DNA microarrays: applications and novel approaches for analysis and interpretation.

Dissertation zur Erlangung des
naturwissenschaftlichen Doktorgrades
der Bayerischen Julius-Maximilians-Universität Würzburg

vorgelegt von
Julia Cathérine Engelmann
aus Wetzlar

Würzburg, 2008

Eingereicht am: .....

Mitglieder der Promotionskommission:
Vorsitzender: Prof. Dr. Martin J. Müller
Gutachter: Prof. Dr. Thomas Dandekar
Gutachter: Prof. Dr. Sven Rahmann

Tag des Promotionskolloquiums: .....

Doktorurkunde ausgehändigt am: .....

Acknowledgments

I would like to express my gratitude to my academic advisors Prof. Thomas Dandekar and Dr. Tobias Müller for their guidance and constant support in helping me to conduct and complete this work. In addition, I want to thank Prof. Sven Rahmann for serving on my advisory committee, as well as for his excellent advice. I am thankful to IZKF B-36 and BMBF project FUNCRYPTA (FKZ 0313838B) for funding. Thanks to my collaborators who have contributed experimental work, advice, fruitful discussions and provided lab facilities to complete the projects described in this thesis. I am also grateful to *Molecular Ecology Resources* journal for the permission to print my publication "Modeling cross-hybridization on phylogenetic DNA microarrays increases the detection power of closely related species" in this thesis.

Many thanks to all the people I have come to know in the Department of Bioinformatics at the University of Würzburg, whose friendship I always enjoyed. Special thanks to the best roommates in town, Torben Friedrich, Philipp Seibel, Juilee Thakar and Stefan Pinkert. Thanks to Karin Schleinkofer, Kornelia Neveling and Frank Förster for proof-reading parts of this thesis. A big thank you goes to Kornelia Neveling for always being there when I needed a friend.

I owe everything to my family who has supported and encouraged me over the years. I especially want to thank Sven for his inspiration and continuous encouragement during my studies. Finally, I want to express my deepest appreciation to my beloved parents for their love, affection, and unlimited support during my life and studies.

Contents

Part I

General Introduction

Thesis Outline

In this thesis, applications of the microarray technology to answer biological questions as well as novel approaches for microarray data analysis are presented. Among others, microarrays can be used to provide a snapshot of the transcription level of thousands of genes simultaneously, which is a main theme of this thesis.

Part I gives an introduction into the microarray technology and analysis. In the individual chapters of part II, the main part of this thesis, the results of the experiments and analyses conducted during my PhD studies are presented. The first publication presented in Part II is concerned with the question how biodiversity studies interrogating the species composition of a certain habitat could be improved using the microarray technology. Existing microarray approaches are typically based on the evaluation of unprocessed signal intensities of the individual species spots. For very closely related species, however, cross-hybridization impedes species detection based on signal intensities alone. I present in this thesis the design of a phylogenetic DNA microarray and a novel approach for its data analysis. Using simple linear regression modeling on the signal intensities, I could show that this analysis approach greatly improves the resolution of phylogenetic DNA microarrays for species detection.

The second publication, chapter 2 of part II, addresses the question of how the large amounts of gene expression microarray datasets which have accumulated in public repositories can effectively be integrated into a coherent analysis. While several studies have integrated microarray data targeting the same biological question, approaches to integrate data from a wide range of experimental conditions are missing. In chapter 2, I present an explorative meta-analysis approach exemplified on *Arabidopsis thaliana* datasets which makes use of the large amounts of microarray gene expression data stored in public databases.

In later chapters of part II, I describe projects using microarrays for gene expression profiling to answer a defined biological question. One experiment was conducted to find out if microwave irradiation has an effect on the transcription levels of an *Arabidopsis thaliana* cell culture (publication/chapter 3). With explorative analysis methods, I found that the irradiation had an effect on gene expression, but this effect was very small and might not have an influence on the physiology of a whole plant.

To answer the question how plant tumors differ from normal inflorescence tissue and how they sustain growth, another gene expression microarray experiment was performed and is described here. Microarray data and solute measurements were used to characterize *Arabidopsis thaliana* tumors by their transcription and solute profiles (publication/chapter 4). Among others, the results showed that the plant tumor cells change from an auxotrophic to a heterotrophic metabolism. The tumor acts like a sink tissue, reducing its photosynthesis to a minimum and accumulating nutrients from the host plant.

Besides using the microarray technology to answer biological questions, I also participated in the development of a new approach to integrate microarray

gene expression data with other data types (publication/chapter 5). Integration of gene expression with chromosomal localization, functional annotation or other high-throughput data can facilitate the interpretation of microarray experiment results. The web application and database GEPAT allows integrating microarray data results with other data types like annotation data on gene function, protein interactions or CGH data. GEPAT has been used to analyze a data set of Mantle Cell Lymphoma (MCL) patients consisting of gene expression microarray and CGH (Comparative Genomic Hybridization) data which is described in publication/chapter 6. The last chapter of the results part presents a re-analysis of Diffuse Large B-cell Lymphoma (DLBCL) gene expression data revealing regulation differences between long and short surviving patients.

The results of part II are discussed in a concluding discussion in part III.

The microarray technology

The central part and common theme of this thesis is the analysis of microarray data. Except the first publication which describes the development and analysis of a phylogenetic DNA microarray, the individual publications of this thesis are concerned with measuring and analyzing gene expression levels to answer a biological question. The reason to study gene expression levels is, that they are fundamentally important for living cells. They are dependent on the cell type, developmental stage and influenced by environmental factors. Transcriptional activity needs to be well-coordinated to assure the proper function of a cell. On the other hand, a dysregulated level of transcription can lead to disease and cancer.

To study whole genome transcription levels, microarrays have become popular in recent years. A DNA microarray can be pictured as a miniture gene-detection assay. The detection is based on the complementary binding properties of DNA to DNA or DNA to RNA. Microarrays hold thousands of spots of different DNA sequences each interrogating a particular gene. Different platforms exist, the most common are cDNA arrays, short oligonucleotide arrays and long oligonucleotide arrays (Figure 1).

For cDNA arrays, as used in publications 6 and 7 of this thesis, first a library of cDNA clones, each containing the sequence of one expressed gene, needs to be constructed. The gene sequences are amplified with PCR and printed on a glass array. Because the size of the individual gene spots and the amount of immobilized DNA varies between spots, usually cDNA from two samples is labelled with different fluorescent dyes and hybridized to the same array. Then the fluorescent signal intensities are read out with a laser scanner. With this technique, one achieves relative measurements of gene expression: expression signal of sample 1 relative to the expression signal of sample 2. Thus, variation in spots size and amount of DNA are evened out.

The probes of short oligonucleotide microarrays are synthesized *in situ* directly on the array and are usually only about 25 bp long. Because they are so short, they are less specific for a gene which is compensated for by

using several probes for the same gene. Some short oligonucleotide arrays also contain probes which have a non-matching base at the central position of the oligonucleotide. The measurements from these probes can be used to estimate cross-hybridization effects. Because the size of the spots can be better controlled, each array is hybridized with cDNA or RNA from only one sample and the signal intensities are read with a laser scanner (Figure 1). A short oligonucleotide microarray commercially available from Affymetrix was used in publications 2, 3 and 4.

The third platform makes use of long oligonucleotides. Gene specific oligonucleotides of about 70 bp length are selected with an oligo design algorithm, synthesized *in situ* and printed on a glass array. Usually, the arrays are then treated like cDNA arrays and hybridized with two differently labeled samples to achieve relative expression measures (Figure 1), (Barrett and Kawasaki, 2003). In principle, this platform was used in publication 1, except that instead of *in situ* synthesized oligonucleotides, PCR fragments of 100-150 bp length were spotted on the microarrays and only one sample was labeled and hybridized at a time.

Applications of the microarray technology

The main application of the microarray technology is to analyze the expression level of thousands of genes simultaneously, finding genes expressed in significantly different patterns in condition or tissue A compared to condition or tissue B. Expression arrays provide a snapshot of cells, tissue or whole organisms at a certain timepoint. Currently several hundred up to millions of cells are needed for one microarray measurement, therefore the expression values give an average estimate over a possibly heterogenous population of cells. Although laser caption dissection allows to collect only cells from the same cell type yielding a homogenous starting material but the amounts are generally so low that the RNA needs to be amplified before it can be used for a microarray hybridization. RNA amplification steps, however, can introduce amplification bias. Although recently it has been shown that single cell expression measurements are feasible for a number of different cell types, they are not in extensive use yet. As of December 2007, in PubMed [1] only a few publications describe experiments where RNA had been extracted from a single cell to hybridize a microarray (Hartmann and Klein, 2006; Kamme *et al.*, 2003; Tietjen *et al.*, 2003; Chiang and Melton, 2003). Besides the complicated protocols to isolate single cells and extract RNA or DNA, care needs to be taken of possible amplification bias. However, data about the transcription levels of single cells would vastly increase the possibilities to learn about cell functions and regulation. In this thesis, samples consisting of a composition of cells were used to estimate differences in transcription levels of two different conditions or tissues, a fact that one might want to bear in mind when interpreting the results of microarray analyses. The differences in expression can be considered as an

[1] www.pubmed.gov

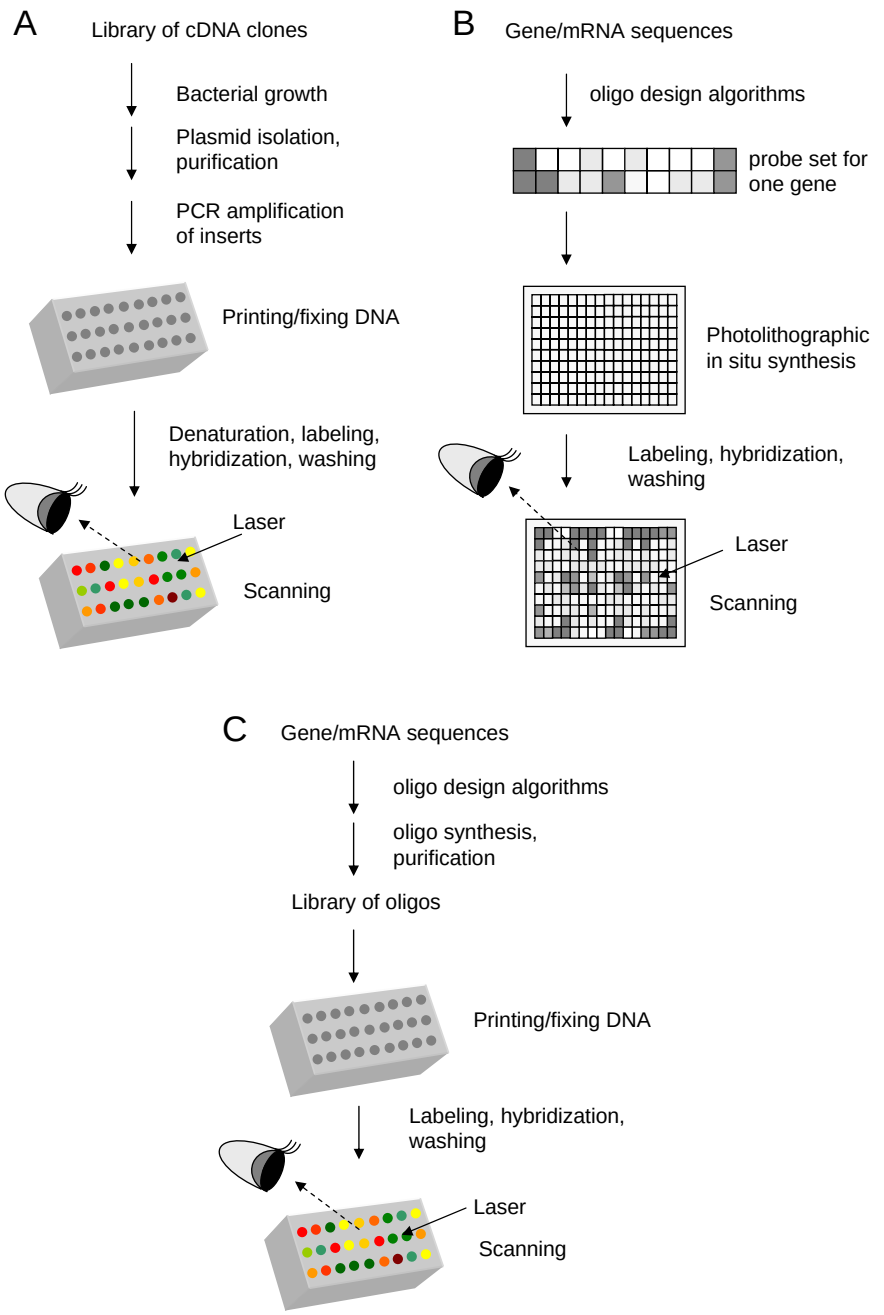


Figure 1: The three most common microarray platforms. A) cDNA arrays. cDNA clones are cultured in bacteria and stored in 96- or 384-well plates. The cDNA sequences are amplified and typically spotted onto glass slides. Sample RNA is labeled, hybridized to an array and scanned by a laser scanner. B) Short oligonucleotide arrays. Perfectly matching oligomers and for some types of arrays also oligomers with one mismatch are synthesized *in situ* directly on the array. Sample RNA is hybridized to an array and scanned. C) Long oligonucleotide arrays. Long specific oligomers of equal length are bioinformatically designed. Printing, hybridization and data analysis is similar to cDNA arrays. The schema is based on Barrett and Kawasaki (2003).

averaged measurement over the cells contained in the sample, while individual cells might have higher or lower changes in expression.

Apart from finding differentially expressed genes between different tissues, treatments, healthy and diseased patients, there are several other areas where expression arrays are applied. Among them are the discovery of biomarkers which can be used to describe a certain tissue or a disease state (Iqbal *et al.*, 2006; Nagata *et al.*, 2003; Tibshirani *et al.*, 2002). Once appropriate biomarkers are found, their transcription level can be measured with low-throughput techniques to characterize the tissue or disease state. For the purpose of diagnostics, also special diagnostic arrays have been developed to classify diseases and their subtypes, e.g. different types of cancer (Golub *et al.*, 1999; Bullinger *et al.*, 2007; Wright *et al.*, 2003), with a limited number of genes.

In clinical applications, gene expression microarrays can be used to find targets for drug development (Clarke *et al.*, 2001; Marton *et al.*, 1998). If the altered transcription of a gene can be associated to a disease, drugs can be developed to recover the transcription level of healthy people. Since this is often difficult, for most cases, it is easier to substitute or influence the product of transcription, the corresponding protein or the cellular processes it influences. A prominent example of altered transcription levels that lead to disease is again cancer. Many so-called oncogenes are transcription factors which act at early stages in signaling pathways (e.g. Ras gene, (Chang *et al.*, 2003)). Their deregulation leads to changes in the transcription of genes further down the pathway and ultimately leads to the development of a cancer cell. For example, the tumor suppressor protein p53 acts as a transcription factor which can activate several independent pathways to sustain a normal cell. If a cell lacks p53 expression, cell growth is unrestrained and tumors can develop (Oren, 2003). In publications 6 and 7 of this thesis, genes could be identified whose expression levels can be used to predict survial of patients, another important clinical application. Additionally, the genes of the predictors might point to possible drug targets.

Further applications of the microarray technology are: DNA arrays for barcoding, comparative genomic hybridization (CGH), genotyping (SNP arrays), chromatin immunoprecipitation (ChIP-chip experiments) and tiling arrays. These will be briefly introduced in the following sections.

Genotyping or SNP (Single Nucleotide Polymorphism) arrays measure single base pair changes, which can be caused by mutations, insertions or deletions. SNP arrays are used to simultaneously identify the genetic variation of numerous single nucleotide positions of individuals and across populations. This process is called genotyping. A famous example of a disease which can be caused by a single nucleotide exchange is sickle cell anemia (Campbell, 1997). In genome wide association studies, the SNPs of populations of healthy individuals are compared to populations of diseased individuals to find base changes associated with the disease. In principle, the analysis of single nucleotide polymorphisms is a classification problem. For each SNP, an algorithm has to decide on the allele frequency. The SNP can be absent, present in one allele or present in both alleles.

Comparative genomic hybridization (CGH) arrays measure absence, presence and amplifications of genomic regions. DNA copy number alterations are key genetic events in the development of cancer. With CGH arrays diseases can be characterized and related to chromosomal abberations (Lichter *et al.*, 2000). Genes that cause the disease can be found, and diseases with known chromosomal abberations can be diagnosed. Additionally, carriers of a disease which might have symptoms in later years of life (e.g. Huntington's disease) can be identified.

To get a holistic view of a system, e.g. a disease, the integration of data from different sources is necessary. For certain types of cancer, it is known that chromosomal abberations increase when the disease progresses. In these cases, the integration of gene expression and DNA copy numbers is indicated. For example, Bussey *et al.* (2006) have integrated CGH data with transcription levels and drug sensitivities for 60 human cancer cell lines (NCI-60). They found a correlation of the gene ERBB2 (v-erb-b2 avian erythroblastic leukemia viral oncogene homologue 2) which induces cancer when over-expressed and the copy number of the genomic region where this gene is found (chromosome 3p). ERBB2 overexpression is observed when 3p is deleted or heterozygosity is lost. From these findings, the authors suggest that the lost regions on chromosome 3 may harbor tumor suppressor genes involved in ERBB2-induced carcinogenesis.

The use of microarray gene expression data to study transcript regulation is limited. A more appropriate but also labor-intensive method to study the regulation of transcription is to determine the locations of binding sites of regulatory proteins on genomic DNA (e.g. transcription factors). Interactions between proteins and DNA can be found by chromatin immunoprecipitation (ChIP) combined with DNA microarray technology into so called ChIP-chip analysis. For a ChIP-chip experiment, a protein is incubated with DNA and then bound protein is cross-linked to the DNA. The protein-DNA complex is pulled out by an antibody specific to the protein. The DNA is eluted, labelled and hybridized to a genomic array which either spans promoter regions or the entire genome at regular intervals. The spots of fragments which are enriched in the DNA sample can be identified as protein binding sites. The drawback of this technique is that only one protein (e.g. transcription factor) can be studied at a time. For an overview of design, analysis and application of ChIP-chip experiments, see Buck and Lieb (2004).

To determine which sections of a genome are transcribed at a certain time or under a certain condition, tiling arrays offer a high-throughput solution. The probes on this kind of array are spaced equally and cover the complete genome, whether they are known to code for a transcript or not. Thus, in contrast to gene expression arrays, these arrays are used to find transcripts which have not been characterized so far (Johnson *et al.*, 2005; Mockler *et al.*, 2005).

A field that has only recently discovered microarray technology for their purposes is DNA barcoding (Moritz and Cicero, 2004). A DNA barcode is a defined region in the genome which allows the identification of a species.

Barcoding approaches sequence a marker region of several species, but for a large number of samples, this is time and cost-intensive. DNA microarrays for species detection have a fundamentally different design than for example, gene expression arrays. While expression arrays contain thousands of genes of one organism, arrays for species detection harbor the complementary sequences of one marker gene for several hundred species. With this design, phylogenetic arrays can be used to distinguish species in environmental samples (Avarre *et al.*, 2007; Hajibabaei *et al.*, 2007). While for most of the other array technologies, a broad spectrum of analysis methods exist, the analysis of DNA arrays for species detection or phylogenetic arrays has been neglected in the past. In this thesis, a data analysis method to improve the detection power of phylogenetic arrays on closely related species is presented in publication 1. Furthermore, applications of microarray technology and approaches for their analysis of finding differentially expressed genes, meta-analysis, and comparative genomic hybridization are presented in later chapters.

Introduction to microarray statistics

Normalization

In the past decade, a number of different microarray technologies and platforms for the analysis of gene expression have been developed. The most prominent ones are: two-color cDNA arrays, where two samples are hybridized simultaneously to the same array with two different colors and a direct comparison of the samples is possible; Affymetrix Gene Chip Technology with short oligomers on a glass wafer and only one color and arrays with long oligomers and usually two colors (Figure 1). The different technologies and platforms require specific algorithms to compute expression values, usually one value per gene. The signal intensities of the individual probes on the microarray are read out with a scanner. Optical noise and non-specific binding requires that the raw signals are preprocessed before statistical data analysis methods are applied. This step is usually called normalization (Quackenbush, 2001). To make the gene expression measurements from the single microarray hybridizations comparable, the gene expression values are first adjusted in respect to the other measurements on the same array (within-array-normalization) and then- if needed- they are adjusted in respect to the measurements of the other arrays (between-array-normalization) (Smyth *et al.*, 2003). The more complex normalization methods usually include both steps. These preprocessing steps should ideally remove any technical variation due to slightly different hybridization conditions, spatial and scanning effects and other factors and conserve the biological signal in the data. Because of the *bias-variance trade-off*, one tries to use a normalization method which adjusts the data just as much as needed to remove the bias and keep as much of the biological signal as possible (Huber *et al.*, 2005; Yang and Paquet, 2005).

A popular normalization method which can be applied to both Affymetrix and cDNA arrays because it works on single channels is VSN (normalization

and variance stabilizing transformation) (Huber *et al.*, 2002). Methods which take into account the multiple probes for each gene on Affymetrix arrays are RMA (Irizarry *et al.*, 2003), gcRMA (Wu *et al.*, 2004), and PLIER (Hubbell *et al.*, 2004). Preprocessing methods mainly used on cDNA arrays are loess, lowess and print-tip loess (Yang *et al.*, 2002). They are applied on the gene expression ratios of the red and green channel (often termed "M") and assume that the majority of genes do not change their expression level over the different conditions. A more detailed review about the multiple normalization methods can be found in Irizarry *et al.* (2006) and Smyth and Speed (2003). Once expression values have been calculated, the statistics for differential expression, clustering, classification and others can be applied to the data independent of their origin.

Explorative analysis and quality control

Quality control is closely intertwined with the preprocessing steps. A microarray consists of a set of DNA sequences often called *probes* which are immobilized on a solid surface (array). A sample contains a complex mixture of nucleic acid sequences often referred to as *targets*, which can bind to the probes on the array (Huber *et al.*, 2005). This binding takes place because complementary nucleic acid sequences hybridize to each other. The amount of labeled and bound sample sequences is read out with a laser scanner. At each of these steps, systematic error can be introduced which must be identified and possibly be reduced during normalization. Therefore, graphical inspection of the data before and after normalization is important to identify problematic samples, spatial effects on one or several arrays and the appropriateness of all preprocessing steps. Problematic hybridizations can be found with density plots and histograms which display the general shape of the data distribution. After normalization, the arrays should have a similar distribution under the assumption that most of the genes do not change under a certain treatment.

Further means to visualize microarray data are box plots, MA-plots (Minus-versus-Add-plots) and methods which reduce the dimensions of data like Principal Components Analysis (PCA) and Correspondence Analysis (CA).

Box plots give a compact overview over a distribution by graphically representing the five-point-summary. The boxplot displays the minimum, first quartile, median, third quartile and maximum of a distribution. The interquartile range is represented by a box, minimum and maximum by whiskers. Thus, several distributions can be aligned in one plot and their median and variance can be easily compared (Yang and Paquet, 2005).

MA-plots display the difference in expression (M) versus the mean expression (A) over two conditions (Huber *et al.*, 2005). With a loess curve drawn on top of the MA-plot, it can be used for quality control. An oscillating loess curve or a large variability in the M values of one array compared to the others indicates problems with this array. If the quality of the arrays can be considered good, MA-plots can also be used to visualize differentially expressed genes (large M values). However, MA-plots cannot estimate significance.

Correspondence analysis (CA) and Principal Component Analysis (PCA) are unsupervised clustering techniques which are used to project high-dimensional data in a low-dimensional space while retaining as much information as possible. In CA, the data is scaled such that rows and columns are treated equivalently and can be represented in the same space. By displaying both rows and columns of a matrix in the same graph, CA allows to connect row- to column vectors and vice versa. In the context of microarray analysis, CA is applied to project the vectors of microarray samples and genes into a lower-dimensional subspace (typically two dimensions) that accounts for the main variance in the data, in a way that distances among points reflect their original distances in the high-dimensional space as closely as possible (Fellenberg *et al.*, 2001).

Like CA, Principal Component Analysis (PCA) tries to reduce the dimensions of a dataset, but the data is not scaled. Therefore, PCA is typically applied on one dimension (rows or columns) of a data matrix. The axes of the new coordinate system are chosen in a way that each axis or principal component explains as much of the (remaining) variance of the data as possible and that all axes after the first are orthogonal to the ones before (Jolliffe, 1986). Visually inspecting CA and PCA graphs, one can assess similarities and differences between microarray samples and/or genes and also discover experimental artifacts.

The statistical software R (R Development Core Team, 2007) and Bioconductor (Gentleman *et al.*, 2004) offer an excellent environment for the production of diagnostic plots for quality control and explorative analysis.

Differential expression measures

Probably the most central question researchers want to answer with microarray experiments is: which genes are differently transcribed in two conditions, tissues, developmental stages or time points. This question also recurs in all but the first chapters of this thesis. A very simple approach is to calculate a fold change, that is the difference in gene expression between two conditions or tissues, for each gene and rank the genes according to the fold change. But the drawback of this approach is that first, the fold change does not include an assessment of significance on gene expression differences and second, the genes with the largest changes in expression might not be the genes which are biologically most interesting. Therefore, statistics which will rank the genes in the order of significance of differential expression are more popular today. In microarray data analysis, the t-test and variations thereof are very common (Cui and Churchill, 2003; Hatfield *et al.*, 2003; Smyth, 2004; Tusher *et al.*, 2001). The t-test is applied on each individual gene and the significance of a change in gene expression given the variance of this gene is calculated. A drawback of this approach is that a large number of hypothesis tests are performed which can lead to many falsely called significant genes. Multiple testing procedures adjust the p-values (measure of significance) of the individual tests such that they are valid considering the total number of tests (Dudoit *et al.*, 2003; Scholtens and von Heydebreck, 2005). Because microarray experiments typi-

cally interrogate a large number of genes and only a small number of samples (large n, small p), applying a classical t-test is not appropriate. The variance of a gene cannot be reliably calculated when there are only a small number of replicate microarray hybridizations. The Bioconductor package *limma* offers a robust solution for small sample sizes and has therefore been used for the differential expression analyses contained in this thesis. It fits a linear model to the expression values of each gene. Because empirical Bayes methods are used to borrow information across genes, the analyses are also stable for a small number of arrays (Smyth, 2004; Smyth *et al.*, 2005). A variance prior is estimated from the data and added to the variance of each gene, lowering the chances that a gene is falsely called differentially expressed because it shows a very low variance. In the linear model, several groups of arrays can be compared and different effects can be included. These effects can be dye or batch effects or correlations of replicate spots on the same array (within-array technical replicates). To set up the linear model, two matrices are required: the *design matrix* which represents which RNA has been used to hybridize each array and the *contrast matrix*, which holds the information of how the coefficients of the design matrix need to be combined to yield the *contrasts* (comparisons) of interest. After the linear models have been fit to the data, an ordered list of differentially expressed genes with logarithmic fold changes and p-values can be obtained.

In the next step, one has to decide on a critical value above which the gene is considered to be significant. While ranking the genes by significance is relatively easy to achieve and several different softwares exist (Smyth, 2004; Tusher *et al.*, 2001), the decision on the p-value criterion has to be made by the individual researcher taking into account any peculiarities about the experiment and what the further analysis and experimental steps will be. Often, only a limited number of genes can be followed up with confirmative experiments in the laboratory, so the number of genes for further research will be kept low. If a more general overview is aspired, Gene Ontology (GO) (Gene Ontology Consortium, 2001) or functional category analysis is appropriate which can handle larger numbers of genes. A functional category analysis approach to characterize *Arabidopsis thaliana* tumors is presented in publication 4.

Integration of different data types improves microarray results

The results of the statistical microarray analysis are the basis for the subsequent biological or medical analysis. Again, bioinformatic methods can aid in this step of the analysis. Further data like functional annotation and classification, chromosomal localization and interaction data can point to the biological processes that are regulated by transcriptional changes.

Enrichment analysis of functional categories, e.g. defined by Gene Ontology (Gene Ontology Consortium, 2001) or MapMan (Usadel *et al.*, 2005) or metabolic or regulatory pathways from KEGG (Kanehisa *et al.*, 2006) help in

getting an overview of the transcriptional changes. An example of a functional analysis based on MapMan is demonstrated in chapter 4, describing the expression profiles of *Arabidopsis thaliana* tumors. In this chapter, the expression profiles are also integrated with solute measurements to round off the characterization of plant tumor cells. Integrating expression data with functional annotation is one of the strengths of the web application GEPAT (publication 5). The chromosomal position of a gene can also influence its expression profile. In bacteria, the organization of genes in operons leads to the collective regulation of an operon (Campbell, 1997). In eukaryotes, the regulation of genes is more complicated, here, methylation can repress transcription over large areas of the genome. Certain types of cancer display typical chromosomal aberrations which also lead to position specific gene expression profiles. Thus, it is worthwile to analyze gene expression data under its chromosomal context, which can also be done in GEPAT. Where data on chromosomal aberrations (CGH) is available, its integration with gene expression data can yield additional insights into tumor biology as demonstrated by the analysis of gene expression and CGH data of Mantle Cell Lymphoma patients in chapter 6. Thus, several examples of how results from gene expression microarray experiments can be integrated with other data types, e.g. functional annotation, protein interactions and chromosomal position are given in the individual chapters of this thesis. This data integration helps in interpreting the results and draw meaningful conclusions.

Part II

Results

Chapter 1

Modeling cross-hybridization on phylogenetic DNA microarrays increases the detection power of closely related species

Modeling cross-hybridization on phylogenetic DNA microarrays increases the detection power of closely related species

Julia C. Engelmann[a] Sven Rahmann[b,c] Matthias Wolf[a] Jörg Schultz[a]
Epameinondas Fritzilas[b] Susanne Kneitz[d] Thomas Dandekar[a]
Tobias Müller[a*]

March 25, 2008

[a]Department of Bioinformatics, Biocenter, University of Würzburg, Am Hubland, D-97074 Würzburg, Germany
[b]Bioinformatics for High-Throughput Technologies, Computer Science 11, TU Dortmund, D-44221 Dortmund, Germany
[c]Genome Informatics, Faculty of Technology, and Graduate School for Bioinformatics and Genome Research, CeBiTec, Bielefeld University, D-33594 Bielefeld, Germany
[d]Microarray Core Facility, Interdisciplinary Center for Clinical Research, University of Würzburg, Versbacher Str. 7, D-97078 Würzburg, Germany

Keywords: DNA microarray, ITS2, green algae (Chlorophyta), biodiversity, barcoding, species identification

Running title: Species detection with DNA microarrays

- accepted by *Molecular Ecology Resources* -

*to whom correspondence should be addressed. email: Tobias.Mueller@biozentrum.uni-wuerzburg.de, fax: +49 931 888 4552

Abstract

DNA microarrays are a popular technique for the detection of microorganisms. Several approaches using specific oligomers targeting one or a few marker genes for each species have been proposed. Data analysis is usually limited to call a species present when its oligomer exceeds a certain intensity threshold. While this strategy works reasonably well for distantly related species, it does not work well for very closely related species: Cross-hybridization of non-target DNA prevents a simple identification based on signal intensity. The majority of species of the same genus has a sequence similarity of over 90%. For biodiversity studies down to the species level, it is therefore important to increase the detection power of closely related species. We propose a simple, cost-effective and robust approach for biodiversity studies using DNA microarray technology and demonstrate it on scenedesmacean green algae. The internal transcribed spacer 2 (ITS2) rDNA sequence was chosen as marker because it is suitable to distinguish all eukaryotic species even though parts of it are virtually identical in closely related species. We show that by modeling hybridization behavior with a matrix algebra approach, we are able to identify closely related species that cannot be distinguished with a threshold on signal intensity. Thus this proof-of-concept study shows that by adding a simple and robust data analysis step to the evaluation of DNA microarrays, species detection can be significantly improved for closely related species with a high sequence similarity.

Introduction

In recent years, DNA barcoding has become popular to study the inventory of natural communities. For DNA barcoding of species, a short standardized genomic region is sequenced and compared to a sequence library of known species (Hebert *et al.*, 2003). DNA microarrays offer an alternative to sequencing and have been shown to perform comparably well in the detection of mammalian species (Hajibabaei *et al.*, 2007a). While a limitation of DNA microarrays is that the sequences to be detected need to be known beforehand, they have the advantage that also complex mixtures of species (e.g. from environmental samples) can be analyzed (Summerbell *et al.*, 2005). In this article, we present a DNA microarray approach which offers the potential to perform large scale biodiversity studies.

The applicability of DNA microarrays for microbial diagnostics has been shown by several publications (He *et al.*, 2007; Kostić *et al.*, 2007; Lehner *et al.*, 2005; Loy *et al.*, 2002, 2005; Peplies *et al.*, 2003). The 16S rDNA has been a popular marker to distinguish bacterial species, but the 18S rDNA is not suited for closely related eukaryotic species when resolution on the species level is desired.

To distinguish mammalian species, mitochondrial marker genes cytochrome *c* oxidase I and cytochrome *b* yielded promising results (Hajibabaei *et al.*, 2007a; Pfunder *et al.*, 2004). While cytochrome *c* oxidase I and cytochrome *b* sequences are available for a wide range of animal taxa, the coverage of plant, protist and fungi sequences is rather poor. The internal transcribed spacer regions 1 and 2 (ITS1, ITS2) cover a wide range of taxonomic levels (animals, plants, protists and fungi); however, fewer ITS than cytochrome

sequences are available for animals (Hajibabaei *et al.*, 2007b).

For DNA microarrays distinguishing fungal species, Leinberger *et al.* (2005) and Nicolaisen *et al.* (2005) have chosen the ITS region of the rRNA gene cassette. Leinberger *et al.* designed an oligonucleotide array to diagnose pathogenic *Candida* and *Aspergillus* species. They used a more complicated microarray design with 51 capture and control probes for 12 species lying in the ITS1, ITS2, 5.8S or 18S rDNA. Hybridizing genomic DNA from only one strain at a time, classification was performed with a threshold on signal intensity. Nicolaisen *et al.* used an oligonucleotide array with ITS2 capture probes to distinguish 12 *Fusarium* species living on cereal grain and of which some produce toxic compounds. While they could group the species in different groups of toxic compound producers and non-producers with a simple intensity cutoff, it was difficult to find specific ITS2 oligonucleotides that would yield resolution on species level. They succeeded for 7 out of 12 species, leaving potential to improve results by a more sophisticated data analysis.

We also use the ITS2 sequence to detect closely related species with a DNA microarray. While the ITS2 sequence has been widely used for phylogenetic reconstruction on the genus and species level in the past, it has also been proposed as a marker for taxonomic classification over a wide range of levels by Coleman (2003); Müller *et al.* (2007); Schultz *et al.* (2006, 2005); Wolf *et al.* (2005). Because the ITS2 sequence is surrounded by the highly conserved 5.8S and 28S rDNA, sequences from different species can be amplified with universal primers, so the ITS2 sequences of all species present in the sample can be amplified in a single PCR reaction. Within the ITS2 sequence, some parts are very conserved, others are highly variable. Choosing the variable parts of the ITS2 sequences of algae species as microarray capture probes, we analyzed whether this microarray is capable of distinguishing between closely related algae species.

While for the analysis of gene expression microarrays, numerous algorithms based on parametric models exist for normalization, differential expression, classification and so on, there are practically no such algorithms for species microarrays. Data analysis of species microarrays has been neglected in the past and was mostly restricted to calling a species present when the signal intensity was above an arbitrary threshold (Bodrossy *et al.*, 2003; Leinberger *et al.*, 2005; Loy *et al.*, 2002; Mitterer *et al.*, 2004; Nicolaisen *et al.*, 2005; Nübel *et al.*, 2004). While this strategy works well for many distantly related species, it does not when very closely related species with high sequence similarity are studied. Cross-hybridization will then inhibit a straightforward analysis. But sometimes, it is of particular interest to identify organisms down to the species or even strain level, for example when one species is pathogenic, blooming or toxic and the other closely related ones are not.

The limitations of the intensity threshold approach become even more apparent when mixtures of species are to be diagnosed. While most DNA microarray applications are capable of identifying the majority of species when they were hybridized in pure culture, they have difficulties in identifying mixtures

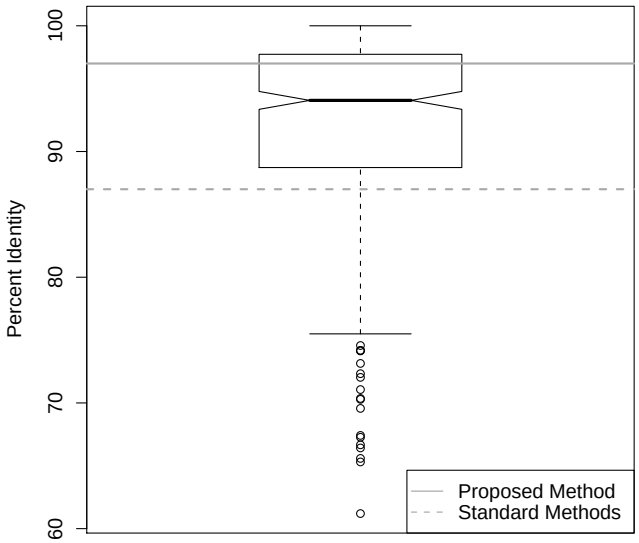


Figure 1: Distribution of the sequence identity of two randomly drawn species from the same genus. From the multiple sequence-structure alignments of four hundred genera (Müller *et al.*, 2007), 10 pairs of sequences were randomly chosen and the percent identity was calculated. Using standard methods (cutoff on signal intensity), species with up to 87% sequence similarity can be distinguished with microarrays. Thus 78.9% of species are missed. With our approach, 70.5% of species can be identified and only 29.5% are missed.

of different species (Wilson *et al.*, 2002), if mixtures were considered at all. Recent theoretical work of Klau *et al.* (2007) and Ragle *et al.* (2007) on the design of non-unique probes, especially in the context of a given phylogenetic tree (Schliep and Rahmann, 2006) suggests that species detection may be possible with a sufficiently large number of oligonucleotide probes, but these approaches have not been applied in practical wet lab work, as far as we know.

In the literature, it has been reported that species with a sequence identity of up to 75-87% in a long capture oligo can be identified with a cutoff on signal intensity. For species with higher sequence similarity, cross-hybridization of non-target probes impedes applying a cutoff criterion. Figure 1 shows that the majority of two species from the same genus has a higher sequence similarity than 87% in the ITS2 rDNA. If assuming a standard microarray approach detecting species based on a cutoff on signal intensity can distinguish species with up to 87% sequence similarity, then, this approach can only identify 21.1% of closely related species. Therefore it is important to develop approaches that raise the value of percent sequence identity to be able to identify closely related species. With the approach proposed here, we could identify species with a sequence similarity of at least 97%. Therefore, we are now able to distinguish at least 70.5% of closely related species (Figure 1).

We chose green algae (Chlorophyta) species for this proof of concept study because they show a high degree of biodiversity comprising a large number of species and because many of them can only be identified by an algae expert using light microscopy or by sequencing. Recently, Johnson *et al.* (2007) described the diversity of *Scenedesmus* Meyen 1829 and *Desmodesmus* (Chodat) An, Friedl & E. Hegewald

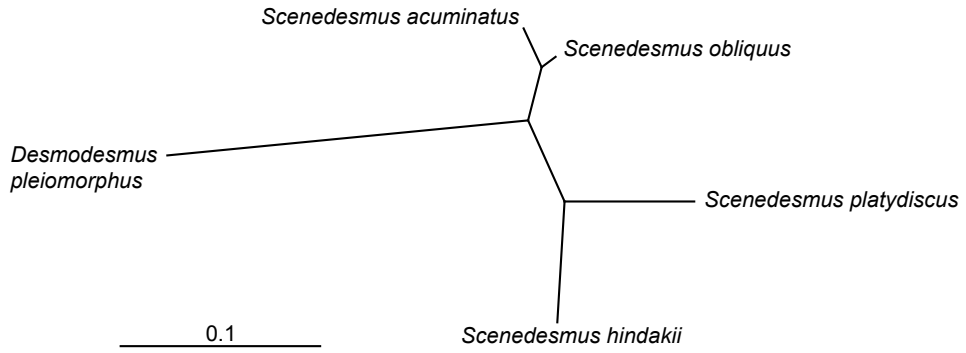


Figure 2: Neighbor-Joining tree (Kimura-2-Parameter model) of amplified ITS2 sequences of the five green algae under study. The tree is in concordance with Hegewald and Wolf (2003).

1999 species in Itasca State Park, Minnesota, USA using light microscopy and ITS2 sequence analysis. For this kind of studies, algae species detection with a microarray would be a sensible alternative because it can be less time consuming on large datasets and more powerful on mixed environmental samples. While low amounts of DNA of one species in a mixture of several species are not detectable with typical Sanger sequencing, a carefully designed DNA microarray coupled with our proposed data analysis can in principle predict any concentration of DNA.

This pilot study resembles a worst case scenario: five closely related green algae classified within the *Scenedesmaceae* Oltmanns 1904 (Sphaeropleales, Chlorophyceae) with sequence similarities up to 97% (Figure 2). Our goal was to keep everything as simple as possible and thus easily transferable to different settings. Therefore the design of the microarray included only one marker gene per species, represented by a capture probe that was spotted in several replicates on the microarray. To take into account the different hybridization and cross-hybridization properties of the different algae, we propose an approach that models the affinities of the capture probes to their targets and non-targets to be able to diagnose which species are present in a sample. If our DNA microarray is capable of distinguishing very closely related species, it will be comparatively easy to extend it to more species, which can either be closely or more distantly related.

Methods

Taxon sampling

Algae cultures were obtained from the Culture Collection of Algae (SAG) at the University of Göttingen, Germany and The Culture Collection of Algae at the University of Texas at Austin, USA (UTEX).

The following five closely related algae species were chosen:

Scenedesmus hindakii E. Hegewald et Hanagata 2000 (SAG47.86),

Table 1: Primers to amplify ITS2 capture probes for the microarray.

Scenedesmus hindakii (SAG47.86)	5'-GCTTTCCCAATCCTTTAGGG-3' (forward); 5'-AAGCCGTTGCTACCTATCCA-3' (reverse)
Scenedesmus acuminatus (UTEX415)	5'-TACCCTCACCCCTCTCTCCT-3' (forward); 5'-CCATATCGGGTCCTTGCTTA-3' (reverse)
Scenedesmus obliquus (UTEX1450)	5'-TACCCTCACCCCTCTCTCCT-3' (forward); 5'-CCATATCGGGTCCTTGCTTA-3' (reverse)
Desmodesmus pleiomorphus (UTEX1590)	5'-ACCCTCACCCCTCTTCCTTA-3' (forward); 5'-CTATCCAGTTGAGCCCGAAT-3' (reverse)
Scenedesmus platydiscus (UTEX2457)	5'-GGCTTGTTAGCCAGCCATAG-3' (forward); 5'-CCATAACGGGTCCTTGCTTA-3' (reverse)

Scenedesmus acuminatus (Lagerheim) R. Chodat 1902 (UTEX415),

Scenedesmus obliquus (Turpin) Kützing 1833 (UTEX1450),

Desmodesmus pleiomorphus (Hindák) E.Hegewald 2000 (UTEX1590),

Scenedesmus platydiscus (G.M. Smith) R. Chodat 1926 (UTEX2457).

Selection of representative DNA sequences as capture probes

ITS2 sequences from the selected *Scenedesmus* and *Desmodesmus* species were retrieved from GenBank (gi|37727740, gi|6625510, gi|6625531, gi|12055733, gi|56122680). The sequences were aligned with Clustal V (Higgins *et al.*, 1992) and for each alga, suitable specific primers were created manually that capture as much of the variable region of the ITS2 sequence as possible with a sequence length between 100-152 bp. Primers were checked for forming dimers and hairpin structures with the primer3 software (Rozen and Skaletsky, 2000), BLAST searches were performed to ascertain that the primers are specific for the ITS2 sequence. The chosen primers are shown in Table 1. PCR fragments of the specific primers were used as capture probes and spotted on the glass arrays.

Primers for sample ITS2 sequences

Universal primers ITS3 and ITS4 (White *et al.*, 1990) were used to amplify the ITS2 sequences of the sample DNA.

ITS3: 5'-GCATCGATGAAGAACGCAGC-3' (forward)

ITS4: 5'-TCCTCCGCTTATTGATATGC-3' (reverse)

DNA extraction and PCR amplification

DNA was extracted from liquid algae cultures following the protocol described in Doyle and Doyle (1990). We used a modified version of this protocol where the concentration of CTAB buffer was increased from 2% to 4%. The cells were not ground in liquid nitrogen but with sea sand in CTAB buffer. These modifications were proposed by Anke Braband (Berlin, Germany, personal communication). DNA from several DNA extractions of the same alga species was pooled before PCR amplification.

For PCR amplification of the ITS2 sequences used as capture probes on the microarray, 40 cycles of denaturation at 94°C, primer annealing at 62°C for 20 s, and elongation at 70°C for 20 s were performed. Algae ITS2 PCR products were sequenced to confirm that the cultures were not contaminated by other species.

The thermal profile used for PCR amplification of the sample DNA (ITS3, ITS4 primers) was as follows: 10 cycles of denaturation at 94°C for 45 s, primer annealing at 55°C for 45 s, elongation at 70°C for 3 min followed by 30 cycles of a ramp protocol increasing the elongation step by 0.5 s per cycle.

Microarrays: Spotting, labeling, scanning

PCR products were spotted using 3xSSC + 1.5 mol betaine as spotting buffer and immobilized (30 min. humid chamber followed by baking for 60 min. at 120°C). Each PCR product was spotted 64 times on the microarray. Blocking and washing was performed according to the Schott Nexterion manual for Nexterion Slide E. 100 ng total of algae PCR product for the hybridization of one alga were labeled using Cy3 following manufacturers instructions (CyScribe Direct Labeling Kit, GE Healthcare, UK). For the simultaneous hybridization of two algae, 50 ng of each algae PCR product was used for hybridization. Hybridization to the microarrays was performed in a hybridization station (Lucidea SlidePro, GE Healthcare) at 42°C, overnight. For scanning a ScanArray 4000 (PerkinElmer, MA, USA) was used. Data acquisition was done using the ScanAlyze Software (M. Eisen, LBNL, CA, USA).

Data preprocessing

All calculations were performed in the statistical programming environment `R` (R Development Core Team, 2007). Spot intensities were normalized with the vsn algorithm (Huber *et al.*, 2002) and the values were mapped back to a non-logarithmic scale ($x \mapsto \mathrm{e}^x$). The median of all 64 spots of one algae per array was used for setting up the measurement (algae $\times$ microarrays) intensity matrix Y; thus $Y_{i,k}$ denotes the median of the 64 alga-i capture probe intensities on array k.

Data and `R` scripts to reproduce our analysis are available at `http://www.biozentrum.uni-wuerzburg.de/phylochips.html`.

