



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: 15th to 18th of November 2008

Delivery Date: 20th November 2008

Version 1

Partner responsible: NTNU

Minutes from 4th EMBO Conference “From Functional Genomics to Systems Biology” conference, October 15-18, 2008 (Heidelberg, Germany) and EMERALD workshop at the conference.

The agenda for the meeting “From Functional Genomics to Systems Biology” included six sessions in addition to the EMERALD session. Two poster sessions were also on the program.

Session I: Genome control I (Saturday 15th)

Session II: Genome Control II (Sunday 16th)

Session III: Variation, phenotype and disease (Sunday 16th)

Session IV: Protein function and regulatory networks I (Monday 17th)

Session V: Protein function and regulatory networks I (Monday 17th)

Session VI: Systems (Tuesday 18th)

Session VII: EMERALD Workshop: Microarray data quality and systems biology (Tuesday 18th).

About 300 people (mainly from Europe) attended the conference and about 60 attended the EMERALD workshop. The aim of the EMERALD workshop was to highlight efforts to increase the quality of microarray data (by introducing guidelines and quality metrics systems) which is an important factor when using such data in systems biology approaches.

The specific agenda for our session was:

13.00 – 13.45 Keynote lecture. Exploring Networks and Stochastics in Genomic Data.
John Quackenbush, Dana-Farber, Harvard, US.

13.45 – 14.00 The EMERALD project.
Martin Kuiper, NTNU, Norway.

14.00 – 14.30 Standards Development: Necessarily a Two-Way Street
Chris F Taylor, EBI, UK.

14.30 – 15.00 Break

15.00 – 15.45 Data quality aspects in meta-analysis of condition-specificity of gene expression.
Misha Kapushesky, EBI, UK.

15.45 – 16.15 Quality Metrics: use cases and applications of quality metrics.
Wolfgang Huber, EBI, UK.

16.15 – 16.45 Outlier detection and application to the ArrayExpress database.
Audrey Kauffmann, EBI, UK.

Summary of the talks:

Talk I

The first speaker was **John Quakenbush** from Dana-Farber Cancer Institute at Harvard (US). John's talk was titled "Exploring Networks and Stochastics in Genomic Data", and he started by stating that quality control is a very important step in microarray technology and essential to generate good data for systems biology approaches.

John talked about what is meant by "Networks" in system biology and how these can be derived from microarray data. He gave some examples of how his group has used Bayesian networks approaches where they include meta data (from e.g. The literature) to model biological networks. Pro's and con's regarding literature mining were discussed. He also discussed the uses of microarray data alone vs. with literature data, concluding that by adding literature data the results were increased. Further he presented some work related to network refinement through perturbation, where they knock down genes by siRNA and uses RT-PCR to assay gene expression of other genes in a specific pathway. They use Bayesian Network analysis to derive network from qRT-PCR data. Primary conclusions of this work were that perturbations of predicted networks using expression as an output, filtered through a BN analysis can lead to deduction of network properties. Multiple perturbations begin to converge on meaningful predictive networks. Analysis using Literature Priors in this case improves over a no-Prior approach. There is a need to add additional perturbations and complete bootstrapping analysis.

An other subject that was described was mesoscopic biology. The basic idea is that gene expression is variable in individual cells. The most likely distribution for gene expression levels is a Poisson distribution. From this it can be calculated how the Poisson distribution will cause variance in measurements to grow as we sample fewer cells. This can again be tested experimentally in the lab. This part of the talk was concluded with: Gene expression is variable in individual cells and the Poisson distribution fits the data extremely well. This is the first time mesoscopic measurements have been used in biology to validate microscopic models. This direct evidence of stochastic influence on biology provides important input into development of network models. These models are likely to be more like predicting the weather than predicting the orbits of the planets and will require intense computational simulation combined with analytical modeling.

Finally John talked about State Space Models of Gene Expression where cells are modeled as complex systems and they try constructing a trajectory in gene expression space by using time series data. It was showed how they hypothesizes that cells that transition from one state to another are influenced by the combination of two types of processes, core and transient groups.

The final conclusions were: There is still a role for biology! We are approaching a time in which we can begin to look at cells and organisms holistically. We also need to begin to think about integrating diverse data types in an intelligent way. This must include cross-species comparisons and inclusion of environmental effects. We may soon be in a position to begin development of a theoretical biology. Theoretical biology will require a transition from a Deterministic to a Stochastic approach.

Talk II

The second speaker was **Martin Kuiper** (coordinator EMERALD project) from NTN,U who gave a presentation of the EMERALD project. The aims of the project were presented. An update of the dissemination activities was given. Attention was given to our webpage, where we have now added a discussion forum and where interested

people can sign up for an EMERALD newsletter. In addition a web survey that is available through our web pages was highlighted, with a question to the audience to return their input.

