology is now available from: [http://obi-ontology.org/page/Main\\_Page](http://obi-ontology.org/page/Main_Page).  
People responsible: Helen Parkinson and James Malone (EBI).

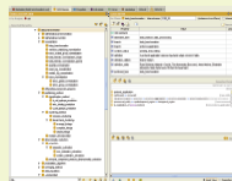


Figure 1. Example of the NTO Normalisation and Transformation Ontology.

Array Name	Median	Quality Metric	Significance	Array ID
1	0.0000000000000000	0.0000000000000000	0.0000000000000000	1
2	0.0000000000000000	0.0000000000000000	0.0000000000000000	2
3	0.0000000000000000	0.0000000000000000	0.0000000000000000	3
4	0.0000000000000000	0.0000000000000000	0.0000000000000000	4
5	0.0000000000000000	0.0000000000000000	0.0000000000000000	5
6	0.0000000000000000	0.0000000000000000	0.0000000000000000	6
7	0.0000000000000000	0.0000000000000000	0.0000000000000000	7
8	0.0000000000000000	0.0000000000000000	0.0000000000000000	8
9	0.0000000000000000	0.0000000000000000	0.0000000000000000	9
10	0.0000000000000000	0.0000000000000000	0.0000000000000000	10
11	0.0000000000000000	0.0000000000000000	0.0000000000000000	11
12	0.0000000000000000	0.0000000000000000	0.0000000000000000	12
13	0.0000000000000000	0.0000000000000000	0.0000000000000000	13
14	0.0000000000000000	0.0000000000000000	0.0000000000000000	14
15	0.0000000000000000	0.0000000000000000	0.0000000000000000	15
16	0.0000000000000000	0.0000000000000000	0.0000000000000000	16
17	0.0000000000000000	0.0000000000000000	0.0000000000000000	17
18	0.0000000000000000	0.0000000000000000	0.0000000000000000	18
19	0.0000000000000000	0.0000000000000000	0.0000000000000000	19
20	0.0000000000000000	0.0000000000000000	0.0000000000000000	20

Figure 2. Summary report.

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. Shows a summary report of arrays identified as having a potential problem or as being an outlier.

Figure 3. Represents MA plot for each array. M and A are defined as:  $M = \log_2(I1) - \log_2(I2)$   $A = 1/2 (\log_2(I1) + \log_2(I2))$  where I1 is the intensity of the array studied and I2 is the intensity of a "pseudo"-array, which have the median values of all the arrays. 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 rank(A).

Figure 4. Shows a false color heatmap of between arrays distances, computed as the median absolute difference of the M-value for each pair of arrays. This plot can serve to detect outlier arrays. Arrays whose distance matrix entries are way different give cause for suspicion. The dendrogram on this plot also can serve to check if, without any probe filtering, the arrays cluster accordingly to a biological meaning.

Figure 5. Is a Normalized Unscaled Standard Error (NUSE) plot. Low quality arrays are those that are substantially elevated or more spread out, relative to the other arrays. NUSE values are not comparable across data sets. Both RLE and NUSE are performed on preprocessed data (background correction and quantile normalization).

### arrayQualityMetrics

The assessment of data quality is a major concern in any microarray analysis. The Bioconductor package arrayQualityMetrics provides a report with diagnostic plots for one or two colour microarray data. The quality metrics assess individual array quality, homogeneity, signal to noise ratio, and it identifies apparent outlier arrays. The tool handles most current microarray technologies and is amenable to use in automated analysis pipelines or for automatic report generation, as well as for use by individuals. Removing outlier arrays from the data set before performing the analysis reduces the noise, and can increase the statistical power and lead to a more accurate biological understanding of the studied system. Recent information can be found at our web page: [http://www.microarray-quality.org/quality\\_metrics.html](http://www.microarray-quality.org/quality_metrics.html).  
People responsible: Wolfgang Huber, Audrey Kauffmann (EBI).

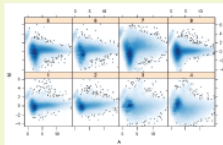


Figure 3. Individual array quality plot.

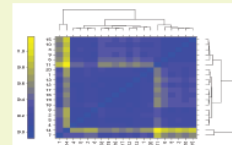


Figure 4. Between array comparison.

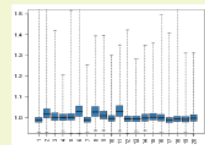


Figure 5. Normalized Unscaled Standard Error plot.

### \*Project partners

Martin Kulper - NTNU, Norway.  
Arne K. Sandvik - NTNU, Norway.  
Arvis Brazma - EBI, United Kingdom.  
Carole Foy - IGC, United Kingdom.

Joaquin Dopazo - CIBERSpain.  
Laszlo Paskas - 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 Kulper (mkulper@ntnu.no)  
Project fellow: Vidar Belting (vidar.belting@ntnu.no)

### Funding

EMERALD is 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 (0ec.europe.eu)

[www.microarray-quality.org](http://www.microarray-quality.org)

## Attachment 2: EMERALD leaflet distributed at AMT 2009.

### Project management

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

<b>Martin Kølper</b>	Norwegian University of Science and Technology, NTNU, Norway.
<b>Arne K. Sandvik</b>	Norwegian University of Science and Technology, NTNU, Norway.
<b>Alvis Brazma</b>	European Bioinformatics Institute, EBI, United Kingdom.
<b>Carole Fey</b>	LEG, United Kingdom.
<b>Joaquín Dopazo</b>	Centro de Investigación Príncipe Felipe, Spain.
<b>László Puskas</b>	Biological Research Center of the Hungarian Academy of Sciences, Hungary.
<b>Helix Schirrmel</b>	Institute for Reference Materials and Measurements, Belgium.
<b>Ulf Landegren</b>	Uppsala University Sweden.

