



# **Cost Action 853**

## **Agricultural Biomarkers for Array Technology**

### **Working Group 3:**

### **Bioinformatics and Information Dissemination**

**September 18-19, 2006**

**Zurich, Switzerland**

**Meeting Report**

The Working Group 3 meeting was held at the University of Zürich, Switzerland. The meeting started on Monday September 18 at 13:30, and it ended on Tuesday September 19 at 17:00.

## Schedule

### Monday, September 18

**13:30** Registration and Welcome

#### Gene Expression Microarray Data Analysis

14:00 – 14:30 Introduction U. Wagner

14:30 – 15:15 J. Jäger, MPI Berlin:

Selecting normalization genes for small diagnostic microarrays.

15:15 – 16:00 S. Barkow, FGCZ:

Clustering methods in gene expression analysis

16:00 – 16:30 Coffee Break

16:30 – 17:15 P. Zimmermann, ETHZ:

Genevestigator: A microarray database and tool box

17:15 – 18:00 Z. Trajanoski, TU Graz :

Can molecular mechanisms be revealed by gene expression profiling?

18:00 – 18:30 Discussion

20:00-22:30 Dinner at the Kantorei in Zürich

## Tuesday, September 19

### Diagnostic Microarrays: Probe Design, Data Analysis and Data Management

**9:00** – 09:30 Introduction J. Frey

09:30 – 10:15 P. Peyret, U Clermont-Ferrand :  
New algorithms development for microarray probe design to identify  
microorganisms and  
characterize metabolic pathways

10:15 – 10:45 Coffee Break

10:45 – 11:30 H. Meier, TU München:  
Phylogenetic Oligonucleotide Probe Design using ARB

11:30 – 12:15 C. Gautier, U Lyon  
Interactive Microarray Data Management Systems.

12:15 – 13:00 H. Rehrauer, FGCZ:  
Analysis of a Diagnostic Microarray- The Burkholderia Phylochip.

13:00 – 14:00 Lunch

14 :00 – 14:45 H. Meier, TU München:  
Chipanalyser: Computational Analysis of Data from Microbial Diagnostic  
Microarrays.

14 :45 – 15:30 Seminar Session:  
Dr. Marusa Pompe Novak:  
Gene expression microarray data analysis in plant – pathogen  
interaction studies  
Dr. Irina S. Druzhinina  
An oligonucleotide barcode for species identification in *Trichoderma*  
the basis for DNA microarray development.

15:30 – 16:15 Coffee Break and Poster Session

16:15 – 17:00 Discussion

**17:00** Closing

## Gene Expression Microarray Data Analysis

### Introduction to Gene Expression Microarray Data Analysis

Ulrich Wagner, Senior Bioinformatic Scientist, Functional Genomics Center, Uni ETH Zurich

The main focus of the first day of the COST853 Working Group 3 Meeting was the usage of microarrays as a tool to measure gene expression and the analysis of the respective data. Microarray techniques are most often used in studies that focus on studying differential gene activity. There are many microarray platforms available that differ in respect to e.g. probe lengths, probe numbers, dyes used for RNA labeling, array surfaces etc. A recent effort (MAQC Consortium, 2006) involving many different institutions has elucidated that in spite of these varying factors the consistency of results is high. This is true for results obtained in different labs with the same platform but also with different platforms. More specifically, this also means that results obtained with one- or two-color arrays are fairly well comparable (Patterson et al., 2006), indicating that the technological aspects of microarray technique start to enter the maturing phase.

In concert with the technological developments of microarrays, the development of tools for the statistical analysis of microarray data has been rapid within the last years. But the question is whether also data analysis is maturing likewise. Efforts have been made in order to streamline workflows and standardize data management and treatment, like MIAME (Brazma et al., 2001), MGED normalization group (Ball & Brazma, 2006), CAMDA conference (Johnson & Lin, 2001) be a more or less good agreement on how a general analysis workflow for microarray data analysis should look like, however there is a multitude of opinions, which methods is the best for which step within this workflow. In the following, the most important workflow elements are shortly listed:

