



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: D3.4

Due Date: 13th to 14th of December 2007

Delivery Date: 24th January 2008

Version 1

Partner responsible: NTNU and VIB

Minutes of EMERALD associated activities at CAMDA07 13-14 December, 2007, Valencia, Spain.

1. Introduction

CAMDA (Critical Assessment of Microarray Data Analysis) is a meeting that is organized annually. CAMDA07 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. For the first time CAMDA had left its traditional location at Duke and has moved overseas to Europe. This year's venue was the Centro de Investigación Príncipe Felipe in Valencia, Spain. Approximately 60-70 people attended the conference. This year the focus was put on two datasets. One dataset, from the CDC Chronic Fatigue Syndrome Research Group, containing gene expression, proteomic, SNP, and clinical data. The hope was that this would foster integrative and biological goal-oriented analysis. Additionally another dataset composed by 6000 arrays of diseased and normal human samples and cell lines collected from ArrayExpress and GEO were discussed. This second dataset possessed challenging questions on large-scale data analysis and visualization. EMERALD contributed with a poster (see attachment 1) a leaflet (see attachment 2), and the organization of an EMERALD session.

The meeting covered different topics with invited keynote speakers, CAMDA Speaker session (where the dataset were discussed), software demos and the EMERALD session where we put a focus on quality control and data preprocessing techniques. The EMERALD session consisted of three talks by invited speakers, three shorter talks by authors of selected posters and a final plenary discussion.

Program EMERALD session (14.30 - 18.00):

14.30 - 14.40 Introduction by Martin Kuiper

14.40 - 15.10 Invited speaker: Federico Goodsaid, FDA, US.

15.10 - 15.30 Selected speaker from submitted abstracts: William Langdon, University of Essex, UK (for abstract, see attachment 3).

15.30 - 16.00 Break

16.00 - 16.20 Invited speaker: Audrey Kauffmann, EBI, UK.

16.20 - 16.40 Selected PhD student abstract from submitted abstracts: Seraya Maouche, INSERM, France (for abstract, see attachment 4).

16.40 - 17.00 Selected PhD student abstract from submitted abstracts: Richard Pearson, University of Manchester, UK (for abstract, see attachment 5).

17.00 - 17.20 Invited speaker: James Malone, EBI, UK.

17.20 - 18.00 Open discussion, lead by Paul Van Hummelen, Katholieke Universiteit Leuven, Belgium.

2. Summary of talks

1. Martin Kuiper presented the project objectives and status. He gave examples of how we will disseminate the results of the project through a series of workshops in relation to European meetings.
2. Invited speaker: Federico Goodsaid presented his talk through a WIFI webcast from US. He talked about FDA's mission with respect to microarray standards and what they do in relation to how microarray experiments and data have to be performed to be accepted by FDA and how it can be possible to transfer the results to a diagnostic platform. There is no or little consensus on protocols, algorithms, and biological interpretations of microarray gene expression data, so one of the "take home messages" was: "we need a consensus on how to generate, analyze, interpret and report microarray data" and through their work (also together with the MAQC project), they have been able to make a draft for such a document (titled: Guidance for Industry. Pharmacogenomic Data Submission - Companion Guidance), that now is out for feedback. One of the comments they got on this draft version was that, "assuming that the pharmacogenomics science continues to evolve rapidly, the utility of the guidance is jeopardized by the risk that its content will not be state-of-the-art". This led to one of the conclusions of this work: Guidance documents follow the science, not the other way around. Goodsaid also summed up their result from MAQC1 where the focus was on class comparison (DEG) and intra and inter lab reproducibility and cross-platform comparability. The major findings were that microarrays are: repeatable within a laboratory, reproducible across laboratories, concordant across platforms, comparable with alternative technologies (e.g. QPCR) and reflective of biology regardless of the differences in technology. Further MAQC-II have now started and this time the focus is on class prediction of e.g. treatment outcome, prognosis, diagnosis and personalized medicine. EMERALD has recently established a connection with MAQC which may constitute a good platform for collaboration when similar initiatives are started in Europe as a part of the EMERALD effort.
3. Selected speaker from submitted abstracts: William Langdon presented a study where they had reanalyzed 5896 human Affymetrix files from ArrayExpress and looked for spatial defects. They found that the mean error rate was low (1.6%), however some locations on the arrays were much more prone to errors than others, with up to 28% of probes being affected. They also found that the "oldest" arrays had the highest numbers of affected probes and that the most "recent" arrays were the best ones, indicating that the whole technology seems to be more stable now, maybe due to more experience and standardization of protocols(?). Finally, Langdon presented some data indicating that removal of erroneous data improved breast cancer survival prediction.
4. Invited EMERALD speaker: Audrey Kauffmann presented the EMERALD Quality diagnostics program (R, Bioconductor) 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. Currently, work to extend this program is in progress and will focus on "numerical" values that indicate quality, in addition to the plots.
The program can be downloaded through this page:
<http://bioconductor.org/packages/2.1/bioc/html/arrayQualityMetrics.html> .