Talk III

The third speaker was **Chris F Taylor** from the European Bioinformatics Institute (EBI, Hinxton, UK). The title of his talk was: Standards Development: Necessarily a Two-Way Street.

Chris talked about some of the mechanisms behind scientific advancements; how different standards, like MIAME have evolved and how the graphical user interface must be user-friendly to get biologist to use tools developed by informaticians. Standards for functional genomics were discussed by dividing them in three aspects including, biology, generic features and technology.

It was further discussed what the MIBBI project is aiming for (minimal information for biological and biomedical investigation). Finally he discussed the motivation of providing data in a standardized way so it can be shared and understood by others.

Talk IV

The fourth speaker was Misha Kapushesky from the European Bioinformatics Institute (EBI, Hinxton, UK). The title of his talk was: Data quality aspects in meta-analysis of condition-specificity of gene expression.

Misha started by describing the ArrayExpress repository of microarray experiments and the ArrayExpress Atlas of gene expression databases. Altogether there are now (in 2008) almost 8000 experiments submitted to the database, and the increases have been exponential in recent years. The same trend has been seen for the number of hybridizations which now is almost 193000. Another general characteristic is that 43% of the data is from human samples and that Affymetrix datasets represent 29% of the data. Most of the data today are expression data, but data from other technologies like CGH, Chip-ChIP, microRNA, genotyping and protein arrays are also present. The question is then, can we uncover and describe this context-specificity from available high throughput gene expression data? The database works regardless of platform/technology and allows to express basic differential expression questions.

Micha then presented a meta study where they had looked into what is common to all 6300+ experiments in the ArrayExpress Repository. The goal was to look at this collection of measurements of gene expression across a variety of conditions, to associate every gene to one or several conditions where it shows strong differential expression. They also looked for if independent gene expression studies support each other. In this way they were able to generate a gene expression atlas metrics. Forty matrices for different condition types, including disease state, organism part, phenotype, compound treatment, developmental stage, etc. was developed. Most genes show no significant differential expression. The ArrayExpress Atlas is now accessible through a web portal as an online resource.

The conclusions of the talks were: Despite relatively poor resolution of microarray data, we can increase statistical power of analysis by aggregating data from multiple independent studies. ArrayExpress Atlas is a resource for discovering context-specific gene expression patterns and a research platform for evaluating hypotheses against the corpus of publicly available transcriptomic data. Initial insights into the complexity of condition-specificity of expression of transcriptionally regulated pathways may be gained by robust meta-analysis of genomic data.

Talk V

The fifth speaker was Wolfgang Huber from the European Bioinformatics Institute (EBI, Hinxton, UK). Wolfgang's talk was titled: Quality Metrics: use cases and applications of quality metrics.

Wolfgang presented the EMERALD Quality diagnostics program developed, 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. 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> .

Talk VI

The sixth and last talk at the EMERALD workshop was given by Audrey Kauffmann from the European Bioinformatics Institute (EBI, Hinxton, UK). The title of her talk was: Outlier detection and application to the ArrayExpress database.

Audrey described the Array Quality Metrics software in more detail than Wolfgang with specific attention on how the outlier detection is done. Audrey then presented some work that was done on the data sets available in ArrayExpress database. They have been looking for relationship between outlier detection and when the data have been added to Array Express, how many arrays there were in a typical experiment, which species the experiments have been performed on and if there was a relation to the impact factor of the journal where the data have been published. They found no linear improvement with time and impact factor, but for species they found that some species were not as good as human, mouse and rat. The reason for this may be that the clone collection for these species is not as good as the better annotated genomes of humans, rat and mice. She, however, concluded that more analysis needs to be done to draw any conclusions. Further work may include analysis where platforms, different protocols, are compared, effect of tissue types and correlations to biological factors are measured.

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 a newsletter and the leaflet describing the EMERALD project, including all contact information for EMERALD (see attachment 2).

Attachment 1: EMERALD poster presented at 4th EMBO Conference “From Functional Genomics to Systems Biology”, 2008.



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 the 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.
People responsible: Helen Parkinson and James Malone (EBI).

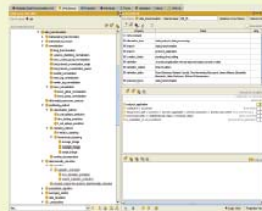


Figure 1. Example of the A Normalisation and Transformation Ontology.