A successful microarray analysis can only be carried out, when the experiment is well designed. Aspects like replication, randomization, blocking, correction for dye bias, sample size and sample pooling should be considered best in agreement with the local statistical support person. After image analysis, the quality of the readouts should be checked. Among others, aspects like background distribution, RNA degradation and replicate consistency should be checked. In the next step data has to be preprocessed and properly normalized. Genes should then be excluded from further analysis based on flagging, intensity threshold, statistical tests and fold change criteria. If the experimental setup is more complex, e.g. involves several conditions, groups of genes with similar expression profiles can be determined. In combination with a priori knowledge (e.g. Gene Ontology categories or established pathways), biological processes can be detected that are involved in the transition from one to the other experimental conditions. Depending on the research goal and the available data, microarray data are being developed to be used for diagnostic purposes. Finally, a microarray experiment can only reproduced and understood by other scientist, when properly stored and annotated under standardized conditions. This high quality of data can be beneficial for higher order data mining.

Ball CA, Brazma A. MGED standards: work in progress. OMICS. 2006 Summer;10(2):138-44.

Brazma A, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet. 2001 Dec;29(4):365-71.

Johnson KF, Lin SM. Critical assessment of microarray data analysis: the 2001 challenge. Bioinformatics. 2001 Sep;17(9):857-8.

The MicroArray Quality Control The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol. 2006 Sep 8;24(9):1151-1161

Patterson TA, et al. Performance comparison of one-color and two-color platforms within the Microarray Quality Control (MAQC) project. Nat Biotechnol. 2006 Sep;24(9):1140-50.

## **Selecting normalization genes for small diagnostic microarrays**

Jochen JÄGER, HAMILTON Life Science Robotics, Bonaduz, Switzerland

Normalization of gene expression microarrays carrying thousands of genes is based on assumptions that do not hold for diagnostic microarrays carrying only few genes. Thus, applying standard microarray normalization strategies to diagnostic microarrays causes new normalization problems.

In this talk I will point out the differences of normalizing large microarrays and small diagnostic microarrays. I suggest to include additional normalization genes on the small diagnostic microarrays and propose two strategies for selecting them from genomewide microarray studies. The first is a data driven univariate selection of normalization genes. The second is multivariate and based on finding a balanced diagnostic signature. Finally, I compare both methods to standard normalization protocols known from large microarrays. The results show that not including additional genes for normalization on small microarrays leads to a loss of diagnostic information. Using house keeping genes from the literature for normalization fails to work for certain datasets.

While a data driven selection of additional normalization genes works well, the best results were obtained using a balanced signature.

## **Biclustering methods for gene expression data analysis**

S. BARKOW, FGCZ, Zürich, Switzerland

The recent advances in generating large amounts of high-throughput data in the life sciences have lead to an increasing demand for computational approaches for data analysis.

Especially data from micro array experiments are no longer analyzable without the help of algorithms. The probably most common technique for data mining in general is clustering. In recent years several approved clustering algorithms as well as newly developed techniques, such as biclustering methods, have been tested for applicability in the life sciences field. In this talk I will give an overview over the traditional and more recently developed clustering algorithms and talk about their strengths and weaknesses in the biological field. Also, I will introduce a self developed software, the Biclustering Analysis Toolbox (BicAT), that allows an easy application of a selection of algorithms to biological data.

## **Array technology for agricultural biomarkers?**

P. ZIMMERMANN, ETH, Institute of Plant Science, Zürich, Switzerland

By definition, a biomarker is a characteristic that is objectively measured and evaluated as an indicator of normal or altered biologic processes. Establishing libraries of biomarkers allows to develop technologies for sample classification or diagnostics, e.g. diagnostic arrays. But how are biomarkers identified in the first place, and how specific are they? In my talk, I will present the recently developed Genevestigator software and show that large-scale transcriptome microarray data can be used to identify biomarkers related to anatomy, development stage and responses to stimuli or mutations. These in turn can be used to develop smaller, specific arrays for various purposes.

## **Can molecular mechanisms be revealed by large-scale expression profiling?**

Zlatko TRAJANOSKI, Institute for Genomics and Bioinformatics, Graz  
University of Technology, Graz, Austria

**Background:** Large-scale transcription profiling of cell models and model organisms has been used to identify new molecular components involved in fat cell development. However, detailed characterization of the sequences of the identified gene products has not been performed and global mechanisms have not been addressed. We asked to which extent can molecular processes be revealed by expression profiling and functional annotation of genes differentially expressed during fat cell development.