5. Selected speaker from submitted abstracts: Seraya Maouche presented a study where they performed a cross-platform comparison. Microarray data from three platforms were compared (Illumina Bead Chip Human-6 V1, Affymetrix HUG133plus 2.0, and the academic RNG/MRC two-color chip). The goal for the study was to choose a platform for the Cardiogenetics project. The results showed that the inter-replicate correlation of absolute intensities was Affymetrix $r > 0.96$, Illumina $r > 0.98$ and RNG-10 $r > 0.64$. Pair-wise correlation of relative intensities showed that both Illumina and Affymetrix provided high inter-replicate reproducibility (Illumina $r > 0.84$ and Affymetrix $r > 0.84$). The RNG-10 gave less consistent results ($r > 0.59$). The results also showed, that the criteria used to select lists of DEGs strongly influenced the degree of concordance among the 3 platforms and a High level of agreement in lists DEGs was observed between Affymetrix and Illumina. The RNG/MRC was less reliable than Affymetrix and Illumina and therefore needs larger sample size to reach the performance obtained on the commercial platforms. Further they showed that combining a non-stringent P-value and a fold change (recommended by MAQC authors) is inappropriate for the Illumina data. However, despite some lack of agreement in gene lists, Gene Ontology analyses revealed that concordant biological conclusions can be drawn using the 3 platforms.

6. Selected speaker from submitted abstracts. Richard Pearson presented AffyDEComp, like Affycomp, a tool for benchmarking methods for analyzing Affymetrix data. Whereas Affycomp concentrates on expression summarization methods, the focus of AffyDEComp is on differential expression (DE) detection methods, or more precisely, on the combination of summarization and DE detection methods. AffyDEComp is currently based on the Golden Spike data set described in the paper of Choe *et al.* 2005. Pearson et al believe that, at present, this is the best publicly-available data set for comparison of Affymetrix DE detection methods. However, they also recognize that the data set used might not be representative of data sets more generally. In particular, just because a method does well here, doesn't necessarily mean the method will do well on your data sets. In the future they plan to extend AffyDEComp to other data sets.

7. Invited EMERALD speaker. James Malone was talking about the EMERALD Ontology effort. First of all, he gave an introduction to what an ontology is and how it is built. Malone then focused on how the Normalization and Transformation (NOT) ontology are built. In conclusion an NTO would give us: 1. Consistency in usage of terms through explicit definitions. 2. Widen reproducibility of microarray experiments. 3. Richer representations, again definitions, but also axioms, relationships, properties to describe the data. 3. Reduction of disparate efforts and 4. (potentially) mappings to external resources.

8. Plenary discussion, lead by Paul van Hummelen. Martin Kuiper, Audrey Kauffmann and James Malone contributed in the panel. The discussion was started by a series of questions about of standardization of microarray preprocessing by Paul van Hummelen.
 1. Is it useful?
 2. Standards or Consensus?
Will standards hinder new developments?
 3. How to insure cross-platform comparisons:
Is it possible?
Standard method or platform dependent
 4. How to perform a quality assessment?

- Quality report
- Minimum QC criteria
- Should it be included in the MIAME guidelines?
- 5. Actions needed
 - Validation data set
 - Guidance document
- 6. Future implementations
 - Standards for other microarray platforms:
aCGH, ChIP-chip, Exon, SNP