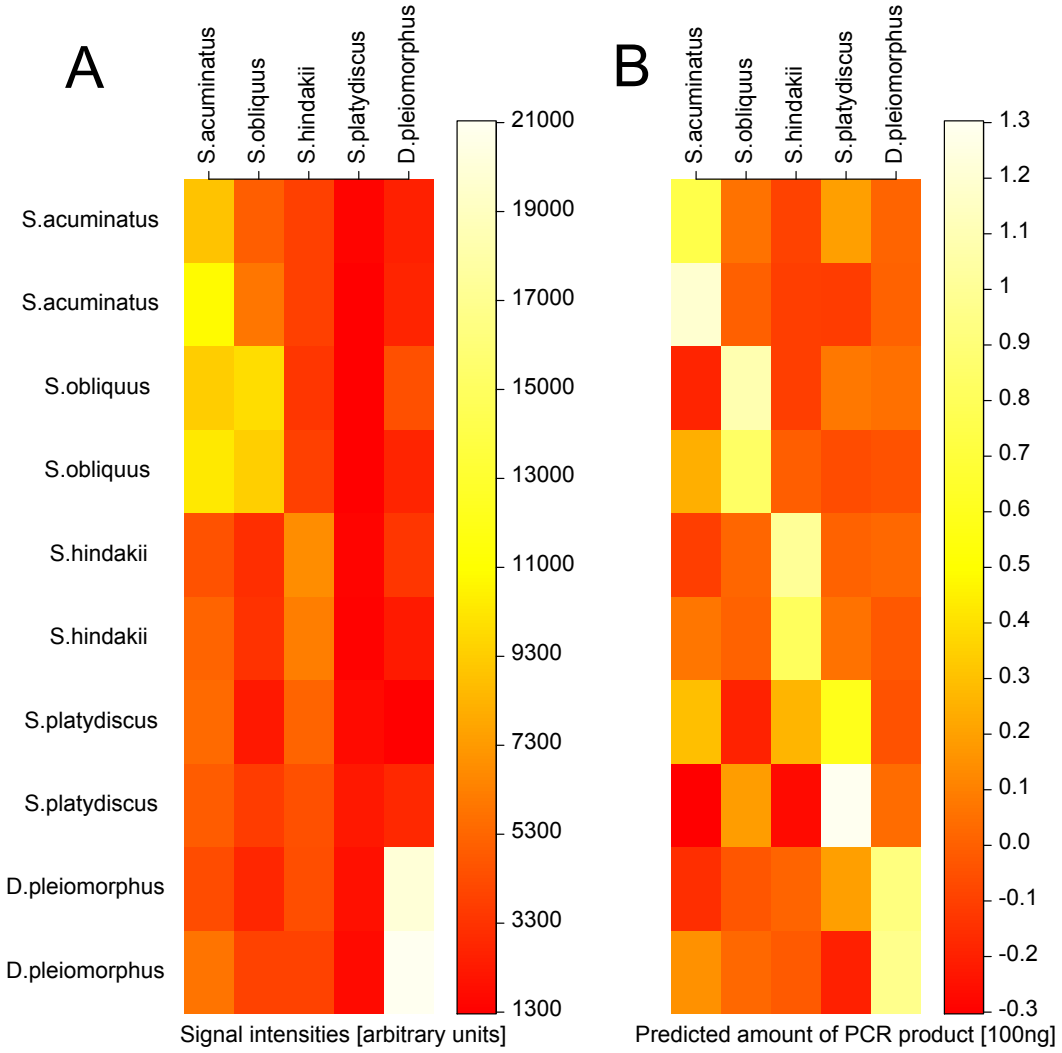


Figure 3: A) Signal intensities of individual algae (columns) when one algae was hybridized to an array (rows). B) Predictions of our model. Predicted amount of PCR product of each algae (columns) when one algae was hybridized (rows). A value of 1 corresponds to 100 ng DNA.

Results

Data analysis with an intensity threshold

For most cases, it is not possible to predict the algae which had been used for hybridization from the signal intensities (Figure 3). If one applies a threshold criterion on the signal intensities, only two algae can be diagnosed correctly. The signal for the alga with the largest sequence divergence, *Desmodesmus pleiomorphus*, is highly specific even when looking at intensity values. This confirms that species detection with microarrays performs well on the genus level. The second alga that can be diagnosed with a threshold on signal intensity is *Scenedesmus acuminatus*. Its capture probes also display the highest intensities when its DNA is hybridized to a microarray.

In concordance with their phylogenetic neighborship (Figure 2), when *Scenedesmus obliquus* is hy-

bridized on an array, the capture probes for both species, *Scenedesmus obliquus* and *Scenedesmus acuminatus* show high intensities (Figure 3). Thus the capture probe of *Scenedesmus obliquus* is not specific. This example shows that for closely related species, cross-hybridization prevents a simple prediction applying a threshold criterion, even when a divergent marker like the ITS2 sequence is used.

The capture probe of *Scenedesmus platydiscus* apparently has a very low binding affinity. When DNA from *Scenedesmus platydiscus* is hybridized, the intensity of its capture probe is lower than the intensities from the other algae (Figure 3).

Looking at exemplary mixtures of two algae, the capture probes of the algae which are present in the sample show the strongest signal, although on one array, the capture probe of an alga not present also displayed a high intensity (signal intensities not shown).

As expected, a simple threshold criterion does not lead to satisfying results when closely related species are involved; therefore we model hybridization behavior.

Modeling hybridization behavior

In DNA microarray analysis, the measured signal intensity of a probe depends on (i) amount of bound target DNA, (ii) amount of bound non-target DNA (cross-hybridization), and (iii) unspecific binding of DNA (background "noise"). In addition, probe affinities to their perfect match targets vary between the different probes making a direct comparison of signal intensities of different probes impossible. Therefore, in our approach the hybridization behavior of each alga was modeled separately. Because the algae species chosen are closely related, cross-hybridization affinities are included in the model. Considering both effects, hybridization and cross-hybridization, we predict the presence of an alga from the probe intensities.

We assume a linear correlation between the amount of algae DNA hybridized to a microarray and the measured fluorescence. For 50-mers of bacterial genes, a linear correlation has been shown by Tiquia *et al.* (2004). Thus, we set up a linear matrix model by

$$Y = A \cdot X \quad \text{or} \quad Y_{ik} = \sum_j A_{ij} X_{jk},$$

where Y is the (algae probes $\times$ microarrays) matrix with $Y_{i,k}$ denoting the median of all spot intensities of the alga-i probes in array k; A is an (algae probes $\times$ algae targets) affinity matrix with $A_{i,j}$ being the affinity of alga-i probes to bind to alga-j target DNA, and X is the unknown (algae targets $\times$ microarrays) design matrix, i.e., $X_{j,k}$ is the amount of DNA of alga j in microarray hybridization experiment k.

Since initially A is not known, we estimate A by carefully designed spike-in experiments with known concentrations of each alga: For each alga, two microarrays with 100 ng PCR product were hybridized (single alga hybridizations). Two microarrays were hybridized with mixtures of two different algae with 50 ng PCR product each (mixed hybridizations). In the absence of cross-hybridization, A would be a

diagonal matrix. In the case of cross-hybridization, A is non-diagonal and its coefficients are estimated by least squares regression as follows.

First, note that probes can be treated independently, since the linear model assumes that there is no competition for PCR product among probes. Fix a row $\alpha := A_i$ of A, corresponding to probe i, and set y to the i-th row of Y. The model for probe i then becomes $y_k = \sum_j \alpha_j X_{jk}$, or $y = \alpha \cdot X$, or equivalently

$$y^\mathsf{T} = X^\mathsf{T} \cdot \alpha^\mathsf{T},$$

where $(\cdot)^\mathsf{T}$ denotes transposition, so y^T and α^T are column vectors. This model is an over-determined linear system in standard form, which we solve for α by the standard least-squares principle using the statistical software `R`, as explained above.

Note that we only assume that X has full rank (equal to the number of different algae), so XX^T is invertible. Other than that we make no specific assumptions about the spike-in experiments, in particular they need not be single-alga experiments, but can be mixture experiments. However, using single alga-experiments is beneficial in the sense that the matrix condition of XX^T is small, yielding more precise estimates.

Since X has full rank, we can express the least-squares solution by the Moore-Penrose inverse of X^T (Penrose, 1955) in the form $\hat{\alpha}^\mathsf{T} = [(X^{\mathsf{TT}} X^\mathsf{T})^{-1} X^{\mathsf{TT}}] \cdot y^\mathsf{T}$. This holds for every row i of A; writing these equations next to each other yields

$$\hat{A}^\mathsf{T} = (XX^\mathsf{T})^{-1} X \cdot Y^\mathsf{T} \quad \text{or} \quad \hat{A} = Y \cdot X^\mathsf{T} \cdot (XX^\mathsf{T})^{-1}.$$

The resulting affinity matrix is shown in Figure 4. The larger part of the matrix is diagonally dominant, demonstrated by the fact that the capture probes of four algae have the highest binding affinities to their target DNA. *Scenedesmus acuminatus* target DNA shows cross-hybridization, it has a strong affinity to both its capture probe and the capture probe of *Scenedesmus obliquus*. The target DNA from *Scenedesmus platydiscus* binds poorly to all of the capture probes, including its perfectly matching probe.

Since we do not constrain $\hat{A}$ to positive values, in principle, we may obtain a solution with negative affinity coefficients. We did not observe this, suggesting that nothing went terribly wrong with the experiments or measurements.

Note that the exact same approach would work if we had two (or a variable number of) probes per alga. In this case, the matrix A would be rectangular.

Now samples with unknown amounts X of PCR product can be hybridized, each probe response can be measured as Y, and the linear system $Y = \hat{A}X$ with estimated A can be solved. If A is square and invertible, as in our case, we can express the solution as $\hat{X} = \hat{A}^{-1} \cdot Y$; if A is rectangular but has full rank,

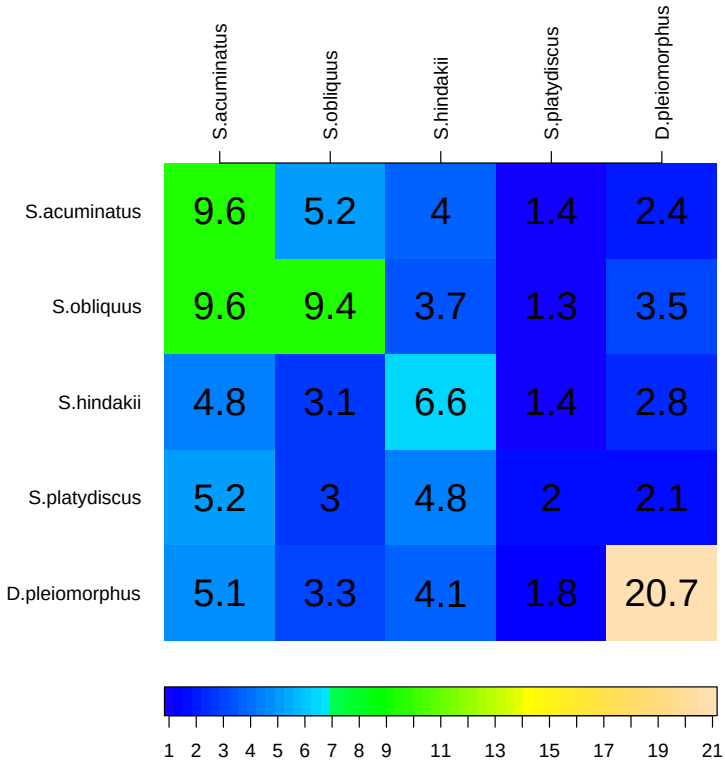


Figure 4: Affinity matrix A. The rows represent the capture probes of the green algae immobilized on the microarray, the columns represent the target DNA which is hybridized to the microarray. The values in the cells need to be multiplied by 10^3 to yield the rounded affinities of alga-i probe to alga-j target DNA. In the case of no cross-hybridization, the matrix would be diagonal. Here the capture probe of *Scenedesmus obliquus* cross-hybridizes with DNA from *Scenedesmus acuminatus*. The DNA from *Scenedesmus platydiscus* has a very low binding affinity, even to its capture probe.

a pseudoinverse (like the Moore-Penrose inverse) can be used to obtain a prediction X for the amount of labelled PCR product used for hybridization for each alga.

For the spike-in experiments, we can evaluate the difference between $\hat{X}$ and the known X to assess the accuracy of the estimated affinity matrix $\hat{A}$. This is discussed in detail below.

Comparison of model versus intensity threshold

By applying a linear model to the probe intensities of our ITS2 microarray, we considerably improve the predictions of which algae are present in a sample. In Figure 5 we show exemplarily how our model outperforms a simple threshold criterion.

For *Desmodesmus pleiomorphus*, whose ITS2 sequence is more divergent from the rest, both the threshold criterion and our model yield the correct result. When DNA from *Scenedesmus obliquus* is hybridized, both the target probe and the probe interrogating *Scenedesmus acuminatus* display the highest intensities and thus the correct alga cannot be detected from the pure intensities. Our model, however, clearly predicts *Scenedesmus obliquus* as the only alga in the sample.

Thus by applying a simple linear model to the microarray data, the five algae species can be perfectly

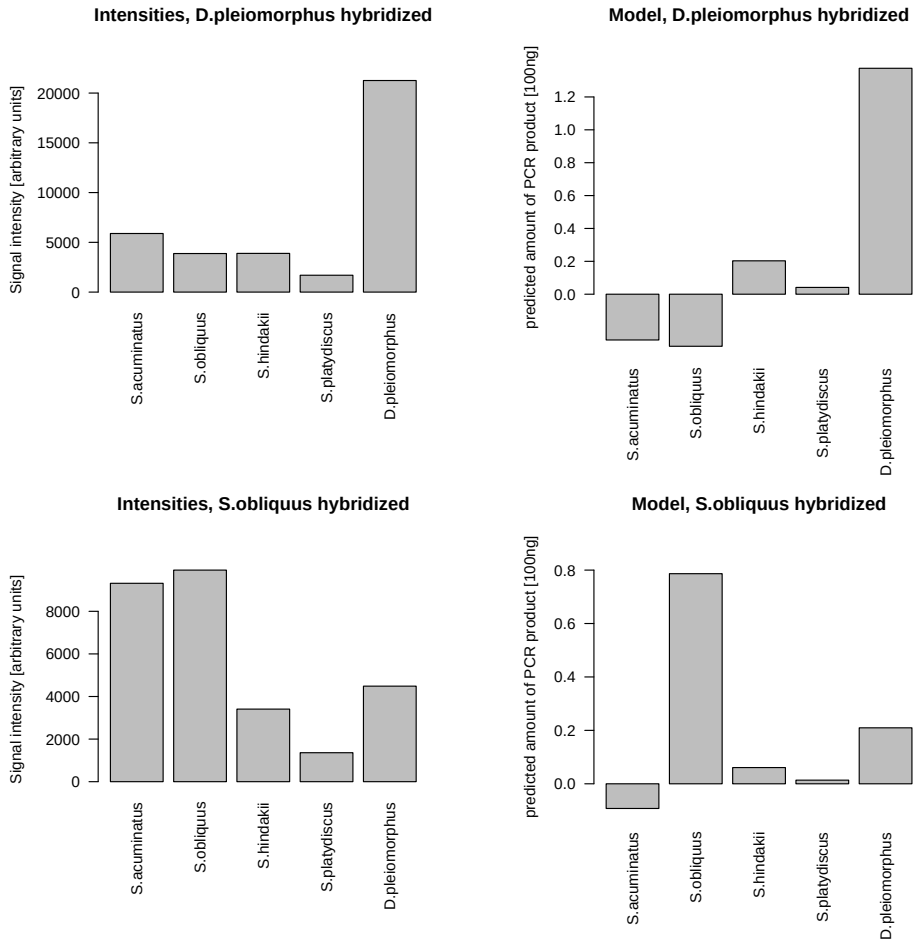


Figure 5: Exemplary comparison of the results of applying a threshold criterion with our model approach. Signal intensities Y of individual algae when one alga (*D. pleiomorphus* or *S. obliquus*, see headers of sub-figures) is hybridized are shown in the left column. A threshold-based approach would only consider these values; note the two high intensities when *S. obliquus* is hybridized. The right column shows the predicted amounts of PCR product $\hat{X}$ of the linear regression model for the same alga. In all cases, Y and $\hat{X}$ should be compared to the true amount of PCR product, which is zero for all algae except the hybridized one which has a true amount of PCR product of 1.

distinguished (Figure 6). The model takes into account the cross-hybridization between the algae and the different binding affinities of the capture probes to their targets. This is particularly evident for *Scenedesmus platydiscus*, where the model takes into account the lower binding properties of its capture probe to correctly predict its presence. The model estimates the affinities such that the overall error is minimized (in a least-squares sense), yielding good predictions for all of the algae at the cost of a less striking separation between the only *Desmodesmus* species and the *Scenedesmus* species, compared to pure signal intensities (Figure 3).

Figure 6A shows the results of our model approach. Training the model on all the microarrays and then predicting the amount of labelled PCR product used for hybridization yields the training error. The amount of DNA for hybridization was standardized, a value of 1 on the vertical axis corresponds to 100

ng of labelled PCR product, 0.5 corresponds to 50 ng. In Figure 6, the predictions for the two replicates of every single alga hybridization and the two mixture hybridizations are shown. For every microarray hybridization (horizontal axis), the alga which was hybridized has the highest predicted amount of PCR product, which varies between approx. 0.6 and 1.3. Furthermore, there is a clear separation between absent and present algae concerning the predicted amount of DNA. Although a quantitative evaluation is possible with our approach, we prefer a qualitative evaluation for this proof-of-concept study. Using a cutoff anywhere between 0.4 and 0.6 to call a species present when it exceeds the cutoff results in 100% correct predictions for the hybridizations of one alga.

The mixtures of two algae are more difficult to predict because here, only 50 ng labelled PCR product of each algae were used for hybridization. The predictions for the microarray hybridized with DNA from *Scenedesmus hindakii* and *Scenedesmus acuminatus* are very good: The predicted amounts are around 0.6 and the predictions for the absent algae are around zero. For the other mixture, the prediction of the present algae are also very good (around 0.5), but the prediction for the absent algae *Scenedesmus hindakii* is too high to speak of a clear separation between present and absent algae.

In Figure 6B, the results of a leave-one-out cross-validation, representing the prediction power of the model on new data, are displayed. Here the model is trained on 11 array hybridizations and the amounts of labelled PCR product of the 12th array are predicted. This procedure is repeated 12 times such that each array is left out once. Looking at the predicted amounts of PCR product of the individual algae, the separation between present and absent algae is still visible, but less clear than in Figure 6A. When using again a cutoff of 0.6 to call a species present, the two predictions of *Scenedesmus platydiscus* are not correct. The amount of this alga is hard to predict, most likely because the signal intensity of its capture probe is always low, even if the alga itself is hybridized to the array. The predictions of the two mixtures of algae are again close to the true DNA amounts used for hybridization, only the amount of *Scenedesmus hindakii* is over-estimated in the mixture where *Scenedesmus acuminatus* and *Scenedesmus obliquus* were hybridized.

Discussion

While several publications describe the development and application of DNA microarrays for species detection, data analysis has been neglected in the past and is typically restricted to calling a species present when the signal intensity exceeds an arbitrary threshold (Loy and Bodrossy, 2006). Determining whether a species is present or not is usually done manually, a tedious work especially when several probes are used for the same species or when many species are studied. Cross-hybridization of non-target DNA even prevents the prediction of closely related species. These major drawbacks can be overcome by applying statistical models, as proposed here, on the microarray data.

For this pilot study, we have chosen a worst case scenario with five closely related green algae from the family of *Scenedesmaceae*. Our aim was to keep the approach simple and easily transferable to different biological problems, therefore only one capture probe per species was chosen for the DNA microarray. In contrast to many former studies, our data analysis does not stop after signal intensities of the single probes have been calculated. From the signal intensities, we model hybridization and cross-hybridization behavior with a simple linear model to estimate which species had been in the sample. With this approach, we get better predictions compared to applying an intensity threshold directly on the fluorescence intensities.

In a cross-validation, for the single alga hybridizations, the accuracy of our model was 80%, considering a species present when the predicted concentration of labelled PCR product used for hybridization of the microarray exceeds 60 ng of DNA. This is a significant improvement compared to a 40% accuracy when an arbitrary but most beneficial threshold was used on the fluorescence intensities. While the predicted amounts of DNA were close to the correct 100 ng for the majority of species present in the sample, the amount of *Scenedesmus platydiscus* was underestimated with a predicted concentration of about 60 ng DNA. Considering that very closely related species were studied here, this is a promising result indicating that even species with a high sequence similarity can be detected with DNA microarrays.

While former studies estimated the threshold for species differentiation around 75-87% sequence similarity (Loy and Bodrossy, 2006, and references therein), we could show that species identification is possible even for sequences of at least 97% sequence similarity in the marker gene.

We have shown here that it is feasible to design a DNA microarray distinguishing closely related species using cost-effective methods. Our system is also sensitive, yielding good results with only 100 ng of PCR products. The shorter PCR fragment used as capture probe for *Scenedesmus platydiscus* (100 bp long) most likely caused the lower overall intensities of that probe. Although we suspected this behavior when choosing the capture probes, we could not find an appropriate primer pair to yield a 140-150 bp fragment like the ones used for the other algae. Using a more accurate but also more expensive technology probing long oligomers on the arrays will make it easier to design oligomers of approximately the same length and with similar binding affinities. This will most likely further improve the predictions and allow quantitative analysis. A combination of both, optimized technology and enhanced data analysis is needed to further increase the sensitivity and accuracy of species microarrays. The proposed linear model is simple but efficient and can easily be applied to different datasets, studying the same or different species.

Other modeling approaches with general linear models or kernel support vector machines with several different kernels were also applied on the microarray data and yielded comparable results. The linear regression approach was finally chosen because it gave the best results in the leave-one-out cross-validation.

In this study, we used specific primers to generate the capture probes on the microarray and a universal primer to amplify the sample ITS2 region. While we are aware of possible biases that might be introduced by PCR amplification when analyzing complex samples, we believe that these biases are small because of

the perfectly matching universal primers. Still, optimization strategies need to be developed to improve specificity and allow quantification of species before this approach can be applied for environmental studies.

To test whether our model is also capable of correctly predicting mixtures of algae species, we hybridized DNA from two different algae in equal amounts to the same array, keeping the total amount of DNA (100 ng) constant. Results are promising: on two test arrays, the present algae had the highest predicted concentrations. On one array, however, the concentration of the non-present *Scenedesmus hindakii* was over-estimated.

An advantage of our model aproach is that not all possible combinations of species have to be hybridized to microarrays to be able to predict them. The only requirement at the moment is that each species has to be hybridized once, either by itself or in a mixture. Then, all possible combinations can in principle be predicted. Nonetheless, future work will include more hybridizations with different mixtures of algae to fully characterize the potential of the model to predict complex mixtures of DNA.

In principle, quantification of algae DNA is possible with our modeling approach using only one color in the hybridization. Nonetheless, quantification would be more robust when using two colors and measuring relative differences between two samples. This would be particularly useful to measure spatial and temporal changes in the composition of species in environmental samples.

With the advent of ultra-fast sequencing technologies, much more precise quantitative measurements become feasible. But because at present, a large initial investment is required for this technique, it will not be an option for many small laboratories. ITS2 DNA microarrays coupled with sound statistical analysis, however, offer the potential to conduct more cost-effective large-scale studies.

Recently, it has been shown that multicopy genes such as rDNA display variation between the individual copies which can confound biodiversity estimates (Thornhill *et al.*, 2007). Unlike DNA barcoding approaches using sequencing, DNA microarray approaches like the one presented here are less affected by this phenomenon. Because the resolution of DNA microarrays is not high enough to discover single base pair changes on a sequence length of about 150 bp, single mutations between different copies of a multicopy gene will not affect hybridization.

With a well-annotated 100k ITS2 sequence database at hand (Schultz *et al.*, 2005, 2006; Wolf *et al.*, 2005), it is fairly easy to design any ITS2 species array with the eukaryotic species of interest. Focused as well as large-scale biodiversity studies with hundreds or even thousands of species can be set up based on this database. Possible applications for these microarrays would be the diagnosis of species in water or soil samples, changes of species composition in these samples over time or the diagnosis of toxic species.

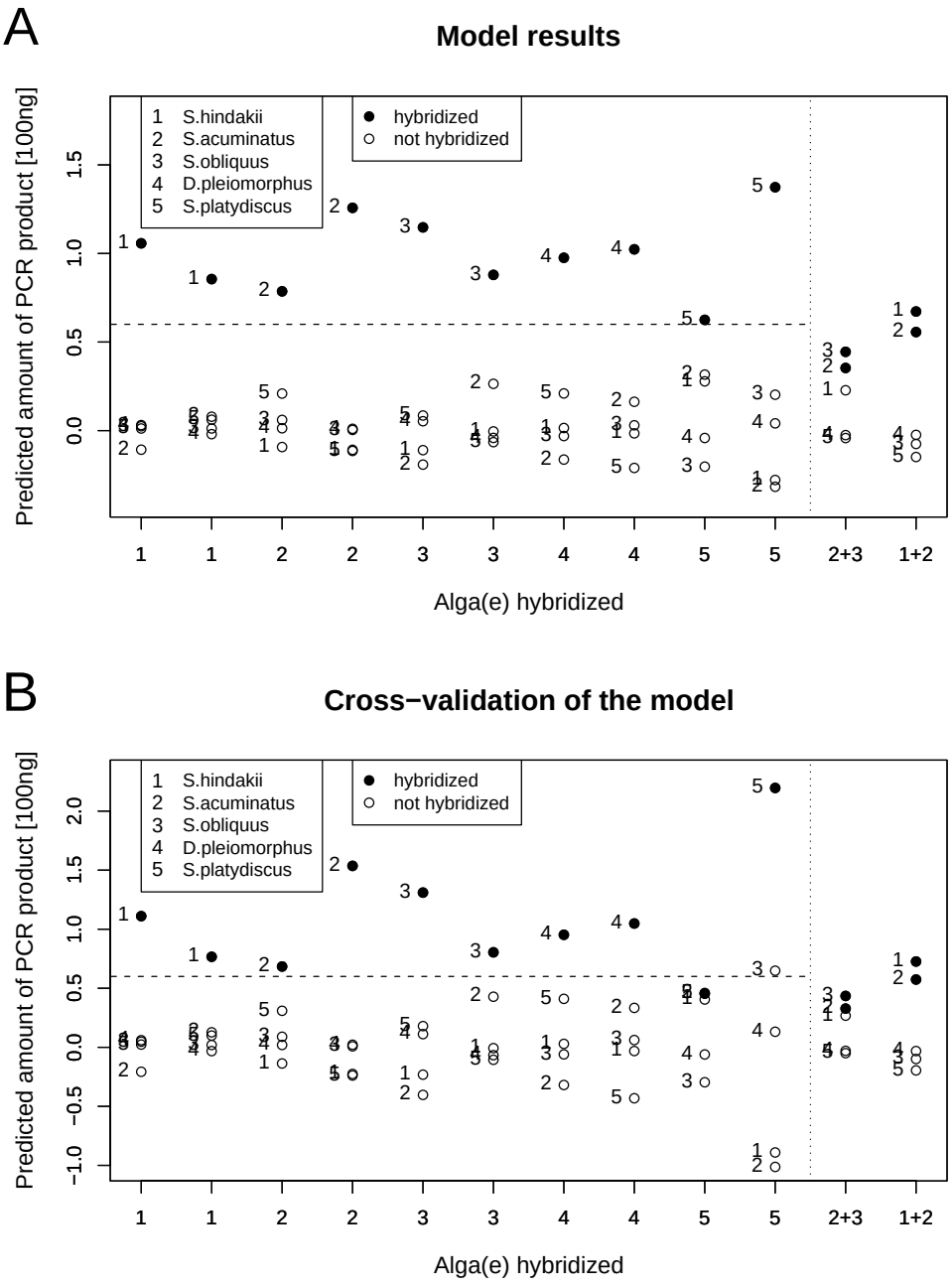


Figure 6: Modeling results. A) Training error. The affinities were derived from all microarrays and used to predict the amounts of PCR product ($\hat{X}$) of each alga. There were two replicates of each single alga hybridization and two microarrays with mixtures of two different algae (x-axis). B) Test error of leave-one-out cross-validation. It represents the prediction power of the model on new data. Affinities were derived from 11 hybridizations and the one left out was predicted such that each array was once predicted. A predicted amount of 1 corresponds to 100 ng DNA.

References

Bodrossy L, Stralis-Pavese N, Murrell JC, *et al.* (2003) Development and validation of a diagnostic microbial microarray for methanotrophs. *Environmental Microbiology*, **5**, 566–582.

Coleman AW (2003) ITS2 is a double-edged tool for eukaryote evolutionary comparisons. *Trends in Genetics*, **19**, 370–375.

Doyle JJ, Doyle L (1990) Isolation of plant DNA from fresh tissue. *Focus*, **12**, 13–15.

Hajibabaei M, Singer GAC, Clare EL, Hebert PDN (2007a) Design and applicability of DNA arrays and DNA barcodes in biodiversity monitoring. *BMC Biology*, **5**, 24.
URL: *http://dx.doi.org/10.1186/1741-7007-5-24*

Hajibabaei M, Singer GAC, Hebert PDN, Hickey DA (2007b) DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends in Genetics*, **23**, 167–172.
URL: *http://dx.doi.org/10.1016/j.tig.2007.02.001*

He Z, Gentry TJ, Schadt CW, *et al.* (2007) GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. *The ISME Journal*, **1**, 6777.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings. Biological sciences / The Royal Society*, **270**, 313–321.
URL: *http://dx.doi.org/10.1098/rspb.2002.2218*

Hegewald E, Wolf M (2003) Phylogenetic relationships of *Scenedesmus* and *Acutodesmus* (Chlorophyta, Chlorophyceae) as inferred from 18S rDNA and ITS-2 sequence comparisons. *Plant Systematics and Evolution*, **241**, 185–191.

Higgins DG, Bleasby AJ, Fuchs R (1992) CLUSTAL V: improved software for multiple sequence alignment. *Computer Applications in the Biosciences*, **8**, 189–191.

Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18** **Suppl 1**, S96–104.

Johnson JL, Fawley MW, Fawley KP (2007) The diversity of *Scenedesmus* and *Desmodesmus* (Chlorophyceae) in Itasca State Park, Minnesota, USA. *Phycologia*, **46**, 214–229.

Klau GW, Rahmann S, Schliep A, Vingron M, Reinert K (2007) Integer linear programming approaches for non-unique probe selection. *Discrete Applied Mathematics*, **155**, 840–856.

Kostić T, Weilharter A, Rubino S, *et al.* (2007) A microbial diagnostic microarray technique for the sensitive detection and identification of pathogenic bacteria in a background of nonpathogens. *Analytical Biochemistry*, **360**, 244–254.
URL: *http://dx.doi.org/10.1016/j.ab.2006.09.026*

Lehner A, Loy A, Behr T, *et al.* (2005) Oligonucleotide microarray for identification of Enterococcus species. *FEMS Microbiology Letters*, **246**, 133–142.
URL: *http://dx.doi.org/10.1016/j.femsle.2005.04.002*

Leinberger DM, Schumacher U, Autenrieth IB, Bachmann TT (2005) Development of a DNA microarray for detection and identification of fungal pathogens involved in invasive mycoses. *Journal of Clinical Microbiology*, **43**, 4943–4953.
URL: *http://dx.doi.org/10.1128/JCM.43.10.4943-4953.2005*

Loy A, Bodrossy L (2006) Highly parallel microbial diagnostics using oligonucleotide microarrays. *Clinica Chimica Acta*, **363**, 106–119.
URL: *http://dx.doi.org/10.1016/j.cccn.2005.05.041*

Loy A, Lehner A, Lee N, *et al.* (2002) Oligonucleotide microarray for 16S rRNA gene-based detection of all recognized lineages of sulfate-reducing prokaryotes in the environment. *Applied and Environmental Microbiology*, **68**, 5064–5081.

Loy A, Schulz C, Lücker S, *et al.* (2005) 16S rRNA gene-based oligonucleotide microarray for environmental monitoring of the betaproteobacterial order "Rhodocyclales". *Applied and Environmental Microbiology*, **71**, 1373–1386.
URL: *http://dx.doi.org/10.1128/AEM.71.3.1373-1386.2005*

Mitterer G, Huber M, Leidinger E, *et al.* (2004) Microarray-based identification of bacteria in clinical samples by solid-phase PCR amplification of 23S ribosomal DNA sequences. *Journal of Clinical Microbiology*, **42**, 1048–1057.

Müller T, Philippi N, Dandekar T, Schultz J, Wolf M (2007) Distinguishing species. *RNA*, **13**, 1469–1472.
URL: *http://dx.doi.org/10.1261/rna.617107*

Nicolaisen M, Justesen AF, Thrane U, Skouboe P, Holmstrøm K (2005) An oligonucleotide microarray for the identification and differentiation of trichothecene producing and non-producing *Fusarium* species occurring on cereal grain. *Journal of Microbiological Methods*, **62**, 57–69.
URL: *http://dx.doi.org/10.1016/j.mimet.2005.01.009*

Nübel U, Schmidt PM, Reiss E, *et al.* (2004) Oligonucleotide microarray for identification of *Bacillus anthracis* based on intergenic transcribed spacers in ribosomal DNA. *FEMS Microbiology Letters*, **240**,

215–223.

URL: *http://dx.doi.org/10.1016/j.femsle.2004.09.042*

Penrose R (1955) A generalized inverse for matrices. *Proceedings Of The Cambridge Philosophical Society*, **51**, 406–413.

Peplies J, Glöckner FO, Amann R (2003) Optimization strategies for DNA microarray-based detection of bacteria with 16S rRNA-targeting oligonucleotide probes. *Applied and Environmental Microbiology*, **69**, 1397–1407.

Pfunder M, Holzgang O, Frey JE (2004) Development of microarray-based diagnostics of voles and shrews for use in biodiversity monitoring studies, and evaluation of mitochondrial cytochrome oxidase I vs. cytochrome b as genetic markers. *Molecular Ecology*, **13**, 1277–1286.

URL: *http://dx.doi.org/10.1111/j.1365-294X.2004.02126.x*

R Development Core Team (2007) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

URL: *http://www.R-project.org*

Ragle MA, Smith JC, Pardalos PM (2007) An optimal cutting-plane algorithm for solving the non-unique probe selection problem. *Annals of Biomedical Engineering*, in press.

URL: *http://www.springerlink.com/content/kp7621877143567n/*

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods in Molecular Biology*, **132**, 365–386.

Schliep A, Rahmann S (2006) Decoding non-unique oligonucleotide hybridization experiments of targets related by a phylogenetic tree. *Bioinformatics*, **22**, e424–e430.

Schultz J, Maisel S, Gerlach D, Müller T, Wolf M (2005) A common core of secondary structure of the internal transcribed spacer 2 (ITS2) throughout the Eukaryota. *RNA*, **11**, 361–364.

URL: *http://dx.doi.org/10.1261/rna.7204505*

Schultz J, Müller T, Achtziger M, *et al.* (2006) The internal transcribed spacer 2 database–a web server for (not only) low level phylogenetic analyses. *Nucleic Acids Research*, **34**, W704–W707.

URL: *http://dx.doi.org/10.1093/nar/gkl129*

Summerbell RC, Lévesque CA, Seifert KA, *et al.* (2005) Microcoding: the second step in DNA barcoding. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **360**, 1897–1903.

URL: *http://dx.doi.org/10.1098/rstb.2005.1721*

Thornhill DJ, Lajeunesse TC, Santos SR (2007) Measuring rDNA diversity in eukaryotic microbial systems: how intragenomic variation, pseudogenes, and PCR artifacts confound biodiversity estimates. *Molecular Ecology*, **16**, 5326–5340.
URL: *http://dx.doi.org/10.1111/j.1365-294X.2007.03576.x*

Tiquia SM, Wu L, Chong SC, *et al.* (2004) Evaluation of 50-mer oligonucleotide arrays for detecting microbial populations in environmental samples. *Biotechniques*, **36**, 664–70, 672, 674–5.

White TJ, Bruns T, Lee S, Taylor J (1990) *PCR protocols: a guide to methods and application*, chap. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. Academic Press, Inc., San Diego, CA, USA, pp. 315–322.

Wilson KH, Wilson WJ, Radosevich JL, *et al.* (2002) High-density microarray of small-subunit ribosomal DNA probes. *Applied and Environmental Microbiology*, **68**, 2535–2541.

Wolf M, Achtziger M, Schultz J, Dandekar T, Müller T (2005) Homology modeling revealed more than 20,000 rRNA internal transcribed spacer 2 (ITS2) secondary structures. *RNA*, **11**, 1616–1623.
URL: *http://dx.doi.org/10.1261/rna.2144205*

Acknowledgments

We gratefully acknowledge the funding from DFG (German Research Foundation) for grant number Mu-2831/1-1, BMBF Projekt FUNCRYPTA (FKZ 0313838B), SFB 630/6, and IZKF Z1. We would like to thank Margarete Göbel for technical assistance and Anke Braband for providing us her modified protocol for DNA isolation. Thanks to the anonymous reviewers for insightful comments.

Chapter 2

Unsupervised meta-analysis on diverse gene expression datasets allows insight into gene function and regulation

Unsupervised meta-analysis on diverse gene expression datasets allows insight into gene function and regulation

Julia C. Engelmann[1] Roland Schwarz[1] Steffen Blenk
Torben Friedrich Philipp N. Seibel Thomas Dandekar
Tobias Müller*

March 25, 2008

Running header: Unsupervised meta-analysis on gene expression datasets

Department of Bioinformatics, Biocenter, University of Würzburg, Am Hubland, D-97074 Würzburg, Germany

[1] Both authors contributed equally

* Corresponding author. Phone: +49 931 888 4563,
mail: Tobias.Mueller@biozentrum.uni-wuerzburg.de

Abstract

Over the past years, microarray databases have increased rapidly in size. While they offer a wealth of data, it remains challenging to integrate data arising from different studies. Here we propose an unsupervised approach of a large-scale meta-analysis on *Arabidopsis thaliana* whole genome expression datasets to gain additional insights into the function and regulation of genes. Applying kernel principal component analysis and hierarchical clustering, we found three major groups of experimental contrasts sharing a common biological trait. Genes associated to two of these clusters are known to play an important role in indole-3-acetic acid (IAA) mediated plant growth and development or pathogen defense. Novel functions could be assigned to genes including a cluster of serine/threonine kinases that carry two uncharacterized domains (DUF26) in their receptor part implicated in host defense. With the approach shown here, hidden interrelations between genes regulated under different conditions can be unraveled.

Keywords: *Arabidopsis thaliana*, microarray, unsupervised meta-analysis, function prediction, database, gene expression

- accepted by *Bioinformatics and Biology Insights* -

Introduction

In the last years, enormous data has been generated with microarray experiments from different organisms, tissues and platforms under various experimental conditions. Databases like the NCBI Gene Expression Omnibus (GEO) (Barrett et al. 2007), ArrayExpress (Parkinson et al. 2007) and NASCArrays (Craigon et al. 2004) have been set up to archive these datasets and to make them available to the scientific community. The size of microarray databases is likely to increase exponentially in the future, as is typical for all molecular databases, increasing the need for sophisticated methods to analyze these large amounts of data appropriately.

Several factors impede a straight-forward analysis of microarray database content: standards for data submission vary between different databases, some microarray datasets do not provide raw data and on the experimental side, protocols and experimental conditions can differ between diverse laboratories conducting microarray hybridizations. However, a major advantage of microarray meta-analysis is that through the integration of a potentially large number of datasets, additional insights into gene regulation can be gained which could have been overseen or not detected in the single experiments. Reasons for this could be that either the signal from a particular gene or group of genes was too weak to be detected in the single experiment or because it can be put into a functional context taking into consideration its regulation under other conditions or treatments.

Several methods for microarray meta-analysis have been proposed in recent years, most of them using models which compute an "effect size" and take care of inter-study variation (Choi et al. 2003; Conlon et al. 2006; Hu et al. 2005; Moreau et al. 2003). Thus, they often resemble procedures applied for the detection of differential expression but add the study as an extra explanatory variable. Several datasets from different microarray experiments are integrated in the meta-analysis to increase the number of replicates and thereby the power to detect differentially expressed genes. Because this design implies that datasets addressing the same topic such as the same cell type or treatment are used, microarray meta-analyses of this kind usually consist of only a small number of studies.

A second approach to supervised microarray meta-analysis is to integrate knowledge of biological functions into the analysis to predict global co-expression relationships and to infer functional relationships between co-regulated genes (Huttenhower et al. 2006).

Nevertheless, all the above methods are based on parametric models which have several biological and statistical assumptions. Similar to classical microarray analysis, in which a first explorative analysis reveals possible signals in the data which can then be verified or disproved by parametrical hypothesis testing, our approach of unsupervised meta-analysis yields insights into the biological structure of the data and may thus lead to precise biological hypotheses. These could then be tested by the parametric models described above. The aim of this study is to compare the results from a large number of microarray experiments on *Arabidopsis thaliana* using the well established

Affymetrix ATH-1 Genome Array [1] as a starting point. We restricted our analysis to this highly-standardized platform to reduce uninformative variability introduced by different technologies.

In this unsupervised meta-analysis, we show how to overcome the challenges posed by the heterogeneity of microarray data and apply exploratory data analysis methods. First, microarray datasets from public web sources were collected and pre-processed to remove noise from the data and build a common data basis for further analyses. Later, exploratory data analysis was applied to the processed datasets, namely kernel Principal Component Analysis (kPCA) and spectral and hierarchical clustering, to group contrasts from different microarray experiments and to find genes regulated in a specific cluster. Identification of regulated genes in a specific cluster was achieved by unsupervised feature subset selection using the kernel principal component loadings. Although gene selection or feature subset selection is a challenging task for classification, many different approaches have been proposed for the same. According to our knowledge, gene selection or feature subset selection has not yet been performed using loadings of features on kernel PCA scores in the context of meta-analysis.

Genes selected to play a role in either plant growth and development (related to indole-3-acetic acid, a plant growth hormone) or pathogen defense were mapped onto physiological processes and functions and could be validated by previous studies. For genes which have not completely been characterized yet, our approach was able to propose a function and a possible regulatory mechanism as shown here for DUF26 (Domain of Unknown Function) kinase genes.

Methods

Data pre-processing

Microarray data were collected from the Gene Expression Omnibus (GEO) database (Barrett et al. 2007). For our analysis, we defined a *dataset* as a GEO entry with a unique GSE series accession number. Each dataset consisted of several Affymetrix CEL-files, each one representing the raw data from one microarray hybridization. The raw data of one microarray is termed a *sample* in the following section. Instead of comparing whole GEO datasets with each other, we broke down each dataset into *contrasts* and used these as 'entities' for our analysis (Fig. 1, (Everitt 2005)). A *contrast* is the difference in gene expression between any two sample groups of the same dataset. A sample group contains all replicate samples from one condition (e.g. treatment, mutant, see Table 2). Therefore, for most GEO datasets, several contrasts were set up. For example, a contrast could be a comparison of an *Arabidopsis thaliana* mutant with a wild type plant.

A contrast was then represented by a vector of the logarithmic (base 2) fold changes of all 22810 probe sets on the ATH1 chip. The majority of probe sets on the ATH1 chip interrogates the expression level of one gene, some match to two or more genes. Before computing the fold changes, raw intensity values of all samples of a contrast

[1] http://www.affymetrix.com/products/arrays/specific/arab.affx

were normalized using the gcRMA algorithm implemented in the *gcrma* package (Wu et al. 2005) which is part of Bioconductor (Gentleman et al. 2004) and runs under the statistical software R (R Development Core Team 2004). Logarithmic fold changes and p-values adjusted for multiple testing using the false discovery rate method (Benjamini and Hochberg 2000) were computed using the *limma* package (Smyth 2004) which is also integrated into Bioconductor.

We imposed the following selection criteria on the datasets: a) Availability of the Affymetrix raw data (CEL-files) for download, b) at least two replicates of each condition are available c) time-course experiments were excluded. 20 GEO datasets fulfilled these criteria as of November 2006. From these datasets, 76 contrasts could be set up on the basis of 424 CEL-files. The final data matrix used for the unsupervised meta-analysis was a 76×22810 matrix, 76 contrasts with 22810 log fold changes.

Outlier removal and transformation

To remove experimental outliers from the data which could negatively influence any further analysis, a filtering criterion was set up as follows. Across all experiments, 15% and 85% quantiles of the distributions of medians and variances of the log fold changes were calculated. Experiments whose medians laid outside the inter-quantile-range or whose variances were below the 15% quantile threshold were excluded from further analysis. This resulted in a reduced data matrix X with 41 remaining contrasts. We randomly inspected the 35 removed contrasts for detectable problems and found several contrasts having a low-variant distribution of multiple-testing corrected p-values with almost all p-values close to one.

When dealing with heterogenous experimental datasets from different laboratories and experimental settings, efficient data transformation methods are necessary to produce a reasonable level of comparability. Log fold changes from microarray experiments deserve special attention in that they implicitly define a "direction" of differential expression by their algebraic sign which is semantically not sustainable when comparing contrasts from divergent settings. We therefore only evaluated the absolute value of the log fold changes and brought all remaining 41 contrasts approximately to a standard normal distribution by applying the *Box-Cox-Transformation* (Eq.1, (Box and Cox 1964)) using Maximum-Likelihood estimated power coefficients.

For a power coefficient p and data x the box-cox-transformed data x' is defined as follows:

$$x' = \begin{cases} (x^p - 1)/p & \text{if } p \neq 0 \\ log(x) & \text{if } p = 0 \end{cases} \quad (1)$$

The average p values were about 0.13, resulting in an approximately logarithmic transformation of the log fold changes. Subsequently, all datasets were standardized to zero mean and unit variance to analyze datasets without regard to their scale and location.

Kernel PCA

Principal Component Analysis (PCA) aims to provide a lower dimensional view of high dimensional data by projecting the data points from a data matrix X onto a new coordinate system retrieved by eigen-decomposition of the associated covariance matrix. The axes of the new coordinate system are thereby chosen in a way that each axis or principal component explains as much of the (remaining) variance of the data as possible and that all axes after the first are orthogonal to the ones before.

Kernel PCA (Schölkopf et al. 1998) is a non-linear extension of the regular PCA, performing the same projection in a possibly even higher dimensional feature space. The data points are implicitly projected from the input space I into the feature space F by replacing the standard Euclidean dot product with a positive-semidefinite symmetric bilinear form, the kernel function κ (Eq. 2). The algorithm is represented in a dual form such that all computation takes place using only the matrix of pairwise dot products XX' (Shawe-Taylor and Cristianini 2004), the Gram or Kernel matrix K (Eq. 3), instead of using the data points or its variances directly.

More precisely, for a row-indexed data matrix X and a mapping $\phi : I \to F$, $x \mapsto \phi(x)$ the kernel function κ and its associated kernel matrix K is defined as

$$\kappa(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \quad (2)$$

$$K_{ij} = \kappa(x_i, x_j). \quad (3)$$

Kernel PCA has the advantage of being able to detect non-linear patterns in the data which might be overlooked or not covered appropriately when using conventional PCA.

For our analysis we used the Kernel PCA algorithm implemented in the "kernlab" package (Karatzoglou et al. 2004), for the kernel function κ we chose a polynomial kernel

$$\kappa(x_i, x_j) = (s \langle x_i, x_j \rangle + k)^d$$

of degree $d = 2$, scale $s = 1$ and offset $k = 0$.

Clustering

Clustering was performed on all remaining contrasts after removal of outliers. For an initial identification of the three main clusters of contrasts, we applied a spectral clustering algorithm from the "kernlab" package (Karatzoglou et al. 2004). Spectral clustering algorithms cluster points using eigenvectors of matrices derived from the data, the kernel matrix K in this case. Similar to k-means clustering for data in the input space, the initial number of clusters has to be specified.