**Results:** Mouse microarrays with >27.000 elements were developed and transcription profiles of 3T3-L1 cells were monitored during differentiation. 780 differentially expressed ESTs were subjected to in-depth bioinformatics analyses. The analysis of 3'UTR sequences from 395 ESTs showed that 71% of the differentially expressed genes could be regulated by miRNAs. Molecular atlas of fat cell development was then constructed by *de novo* functional annotation on a sequence segment/domain-wise basis of 659 protein sequences, and subsequent mapping onto known pathways, possible cellular roles and subcellular localizations. We found that key enzymes in 27 out of 36 investigated metabolic pathways were regulated at the transcriptional level, typically at the rate limiting steps in these pathways. We also found that co-expressed genes rarely shared consensus transcription factor binding sites, and were typically not clustered in adjacent chromosomal regions, but were instead rather widely dispersed throughout the genome.

**Conclusion:** This study shows that large-scale transcription profiling in conjunction with sophisticated bioinformatics analyses can provide not only a list of novel players in a particular setting, but also a global view on biological processes and molecular networks.

## Diagnostic Microarrays: Probe Design, Data Analysis and Data Management

### Diagnostic Microarrays – how to get there ?

JE Frey Agroscope Changins Wädenswil, Plant Protection/ Molecular Diagnostics, Wädenswil, Switzerland

Developing diagnostic microarrays seems at first sight to be a fairly straightforward and simple task. Initially, bioinformatics issues are not in the centre of interest. The first focus is on producing slides that show acceptable hybridization patterns. Issues such as surface chemistry, type of probe, probe attachment, spotting and hybridization buffer, and assay sensitivity do no more present intractable problems. Next is probe design. Again, it seems quite simple. Again, at least for low-density chips, not much bioinformatics seems to be involved beyond  $T_m$  calculations. However, soon after the first success, problems pop up that boil down to questions to which bioinformaticians most probably have answers: What is the DNA fragment best suited for the job? What do we do if only low levels of differentiation between some taxa exist? Will increasing redundancy of probes increase the robustness of the diagnosis? Will it also increase the potential for false-positives? How can false-positives be avoided? If they cannot be avoided, how do we treat them? Finally, to date there is in general only little information on the genetic make-up of the target sequence in agriculturally relevant groups of organisms. How can we in this context estimate the robustness of our diagnosis? I will give a short frame to these questions and hope for many answers.

### New algorithms development for microarray probe design to identify microorganisms and characterize metabolic pathways.

P. PEYRET Protist Biology Laboratory UMR CNRS 6023 Aubiere. France

The development of a SSU rRNA-gene based microarray approach, as high throughput tool to investigate microbial dynamics in complex environments, relies on highly efficient probes set. These biosensors have to be sensitive enough to detect all microbial communities components, even in low abundance, and highly specific to recognize targeted groups only. Moreover, these probes must be explorative in order to obtain a snapshot of microbial communities close to the reality (known and unknown parts). However, these criteria are far to be common places especially the last one. To date, the majority of oligonucleotides design softwares generate probes only targeting microorganisms from DNA sequences found in international databases. However, the majority of microorganisms remains unknown and cannot therefore be found in those databases. In order to solve this problem and to obtain efficient microarray probes, new probe design algorithms have been developed. These original approaches (algorithmically and biologically) are called PhylArray and GoArrays. To evaluate metabolic pathways in complex ecosystems we also developed the new algorithm MetabolicDesign. This software allowed us an *in silico* reconstruction of metabolic pathways and a probes design for functional microarrays that target mRNA encoding enzymes.

## **Phylogenetic Probe Design using ARB**

H. MEIER, TU München, Germany

The design of oligonucleotide probes or probe sets is an important step in the process of developing a diagnostic microarray. This presentation includes a comprehensive overview on probe design by discussing the factors influencing the quality of a designed probe or probe set. In the second part of this talk ARB is introduced. While following all the steps from sequence retrieval to probe selection the advantages of using ARB in particular for the design of phylogenetic oligonucleotide probes are pinpointed. The most important algorithms and methods are explained in detail. Finally the latest developments in ARB are depicted and future plans introduced.

## **Some bioinformatics tools for studying microbial diversity**

C. GAUTIER, PRABI & LBBE (Université Lyon 1), France

The evaluation of biodiversity cannot be limited to the enumeration of present species. Firstly such a task would be impossible particularly due to numerous unknown species, secondly because phylogenetic and functional position of species must be taken into account. We present here some bioinformatic and statistic tools that could help description of bacterial biodiversity by analysing data from genome sequencing and DNA chips.

