



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: ???

Due Date: 5th of December 2008

Delivery Date: 20th November 2008

Version 1

Partner responsible: NTNU

Minutes from CAMDA08 Conference, December 4-6, 2008 (Vienna, Austria).

CAMDA is a conference that offers researchers from computer science, statistics, molecular biology, and other areas an opportunity to benefit from the critical evaluation of various techniques in microarray data analyses. In advance of the conference a data set was released and the participants were asked to analyze this and submit a report to be presented at the workshop. Selected abstract were chosen for shorter oral presentations or poster presentations. This year two data sets were offered. The primary, general CAMDA dataset was offered by the laboratory of Prof. Christine and collaborators. They made available raw and processed data from a small microarray gene expression time-course experiment that is typical of gene expression time-course data sets yet provides an unusual opportunity for pushing the performance of analysis methods. The second dataset was the 'EMERALD' dataset provided by Marc Salit and colleagues from NIST (National Institute of Standards and Technology) (Liggett *et al.*, 2008). This study and dataset aimed to make a microarray experiment to study the relative magnitudes of technical and biological variation.

The EMERALD workshop/session took place on the 5th of December and it was the major contribution to the conference. The aim of the EMERALD workshop at CAMDA was to put focus on and discuss issues related to preprocessing and quality of microarray data. About 40-50 people (mainly from Europe) attended the conference.

The specific agenda for the EMERALD session was:

Friday, 5 Dec – Emerald sessions

- 09.00 – Introduction to the EMERALD dataset, Ron Peterson, Novartis Institute of Biological Research, Cambridge, Massachusetts, U.S.A.
- 09.30 – Research, Cambridge, Massachusetts, U.S.A.
- 09:30 – Keynote, 'Muddling or modelling your way through normalization?', Ernst Wit, University of Groningen, The Netherlands.
- 10:15
- 10.35 – 'Metrology for Gene Expression: Measurement Batch Effects, Probe Sensitivity, Gene-List Reproducibility', Walter Liggett, NIST, Gaithersburg, Maryland, U.S.A.
- 11.20
- 14:45 – 'EMERALD microarray platform comparison based on hypothesis tests under order restrictions', Florian Klingmüller and Thomas Tuechler, Department of Statistics and Probability Theory, University of Technology Vienna, Austria.
- 15:15
- 15:45 – 'Exploiting the EMERALD mixture design for model based microarray platform comparisons by Bayesian inference of technical and biological variance components'. Thomas Tuechler, et al, Boku University Vienna, Austria.
- 16:15
- 16:15 – 'Progress on transformation and normalization ontology', James Malone, European Bioinformatics Institute (EMBL-EBI), Cambridge, U.K.
- 16:45
- 16.45 – Panel discussion
- 17.15

Summary of the talks:

First talk: Introduction to EMERALD dataset (invited speaker)

An introduction to the dataset was given by Ron Peterson, Novartis.

There is a growing understanding of the sources of variability in microarray experiments, and ways to control that variability are propagating. In part because the technical variability observed in contemporary microarray experiments has become better controlled, statistically significant lab-to-lab and batch-to-batch effects have been observed. A number of experiments which study the same samples across a variety of laboratories and platforms have reported this. The essential question is whether these effects are significant with respect to the biological variability observed amongst the samples. This question lies at the heart of establishing the fitness for purpose of microarrays for biological studies.

Data was produced by *three different laboratories measuring the same samples on three different platforms* – each with their own *batch factors* (Liggett *et al.*, 2008). The platforms were the *Affymetrix Rat Genome 230 2.0* array, the *Illumina RatRef-12* array, and the *Agilent Whole Rat Genome* array. The samples were a *titration mixture* of RNA isolated from kidney and liver, from 6 different normal control rats from an earlier experiment at Novartis. This titration presents a series of 4 samples from each rat: RNA from the kidney, a mixture of 75% RNA from kidney and 25% from liver, a mixture of 25% RNA from kidney and 75% from liver, and RNA from the liver. These samples were measured in replicate, for each animal. Pooled samples from the various animals were also measured, for a nominal 96 arrays from each platform.

The relationship amongst these samples enables model-based analysis, amongst other approaches. Model-based approaches can be compelling because they permit observation and apportionment of variation in the residuals. The titration samples present interesting opportunities for alternative analyses as well, with the titration fraction as a surrogate or proxy for RNA concentration. A particular interest for this CAMDA dataset was its use for evaluating the performance of different preprocessing approaches and techniques.

Second talk: Keynote talk (invited speaker)

The keynote talk was held by Ernst Wit, University of Groningen: Muddling or modelling your way through normalization?

Ernst talked about two attitudes to “normalization”: the Computer Scientist’s Attitude: Muddling a preprocessing activity, whereby data are cleaned before further analysis; and the Statistician’s Attitude: Modelling, a joint modelling activity, whereby analysis and accounting for nuisance effects are combined.