To gain structured clustering results, we applied hierarchical clustering using Ward's minimum variance method, which aims to find compact and spherical clusters based on Euclidean distance (Ward 1963). Decomposition of the symmetric kernel matrix K

$$K = S\Lambda S' \quad (4)$$

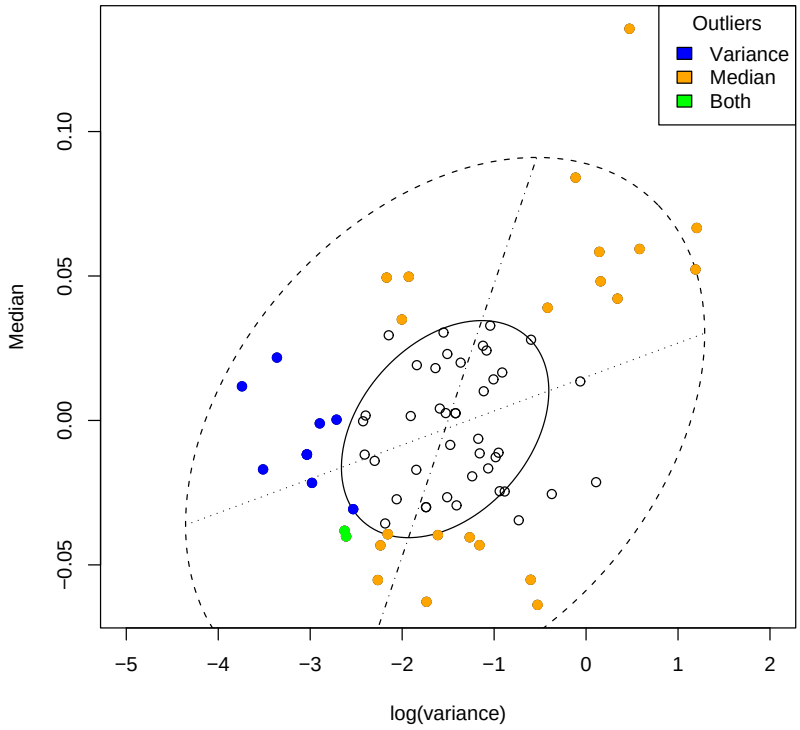


Figure 1: Outlier removal. Median vs. log(variance) plot of all 76 contrasts and the associated bivariate box plot, colors indicate the type of outlier (see legend). The bivariate box plot is the two-dimensional analog of the familiar box plot of univariate data and consists of a pair of concentric ellipses, the hinge and the fence (Everitt 2005). This box plot is based upon a robust estimator for location, scale and correlation. Uncolored contrasts were kept for further analysis.

leads to a product of the orthogonal matrix S of its eigenvectors, a diagonal matrix Λ consisting of its eigenvalues and the transpose of S, S'. As the eigenvalues of K are directly linked to the proportion of explained variance of the principal component axes, the axes were scaled by the square roots of their respective eigenvalues, i.e.

$$\widetilde{X} = S\Lambda^{1/2}. \tag{5}$$

The result is a Euclidean distance

$$d(x_i, x_j) = \sqrt{\langle \widetilde{x}_i, \widetilde{x}_j \rangle} \tag{6}$$

weighted by the information content of each of the vector coefficients, thus scaling down axes that were given a low information content in the previous kPCA analysis.

Uncertainty of the predicted clusters was estimated by a 1000-fold multi-scale bootstrap resampling using the "pvclust" algorithm (Suzuki and Shimodaira 2006).

Table 1: Variance of kernel principal components. Variance of the first 15 principal components on the 41×22810 data matrix of *Arabidopsis thaliana* microarray data, explaining close to 60% of the variance of the data. Abbreviations: PV = Proportion of Variance, CP = Cumulative Proportion of variance.

	PC1	PC2	PC3	PC4	PC5
PV	0.10035	0.05383	0.05003	0.04640	0.03887
CP	0.10035	0.15418	0.20422	0.25062	0.28949
	PC6	PC7	PC8	PC9	PC10
PV	0.03725	0.03250	0.03226	0.03142	0.02973
CP	0.32674	0.35925	0.39151	0.42293	0.45267
	PC11	PC12	PC13	PC14	PC15
PV	0.02793	0.02699	0.02647	0.02606	0.02470
CP	0.48061	0.50761	0.53409	0.56016	0.58486

Results

Dimension reduction by kernel principal component analysis (kPCA)

The ATH-1 whole genome chip consists of 22810 probe sets, this led to a 41×22810 data matrix (contrasts $\times$ log fold changes of probe sets) after outlier removal. To reduce the dimension of the data matrix, a kernel PCA algorithm was applied which was able to cover virtually the complete information content by defining an orthonormal system of 38 principal component axes. The 22810 log fold changes could therefore be represented by a 41×38 data matrix without any measurable loss of information. Using only the first 25 principal components, 80.585% of the variance could be described. If we state that the remaining 20% of the variance in the data describe noise, an estimation which is certainly not too strict in the context of large-scale gene expression measurements, an effective de-noising can be reached by considering only the first 25 principal components in further steps of the analysis. For a detailed overview of the variance distribution on the first 15 principal components, see Table 1.

Unsupervised analysis reveals three clear clusters of contrasts

The principal component plot (Fig. 2) revealed three major clusters of contrasts and several minor ones. In contrast to typical meta-analyses these clusters were not a priori defined, but detected by the proposed unsupervised meta-analysis. Based on this clustering we used an implementation (Karatzoglou et al. 2004) of the spectral clustering algorithm proposed by Ng et al. (2001), a variant of the k-means clustering algorithm in a kernel defined feature space, to support the clusters shown in Fig. 2. According

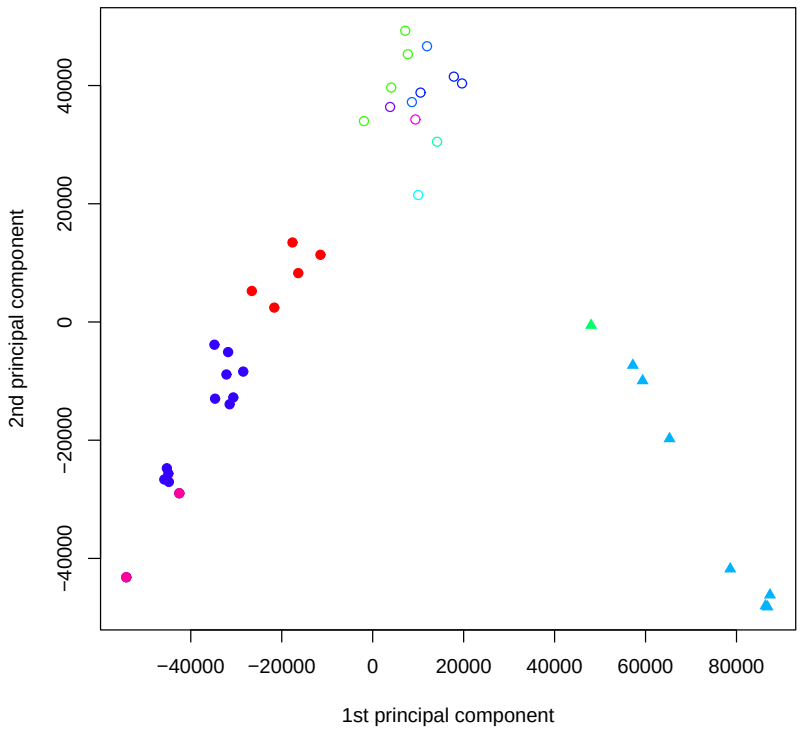


Figure 2: Kernel PCA on 41 *Arabidopsis thaliana* contrasts. Plot of all 41 contrasts using the first two principal component axes. Comparisons are colored according to the experiment they originated from and correspond to the colors used in Figure 3, different shapes indicate the three different clusters obtained from spectral clustering: Indole-3-acetic acid (IAA) related contrasts (solid circle), pathogen related contrasts (triangles) and others (outlined circle).

to the annotation of the datasets retrieved from GEO, the three clusters were related to indole-3-acetic acid (IAA) addition or inhibition (cluster 1, triangles), pathogen defense activation (cluster 2, solid circles) and "others" (cluster 3, outlined circles). For a detailed biological interpretation, see section "Biological interpretation of clusters". Additionally, inspection of the pairwise plots of the other principal components contributing to a lower extent to the variance of the data revealed more contrast clusters.

To get further structural insights into the relationships between contrasts and the experimental settings, we performed hierarchical clustering assessed by multi-scale bootstrapping (Fig. 3). In agreement with the spectral clustering performed earlier and the graphical inspection of the pairwise scatterplots of contrasts on the kPCA axes, the three main clusters of contrasts could also be found as the first two splits in the resulting dendrogram with high bootstrap support.

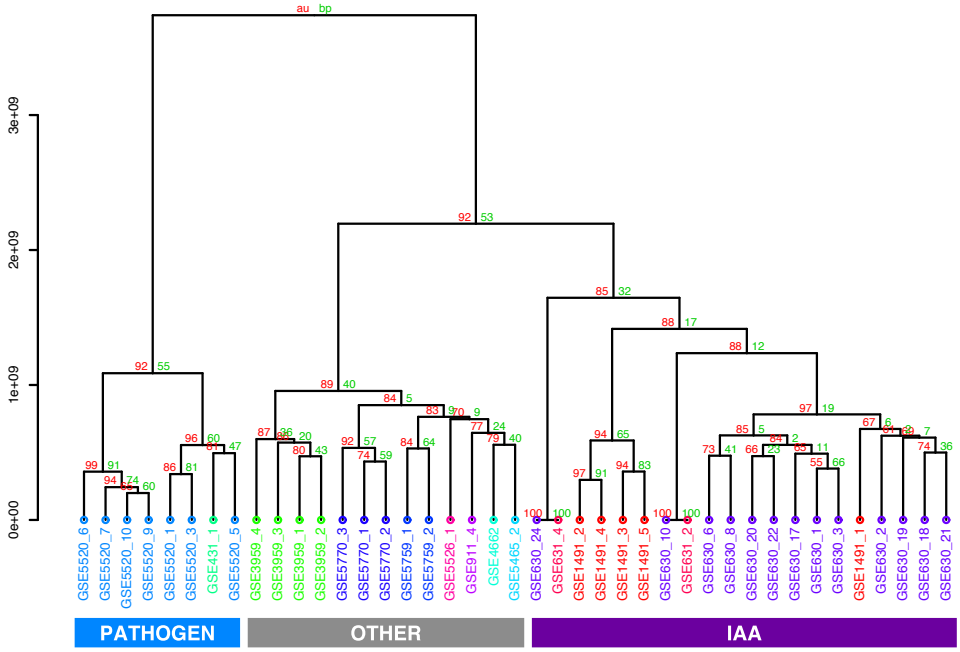


Figure 3: Hierarchical clustering on 41 Arabidopsis contrasts. Cluster dendrogram using hierarchical ward clustering on all 38 principal component vectors resulting from kernel PCA. Contrasts are colored according to their experimental affiliation. Approximately unbiased (au, (Suzuki and Shimodaira 2006)) and standard bootstrap (bp) values are given for all splits and support the results from the previous spectral clustering (Fig. 2).

As the three clusters were mainly separable through the x-axis on the kPCA scatterplot using the first two axes (Fig. 2), we postulated that the first principal component alone might be enough to select genes whose co-regulation patterns could clearly distinguish between IAA related, pathogen-defense related and other contrasts.

Gene selection with kPCA loadings

To accomplish an efficient feature subset selection, i.e. to identify genes that are responsible for the clustering, a variety of methods have been described, e.g. Self-Organizing Maps (SOMs) (Tamayo et al. 1999), Maximal Margin Linear Programming (MAMA) (Antonov et al. 2004), Correlation Based Feature Selection (CFS) (Hall 1999) or Recursive Feature Elimination (RFE) using Support Vector Machines (SVM) (Guyon et al. 2002; Zhang et al. 2006). In consequent continuation of our approach of exploratory meta-analysis, we looked for genes that have a strong association with the first kPCA axis, i.e. we calculated the loadings of each of the genes onto the principal components.

To achieve this with respect to the kernel defined feature space we projected single artificial contrasts containing only one de-regulated gene onto the new coordinate system. Each of the 22810 artificial contrasts was set up in a way that it showed a high absolute fold change value in one of the genes and all others being set to zero. From the resulting 22810×38 matrix of loadings of each of the genes onto the 38 principal components, we selected the 500 top genes for both positive (IAA related) and negative (pathogen related) extrema. To assess the accuracy of the gene selection process exploratively, we repeated the previous kernel PCA analysis using only the selected genes, i.e. on the remaining 41×500 data matrices, and inspected pairwise scatterplots of the first 20 principal components for each dataset of either IAA-related or pathogen-associated genes. All kPCA plots of the IAA-related gene set, even the one of the first two axes which contribute most to the overall variance of the data, showed a wide spread of IAA contrasts along the principal component axes. This indicated a high variance of the selected genes in IAA-related contrasts. All other contrasts were projected onto a compact local cluster by kPCA, demonstrating that the selected genes do not vary in these contrasts. The same was found in the kPCA plots of the matrix with pathogen-associated genes (data not shown). These findings indicate that expression patterns related neither to IAA nor pathogen treatment were efficiently stripped off by the gene selection process.

Biological interpretation of clusters

The hierarchical clustering on all kPCA scores in Figure 3 revealed three main clusters of contrasts: contrasts studying pathogen defense (blue), contrasts analyzing indole-3-acetic acid (IAA) effects (violet) and other contrasts studying various effects (gray). These three clusters were well-supported by high bootstrap values. The labels at the edges include the GEO accession number followed by an index indicating the contrast number. For a detailed description of contrasts see Table 2. For each contrast, two groups of samples were compared and for each group, the genetic background and treatment is listed. The last column of Table 2 indicates the cluster this contrast was assigned to in kernel PCA clustering.

Zooming into the IAA cluster, a cluster containing only contrasts with IAA inhibition (GSE1491_2, GSE1491_3, GSE1491_4 and GSE1491_5) was well-separated from the remaining contrasts, including GSE1491_1, a contrast from the same dataset, but where IAA instead of an IAA inhibitor was added to one sample group. The remaining contrasts in the IAA cluster mainly studied the effect of IAA on different mutants with defects in IAA biosynthesis or signaling. Indole-3-acetic acid (IAA) belongs to a group of plant growth hormones called auxins. The "others"- cluster consisted of contrasts studying various effects like the effect of lincomycin which is an inhibitor of plastid protein translation, regulation changes of an embryogenesis transcription factor mutant or of stress tolerant mutants. Naturally, in this cluster of divergent contrasts, contrasts from the same dataset clustered closely together. The architecture of the hierarchical cluster tree shows that data preprocessing followed by kernel PCA adjusted the data in such a way that contrasts stemming from biologically similar experiments are indeed more similar to each other than to other contrasts. Thus, with our analysis, we were able

to achieve comparability of microarray datasets from different laboratories addressing different biological questions. This is nontrivial and important considering the numerous sources of variation that affect the nature of the datasets underlying this analysis.

Arabidopsis thaliana genes regulated by indole-3-acetic acid (IAA)

To get an overview of the functions of the selected genes representative for the contrast clusters "IAA" or "pathogen", the *Arabidopsis thaliana* pathway analysis program MapMan (Usadel et al. 2005) was used. With MapMan, gene expression values can be displayed onto diagrams of functional categories and metabolic and regulatory pathways. In this study, MapMan was used to visualize the representative genes for the two clusters "IAA" and "pathogen".

Among the genes representative for IAA contrasts, the functional category "hormones" with the subgroup "IAA" defined by MapMan showed the highest proportion of regulated genes (diagram not shown). The subgroup "IAA" consists of 215 genes in MapMan. We selected 500 genes representative for IAA with our approach and out of these, 43 genes are cataloged in the MapMan subgroup "IAA". Thus, by selecting 500 genes from the ATH1 microarray which comprises roughly 2% of the array, we were able to capture 20% of the genes annotated as IAA-related in MapMan.

In the "hormones" subgroup "ethylene", and in the category "transcription factor" many genes are regulated under IAA treatment, while a smaller number of genes is regulated in the categories "Cytochrome P450" and "cell wall" (data not shown).

Regulated genes in the subgroup "ethylene" are either involved in ethylene synthesis or signal transduction. Ethylene plays a role in the regulation of a number of developmental processes, often in interaction with other plant hormone signals. For example, auxins can induce ethylene formation and in turn ethylene can trigger an auxin increase. Some processes such as root elongation, differential growth in the hypocotyl and root hair formation and elongation are regulated by both auxin and ethylene in *Arabidopsis thaliana* (Stepanova et al. 2005). All the GEO datasets we annotated as IAA-related originate from seedling RNA extracts. Since IAA belongs to the group of auxins, the aforementioned processes are likely to be regulated under IAA treatment.

Cytochrome P450 monooxygenases are involved in various biosynthetic reactions which synthesize for example plant hormones or defense compounds. Regulation of cell wall genes is also expected as auxins mediate cell elongation by stretching of the cell wall which requires restructuring processes.

In conclusion, the gene selection of our unsupervised meta-analysis approach chose many genes which are annotated and independently validated as being IAA regulated.

Arabidopsis thaliana genes regulated by pathogen exposure

Gene selection for contrasts studying plant response to pathogens revealed a high number of regulated genes in the following functional categories of MapMan (Usadel et al. 2005): "biotic stress", "receptor kinases", "photosynthesis" (light reactions), "alkaloid-like proteins" from "secondary metabolism", "nitrilases", "cell wall" genes and "WRKY

Table 2: Overview of all contrasts included in the explorative meta-analysis. Each contrast consists of two groups which are described by their genetic background (genotype) and treatment. The last column "Cluster" derives from the clustering on the kernel PCA scores. Contrasts are labeled with the GEO series number followed by a contrast index.

	Sample Group 1		Sample Group 2		
Contrast	Genotype	Treatment	Genotype	Treatment	Cluster
GSE1491_1	WT Col-0	IAA	WT Col-0	non	IAA
GSE1491_2	WT Col-0	IAA inhibitor A	WT Col-0	non	IAA
GSE1491_3	WT Col-0	IAA inhibitor B	WT Col-0	non	IAA
GSE1491_4	WT Col-0	IAA/IAA inhibitor A	WT Col-0	non	IAA
GSE1491_5	WT Col-0	IAA/IAA inhibitor B	WT Col-0	non	IAA
GSE3959_1	MU LEC2GR	1h LEC2 induction	MU LEC2GR	no LEC2 induction	other
GSE3959_2	MU LEC2GR	4h LEC2 induction	MU LEC2GR	no LEC2 induction	other
GSE3959_3	MU LEC2GR	1h LEC2 induction	WT WS-0	4h LEC2 induction	other
GSE3959_4	MU LEC2GR	4h LEC2 induction	WT WS-0	NA	other
GSE431_1	pmr4-1 MU	non	pmr4-1 MU	powdery mildew	pathogen
GSE4662_1	MU STA1	non	WT	NA	other
GSE5465_2	MU OETOP6B	non	WT	NA	other
GSE5520_1	WT Col-0	DC1318 Cor 10e6	MU STA1	non	pathogen
GSE5520_10	WT Col-0	EcTUV86-2 fliC 10e8	WT Col-0	non	pathogen
GSE5520_3	WT Col-0	DC3000 10e6	WT Col-0	non	pathogen
GSE5520_5	WT Col-0	DC1318 Cor 5x10e7	WT Col-0	non	pathogen
GSE5520_6	WT Col-0	DC3000 hrpA-fliC 10e8	WT Col-0	non	pathogen
GSE5520_7	WT Col-0	DC3000 hrpA 10e8	WT Col-0	non	pathogen
GSE5520_9	WT Col-0	EcO157H7 10e8	WT Col-0	non	pathogen
GSE5526_1	WT?	non	WT?	non	other
GSE5759_1	WT Col-0	dark plus lincomycin	WT Col-0	dark	other
GSE5759_2	WT Col-0	red light plus lincomycin	WT Col-0	red light	other
GSE5770_1	WT Col-0	lincomycin	WT Col-0	non	other
GSE5770_2	abi4-102 MU	lincomycin	abi4-102 MU	non	other
GSE5770_3	gun1-1 MU	lincomycin	gun1-1 MU	non	other
GSE630_1	WT Col-0	IAA (2h 5μM)	WT Col-0	EtOH (2h)	IAA
GSE630_10	MU arf2-6	IAA (2h 5μM)	MU arf2-6	EtOH (2h)	IAA
GSE630_17	MU IAA17-6	EtOH (2h)	WT Col-0 I	EtOH (2h)	IAA
GSE630_18	MU arx3-1	EtOH (2h)	WT Col-0 I	EtOH (2h)	IAA
GSE630_19	MU i5i6i19	EtOH (2h)	WT Col-0 I	EtOH (2h)	IAA
GSE630_2	MU nph4-1	IAA (2h 5μM)	MU nph4-1	EtOH (2h)	IAA
GSE630_20	MU IAA17-6	IAA (2h 5μM)	WT Col-0 I	IAA (2h 5μM)	IAA
GSE630_21	MU arx3-1	IAA (2h 5μM)	WT Col-0 I	IAA (2h 5μM)	IAA
GSE630_22	MU i5i6i19	IAA (2h 5μM)	WT Col-0 I	IAA (2h 5μM)	IAA
GSE630_24	MU arf2-6	IAA (2h 5μM)	WT Col-0 A2	IAA (2h 5μM)	IAA
GSE630_3	MU arf19-1	IAA (2h 5μM)	MU arf19-1	EtOH (2h)	IAA
GSE630_6	MU IAA17-6	IAA (2h 5μM)	MU IAA17-6	EtOH (2h)	IAA
GSE630_8	MU i5i6i19	IAA (2h 5μM)	MU i5i6i19	EtOH (2h)	IAA
GSE631_2	MU arf2-6	IAA (2h 5μM)	MU arf2-6	non	IAA
GSE631_4	MU arf2-6	IAA (2h 5μM)	WT Col-0	IAA (2h 5μM)	IAA
GSE911_4	35S::LFY	non	WT ler	35S::LFY	other

transcription factors". For all of the functional categories mentioned above, it has been reported that genes in these categories are regulated after pathogen attack and play a role in plant defense. Figs 4 and 5 show details of the MapMan maps which harbor these categories. In the figures, gray areas inside the diagrams represent all the individual genes present on the ATH1 chip and annotated in MapMan. The selected genes representative for contrasts studying the effects of pathogen exposure are highlighted by small dark blue squares. For example, Fig. 4 C shows that there are 41 DUF26 receptor kinases present on the ATH1 chip, of which 9 are regulated after pathogen exposure. In the following, we give a short description of the functions of the genes regulated after pathogen exposure.

A change in carbohydrate metabolism after pathogen attack as observed here (Fig. 4 A, upper right: "light reactions") has also been reported by Berger et al. (2004) for the pathogens *Pseudomonas syringae* or *Botrytis cinerea*. The authors have shown a co-regulation of defense, sink and photosynthetic gene expression in response to the pathogens under study.

As the cell wall is a natural barrier for plant pathogens, plant defense includes cell wall modifications and biosynthesis to thicken cell walls and impede further pathogen attack (Cheong et al. 2002). Figure 4 A shows that several genes of the cell wall metabolism are regulated after pathogen exposure.

The regulation of WRKY transcription factors (Fig. 4B, upper left) is also described in the publication accompanying the GEO dataset GSE5520 (Thilmony et al. 2006). Our findings confirm their suggestion that these transcription factors regulate plant response to bacteria.

Alkaloids (Fig. 4 A, lower left) are secondary metabolites listed in the "N-misc." category of MapMan. They are generally not essential for the basic metabolic processes of the plant but play an important role in plant defense (Dixon 2001). They are produced by the plant to restrict pathogen feeding. The accumulation of antimicrobial substances is often regulated by signal-transduction pathways which require the perception of the pathogen by a plant receptor encoded by host resistance genes (Dangl and Jones 2001; Piroux et al. 2007). Thus, the regulation of DUF26 containing genes postulated by our analysis of the *Arabidopsis thaliana* transcriptome (Fig. 4 C) might reflect their function in pathogen recognition. Receptor kinases are discussed in more detail in the next section.

The functional category "biotic stress" (Fig. 5 A) comprises a number of different genes which are annotated to be pathogen related.

Nitrilases (Fig. 5 B, upper right) are involved in IAA biosynthesis and catalyze the conversion of indole-3-acetonitrile to IAA. The induction of four *Arabidopsis thaliana* nitrilases by the pathogen *Pseudomonas syringae* has been shown by Bartel and Fink (1994).

Thus, gene selection by unsupervised meta-analysis was able to pinpoint biologically important genes of which many are experimentally validated to be regulated by pathogen attack. Clearly, one could postulate that the remaining genes of unknown function are also associated with responses to pathogen attack.

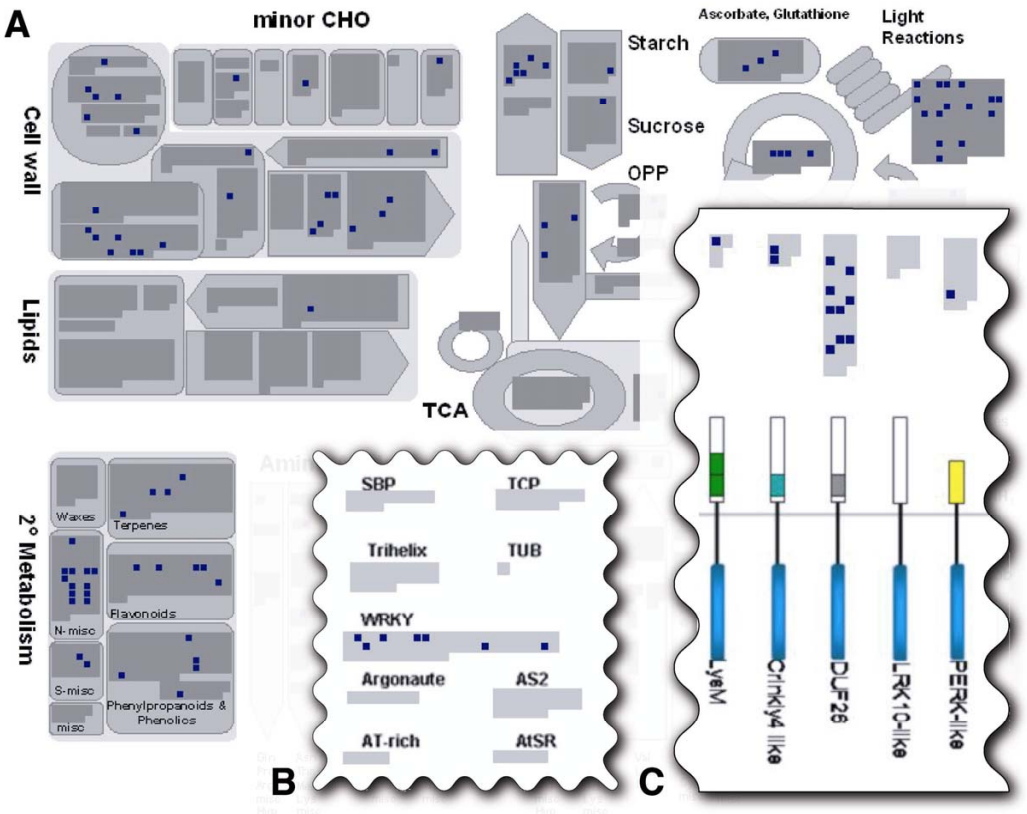


Figure 4: Overview of genes regulated in pathogen associated contrasts. The gray areas inside the individual diagrams of the functional categories represent all genes present on the ATH1 chip. Dark blue squares highlight genes regulated in contrasts of the "pathogen" cluster. Regulation of cell wall genes (upper left), alkaloids which fall into the category "N-misc." of "secondary metabolism" and "Light Reactions" of photosynthesis (upper right) is apparent. B) Part of the "transcription" map indicating regulation of WRKY transcription factors. C) Section of the "receptor like kinases" map indicating regulation of DUF26 kinases. Figure reading example: In subfigure C, a total of 41 DUF26 kinases are represented on the ATH1 chip of which 9 are regulated after pathogen exposure. The figure is based on maps from the pathway analysis program MapMan (Usadel et al. 2005).

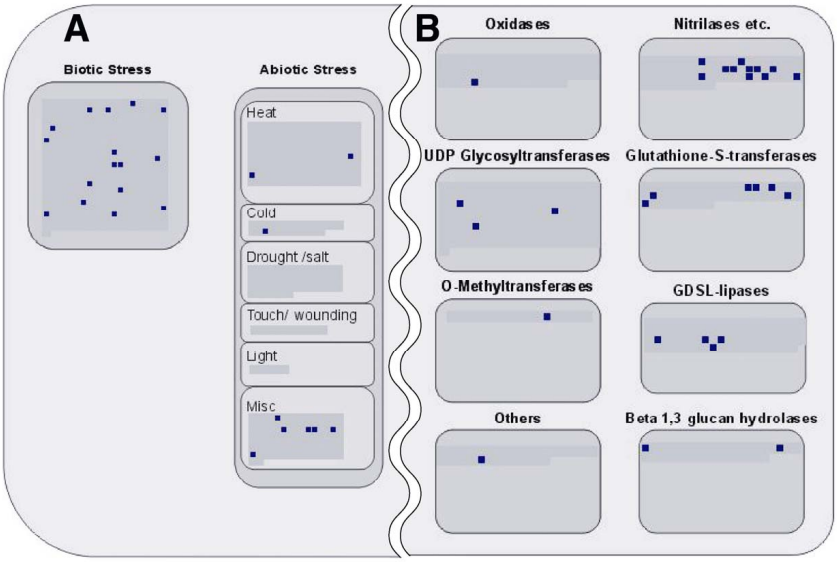


Figure 5: Overview of (A) stress genes and (B) genes of large enzyme families regulated in pathogen-associated contrasts. The gray areas inside the individual diagrams of the functional categories represent all genes present on the ATH1 chip. Dark blue squares indicate regulated genes. Subcategories "Biotic Stress" (A) and "Nitrilases etc." (B) contain a high number of genes regulated after pathogen exposure. The figure is based on maps from the pathway analysis program MapMan (Usadel et al. 2005).

Serine-threonine kinases involved in plant response to pathogens

As presented in Figure 4 C, the extracted set of genes deregulated in response to pathogens includes a number of receptor kinases. Many kinases belong to the group of serine/threonine kinases of the DUF26 subfamily. They all share the same domain composition and order consisting of a signal peptide, an extracellular region containing two domains of unknown function (DUF26, PF01657) and a cytosolic serine/threonine kinase domain (pkinase, PF00069). According to the SMART database (Letunic et al. 2006), proteins of this family are exclusively found in Streptophyta. The 9 putative receptor kinases exhibit high similarity in domain composition and nucleotide sequence with the receptor-like kinase 4 of *Arabidopsis thaliana* (Swiss-Prot-ID Q9C5T0). This enzyme is reported to be a member of the systemic acquired resistance pathway in higher plants. Its expression can be activated by a regulatory protein induced via pathogen and salicylic acid interaction (Du and Chen 2000). Salicylic acid is a signaling molecule which induces systemic acquired resistance in the host plant (Ryals et al. 1996). These findings suggest a function for the putative receptor-like kinases in host defense processes.

Two of the DUF26 kinase genes (At4g21400, At4g21410) were also regulated in the contrasts from dataset GSE3959 and in one contrast from the dataset GSE5770. In the former dataset, the function of B3 domain protein LEAFY COTYLEDON2 (LEC2) was studied. This transcription factor is required for several aspects of embryogenesis

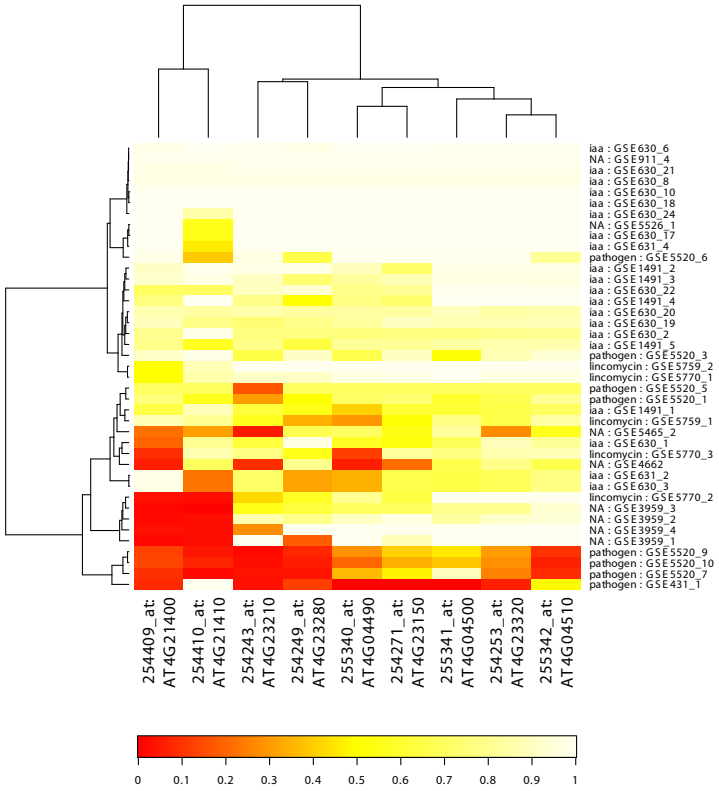


Figure 6: Regulation of DUF26 kinase genes. Red cells indicate low p-values for a gene in a particular contrast, light yellow cells represent high p-values. The DUF26 kinase genes are strongly regulated in four pathogen-associated contrasts.

including the maturation phase. In the latter contrast, *abi4* mutant plants were treated with lincomycin and compared to untreated mutants. ABI4 is a transcription factor, lincomycin inhibits plastid protein translation. From this finding it may be concluded that these two DUF26 kinase genes either play a role in more than one signaling pathway or that the same pathway is used to regulate several functions. This might be an interesting starting point to study these pathways in more detail.

As can be seen from Figure 6, the DUF26 kinase genes were not regulated in all of the contrasts involving pathogen exposure. This could be due to several reasons. For example either the variance in the single microarray intensities was so high that differential expression could not be detected in the contrast or the difference in expression levels (i.e. the logarithmic fold change) was too low to be significant because of biological reasons. Again, this finding might be an interesting starting point to analyze the function and regulation of the DUF26 kinase genes.

Discussion

Public microarray data repositories accumulate large amounts of data which have so far rarely been used for large-scale analyses. Using this wealth of information, additional implications for the function and regulation of genes can be made which could not be derived from single microarray datasets. This stresses the importance of meta-analyses and their benefit over classical microarray experiments.

In this study, we apply a novel approach of an unsupervised meta-analysis on a large number of gene expression microarrays. Before conducting the analysis, we performed a pre-processing which included a conservative outlier removal. Kernel PCA, followed by hierarchical clustering, revealed robust and significant clusters of contrasts which reflect similar experimental conditions. Thus we were able to detect biologically important known and unknown factors (e.g. IAA- or pathogen-associated) through an unsupervised analysis.

To find genes specifically regulated in these clusters, a novel approach of gene selection was conceived. Gene selection was performed using loadings of features on kernel PCA scores, which has to our knowledge not been performed in the context of meta-analysis before. Gene selection based on loadings of features on kernel PCA scores circumvents a major drawback of most proposed methods of feature selection: They tend to find linear combinations of features, i.e. genes, that separate the given experimental classes best (e.g. different cancer types, etc.). This is challenging as the search space for all possible linear combinations is too large to be searched exhaustively and sophisticated heuristics and optimization methods have to be chosen which likely yield differing results, see e.g. Zhang et al. (2006). An unsupervised analysis as proposed here circumvents this problem efficiently by working directly on the loadings from the kPCA analysis. Eigen-decomposition of the kernel matrix is deterministic and so are the results from our gene selection process, provided the projection is capable of clustering the contrasts appropriately. The genes selected by our feature extraction were found to be representative of a group of contrasts and could in part be experimentally validated.

Furthermore, adding random noise to the data did not change the set of selected genes, proving the robustness of the proposed gene selection method.

It is the gene-selection in the first place that benefits most from an analysis across several datasets. Weak regulation signals can easily be overlooked in a single dataset, i.e. the genes will likely receive an insignificant p-value due to their low fold changes compared to a relatively high variance. The situation becomes even worse after a correction for multiple testing has raised the overall p-value level, efficiently removing those subtle signals. In a meta-analysis approach which integrates many datasets, even a small signal that is consistent across several contrasts can be detected. To ensure this surplus and to prevent early losses of information, we used fold changes and not p-values for our analysis. We performed the unsupervised meta-analysis on absolute fold changes to reduce variation introduced by different experimental settings. For example, when there are contrasts in the dataset which compare a surplus of a factor with a control and other contrasts comparing a lack of a factor with another control, we might expect fold changes with opposite signs but still want the contrasts to cluster closely together because the same factor was studied in both. In some cases the direction of the experimental setup was not even apparent from the description of the dataset.

To ensure that results of similar quality could not be obtained by a simpler model and thus to prevent overfitting of the data we compared the results to the ones obtained from traditional linear PCA. Even though linear PCA was also able to detect some of the major clusters in principle, its accuracy as assessed by hierarchical clustering as well as by the gene selection process fell far short of the results from the kernelized version. Additionally, it should be noted that kernel PCA outperforms the traditional approach significantly, considering that the dimension of the kernel matrix as a matrix of pairwise scalar products between the data points is independent of the dimension of the data, which is 22810 (the number of probe sets) in the case of the ATH-1 arrays.

For a large *Arabidopsis thaliana* microarray dataset, we demonstrate here that gene selection, based on the study of principal components, proposed genes typical for either IAA- or pathogen-associated contrasts. These genes were proved to be related to either IAA effects or plant reactions in response to pathogen exposure by previous studies. Furthermore, starting from our finding that DUF26 kinases are regulated in pathogen-associated contrasts, we applied homology modeling to propose that DUF26 kinases have a function in plant pathogen defense. Further experiments are needed to confirm this hypothesis. Nonetheless, this example demonstrates how unsupervised analysis can aid and guide the next steps of such an analysis.

In general, unsupervised meta-analysis embracing several highly divergent experimental settings can suggest novel gene functions by revealing the regulation of a gene under different conditions. It is noteworthy that these analyses are not restricted to datasets addressing the same topic, but that they profit from the divergence of the experimental settings.

However, it has to be mentioned that an unsupervised meta-analysis is suggestive rather than definitive. But since it is common in classical statistics to precede a supervised, parametric analysis with an explorative approach to check the integrity and

quality of the data, we recommend the same here for microarray meta-analyses. Hypotheses from unsupervised analyses can then be tested with supervised methods and biological experiments.

We have shown here that it is feasible to integrate various datasets spanning a large range of experimental questions and originating from various laboratories into a coherent unsupervised analysis. This analysis can be applied to find genes representative of a cluster of related contrasts. Based on expression changes between clusters, the function and regulation of genes can be predicted. Our study is based on the Affymetrix ATH1 Genome Array platform here, but our approach can be transferred to any platform, organisms and experimental design which allows one to compute a logarithmic fold change, e.g. human or mouse microarray datasets. To achieve easy access to our unsupervised meta-analysis results, we intend to set up a database web server where new datasets can easily be added and compared to our curated database of *Arabidopsis thaliana* ATH-1 microarrays.

Availability

R code is available on request from the authors.

Acknowledgments

We gratefully acknowledge the funding from the Impuls- & Vernetzungsfonds Helmholtz-Gemeinschaft deutscher Forschungszentren e.V. (VH-VI-023), BMBF Projekt FUNCRYPTA (FKZ 0313838B), DFG (SPP 1150 / Da208/7-1 / TR 34-TPA5) and Land Bayern (Foringen TP D1, IZKF B-36). We would like to thank Dr Alan Horowitz (Dep. of Bioinformatics, Univ. of Würzburg) and Dr Biju Joseph (Dep. of Hygiene and Microbiology, University of Würzburg) for proof-reading the manuscript.

References

Antonov, A.V., Tetko, I.V., Mader, M.T. et al. 2004. Optimization models for cancer classification: extracting gene interaction information from microarray expression data. *Bioinformatics*, 20:644–652.

Barrett, T., Troup, D.B., Wilhite, S.E. et al. 2007. NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res*, 35:D760–D765.

Bartel, B. and Fink, G.R. 1994. Differential regulation of an auxin-producing nitrilase gene family in *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A*, 91:6649–6653.

Benjamini, Y. and Hochberg, Y. 2000. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, 25:60–83.

Berger, S., Papadopoulos, M., Schreiber, U. et al. 2004. Complex regulation of gene expression, photosynthesis and sugar levels by pathogen infection in tomato. *Physiologia Plantarum*, 122:419–428.

Box, G.E.P. and Cox, D.R. 1964. An analysis of transformations. *Journal Of The Royal Statistical Society Series B-Statistical Methodology*, 26:211–252.

Cheong, Y.H., Chang, H.S., Gupta, R. et al. 2002. Transcriptional profiling reveals novel interactions between wounding, pathogen, abiotic stress, and hormonal responses in *Arabidopsis*. *Plant Physiol*, 129:661–677.

Choi, J.K., Yu, U., Kim, S. et al. 2003. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, 19:84–90.

Conlon, E.M., Song, J.J. and Liu, J.S. 2006. Bayesian models for pooling microarray studies with multiple sources of replications. *BMC Bioinformatics*, 7:247.

Craigon, D.J., James, N., Okyere, J. et al. 2004. NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. *Nucleic Acids Res*, 32:D575–D577.

Dangl, J.L. and Jones, J.D. 2001. Plant pathogens and integrated defence responses to infection. *Nature*, 411:826–833.

Dixon, R.A. 2001. Natural products and plant disease resistance. *Nature*, 411:843–847.

Du, L. and Chen, Z. 2000. Identification of genes encoding receptor-like protein kinases as possible targets of pathogen- and salicylic acid-induced WRKY DNA-binding proteins in *Arabidopsis*. *Plant J*, 24:837–847.

Everitt, B. 2005. An R and S-PLUS Companion to Multivariate Analysis. Springer-Verlag London Limited.

Gentleman, R.C., Carey, V.J., Bates, D.M. et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5:R80.

Guyon, I., Weston, J., Barnhill, S. et al. 2002. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422.

Hall, M. 1999. Correlation-based Feature Selection for Machine Learning. Ph.D. thesis, Hamilton NZ: Waikato University, Department of Computer Science.

Hu, P., Greenwood, C.M.T. and Beyene, J. 2005. Integrative analysis of multiple gene expression profiles with quality-adjusted effect size models. *BMC Bioinformatics*, 6:128.

Huttenhower, C., Hibbs, M., Myers, C. et al. 2006. A scalable method for integration and functional analysis of multiple microarray datasets. *Bioinformatics*, 22:2890–2897.

Karatzoglou, A., Smola, A., Hornik, K. et al. 2004. kernlab - an S4 package for kernel methods in R. *Research Report Series / Department of Statistics and Mathematics*.

Letunic, I., Copley, R.R., Pils, B. et al. 2006. SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res*, 34:D257–D260.

Moreau, Y., Aerts, S., Moor, B.D. et al. 2003. Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends Genet*, 19:570–577.

Ng, A., Jordan, M. and Weiss, Y. 2001. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, 14.

Parkinson, H., Kapushesky, M., Shojatalab, M. et al. 2007. ArrayExpress–a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res*, 35:D747–D750.

Piroux, N., Saunders, K., Page, A. et al. 2007. Geminivirus pathogenicity protein C4 interacts with *Arabidopsis thaliana* shaggy-related protein kinase AtSKeta, a component of the brassinosteroid signalling pathway. *Virology*, 362:428–440.

R Development Core Team 2004. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ryals, J.A., Neuenschwander, U.H., Willits, M.G. et al. 1996. Systemic acquired resistance. *Plant Cell*, 8:1809–1819.

Schölkopf, B., Smola, A. and Müller, K.R. 1998. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10:1299–1319.

Shawe-Taylor, J. and Cristianini, N. 2004. Kernel Methods for Pattern Analysis. Cambridge University Press.

Smyth, G.K. 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3:Article 3.

Stepanova, A.N., Hoyt, J.M., Hamilton, A.A. et al. 2005. A Link between ethylene and auxin uncovered by the characterization of two root-specific ethylene-insensitive mutants in *Arabidopsis*. *Plant Cell*, 17:2230–2242.

Suzuki, R. and Shimodaira, H. 2006. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, 22:1540–1542.

Tamayo, P., Slonim, D., Mesirov, J. et al. 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A*, 96:2907–2912.

Thilmony, R., Underwood, W. and He, S.Y. 2006. Genome-wide transcriptional analysis of the *Arabidopsis thaliana* interaction with the plant pathogen *Pseudomonas syringae* pv. *tomato* DC3000 and the human pathogen *Escherichia coli* O157:H7. *Plant J*, 46:34–53.

Usadel, B., Nagel, A., Thimm, O. et al. 2005. Extension of the visualization tool MapMan to allow statistical analysis of arrays, display of corresponding genes, and comparison with known responses. *Plant Physiol*, 138:1195–1204.

Ward, J.H. 1963. Hierarchical grouping to optimize an objective function. *Journal Of The American Statistical Association*, 58:236–&.

Wu, J., Irizarry, R. and with contributions from James MacDonald and Jeff Gentry 2005. gcrma: Background Adjustment Using Sequence Information. R package version 2.2.1.

Zhang, X., Lu, X., Shi, Q. et al. 2006. Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics*, 7:197.

Chapter 3

Is gene activity in plant cells affected by UMTS irradiation? A whole genome approach.

Is gene activity in plant cells affected by UMTS-irradiation?

A whole genome approach.

Julia C. Engelmann[3*], Rosalia Deeken[1*], Tobias Müller[3], Günter Nimtz[2], M. Rob G. Roelfsema[1] and Rainer Hedrich[1§]

[1]Molecular Plant Physiology and Biophysics, Julius-von-Sachs Institute for Biosciences, Biocenter, University of Würzburg, Julius-von-Sachs-Platz 2, D-97082 Würzburg, Germany

[2]Institute of Physics II, University of Cologne, Zülpicher Straße 77, D-50937 Cologne, Germany

[3]Department of Bioinformatics, Biocenter, University of Würzburg, Am Hubland, D-97074 Würzburg, Germany

* Theses authors contributed equally to this work

[§]Corresponding author:

hedrich@botanik.uni-wuerzburg.de

phone: +49 931 888-6100

fax: +49 931 888-6157

Running header: Effects of UMTS-irradiation on plant cells

Keywords: suspension cultured plant cells, radio frequency electromagnetic fields, microarrays, *Arabidopsis thaliana*

-Submitted to "Computational Biology and Chemistry: Advances and Application"-

Abstract

Mobile phone technology makes use of radio frequency (RF) electromagnetic fields transmitted through a dense network of base stations in Europe. Possible harmful effects of RF fields on humans and animals are discussed, but their effect on plants has received little attention. In search for physiological processes of plant cells sensitive to RF fields, cell suspension cultures of *Arabidopsis thaliana* were exposed for 24 h to a RF field protocol representing typical microwave exposition in an urban environment. mRNA of exposed cultures and controls was used to hybridize Affymetrix-ATH1 whole genome microarrays. Differential expression analysis revealed significant changes in transcription of 10 genes, but they did not exceed a fold change of 2.5. Besides that 3 of them are dark-inducible, their functions do not point to any known responses of plants to environmental stimuli. The changes in transcription of these genes were compared with published microarray datasets and revealed a weak similarity of the microwave to light treatment experiments. Considering the large changes described in published experiments, it is questionable if the small alterations caused by a 24 h continuous microwave exposure would have any impact on the growth and reproduction of whole plants.