ArrayID	ArrayName	Weight	Synthetic	Platform	Design	MA	MS
1	18_423584-01.F1001.C1.C						
2	18_423584-01.F1001.C1.C						
3	18_423584-01.F1001.C1.C						
4	18_423584-01.F1001.C1.C						
5	18_423584-01.F1001.C1.C						
6	18_423584-01.F1001.C1.C						
7	18_423584-01.F1001.C1.C						
8	18_423584-01.F1001.C1.C						
9	18_423584-01.F1001.C1.C						
10	18_423584-01.F1001.C1.C						
11	18_423584-01.F1001.C1.C						
12	18_423584-01.F1001.C1.C						
13	18_423584-01.F1001.C1.C						
14	18_423584-01.F1001.C1.C						
15	18_423584-01.F1001.C1.C						
16	18_423584-01.F1001.C1.C						
17	18_423584-01.F1001.C1.C						
18	18_423584-01.F1001.C1.C						
19	18_423584-01.F1001.C1.C						
20	18_423584-01.F1001.C1.C						

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).

arrayQuality Metrics

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.
People responsible: Wolfgang Huber, Audrey Kauffmann (EBI).

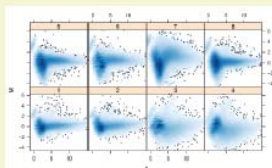


Figure 3. Individual array quality plot.

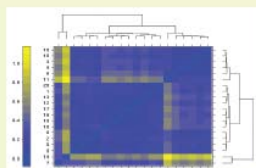


Figure 4. Between array comparison.

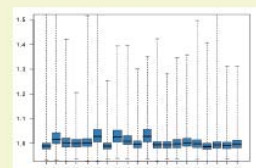


Figure 5. Normalized Unscaled Standard Error plot.

*Project partners

Martin 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@psugent.be)

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 (@ec.europa.eu)

www.microarray-quality.org

Attachment 2: EMERALD leaflet distributed at Proust "Gene at Work at Time", 2008.

Project management

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

Martje Kulper	Flanders Institute for Biotechnology, VIB, Gent, Belgium and Norwegian University of Science and Technology, NTNU, Norway.
Arne K. Sandvik	Norwegian University of Science and Technology, NTNU, Norway.
Alvis Brazma	European Bioinformatics Institute, EBI, United Kingdom.
Carole Fey	LEG, United Kingdom.
Joaquín Dopazo	Centro de Investigación Príncipe Felipe, Spain.
László Puskas	Biological Research Center of the Hungarian Academy of Sciences, Hungary.
Helga Schramel	Institute for Reference Materials and Measurements, Belgium.
Ulf Landegren	Uppsala University Sweden.

The project management is assisted by a scientific advisory board:

Frank Holtege	Utrecht University Netherlands.
Helen Causton	Imperial College London, United Kingdom.
Barbel Irizarry	Johns Hopkins University, United States.
Joerg Hohenseil	German Cancer Research Center, DKFZ, Germany.
Astrid Lagnrad	Norwegian University of Science and Technology, Norway.
Marc Salit	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:
Martje 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 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
and
Ewa Sugajska
VIB Department of Plant Systems Biology,
UGent-VIB Research Building FSVM,
Technologiepark 927,
BE-9052 GENT, Belgium
Email: ewusug@psb.ugent.be
Phone +32 93313823
Fax +32 93313809

www.microarray-quality.org



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

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 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).

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, identifying arrays identified as having a potential problem or as being an outlier.

People responsible are Audrey Kauffmann and Wolfgang Huber at EBI, Hinxton, UK.

More information about the *arrayQualityMetrics* can be found at our web page: 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

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.

People responsible: Helen Parkinson and James Malone (EBI).

Participate in EMERALD QC web survey

The goal of this survey is to gain insight into procedures, platforms, and needs of the microarray users community. The focus is both on commercial GeneChip arrays and all sources of cDNA and oligonucleotide microarrays. The survey is geared to gather information anonymously from academic, pharmaceutical, and commercial laboratories, which use microarray technologies routinely. The survey is now open and can be found on our web page: www.microarray-quality.org or directly at: <http://kvas.itea.ntnu.no/eva/login.do?externalId=2021-14551-abbr>. The results of the survey will be made freely available to the microarray community through our web page, following analysis of the data. We appreciate your participation in this study.

EMERALD workshops

- WS7:** Data quality and Systems Biology (in collaboration with the 4th EMBO Conference: From Functional Genomics to Systems Biology, 15-18 November 2006, Heidelberg, Germany).
- WS8:** 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, 2006).
- WS9:** Towards federal standards (planned Spring 2008).
- 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.cst.nist.gov/biotech/CellTissueMeasurements/GeneExpression/ERCC.htm)
Microarray Quality Control (MAQC) project (www.ebi.gov/ncf/science/centers/toxicoinformatics/maqc/)