Genomic sequences could help both for phylogenetic reconstruction and for species identification. Hogenome is a database that gives access to gene families of less than 500 members (will be increased to 1000 very soon), to their alignment and phylogenetic trees. The clustering of genes in families is made by Blast, alignment by Muscle and phylogeny is estimated by maximum likelihood (Phyml). A graphical query language is available to select family by phylogenetic tree patterns. An example is shown to select families in which an horizontal transfer may have occurred between bacteria and archaeae. Moreover, starting with a new sequence Hoseql allows determining its family and positioning it in the phylogeny of the family. Sequence of rRNA is an efficient tool to identify bacteria; BiBi is a design database and software for such determination for pathogenic bacteria to be made in hospital department.

Concerning DNA chips we emphasize on both data management and data analysis. The building of a designed database is presented. This database integrated ARB for the design of the chip and used an UML modelling. It associates data on chip design, hybridization experiment and ecological environment. Data analysis is focused on the use of approaches as principal component analysis, co-inertia analysis. Particularly this last method is efficient to associate data from DNA chips to environmental information.

## **Analysis of a Diagnostic Microarray: The Burkholderia Phylochip**

H. REHRAUER, FGCZ, Zürich, Switzerland

Members of the Genus *Burkholderia* play an important role as human pathogens, biocontrol agents, soil bacteria exhibiting different types of non-pathogenic interactions with plants (like growth promotion), bioremediation of recalcitrant xenobiotics, nitrogen fixation & are also known as plant pathogens. We present a diagnostic microarray for the detection and identification of *Burkholderia* bacteria in environmental samples. We describe how the experimental protocol has been developed and optimized. Further we give details on the data processing and validation of the chip.

## **Computational Analysis of Data from Microbial Phylochip Microarrays - ChipAnalyser**

H. MEIER, TU München, Germany

For processing image analysis data from microarray gene expression experiments a variety of tools and programs are available. This doesn't hold true for data from diagnostic microarray experiments.

In this presentation an overview is given on the topic of analysing data from diagnostic microarray experiments. The characteristics of microarray experiments for profiling microbial communities in complex samples are mentioned thereby. We discuss problems concerning the development of suitable algorithms for data analysis as well as proper microarray design. The software "ChipAnalyser" is introduced as a GUI-based environment for semiautomatic interpretation of microarray data. Already implemented functions are explained and their usability for analysing real world data is shown.

## Gene expression microarray data analysis in plant – pathogen interaction studies

M. POMPE-NOVAK, Š. BAEBLER, H. KREČIČ-STRES, A. ROTTER, K. GRUDEN, M. RAVNIKAR, National Institute of Biology, Ljubljana, Slovenia

In recent years cDNA microarrays have been employed to monitor gene expression in plant – pathogen interaction studies on a scale much larger than previously possible. Nevertheless, data analysis still remains a challenge. Several computer programs and statistical packages are being used in data analysis. ArrayPro, R, MEV and MapMan turned out to be the most useful ones for analyzing and visualizing large data sets obtained with TIGR 10000 clones potato cDNA microarrays. MapMan is a data visualization tool developed by Max-Planck-Institute of Molecular Plant Physiology (Potsdam, Germany). The use of MapMan offers the possibility to paint out microarray profiling experiments onto diagrams of metabolic pathways or processes, and to visualize the responses of gene expression in a biological context. MapMan is supported by a plant specific ontology. The principle of the MapMan ontology is a hierarchical BIN-based structure. Each BIN comprises items of similar biological function, and can be further split into subBINs corresponding to submodes of the biological function. Our recent and very important goal, in agreement with our collaborators from Max-Planck-Institute of Molecular Plant Physiology (Potsdam-Golm, Germany), is to extend this software platform to allow it to be applied to potato and grapevine plants.