Introduction

The use of radio frequency (RF) electro magnetic fields in mobile phone technology has led to a discussion on possible harmful effects on humans and animals (Scientific Committee on emerging and newly identified health risks (SCENIHR) 2006; European Commission-Research Directorate-General-European Communities 2005). A number of studies suggested that RF fields can affect living organisms by increasing the occurrence of brain tumors (Hardell et al 2005) and leukemia (Hocking et al 1996). Comparable studies, however, did not confirm these results and the possibility of carcinogenic risks imposed by these non-ionizing electromagnetic fields therefore is still a matter of debate (Moulder et al 2005). In contrast to ionizing radiation, it is unclear how non-ionizing fields can trigger physical events that will affect small biological structures such as organelles (Adair 2003). The energy absorbed by organelles or small cells from RF fields seems to be too small to force changes in their physiology. However, larger biological structures may sense weak electrical fields. This is obvious from the electroreceptors found in a number of fish species, such as sharks and rays, which enables them to communicate or localize their prey (Hopkins 1995; Kalmijn 1966). Likewise, migrating birds are sensitive to the earth magnetic field, using a sensory system that probably involves cryptochrome blue light receptors (Mouritsen and Ritz 2005). The latter group of photo-receptors is also found in plants (Cashmore 2003) and an effect of electromagnetic fields on cells of animals and plants therefore should not be ruled out, a priori.

In comparison to humans and animals, the possible effect of RF fields on plants has received very little attention. In a study on cuttings of *Tradescantia*, increased numbers

of micronuclei were determined, suggesting that RF fields enhanced breakage of DNA strands (Haider et al 1994). In search for possible targets of high frequency electromagnetic fields in plant cells, we undertook a whole genome approach. Many cellular processes will feed in on gene regulation and thus will alter gene activity. In case the electromagnetic fields used in mobile phone technology alter such a cellular process, it is likely that gene activity is also altered. The activity of approximately 23.000 genes in *Arabidopsis thaliana*, the model plant for molecular biology, can be determined with the Affymetrix ATH1 genome microarray. The application of microarrays thus provides a means to identify possible molecular targets of RF electromagnetic fields in plants.

Materials and Methods

Growth of cell culture

Arabidopsis thaliana suspension-cultured cells were derived from a callus culture originally gained from Col-0 seeds (Deeken et al 2003) and grown in media containing 1x MS+MES salts (Duchefa, Haarlem, The Netherlands), 0.56 mM myo-inositol, 0.1 mM $FeSO_4$, 0.13 mM EDTA, 2.26 µM 2,4-Dichlorophenoxyacetic acid, 4.06 µM nicotinic acid, 2.5 µM pyridoxal hydrochloride, 0.3 µM thiamine hydrochloride and 2% D-sucrose, pH 5.7. The suspension-cultured cells were grown at 26°C on a rotary shaker (140 rpm) and sub-cultured weekly by transferring 20 ml cells into 50 ml fresh medium.

Exposition of suspension-cultured cells to electromagnetic fields

For the irradiation experiment in which microwave exposition in an urban environment was simulated, a stock of suspension-cultured cells was divided into fourteen 50 ml sub-

cultures that were kept in 250 ml Erlenmeyer flasks. After one day, eight of these sub-cultures were transferred to a temperature-controlled dark room at 25°C. All eight sub-cultures were placed on a single rotary shaker (type 3015, GFL, Burgwedel, Germany) rotating with 140 rpm (*Figure 1*). The rotary shaker was covered with reflection free absorber in order to avoid standing wave patterns or magnetic fields, which may be caused by the motion of the rotary shaker. Extremely low frequency fields (ELF) and their magnetic components were found below the 50 Hz noise level in the laboratory. Four sub-cultures were positioned in the far field of an antenna that irradiated microwaves with a frequency of 1.9 GHz UMTS (universal mobile telecommunication system) modulation. The UMTS electromagnetic field was produced by the following equipment: A Signal Generator (SMIQ 03B, Rhode & Schwarz, München, Germany) operating at 300 kHz - 3.3 GHz and a pulse modulator at 5 MHz (Model 184, Wandel & Goltermann, Eningen, Germany). The operation modus was FDD and a periodic modulation CDMA at a carrier frequency of 1.9 GHz. There was one control channel with a 1.5 kHz modulation and 6 data channels. The power supply of the signal was controlled by a computer, which simulated a scenario in an urban environment (Bilz et al 2001). In this scenario, there was a 3 dB up and down power modulation for 45 s and during 15.3 s there was a 30 dB periodic attenuation, resulting in a total period of 60.3 s. The RF field had an average power of 8 mW/cm^2 and a peak power of 20 mW/cm^2, measured at the samples' locality with an EM radiation monitor (EMR-20, Wandel & Goltermann). During the periodical exposure time, the peak power was transmitted for 37.5 % of the time. The total time of exposure was 24 h. The wavelength was much larger than the sample size and the bottle walls, therefore the irradiated inhomogeneous dielectric system behaved as an effective medium. The effective electric field is

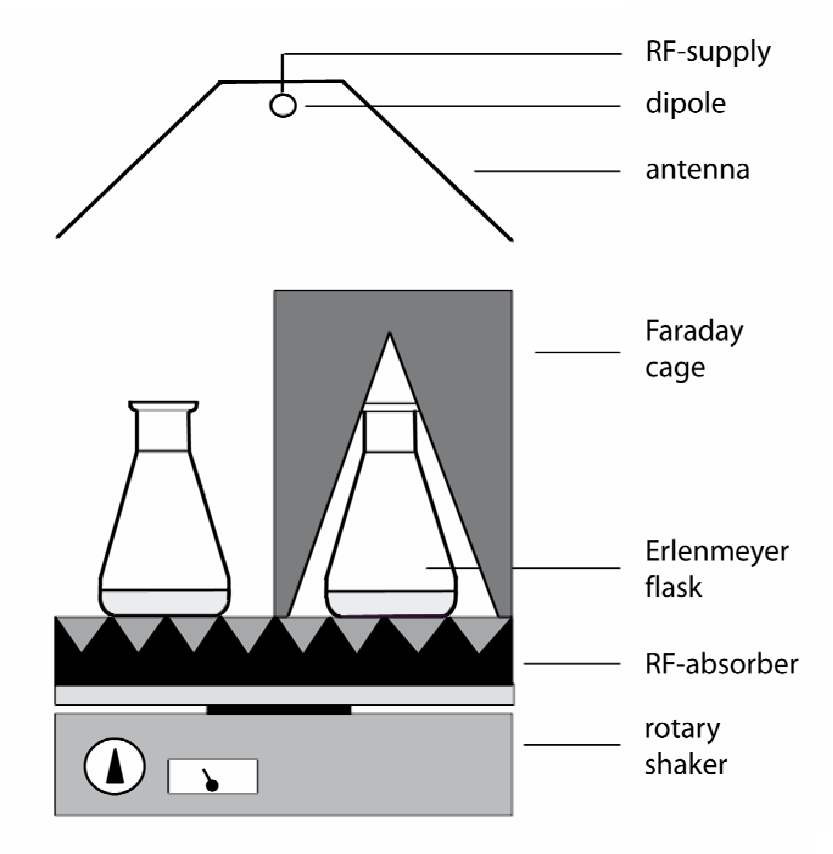


Figure 1. Schematic representation of the experimental setup used for UMTS field exposition of suspension-cultured cells.

The distance between the dipole antenna and the sample solution was 1 m. The dipole was placed in front of a metallic reflector. The linearly polarized microwave carrier frequency of 1.9 GHz was modulated with a special UMTS signal (Bilz et al. 2001).

therefore some percentage higher in the microscopic dielectric heterostructure than it would be in the bulk material. Four control sub-cultures were shielded from electromagnetic fields by a Faraday cage attenuating the field with >> 30dB (*Figure 1*). The aluminium cage was wrapped by an anti-reflecting layer to avoid reflection and thus suppress standing wave patterns. Taking into consideration a refractive index of 9 at this frequency and an absorption coefficient of 0.5 1/cm, the peak and the average

SAR values are 2 and 0.75 W/kg, respectively. A possible rise in the liquid temperature during exposure was ruled out by comparing additional flasks with culturing medium with a liquid-based thermometer.

The remaining six sub-cultures were divided into two groups, one of which was tested for sensitivity to 50 µM abscisic acid, while the others were kept as controls. The latter six cultures were incubated for 3 h in a growth chamber on a rotary shaker at 25°C and 140 rpm. All *Arabidopsis thaliana* cultures were quickly harvested on a filter paper, frozen in liquid nitrogen and stored at -80°C.

RNA-extraction, microarray hybridization and quantitative RT-PCR

RNA-extraction and digestion of contaminating DNA was carried out with the Plant RNeasy Extraction kit (Qiagen, Hilden, Germany). The hybridization of a total of eight microarrays (ATH1) was performed according to the manufacturer's protocol (Affymetrix, Santa Clara, CA, USA). Four arrays were hybridized with RNA from microwave treated and another four arrays with RNA from microwave shielded sub-cultured *Arabidopsis* suspensions-cultured cells. For each array, RNA extracted from one sub-culture was used for hybridization, resulting in four replicates for each treatment group.

For quantitative real time RT-PCR, the contaminating DNA was digested using RNase-free DNase (Amersham, Freiburg, Germany) according to the manufacturer's protocol. First-strand cDNA was prepared using the M-MLV-RT kit (Promega, Mannheim, Germany) and diluted for PCR 20-fold with water. Quantitative PCR was performed in a LightCycler (Roche, Mannheim, Germany) with the LightCycler-Fast Start DNA Master SYBR Green I kit (Roche, Mannheim, Germany). The following primers were used: *AtACT*fwd (5'-GGT GAT GGT GTG TCT), -rev (5'-ACT GAG CAC AAT GTT AC);

*At3g47340*fwd (5'-ACT CTG CGA GAC TAA C), -rev (5'-CAA AAC ACT TCA CCC A); *At3g15460*fwd (5'-GAT TTA GCA CAG CCT T), -rev (5'-ACT GTA TGT TTC TAG GG); *At4g39675*fwd (5'-TTG GAG CAA GTT ACG C), -rev (5'- CGA CCA AGA TAC GTT T); *At4g26260*fwd (5'- GTG CAT TTG ATG AAT CT), -rev (5'- GTA GTA AGG CTT GAC C); *AtCg00630*fwd (5'- ATA TCT TTC CGT AGC A), -rev (5'- AGG GAA ATG TTA ATG C); *At3g60140*fwd (5'- AGG ATA TTA CGC ATG G), -rev (5'-CAA AGG AGC AAC GAT TA); *At3g24500*fwd (5'-AGT AAC ACA AGA CTG G), -rev (5'-ACA GCC TGA TTA GGA A); *At5g10040*fwd (5'- GTG AAT ACA ACG GCA G), -rev (5'- GGT GAT TAG AGA AGC AA); *AtCg00120*fwd (5'- AAG CTA TGA AAC AGG T), -rev (5'- CTT GGT AGA GGC TAT GA). All mRNA quantifications were normalized to 10,000 molecules of actin cDNA fragments amplified by AtACTfwd and AtACTrev. Each type of transcript was quantified by using its individual standard. In order to detect contaminating genomic DNA, quantitative RT-PCR was performed with the same RNA template used for cDNA synthesis. To compute a p-value for the fold changes of each gene, the Student's t-test was applied on the normalized transcript numbers from quantitative RT-PCR.

Normalization of microarray data

The microarray data were analyzed using the Bioconductor software (Gentleman et al 2004) designed for genomic data analysis running under the statistical programming environment R (Ihaka and Gentleman 1996). To obtain a normalized gene expression value from Affymetrix probe intensities for each gene of each microarray, variance stabilization (VSN) within the Bioconductor software (Gentleman et al 2004; Huber et al 2002) was applied. As recommended in the VSN manual, no background correction was performed on the Affymetrix probe intensities prior to VSN-normalization. Only the perfect match (PM) probes were used to compute an expression value for each gene.

For summarization of probe intensities into gene expression values, the median polish algorithm was applied which is also incorporated in the commonly used Robust Multiarray Analysis (RMA) (Irizarry et al 2003).

Correspondence Analysis

Correspondence analysis (CA) was conducted using the R package *MASS* (Venables and Ripley 2002). It was applied on the data matrix of 22810 genes (in the rows) and 6 array samples (in the columns). We used CA to project the vectors of array samples into a lower-dimensional subspace (typically two dimensions) that accounts for the main variance in the data, in a way that distances among points reflect their original distances in the high-dimensional space as closely as possible (Fellenberg et al 2001). The same reduction of dimensions was carried out for all genes at the same time. In the CA graph, dissimilar objects are separated along the component axes while similar objects cluster close to each other.

Hierarchical cluster analysis

Hierarchical cluster analysis was performed in R using the *stats* package (Venables 2002). We applied complete linkage clustering on Euclidian distances between objects to form hierarchical cluster trees. The bootstrapping algorithm for judging the robustness of the estimated tree was programmed in R as described by Efron and Tibshirani (1993). To calculate bootstrap values, 100 single trees were calculated drawing genes uniformly with replacement from the selected genes. In this procedure one gene may appear more than once while others do not appear at all. The function "consense" of the PHYLIP software (Felsenstein 1989) was applied to calculate a consensus tree with bootstrap values out of the single trees. The bootstrap value

indicates how often each split was found in the single trees indicating the strength of the cluster signal to separate the groups (here: arrays). In principle, the procedure described above is equivalent to the well-known bootstrap method in phylogenetic analysis. Here, microarray hybridizations represent sequences and genes replace the sites of the multiple sequence alignment (Efron et al 1996).

Differential expression of genes

Differential expression of genes between microwave exposed and control cultured cells was performed by applying a moderate t-statistic implemented in the Linear Models for Microarray data package (*limma* (Smyth 2004)) which is part of the Bioconductor software project. The linear models were fitted on the expression values of each gene with the factor "microwave-exposure" or "no treatment". The function *eBayes* was used to compute moderated t-statistics by empirical Bayes shrinkage of the standard errors towards a common value. The null hypothesis of differences between treatments being equal to zero was tested under the assumption of independent errors following a normal distribution. For each gene, a fold change and a p-value measuring the statistical significance of differential expression was calculated. P-values were corrected for multiple testing by applying "False Discovery Rate" (FDR) (Benjamini and Hochberg 1995).

Comparison of different ATH1-microarray experiments using Principal Component Analysis

Principal Component Analysis (PCA) was applied to compare the microwave dataset with other *Arabidopsis thaliana* microarray datasets of several categories, available from Genevestigator online (Zimmermann 2004). Since all microarray datasets stored in Genevestigator are normalized with the MAS5 algorithm (Affymetrix 2002), the

microwave dataset was also normalized with this algorithm to achieve comparability, but these values were only used for PCA. Principal Component Analysis was used to reduce the dimensionality of the dataset without a significant loss of information to better recognize patterns in the data (Jolliffe 1986). The top ten genes with lowest p-values of the microwave dataset were selected and their fold changes were compared to the fold changes of the same genes in the Genevestigator datasets. Therefore the vectors of 10 genes of each microarray dataset were projected into two dimensions which contain the main variance of the data. Thus, each experiment was represented by one point in a two-dimensional space. PCA was performed in R using functions from the *stats* package (Venables 2002).

Results

UMTS irradiation and preliminary data analysis

For this study, a single batch of suspension-cultured cells of *Arabidopsis thaliana* was used, providing a homogeneous starting material. The starting batch was divided into sub-cultures, to ascertain a minimal degree of biological variation between control and RF-exposed sub-cultures. Because of the identical starting cultures, a maximal sensitivity for stimulus-induced changes in transcription was obtained. This experimental approach thus allowed the detection of very small changes in transcription. Such small transcriptional changes may be superimposed by natural variation, in case of cell suspensions cultured separately or in experiments carried out with whole plants. Four of the sub-cultures were exposed for 24 h to microwaves with a frequency of 1.9 GHz, a field strength considerably higher than the international recommended exposure for UMTS mobile communication (1 mW/cm^2, (International Commission on Non-

Ionizing Radiation Protection 1998)). The other four sub-cultures were shielded from the RF field and served as controls for UMTS exposure. In addition to the first eight sub-cultures, six sub-cultures were divided into two groups, of which one group was tested for responsiveness to stress signals by exposing them to the stress hormone abscisic acid (50 μM) and the other three served as controls for the hormone treatment. Then, transcript numbers of the potassium channel gene *GORK*, which has been shown to be very sensitive to abscisic acid treatment (Becker et al 2003), were quantified applying real time RT-PCR. After an incubation period of 3 h, abscisic acid induced an 11-fold increase in the transcript number of the GORK gene. This indicated that the suspension-cultured cells used for the microwave experiment were sensitive to stress signals.

After termination of the RF field exposure, the analysis was carried out blinded, the code on the cultures was neither known by the experimenters handling the samples nor by those that performed the initial data analysis. A first analysis of the data indicated that the hybridization procedure had failed for two of the eight microarrays. Since this was due to technical problems, the RNA from these samples was hybridized to two new microarrays. To avoid any impact of differences due to hybridization conditions, the newly hybridized arrays were excluded from the initial analysis. At this point of analysis, at which the grouping was still unknown, the data of all genes of an array were incorporated and possible small changes caused by altered hybridization conditions thus would have caused a loss of sensitivity.

Cluster analysis reveals grouping of microwave treated and untreated samples

At the beginning of our analysis, the grouping of the microarray hybridizations was still unknown to the data analysts. In order to uncover the so far unknown “group labels” of

the 6 remaining microarrays (Arrays 1 and 5 were left out), a correspondence analysis (CA) was performed with all genes of the Affymetrix microarray. When all genes were taken into account, a separation of arrays into two distinct groups along the first or second component axis was not found (*Figure 1S*). This indicates that the electromagnetic fields did not alter the expression of the majority of genes.

Since a major effect of microwave exposition on the transcription levels of *Arabidopsis thaliana* genes could not be found by CA, in the next step it was studied whether microwaves had a notable effect on the expression of a small number of genes. To perform a hierarchical cluster analysis, genes were arranged according to the degree of variance in signal intensity between the 6 arrays. The variation in signal intensity might come from differences between the microwave-treated and untreated RNA-samples or from variation that is unrelated to this grouping. In case of an influence of RF fields, differentially expressed genes should be among the most variant genes and hierarchical clustering should result in a clear separation between these groups. In the case of no differential expression, a clear split between microwave treated and untreated samples should not be found.

In the first step of this hierarchical cluster analysis, the two genes with the highest variance were used to construct a hierarchical cluster tree of microarray samples and in each following step one gene was added. In case the transcription levels of these genes would vary randomly over the microarray measurements, frequent changes in the cluster tree topology would be expected when adding more genes to the dataset. The analyses of 2 to 20 genes consistently revealed the same clear split between the cluster of arrays 3 and 7, and the remaining four arrays (*Figure 2 A*). In case 21 to 30 genes were used for the analyses, no consistent group of two arrays could be detected. From the clear

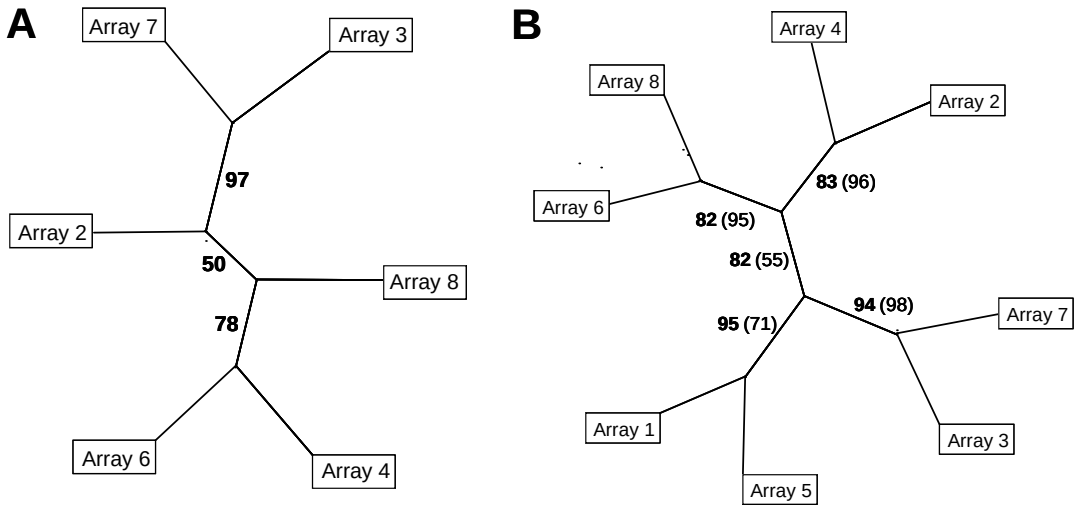


Figure 2. Hierarchical cluster trees of arrays hybridized with cDNA of control (uneven numbers) or microwave-exposed (even numbers) samples

Numbers on the edges indicate bootstrap values. **A** Clustering of 6 arrays using 10 genes with highest variance in signal intensity. **B** Clustering of 8 arrays (6 original arrays and 2 arrays hybridized later), bootstrap values from trees with 10 genes are given in bold numbers, those of 20 genes in normal numbers in parentheses.

split between arrays 3 and 7, and the remaining arrays when using 2 up to 20 genes (*Figure 2 A*), it was concluded that arrays no. 3 and 7 had been probed with different samples than the other four arrays. This grouping of arrays 3 and 7 versus the remaining arrays must have been due to genes differentially expressed between the two sample groups. These differentially expressed genes were among the uppermost variant genes. Adding more genes to the dataset eventually diluted the signal until it disappeared when using more than 20 genes for the hierarchical cluster tree. Therefore, the robustness of the hierarchical cluster tree was examined with a bootstrap algorithm based on the 10 most variant genes. This revealed a cluster of arrays 3 and 7 separated from the other arrays in 97 of 100 cluster trees, indicating a strong difference between both sets of

arrays considering those 10 genes (*Figure 2 A*, bold numbers on the lines). As expected, the separations between the remaining 4 arrays were less clear, indicating a stronger similarity of these arrays (*Figure 2 A*). The hierarchical cluster analysis was repeated, incorporating the data of arrays 1 and 5, which were hybridized later than the other 6 arrays. Again, the 10 genes with highest variance (*Table 1*) were used for constructing a hierarchical cluster tree with bootstrap values (*Figure 2 B*).

Table 1. Genes with the highest variance in expression signals

The variance was determined over all 8 arrays hybridized with control or microwave exposed samples. The fold change and corresponding p-values are given for the microarray assay as well as for quantitative real time RT-PCR.

* The Affymetrix probeset for this gene also hybridizes with At3g15450.

	AGI Code	Fold change microarray	p-value microarray	Fold change RT-PCR	p-value RT-PCR	**Protein function**
1.	At3g47340	0.4	$0.91\ 10^{-4}$	0.4	0.05	**glutamine-dependent asparagine synthetase**
2.	At3g15460*	0.5	$0.27\ 10^{-2}$	0.5	0.09	**brix domain protein**
3.	AtCg00590	1.7	0.22	n.d.	n.d.	**orf31 hypothetical protein**
4.	At4g39675	1.5	0.30	1.9	0.14	**expressed protein**
5.	At4g26260	0.5	$0.12\ 10^{-3}$	0.3	0.04	**protein similar to myo-inositol oxygenase**
6.	AtCg00630	1.5	0.28	0.9	0.70	**PSI J protein (chloroplast)**
7.	At3g60140	0.6	$0.68\ 10^{-3}$	0.6	0.19	**beta-glucosidase-like protein**
8.	At3g24500	1.1	0.73	1.0	0.92	**ethylene-responsive transcriptional coactivator**
9.	At5g10040	1.4	0.31	1.4	0.17	**expressed protein**
10.	AtCg00120	1.4	0.28	0.9	0.73	**ATPase alpha subunit (chloroplast)**

A partition into evenly (arrays 2, 4, 6 and 8) and unevenly (arrays 1, 3, 5 and 7) numbered arrays was found, which reflected the true sample grouping. It was supported

by a bootstrap value of 82. The bootstrap values dropped when the tree was based on the signals of 20 genes (*Figure 2 B*, numbers in parentheses). Apparently, the two clusters of arrays were found as long as only a small group of genes with a high variance was taken into account. Since the hierarchical cluster analysis correctly identified two distinct groups of arrays, their code was disclosed. Microarray samples with even numbers had been hybridized with RNA of microwave-treated cell cultures and those with uneven numbers represented the untreated controls.

Quantitative RT-PCR analysis confirms gene expression changes

The differences in microarray signals of the genes which were used in the hierarchical cluster analysis (*Table 1*) could reflect either biological meaningful differences in transcript numbers between the microwave-exposed and control samples, or technical variations due to slightly differing hybridization properties of the arrays. The transcript numbers of the 10 genes in *Table 1* were determined with a second technique. For 9 of the 10 genes listed, the fold change in transcription number was measured applying real time RT-PCR and tested for significance with a student's t-test. No PCR product could be obtained for ORF 31 using several primer pairs designed after the published sequence (TAIR-database). Three out of four significant changes in transcription ($p < 0.05$) observed with microarrays, were confirmed with quantitative RT-PCR (At4g26260, At3g47340, At3g15460; *Table 1*). However, the degree of variation was higher with the latter method and revealed p-values < 0.05 only for two genes (At4g26260, At3g47340; *Table 1*). Although the third gene (At3g15460) had a non-significant p-value ($p = 0.09$), we considered it confirmed claiming that the higher p-value is due to higher variance in the qRT-PCR measurements.

Independent from p-values, agreement between the microarray assay and quantitative RT-PCR, can be seen when ordering the genes measured by qRT-PCR by their p-value: The first three genes with smallest p-values (At4g26260, At3g47340; At3g15460, *Table 1*) are among the 4 most significant differentially expressed genes in the microarray measurements (*Table 2*).

Table 2. Genes with most significant p-values ($p < 0.05$)

Fold changes and corresponding p-values for genes differentially expressed between microwave exposed and control samples in the microarray assay.

* The Affymetrix probesets for these genes also hybridize with At3g15450 (1) and At5g34780 (2).

Nr.	AGI Code	Fold change microarray	p-value	**Protein function**
1.	At3g47340	0.4	$0.91\ 10^{-4}$	**glutamine-dependent asparagine synthetase 1**
2.	At4g26260	0.5	$0.12\ 10^{-3}$	**protein similar to myo-inositol oxygenase**
3.	At3g60140	0.6	$0.68\ 10^{-3}$	**beta-glucosidase-like protein**
4.	At3g15460*[1]	0.5	$0.26\ 10^{-2}$	**brix domain protein**
5.	At1g62480	0.6	$0.66\ 10^{-2}$	**vacuolar calcium-binding protein-related**
6.	At1g15380	0.8	0.010	**lactoylglutathione lyase family protein**
7.	At1g21400*[2]	0.8	0.027	**putative 2-oxoisovalerate dehydrogenase**
8.	At1g80160	0.8	0.027	**lactoylglutathione lyase family protein**
9.	At2g05540	0.7	0.027	**glycine-rich protein**
10.	At4g35770	0.7	0.027	**senescence-associated protein**

A small number of genes is differentially expressed between microwave-treated and shielded samples

After disclosing the group labels of the microarray samples which had been correctly predicted by hierarchical cluster analysis, differential expression could be analyzed.

Using a moderate t-test, the genes were tested for differential expression between the four microwave-treated and four untreated microarray samples. This revealed 3 genes that were highly significant differentially expressed ($p < 0.001$), 2 genes significant differentially expressed ($p < 0.01$) and 5 genes weakly significant differentially expressed ($p < 0.05$) (*Table 2*) after multiple testing correction.

To further confirm that differentially expressed genes exist in the microarray dataset of microwave-treated and untreated samples, the distribution of uncorrected p-values was analyzed and contrasted to the distribution of uncorrected p-values of a random grouping of arrays into two groups. For the random grouping, the array dataset was split into two groups irrespective of microwave treatment and tested for differential gene expression. In this case, the analysis revealed no significant differentially expressed genes. This finding is confirmed by the distribution of uncorrected p-values (see *Figure 2S*). In case of no differential expression, uncorrected p-values follow a uniform distribution (Wassermann 2004). This can be observed for random sample groupings irrespective of microwave treatment (*Figure 2S A*). However, for the correct sample grouping into microwave treated and untreated microarray samples, the p-value distribution differs from the uniform distribution, having a higher number of genes at low p-values, indicating differential expression (*Figure 2S B*).

Comparison of significant genes with other gene expression datasets

In order to dissect stimuli acting in a similar manner on the activity of these genes and since the physiological role of most of the genes is not known yet, gene expression changes found in the microwave dataset were compared to publicly available microarray data. Seventy-four *Arabidopsis thaliana* Affymetrix ATH1-datasets available at Genevestigator (Zimmermann et al 2004), belonging to one of the following

categories were selected for comparison: “biotic”, “chemical”, “hormone”, “light”, “nutrient”, and “stress”.

From these datasets, the logarithmic fold changes of the 10 differentially expressed genes of the microwave dataset (*Table 2*) were extracted and compared in a principal component analysis (PCA). With the analysis of the selected gene expression values in a single PCA, the microwave dataset could be related to the datasets and categories provided by Genevestigator. The PCA-plot (*Figure 3*) shows similar objects situated close to each other while dissimilar objects are separated along the principal component axes. The strongest factor of variance is represented by the horizontal axis, the second strongest factor by the vertical axis. For interpretation of the PCA-plot, the experiment categories “hormone” (blue)**,** “light” (turquoise), “nutrient” (magenta) and “stress” (yellow) were highlighted by convex hulls in the same color as the data points (Fig. 3). The convex hull was drawn such that all points lie either within or on the line of the hull (Everitt 2005) except for the large categories “hormone” and “stress”, for which a robust convex hull less sensible to outliers was drawn. For these categories, the convex hull was computed twice: after the first computation it was again computed on the remaining points resulting in shaded areas (*Figure 3*).

The microwave dataset is not located close to any of the clusters formed by the experimental categories “hormone”, “light”, “nutrient” or “stress”, implying that the genes differentially expressed in the microwave dataset are differently regulated in the Genevestigator datasets.

Considering PC1, the microarray datasets in which light conditions were altered (turquoise), comprise the closest cluster to the microwave dataset. This suggests similarities in gene regulation of the genes used for principal component analysis. The

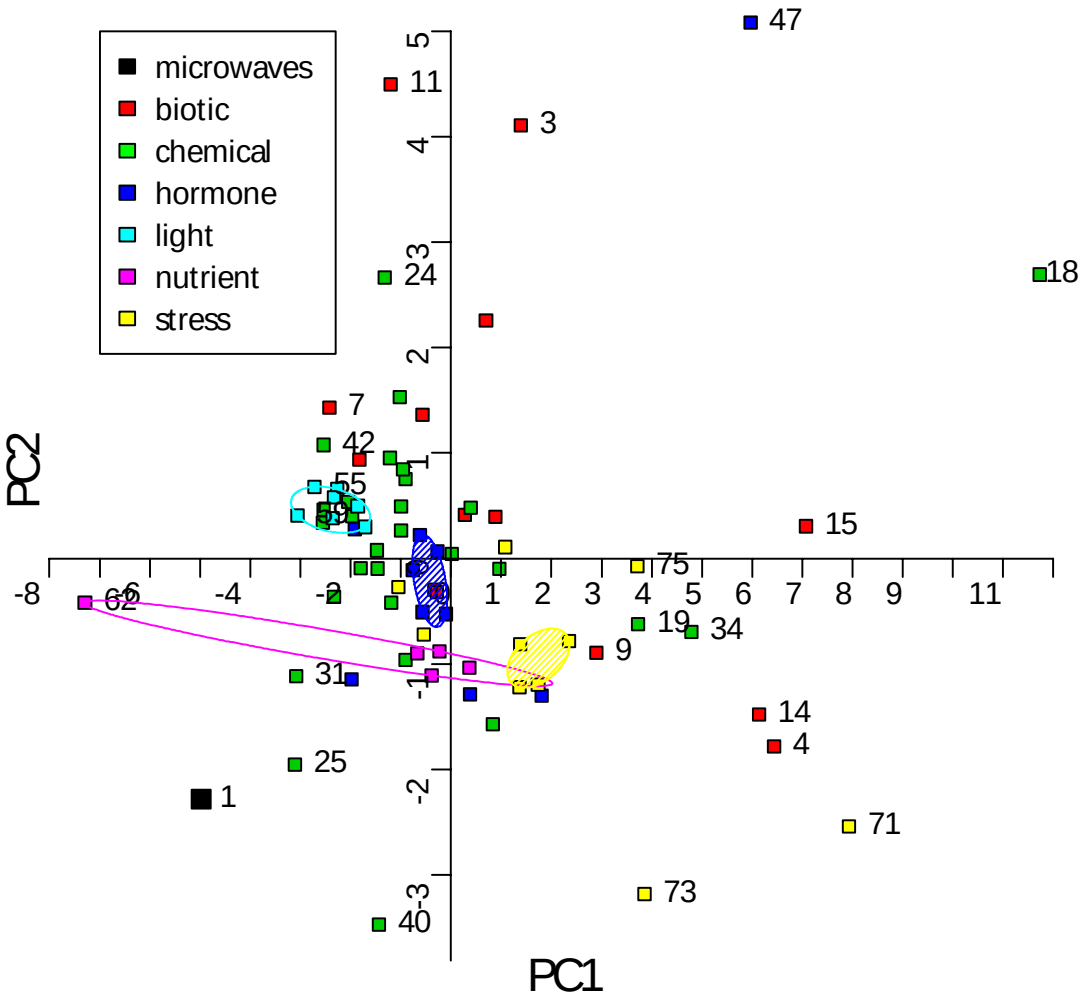


Figure 3. Principal Component Analysis of 75 ATH1 microarray datasets.

The fold changes of 10 genes differentially expressed in the microwave dataset (*Table 2*, no. 1 in Figure 3) were compared with fold changes of 74 ATH1 microarray datasets of Genevestigator (no. 2 to 75). The categories based on several datasets are: “biotic”, “chemical”, “hormone”, “light”, “nutrient”, and “stress” conditions, and are shown by symbols as indicated in the graph. Convex hulls encircle datasets treated with different light conditions (turquoise) or nutrient availability (magenta). Datasets treated with phytohormones (blue) or stress conditions (yellow) are encircled by a robust convex hull, disregarding data points on the outer convex hull and encompassing the remaining data points.

The following datasets are displayed but not all of them are numbered in the graphic: 1. microwaves, 2. *A. brassiciola*, 3. *A. tumefaciens*, 4. *B. cinerea*, 5. *E. cichoracearum*, 6. *E. orontii*, 7. *F. occidentalis*, 8. *M. persicae*, 9. *M. persicae*, 10. mycorrhiza, 11. nematode, 12. *P. infestans*, 13. *P. rapae*, 14. *P. syringae*, 15. *P. syringae*, 16. 2,4,6-trihydroxybenzamide, 17. 4-thiazolidinone / acetic acid, 18. 6-benzyl adenine, 19. $AgNO_3$, 20. aminoethoxyvinylglycine (AVG), 21. brassinazole 220, 22. brassinazole 91, 23. chitin, 24. high CO_2, 25. cycloheximide, 26. daminozide, 27. furyl acrylate ester, 28. hydrogen peroxide, 29. ibuprofen, 30. isoxaben, carbobenzoxyl-leucinyl-leucinyl-leucinal (MG13), 31. norflurazon, 33. naphthylphthalamic acid (NPA), 34. ozone, 35. paclobutrazole, 36. p-chlorophenoxyisobutyric acid (PCIB), 37. n-octyl-3-nitro-2,4,6-trihydroxybenzamide (PNO8), 38. prohexadione, 39. propiconazole, 40. syringolin, 41. 2,3,5-triiodobenzoic acid (TIBA), 42. uniconazole, 43. zearalenone, 44. absisic acid, 45. 1-aminocyclopropane-1-carboxylic acid (ACC), 46. brassinolide, 47. brassinolide / H_3BO_3, 48. ethylene, 49. giberellic acid (GA_3), 50. indole acitic acid, 51. methyl-jasmonate, 52. salicylic acid, 53. zeatin, 54. white light , 55. blue light, 56. far red light, 57. red light, 58. UV-A-irradiation, 59. UV-AB-irradiation, 60. white light, 61. Cs^+, 62. glucose/sucrose, 63. (-) potassium, 64. (-) nitrogen, 65. (-) sulfur, 66. cold, 67. drought, 68. genotoxic, 69. heat, 70. hypoxia, 71. osmotic, 72. oxidative, 73. salt, 74. UV-B, 75. wounding.

datasets of the stress experiments (yellow symbols and hull), behave differently compared to the microwave dataset because they have positive values on PC1. Thus, there is a clear separation between “light” and “stress” experiments along PC1. One dataset in the category “nutrient” (magenta) has a large negative value of PC1 indicating some similarity to the microwave experiment, but the remaining “nutrient” experiments form a cluster around the center of PC1, taking an intermediate position between the “light” and “stress” cluster. The datasets of the category “hormone” (blue) are spread over positive and negative values of PC1, but most experiments are situated around zero. They also take an intermediate position between the “light” and the “stress” datasets. Both categories, “biotic” (red) and “chemical” (green), are spread over the

whole range of values of PC1, indicating that their gene expression values concerning the selected genes differ between the single datasets.

While the first principal component axis (PC1) accounts for the majority of variation (51%), and thus conveys a large part of the information contained in the data, the second principal component axis (PC2, *Figure 3*) which holds the second strongest factor of variance, only accounts for 12% of the variation. Here, no obvious separation of groups is identifiable. The Principal Component Analysis did not unequivocally reveal which environmental factors or signaling pathways are involved in the regulation of the 10 genes listed in *Table 2*.

Discussion

The question if electromagnetic fields have an influence on gene expression in plant cells was addressed by a 24 h treatment of *Arabidopsis* cell suspensions with a microwave protocol which represents a worst case scenario of a pedestrian walking around in an urban area. This study was carried out with cell suspensions to ensure a minimal variation in the starting material. The experiments were performed double blinded, in which neither the experimenters handling the cell cultures nor the data analysts performing the initial microarray gene expression analysis knew which samples had been treated with microwaves. This procedure ensured an unbiased and unprejudiced analysis of the data. Exploratory analysis of Affymetrix ATH1 microarray data revealed that high frequency electromagnetic fields did not cause any major changes in gene activity (*Figure 1S*). This indicates that a 24 h period of exposure to electromagnetic fields as used in UMTS-technology does not have a major impact on

gene expression in plant cells. Hierarchical cluster analysis, however, revealed that microwave exposure could alter the activity of just a few genes. After disclosing which microarray samples had been treated with microwaves, differential gene expression analysis revealed significant changes in the transcription of 10 genes. Although the changes in gene activity were small, they were statistically significant. Real time RT-PCR experiments confirmed these changes in transcript numbers, but the degree of variation was much higher due to the higher sensitivity of this technique.

Of the 10 genes with significant p-values ($p < 0.05$), highest fold changes were maximal 2.5 fold down-regulated between microwave-exposed and control-cultured cells (*Table 2*). Compared, for example, to the elevation of the K^+ channel transcripts *GORK* after treatment of these suspension cultured cells with the stress phytohormone ABA (11-fold), this is very moderate. It indicated that the cells of the cell suspensions were able to react very sensitively to stress signals. Since the functions of several of the 10 genes differentially transcribed in the microwave experiment did not directly point to known responses of plants to any other environmental factors, their fold changes were compared to those of 74 ATH1-microarray experiments available online (Genevestigator, *Figure 3*). The microwave dataset clustered most closely to experiments in which plants were exposed to different light conditions. This is in concordance with the annotation of three out of the 10 genes listed in *Table 2*, which are known to be dark-inducible (At3g47340, At3g60140, At3g15450). In contrast, stress experiments formed the most distant cluster to the microwave dataset, implying that the genes studied here (*Table 2*) are regulated in a different way under stress conditions. If at all, radio frequency fields as used in UMTS-communication might be perceived by plants as irradiation, but are not recognized as a stress signal.

The microwave experiment was designed to achieve a maximal sensitivity. For this purpose, the biological variation was kept at a minimum, since it might otherwise have hidden small changes in gene transcription caused by RF fields. To confirm that the microwave dataset displays a low variability of gene expression, it was compared to publicly available datasets from the NCBI GEO database (http://www.ncbi.nlm.nih.gov/geo). Of these datasets which originally consisted of treated and untreated (= control) samples only the microarray data of the control hybridizations were selected. This comparison underlined that indeed the variation of the microwave dataset (exposed and shielded samples) was small compared to the variability of the publicly available control hybridizations (*Figure 3S*). For example, the degree of variation between control leaves of *Arabidopsis* plants (controls from GEO dataset GSE5611) was much larger than that between suspension cultured cells exposed to or shielded from RF fields.

Furthermore, we found that the transcript numbers of the genes listed in *Table 2* varies considerably between the different untreated control hybridizations of published microarray datasets (*Figure 3S*). Even between suspension cultured cells that were used as controls (GEO dataset GSE5748), the variation in transcript number of several genes listed in *Table 2* was larger than their variation due to RF field exposure. Therefore, the significant changes found in the microwave experiment would most likely be hidden by biological variation if cell cultures were cultured separately (e.g. at different times of the year) or if whole plants were used.

Because of the limited number of genes altered in RF-exposed cells and because their physiological functions are not well-annotated, it is difficult to predict what the impact of the observed changes in transcription would be in intact plants. Based on the

comparisons of variability between the microwave dataset and controls of different other datasets, it is very unlikely that the small changes in transcript numbers found in our analysis would have been observed when whole plants or tissues would have been used as starting material.

Conclusions

Overall, we conclude that RF fields used in mobile phone communication have no dramatic effect on the gene activity of plant cells in suspension culture. Only few genes displayed an altered transcription level after 24 h of exposure to high frequency electromagnetic fields and the alterations did not exceed a 2.5 fold reduction or increase in gene activity. It is unlikely that these small changes in gene activity of very few genes will have pronounced effects on the physiology of plant cells. Cells of a suspension culture, however, do not resemble autotrophycally growing plants in every respect and their responses to RF fields may differ from those of intact plants. Future experiments may be set out to test responses of whole plants, including trees, to further estimate the impact of UMTS technology on the green environment.

Acknowledgements

We thank K. Neuwinger, J. Arnold, N. Hong and U. Taggeselle for technical assistance and are grateful to K. Philippar, Department of Biology I, Ludwig-Maximilians-University of Munich (Germany) for carrying out the microarray hybridizations.

The work was supported by E-plus (Germany), BMBF project FUNCRYPTA (FKZ 0313838B), the state of Bavaria (IZKF B-36) and by technical equipment and a grant from Swisscom Innovations (Switzerland).

References

Adair RK. 2003. Biophysical limits on athermal effects of RF and microwave radiation. *Bioelectromagnetics,* 24:39-48.

Affymetrix. 2002. Statistical algorithms description document. Technical Report. Affymetrix. Santa Clara, CA, USA.

Becker D, Hoth S, Ache P, et al. 2003. Regulation of the ABA-sensitive Arabidopsis potassium channel gene GORK in response to water stress. *FEBS Letters,* 554:119-126.

Benjamini Y and Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B-Biological Sciences*, 57:289-300.

Bilz A, Bökelmann V, Gerhardt D, et al .2001. A generic UMTS test signal for bio-experiments. 5th. International Congress of the European BioElectromagnetics Association, Helsinki 173-174.

Cashmore AR .2003. Cryptochromes: enabling plants and animals to determine circadian time. *Cell*, 114:537-543.

Deeken R, Ivashikina N, Czirjak, et al .2003. Tumour development in *Arabidopsis thaliana* involves the Shaker-like K+ channels AKT1 and AKT2/3. *Plant Journal*, 34:778-787.

Efron B, Halloran E, Holmes S. 1996. Bootstrap confidence levels for phylogenetic trees. *Proc Natl Acad Sci USA*, 93:7085-7090.

Efron B and Tibshirani R. 1993. An Introduction to the Bootstrap. London: Chapman and Hall.

European Commission-Research Directorate-General-European Communities. Health and electromagnetic fields. 2005. EU-funded research into the impact of electromagnetic fields and mobile telephones on health. ISBN 92-79-00187-6. URL: http://ec.europa.eu/research/quality-of-life/pdf/emf_brochure_and_sheets_en.pdf.

Everitt B. 2005. An R and S-PLUS Companion to Multivariate Analysis. London: Springer. p 22.

Fellenberg K, Hauser NC, Brors B, et al .2001. Correspondence analysis applied to microarray data. *Proc Natl Acad Sci USA*, 98:10781-10786.

Felsenstein J .1989. PHYLIP- Phylogeny Inference Package (Version 3.2). *Cladistics*, 5:164-166.

Gentleman R, Carey V, Bates M, et al .2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5:R80.

Haider T, Knasmueller S, Kundi M, et al .1994. Clastogenic effects of radiofrequency radiations on chromosomes of Tradescantia. *Mutat Res,* 324:65-68.

Hardell L, Carlberg M, Mild KH. 2005. Case-control study on cellular and cordless telephones and the risk for acoustic neuroma or meningioma in patients diagnosed 2000-2003. *Neuroepidemiology*, 25:120-128.

Hocking B, Gordon IR, Grain HL, et al .1996. Cancer incidence and mortality and proximity to TV towers. *Med J Aust,* 165:601-605.

Hopkins CD. 1995. Convergent designs for electrogenesis and electroreception. *Curr Opin Neurobiol*, 5:769-777.

Huber W, Von Heydebreck A, Sültmann H, et al. 2002. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics,* 18:S96-S104.

Ihaka R and Gentleman R. 1996. R: A Language for Data Analysis and Graphics. *J Comput Graph Stat*, 5:299-314.

International Commission on Non-Ionizing Radiation Protection. 1998. Guidelines for Limiting Exposure to Time-Varying Electric, Magnetic and Electromagnetic Fields (up to 300 GHz). *Health Physics,* 74:494-522.

Hobbs B, Collin F, Beazer-Barclay YD, et al. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostat,* 4:249-264.

Jolliffe I. 1986. Principal Component Analysis. New York: Springer.

Kalmijn AJ. 1966. Electro-Perception in Sharks and Rays. *Nature*, 212:1232-1233.

Moulder JE, Foster KR, Erdreich LS, et al. 2005. Mobile phones, mobile phone base stations and cancer: a review. *Int J Radiat Biol,* 81:189-203.

Mouritsen H and Ritz T. 2005. Magnetoreception and its use in bird navigation. *Curr Opin Neurobiol*, 15:406-414.

Scientific Committee on emerging and newly identified health risks (SCENIHR). 2006. Preliminary opinion on possible effects of electromagnetic fields (EMF) on human health. URL: http://ec.europa.eu/health/ph_risk/committees/04_scenihr/docs/scenihr_o_006.pdf.

Smyth GK. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Applic Genet Mol Biol*, 3:Article 3.

Venables WN and Ripley BD. 2002. Modern Applied Statistics with S. Fourth edition. Springer.

Wasserman L. 2004. All of Statistics. A Concise Course in Statistical Inference. Springer. p. 158.

Zimmermann P, Hirsch-Hoffmann M, Hennig L, et al. 2004. GENEVESTIGATOR. *Arabidopsis* microarray database and analysis toolbox. *Plant Physiol,* 136:2621-2632.

Additional files

Figure 1S. Correspondence Analysis of expression signals of all genes on the 6 ATH1 microarrays of the microwave dataset

Smoothed color density representation of genes. Dark blue areas reflect high densities of genes and light blue areas represent low gene densities. Single genes in the outer area are marked by small black points. Single microarrays are marked with black squares. There is no clustering of two groups of arrays along the first or second component axis (Array 1 and 5 were left out, since these were hybridized later).