Examples of a computer scientist approach include normalizing all local features first, then progress to normalizations that involve several and, finally, all arrays, were shown.

What are the drawbacks of “muddling”? False believe that the normalized data are clean (and typically no way of checking whether this is true). The uncertainty inherent in the normalization is not carried forward to the analysis: results can be too liberal. Most preprocessing methods can’t deal with additional structure in the data.

As an alternative Ernst proposed a statistical model, in order to check the validity of their normalization model. A model carries the uncertainty in the normalization over to inference and is able to deal with the peculiar structure of the EMERALD dataset.

Ernst asked: What are the essential features of the EMERALD data? Comparison of interest: 2 tissue types: kidney and liver. Measured in 0/1, 0.25/0.75, 0.75/0.25, 1/0 mixtures. Each repeated 3 times (per rat, per platform), plus some additional pools. 3 different laboratories each with their own platform. 6 normal rats, repeatedly used in each lab. 96 arrays in each platform. Therefore, the platform is confounded with laboratory. Low replication number: only 6 degrees of freedom for comparing kidney/liver across thousands of genes; deal with lots of technical replication. Mixtures are introduced, which need to be modelled.

It was also asked: What are the nuisance (but relevant) features of the EMERALD data?

There might be spatial variation across the slides. Depending on the platform, there is information about, Fluidics station, Fluidics Machine and Scanner that was used in the experiment on each array.

We want to learn which genes behave differently in the liver and the kidney, and for this a random effect model is useful. The advantage over a usual regression model is that we require only 4 parameters instead of 40,000! We can still do inference on the basis of the random effects and it allows a more subtle normalization model. Further more data were added to the model like: hybridization artifacts, technical replication. And finally, scale and variation differences between platforms were added (probably the most challenging part).

The bad news with this approach, it that it takes several hours to process the data (approximately 500,000 data points) and to fit the model, but the good news is that the method can be run in any package with mixed model capabilities. In conclusion, the muddling approach to normalization has and will have a role to play in large datasets. Mixed effects models make it possible to replace the muddling approach by a modelling approach, which

means that quality of the inference improves. And finally, the EMERALD dataset is a fantastic dataset for the development of intra-platform methods.

Third talk: Metrology for Gene Expression: Measurement Batch Effects, Probe Sensitivity, Gene-List Reproducibility by (invited speaker) Walter Ligget

They have obtained insight into the relative size of measurement batch effects and biological variation as represented by the animal-to-animal differences. These differences provide a practical benchmark because the animals were all subject to the same control-group treatment.

Although calibration curves for individual probes are unknown, insight into calibration can be obtained from a platform-to-platform correspondence that identifies probes that measure the same transcript. This identification allows insight into the relative sensitivity of probes from different platforms.

They found that, although the animal-to-animal variation is generally larger than the measurement batch effects, their measurements do lead to the conclusion that these effects should not be ignored in experimental design and analysis. It is moreover the case that the measurement batch effects might be larger in a different experiment. Second, over the set of transcripts for which liver expression is appreciably different from kidney expression, no platform is undeniably more sensitive than another. However, the difference in probe sensitivity between two platforms varies appreciably from transcript-to-transcript. That is, one platform seems more sensitive for some transcripts, and the other platform more sensitive for other transcripts. This observation suggests considerations in the interpretation of single-platform studies. Third, they found that gene list reproducibility is likely to be worse than might be expected.

In conclusion, their investigation provides full coverage only of the probes for which liver expression differs from kidney expression. Inclusion of more animal organs would lead to better coverage of the probes.

Forth talk (selected from submitted paper): EMERALD microarray platform comparison based on hypothesis tests under order restrictions, by Florian Klingmüller University of Technology, Vienna.

Klingmüller asked: Do the measured intensities reflect the titration? They looked for agreement across platforms and influence of normalization. For this they used a method of tests against ordered alternatives based on isotonic regression. What they then did was to compute one sided permutation test p-values for each animal, on each platform separately with Quantile and Baseline normalized data, and combine per animal tests from each platform and finally combine per platform tests from each normalization. They then found a location-shift, and speculated if there was a higher messenger-RNA content in kidney? In addition, both normalization methods remove any visible trends in location. Further, they then found around 2 times more significant genes exclusive to baseline compared to quantile normalized data, and that more than 97% of genes exclusive to baseline normalized data are upregulated. The up-down in quantile was measured to exclusive genes 40:60 (up/down).

In summary they concluded: Substantial number of genes shows significant monotonicity. Across platform agreement exceeds chance levels. Agreement on baseline normalized data is worse. Baseline normalized data shows more upward trends. Genes exclusively significant in baseline data are mostly showing upward trends. Regarding the methods they concluded that isotonic regression provides a means to detect monotonic trends. p-Value combination as a means to compare results from different platforms.