We had a lively discussion with the audience. The general consensus of the discussion was that standards would be very useful. However, because microarrays is still a new and developing technology, a “guidance” document was preferred over fixed standards. A standard may be seen as a constraint, and limiting, whereas a guidance document is a voluntary assessment of the merits of different approaches. Also, it will be to demanding to develop standards for a rapidly evolving technology and therefore also hinder new developments. The audience also agreed that some kind of quality report of the microarray data is essential to reanalyze and to re-assess the quality of the data in peer reviews. The audience also proposed to set clear minimum quality criteria. However, after a vivid discussion it was agreed that it is still too early in the development of the technology to do so. Again, a detailed QC report containing all quality metrics and a guidance as for how to interpret them would be a better idea. Such QC report should be integral part of the MIAME guidelines and MIAME data submission. What also was discussed is the need of some benchmark samples and data sets that can be used for evaluation of quality. Benchmark samples and dataset may be the best solution for evaluating the technology and data analysis. Finally, people agreed that this effort for gene expression “standardization” will be useful for similar technologies like CGH, ChIP-chip, exon and SNP arrays.

4. EMERALD travel bursary for CAMDA07

EMERALD announced three travel bursaries (for up to Euro 1500) for PhD students from the European Union member countries. Our intention was to provide support for selected students to participate. Two PhD candidates were selected for travel bursary and an oral presentation of their work. The selected candidates Seraya Maouche and Richard Pearson, both gave very good talks and contributed to the general discussion.

5. Summary

All together EMERALD was very visible at the conference and we were able to communicate our mission and results of our work to the microarray data analysis community. The general feedback was that the community appreciates our initiative and the main conclusion of the discussion was that we probably need to work for a common consensus related to these issues, and not strive for specific standard that may limit further development.



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 allow 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).**

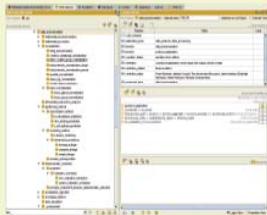


Figure 1. A Normalisation and Transformation Ontology (NTO).

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.

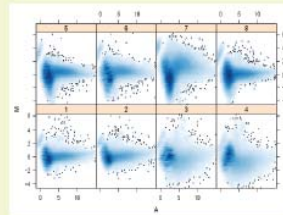


Figure 2. Represents MA plot for each array.

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 rank(A).

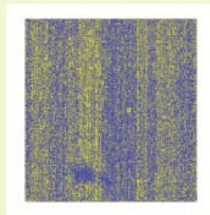


Figure 3. Intensity representation on the array (spatial plots).

False color representations of the spatial intensity distributions of each arrays. The color scale is shown in the panel on the right. The color scale was chosen proportional to the ranks. These graphical representation permit to show problems during the experimentation such as fingerprints, artifactual gradient or dye specific failure for instance.

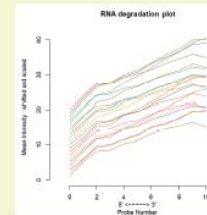


Figure 4. RNA digestion plot.

In this plot each array is represented by a single line. It is important to identify any array(s) 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 computability 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.
Arne K. Sandvik - NTNU, Norway.
Alvis Brazma - EBI, United Kingdom.
Carole Foy - LGC, United Kingdom.

Joaquín Dopazo - CIPF, Spain.
László 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 (martin.kuiper@psb.ugent.be)

Project fellows:
Vider Beisvåg (vidar.beisvag@ntnu.no)
Ewa Sugijska (ewsug@psb.ugent.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 CAMDA07, Valencia, Spain, 2007

Project management

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

Martin Kuiper	Flanders Institute for Biotechnology, VIB, Gent, Belgium.
Arae 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.
László Puskas	Biological Research Center of the Hungarian Academy of Sciences, Hungary.
Heliz Schirrmal	Institute for Reference Materials and Measurements, Belgium.
Ulf Landegren	Uppsala University Sweden.

The project management is assisted by a scientific advisory board:

Frank Holteaga	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.
Janet Warrington	Affymetrix, Inc., 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 Kuiper
VIB Department of Plant Systems Biology,
UGent-VIB Research Building FSVM,
Technologiepark 927
BE-9052 GENT, Belgium
Email: martin.kuiper@psb.ugent.be
Phone +32 93313805
Fax +32 93313809

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: vidarbelsvag@ntnu.no
Phone +47 73598615
Fax +47 72576400
or
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

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 content, and has only recently been extended to include 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 normalisation.

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 analyses 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).

EMERALD workshops

WS1: Introducing EMERALD to the community (EMERALD session in conjunction with MGED10, September 2007, Brisbane, Australia).

WS2: Ontology Workshop (5-9 November 2007, Hinxton, UK).

WS3: Data Quality Control and Transformation workshop at CAMDA07 (13-14 December 2007, Valencia, Spain).