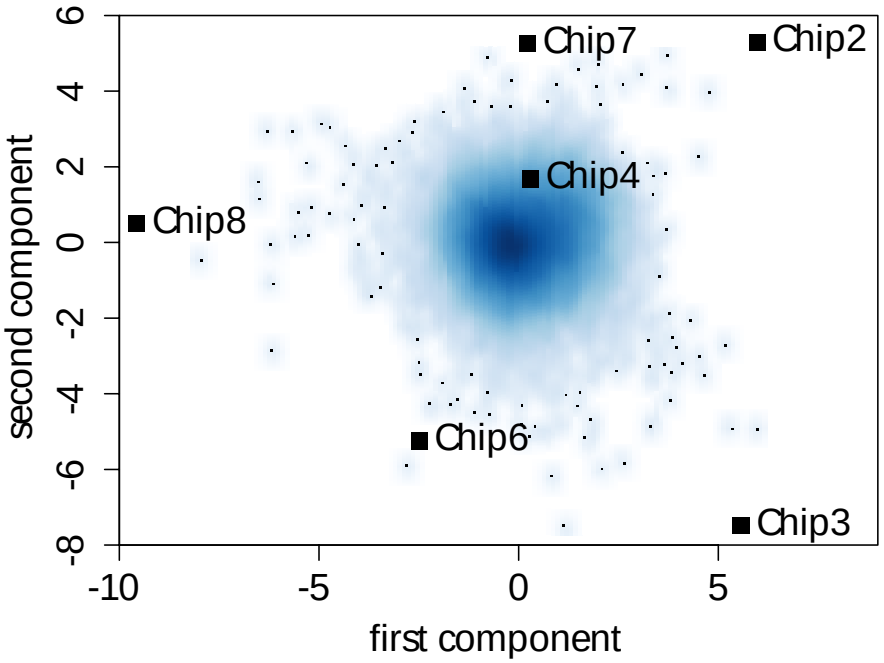


Figure 2S. Distribution of unadjusted p-values of differential gene expression

(A) Unadjusted p-values for a sample grouping irrespective of microwave treatment.

(B) Unadjusted p-values for true sample grouping: microwave treated vs. untreated samples. Shaded red areas represent the uniform distribution of p-values of no differential expression. For the true grouping, blue bars reaching out of the shaded area represent differentially expressed genes. Naturally, after multiple testing correction of p-values, the number of genes with significant p-values (Table 2) is substantially lower than what could be estimated from the distribution of unadjusted p-values.

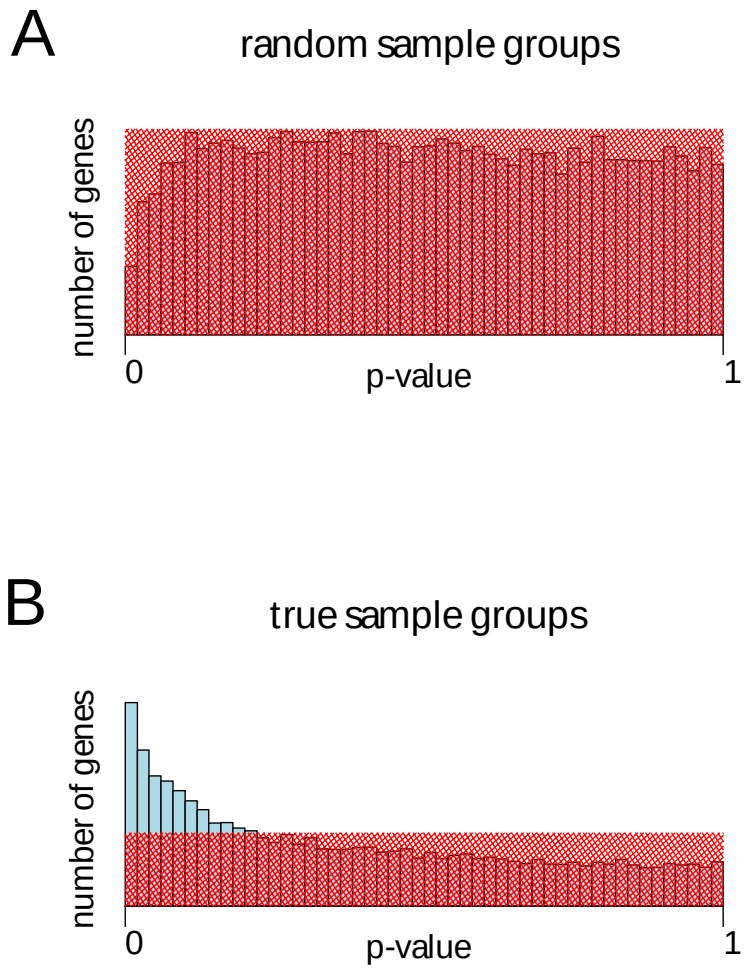
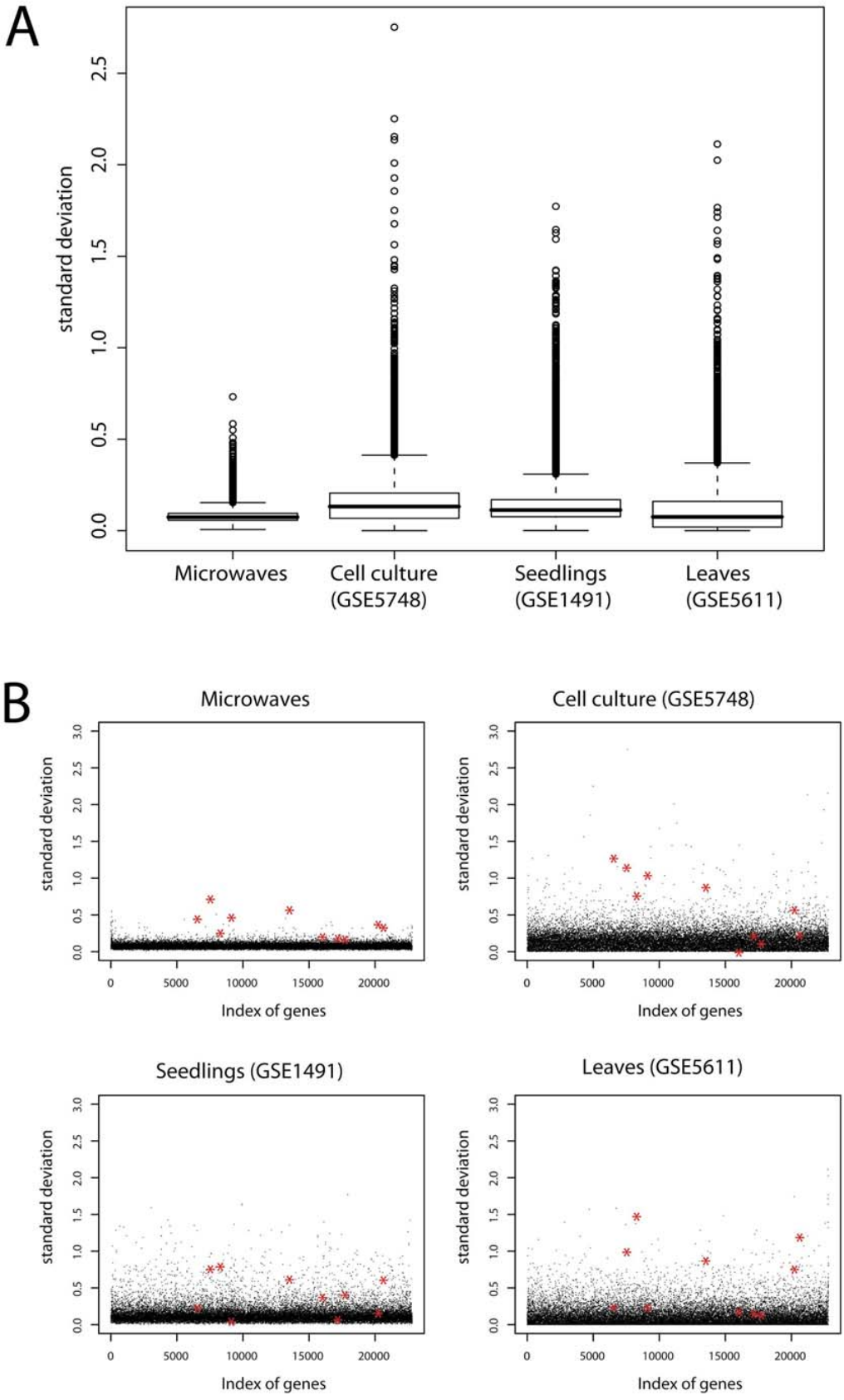


Figure 3S. Comparison of gene expression variability of the microwave dataset with untreated control microarrays from publicly available datasets.

(A) Box-plots of standard deviations of all genes on the ATH1 microarray. The microwave dataset, the controls of a cell culture dataset (GSE5748), those of seedlings (GSE1491), and of leaves (GSE5611) are shown.

(B) Scatter-plots of standard deviations of all genes on the ATH1 microarray. The 10 differentially expressed genes of the microwave dataset are highlighted with red stars in each of the datasets.

The controls of the cell culture, seedlings and leaves dataset are accessible at NCBI GEO database, (http://www.ncbi.nlm.nih.gov/geo) with their GSE identifier. Raw data of the microarray hybridizations were normalized with the same methods as the microwave microarrays, as described in the Methods section.


A
standard deviation
0.0
0.5
1.0
1.5
2.0
2.5
Microwaves
Cell culture (GSE5748)
Seedlings (GSE1491)
Leaves (GSE5611)
B
Microwaves
Cell culture (GSE5748)
Seedlings (GSE1491)
Leaves (GSE5611)
standard deviation
Index of genes
0
5000
10000
15000
20000

Chapter 4

An integrated view of gene expression and solute profiles of *Arabidopsis* tumors: A genome-wide approach

The Plant Cell, Vol. 18, 3617–3634, December 2006, www.plantcell.org © 2006 American Society of Plant Biologists

An Integrated View of Gene Expression and Solute Profiles of *Arabidopsis* Tumors: A Genome-Wide Approach [W]

Rosalia Deeken,[a] Julia C. Engelmann,[b] Marina Efetova,[a] Tina Czirjak,[a] Tobias Müller,[b] Werner M. Kaiser,[a] Olaf Tietz,[c] Markus Krischke,[d] Martin J. Mueller,[d] Klaus Palme,[c] Thomas Dandekar,[b] and Rainer Hedrich[a,1]

[a] Julius-von-Sachs-Institute, Department of Molecular Plant Physiology and Biophysics, University of Wuerzburg, D-97082 Wuerzburg, Germany

[b] Theodor-Boveri-Institute, Department of Bioinformatics, University of Wuerzburg, D-97074 Wuerzburg, Germany

[c] Institute of Biology II, Cell Biology, University of Freiburg, 79104 Freiburg, Germany

[d] Julius-von-Sachs-Institute, Department of Pharmaceutical Biology, University of Wuerzburg, D-97082 Wuerzburg, Germany

Transformation of plant cells with T-DNA of virulent agrobacteria is one of the most extreme triggers of developmental changes in higher plants. For rapid growth and development of resulting tumors, specific changes in the gene expression profile and metabolic adaptations are required. Increased transport and metabolic fluxes are critical preconditions for growth and tumor development. A functional genomics approach, using the Affymetrix whole genome microarray (~22,800 genes), was applied to measure changes in gene expression. The solute pattern of *Arabidopsis thaliana* tumors and uninfected plant tissues was compared with the respective gene expression profile. Increased levels of anions, sugars, and amino acids were correlated with changes in the gene expression of specific enzymes and solute transporters. The expression profile of genes pivotal for energy metabolism, such as those involved in photosynthesis, mitochondrial electron transport, and fermentation, suggested that tumors produce C and N compounds heterotrophically and gain energy mainly anaerobically. Thus, understanding of gene-to-metabolite networks in plant tumors promotes the identification of mechanisms that control tumor development.

INTRODUCTION

Integration and expression of oncogenes, encoded by the T-DNA of the *Agrobacterium tumefaciens* Ti plasmid, induce the development of plant tumors, also referred to as crown galls (Van Larebeke et al., 1974; Chilton et al., 1977). Rapid cell proliferation of tumors is promoted by high concentrations of cytokinin and auxin, which are synthesized by bacterial enzymes encoded by genes of the T-DNA (Kado, 1984). These plant growth factors not only control dedifferentiation of plant cells into primary tumor cells but also induce tumor cells to differentiate a vascular network of vessels and sieve elements. This network connects to vascular bundles of the host plant and thus sustains a rapid supply of nutrients and water (Malsy et al., 1992). Moreover, inevitable transpiration of noncutinized tumors with a disrupted epidermal layer drives nutrient flow into the tumor and mediates the accumulation of nutrients (Schurr et al., 1996; Wachter et al., 2003).

The growth of solid animal and human tumors also depends on neovascularization (Folkman, 1971; Gimbrone et al., 1974). Human tumors induce a dense network of blood vessels that supply the tumor with nutrients, water, and oxygen. Likewise, plant tumor cells proliferate only in vascularized regions, whereas in nonvascularized areas they necrotize (Ullrich and Aloni, 2000). Animal tumors overexpress angiogenic growth factors, such as tumor necrosis factor, fibroblast growth factor, and vascular endothelial growth factor, the latter of which is considered to be a major mediator in tumor angiogenesis (Risau, 1990; Carmeliet and Jain, 2000). In plants, gradients of growth factors such as cytokinins and auxin are established, inducing and controlling vascular differentiation (Aloni et al., 2003; Scarpella et al., 2006).

A common property of cancer cells is their capacity to metabolize glucose at high rates (Warburg, 1930; Aisenberg, 1961; Pedersen, 1978). Tumor mitochondria show impaired respiration, which is compensated for by an unusually high contribution of glycolysis to total ATP production. Some types of cancer have increased activity of the glucose transporter-1 (Chang et al., 2000), and its activity correlates inversely with survival (Wachsberger et al., 2002). The aberrant glucose metabolism provides a constant supply of energy even when oxygen levels decrease; as a result, the tumor has a metabolic growth advantage over normal tissues.

To attain a comprehensive picture of a T-DNA–induced plant tumor, we combine here bioinformatics and genome-wide expression analysis with direct analysis of metabolites, ions, oxygen consumption, and photosynthesis. The different data and approaches complement and strengthen each other and allow a detailed picture of the induced changes in the host cell from an

[1] To whom correspondence should be addressed. E-mail hedrich@botanik.uni-wuerzburg.de; fax 49-391-888-6157.
The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: Rosalia Deeken (deeken@botanik.uni-wuerzburg.de).

[W] Online version contains Web-only data.

www.plantcell.org/cgi/doi/10.1105/tpc.106.044743

auxotrophic, aerobic metabolism to a heterotrophic, transport-dependent, sugar-degrading, and cell wall–synthesizing (gall formation) anaerobic metabolism.

RESULTS

In this work, we have studied alterations in the gene expression and solute content of *Arabidopsis thaliana* tumors and compared them with tumor-free inflorescence stalk tissue. When we analyzed the composition of metabolites of *Arabidopsis* tumors, we realized that their contents differed substantially from those of tumor-free tissue (see below). To understand the molecular mechanism causing these changes, we determined the gene expression profile of the transcriptome of *Arabidopsis* tumors. Because *Agrobacterium*-induced tumors might be composed of T-DNA–transformed and nontransformed cells responding to the altered milieu, the alterations in transcript levels and metabolite contents may reflect responses of both cell types. To interpret gene expression and metabolite data concerning tumor physiology correctly, we have used an in situ hybridization technique to calculate the percentage of transformed cells within *Arabidopsis* tumors. Using nopaline synthase (*NOS*) antisense RNA as a probe, >95% of tumor cells revealed a hybridization signal (reddish color in Figures 1A to 1D), indicating that they express *NOS* mRNA encoded by the T-DNA. However, this estimation differs from previous reports (see Discussion). Strong hybridization signals were observed in small and plasma-rich tumor cells (Figures 1A and 1C) but not in inflorescence stalk cells adjacent to the tumor (Figure 1B). In tumor cells with a large central vacuole, the hybridization signal was visible in the cytoplasmic border layers (Figure 1D, arrows). No signal was found in small or large tumor cells hybridized with the *NOS* sense RNA probe (Figures 1E and 1F, respectively). We cannot exclude the possibility that *Arabidopsis* tumors also contain a few nontransformed cells, but most cells appear transformed and express genes located on the T-DNA. Therefore, our studies of changes in gene expression and solute content reveal the results of the T-DNA integration event.

To analyze tumor gene expression data gained by microarrays (ATH1; Affymetrix), we checked the reliability of microarray data applying bioinformatics tools. After bioinformatic analysis (see below) of the transcriptome and biochemical analysis of the metabolome, we found concerted changes from autotrophic to heterotrophic metabolism in the tumor tissue. These involve the upregulation of genes involved in transport, glycolysis, sucrose degradation, and cell wall synthesis (for gall formation) as well as the downregulation of genes for photosynthesis, lipid metabolism, N metabolism, and amino acid synthesis. These results are described in detail below and are integrated into a comprehensive model (see Figure 9 below).

Bioinformatic Analysis of Affymetrix Microarrays

Data Acquisition

The Affymetrix microarray (ATH1 121501) was used to explore the differential expression profiles of genes in plant tumors induced by the nopaline-using *Agrobacterium* strain C58. Differentially expressed genes were identified from the expression data acquired from eight independent microarray hybridizations. Four replicates of tumor RNA and four of injured but not infected inflorescence stalks as reference RNA were used to calculate the expression value for each gene. Each replicate of four contained tissue fragments from at least 10 to 12 individual plants. For analysis of the expression profile, the fold change of normalized signals derived from tumor versus reference stalk tissue was calculated. Only fold changes of genes that met the significance criterion of $P < 0.01$ are presented here. Of 22,810 spotted genes on the *Arabidopsis* ATH1 microarray, 5054 (22%) met this criterion. Among them, 2340 genes (10%) were higher expressed in tumors (see Supplemental Table 1 online), and 2714 genes (12%) were higher expressed in reference inflorescence stalk tissue (see Supplemental Table 2 online). Of the 2340 genes with higher expression in tumors, 551 had a more than threefold difference, and of the 2714 with lower transcription, 608 were reduced at least threefold. The largest fold changes among all of the genes was a 56-fold ($P = 2.3E-04$) upregulation of an auxin-responsive GH3 family gene (At2g23170) and a 49-fold ($P = 1.2E-04$) downregulation in tumor tissues of the branched chain amino acid aminotransferase gene (At3g19710).

Clustering Microarray Data by Correspondence Analysis

Correspondence analysis revealed that the main difference between the eight microarray hybridization assays is attributable to differential expression between the two tissue types, tumor (T) and noninfected inflorescence stalk (N) tissue (Figure 2). This can be seen by the clear separation of the microarray assays from tumor and noninfected inflorescence stalk along the axis of the first component (*x* axis), confirming the high quality of the data. To examine whether the genes were also dispersed along the axis of the first component according to differential expression between tissue types, the locations of the 10 differentially expressed genes with lowest P values (Figure 2, circles) or highest fold changes (Figure 2, crosses) according to Linear Models for Microarray (LIMMA) analysis (Smyth, 2004) are highlighted in the correspondence analysis plot. Their positions at extreme values of the first component axis reflect differential expression of genes in different tissue types. Genes with the highest fold change are located at the outmost range of the first component, whereas genes with lowest P values have less extreme levels. The latter effect is attributable to the fact that genes with higher variance receive lower P values in the moderate *t* test analysis. This again indicates that genuine differential expression between tissue types is the strongest factor of variation in the data.

Functional Categorization

To structure the genes present on the *Arabidopsis* whole genome microarray, they were assigned to functional categories using the pathway analysis program MapMan (http://gabi.rzpd.de/projects/MapMan, version 1.8.0 [January 30, 2006]). MapMan is a user-driven tool that displays large data sets such as gene expression data from *Arabidopsis* Affymetrix microarrays onto diagrams of metabolic pathways or other processes (Usadel et al., 2005). Of 22,810 genes on the Affymetrix ATH1 chip,

13,642 (59.8%) could be assigned to categories with known biological functions (see Supplemental Figure 1 and Supplemental Table 3 online). The largest number of annotated genes of the ATH1 chip fall into the functional categories of protein and RNA.

To functionally categorize differentially expressed genes, we chose the 5046 genes with $P < 0.01$, which constituted ~22% of the 22,810 genes and gave a robust overview of differential gene regulation within functional categories. Among the 5046 genes selected, 3357 genes could be classified by MapMan into functional categories (see Supplemental Table 3 online). The distribution of these genes (total of 3357) within functional categories was compared with the distribution of all classified genes of the microarray (total of 13,642) and is presented as factorial changes in Figure 3. The factorial change describes the natural logarithm of the ratio of the percentage of differentially expressed genes in a functional category and the percentage of all annotated genes on the microarray in that category. For better comparison, the logarithmic factorial change of each category was plotted according to the following equation:

$$\text{factorial change} = \ln[(DE_{cat}/DE_{all})/(A_{cat}/A_{all})],$$

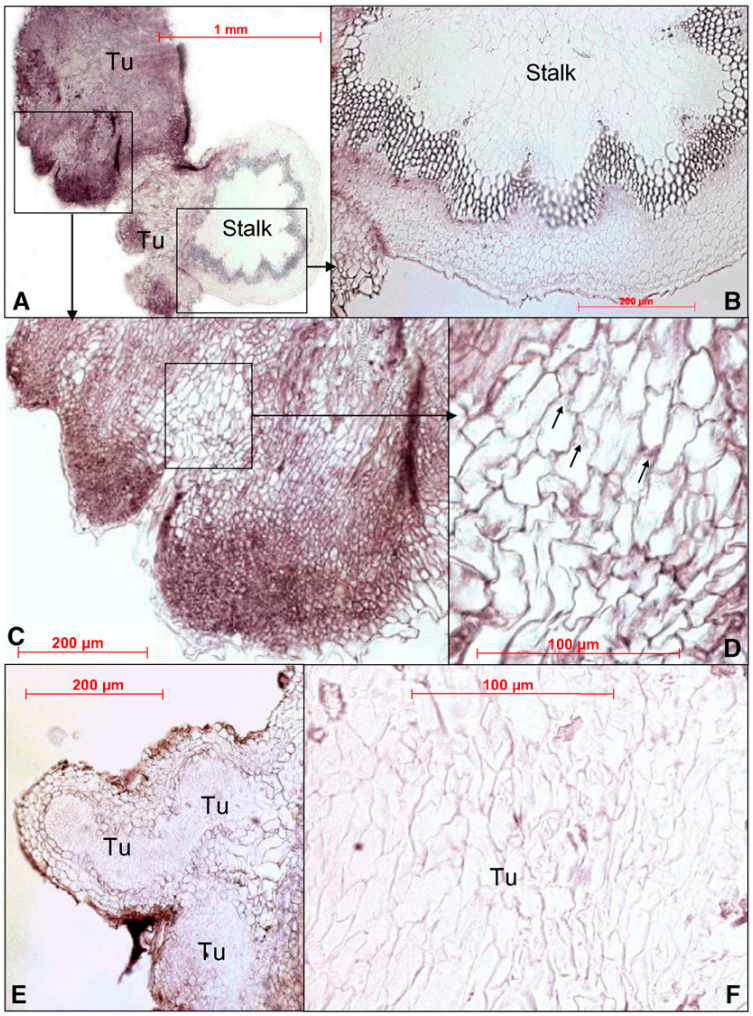


Figure 1. Detection of T-DNA–Transformed Cells in *Arabidopsis* Tumors.

(A) to **(D)** In situ hybridization using *NOS* antisense RNA as a probe.
(E) and **(F)** Hybridization with the sense RNA as a control. Positive hybridization signal appears as reddish color.
(A) Cross section through an inflorescence stalk (Stalk) and a tumor (Tu) attached to it.
(B) and **(C)** Enlargements of the marked areas in **(A)** of the inflorescence stalk and tumor.
(D) Enlargement of a tumor area with large cells marked in **(C)** showing hybridization signals close to the cell wall.
(E) Cross section of a tumor.
(F) Enlargement of a tumor area with large cells.

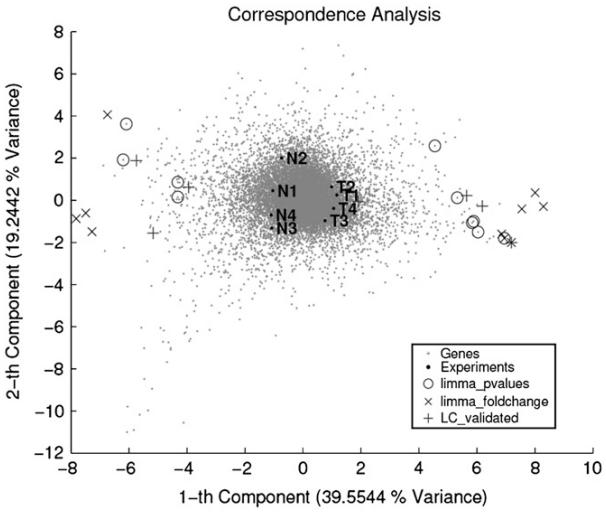


Figure 2. Correspondence Analysis of the Gene Expression Levels of the *Arabidopsis* Transcriptome.

Correspondence analysis shows that the main difference between the expression values of the different microarray hybridizations is attributable to differential gene expression between tumor and uninfected tissue. This can be seen from the separation of tumor and uninfected tissue microarray assays and differentially expressed genes along the horizontal axis. Genes are represented by tiny gray spots, and single microarray slides are indicated by black dots. N stands for reference slides, and T stands for tumor slides. Genes with lowest P values and highest fold change are marked with circles and crosses, respectively. Genes verified by RT-PCR are marked by plus signs.

where DE_{cat} represents differentially expressed genes in a functional category, DE_{all} represents all differentially expressed genes, A_{cat} represents all annotated genes in a functional category, and A_{all} represents all annotated genes on the array.

Thus, the factorial change represents a relative measure for overall gene regulation in a category. Functional categories with a positive factorial change have a higher fraction of differentially expressed genes than would be expected from the total number of genes assigned to that category. A negative value indicates a lower number of differentially regulated genes in the respective category than expected. Of the 17 functional categories shown in Figure 3, 11 contained a higher number of differentially expressed genes, whereas in six categories the number was lower. Categories 1 (photosynthesis [PS]), 36 (primary metabolism), and 34 (transport) were the three with a higher number of differentially expressed genes, in contrast with category 28 (DNA), in which a larger number of genes were not differently transcribed in the tumor and reference tissue (Figure 3; see Supplemental Table 4 online). The mean percentage of genes differentially expressed in a functional category was 28%, whereas in the DNA category, only 8% of the total gene number (882 genes) were significantly differentially expressed ($P < 0.01$; 71 genes). However, among the 71 significantly differentially expressed genes in the DNA category, 70% were activated in tumors. In the subcategory of DNA synthesis, even 73% of the genes involved in cell proliferation showed increased expression levels (cf. the subtables DNA_all and DNA_P < 0.0.1 in Supplemental Table 4 online).

Because the categories PS, primary metabolism, and transport were the three with the greatest number of differentially expressed genes, all annotated genes of the complete microarray belonging to these categories were assigned to subcategories (see Supplemental Figures 2A to 2C and Supplemental Tables 5 to 7 online). The greatest number of genes in the photosynthesis category was formed by the subcategory light reaction, and within the category of primary metabolism were major and minor CHO (carbohydrate metabolism) and mitochondrial e^- (electron) transport. In the category of transport, the subcategories ABC (for ATP binding cassette), metal, sugars, and metabolite contained the greatest number of differentially expressed genes.

The distribution of genes in these functional subcategories was again compared with the distribution of all differentially expressed genes of the microarray via factorial changes (Figure 4). Subcategories 14 (S assimilation), 12 (N metabolism), and 7 (oxidative pentose phosphate [OPP]) revealed the greatest number of differentially regulated genes, followed by subcategories 2 (major CHO), 5 (fermentation), and 4 (glycolysis) of the primary metabolism category (Figure 4A; see Supplemental Table 5 online). The greatest number of differentially expressed genes of the photosynthesis category belonged to subcategories 1.1 (light reaction) and 1.3 (Calvin cycle) (Figure 4B; see Supplemental Table 6 online). Among the 15 subcategories of the transport category, only two, 34.1 (P- and V-ATPases) and 34.8 (metabolite) contained a lower number of differentially regulated genes. It has to be mentioned that in the transporter subcategory metabolite, only mitochondrial membrane transporters are listed (see Supplemental Table 7 online). Because the four subcategories 9

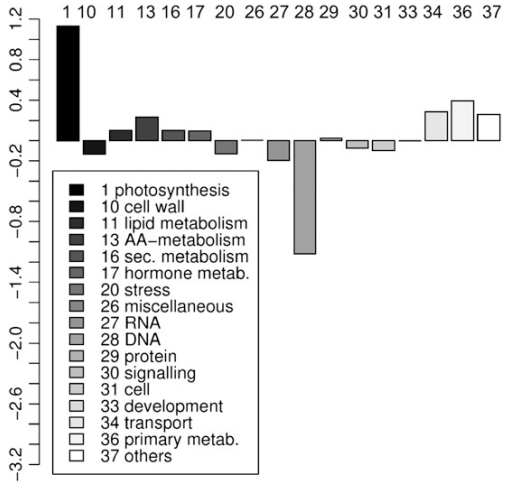


Figure 3. Factorial Changes within Functional Categories of All Differentially Expressed Genes.

The natural logarithm of factorial changes is plotted against each functional category. Positive factorial changes indicate a larger fraction of differentially expressed genes; negative factorial changes represent categories with a smaller fraction of regulated genes than expected from the total number of genes in the respective category. AA, amino acid.

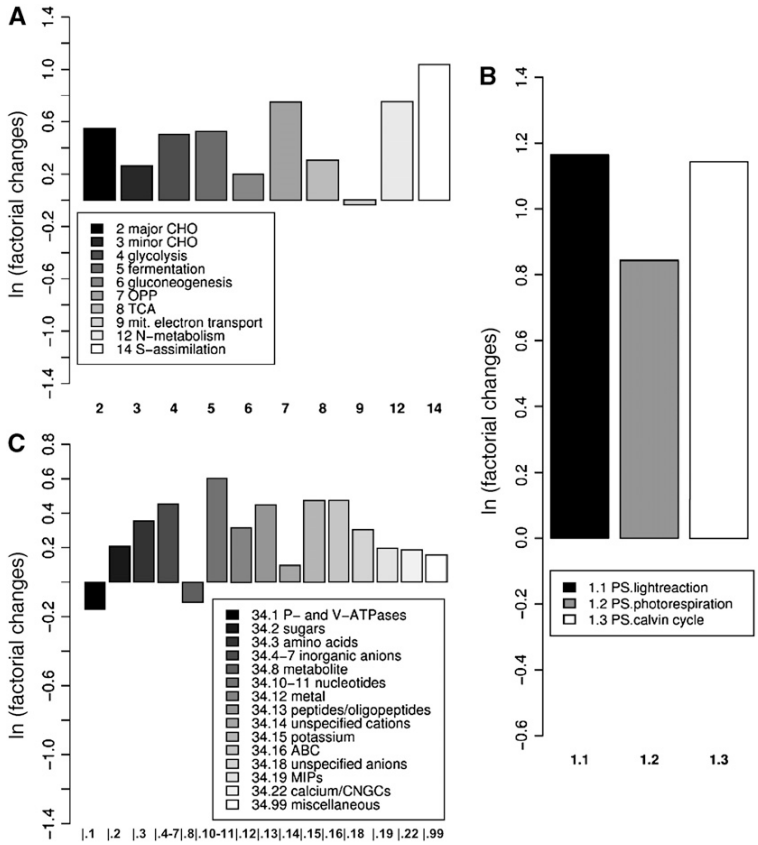


Figure 4. Factorial Changes of Functional Subcategories with the Highest Number of Differentially Expressed Genes.

The natural logarithm (ln) of factorial changes is plotted against the functional categories of photosynthesis **(A)**, primary metabolism **(B)**, and transport **(C)**. Positive factorial changes indicate more regulated genes; negative factorial changes indicate a smaller number of regulated genes in that category than expected from the size of the category. TCA, tricarboxylic acid; MIP, major intrinsic protein family; CNGC, cyclic nucleotide gated channel.

(mitochondrial e^- transport), 14 (S assimilation), 12 (N metabolism), and 7 (OPP) contain only a small numbers of genes, their factorial changes displayed in Figure 4A and in Supplemental Table 5 online are less reliable than those from the larger subcategories, such as 2 (major CHO) and 4 (glycolysis).

Members of the functional category transport are highlighted (black dots) in a MA plot (see Methods; Figure 5), because this category contained a large number of differentially expressed genes (Figure 3; see Supplemental Table 4 online). This plot shows that a considerable number of transporter genes with high mean intensities (high A values) are differentially expressed. These changes are referred to in more detail below.

Verification of Microarray Data by Quantitative RT-PCR

Numbers of transcripts of selected genes with either moderate or high differential expression values from microarray analysis were independently quantified by quantitative RT-PCR. They include a cytochrome oxidase (At5g56970), a wound-induced protein (At4g10270), a glycosyl hydrolase (At1g66280), a receptor protein kinase (At1g51805), a 2,4-D–inducible glutathione *S*-transferase (At1g78370), and a Ser carboxypeptidase I (At2g22990). The quantitative RT-PCR results of the latter six genes showed similar differential expression patterns as obtained by microarrays. Comparing the results of previous quantitative RT-PCR studies of *Arabidopsis* ion channel genes (Deeken et al., 2003) with those derived from the ATH1 microarray (KCO1, At5g55630; KCO2, At5g46370; KCO5, At4g01840; KCO6, At4g18160; KAT1, At5g46240; KAT2, At4g18290; AKT1, At2g26650; AKT2/3, At4g22200; At KC1, At4g32650; and GORK, At5g37500), a similar differential gene expression profile was obtained. Quantitative RT-PCR confirmed the microarray data. Of nine potassium channel genes of the *Arabidopsis* genome, two, AKT2/3 and GORK, were repressed in tumor tissues, whereas AKT1 and At KC1 were induced. Thus, both methods, microarray analysis and quantitative RT-PCR, revealed a high correlation between their identified fold changes of gene expression. The Pearson's correlation coefficient of microarray gene expression and quantitative RT-PCR data was 0.9453 ($P = 3.4E-8$) (see Supplemental Figure 3 online).

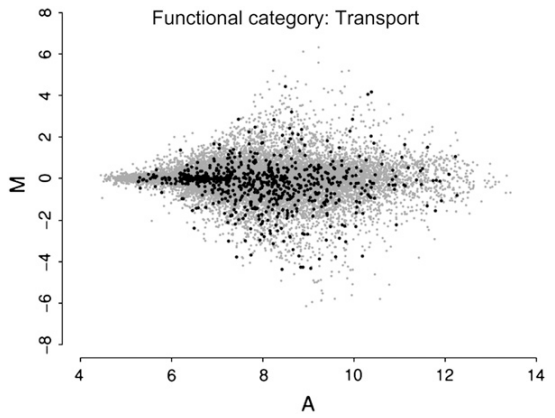


Figure 5. MA Plot of Genes of the Functional Category of Transport.

Many genes of the functional category transport are regulated between tumor and uninfected tissue. Regulated genes have large M values in the MA plot. The M values on the vertical axis represent differential expression between the two tissue types, and the A values on the horizontal axis represent average expression over all of the microarray assays. All genes are shown as gray dots, and genes of the functional category of transport are marked with black dots.

Solute Patterns and Gene Expression Profiles

Carbohydrate Metabolism and Photosynthetic Light Reactions

The carbohydrate pool (Figure 6) in tumor tissue was dominated by glucose (14.9 μmol/g fresh weight) compared with sucrose (2.8 μmol/g fresh weight) and fructose (1.7 μmol/g fresh weight) levels. Glucose concentration in tumors was 3.3 times higher than that in reference stalks (4.4 μmol/g fresh weight). Sucrose was found only in tumors and was undetectable in reference stalks.

Glucose accumulation in *Arabidopsis* tumors is most likely not derived from photosynthesis de novo, as transcript levels of the majority of differentially expressed genes ($P < 0.01$) related to photosynthesis were decreased (see Supplemental Table 4 online). This reflects the reduced number of chloroplasts in tumor cells that were smaller than those of mesophyll cells and the reduced expression of all genes encoded in the chloroplast genome (see Supplemental Table 6 online). A comparison of chlorophyll content revealed a three times reduced level in tumors compared with that in inflorescence stalk tissues (161 versus 499 μg chlorophyll *a/b* per gram fresh weight). This was paralleled by the downregulation of six genes out of seven involved in tetrapyrrole synthesis (see Supplemental Table 6 online). However, the relative quantum efficiency of chlorophyll fluorescence (Schreiber et al., 1986) was very similar in both tissues (0.67 ± 0.05 in tumors and 0.65 ± 0.04 in controls), which indicates that the still existing photosynthetic membranes were functional. Of 100 genes with differential expression ($P < 0.01$) involved in photosynthetic light reactions, all were significantly lower expressed in tumors. The same holds true for Calvin cycle genes (24 of 27; see Supplemental Table 6 online).

In tumor tissues, we found a strong activation of sucrose-degrading enzymes, accompanied by activation of *STP4* (At3g19930), a sink-specific monosaccharide transporter (see Supplemental Table 7 online), and increased glucose levels (Figure 6). The cluster of sucrose-degrading enzymes comprised 14 differentially regulated genes, of which 9 were activated. Most pronounced was the induction of two sucrose synthase genes (*SuSy3* [At3g43190] and *SuSy5.2* [At5g20830]), a fructokinase (At2g31390), and a cell wall invertase gene (At3g13790), whereas the genes of two vacuolar invertases were expressed to a much lower level in tumors (At1g12240 and At1g62660). Transcription of genes of the major CHO pathway coding for starch-synthesizing enzymes (7 of 10) and starch-degrading enzymes (13 of 14) was also reduced, again reflecting the reduced number of chloroplasts in tumor cells (see Supplemental Table 5 online).

Energy Production

Genes coding for proteins of mitochondrial electron transport were mainly unchanged in tumors, with the exception of an uncoupling protein, At PUMP1 (At3g54110), involved in the alternative respiratory chain, which was 2.6-fold ($P = 4.6E-04$) induced (see Supplemental Figure 5 and Supplemental Table 5 online). Genes required for fermentation were strongly upregulated: *Pyruvate Decarboxylase* (*PDC1* [At4g33070], $P = 1.2E-05$) and *Alcohol Dehydrogenase* (*ADH* [At1g77120], $P = 9.3E-05$). The increased transcript level of the latter two genes was paralleled by a threefold increase of ethanol concentrations in tumors (4.4 ± 1.5 μmol/g fresh weight) compared with reference stalk tissue (1.4 ± 0.4 μmol/g fresh weight), whereas lactate contents were not significantly different in both tissues. In addition, the oxygen uptake rate at air saturation of tumor tissue was 5.7 times higher per gram fresh weight (71 ± 31 versus 12 ± 3 $\mu mol \cdot g^{-1}$ fresh weight$\cdot h^{-1}$) and 3.8 times higher per gram of soluble protein (1622 ± 207 versus 427 ± 117 $\mu mol \cdot g^{-1} \cdot h^{-1}$).

Cell Wall Formation

The majority of genes (87 of 103) involved in cell wall synthesis, degradation, or modification showed increased expression in tumor tissue, of which 25 genes were more than threefold higher

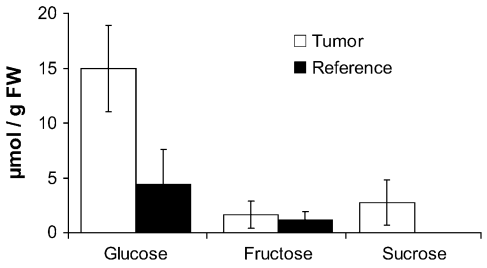


Figure 6. Sugar Content of *Arabidopsis* Tumors Induced by *Agrobacterium*.

Glucose, fructose, and sucrose were determined from tumors and tumor-free main inflorescence stalk segments (means ± SD; $n = 3$). FW, fresh weight.

expressed (see Supplemental Table 4 online). Among these were genes of the expansin family (At *EXP1* [At1g69530], P = 1.0E-04; At *EXP10* [At1g26770], P = 2.6E-03; At *EXP6* [At2g28950], P = 3.7E-03), the xyloglucosyl transferase family (At5g48070, P = 5.8E-04; At2g06850, P = 3.6E-04), the β-glucanase family (At1g70710, P = 4.6E-04; At4g02290, P = 3.4E-03), pectate lyases (At4g24780, P = 3.7E-03; At3g53190, P = 6.9E-04; At4g13210, P = 4.5E-05), cellulose synthase isoforms (At1g02730, P = 5.0E-04; At5g22740, P = 1.1E-03), members of the pectin esterases (At1g11580, P = 1.9E-04; At2g47550, P = 7.0E-04), polygalacturonase inhibitors (*PGIP1* [At5g06860], P = 3.1E-04; *PGIP2* [At5g06870], P = 2.9E-03), an α-xylosidase (At1g78060, P = 1.4E-03), a UDP-glucose-4-epimerase (At1g63180, P = 1.1E-03), and a polygalacturonase (At1g70500, P = 1.5E-04), indicating an increased reorganization and growth of cell walls.

Lipid Metabolism

Expression levels of the majority of genes involved in lipid metabolism were lower (68 of 101) in tumors than in inflorescence stalk tissue, except those belonging to the gene family of fatty acid desaturation. Genes of this family were higher transcribed in tumors, and three of five were induced even more than threefold (see Supplemental Table 4 online): a stearoyl acyl carrier protein desaturase (At1g43800, P = 7.5E-05), an ω-3 fatty acid desaturase (At2g29980, P = 4.2E-04), and a Δ-9 fatty acid desaturase (At2g31360, P = 4.4E-04). In addition, one gene of the lipid transfer protein family (LPT) showed 10-fold increased transcript levels (*LTP2* [At2g38530], P = 2.3E-04). The fact that tumors lack an intact epidermal cell layer covered by a cuticle is reflected by a 13-fold lower expression of *CUT1* (At1g68530, P = 1.5E-04) and a 4-fold lower expression of *WAX2* (At5g57800, P = 2.2E-04), two genes involved in cutin biosynthesis (see Supplemental Table 4 online).

N Metabolism

The total amino acid content in tumors was 8.4-fold higher than in inflorescence stalk tissue (Figure 7A). Among the proteinaceous amino acids, Gln was most prominently increased (14.7-fold). Furthermore, Ser, Asp, Glu, Thr, Pro, and Asn were increased severalfold in tumor tissues (6.4-, 6-, 3.1-, 6.5-, 11.7-, and 31-fold, respectively), and Ala, Val, Ile, Leu, His, and Arg were 25- to 7-fold higher in tumors than in controls.

Amino acids may accumulate in the tumor by import, de novo synthesis, and/or protein degradation. Uptake of amino acids into tumor cells might be mediated by two amino acid transporters (At1g47670, threefold, P = 8.9E-04; At1g25530, threefold, P = 5.0E-03) in addition to two H^+-dependent oligopeptide transporters (At4g21680, 17-fold, P = 1.6064E-05; At1g59740, 11-fold, P = 4.8E-04), the genes of which were strongly upregulated in tumor tissues (see Supplemental Table 7 online).

The first step in autotrophic N metabolism is the reduction of nitrate, which may be taken up by nitrate transporters into tumor cells. Transcription of two-high affinity nitrate transporter genes (At3g45060, 6-fold, P = 3.1E-04; At5g60780, 2-fold, P = 5.3E-03) was significantly induced in tumors, whereas that of low-affinity nitrate transporters, active in the millimolar range (*NTP2* [At2g26690], 0.13-fold, P = 9.5E-04; *NTP3* [At3g21670], 0.08-fold, P = 4.6E-05), was reduced severalfold (see Supplemental Table 7 online). The nitrate concentration (Figure 8A) was low in tumors (8 μmol/g fresh weight) and fivefold higher in reference tissues (41 μmol/g fresh weight). Expression of the two *Arabidopsis* nitrate reductase genes, *NR1* (At1g77760) and *NR2* (At1g37130), was 0.88 (P = 8.4E-02) and 1.45-fold (P = 6.0E-02), respectively, and thus not significantly different among both tissues. Corresponding to the nitrate content, the actual nitrate reductase enzyme activity was reduced fivefold in tumors. The maximal nitrate reductase activity was even 11-fold lower (Figure 8B). One gene encoding a mitochondrial Gly decarboxylase complex H (At2g35120), a source for photorespiratory ammonia, was 2-fold (P = 5.6E-03) higher in tumors, but two others (At2g35370, 0.13-fold, P = 9.9E-05; At1g32470, 0.53-fold, P = 7.9E-03) were strongly repressed (see Supplemental Table 6 online). Uptake of ammonium into tumor cells was most likely not facilitated by transporters, because the two differentially expressed ammonium transporter genes (*AMT2* [At4g13510], 0.54-fold, P = 2.5E-03; *AMT1.1* [At2g38290], 0.59-fold, P = 5.1E-03) and the only differentially expressed gene encoding a tonoplast-located aquaporin (*TIP2.2* [At4g17340], 0.13-fold, P = 2.8E-04) showed decreased expression levels in tumors (see Supplemental Table 7 online).

Among the 73 differentially expressed genes (P < 0.01) involved in amino acid metabolism (see Supplemental Table 4 online), only those coding for enzymes of Trp (*ASA1* [At5g05730], 3-fold, P = 5.2E-04; At5g38530, 3-fold, P = 3.2E-04) and Asp (At5g19550, 4.5-fold, P = 1.3E-04) biosynthesis were strongly induced in tumors, correlating with the increased level of Asp, but not that of Trp, which was below the detection level (Figure 7A). Although transcript levels do not necessarily correlate with enzyme activity, or with the accumulation of metabolites, this finding suggests an increased consumption of Trp in tumors, most likely as a precursor for auxin biosynthesis through enzymes expressed by the bacterial T-DNA. None of the membrane permease genes participating in auxin uptake (*AUX1* [At2g38120], *LAX1* [At5g01240], *LAX2* [At2g21050], *LAX3* [At1g77690]) or release (*PIN1* [At1g73590], *PIN2* [At5g57090], *PIN3* [At1g70940], *PIN4* [At2g01420], *PIN5* [At5g15100], *PIN6* [At1g77110], *PIN7* [At1g23080]) were significantly (P < 0.01) differentially expressed (see Supplemental Table 4 online). Genes of the ABC superfamily, coding for *p*-glycoproteins (PGPs), were recently shown to actively transport auxin (Geisler et al., 2005; Geisler and Murphy, 2006). Two genes of the PGP subfamily (At *PGP1* [At2g36910], 3-fold, P = 1.3E-03; At *PGP4* [At2g47000], 3-fold, P = 7.3E-04) were significantly expressed to a higher level in tumors (see Supplemental Table 2 online), indicating that auxin relocation in tumors might be controlled by this type of transporter.