The project management is assisted by a scientific advisory board:

<b>Frank Holtege</b>	Utrecht University Netherlands.
<b>Helen Causton</b>	Imperial College London, United Kingdom.
<b>Rafael Irizarry</b>	Johns Hopkins University United States.
<b>Joerg Hohensee</b>	German Cancer Research Center, DKFZ, Germany.
<b>Astrid Lægreid</b>	Norwegian University of Science and Technology, Norway.
<b>Marc Salit</b>	National Institute of Standards and Technology, NIST, United States.



[www.microarray-quality.org](http://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 Kølper  
Norwegian University of Science and Technology (NTNU),  
Department of Biology, Realfagbygget, NTNU, 7491 Trondheim, Norway.  
Email: kolver@nt.ntnu.no  
Phone +47 73550348  
Fax +47 73596100

**Project fellow:**  
Vidar Belsvåg  
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.belsvag@ntnu.no  
Phone +47 73598615  
Fax +47 72576400



[www.microarray-quality.org](http://www.microarray-quality.org)



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

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



[www.microarray-quality.org](http://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](http://www.microarray-quality.org)

### A Tool for Quality Assessment

We are developing a new Bioconductor package, named *arrayQualityMetrics*, 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, and 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 normalisation. The most recent version, available this autumn, will provide an overview table added, identify arrays identified as having a potential problem or as being an outlier. People responsible are Audrey Kaufmann and Wolfgang Huber at EBI, Hinxton, UK.  
More information about the *arrayQualityMetrics* can be found at our web page: [www.microarray-quality.org](http://www.microarray-quality.org) or at the Bioconductor web page: <http://bioconductor.org/packages/2.1/bioc/html/arrayQualityMetrics.html>

### Sign up for new issues of the EMERALD Newsletter at our web page:

[www.microarray-quality.org](http://www.microarray-quality.org)

### Normalisation and Transformation ontology (NTO)

The diversity in microarray experiment designs and applications requires that a large number of pre-processing approaches are available. In order to facilitate unambiguous and consistent descriptions of experimental data transformation the development of a 'normalisation and transformation ontology' (NTO) has been undertaken. This provides a means to conceptualize and classify the approaches used, describe relationships between these concepts and store these in a machine readable form. Such a representation can offer a useful checking mechanism to ensure that data is correctly modelled as well as a more powerful querying mechanism. The NTO has been developed as part of the Ontology for Biomedical Investigations (OBI), a large, multi-national, collaborative community development project. A Beta version of the ontology is now available from [http://obi-ontology.org/page/Main\\_Page](http://obi-ontology.org/page/Main_Page).  
People responsible: Helen Parkinson and James Malone (EBI).

### New on the web page

#### Discussion forum

Through our webpage (<http://www.microarray-quality.org>) a new discussion forum, related to the aims of the project is now open. Categories are related to quality metrics, ontology development, external standards, best laboratory practices and new technologies. The forum can be found at this address: <http://www.microarray-quality.org/phpBB3/index.php>

#### The Molecular Methods Database (MolMeth)

MolMeth is a structured database that provides free access to methods used in molecular biology and molecular medicine, and it allows the user to print user-friendly manuals. A unique accession number is assigned by the database which permanently identifies the protocol submitted. Each method includes a short description of the method and a list of the required reagents and equipment. The methods also include a detailed step-by-step protocol that the user can download as a pdf print or view on the screen. Submitted methods and contributions are subject to manual curation.  
One key aspect of the database is that new methods can be created that are combinations of methods already available in MolMeth. Thereby, methods available in the database can be used in different combinations. We also link reagents to the suppliers and aim to link to databases with chemical and structural information (e.g. PubChem). MolMeth methods that have been published in scientific journals will also be cross-linked to PubMed.  
The Molecular Methods Database (MolMeth) is currently being developed to provide the research community with a reliable source of methods and protocols used in molecular biology and molecular medicine. The beta version of the database can be found through this web page: <http://www.molmeth.org>.

### Web pages relevant for the project

**EMERALD** ([www.microarray-quality.org](http://www.microarray-quality.org))  
**Microarray Gene Expression Data (MGED) Society** ([www.mged.org](http://www.mged.org))  
**National Institute of Standards and Technology (NIST)** ([www.nist.gov/](http://www.nist.gov/))  
**External RNA Control Consortium (ERCC)** ([www.cst.nist.gov/biotech/Cell8TissueMeasurements/GenExpression/ERCC.html](http://www.cst.nist.gov/biotech/Cell8TissueMeasurements/GenExpression/ERCC.html))  
**MicroArray Quality Control (MAQC) project** ([www.fda.gov/oc/ct/science/centers/toxicoinformatics/maqc/](http://www.fda.gov/oc/ct/science/centers/toxicoinformatics/maqc/))

### 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).