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

WS5: Towards federal standards (planned 2008).

WS6: Data quality and Systems Biology (planned Autumn 2008 / Spring 2009).

WS7: Implications for new technologies (planned Spring 2009).

WS8: Dissemination of results to larger community (planned Autumn 2009).

www.microarray-quality.org

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/Cell&TissueMeasurements/GeneExpression/ERCC.htm)
MicroArray Quality Control (MAQC) project (www.fda.gov/oc/toxscience/centers/toxicoinformatics/maqc/)

Attachment 3. Abstract from selected submitted abstracts: William Langdon.

Spatial Defects in 5896 HG-U133A GeneChips.

W. B. Langdon WLangdon@essex.ac.uk, **R. da Silva Camargo** and **A. P. Harrison**
Departments of Mathematical Sciences and Biological Sciences,
University of Essex, CO4 3SQ, UK

Abstract

Motivation: Modern biology has moved from a science of individual measurements to a science where data are collected on an industrial scale. Foremost amongst the new tools for biochemistry are chip arrays which, in one operation, measure hundreds of thousands or even millions of DNA sequences or RNA transcripts. Whilst this is impressive, increasingly sophisticated analysis tools have been required to convert gene array data into gene expression levels. Despite the assumption that noise levels are low, since the number of measurements for an individual gene is small, identifying which signals are affected by noise is a priority.

Results: 5896 raw data (Affymetrix CEL) files were obtained from <ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/TABM/E-TABM-185/>. Each CEL file was checked for spatial errors. In HG-U133A high-density oligonucleotide array (HDONAs) the mean error rate is only 1.6% which is amongst the best for human GeneChips. However some locations are much more error prone than others, with up to 28% of probes being affected. Removal of erroneous data improves breast cancer survival prediction.

Attachment 4. Abstract from submitted abstracts: Seraya Maouche

Explaining the Sources of Discrepancies in Gene Expression Profiles Generated on three Whole-Genome Gene Expression Microarray Platforms

Maouche Seraya, INSERM UMR S525, Faculté de Médecine Pitié-Salpêtrière, 91 Boulevard de l'Hôpital, Paris, 75013, France, maouche@chups.jussieu.fr

Abstract

Previous small-scale cross-platforms comparative studies have discussed several issues of microarraybased gene expression data, including comparability between platforms, repeatability between labs, performance, and concordance to non-array based gene expression. Recently, results of the MicroArray Quality Control (MAQC) project, the first large-scale crossplatforms study conducted with the goal of establishing quality control metrics for microarray data and of assessing the reliability of gene expression profiles generated on different platforms, showed that using standardized procedures, microarray results from different platforms are reproducible. We conducted a study to compare gene expression data generated on three platforms: Illumina Bead Chip Human-6 V1, Affymetrix HGU133plus 2.0, and the academic RNG/MRC two-color chip. 10 RNA samples from human monocyte and monocyte-derived macrophage were hybridized in parallel to the 3 platforms. In addition, a list of differentially expressed genes generated using a larger number of hybridizations to the RNG/MRC platform was included in the cross-platforms comparisons and used as a reference to assess the 3 platforms.

AffyDEComp: towards a benchmark for differential expression methods

Richard D Pearson (richard.pearson@postgrad.manchester.ac.uk), **University of Manchester**

Abstract

The issue of method validation is of great importance to the microarray community; arguably more important than the development of new methods [Allison et al., 2006]. The microarray analyst is faced with a seemingly endless choice of methods, many of which give evidence to support their claims of being superior to other approaches, which at times can appear contradictory. Method validation is a difficult problem in microarray analysis because, for the vast majority of microarray data sets, we don't know what the "right answer" really is. For example, in a typical analysis of differential gene expression, we rarely know which genes are truly differentially expressed (DE) between different conditions. Perhaps the most well-known and widely used benchmark for Affymetrix analysis methods is Affycomp [Cope et al., 2004]. While a very valuable tool of summarization method validation, Affycomp is not ideal for comparison of DE methods because:

1. It uses data sets which only have a small number of DE spike-in probesets.
2. It only uses fold change (FC) as a metric for DE detection, and hence cannot be used to compare other competing DE methods. The "Golden Spike" data set of Choe et al. [2005] includes many differentially expressed spike-in probesets, making it potentially very valuable as a benchmark data set. There have, however been a number of criticisms of this data set.