The accumulation of Arg in tumors might be attributable in part to the differential expression (P < 0.01) of four genes involved in Arg metabolism. Three genes (At3g27740, P = 7.1E-03; At1g75330, P = 3.7E-03; At3g57560, P = 8.1E-03) coding for enzymes of Arg synthesis were upregulated (1.3- to 1.4-fold), but the only gene for Arg degradation, Arg decarboxylase 2 (At4g34710), was downregulated (0.72-fold; P = 4.4E-03) in

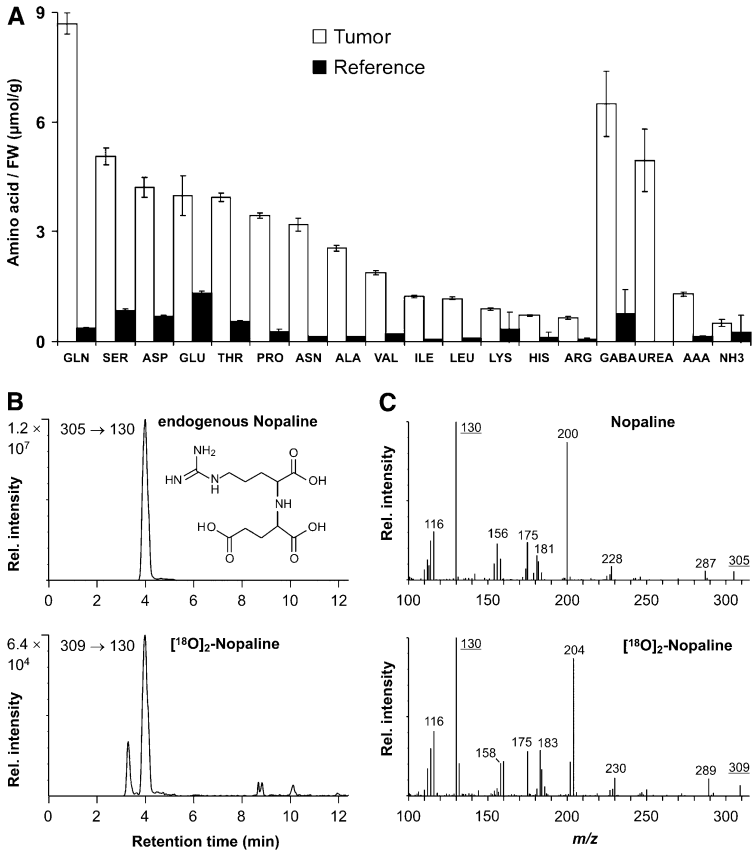


Figure 7. Content of N Components and Nopaline Determination of *Arabidopsis* Tumors Induced by *Agrobacterium*.

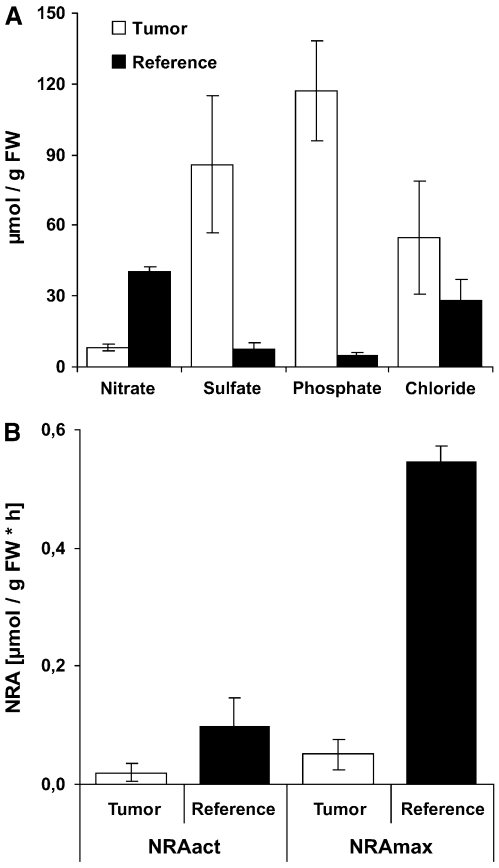
(A) Proteinaceous and nonproteinaceous amino acids were determined from tumors and tumor-free main inflorescence stalk segments (means ± SD; $n = 3$). FW, fresh weight.

(B) Multiple reaction monitoring ion chromatograms of an extracted *Arabidopsis* tumor sample. Mass chromatograms for the multiple reaction monitoring transitions at *m/z* 305→139 (endogenous nopaline) and *m/z* 309→139 ($[^{18}O]_2$-nopaline, internal standard) are shown. Endogenous nopaline levels were calculated from the ratio of the peak areas (nopaline:$[^{18}O]_2$-nopaline].

(C) Product ion spectra of $[M+H]^+$ of endogenous nopaline (*m/z* 305) and $[^{18}O]_2$-nopaline (internal standard, *m/z* 309). Retention times and spectra obtained from liquid chromatography–tandem mass spectrometry (LC-MS/MS) runs of nopaline extracted from tumor samples were identical to spectra of authentic reference compounds.

tumors (see Supplemental Table 4 online). Levels of nopaline, which is synthesized in transformed tumor cells from Arg, were determined in tumors by applying HPLC–electrospray ionization–mass spectrometry. Endogenous nopaline was unambiguously identified by its retention time and product ion spectrum of $[M+H]^+$ at *m/z* 305 (Figures 7B and 7C). Quantification in the multiple reaction monitoring mode was performed using $[^{18}O]_2$-nopaline as internal standard. Nopaline was found in *Arabidopsis* crown gall tumors in the millimolar concentration range (5.58 ± 1.89 μmol/g fresh weight; mean ± SD [$n = 4$]).

Most abundant among nonprotein N compounds were levels of γ-aminobutyric acid (GABA), α-aminoadipinic acid (AAA), and urea, which were either present in reference tissues in very low concentrations or not detectable (Figure 7). GABA accumulation might be explained by the strong induction of the Glu decarboxylase gene *GAD1* (At5g17330, sixfold, P = 3.3E-05). Urea is a product of nopaline degradation, catalyzed by arginase, an enzyme of *Agrobacterium*. Agrobacteria are present in the apoplast of growing tumors, and expression of the bacterial arginase gene within *Arabidopsis* tumors was confirmed by RT-PCR (data not shown). Uptake of urea into *Arabidopsis* tumor cells could be mediated by aquaporins, because it has been shown that at least the tobacco (*Nicotiana tabacum*) plasma membrane–located water channel, Nt AQ1, in addition to water, facilitated the transport of urea (Otto and Kaldenhoff, 2000). Two genes encoding plasma membrane–located aquaporins were severalfold higher expressed in tumors (*PIP2.5* [At3g54820], 12-fold, P = 3.3E-05; *PIP1.3* [At1g01620], 6-fold, P = 8.1E-05), whereas the urea transporter (At5g45380) was downregulated (see Supplemental Table 7 online).



C

% Anions Located in the Apoplast		
Anion	**Tumor**	**Reference**
Sulfate	20.1 ±9.6	10.8 ±4.2
Nitrate	5.9 ±10,2	6.5 ±11.2
Phosphate	21.0 ±13.3	12.3 ±6.9
Chloride	25.8 ±18.3	9.7 ±3.6

Figure 8. Anion Content and Nitrate Reductase Activity of *Arabidopsis* Tumors Induced by *Agrobacterium*.

(A) Sulfate, nitrate, phosphate, and chloride were determined from tumors and tumor-free main inflorescence stalk segments (means ± SD; $n = 3$). FW, fresh weight.

(B) Comparison of the actual nitrate reductase activity (NRAact) and maximal nitrate reductase activity (NRAmax) of tumor and inflorescence stalk tissue (means ± SD; $n = 3$).

(C) Relative proportions of apoplastic anion to total anion contents of equally sized tumor and reference stalk tissue fragments.

Among the genes for amino acid degradation, only one, an osmotic stress–induced Pro dehydrogenase (At3g30775, four-fold, P = 1.0E-03) was prominently increased in tumors. In addition, more than half of the differentially expressed genes (65 of 126; P < 0.01), encoding enzymes of ubiquitin-dependent protein degradation, showed increased transcription in tumors (see Supplemental Table 4 online).

Inorganic Anions

Concentrations of sulfate and phosphate were increased in tumors (Figure 8A). The sulfate concentration was 12-fold higher in tumors than in reference stalk tissue. The phosphate content in tumor tissue reached 117 μmol/g fresh weight and was 23-fold enriched compared with that in the stalk. Chloride levels were increased in tumors (55 μmol/g fresh weight), but not significantly compared with the control stalks (28 μmol/g fresh weight). Because tumors show increased water loss and solute flow, certain anions might accumulate in the apoplast of tumors to a higher degree than in reference stalk tissue. As a crude approximation to apoplastic anion concentrations, we determined the relative anion content of the apoplast and found that the percentage of anions (except for nitrate) washed out of the apoplast within 10 min was approximately two times higher in tumors compared with inflorescence stalk segments of the same size (Figure 8C). All genes for sulfate transporters with P < 0.01 (At3g12520, At5g13550, At1g77990, At5g10180, At1g23090, At3g51895) were downregulated, whereas two of six coding for phosphate transporters (At1g26730, At1g14040) showed increased transcript levels (see Supplemental Table 7 online). This finding could indicate that tumor cells take up phosphate preferentially.

Finally, in a bioinformatic comparison, the tumor gene expression profile of the functional categories described above (photosynthesis, cell wall, lipid metabolism, amino acid metabolism, secondary metabolism, hormone metabolism, transport, primary metabolism, and tetrapyrrole synthesis from the category of others) was compared with the transcriptome of indole acetic acid– or zeatin-treated plant tissues (see Supplemental Table 8 online). These microarray data from the RIKEN Laboratory (Japan) are available at the AtGenExpress database (The Arabidopsis Information Resource [TAIR] accession: expression sets 100796604 and 1007965859) and revealed that a number of tumor genes (11 genes, P = 0.01; see Supplemental Table 8 online) were also differentially expressed after a 3-h treatment with indole acetic acid. The transcription of only two differentially expressed tumor genes (P = 0.01) was similarly regulated by zeatin, suggesting that auxin dominates the transcriptional regulation of tumor genes within the functional categories analyzed here.

DISCUSSION

This study shows that transformation of plant cells with T-DNA of the virulent *Agrobacterium* strain C58 results in genome-wide effects reflecting the adaptation of transport and metabolism. To our knowledge, a comprehensive transcriptome analysis of a crown gall tumor that integrates data of the tumor metabolome

had not been performed previously. A transcriptome study that focused on the timing of plant responses to short-term *Arabidopsis–Agrobacterium* interactions was performed recently. It shows that already at 48 h after inoculation of *Arabidopsis* cell suspensions with *Agrobacterium*, genes of the functional categories cell wall, primary metabolism, and protein/amino acid metabolism as well as transport were differentially expressed (Ditt et al., 2006). A comparison of the 303 regulated genes found by Ditt and coworkers with those of crown gall tumors revealed that 12% of the genes, irrespective of functional clusters, were regulated in both experiments. However, only 7% of these genes were similarly regulated, either activated or repressed ($P < 0.01$), in both data sets, when comparing only the genes belonging to the same functional clusters. This divergence in gene expression profile reflects the facts that (1) the study by Ditt et al. (2006) used a different plant system, (2) a different approach for the gene expression analysis was used, and (3) the gene set involved in the metabolism and transport of crown gall tumors differs even more from those of early signals of plant–pathogen interaction.

The question whether all tumor cells are transformed or most of them are only adapted to increased phytohormone levels produced by a few transformed cells has been addressed for almost 30 years. In earlier work, when no molecular markers were available, it was found that 1.2% of the cells were transformed (Sacristan and Melchers, 1977; Ooms et al., 1982). Later, Van Slogteren (1983) calculated by cloning of isolated axenic tumor tissues that 10 to 25% of the tumor cells were transformed. Recent studies, using β-glucuronidase (gus) gene–containing wild-type bacteria (A281p35S gus-int) and RT-PCR, provided strong evidence that in *Agrobacterium*-induced tumors, most cells, or even all cells (i.e., ~100%), were transformed (Rezmer et al., 1999). Here, using the *in situ* hybridization technique, a similar result was obtained. When several images, like those presented in Figure 1, were assessed visually by staining, >95% of the tumor cells were shown to express *NOS* mRNA. However, even if all cells are transformed, expression may be prevented by epigenetic phenomena as well. In fact, it has been shown that T-DNA–encoded genes can be inactivated through methylation (Gelvin et al., 1983; Amasino et al., 1984).

In the studies presented here, the metabolite and anion concentrations determined for whole tumors do not permit conclusions about their localization in bacteria, plant cell apoplast, or symplast, or about their subcellular distribution. However, gene expression analysis reflects exclusively the response of the plant cell. This study, using two different approaches, gene expression and solute analysis, indicates changes in the whole plant cell physiology. In a model of a plant tumor cell, we summarize the major changes in gene expression of transporters and metabolic pathways (Figure 9).

Nutrient Accumulation

Rapid growth of plant tumors creates strong metabolic sinks on host plants (Malsy et al., 1992; Pradel et al., 1996, 1999; Mistrik et al., 2000). In *Arabidopsis* tumors, almost all major nutrients were at higher concentrations than in the respective host tissues (Figures 6 to 8). However, some nutrients may appear specifically accumulated as a result of agrobacteria existing in developing crown gall tumors. Agrobacteria metabolize nopaline, synthesized by the transformed tumor cell, to Glu via Arg, Orn, and Pro (Dessaux et al., 1986). With the exception of Orn, the other three amino acids are markedly accumulated in tumors. However, in 1 mL of a suspension from the *Agrobacterium* strain C58 (OD = 0.873), a concentration that we have used for the induction of tumors, all amino acids were below the detection level. Even in 1 g of a bacterial pellet, the concentrations of the stress metabolites Pro, AAA, and GABA (see below), which accumulate in tumors, were not measurable (data not shown). Arg, the precursor of nopaline, was found in the bacterial pellet at a concentration of 0.44 μmol/g bacterial pellet, still slightly lower than in tumors (0.63 μmol/g fresh weight). Because the bacterial biomass in crown galls is much lower compared with that of plant cells, the contribution of bacterial Arg to the total content of Arg should be negligible in crown galls. Nopaline, which is synthesized by the T-DNA–encoded enzyme NOS, was found in tumors in the millimolar range. Because genes encoding enzymes of Arg synthesis were only slightly upregulated in *Arabidopsis* crown galls, their Arg content results most likely from translocation by the host plant via the transpiration stream. It has been shown that the content of several amino acids, including Arg, increases in the xylem sap of tumorized plants (Mistrik et al., 2000). Thus, amino acids do not reach the tumor only via the phloem but also via the transpiration stream. Moreover, Arg accumulation appears necessary, because the K_m of purified NOS for Arg is 0.74 mM (Kemp et al., 1979). This concentration is close to that measured in *Arabidopsis* tumors (0.63 mM).

Recently it was shown that the higher potassium concentration of *Arabidopsis* tumors was accompanied by the induction of root-specific K^+ channels in favor of shoot-specific channels (Deeken et al., 2003). A three times higher total anion concentration in tumors (266 versus 81 μmol/g fresh weight) might be the result of excessive transpiration and/or a lack of retranslocation from the tumor back to the plant. It has been shown that tumors are not covered by a cuticle and have a higher transpiration rate (Schurr et al., 1996), which might also cause the strong induction of the two water channel genes (Figure 9A). Thus, a high transpiration–driven solute movement may cause the accumulation of solutes in tumors (Wachter et al., 2003). Interestingly, most differentially expressed genes of anion transporters were downregulated in tumors (Figures 9B and 9C), indicating that part of the anions might actually be located within the apoplast. This hypothesis is supported by our observation that tumors lost a higher percentage of anions after 10 min of washing (Figure 8C). In addition, a higher protein content (55 ± 12 versus 33 ± 8 mg/g fresh weight) and neutral red staining (data not shown) indicated that tumors very likely possess more plasma-rich cells with smaller vacuoles. This suggests that anions most likely accumulate in the apoplast rather than in vacuoles of tumor cells. The uptake of nutrients into tumor cells is controlled by membrane transporters, of which a number of genes were differentially expressed between *Arabidopsis* tumors and tumor-free stalk tissues (Figures 3 and 4). Active transport is fueled by ATP hydrolysis, and in fact, two P-type H^+-ATPase genes (At4g30190, At1g80660) were upregulated, whereas all vacuolar ATPases were expressed at a lower level in tumors (see Supplemental Table 7 online). The latter might again support the

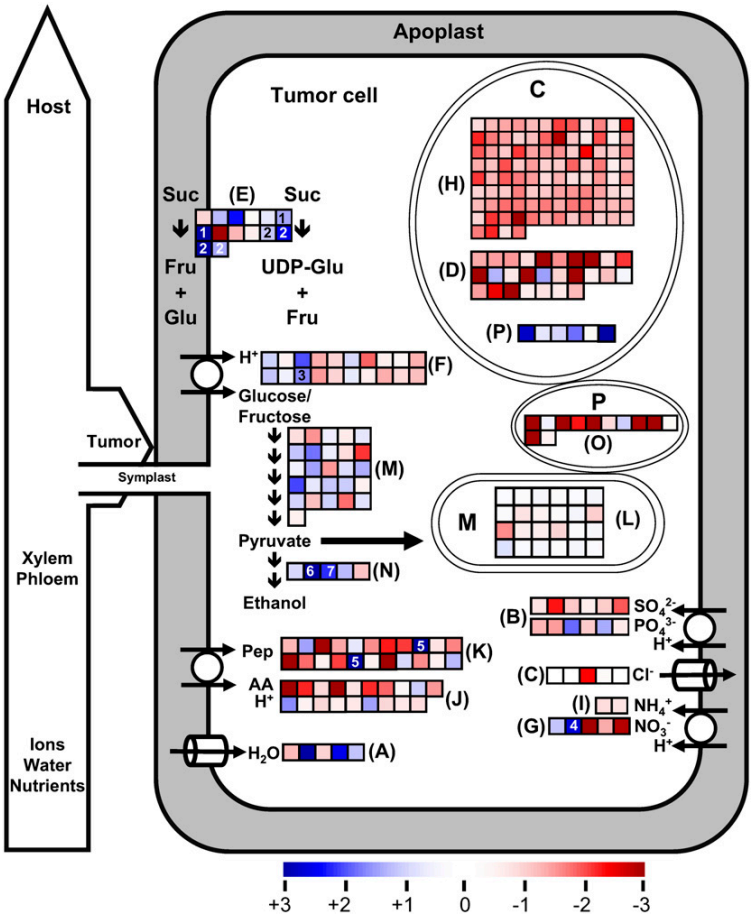


Figure 9. Scheme of Gene Expression Profiles of Transporters and Metabolic Pathways in *Agrobacterium*-Induced *Arabidopsis* Tumors.

Fold changes (log_2) of expression values of tumor versus inflorescence stalk tissues with $P < 0.01$ are presented as red (downregulated), blue (upregulated), and white (unchanged) squares for each gene, based on the pathway analysis program MapMan (https://gabi.rzpd.de/projects/MapMan/). Differential gene expression of water channels **(A)**, anion transporters **(B)**, chloride channels **(C)**, Calvin cycle enzymes **(D)**, sucrose degradation enzymes **(E)**, sugar transporters **(F)**, nitrate transporters **(G)**, light reaction enzymes **(H)**, ammonium transporters **(I)**, amino acid transporters **(J)**, peptide transporters **(K)**, electron transport enzymes **(L)**, glycolysis enzymes **(M)**, fermentation enzymes **(N)**, photorespiration enzymes **(O)**, and fatty acid desaturation enzymes **(P)** is shown. Numbers $+3$ to -3 on the color scale represent log_2 of the fold change. The flow of ions, water, and nutrients from xylem and phloem of the host plant into the tumor is indicated by the large open arrow. AA, amino acids; C, chloroplast; Fru, fructose; Glu, glucose; M, mitochondrion; P, peroxisome; Pep, peptide; Suc, sucrose; UDP-Glu, UDP-glucose.

notion that tumors possess a smaller vacuolar compartment. Another reason for metabolite accumulation might be that tumors are deprived of oxygen. It has been shown that roots that endure hypoxic stress accumulate sugars, amino acids, and GABA, as do the T-DNA–transformed tumor cells (Sousa et al., 2002).

Heterotrophic Metabolism

Glucose and amino acids appear to be the main carbon and nitrogen sources of the tumor. The C content calculated from sugars was 3.7 times higher in tumors than in tumor-free inflorescence stalks. De novo carbohydrate production should be low, as the transcription of the vast majority of the Calvin cycle genes was repressed (Figure 9D). Therefore, carbohydrates have to be supplied by the host plant, probably as sucrose via the phloem. The uptake of glucose into tumor cells is substantiated by an increased transcription of cell wall invertase and sucrose synthase (Figure 9E, 1 and 2) in addition to *STP4* (Figure 9F, 3), a sink- and pathogen-induced member of the monosaccharide transporter gene family (Truernit et al., 1996). An increased enzyme activity of acid cell wall invertase was described for tumors of *Kalanchoë*, tobacco, and *Ricinus* (Weil and Rausch, 1990; Pradel et al., 1999). It is a general phenomenon in the physiology of sink tissues to accumulate nutrients, although the expression of degrading enzymes is activated. This has been

reported by Wachter et al. (2003) for crown gall tumors of *Ricinus communis*, of which the high sucrose level in the periphery of the large tumor was accompanied by high cell wall invertase enzyme activity. These observations and our findings in *Arabidopsis* suggest that the influx of metabolites exceeds consumption in tumors.

Amino acid levels were eightfold higher in tumors compared with tumor-free tissue, whereas nitrate content was very low (8 ± 1 versus 41 ± 2 μmol/g fresh weight). Uptake of nitrate was not facilitated by anion transporters, because genes were downregulated except for two high-affinity-type transporters (Figure 9G, 4). The low nitrate content of tumors reflected the almost immeasurable nitrate reductase enzyme activity (Figure 8B). These findings confirm earlier data found by Mistrik et al. (2000), who have shown that tumors of *R. communis* have no detectable nitrate reductase activity, because of high levels of ethylene that inhibit nitrate reductase activity. In addition, abscisic acid, known to inhibit NO_3^- and PO_4^- uptake (Suleiman et al., 1990), was, as in *Ricinus* tumors (Mistrik et al., 2000), ~10 times higher in *Arabidopsis* tumors compared with uninfected inflorescence stalks (data not shown). Thus, a conclusion would be that increased amino acid levels probably are not attributable to higher nitrate assimilation. This is again substantiated by a reduced expression of genes involved in photosynthetic light reactions (Figure 9H) and by transporters for ammonium uptake (Figure 9I). The formation of ammonium from urea is most likely not favored in tumors, because transcripts of both urea-degrading enzymes appear not to be increased. Thus, nitrogen supply for amino acid and protein biosynthesis in tumors is most likely derived from Gln and Glu, which are translocated by the host plant through the vascular system and seem to be imported into tumor cells by amino acid transporters, of which three genes were induced in tumors (Figure 9J). In addition, peptide transport may provide another source of organic nitrogen. In tumors, two genes of oligopeptide transporters, one homologous with the *PTR1* gene from barley (*Hordeum vulgare*), were expressed (Figure 9K, 5). The PTR1 transporter was associated with peptide transport in germinating barley grains (West et al., 1998), which represent a sink tissue, like a tumor.

Anaerobic Energy Production in Tumors

Within the functional group photosynthesis, which includes photosynthetic light reactions, the majority of genes were strongly downregulated (Figure 9H). This implies less light-dependent oxygen production within the tumor, as confirmed by chlorophyll fluorescence. Expression of genes encoding the respiratory electron transport chain was mainly unchanged (Figure 9L) except for an uncoupling gene, At *PUMP1* (At3g54110). Suspensions of small tissue fragments from tumors gave a 5.7-fold higher oxygen uptake rate per gram fresh weight, or 3.8-fold higher on a protein basis, compared with stalk fragments. Because of their small size, these fragments in stirred solution probably were not limited by oxygen diffusion, and the higher oxygen uptake of tumor fragments over stalk fragments might reflect an uncoupling of respiratory electron transport. However, an intact tumor lacks intracellular air spaces. Thus, as a result of diffusional limitation together with an increased respiratory electron transport capacity, cells in the tumor core may easily become hypoxic. Under such conditions, plant cells switch to fermentative energy metabolism (Tadege et al., 1999). Genes coding for enzymes of the glycolytic pathway (Figure 9M) and ethanolic fermentation (Figure 9N) appear to prevail in tumors. Fructose seems to be the carbohydrate fed into the glycolytic pathway of tumors, because transcripts of two of three fructokinases (At2g31390, At4g10260) were 5.2-fold (P = 3.8E-04) and 2.4-fold (P = 1.3E-04) increased, whereas none of the genes that encode glucose-using enzymes was differentially expressed (see Supplemental Table 6 online). In addition, glucose was highly enriched in tumors, but fructose and sucrose contents were low (Figure 6). An increased ethanol level and the induction of *PDC1* transcripts (Figure 9N, 6) and *ADH* (Figure 9N, 7) confirm the switch to fermentation. The *PDC1* gene encodes the main regulatory enzyme of ethanolic fermentation and is also induced by abscisic acid (Kursteiner et al., 2003). In addition, it has been shown that genes involved in alcoholic fermentation, such as *ADH* (At1g77120), *PDC1* (At4g33070), and *PDC2* (At5g54960), showed a dramatic increase in expression under low-oxygen conditions in *Arabidopsis* roots (Klok et al., 2002). Both hypoxia and high abscisic acid levels might add to the induction of *PDC1* in tumors. ADH is also strongly induced by abscisic acid. These results imply that transformation of plant cells with T-DNA of the virulent *Agrobacterium* strain C58 is accompanied by a change from autotrophic to heterotrophic metabolism, in which ATP production is powered mainly by glycolysis and fermentation.

Stress Metabolites in Tumors

In addition to increased abscisic acid levels (see above), other stress metabolites, such as Pro, GABA, and AAA, were also strongly accumulated in *Arabidopsis* tumor cells (Figure 7) but were not measurable in pure agrobacteria. AAA might accumulate through the catabolism of Lys by the saccharopine pathway, which is important for the regulation of Lys homeostasis (Karchi et al., 1994). AAA is supposed to regulate growth, development, and responses to environmental changes by regulating the expression of genes involved in nitrogen metabolism (Arruda et al., 2000). In the case of osmotic stress responses, Glu, which is generated during Lys catabolism, might also act as a precursor of enhanced Pro biosynthesis. The increased Pro content (12-fold) was correlated with a strong repression of Pro oxidase (At5g38710). However, Pro dehydrogenase (At3g30775), another Pro-degrading enzyme, was induced, most likely as a result of increased Pro concentrations. Pro levels appear tightly controlled through feedback regulation (Kiyosue et al., 1996; Peng et al., 1996). GABA is probably increased as a consequence of anaerobic conditions within the tumor tissue, because oxygen deprivation and the resulting cellular acidosis strongly induce GABA accumulation (Kinnersley and Turano, 2000).

Signals derived from increased sugar levels lead to the inhibition of genes involved in photosynthesis, the Calvin cycle, and chlorophyll synthesis (Sheen et al., 1999; Pego et al., 2000; Smeekens, 2000). In *Arabidopsis* tumors, the vast majority of genes for light reactions (Figure 9H), the Calvin cycle (Figure 9D), and photorespiration (Figure 9O) show reduced transcription, in addition to genes encoded by the chloroplast genome (see

Supplemental Table 6 online). The latter reflects the reduced number of chloroplasts in tumor tissue. Whether that was attributable to the increased glucose remains unclear. The recent discovery of genes encoding enzymes of trehalose metabolism in higher plants has revealed trehalose-6-phosphate synthase and/or its product, trehalose-6-phosphate, as another potential player in sugar sensing (Rolland et al., 2001; Eastmond and Graham, 2003; Eastmond et al., 2003). Trehalose-6-phosphate synthase, of which the transcript levels of two members were strongly increased in *Arabidopsis* tumors (At1g23870, At2g18700), is required for embryo maturation and might also control developmental processes.

Expression of genes for the modification of fatty acids, such as desaturases (Figure 9P), and the transport of lipids, such as the lipid transfer protein LTP2 (At2g38530), was strongly increased in tumors. Within this group, some are involved in pathogen defense signaling. Stearoyl acyl-carrier protein desaturase (S-ACP-Des) catalyzes the initial step in fatty acid desaturation to form oleic acid. This monounsaturated fatty acid serves as a stimulator of phospholipase Dδ, which was shown to prevent cell death in parsley (*Petroselinum crispum*) suspension cells upon pathogen elicitation (Kirsch et al., 1997; Ryu, 2004; Wang, 2004) and modulates the activation of defense signaling pathways in plants (Kachroo et al., 2001, 2003, 2004). Activation of the S-ACP-Des gene in *Arabidopsis* tumors may help to prevent a hypersensitive response in defense against *Agrobacterium. Defective in Induced Resistance1* encodes a putative apoplastic LTP that is involved in systemic, but not local, resistance to pathogens (Maldonado et al., 2002). LTPs could also play a major role in cell wall modification. In tumors, they might shuttle lipids such as suberine monomers from their sites of biosynthesis through the plasma membrane into the cell wall, as suggested by Kunst and Samuels (2003), to minimize loss of water.

Auxin and Cytokinin May Control the Expression of Genes Involved in Tumor Metabolism

The differential expression of several of the genes discussed here may be regulated by auxin and cytokinin, two phytohormones that are known to be increased in crown gall tumors. Tumor cells are not only exposed to increased auxin and cytokinin levels for weeks but also to high levels of abscisic acid, ethylene, or jasmonic acid (Veselov et al., 2003). However, a comparison with the transcriptome of plant cells, treated with auxin and cytokinin for 3 h, indicated that the expression of at least 13 tumor genes (see Supplemental Table 8 online) might be regulated by auxin and cytokinin. The majority of these genes (10 of 13) are involved in phytohormone metabolism or signaling.

In conclusion, we have shown that plant tumors are characterized by anaerobic and heterotrophic metabolism and display an altered differentiation with modified, tissue type–specific gene expression patterns for photosynthesis, amino acid, cell wall, and lipid metabolism as well as for solute transporters. The transcription of several of these genes might be regulated by auxin and cytokinin. Metabolic changes and altered metabolite signaling seem to maintain vigorous growth of plant tumors after intrusion and successful transformation by agrobacteria.

METHODS

Plant Material and RNA Preparation

Arabidopsis thaliana plants (ecotype Ws-2) were cultivated in growth chambers under short-day conditions (8 h of light) at 22°C and 16°C during the dark period (16 h). Tumors were induced by applying the nopaline-using *Agrobacterium tumefaciens* strain C58 (noc^c) to the base of a wounded, very young inflorescence stalk (2 to 5 cm). At 35 d after infection, tumor tissue was separated from the host inflorescence stalk using a scalpel. Wounded but uninfected tumor-free inflorescence stalk segments of the same age served as reference tissues. To reduce data variation, total RNA was prepared from four independent biological replicates and used in four separate microarray hybridizations. Each replicate consisted of material from 10 to 12 individual plants. Total RNA was extracted from tumor and inflorescence stalk tissues and treated with DNase using the RNeasy plant mini kit (Qiagen) according to the manufacturer's protocol.

Probe Synthesis and in Situ Hybridization

Probes were generated by PCR using a 460-bp fragment of *NOS* cDNA and the following primers carrying a T7- or T3-RNA polymerase binding site at the 5′ end: NOSas-T7, 5′-CTTCTTTACCTATTTCCGCC-3′; NOSs-T3, 5′-TGATCCGATAGCTTAGACG-3′. Labeling with digoxygenin-11-dUTP of sense and antisense RNA strands was performed with the DIG-RNA labeling mix, applying either T7- or T3-RNA polymerase, respectively, according to the manufacturer's protocol (Roche Diagnostics). Labeled probes were dissolved in 100 μL of water. For hybridization, pieces of tumors with adjacent stalks were fixed in PBS + 4% paraformaldehyde at 4°C overnight, dehydrated in a series of increasing concentrations of ethanol and Histo-Clear (National Diagnostics), and finally embedded in paraffin at 60°C. Embedded material was cut with a microtome (Leica RM2245) in 9-μm sections and transferred to charged slides (Cnops et al., 2006). Samples were inspected with an inverted microscope (Axiovert 200M; Zeiss) and photographed with a digital camera (Axiocam MRC; Zeiss), applying the AxioVision LE software (Zeiss).

Microarrays and Data Preprocessing

A total of eight microarray slides (ATH1 121501; Affymetrix) containing the almost complete genome of *Arabidopsis* were used to monitor differentially expressed genes in tumor and inflorescence stalk tissue. Two different laboratories conducted two microarray hybridizations of each tissue type: (1) Nottingham Arabidopsis Stock Centre, Plant Science Division School of Biosciences, University of Nottingham, UK (http://www.york.ac.uk/res/garnet/providers); and (2) VBC-Genomics Bioscience Research, Vienna, Austria (www.vbc-genomics.com). Altogether, four arrays were hybridized with four different samples of tumor RNA and four with four different samples of inflorescence stalk RNA.

Data preprocessing was performed using Bioconductor software (Huber et al., 2002; Gentleman et al., 2004) running under the statistical programming environment R (Ihaka and Gentleman 1996). To obtain a normalized gene expression value from oligomer intensities for each gene of each microarray slide, variance stabilization (Huber et al., 2002) was applied. Variance stabilization calibrates for variations between the arrays through shifting and scaling and transforms the intensities to a scale on which the variance is approximately independent of the mean intensity. Before applying variance stabilization, no background correction was performed to the Affymetrix probe intensities, according to recommendations in the variance stabilization manual, and only the perfect match probes were used to compute the expression values for individual genes.

For summarization of probe intensities into gene expression values, we applied the median polish algorithm (D. Holder, R.F. Raubertas, V.B. Pikounis, V. Svetnik, and K. Soper, unpublished data), which is also incorporated into the commonly used robust multiarray analysis by Irizarry et al. (2003).

Assessing the Quality of the Data

To examine the quality of the microarray data, we applied three independent methods: two statistical methods (scatterplot and correspondence analysis) and one biological method (quantitative RT-PCR). First, the reproducibility of the individual Affymetrix microarray hybridizations was checked by scatterplot analyses. Normalized expression values of all genes of one microarray were plotted versus the expression values of another microarray. The scatterplots comparing two uninfected inflorescence stalk or tumor tissues with each other (see Supplemental Figure 4 online) displayed the variability of repeated measurements. The average correlation coefficients for the six scatterplots each of tumor and reference stalk arrays were 0.9725238 and 0.9623506, respectively. Thus, expression signals of genes from the same tissue type showed high consistency. The 12 scatterplots resulting from a comparison of tumor versus reference arrays indicated differentially expressed genes (see Supplemental Figure 5 online, red dots). Here, the average correlation coefficient was 0.86326 and indicated that many more genes differ between two microarray hybridizations of different tissue types.

In the next step, the reproducibility of chip hybridizations was confirmed by applying correspondence analysis. Correspondence analysis was conducted using a self-made script within MATLAB (MathWorks). Correspondence analysis represents genes as numerical vectors, with the number of elements of a vector being the number of microarray assays considered. Those vectors are projected into a lower dimensional subspace (typically, two dimensions) that accounts for the main variance in the data such that distances among points reflect their original distances in the high-dimensional space as closely as possible (Fellenberg et al., 2001). The same reduction of dimensions is done for the microarray assays; here, a 22,810-dimensional vector (of genes) is reduced to lower dimensions. By embedding both genes and assays in the same graph, correspondence analysis finds the most pronounced factor of differences between genes and microarray hybridizations.

Differential Gene Expression Analysis

To give a first graphic overview of differential gene expression between tumor and inflorescence stalk tissue, we performed an MA plot on gene expression data (Figure 5). In such an MA plot, the difference of log expression values (Minus) of the two tissue types [$M = \log(\text{tumor}) - \log(\text{reference})$] is plotted against the sum (Add) of the log expression values divided by 2 [$A = \{\log(\text{tumor}) + \log(\text{reference})\}/2$]. Thus, the *x* axis represents the extent of expression levels and the *y* axis represents differential gene expression.

For the statistical evaluation of differential expression between the two tissue types, we used a moderate *t* statistic implemented in the LIMMA package, which is part of the Bioconductor software project (Gentleman et al., 2004; Smyth, 2004). We applied the function lmFit() in the LIMMA software package to fit linear models on the expression values of each gene with the factors tissue type and laboratory. The function eBayes() was used to compute moderated *t* statistics by empirical Bayes shrinkage of the standard errors toward a common value. The advantage of the LIMMA package is its robustness and suitability for experiments with small sample numbers. Four repeated microarray hybridizations of each tissue type are not enough repeats for stable predictions using standard statistical *t* test analyses. To circumvent this limitation, the Bayesian functions were applied, exploiting information across genes and balancing the lack of more repeats needed for a classical *t* test. Thus, analyses with the LIMMA package are still stable with a small number of arrays (Smyth, 2004). The null hypothesis of differences between tissues being equal to zero was tested under the assumption of independent errors following a normal distribution. For each gene, we calculated a fold change and a P value measuring the statistical significance of differential expression. The significance level was corrected for multiple testing by applying the false discovery rate from Benjamini and Hochberg (2000). All of the P values given are corrected for multiple testing.

Fold changes of significantly differentially expressed genes ($P < 0.01$) were analyzed with the pathway analysis program MapMan. MapMan is a user-driven tool that displays large data sets (e.g., gene expression data from Affymetrix microarrays) onto diagrams of metabolic pathways or other processes (Thimm et al., 2004; https://gabi.rzpd.de/projects/MapMan/). A color code symbolizes the fold change of differential gene expression, where blue indicates higher expression in tumors and red indicates higher expression in inflorescence stalk tissue (Figure 9).

Comparison of Tumor- and Phytohormone-Dependent Gene Expression

Differential gene expression discussed for crown gall tumors was compared with two Affymetrix microarray data sets addressing differential expression arising from phytohormone treatments. Both data sets were produced by the RIKEN Laboratory (Japan) and are available at the AtGenExpress database. One data set includes the comparison of seedlings treated with indole acetic acid for 3 h with untreated seedlings (TAIR accession: expression set 100796604), the other set compares gene expression of seedlings treated with zeatin for 3 h with untreated seedlings (TAIR accession: expression set 1007965859). In both microarray data sets, there are two biological replicates of each treatment group, leading to a total of four microarray assays per data set. We analyzed the raw data (Affymetrix CEL files) in the same way as the tumor data, using variance stabilization normalization and LIMMA for differential gene expression analysis. Consistent with the P criterion for the crown gall tumor gene expression data set, genes with a multiple testing corrected $P < 0.01$ were considered differentially expressed. Although the lower number of replicates in the phytohormone data sets results in higher P values in the differential gene expression analysis, the P criterion was kept constant for consistency.

Quantitative Real-Time RT-PCR

Total RNA was extracted from tumor and stalk tissue with the plant RNeasy extraction kit (Qiagen). Poly(A)$^+$ RNA was isolated from total RNA with Dynabeads according to the protocol of the Dynabeads mRNA Direct kit (Dynal). To eliminate contamination with genomic DNA, poly(A)$^+$ RNA samples were treated twice with Dynabeads. First-strand cDNA synthesis and quantitative real-time RT-PCR experiments were performed as described previously (Szyroki et al., 2001) using LIGHTCYCLER 3.1 (Roche). Primers used were as follows: cytokinin oxidase (At5g56970), 5′-GATAGTTTAAACCATGT-3′ (forward), 5′-CAAAACTTCAGTATTTCC-3′ (reverse), 390 bp; wound-induced protein (At4g10270), 5′-TGGAACTACATACTCCG-3′ (forward), 5′-AATTTGAGTCACATTGAT-3′ (reverse), 316 bp; glycosyl hydrolase (At1g66280), 5′-GACACAACTACATTTGGA-3′ (forward), 5′-AACAGCAACAGAATCT-3′ (reverse), 390 bp; receptor protein kinase (At1g51805), 5′-TGGTTCTGTGTGGAAA-3′ (forward), 5′-AATCTACCTAGCCATTG-3′ (reverse), 214 bp; 2,4-D–inducible glutathione *S*-transferase (At1g78370), 5′-TTATTGAGGCAGTGAAG-3′ (forward), 5′-CGCATTATTAGGGGAA-3′ (reverse), 352 bp; Ser carboxypeptidase I (At2g22990), 5′-GGATCCATCTAACACAC-3′ (forward), 5′-AAGCTCTCGTGTATCCA-3′ (reverse), 446 bp. The number of transcripts was normalized to the constitutively expressed Actin2/8 mRNA (An et al., 1996).

Measurements of Solutes

Contents of amino acids, sugars, and anions were determined from aqueous extracts of 20 mg of fresh tumor or inflorescence stalk tissue. Amino acids were quantified with an amino acid analyzer (LC5001, Biotronic; Eppendorf-Nethler-Hinz), sugars by HPLC (Dionex series 4500i chromatography system), and anions by ion chromatography (IC 1000; Biotronic). For extraction of apoplastic anions, fresh tumor or inflorescence stalk tissues were cut into small pieces (2 to 3 mm) with a razor blade and briefly incubated in a large volume of water for 10 to 15 s to remove contamination of destroyed cells on the surface of the tissue. After incubation for 10 min in 1 mL of deionized water, the anion content of the washing solution was measured. Ethanol and lactate were determined enzymatically from 70 to 100 mg of fresh plant tissue (Roche Diagnostics).

Quantification of Nopaline

Synthesis of $[^{18}O]_2$-Nopaline Standard

Labeled nopaline was synthesized through an acid-catalyzed oxygen-exchange reaction according to a previously described procedure (Mueller et al., 2006). Briefly, unlabeled nopaline (25 μg) was dissolved in 50 μL of $H_2^{18}O$ (99 atom % ^{18}O; Isotec). After addition of 50 μL of a 4 M HCl solution in 1,4-dioxane (premade solution; Aldrich Chemicals), the sample was incubated in a tightly closed screw-cap vial for 1 h at 60°C. Thereafter, the sample was dried in vacuum, dissolved in methanol, and stored at −20°C. Theoretically, all six oxygens of the three carboxyl groups of nopaline can be exchanged by ^{18}O through this procedure. Because of the instability of nopaline in the acid-exchange medium, incubation was terminated after 1 h when the majority of the recovered nopaline was labeled with two ^{18}O atoms and unlabeled nopaline became undetectable. In addition to $[^{18}O]_2$-nopaline, the mixture also contained labeled nopaline molecules with one to four ^{18}O atoms; these, however, did not interfere with the analysis. $[^{18}O]_2$-Nopaline was quantified against unlabeled nopaline and used as an internal standard.

Plant Extraction and LC-MS/MS Quantitation of Nopaline

For nopaline analysis, frozen plant material (50 mg) was mixed with 750 μL of a pre-warmed mixture (75°C) of methanol:water (75:25, v/v). After addition of a ceramic bead (6 mm in diameter), the tissue was homogenized and extracted using a vibrating ball mill for 1 min. After an incubation period of 1 min at 75°C, 250 ng of the internal standard, $[^{18}O]_2$-nopaline, was added and homogenization was repeated. Thereafter, the sample was centrifuged (1000*g* for 10 min), and the supernatant was dried in a vacuum centrifuge at 50°C. The residue was suspended in acetonitrile:water (20:80, v/v). After centrifugation (1000*g* for 2 min), the supernatant was transferred into an HPLC vial, and 10 μL was injected into the LC-MS/MS system. LC-MS/MS analyses were performed using a 1200 Agilent HPLC system coupled to a Micromass Quattro Premier triple-quadrupole mass spectrometer (Waters). The column (Phenomenex Synergi Hydro-RP, 150 × 4.6 mm, particle size, 4 μm) was eluted with a linear mobile phase gradient (0.5 mL/min flow rate) starting from water containing 0.1% formic acid at 0 min to acetonitrile:water:formic acid (20:80:0.1, v/v) at 10 min. The mass spectrometer was operated in the ESI+ mode using multiple reaction monitoring. Argon was used as collision gas (22 eV of collision energy).

Pulse Amplitude–Modulated Measurements and Determination of Chlorophyll Content

The relative quantum efficiency of chlorophyll fluorescence was measured with 30-d-old tumors and inflorescence stalks using a MINI-PAM photosynthesis yield analyzer (Heinz Walz). Chlorophyll content was determined according to Arnon (1949).

Measurements of Oxygen Uptake and Protein Content

Polarographic measurements of oxygen uptake were performed with a Clark-type oxygen electrode (Hansatech Instruments). Tumor and inflorescence stalk tissue fragments of 2 to 3 mm were submerged in 10 mM $CaSO_4$, and their O_2 consumption was recorded for 8 min. Three samples were measured, containing tissue fragments from at least three plants. Protein content of tumor and inflorescence stalk tissue was determined using the BSA protein assay (Pierce).

Determination of Nitrate Reductase Activity

Nitrate reductase activity was determined from ~300 mg of frozen tissue of tumors without stalk or tumor-free stalks as described previously (Kaiser and Brendle-Behnisch 1995).

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure 1. Distribution of All Differentially Expressed Genes within Functional Categories.

Supplemental Figure 2. Distribution of Differentially Expressed Genes of Three Functional Categories within Functional Subcategories.

Supplemental Figure 3. Verification of Differentially Expressed Genes by Quantitative RT-PCR after Affymetrix ATH1 Microarray Analysis.

Supplemental Figure 4. Verification of the Reproducibility of Affymetrix ATH1 Microarray Analysis.

Supplemental Figure 5. Comparison of Expression Values of All Affymetrix ATH1 Microarray Slides Applying Scatterplot Analysis.

Supplemental Table 1. Differentially Expressed Genes Upregulated in *Agrobacterium*-Induced Tumors of *Arabidopsis*.

Supplemental Table 2. Differentially Expressed Genes Downregulated in *Agrobacterium*-Induced Tumors of *Arabidopsis*.

Supplemental Table 3. Three Subtables of Differentially Expressed Genes of *Agrobacterium*-Induced *Arabidopsis* Tumors versus Tumor-Free Inflorescence Stalk Tissue.

Supplemental Table 4. Functional Categories of Differentially Expressed Genes Calculated from *Arabidopsis* Tumors Induced by *Agrobacterium* versus Reference Inflorescence Stalks.

Supplemental Table 5. Subcategories of Differentially Expressed Genes of the Functional Category of Primary Metabolism.

Supplemental Table 6. Subcategories of Differentially Expressed Genes of the Functional Category of Photosynthesis.

Supplemental Table 7. Subcategories of Differentially Expressed Genes of the Functional Category of Transport.

Supplemental Table 8. Comparison of Tumor- and Phytohormone-Dependent Gene Expression.

ACKNOWLEDGMENTS

We are grateful to Sean T. May (University of Nottingham, UK) for performing Affymetrix (ATH1) gene chip analyses. Special thanks go to

M. Lesch, E. Reissberg, E. Wirth, and J. Schwartz (Julius-von-Sachs-Institute, University of Wuerzburg) for excellent technical support on enzyme activity and metabolite measurements as well as to Biju Joseph (Department of Microbiology, University of Wuerzburg) for reading the manuscript. Finally, we thank M. Stitt (Max-Planck-Institute for Molecular Plant Physiology, Golm) for fruitful discussions. For generous financial support, we thank the Deutsche Forschungsgemeinshaft (SPP1063 project BO 1099/5 and SFB567 project B5), the Bundesministerium für Bildung und Forschung (IZKF B-36), and the European Union Biotechnology Program.

Received June 8, 2006; revised October 20, 2006; accepted November 10, 2006; published December 15, 2006.

REFERENCES

Aisenberg, A.C. (1961). The Glycolysis and Respiration of Tumors. (New York: Academic Press).

Aloni, R., Schwalm, K., Langhans, M., and Ullrich, C.I. (2003). Gradual shifts in sites of free-auxin production during leaf-primordium development and their role in vascular differentiation and leaf morphogenesis in *Arabidopsis*. Planta **216,** 841–853.

Amasino, R.M., Powell, A.L.T., and Gordon, M.P. (1984). Changes in T-DNA methylation and expression are associated with phenotypic variation and plant-regeneration in a crown gall tumor line. Mol. Gen. Genet. **197,** 437–446.