## An oligonucleotide barcode for species identification in *Trichoderma*: the basis for DNA microarray development

S. DRUZHININA, A. G. KOPCHINSKI, M. KOMON, and C. P. KUBICEK, Institute of Chemical Engineering, Vienna University of Technology, Austria

One of the biggest obstructions to apply knowledge on *Hypocrea/Trichoderma* industrial importance has been the incorrect and confused application of species names to isolates used in biotechnology, biocontrol of plant pathogens and ecological surveys, thereby making the comparison of results questionable. Here we provide a convenient, on-line method for the quick molecular identification of *Hypocrea/Trichoderma* at the genus, clade and species levels based on an oligonucleotide barcode: a diagnostic combination of several oligonucleotides (hallmarks) specifically allocated within the internal transcribed spacer 1 and 2 (ITS1 and 2) sequences of rRNA repeat. The first barcode version was developed on the basis of 979 sequences of 88 genetically characterized species which displayed in total 135 ITS1 and 2 haplotypes. Oligonucleotide sequences which are constant in all known ITS1 and 2 of *Hypocrea/Trichoderma* but different in closely related fungal genera, were used to define genus specific hallmarks. Verification of the DNA-barcode was done by a blind test on 53 unknown isolates of *Trichoderma*, collected in Central and South America. The obtained results were in a total agreement with phylogenetic identification based on *tef1* (large intron), NCBI BLAST of vouchered records and *postum* morphological analysis. The library of species-, clade- and genus-specific hallmarks is stored in the MySQL database and integrated in the *TrichOKey* v. 2 - BarCode sequence identification program with the web interface located on [www.isth.info](http://www.isth.info). The original *TrichOKey* v. 1.0 identified 76 single species and 6 species pairs. The current version of the program has a module for the identification of multiple sequence and also includes newly discovered species. We conclude that oligonucleotide barcode is a powerful tool for the routine identification of *Hypocrea/Trichoderma* species and should be useful as a complement to traditional methods.

## List of Participants

| Last Name     | First Name | Country | Institute  |
|---------------|------------|---------|--|
| Barkow        | Simon      | CH      | Functional Genomics Center, Uni-ETH Zürich                       |
| Druzhinina    | Irina      | AT      | Vienna University of Technology, Vienna                          |
| Duc           | Laurence   | CH      | Functional Genomics Center, UNI-ETH Zürich, Zürich               |
| Eggenschwiler | Jenny      | CH      | University of Applied Sciences, Wädenswil                        |
| Frey          | Jürg       | CH      | Agroscope Changins-Wädenswil Research Station, Wädenswil         |
| Gautier       | Christian  | FR      | CNRS Lyon  |
| Jäger         | Jochen     | CH      | Hamilton AG, Bonaduz   |
| Kljun         | Sasa       | SI      | National Institute of Biology, Ljubljana                         |
| Meier         | Harald     | DE      | Technical University München                                     |
| Migheli       | Quirico    | IT      | University of Sassari, Sassari                                   |
| Noll          | Matthias   | CH      | Functional Genomics Center, UNI-ETH Zürich, Zürich               |
| Oggenfuss     | Markus     | CH      | Agroscope Changins-Wädenswil Research Station, Wädenswil         |
| Pasqür        | Frédériqü  | CH      | Agroscope Changins-Wädenswil Research Station, Wädenswil         |
| Pelludat      | Cosima     | CH      | Agroscope Changins-Wädenswil Research Station, Wädenswil         |
| Petrzik       | Karol      | CZ      | Institute of plant molecular biology Acad.Sci., Ceské Budějovice |
| Peyret        | Pierre     | FR      | University of Clermont Ferrand                                   |
| Pompe Novak   | Marisa     | SI      | National Institute of Biology, Ljubljana                         |
| Regier        | Nicole     | CH      | Swiss Federal Research Institute WSL, Birmensdorf                |
| Rehraür       | Hubert     | CH      | Functional Genomics Center, Uni-ETH Zürich                       |
| Salcher       | Michäla    | CH      | Institute for Plant Biology, University of Zurich                |
| Schönmann     | Sü         | CH      | Functional Genomics Center, UNI-ETH Zürich, Zürich               |
| Slaviak       | Monika     | PL      | University of Gdansk   |
| Timm          | Alicia     | SA/CH   | Agroscope Changins-Wädenswil Research Station, Wädenswil         |
| Trajanoski    | Zlatko     | AT      | University of Graz   |
| von Balthasar | Leopold    | CH      | University of Applied Sciences, Wädenswil                        |
| Wagner        | Ulrich     | CH      | Functional Genomics Center, UNI-ETH Zürich, Zürich               |
| Zeder         | Michäl     | CH      | Institute for Plant Biology, University of Zurich                |
| Zimmermann    | Philip     | CH      | ETH Zürich   |