Fifth talk (selected from submitted paper): Exploiting the EMERALD mixture design for model based microarray platform comparisons by Bayesian inference of technical and biological variance components'. Thomas Tuechler, Boku University Vienna.

Tuechler introduced and applied a fully Bayesian model for the interference of the variance component which explicitly exploits the tissue mixtures featuring in the EMERALD experiment/dataset. They observed intensity dependent differences specific to each platform

and determined that biological variance amounts to about 30% of the signal variance in the data set. They then concluded that variation between individual rats in the EMERALD data set is smaller in relation to the technical noise and that, intriguingly, the three platform's ability to detect this biological component differs remarkably with preprocessing and signal intensity.

Sixed talk (invited speaker): Progress on transformation and normalization ontology', James Malone, European Bioinformatics Institute, EMBL-EBI.

This talk had to be cancelled.

Panel discussion

The session was finished by an open discussion. The discussion focused on the possibility of advising the research community preferred ways of pre-processing microarray data. The CAMDA session basically compared two major approaches: modelling of the data, where a statistical model is used to explain all the variation (both technical, biological, and noise) in the data; and basic pre-processing where data is normalised. Obviously the pre-processing approach is one that is often readily handled by bioinformaticians, whereas the modelling approach is best done when skilled statisticians are available. The availability of these skills is therefore a major factor in deciding which approach to take. It was agreed that it would be difficult to come up with an authoritative advice, it would seem better to just point to the differences between these two approaches. It was mentioned that the four data analysis examples presented in the workshop might offer an interesting example to the users community, and the Emerald consortium should pursue a back-to-back publication of these papers with the authors of each of these papers commenting on the other approaches.

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).



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).

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).

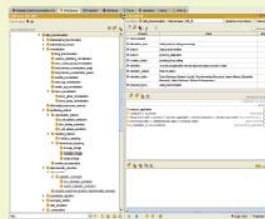


Figure 1. Example of the A Normalisation and Transformation Ontology.

Import	Array/Probe	Major	Subclassifiable	English	Biological	Missing	RLE	NUSE
1	IS:AD2201-01.FP50A1.CEL							
2	IS:AD2201-01.FP50A1.CEL							
3	IS:AD2201-01.FP50A1.CEL							
4	IS:AD2201-01.FP50A1.CEL							
5	IS:AD2201-01.FP50A1.CEL							
6	IS:AD2201-01.FP50A1.CEL							
7	IS:AD2201-01.FP50A1.CEL							
8	IS:AD2201-01.FP50A1.CEL							
9	IS:AD2201-01.FP50A1.CEL							
10	IS:AD2201-01.FP50A1.CEL							
11	IS:AD2201-01.FP50A1.CEL							
12	IS:AD2201-01.FP50A1.CEL							
13	IS:AD2201-01.FP50A1.CEL							
14	IS:AD2201-01.FP50A1.CEL							
15	IS:AD2201-01.FP50A1.CEL							
16	IS:AD2201-01.FP50A1.CEL							
17	IS:AD2201-01.FP50A1.CEL							
18	IS:AD2201-01.FP50A1.CEL							
19	IS:AD2201-01.FP50A1.CEL							
20	IS:AD2201-01.FP50A1.CEL							

Figure 2. Summary report.

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).

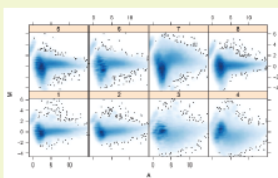


Figure 3. Individual array quality plot.

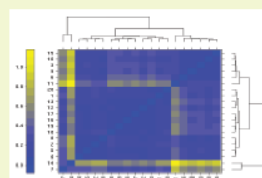


Figure 4. Between array comparison.

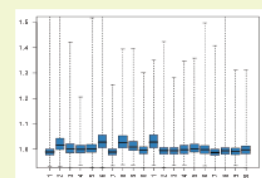


Figure 5. Normalized Unscaled Standard Error plot.

*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@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 CAMDA 2008.

Project management

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

Martin Kulpér	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.
Helax Schimmel	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 Laegreid	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 Kulpér
Norwegian University of Science and Technology (NTNU),
Department of Biology, Realfagbygget, NTNU, 7491 Trondheim, Norway.
Email: kulpert@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 Forskningsenter
Olav Kyrres gt.5,
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 F5VM,
Technologiepark 927,
BE-9002 GENT, Belgium
Email: ewasugaj@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



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

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 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>

www.microarray-quality.org

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.microarrayquality.org or directly at: <http://kvs.itea.ntnu.no/eva/login.do?textemailid=2021-14551abbr>. 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 2008, 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, 2008.

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.eia.gov/nct/science/centers/toxicoinformatics/maqc/)