An, Y.Q., McDowell, J.M., Huang, S.R., McKinney, E.C., Chambliss, S., and Meagher, R.B. (1996). Strong, constitutive expression of the *Arabidopsis* ACT2/ACT8 actin subclass in vegetative tissues. Plant J. **10,** 107–121.

Arnon, D.I. (1949). Copper enzymes in isolated chloroplasts—Polyphenoloxidase in *Beta vulgaris*. Plant Physiol. **24,** 1–15.

Arruda, P., Kemper, E.L., Papes, F., and Leite, A. (2000). Regulation of lysine catabolism in higher plants. Trends Plant Sci. **5,** 324–330.

Benjamini, Y., and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. Journal of Educational and Behavioral Statistics **25,** 60–83.

Carmeliet, P., and Jain, R.K. (2000). Angiogenesis in cancer and other diseases. Nature **407,** 249–257.

Chang, S., Lee, S., Lee, C., Kim, J.I., and Kim, Y. (2000). Expression of the human erythrocyte glucose transporter in transitional cell carcinoma of the bladder. Urology **55,** 448–452.

Chilton, M.D., Drummond, M.H., Merio, D.J., Sciaky, D., Montoya, A.L., Gordon, M.P., and Nester, E.W. (1977). Stable incorporation of plasmid DNA into higher plant cells: The molecular basis of crown gall tumorigenesis. Cell **11,** 263–271.

Cnops, G., et al. (2006). The TORNADO1 and TORNADO2 genes function in several patterning processes during early leaf development in *Arabidopsis thaliana*. Plant Cell **18,** 852–866.

Deeken, R., Ivashikina, N., Czirjak, T., Philippar, K., Becker, D., Ache, P., and Hedrich, R. (2003). Tumour development in *Arabidopsis thaliana* involves the Shaker-like K^+ channels AKT1 and AKT2/3. Plant J. **34,** 778–787.

Dessaux, Y., Petit, A., Tempe, J., Demarez, M., Legrain, C., and Wiame, J.M. (1986). Arginine catabolism in *Agrobacterium* strains—Role of the Ti-plasmid. J. Bacteriol. **166,** 44–50.

Ditt, R.F., Kerr, K.F., de Figueiredo, P., Delrow, J., Comai, L., and Nester, E.W. (2006). The *Arabidopsis thaliana* transcriptome in response to *Agrobacterium tumefaciens*. Mol. Plant Microbe Interact. **19,** 665–681.

Eastmond, P.J., and Graham, I.A. (2003). Trehalose metabolism: A regulatory role for trehalose-6-phosphate? Curr. Opin. Plant Biol. **6,** 231–235.

Eastmond, P.J., Li, Y., and Graham, I.A. (2003). Is trehalose-6-phosphate a regulator of sugar metabolism in plants? J. Exp. Bot. **54,** 533–537.

Fellenberg, K., Hauser, N.C., Brors, B., Neutzner, A., Hoheisel, J.D., and Vingron, M. (2001). Correspondence analysis applied to microarray data. Proc. Natl. Acad. Sci. USA **98,** 10781–10786.

Folkman, J. (1971). Tumor angiogenesis: Therapeutic implications. N. Engl. J. Med. **285,** 1182–1186.

Geisler, M., et al. (2005). Cellular efflux of auxin catalyzed by the *Arabidopsis* MDR/PGP transporter AtPGP1. Plant J. **44,** 179–194.

Geisler, M., and Murphy, A.S. (2006). The ABC of auxin transport: The role of p-glycoproteins in plant development. FEBS Lett. **580,** 1094–1102.

Gelvin, S.B., Karcher, S.J., and Dirita, V.J. (1983). Methylation of the T-DNA in *Agrobacterium tumefaciens* and in several crown gall tumors. Nucleic Acids Res. **11,** 159–174.

Gentleman, R., et al. (2004). Bioconductor: Open software development for computational biology and bioinformatics. Genome Biol. **5:R80** (http://genomebiology.com/2004–5/10/R80).

Gimbrone, M.A., Jr., Cotran, R.S., Leapman, S.B., and Folkman, J. (1974). Tumor growth and neovascularization: An experimental model using the rabbit cornea. J. Natl. Cancer Inst. **52,** 413–427.

Huber, W., Von Heydebreck, A., Sültmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics **18,** S96–S104.

Ihaka, R., and Gentleman, R. (1996). R: A language for data analysis and graphics. J. Comput. Graph. Stat. **5,** 299–314.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics **4,** 249–264.

Kachroo, A., Lapchyk, L., Fukushige, H., Hildebrand, D., Klessig, D., and Kachroo, P. (2003). Plastidial fatty acid signaling modulates salicylic acid- and jasmonic acid-mediated defense pathways in the Arabidopsis *ssi2* mutant. Plant Cell **15,** 2952–2965.

Kachroo, A., Venugopal, S.C., Lapchyk, L., Falcone, D., Hildebrand, D., and Kachroo, P. (2004). Oleic acid levels regulated by glycerolipid metabolism modulate defense gene expression in *Arabidopsis*. Proc. Natl. Acad. Sci. USA **101,** 5152–5157.

Kachroo, P., Shanklin, J., Shah, J., Whittle, E.J., and Klessig, D.F. (2001). A fatty acid desaturase modulates the activation of defense signaling pathways in plants. Proc. Natl. Acad. Sci. USA **98,** 9448–9453.

Kado, C.I. (1984). Phytohormone-mediated tumorigenesis by plant pathogenic bacteria. In Genes Involved in Microbe-Plant Interactions, D.P.S. Verma and T. Hohn, eds (Heidelberg, Germany: Springer Verlag), pp. 311–336.

Kaiser, W.M., and Brendle-Behnisch, E. (1995). Acid-base-modulation of nitrate reductase in leaf tissues. Planta **196,** 1–6.

Karchi, H., Shaul, O., and Galili, G. (1994). Lysine synthesis and catabolism are coordinately regulated during tobacco seed development. Proc. Natl. Acad. Sci. USA **91,** 2577–2581.

Kemp, J.D., Sutton, D.W., and Hack, E. (1979). Purification and characterization of the crown gall specific enzyme nopaline synthase. Biochemistry **18,** 3755–3760.

Kinnersley, A.M., and Turano, F.J. (2000). Gamma aminobutyric acid (GABA) and plant responses to stress. CRC Crit. Rev. Plant Sci. **19,** 479–509.

Kirsch, C., Takamiya-Wik, M., Reinold, S., Hahlbrock, K., and Somssich, I.E. (1997). Rapid, transient, and highly localized induction of plastidial omega-3 fatty acid desaturase mRNA at fungal infection sites in *Petroselinum crispum*. Proc. Natl. Acad. Sci. USA **94,** 2079–2084.

Kiyosue, T., Yoshiba, Y., Yamaguchi-Shinozaki, K., and Shinozaki, K. (1996). A nuclear gene encoding mitochondrial proline dehydrogenase, an enzyme involved in proline metabolism, is upregulated by proline but downregulated by dehydration in *Arabidopsis*. Plant Cell **8,** 1323–1335.

Klok, E.J., Wilson, I.W., Wilson, D., Chapman, S.C., Ewing, R.M., Somerville, S.C., Peacock, W.J., Dolferus, R., and Dennis, E.S. (2002). Expression profile analysis of the low-oxygen response in *Arabidopsis* root cultures. Plant Cell **14,** 2481–2494.

Kunst, L., and Samuels, A.L. (2003). Biosynthesis and secretion of plant cuticular wax. Prog. Lipid Res. **42,** 51–80.

Kursteiner, O., Dupuis, I., and Kuhlemeier, C. (2003). The Pyruvate decarboxylase1 gene of Arabidopsis is required during anoxia but not other environmental stresses. Plant Physiol. **132,** 968–978.

Maldonado, A.M., Doerner, P., Dixon, R.A., Lamb, C.J., and Cameron, R.K. (2002). A putative lipid transfer protein involved in systemic resistance signalling in *Arabidopsis*. Nature **419,** 399–403.

Malsy, S., Van Bel, A.J.E., Kluge, M., Hartung, W., and Ullrich, C.I. (1992). Induction of crown galls by *Agrobacterium tumefaciens* (strain C58) reverse assimilate translocation and accumulation in *Kalanchoë daigremontiana*. Plant Cell Environ. **15,** 519–529.

Mistrik, I., Pavlovkin, J., Wachter, R., Pradel, K.S., Schwalm, K., Hartung, W., Mathesius, U., Stohr, C., and Ullrich, C.I. (2000). Impact of *Agrobacterium tumefaciens*-induced stem tumors on NO_3-uptake in *Ricinus communis*. Plant Soil **226,** 87–98.

Mueller, M.J., Mene-Saffrane, L., Grun, C., Karg, K., and Farmer, E.E. (2006). Oxylipin analysis methods. Plant J. **45,** 472–489.

Ooms, G., Bakker, A., Molendijk, L., Wullems, G.J., Gordon, M.P., Nester, E.W., and Schilperoort, R.A. (1982). T-DNA organization in homogeneous and heterogeneous octopine-type crown gall tissues of *Nicotiana tabacum*. Cell **30,** 589–597.

Otto, B., and Kaldenhoff, R. (2000). Cell-specific expression of the mercury-insensitive plasma-membrane aquaporin NtAQP1 from *Nicotiana tabacum*. Planta **211,** 167–172.

Pedersen, P.L. (1978). Tumor mitochondria and the bioenergetics of cancer cells. Prog. Exp. Tumor Res. **22,** 190–274.

Pego, J.V., Kortstee, A.J., Huijser, G., and Smeekens, S.G.M. (2000). Photosynthesis, sugars and the regulation of gene expression. J. Exp. Bot. **51,** 407–416.

Peng, Z., Lu, Q., and Verma, D.P. (1996). Reciprocal regulation of delta 1-pyrroline-5-carboxylate synthetase and proline dehydrogenase genes controls proline levels during and after osmotic stress in plants. Mol. Gen. Genet. **253,** 334–341.

Pradel, K.S., Rezmer, C., Krausgrill, S., Rausch, T., and Ullrich, C.I. (1996). Evidence for symplastic phloem unloading with concomitant high activity of acid cell wall invertase in *Agrobacterium tumefaciens*-induced plant tumors. Bot. Acta **109,** 397–404.

Pradel, K.S., Ullrich, C.I., Santa Cruz, S., and Oparka, K.J. (1999). Symplastic continuity in *Agrobacterium tumefaciens* induced tumours. J. Exp. Bot. **50,** 183–192.

Rezmer, C., Schlichting, R., Wachter, R., and Ullrich, C.I. (1999). Identification and localization of transformed cells in *Agrobacterium tumefaciens*-induced plant tumors. Planta **209,** 399–405.

Risau, W. (1990). Angiogenic growth factors. Prog. Growth Factor Res. **2,** 71–79.

Rolland, F., Winderickx, J., and Thevelein, J.M. (2001). Glucose-sensing mechanisms in eukaryotic cells. Trends Biochem. Sci. **26,** 310–317.

Ryu, S.B. (2004). Phospholipid-derived signaling mediated by phospholipase A in plants. Trends Plant Sci. **9,** 229–235.

Sacristan, M.D., and Melchers, G. (1977). Regeneration of plants from habituated and *Agrobacterium*-transformed single-cell clones of tobacco. Mol. Gen. Genet. **152,** 111–117.

Scarpella, E., Marcos, D., Friml, J., and Berleth, T. (2006). Control of leaf vascular patterning by polar auxin transport. Genes Dev. **20,** 1015–1027.

Schreiber, U., Schliwa, U., and Bilger, W. (1986). Continuous recording of photochemical and nonphotochemical chlorophyll fluorescence quenching with a new type of modulation fluorometer. Photosynth. Res. **10,** 51–62.

Schurr, U., Schuberth, B., Aloni, R., Pradel, K.S., Schmundt, D., Jaehne, B., and Ullrich, C.I. (1996). Structural and functional evidence for xylem-mediated water transport and high transpiration in *Agrobacterium tumefaciens*-induced tumors of *Ricinus communis*. Bot. Acta **109,** 405–411.

Sheen, J., Zhou, L., and Jang, J.C. (1999). Sugars as signaling molecules. Curr. Opin. Plant Biol. **2,** 410–418.

Smeekens, S. (2000). Sugar-induced signal transduction in plants. Annu. Rev. Plant Physiol. Plant Mol. Biol. **51,** 49–81.

Smyth, G.K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat. Applic. Genet. Mol. Biol. **3,** Article 3 (http://www.bepress.com/sagmb/vol3/iss1/art3/).

Sousa, C., De Ferreira, A., and Sodek, L. (2002). The metabolic response of plants to oxygen deficiency. Braz. J. Plant Physiol. **14,** 83–94.

Suleiman, S., Hourmant, A., and Penot, M. (1990). Influence de l'acide abscissique sur le transport d'ions inorganiques chez la pomme de terre (*Solanum tuberosum* cv. Bintje). Etude comparée avec quelques autres phytohormones. Biol. Plant. (Praha) **32,** 128–137.

Szyroki, A., Ivashikina, N., Dietrich, P., Roelfsema, M.R.G., Ache, P., Reintanz, B., Deeken, R., Godde, M., Felle, H., Steinmeyer, R., Palme, K., and Hedrich, R. (2001). KAT1 is not essential for stomatal opening. Proc. Natl. Acad. Sci. USA **98,** 2917–2921.

Tadege, M., Dupuis, I., and Kuhlemeier, C. (1999). Ethanolic fermentation: New functions for an old pathway. Trends Plant Sci. **4,** 320–325.

Thimm, O., Blasing, O., Gibon, Y., Nagel, A., Meyer, S., Kruger, P., Selbig, J., Muller, L.A., Rhee, S.Y., and Stitt, M. (2004). MAPMAN: A user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J. **37,** 914–939.

Truernit, E., Schmid, J., Epple, P., Illig, J., and Sauer, N. (1996). The sink-specific and stress-regulated *Arabidopsis STP4* gene: Enhanced expression of a gene encoding a monosaccharide transporter by wounding, elicitors, and pathogen challenge. Plant Cell **8,** 2169–2182.

Ullrich, C.I., and Aloni, R. (2000). Vascularization is a general requirement for growth of plant and animal tumours. J. Exp. Bot. **51,** 1951–1960.

Usadel, B., et al. (2005). Extension of the visualization tool MapMan to allow statistical analysis of arrays, display of corresponding genes, and comparison with known responses. Plant Physiol. **138,** 1195–1204.

Van Larebeke, N., Engler, G., Holsters, M., Van den Elsacker, S., Zaenen, I., Schilperoort, R.A., and Schell, J. (1974). Large plasmid in *Agrobacterium tumefaciens* essential for crown gall-inducing ability. Nature **252,** 169–170.

Van Slogteren, G.M.S., Hoge, J.H.C., Hooykaas, P.J.J., and Schilperoort, R.A. (1983). Clonal analysis of heterogeneous crown gall tumor tissues induced by wild-type and shooter mutant strains of *Agrobacterium tumefaciens* expression of T-DNA genes. Plant Mol. Biol. **2,** 321–333.

Veselov, D., Langhans, M., Hartung, W., Aloni, R., Feussner, I., Gotz, C., Veselova, S., Schlomski, S., Dickler, C., Bachmann, K., and Ullrich, C.I. (2003). Development of *Agrobacterium tumefaciens*

C58-induced plant tumors and impact on host shoots are controlled by a cascade of jasmonic acid, auxin, cytokinin, ethylene and abscisic acid. Planta **216,** 512–522.

Wachsberger, P.R., Gressen, E.L., Bhala, A., Bobyock, S.B., Storck, C., Coss, R.A., Berd, D., and Leeper, D.B. (2002). Variability in glucose transporter-1 levels and hexokinase activity in human melanoma. Melanoma Res. **12,** 35–43.

Wachter, R., et al. (2003). Vascularization, high-volume solution flow, and localized roles for enzymes of sucrose metabolism during tumorigenesis by *Agrobacterium tumefaciens*. Plant Physiol. **133,** 1024–1037.

Wang, X.M. (2004). Lipid signaling. Curr. Opin. Plant Biol. **7,** 329–336.

Warburg, O. (1930). The Metabolism of Tumors. (London: Arnold Constable).

Weil, M., and Rausch, T. (1990). Cell wall invertase in tobacco crown gall cells: Enzyme properties and regulation by auxin.. Plant Physiol. **94,** 1575–1581.

West, C.E., Waterworth, W.M., Stephens, S.M., Smith, C.P., and Bray, C.M. (1998). Cloning and functional characterisation of a peptide transporter expressed in the scutellum of barley grain during the early stages of germination. Plant J. **15,** 221–229.

Chapter 5

Genome Expression Pathway Analysis Tool - Analysis and visualization of microarray gene expression data under genomic, proteomic and metabolic context

BMC Bioinformatics

BioMed Central

Software

Open Access

Genome Expression Pathway Analysis Tool – Analysis and visualization of microarray gene expression data under genomic, proteomic and metabolic context

Markus Weniger*, Julia C Engelmann and Jörg Schultz

Address: Department of Bioinformatics, Biocenter, University of Würzburg, 97074 Würzburg, Germany

Email: Markus Weniger* - markus.weniger@biozentrum.uni-wuerzburg.de; Julia C Engelmann - julia.engelmann@biozentrum.uni-wuerzburg.de; Jörg Schultz - joerg.schultz@biozentrum.uni-wuerzburg.de

* Corresponding author

Published: 2 June 2007

BMC Bioinformatics 2007, **8**:179 doi:10.1186/1471-2105-8-179

Received: 20 February 2007
Accepted: 2 June 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/179

Abstract

Background: Regulation of gene expression is relevant to many areas of biology and medicine, in the study of treatments, diseases, and developmental stages. Microarrays can be used to measure the expression level of thousands of mRNAs at the same time, allowing insight into or comparison of different cellular conditions. The data derived out of microarray experiments is highly dimensional and often noisy, and interpretation of the results can get intricate. Although programs for the statistical analysis of microarray data exist, most of them lack an integration of analysis results and biological interpretation.

Results: We have developed GEPAT, Genome Expression Pathway Analysis Tool, offering an analysis of gene expression data under genomic, proteomic and metabolic context. We provide an integration of statistical methods for data import and data analysis together with a biological interpretation for subsets of probes or single probes on the chip. GEPAT imports various types of oligonucleotide and cDNA array data formats. Different normalization methods can be applied to the data, afterwards data annotation is performed. After import, GEPAT offers various statistical data analysis methods, as hierarchical, k-means and PCA clustering, a linear model based t-test or chromosomal profile comparison. The results of the analysis can be interpreted by enrichment of biological terms, pathway analysis or interaction networks. Different biological databases are included, to give various information for each probe on the chip. GEPAT offers no linear work flow, but allows the usage of any subset of probes and samples as a start for a new data analysis. GEPAT relies on established data analysis packages, offers a modular approach for an easy extension, and can be run on a computer grid to allow a large number of users. It is freely available under the LGPL open source license for academic and commercial users at http://gepat.sourceforge.net.

Conclusion: GEPAT is a modular, scalable and professional-grade software integrating analysis and interpretation of microarray gene expression data. An installation available for academic users can be found at http://gepat.bioapps.biozentrum.uni-wuerzburg.de.

BMC Bioinformatics 2007, **8**:179 http://www.biomedcentral.com/1471-2105/8/179

Background

Introduction

Gene expression analysis using microarrays opened new insights into the living cell, revolutionizing biological research in many fields. Gene expression of a whole system can be measured at once, yielding information about the mRNA level of every gene. Microarrays have become a standard tool for gene expression measurement in biology and medicine. Their application ranges from identification of gene expression changes in different states of the cell cycle over the classification of disease types to drug development. Although microarrays are widely used, a fundamental challenge is to cope with the immense amount of data generated. Therefore special software packages have been developed, capable of handling the analysis of microarray data. Still, we think that many of the existing tools are not optimal in respect of usability and integration. To date, most freely-available programs split the data analysis into two parts: In the first, statistical methods are used to identify lists of 'interesting' genes, in the second these lists are searched for biological relevance. Although these two steps are dependent on each other and should be highly interconnected, currently most analysis tools lack an integration of these steps. In the following, we will give an overview of selected tools.

Existing Tools

One of the most sophisticated software for microarray data analysis is the Bioconductor toolkit [1], based on the R statistical programming language [2]. Most algorithms developed for microarray data analysis are available within this package. Unfortunately, Bioconductor is a text-driven command line tool and does not provide an easy-to-use graphical interface. Therefore, it offers advanced analysis methods and the possibility of easy extension only for professional users, and is difficult to use for people unskilled in R. Results could be misinterpreted if people are not understanding the data they are working with or the way to perform the analysis. To solve this problem, different tools were developed wrapping the Bioconductor toolkit for an easier usage. *AMDA* [3] is an R package, providing a graphical user interface and a workflow for the analysis of Affymetrix microarray data. *CARMAWeb* [4] acts as a web-based user interface, making the Bioconductor modules available for data analysis over the internet.

Besides Bioconductor, other data analysis tools are available. *Expression Profiler* [5] offers an integrated, web based approach for microarray data analysis. Various normalization, filtering, between-group-testing, clustering, cluster comparison and GO term enrichment analysis methods are available. Expression Profiler integrates analysis methods in an application-like web interface. *GEPAS* [6] is also a widely used web-based approach for microarray data analysis. In addition to the functionality of Expression Profiler, it also offers class prediction methods, survival analysis and multiple tree viewers. GEPAS functionality is split up into a number of tools, connected by the same file format. The user interface is more web-styled than Expression Profiler, making the usage more complicated for untrained users.

Other Tools are not web-based, but installed on the local machine. *EXPANDER* [7] includes biclustering methods and analysis methods regarding regulatory elements. *TM4* [8] is a collection of 4 programs, covering all computational steps for microarray analysis. *TM4* includes spot detection/image analysis, data normalization and data analysis, linked together by the same file format. The data analysis part includes, beside other analysis methods, support vector machines, gene shaving and relevance networks.

All these programs share the focus on the data analysis part, but most of them lack tools for the interpretation of the results. Only *GEPAS* offers with *Babelomics* [9] an approach into data interpretation. On the other hand there exist tools focusing on the interpretation of analysis results. Besides many others, *WebGestalt* [10] offers biological term enrichment analysis, protein domain tables, tissue expression analysis, links to chromosome location and textmining analysis. The widely used *DAVID* [11] allows an enrichment analysis for GO categories, pathway enzymes, protein domains and other biological terms. *Cytoscape* [12] supports the integration of network information with microarray gene expression data. Other tools for acquiring gene set information are *MAPPFinder* [13], *GFINDer* [14] and *Pathway Explorer* [15]. The *Ensembl* [16] annotation system *ENSMART* allows the user to perform a genome information search and retrieval for sets of genes, but does not help in exploring the information associated with the gene set. All these tools provide annotation ability, the drawback of these tools is the inability of an integrated analysis. They require precalculated gene sets as input, needing other tools for normalization, clustering and subset determination.

GEPAT

For interpreting microarray analysis results with the tools described above, researchers need first to obtain a list of differential expressed genes from an analysis program, and use this list in an interpretation program to get biological information for the results. This might prove feasible for smaller number of experiments, but is time-consuming and complicated if used for larger numbers.

As we were unhappy with the separation of analysis and interpretation, we developed our own tool, GEPAT. GEPAT offers combined genome-, expression- and pathway analysis and interpretation methods. Our idea was the integration of gene expression data evaluation with the cellular regulation and interaction network. Therefore, we provide gene annotation for the probes on the micro-

array and allow the visualization of analysis results on metabolic pathways and gene interaction networks. GEPAT includes different biological databases, making them directly usable in data analysis and interpretation. As a large number of databases require lots of disk space and the analysis methods demand much computation power, we developed GEPAT as a web-based toolkit. GEPAT offers an application-like user interface with menu bar and dialog boxes for easy usage. The installation as server system allows either installation and usage on a single computer, installation on a web server for use within a workgroup, or installation on a web server connected to a computer cluster for large user groups. GEPAT is distributed under LGPL and can be freely downloaded [17], an installation on our server can be used by academic users [18]. For an easy start, GEPAT provides a video tutorial for the basic steps, and offers online help for most functions. For a first impression of GEPAT, a guest login is available, preloaded with microarray data from cancer type classification [19] and cancer subgroup profiling of diffuse large B-cell lymphoma [20], including chromosomal alteration information [21]. All figures in this paper are based on the B-cell lymphoma dataset.

Implementation

Web Server

GEPAT is implemented in the Java programming language [22] and requires a J2EE-compatible servlet container to run. Our server installation uses Apache Tomcat [23] as base. The JavaServer Faces technology is used for the generation of web pages. This technology offers a Model-View-Controller-based programming approach for internet applications, allowing application development similar to desktop applications. Access control and image generation are implemented using Java Servlets. All databases used and algorithms implemented in GEPAT are wrapped in modules. The program itself provides only user management and data management capabilities, all other functionality is modularly implemented. This allows an easy extension with new databases or new analysis methods. Modules are used for import of gene expression data, subset selection of probes or samples, gene information, analysis and interpretation methods. The currently implemented modules for data analysis can either run calculation on the server itself or calculation can be directed to a computer grid running a DRMAA-compatible grid engine [24]. In our case, the computation scripts are run on our 10-node cluster system, based on the Sun Grid Engine [25]. For data analysis, we used the powerful abilities of the Bioconductor toolkit combined with an easy-to-use interface. For graph layout and visualization, the JUNG [26] graph library is used.

Databases

The modular approach of GEPAT allows the usage of any database by developing new modules. We have already integrated modules for the access of some important biological databases as Ensembl [16]. As the format of most databases was not suitable for our purposes, we reformatted these databases for our needs. For storage a mySQL 5 [27] database server is used. GEPAT provides scripts for the creation of the database tables and the conversion of already existing databases into these tables.

For gene annotation, we found no available database for all clone identifier mapping purposes needed. Therefore we created our own database. We used the UniGene database (Build #197, 12/2006) [28] to provide a mapping from cDNA Clone identifiers (ids) and Genbank ids to UniGene clusters, and used the UniGene information for Ensembl gene entries to map each probe to an Ensembl gene ID. Affymetrix probe identifiers are directly annotated with the information provided from the Ensembl database (41_36c). At the moment, our database is focused on human datasets, support for other organisms will follow in the future.

Unluckily, Ensembl-identifiers do not exist for all probes, as some probes are derived from EST tags for which no gene is annotated, or some probes may bind to more than one mRNA. If an Ensembl identifier is available for a probe, the Ensembl database entry is used to gain information about gene name, chromosomal location, proteins, GO Annotation and enzymatic activity. All data annotation in GEPAT is performed via the Ensembl identifier. The identifier used for annotation is selected automatically out of the array files, or can be selected by the user for tabular file input.

Linking a probe to a gene is necessary for interpretation of the results, but may lead to various problems. Microarray probes may not only hybridize with one specific mRNA, but crosshybridize with mRNAs of different genes. It is also possible that a probe detects only one specific splice variant of a gene, while another probe detects all splice variants. Different probes may hybridize more or less efficiently with the mRNA they were designed for. And at last, it is not always sure if the probes contain the cDNA-material they are supposed to. Therefore it is necessary to compare the sequence of somehow interesting probes with a sequence database, to make sure annotation was right, and to verify the results of the microarray analysis by other experimental methods.

Results

Microarray experiments generate a large amount of data in a very short time. In most cases it is not desirable to work with all these data at once. Only few probes may be differentially expressed, so in some cases it is useful to limit data interpretation to only these probes. The array dataset may consist of numerous subsets of somehow different samples. For the probe and sample set, subsets may be used to focus only on a specific group, or to compare two groups. Defining and working with different subsets for

BMC Bioinformatics 2007, **8**:179 http://www.biomedcentral.com/1471-2105/8/179

any kind of analysis and interpretation is one of the main concepts of GEPAT. For visualization purposes, a working set can be defined, and all output is generated for this working set. For example, as it is not always desirable to have all data mapped to a metabolic pathway map, by limiting the working set to a subset of all probes, the amount of displayed probes shown on a pathway map can be limited.

The subsets used in GEPAT can be selected by different characteristics. For an easy access in analysis, a subset can be named and stored as "group". For example, in a clinical study, all samples belonging to a specific type of disease may be stored in a group with the disease name. This allows quick data analysis by just selecting the desired disease groups. As source for the selection of subsets, either the whole dataset or subsets defined as a group can be used. It is also possible to use a previous subset as source for the selection, allowing to subset subsets. An overview of possible criteria for subset selection is given in Table 1. As an example, it is possible to select all differentially expressed genes, to limit this set to all genes located in the nucleus, and to limit this set further to all genes that originate from a specific chromosome. Any other combination of subset selection criteria is possible. The probe and sample subset selection process is handled modular, allowing an easy extension with yet unimplemented selection modules for other criteria.

GEPAT includes the following processing steps for microarray data:

• Import and normalization of microarray gene expression data

• Information for specific genes in the dataset

• Various analysis methods for microarray data, including a moderate t-test and clustering

• Interpretation methods for subsets of the data

Table 1: Possible criteria for selection of probe and sample subsets

Probe set	Sample set
Name Search	Name Search
Groups	Groups
GO Category	k-Means-Cluster Analysis Results
k-means-Cluster Analysis Results	Principal Component Analysis Results
t-test Results	
KEGG Maps	
Chromosomal location	

The analysis and interpretation steps can be performed in any order, allowing the usage of interpretation results as a start for further analysis. The following text describes the processing steps in detail.

Data Import

Data Input

Data input is an important step in data analysis. Most existing programs require processed data in a specific format, frequently tab-separated tables, or support only a limited amount of formats. To allow broad usage of different input file formats, we decided to use a modular system allowing the extension for any type of file format. All input files are handled by a specific module, and following the import the data is stored in an internal, format-independent and fast-accessible format on the server.

At the moment, three different modules are available for data import. The first module enables data import for tab-separated data files, containing either already normalized data or unnormalized single- or dual-channel data. The other two modules allow the import of oligonucleotide and cDNA microarray data. Affymetrix oligonucleotide arrays are handled by read.affybatch, the cDNA-import uses read.maimages R methods. All formats supported at the moment are listed in Table 2.

For saving bandwidth and mouseclicks, multiple array files are imported wrapped in a Zip-File. This allows the upload of a large amount of arrays without separate uploading of each single file. Upload of tab-separated microarray files provides an easy selection of identifier and data columns, shown in figure 1a. After upload, the data channels of the arrays and the data characteristics can be inspected visually to allow a quick identification of blurred or otherwise erroneous arrays. The microarray selection process is shown in figure 1b. Here arrays can be

Table 2: Supported microarray input file formats

Oligonucleotide
Affymetrix CEL Files (Human)
cDNA
Agilent Feature Extraction
ArrayVision
BlueFuse
GenePix
ImaGene
QuantArray
SPOT
Stanford Microarray Database
Tabular
Unnormalized Dual-Channel Data
Unnormalized Single-Channel Data
Normalized Data

BMC Bioinformatics 2007, **8**:179 http://www.biomedcentral.com/1471-2105/8/179

skipped, removing them from further processing. After the selection of microarrays, data normalization methods can be applied to the data.

Normalization

Normalization of microarray data is needed to remove variations in gene expression levels caused by the measurement process, enabling the comparison of different microarrays with each other. It aims to remove the systematic effects while keeping the most of the signal, and brings the data from different microarrays onto a common scale.

Before normalization, missing value imputation can be performed to fill outmasked probes with the k nearest neighbors averaging method provided by the impute package [29] of Bioconductor. Missing value imputation offers an established method to compute values for flagged probes. This allows the usage of analysis methods not capable of handling unknown data values, but may lead to false results, as imputed values may not reflect the real gene expression levels.

After missing value imputation, a normalization method must be chosen. Most normalization methods distinguish between within- and between-array normalization. Within-array normalization normalizes the expression log-ratios of two-color spotted microarray experiments so that the log-ratios average to zero within each array or sub-array. Between-array normalization normalizes expression intensities so that the intensities or log-ratios

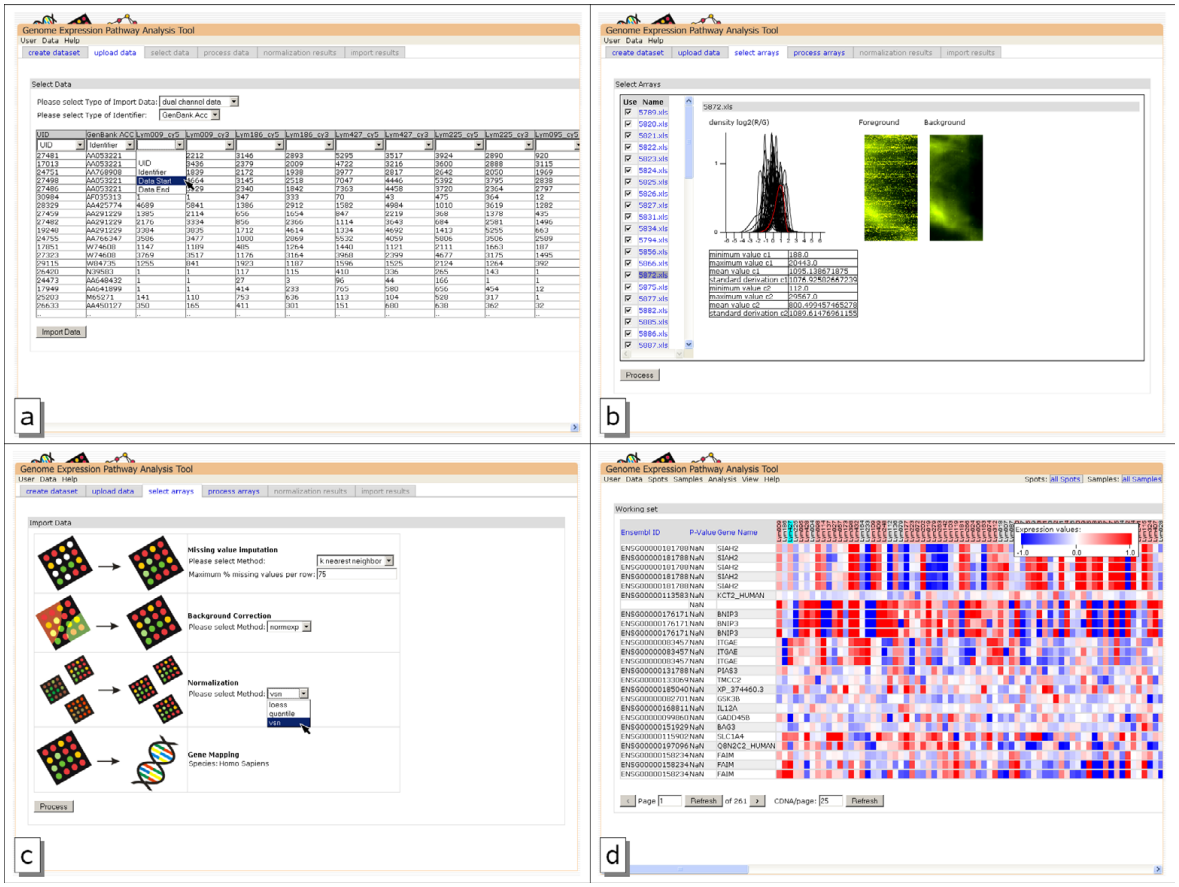


Figure 1

Data import pages. a)import of tabular dual channel unnormalized data. Below the heading of the columns drop-down boxes are used to provide information for import. b)Microarray import view. The table on the left side can be used to select microarrays by name, the right side shows the scanned microarray image and data characteristics. The value distribution of all arrays is given in black, the selected array is marked in red. c)Normalization parameter selection page d)overview of imported and annotated microarray expression data. Probes are shown in the rows, the columns show gene information and sample expression.

BMC Bioinformatics 2007, **8**:179 http://www.biomedcentral.com/1471-2105/8/179

have similar distributions across a series of arrays. Figure 1c shows the normalization configuration page of GEPAT.

GEPAT uses the package *limma* [30] for normalization of two-color microarrays. Different methods are available: One method combines loess within-array normalization [31] and scale between array normalization. The loess method fits the arrays to a polynomial surface, the scale-method scales the log-ratios to have the same median-absolute-deviation across arrays. The other methods use quantile [32] to ensure that the intensities have the same empirical distribution across arrays and across channels or vsn [33] for a robust estimation of variance-stabilizing and calibrating transformations for microarray data. Background correction can be performed via the normexp-method. This method results in a smooth monotonic transformation of the background subtracted intensities such that all the corrected intensities are positive.

For the normalization of Affymetrix arrays the expresso-function of the affy-package is used. Perfect match adjustment ensures that only perfect match oligonucleotides are used for further calculation. For the calculation of the expression values, medianpolish is used. No background-correction is performed. As normalization methods loess, quantile and vsn can be chosen. After normalization, annotation is performed, and data is ready for further analysis. After import, the dataset is shown in an overview table, giving annotation information for the spots and showing the gene expression values for the samples. Figure 1d shows an overview table of the B-cell lymphoma test dataset.

Gene Information

To gain insight about the biological function of the genes on the microarry chip, different sources can be used for gene information. We include some of the most important sources in GEPAT. Gene information is available in most analysis and interpretation views. A click on a probe or gene opens a new window, giving all available gene information. A tab-bar at the top of the page can be used for changing between the different types of information. Gene Information is also modularized and therefore easily expandable with additional information.

Gene Information

For each gene in GEPAT a quick overview showing gene information can be accessed. We offer a subset of the Ensembl gene information, and link to the corresponding Ensembl page for further information. Besides gene name a short description of gene function, chromosomal location, expression values, GO identifier and enzymatic activity are shown if available and link directly to the corresponding pages in GEPAT. An example of an Ensembl information page for the *MYC* gene is given in figure 2a. The information given on the dataset overview page is a subset of the information given on the gene information page, and is modularly expandable.

Gene Associations

The STRING database [34] provides an overview of the physical and functional associations and interactions between proteins. STRING integrates known and predicted protein interaction data from a variety of sources. These associations can be shown in a summary network, displaying the genes as nodes, and different kinds of associations as edges. In GEPAT, we adopted this kind of view. A local instance of the STRING database can be used with GEPAT, and we provide a mapping from Ensembl genes to STRING proteins.

To give an overview of genes interacting with the selected gene, a graph view displaying associated genes can be generated. Similar to the STRING database, possible gene associations are gene neighborhood, gene fusion, co-occurrence, co-expression, experiment, databases and text mining. To keep the graph understandable, the maximal count of nodes can be limited by score and number. For an easier interpretation of the data, differential expression results can be overlaid. A mouse click on a node selects the new gene as center of the graph, allowing browsing through the gene interaction network. The gene association graph for *MYC* is shown in figure 2b.

Literature References

Literature about genes can be found in various journals. To give a quick overview of scientific articles related to a gene, we implemented a literature reference view. We used the RefSeq [35] database from NCBI and the Ensembl RefSeq annotation for the genes to find literature references. For each reference, journal, author and title are provided, and a pubmed outlink offers quick access to abstract and full text. If available in RefSeq, a short summary of the gene function is given, as shown in figure 2c for the *MYC* gene.

Protein Information

Although microarrays designed for resolving different splice variants of genes [36] are starting to get available, most actual microarray techniques provide information on gene level. Nevertheless, sometimes it is necessary to gain information about the proteins derived from these genes. This information is provided in the protein information table. The protein information is drawn out of the Ensembl database, a direct link to the Ensembl website is provided for each protein. The protein information table shows the different possible transcripts of a gene, and provides information about the features, e.g. protein domains, of each protein build out of these transcripts. Figure 2d shows the protein information page for *MYC*.

BMC Bioinformatics 2007, **8**:179 http://www.biomedcentral.com/1471-2105/8/179

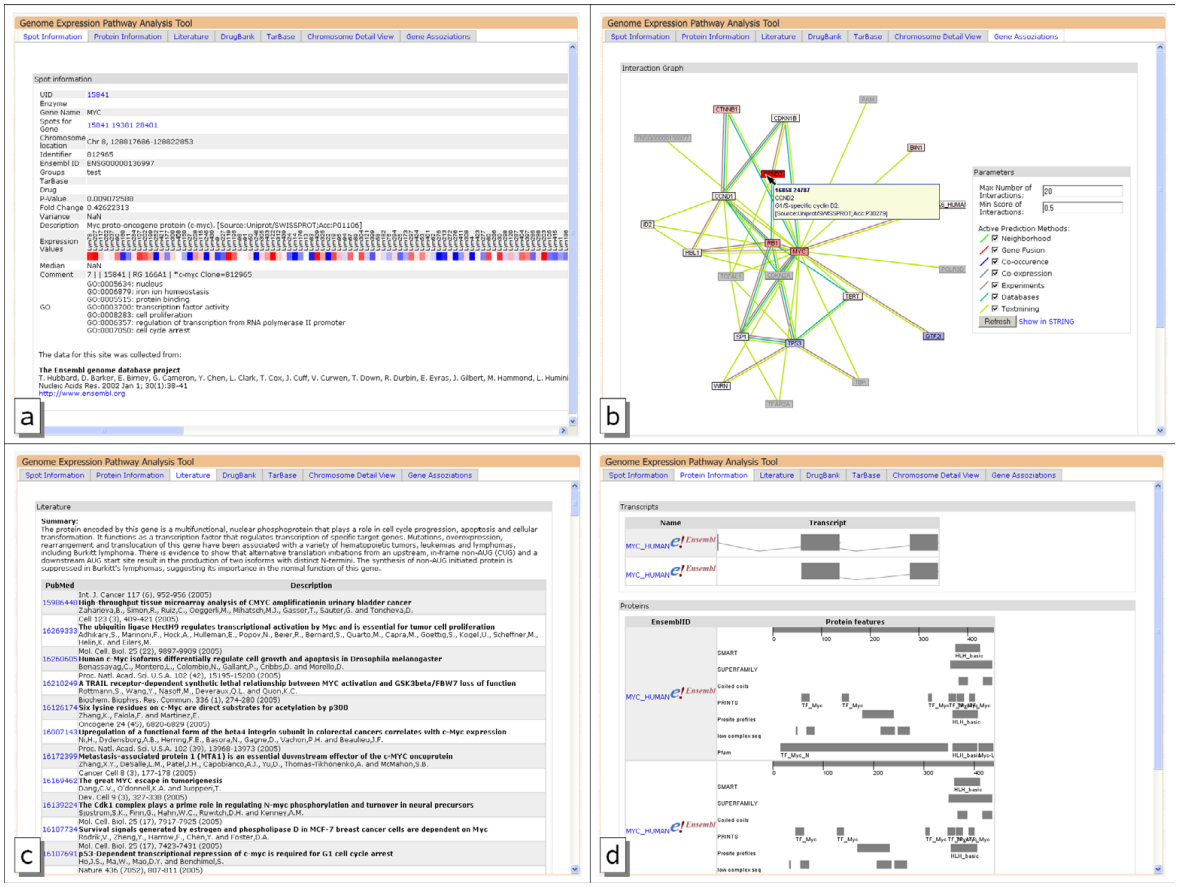


Figure 2
Gene information pages. a)Shows an overview of probe 15841 that measures expression level of the *MYC* gene. The information shown can be modularly extended. b)associated genes for this probe, overlaid with differential gene expression results c)shows literature references and a short description derived from RefSeq, d)shows protein information for the gene. The upper part shows the coding regions, the lower part shows features for the different transcripts.

Data Analysis

A wide variety of data analysis methods is available for gene expression data. We decided to implement differential expression analysis, clustering methods and an analysis of chromosomal alterations in GEPAT. As all analysis methods are implemented as modules, new analysis methods can be added quite easily. With our subset-selection procedure, it is possible to take any probe or sample subset as input for the data analysis methods. The results of the analysis can again be used as criteria for subset selection.

Differential Expression

An important analysis of microarray data is the comparison of expression profiles from different sample groups. Different kinds of tests are available; one of the most advanced is the moderate t-statistics, as it provides stable results even for experiments with small numbers of arrays. We use the limma package of Bioconductor for this analysis [30]. Two sample subsets can be specified and compared. For each probe, the $\log_2$ fold change and p-value are calculated. Benjamini-Hochberg and Benjamini-Yekutieli multiple testing adjustment methods can be applied on raw p-values. These multiple testing correction methods control the false discovery rate, the expected proportion of false discoveries amongst the rejected hypotheses. The false discovery rate is a less stringent condition than the family wise error rate, so these methods are more powerful compared to other methods, e. g. the Bonferroni correction.

The results can be visualized in an M/A-Plot, allowing an overview of the data distribution. The Y Axis shows the M value, the $\log_2$-fold change of probe values in the different

BMC Bioinformatics 2007, **8**:179 http://www.biomedcentral.com/1471-2105/8/179

groups. The X Axis shows the A value, the average expression level for the probe across all the arrays and channels. Additional information is provided via mouse cursor tooltips; a click on a spot provides full information for a probe. An example of an M/A plot is given in figure 3a.

The fold change of differential expression of the compared groups can be mapped onto the visualization components on GEPAT. This allows a direct view of the differential expression on pathways or interaction networks. An important aspect of the t-test is its usage in testing a hypothesis, as it provides error probability values for each tested probe.

Clustering

Clustering means the partitioning of data into subsets (clusters), so that each element of the subset shares a common feature. Clustering methods allow visual insight into the data and can be used for class discovery, e.g. for finding disease categories among experiment samples. GEPAT offers the widely used hierarchical clustering, principal component analysis (PCA) and k-means clustering as unsupervised clustering methods.

The hierarchical clustering method is based on the dist and hclust commands of R. Clustering methods include the widely used unweighted pair-group method using arithmetic averages (UPGMA), single linkage, complete linkage and Ward's algorithm. The single linkage method,

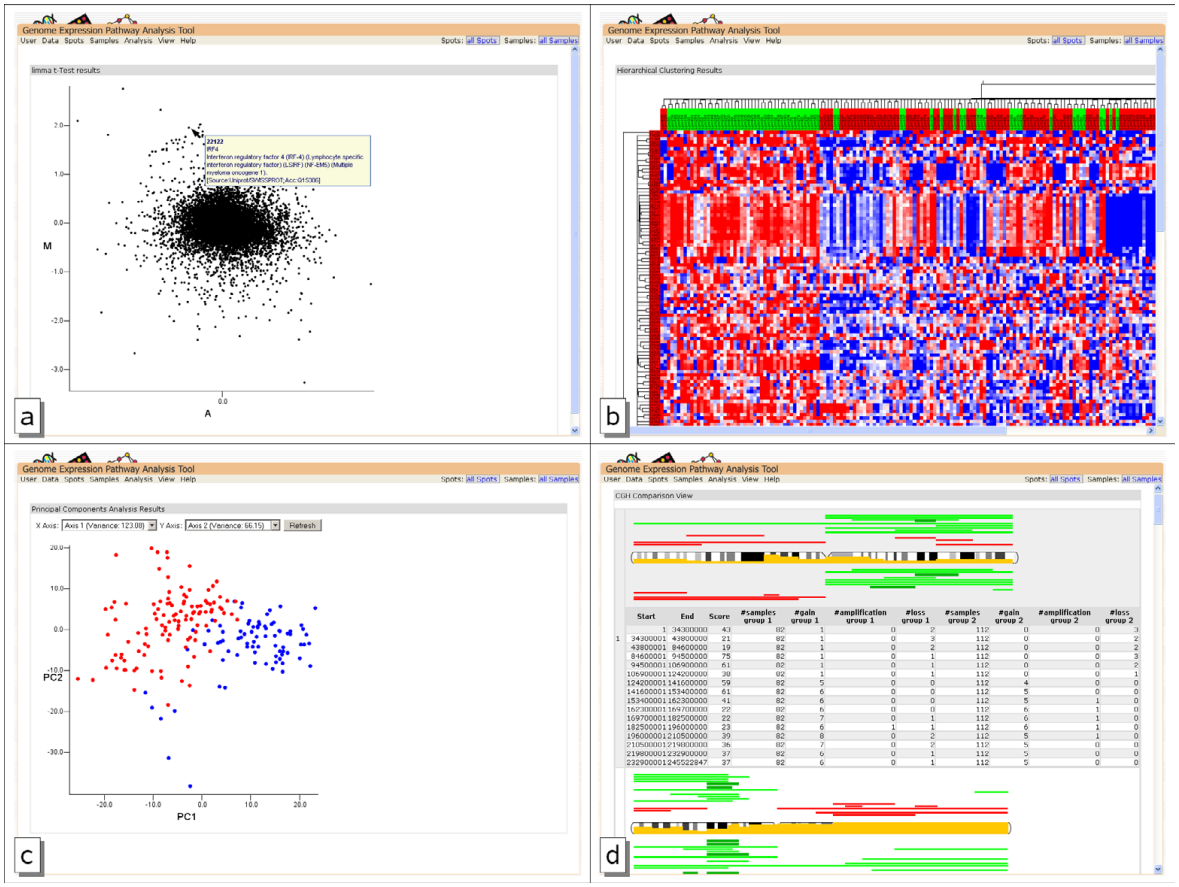


Figure 3

Data Analysis result views. Results are shown for activated B-cell (ABC) type cancer samples and germinal center B-cell (GCB) type cancer samples: a)M/A plot of moderate t-test result comparing ABC with GCB b)hierarchical clustering results. The color of the samples marks the different disease types. c)PCA analysis results. Characteristical probes for disease were used as source for clustering. d)CGH profile comparison. The yellow bar in the chromosome shows the difference between the profiles. Above the chromosome the CGH Profile of the ABC group is shown, the GCB group is shown below.

BMC Bioinformatics 2007, **8**:179 http://www.biomedcentral.com/1471-2105/8/179

closely related to the minimal spanning tree, adopts a 'friends of friends' clustering strategy, the complete linkage methods finds similar clusters, whereas Ward's minimum variance method aims at finding compact, spherical clusters. Figure 3b shows an example of hierarchical clustering results.

PCA is a technique for retrieving information out of a dataset by dimensionality reduction, retaining those characteristics of the dataset that contribute most to its variance. GEPAT can perform PCA on the sample data of the dataset. The samples are shown in a two-dimensional plot, where the principal components for each direction can be chosen freely. A lasso-like selection function provides easy subset selection based on the clustering results. The results of PCA clustering are shown in Figure 3c.

The k-means clustering requires a user input, the expected number k of clusters. GEPAT uses the kmeans command of R to perform a clustering based on the Hartigan-Wong algorithm. As a result, k clusters are returned, and can be used in subset selection for further analysis. These subsets can even be used as base for further clustering, allowing the analysis of complex datasets step by step.

Value Calculation

Other characteristics of the microarray data can be calculated using the expression values. Median and variance can be calculated for all probes and samples, or only for specific subsets of the data. This allows using probes with the highest variance across samples for further analysis.

CGH Data Analysis

GEPAT not only handles microarray data, but is also able to handle additional information for each sample. In cancer datasets, most samples not only differ by gene expression, but have a specific profile of genetic alterations. Comparative genomic hybridization (CGH) is a well-established method that allows the detection of chromosomal imbalances in entire genomes. This technique is widely used in routine molecular diagnostics [37], and many experiments combine CGH and microarray data. We developed a data analysis module capable of comparing the CGH-profile of two sample groups. An unpaired Wilcoxon-Rank test is performed on each chromosomal segment, for comparison of both sample groups. The resulting p-value is plotted directly on the chromosome view, along with the CGH profiles of every group, allowing a quick identification of differing parts. Figure 3d shows a CGH profile comparison example for the lymphoma test dataset.

Data Interpretation

While performing the analysis steps on the data, sets of interesting genes will be found. Methods for correlating these data with prior biological knowledge are necessary. We developed different modules to facilitate the interpretation of these genes and gene sets in a cellular context. The modules are fully integrated with the analysis steps described above and with each other. Therefore, an interpretation can be performed on any subset of data. This integration is a major focus of GEPAT and distinguishes it from many other available tools for the analysis of gene expression data. Data Interpretation in GEPAT is modularly extensible, allowing implementation of any yet unimplemented interpretation method. Out of each Data Interpretation view, gene information can be provided for each probe.

GO Term Enrichment Analysis

At the moment, an automatic ontological approach is one of the most popular methods to gain insights into a set of differentially expressed genes. The Gene Ontology project [38] provides a set of structured vocabularies to describe molecular function, biological process, and cellular component in a hierarchical manner. For interpretation of the data, the GO profile of a subset of genes is compared to the GO profile of a reference set, in most cases all genes of the microarray. The change in the relative frequency of GO terms is used to measure enrichment of GO terms in the subset. A large number of tools exists for performing these analysis for a given list of genes [39]. Out of the different statistical tests used by these tools, we chose an analysis based on a hypergeometrical distribution for GEPAT, as it is an appropriate model for the probability that a certain category occurs x times just by chance in the list of differentially expressed genes. Because of the directed acyclic graph structure of GO multiple testing correction for GO term enrichment analysis is not easy to perform and is still discussed [40], and therefore not provided at the moment.

The results of the GO term enrichment analysis are shown in a tree, representing the direct acyclic graph organization of GO. The tree view of the graph is clearer and enables an easier navigation, but leads to multiple entries of GO categories in different branches of the tree. The tree can be searched for GO Identifiers or GO category names. For each node, the number of genes belonging to the category in the subset, in the reference set, the ratio and p-value is shown. An example for the GO term enrichment view is given in figure 4a. We additionally provide a results table for a quick, sortable overview over all categories.

Pathway Analysis

The GO term enrichment analysis provides information about the biological process genes are involved in, but does not tell how genes interact. Therefore, another important task in microarray analysis is the identification of regulated pathways. The KEGG PATHWAY [41] data-

BMC Bioinformatics 2007, **8**:179 http://www.biomedcentral.com/1471-2105/8/179

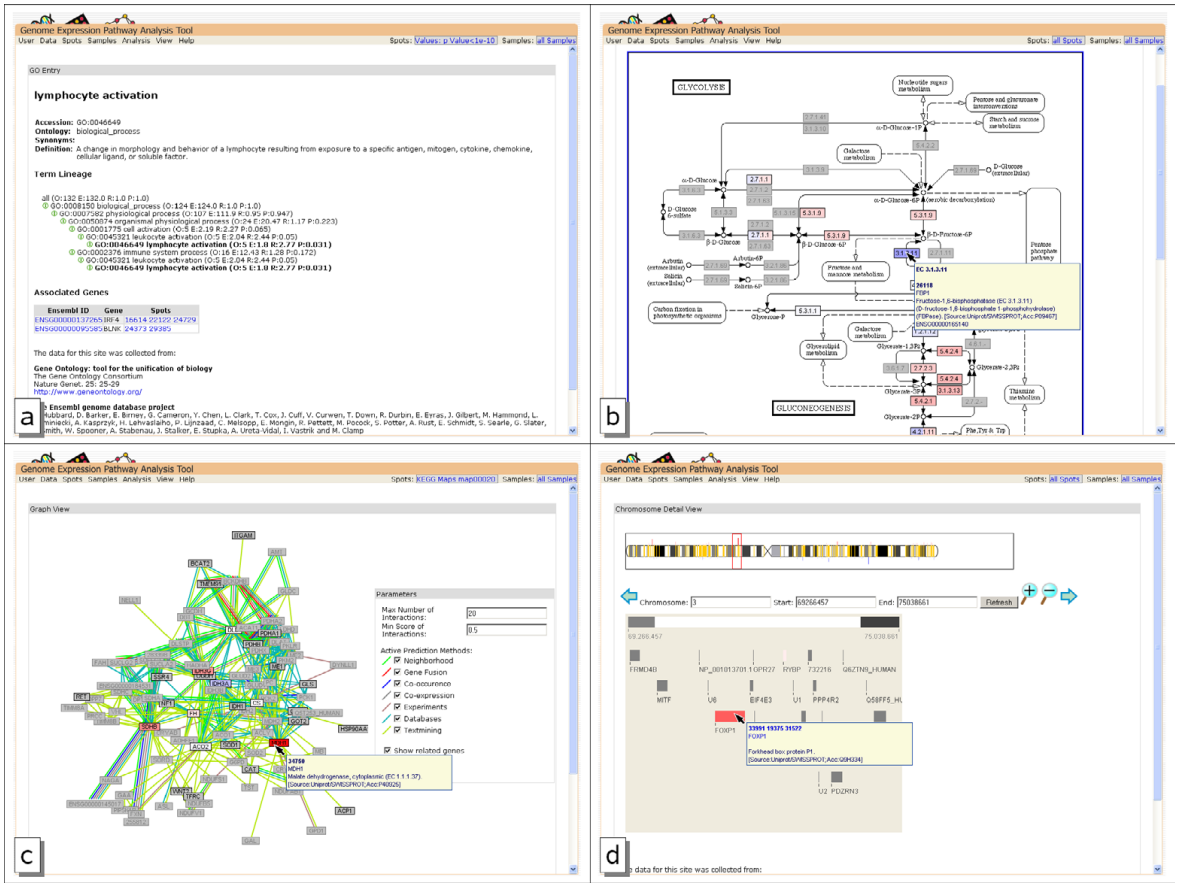


Figure 4

Views of data interpretation. The overlaid differential expression values are the result of the t-test shown in figure 3a. Node colors reflect the differential gene expression. The light gray nodes represent associated genes not on the array. a)Comparing the genes with the lowest p-value shows enriched lymphocyte activation in the GO Term enrichment analysis. b)Glucolysis KEGG map overlaid with differential expression result. c) gene association network of the glucolysis genes. d)Chromosomal location detail view of a chromosome part containing a differential expressed gene. Genes measured on the chip are marked in yellow on the chromosome, differential gene expression is shown above and below the chromosome.

base represents networks of molecular interactions and reactions in the cell in a graphical manner. The available pathways provide key information of the functional and metabolic systems within a living cell. We use this database and color differential gene expression of the current working set onto a pathway, allowing the exploration of functional relationships between genes. The enzymatic activity, described by EC numbers [42] in Ensembl, is used for connecting KEGG maps to the probes on the chip. As an enzymatic activity can be catalyzed by more than one gene the pathway view shows different expression values for each different enzyme. If multiple probes exist for one gene, the median value is calculated and used for coloring.

Figure 4b shows an example of a KEGG map overlaid with differential expression results. To give a fast overview on which maps are containing what amount of genes for the selected working set, a sortable overview table can be displayed.

For each probe on the microarray all KEGG Maps associated with this probe, if there exist some, can be listed. On the other hand, all probes given on a specific map can be used as probe subset in analysis. All enzymes and genes on a KEGG map can be selected by mouse click, giving detailed information about the corresponding genes in the dataset.

Graph View

KEGG pathway information is not available for all genes in a gene subset, because they are not part of a specific pathway, or they are part of a pathway not included in KEGG. However, information about functionally relevant protein interactions is essential for understanding cell behavior. Therefore, an automated display of a STRING summary graph for a subset of genes is implemented in GEPAT. For an easier understanding the differential expression of genes can be mapped onto the graph, giving a fast overview of the expression profile of connected genes. If more than one probe exists for a given gene in the current working subset, the median value is used for visualization. The summary graph can be limited by different types of associations and by the association score provided by STRING.

For each node in the graph, tooltip information is available, and a mouse click on a node provides more information of the selected gene. However, because of the scale-free properties of the gene interaction graph the view is not suitable for larger subsets, as too many nodes do not allow a proper graph layout. An association graph example is shown in figure 4c.

Chromosome Location

To investigate the relationship between gene expression changes and physical gene location, a combined view of gene expression and chromosomal location of the probes is available. The mouse cursor can be used to zoom into a specific genomic region. Inside the zoom view, tooltips are provided for each gene, allowing a quick detail investigation at interesting points of the genome, as shown in figure 4d.

Conclusion

Despite the availability of many programs for microarray data analysis, most of them lack an integration of analysis and interpretation. To understand the effects of differential gene expression an isolated look at genes is not sufficient. It is rather necessary to interpret the results in the context of the cellular network. With the analysis of metabolic or signaling pathways integrating differentially expressed genes, the effects of gene expression on the conditions of cells or tissues can be understood.

GEPAT serves as a toolkit capable of handling the whole progress of microarray data analysis and interpretation in one program. It provides algorithms for the main steps in data analysis, as data import, clustering and differential expression analysis, and offers different methods for data interpretation and visualization, as gene set enrichment analysis or gene association overview. A modular probe and sample selection system allows the usage of analysis and interpretation results as start for new analysis or interpretation methods, facilitating an easy validation of hypotheses or the development of new ones. These integrated capabilities and the build-in annotation support for human microarrays makes GEPAT a powerful tool for microarray data analysis.

It is necessary to be open for new technologies, as biological research develops at fast pace. We implemented large parts of our software in a modular way. Data handling functions serve as a framework that can be extended with various modules for data import, data analysis, data interpretation, subset selection and gene information. As nearly any analysis method can be implemented in this framework, we hope for a future growth of our open-source system. Modules focusing on microRNAs and drug development are currently worked on.

We developed an internet application, focused on easy usage, with a desktop-application like design. This allows a platform-independent remote usage with no need of installation on a local system. With the free availability of the web server, local workgroup installation is possible. To support users untrained in GEPAT, a video tutorial, an online help and test datasets are provided.

Availability and Requirements

Project Name: GEPAT

Project Home Page: http://gepat.sourceforge.net

Operating Systems: Platform independent, tested on windows and linux

Web browser: tested with Internet Explorer 6 and Mozilla Firefox [43]

Programming language: Java > 1.5, R > 2.2

Other requirements: MySQL 5.0, Apache Tomcat 5.0, JSF 1.1

Licence: Free for academic or commercial users under the GNU Lesser General Public Licence (LGPL)

Example Webserver Home Page: http://gepat.bioapps.biozentrum.uni-wuerzburg.de

Authors' contributions

MW created the software and web interface and wrote the manuscript. JE created the data analysis R scripts and gave advice in microarray analysis. JS supervised the project and revised the manuscript. All authors read and approved the final version of the manuscript.

Acknowledgements

MW was funded by the IZKF of the University of Würzburg (IZKF B-36) and by the SFB 688. JE was funded by the IZKF of the University of Würzburg and by the BMBF project FUNCRYPTA (FKZ 0313838B). MW wants to thank P. Seibel for help with programming issues, S. Kneitz for advices concerning the microarray technology and S. Maisel and S. Blenk for testing GEPAT.

References

1. **Bioconductor** [http://www.bioconductor.org/]
2. **The R Project For Statistical Computing** [http://www.r-project.org/]
3. Pelizzola M, Pavelka N, Foti M, Ricciardi-Castagnoli P: **AMDA: an R package for the automated microarray data analysis.** *BMC Bioinformatics* 2006, **7:**335.
4. Rainer J, Sanchez-Cabo F, Stocker G, Sturn A, Trajanoski Z: **CARMAweb: comprehensive R- and bioconductor-based web service for microarray data analysis.** *Nucleic Acids Res* 2006, **34:**W498-503.
5. Kapushesky M, Kemmeren P, Culhane AC, Durinck S, Ihmels J, Körner C, Kull M, Torrente A, Sarkans U, Vilo J, Brazma A: **Expression Profiler: next generation–an online platform for analysis of microarray data.** *Nucleic Acids Res* 2004, **32:**W465-70.
6. Vaquerizas JM, Conde L, Yankilevich P, Cabezón A, Minguez P, Díaz-Uriarte R, Al-Shahrour F, Herrero J, Dopazo J: **GEPAS, an experiment-oriented pipeline for the analysis of microarray gene expression data.** *Nucleic Acids Res* 2005, **33:**W616-20.
7. Shamir R, Maron-Katz A, Tanay A, Linhart C, Steinfeld I, Sharan R, Shiloh Y, Elkon R: **EXPANDER–an integrative program suite for microarray data analysis.** *BMC Bioinformatics* 2005, **6:**232.
8. Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, Sturn A, Snuffin M, Rezantsev A, Popov D, Ryltsov A, Kostukovich E, Borisovsky I, Liu Z, Vinsavich A, Trush V, Quackenbush J: **TM4: a free, open-source system for microarray data management and analysis.** *Biotechniques* 2003, **34:**374-378.
9. Al-Shahrour F, Minguez P, Vaquerizas JM, Conde L, Dopazo J: **BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments.** *Nucleic Acids Res* 2005, **33:**W460-4.
10. Zhang B, Kirov S, Snoddy J: **WebGestalt: an integrated system for exploring gene sets in various biological contexts.** *Nucleic Acids Res* 2005, **33:**W741-8.
11. Dennis GJ, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery.** *Genome Biol* 2003, **4:**P3.
12. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13:**2498-2504.
13. Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, Lawlor SC, Conklin BR: **MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data.** *Genome Biol* 2003, **4:**R7.
14. Masseroli M, Galati O, Pinciroli F: **GFINDer: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists.** *Nucleic Acids Res* 2005, **33:**W717-23.
15. Mlecnik B, Scheideler M, Hackl H, Hartler J, Sanchez-Cabo F, Trajanoski Z: **PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways.** *Nucleic Acids Res* 2005, **33:**W633-7.
16. Hubbard TJP, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Ouverdin B, Parker A, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A, Birney E: **Ensembl 2007.** *Nucleic Acids Res* 2007, **35:**D610-7.
17. **GEPAT** [http://gepat.sourceforge.net]
18. **GEPAT at the University of Wuerzburg** [http://gepat.bioapps.biozentrum.uni-wuerzburg.de]
19. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson JJ, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.** *Nature* 2000, **403:**503-511.
20. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, Gascoyne RD, Muller-Hermelink HK, Smeland EB, Giltnane JM, Hurt EM, Zhao H, Averett L, Yang L, Wilson WH, Jaffe ES, Simon R, Klausner RD, Powell J, Duffey PL, Longo DL, Greiner TC, Weisenburger DD, Sanger WG, Dave BJ, Lynch JC, Vose J, Armitage JO, Montserrat E, López-Guillermo A, Grogan TM, Miller TP, LeBlanc M, Ott G, Kvaloy S, Delabie J, Holte H, Krajci P, Stokke T, Staudt LM: **The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma.** *N Engl J Med* 2002, **346:**1937-1947.
21. Bea S, Zettl A, Wright G, Salaverria I, Jehn P, Moreno V, Burek C, Ott G, Puig X, Yang L, Lopez-Guillermo A, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Gascoyne RD, Connors JM, Grogan TM, Braziel R, Fisher RI, Smeland EB, Kvaloy S, Holte H, Delabie J, Simon R, Powell J, Wilson WH, Jaffe ES, Montserrat E, Muller-Hermelink H, Staudt LM, Campo E, Rosenwald A: **Diffuse large B-cell lymphoma subgroups have distinct genetic profiles that influence tumor biology and improve gene-expression-based survival prediction.** *Blood* 2005, **106:**3183-3190.
22. **Java Technology** [http://java.sun.com]
23. **Apache Tomcat** [http://tomcat.apache.org/]
24. **Distributed Resource Management Application Api** [http://drmaa.org/]
25. **Sun Grid Engine** [http://gridengine.sunsource.net/]
26. **Java Universal Network/Graph Framework** [http://jung.sourceforge.net]
27. **mySQL** [http://www.mysql.com]
28. **NCBI UniGene** [http://www.ncbi.nlm.nih.gov/UniGene]
29. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB: **Missing value estimation methods for DNA microarrays.** *Bioinformatics* 2001, **17:**520-525.
30. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments.** *Stat Appl Genet Mol Biol* 2004, **3:**Article 3.
31. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: **Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.** *Nucleic Acids Res* 2002, **30:**e15.
32. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** *Bioinformatics* 2003, **19:**185-193.
33. Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M: **Variance stabilization applied to microarray data calibration and to the quantification of differential expression.** *Bioinformatics* 2002, **18(Suppl 1):**S96-104.
34. von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Krüger B, Snel B, Bork P: **STRING 7–recent developments in the integration and prediction of protein interactions.** *Nucleic Acids Res* 2007, **35:**D358-62.
35. **NCBI RefSeq** [http://www.ncbi.nlm.nih.gov/RefSeq/]
36. **Genechip Exon Array System** [http://www.affymetrix.com/products/arrays/exon_application.affx]
37. Lichter P, Joos S, Bentz M, Lampel S: **Comparative genomic hybridization: uses and limitations.** *Semin Hematol* 2000, **37:**348-357.
38. Gene Ontology Consortium: **Creating the gene ontology resource: design and implementation.** *Genome Res* 2001, **11:**1425-1433.
39. Khatri P, Drăghici S: **Ontological analysis of gene expression data: current tools, limitations, and open problems.** *Bioinformatics* 2005, **21:**3587-3595.
40. Gentleman R, Carey V, Huber W, Irizarry R, Dudoit R: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer* 2005.
41. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M: **From genomics to chemical genomics: new developments in KEGG.** *Nucleic Acids Res* 2006, **34:**D354-7.
42. Kotera M, Okuno Y, Hattori M, Goto S, Kanehisa M: **Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions.** *J Am Chem Soc* 2004, **126:**16487-16498.
43. **Mozilla Firefox** [http://www.mozilla.com/firefox/]

Chapter 6

Explorative data analysis of MCL reveals gene expression networks implicated in survival and prognosis supported by explorative CGH analysis

Explorative data analysis of MCL reveals gene expression networks implicated in survival and prognosis supported by explorative CGH analysis

Steffen Blenk[1], Julia C. Engelmann[1], Stefan Pinkert[1], Markus Weniger[1], Jörg Schultz[1], Andreas Rosenwald[2], Hans K. Müller-Hermelink[2], Tobias Müller[1] and Thomas Dandekar[1§]

[1]Department of Bioinformatics, University of Würzburg, Biozentrum, Am Hubland, D-97074 Würzburg, Germany

[2]Institute for Pathology, University of Würzburg, Josef-Schneider-Str. 2, D-97080 Würzburg, Germany

Email addresses:

SB: steffen.blenk@biozentrum.uni-wuerzburg.de

JCE: julia.engelmann@biozentrum.uni-wuerzburg.de

SP: stefan.pinkert@biozentrum.uni-wuerzburg.de

MW: markus.weniger@biozentrum.uni-wuerzburg.de

JS: joerg.schultz@biozentrum.uni-wuerzburg.de

AR: Rosenwald@mail.uni-wuerzburg.de

HKMH: path062@mail.uni-wuerzburg.de

TM: Tobias.Mueller@biozentrum.uni-wuerzburg.de

TD: dandekar@biozentrum.uni-wuerzburg.de

[§]corresponding author

Tel. +49-(0)931-8884558, 888-4551

Fax. +49-(0)931-8884552

-accepted by "BMC Cancer"-

Abstract

Background: Mantle cell lymphoma (MCL) is an incurable B cell lymphoma and accounts for 6% of all non-Hodgkin's lymphomas. On the genetic level, MCL is characterized by the hallmark translocation t(11;14) that is present in most cases with few exceptions. Both gene expression and comparative genomic hybridization (CGH) data vary considerably between patients with implications for their prognosis.

Methods: We compare patients over and below the median of survival. Exploratory principal component analysis of gene expression data showed that the second principal component correlates well with patient survival. Explorative analysis of CGH data shows the same correlation.

Results: On chromosome 7 and 9 specific genes and bands are delineated which improve prognosis prediction independent of the previously described proliferation signature. We identify a compact survival predictor of seven genes for MCL patients. After extensive re-annotation using GEPAT, we established protein networks correlating with prognosis. Well known genes (CDC2, CCND1) and further proliferation markers (WEE1, CDC25, aurora kinases, BUB1, PCNA, E2F1) form a tight interaction network, but also non-proliferative genes (SOCS1, TUBA1B CEBPB) are shown to be associated with prognosis. Furthermore we show that aggressive MCL implicates a gene network shift to higher expressed genes in late cell cycle states and refine the set of non-proliferative genes implicated with bad prognosis in MCL.

Conclusions: The results from explorative data analysis of gene expression and CGH data are complementary to each other. Including further tests such as Wilcoxon rank test we point both to proliferative and non-proliferative gene networks implicated in inferior prognosis of MCL and identify suitable markers both in gene expression and CGH data.

Background

Mantle cell lymphomas (MCL) make up about 6% of all cases of non-Hodgkin's lymphomas. They occur at any age from the late 30s to old age, are more common in the over 50 years old population and three times more common in men than in women. Morphologically, MCL is characterized by a monomorphic lymphoid proliferation of cells that resemble centrocytes. MCL is associated with a poor prognosis and remains incurable with current chemotherapeutic approaches. Despite response rates of 50-70% with many regimens, the disease typically relapses and progresses after chemotherapy. The median survival time is approximately 3 years (range, 2-5 y); the 10-year survival rate is only 5-10%.

The characteristic translocation t(11;14) leads to overexpression of Cyclin D1 in the tumor cells which therefore comprises an excellent marker in the diagnostic setting [1]. The present study is an effort to improve molecular insights and markers of the disease [2, 3, 4, 5, 6] to improve the diagnosis and potential therapeutic strategies. We used gene expression data from 71 cyclin D1-positive patients and coupled these to data on their corresponding chromosomal aberrations (n=71). We found molecular markers in addition to cyclin D1 and characteristic antigens (shared with blood cells from which the tumor may develop) CD5, CD20 and FMC7 with the aim to better delineate the regulatory network regulated differently in MCL.

Starting from the proliferation signature [6] we compare long and short living patients subgroups "s" (survivor, above median of survival) and "b" (bad prognosis, below median of survival). Exploratory analysis of gene expression and CGH-data shows new genes differentiating both subgroups, proliferation associated genes and non-proliferative genes. For clinical application a seven gene predictor is derived from these gene markers, distinguishing patients with good or bad survival prognosis. A Wilcoxon rank-sum test on CGH data identifies specific changes on chromosome 9 and 7.

Methods

Data and Materials:

MCL gene expression data (n=71) were obtained from cDNA arrays containing genes preferentially expressed in lymphoid cells or genes known or presumed to be part of cancer development or immune function ("Lymphochip" microarrays [7]; data have been deposited at NCBI's Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) under GEO series accession number GSE10793). The dataset is completed by comparative genomic hybridization (CGH) data for each patient (n=71). The samples were collected from cyclin D1-positive patients of several hospitals in the "Lymphoma and Leukemia Molecular Profiling Project" (LLMPP) [6].

Statistical analysis

Most of the statistical analyses were performed using the "Genome Expression Pathway Analysis Tool" (GEPAT). This is a web-based platform for annotation (allowing also extensive re-annotation of the data), analysis and visualization of microarray gene expression data [8] including genomic, proteomic and metabolic features.

The database performs the analyses applying Bioconductor [9], an open source software for the analysis and comprehension of genomic data, based on the R programming language [10].

For identification of differentially expressed genes, GEPAT uses the "limma" package which offers moderate t-statistics [11, 12]. It fits linear models on the gene expression values of each gene with respect to the groups which are compared. After that empirical Bayes shrinkage of the standard errors is performed. Due to its robustness the method can be applied to experiments with a small number of samples. To correct for multiple testing it offers three options, we chose the method by Benjamini and Hochberg [13].

For identifying all protein-protein network interactions GEPAT uses the "Search Tool for the Retrieval of Interacting Genes/Proteins" (STRING) [14]. The STRING database comprises known and predicted protein-protein interactions. The interaction information arises from genomic context,

experiments, other databases, coexpression and textmining.

For explorative correspondence analysis and principal component analysis, functions from the R package “Modern Applied Statistics with S” (MASS) was applied [15]. A constrained or canonical correspondence analysis (CCA) [16] was performed using the vegan package [17].

The Wilcoxon rank-sum test [18], a non-parametric statistical test, was applied to the CGH data. It tests here each of the chosen bands against the null hypothesis that there is no statistically significant difference between our proposed two MCL patients “b” and “s”. The R package “survival“ is used to calculate all Cox regression hazard models [19, 20]. It examines the correlation between the given measurements and the survival data. For the exploratory analysis of the CGH-data as well as for the new predictor of MCL overall survival, we used the Wald test to determine the significance of the association between the model and the outcome.

Results

Exploratory analysis and lymphoma prognosis

The survival time itself is the most obvious and biological meaningful parameter in which subgroups should show a big difference for realising individual clinical treatment. We selected 3.000 genes with the highest variance and applied correspondence analysis (**Figure 1**). We found (71 MCL patients) that already the second axis separated almost perfectly the longer and the shorter living patients above and below the median of survival. Furthermore, this coincides well with the median of the proliferation signature [6] values in a multidimensional data space (see Methods). This finding was re-examined by exploratory data analysis of the genes of the proliferation signature and a huge amount of further genes. We ranked a total of 71 MCL patients according to their proliferation signature values and separated them according to the median. We define two groups - “s” for small and “b” for big proliferation signature with big difference in the survival time. Patients with a high proliferation signature value live shorter on average, than patients with a low proliferation signature value.

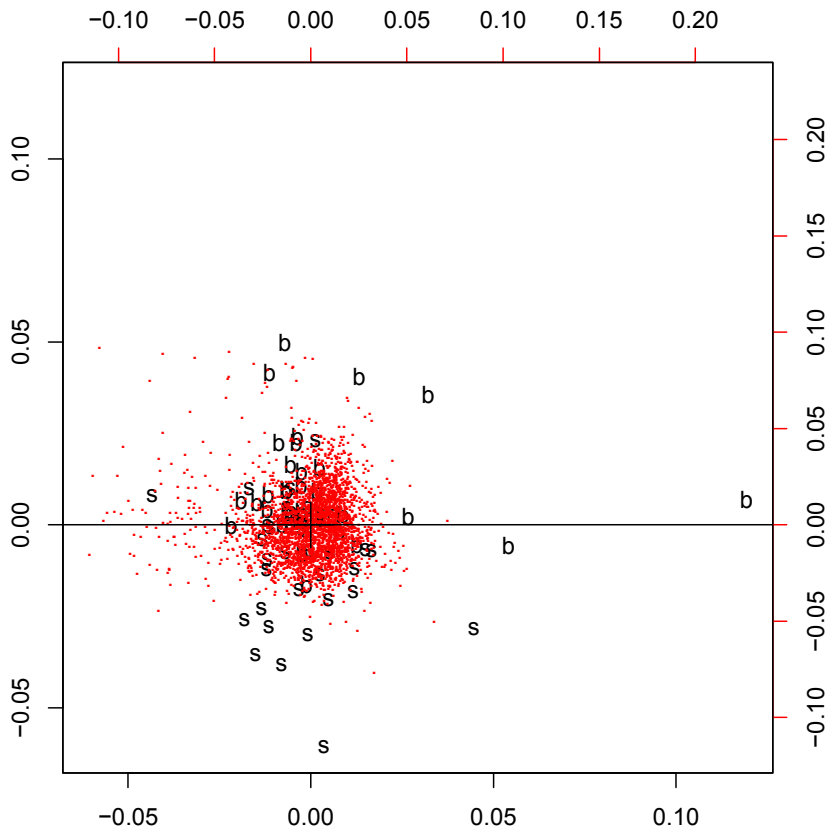


Figure 1: Correspondence analysis identifies the two Mantle cell lymphoma subgroups. The gene expression data are projected on the first two principal axes. The patients can be clearly separated by this exploratory analysis considering the 3.000 genes (red dots) of the highest variance. In the correspondence plot this is indicated by the horizontal separation line. The patients are labelled with "s" and "b" which represent the separation by the *median of the proliferation signature* into two different entities. Patients with a proliferation signature value smaller than the median are marked with „s" and the other patients with „b".

To each single chromosome of the CGH data exploratory data analysis was applied, correspondence analysis (**Supplement**: **Figure 1S**) and principal component analysis (**Figure 2**). Both methods are useful for exploring information and structures in data in order to get a first and unbiased impression. Principal components analysis reduces multidimensional data sets to lower dimensions for analysis. Correspondence analysis works similarly, but scales the data, such that both rows and columns can be visualized in one plot. Results show a strong correlation for four bands of chromosome 9, 9p24, 9p23, 9p22 and 9p21 and above median ("s") or below median patient survival ("b").

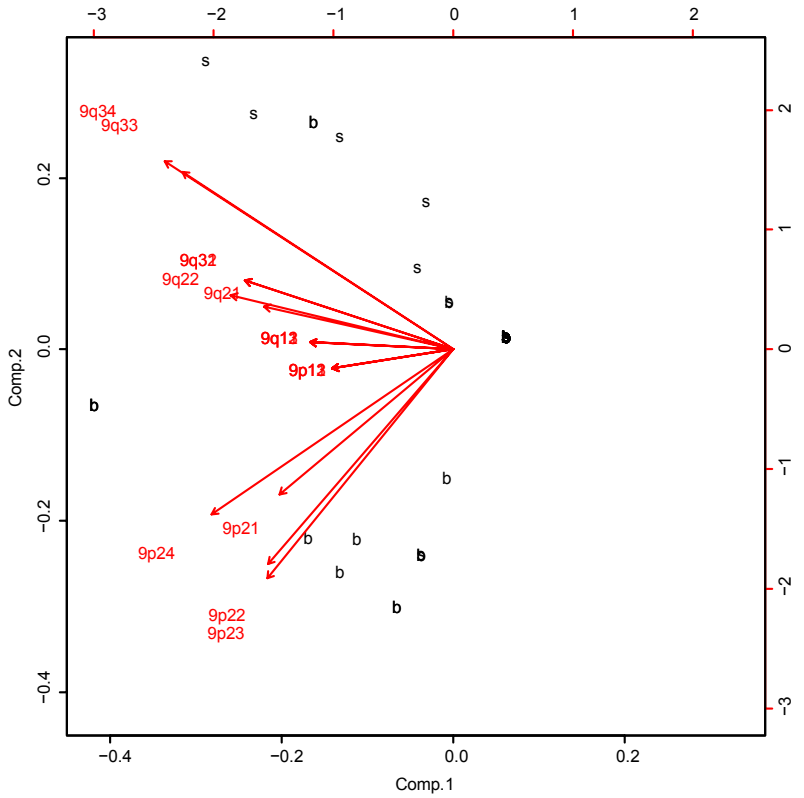


Figure 2: Principal Components Analysis of chromosome 9 bands separating the "s" and "b" group. The second principal component separates almost all patients of the subgroup “b” from the remain. They are grouped together close to the first four vectors, corresponding to the first four bands 9p24, 9p23, 9p22, 9p21, which go into the same direction and are of similar length. Remarkable are the vectors of the bands 9q33 and 9q34. They also are of similar length and go exactly into the same direction. Along their length, they congregate almost all patients of the type “s”. This leads to the assumption, that the first four and the last two bands of chromosome 9 play a crucial role for “s” and “b” classification.

In the *correspondence analysis* plot **Figure 1S (supplement)**, the four bands mentioned before attract most patients of the subgroup “b” and the 1st factor axis separates almost completely the two groups. Bands 9q33 and 9q34, are located relatively far away from the remaining ones. In **Figure 2** the second principal component groups almost all the “b” - patients near the four bands 9p24, 9p23, 9p22, 9p21 with vectors of similar length and similar direction. The vectors of 9q33 and 9q34 include along their lengths almost all “s” samples. These results indicate that these 6 bands of chromosome 9 correlate with good and bad survival between patients. The principal component 1 is

an interesting main component, carrying 51% of the variance, but non-trivial to link to a known phenotype (we investigated different possibilities including sex differences, cancer sub-types, patient accrual and correlation with different gene signatures).

Further exploratory data analysis was performed to merge the survival time and the CGH-data by the Cox regression hazard model. A univariate Cox regression hazard model was performed on all available bands of the CGH-data of all 71 patients. The mentioned four bands of chromosome 9 delivered amongst others the most significant results. The resulting bands are “9p24”, “9p23”, “9p22”, “9p21”, “9q31” and “9q32”. These comprise the first four bands found on chromosome 9 by the analyses before.

A compact predictor of survival with seven genes

Exploratory analysis pointed to differences between longer and shorter living MCL patients, but rather than forming two distinct subgroups, the patients constitute a coherent continuum. Therefore, the results of the exploratory analysis above were not additionally confirmed by classification tools. However, the differences in gene expression above and below the median of survival correlate well with different gene signatures identified before (proliferation signature) as well as with the new ones described in our study (non-proliferative signatures, see below). To improve survival predictions we further searched with univariate Cox regression hazard analysis for highly significant genes, which correlate strongly with the overall survival time. The cox regression was applied to all data points. However, the first 50 MCL samples served as training set for classification by gene signatures and the remaining data (21 patients) for validation. The idea was here to have a large training data set, but still keep a third of the available data for validation.

A four gene predictor with the genes CDC2, ASPM, tubulin-α and CENP-F reported in [6] could not be tested, as after reannotation by GEPAT [8], mapping of CENP-F seemed uncertain. Predictors with 4, 5 or 6 genes delivered not the same predictive power as the proliferation signature [6] (data not shown). The prediction power was calculated from the correct classification

and misclassification for patients over or below the median of survival for 69 patients (the two patients with the median value were excluded).

However, we identified a set of seven genes delivering similar good prognosis separation. It includes (i) the well known key cell cycle kinase CDC2 [21, 22], (ii) the "cell division cycle 20 homolog" (CDC20) required for anaphase and chromosome separation [23] and (iii) the salvage pathway gene HPRT1 (hypoxanthine phosphoribosyltransferase 1), three genes from the 20 genes proliferation signature of Rosenwald [6]. We get improved prediction power including four additional genes (**Table 1**): (i) centromere protein E (CENPE), a kinesin-like motor protein; it accumulates during G2 phase of cell cycle for chromosome movement or spindle elongation (24). (ii) BIRC5 (baculoviral IAP repeat-containing 5 gene), an inhibitor of apoptosis (IAP gene family) is expressed in most tumours and in lymphoma (25), participates in the spindle checkpoint and associates with AURKB (26). (iii) ASPM (abnormal spindle homolog) is essential for normal mitotic spindle function (27). (iv) Insulin-like growth factor 2 mRNA binding protein 3 (IGF2BP3), is found in the nucleolus, is over-expressed in human tumours and represses IGF2 during late development (28, 29, 30).

Table 1: The genes of the survival predictor. Univariate Cox regression hazard analysis revealed these seven genes best correlating with the survival time (see Material and Methods). The first column indicates the gene accession number in the data set (Acc), the second the gene name, followed by the Ensembl identifier and the official full name. The genes are ordered by their significance in decreasing order. CENPE is the most significant gene.

Acc	**gene**	**EnsemblID**	**official full name**
6558	CENPE	ENSG00000138778	Centromeric protein E
7495	CDC20	ENSG00000117399	Cell division cycle protein 20 homolog
7892	HPRT1	ENSG00000165704	Hypoxanthine-guanine phosphoribosyltransferase
7019	CDC2	ENSG00000170312	Cell division control protein 2 homolog
7376	BIRC5	ENSG00000089685	Baculoviral IAP repeat-containing protein 5
6422	ASPM	ENSG00000066279	Abnormal spindle-like microcephaly-associated protein
5923	IGF2BP3	ENSG00000136231	IGF-II mRNA-binding protein 3

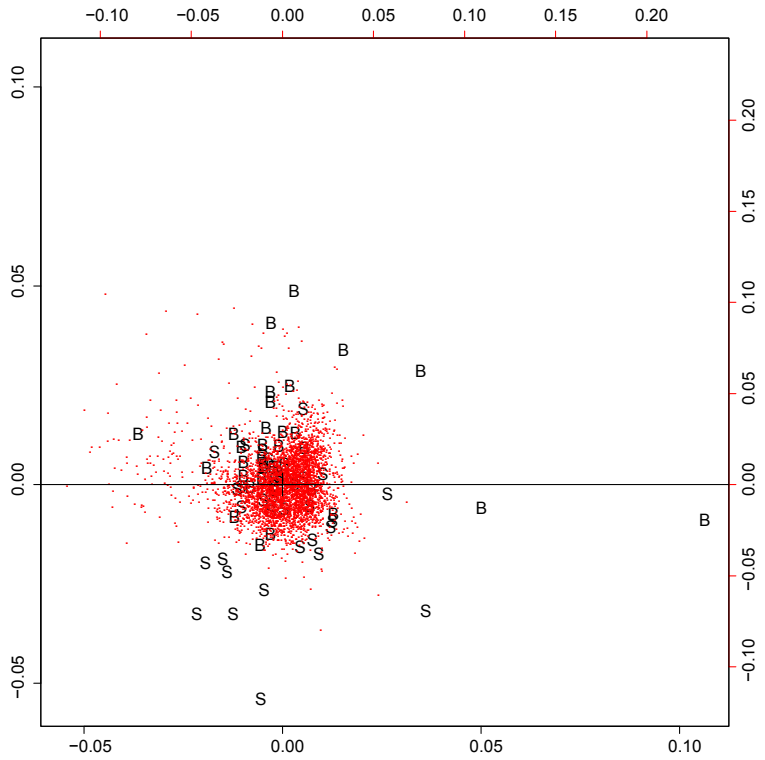


Figure 3: Correspondence analysis separates two MCL subgroups derived by the 7 genes survival predictor. The 3.000 genes with highest variance (red dots) separate between the two subgroups, which were delivered by the seven gene predictor and are drawn as “S” and “B”. They were separated by the median of the predictor values. In contrast to the proliferation signature based predictor (Figure 1), the patients here show a little more overlap, but cluster clearly.

The seven genes were used to calculate a multivariate Cox regression hazard model and with its coefficients, a gene expression based survival estimator separated all 71 patients into two subgroups. Two patients had exactly the median of survival and were excluded in this comparison, 56 agreed with the classification according to the gene signatures, 13 did not. Compared to the proliferation [6] signature’s ability to distinguish patients with good and bad survival prognosis, the seven gene predictor does it similarly well (**Figure 4**). The correlation between this classification and the “s” and “b” groups of the proliferation signature is overall about 0.62 and in our validation set (patients 51 - 71) it is 0.81.

A correspondence analysis of the 3.000 genes with the highest variance showed clear clustering of patients with good or bad prognosis, respectively (**Figure 1**). Using proliferation signature [6]

(**Figure 3**), samples show a little overlap, but are again separated clearly.

Taken together, these results show that the seven gene predictor is able to distinguish patient prognosis as well as the complete proliferation signature, but with less effort.

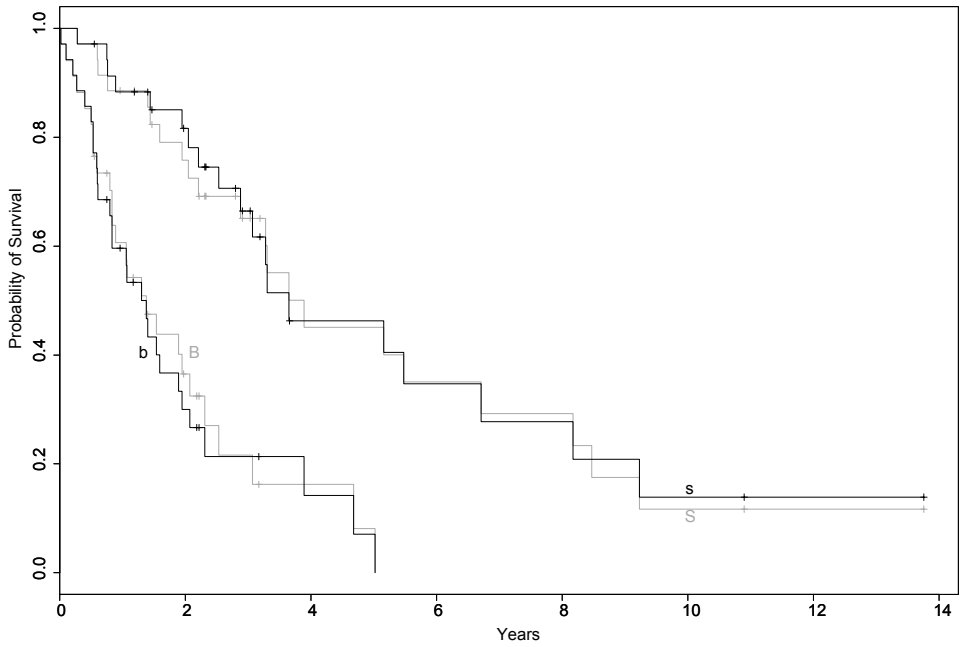


Figure 4: Kaplan Meier plot of survival data in MCL subgroups. The x-axis denotes the course of time in years and the y-axis marks the probability of survival. Both, the proposed proliferation signature (black) and the seven genes predictor (grey) separate clearly two risk groups in the survival data. The overlap between the patients of the two classifications is relatively high.

Protein networks and interactions differently regulated in good and bad prognosis tumors

We found a dense regulatory network of interacting genes correlated with prognosis. Applying a moderate t-test, the well known cell division cycle 2 gene (CDC2 / CDK1) for G1 to S and G2 to M transition [31, 32] shows the most significant difference between the longer living “s” and the shorter living “b” patients (**Table 2**). Furthermore, its interaction partners according to HPRD

Table 2: Most significant genes separating good (s) and bad (b) prognosis. The most significant differentially expressed genes regarding good ("s") or bad ("b") prognosis determined with a moderate t-test. P-values were corrected for multiple testing [13]. The gene "cell division cycle 2" (CDC2), which is important for the transition G1 to S and G2 to M shows the biggest difference in gene expression between the two groups. This indicates that these cell cycle transitions are part of the difference between the two groups.

Acc	gene	fold change	p-value	EnsemblID
7019	CDC2	1.3737029	1.8651454E-13	ENSG00000170312
6632	NP_057427.3	0.94384	3.4574367E-13	ENSG00000117724
3399	UHRF1	1.1446086	1.5513529E-12	ENSG00000034063
5112	NP_060880.2	1.0916529	1.5513529E-12	ENSG00000123485
6994	AURKB	1.4594886	1.5513529E-12	ENSG00000178999
6388	MKI67	1.5062114	1.7304206E-12	ENSG00000148773
6721	Q9Y645_HUMAN	1.2185314	3.2408542E-12	ENSG00000140451
7024	BUB1	1.2488679	3.2408542E-12	ENSG00000169679
6392	NP_057427.3	1.3208085	3.2902188E-12	ENSG00000117724
5726	MKI67	1.4871315	3.6012686E-12	ENSG00000148773
6029	NP_057427.3	1.2980943	5.249176E-12	ENSG00000117724
7423	BIRC5	1.3726515	6.49239E-12	ENSG00000089685
4985	ASPM	1.3310171	7.281489E-12	ENSG00000066279
5754	KIF23	1.2461857	1.6424877E-11	ENSG00000137807
5271	ASPM	1.3205649	2.2259293E-11	ENSG00000066279
6104	KIF23	1.1683029	2.4981522E-11	ENSG00000137807

database [33] show a significant up or down regulation comparing good and bad surviving patients (**Figure 5**), e.g. WEE1 and CDC25. Moreover, aurora kinases A, B [34] and BUB1 kinase (activating the spindle checkpoint, [35]) are differently regulated between shorter and longer living patients. However, there are further genes involved in this network of directly interacting genes differently regulated in good or bad prognosis patients (**Figure 5**) such as (i) proliferating cell nuclear antigen" (PCNA), a cofactor of DNA polymerase delta, helps to increase the processivity of leading strand synthesis during DNA replication in group "b". Because of its ability to interact with multiple partners, it is involved in Okazaki fragment processing, DNA repair, translation, DNA

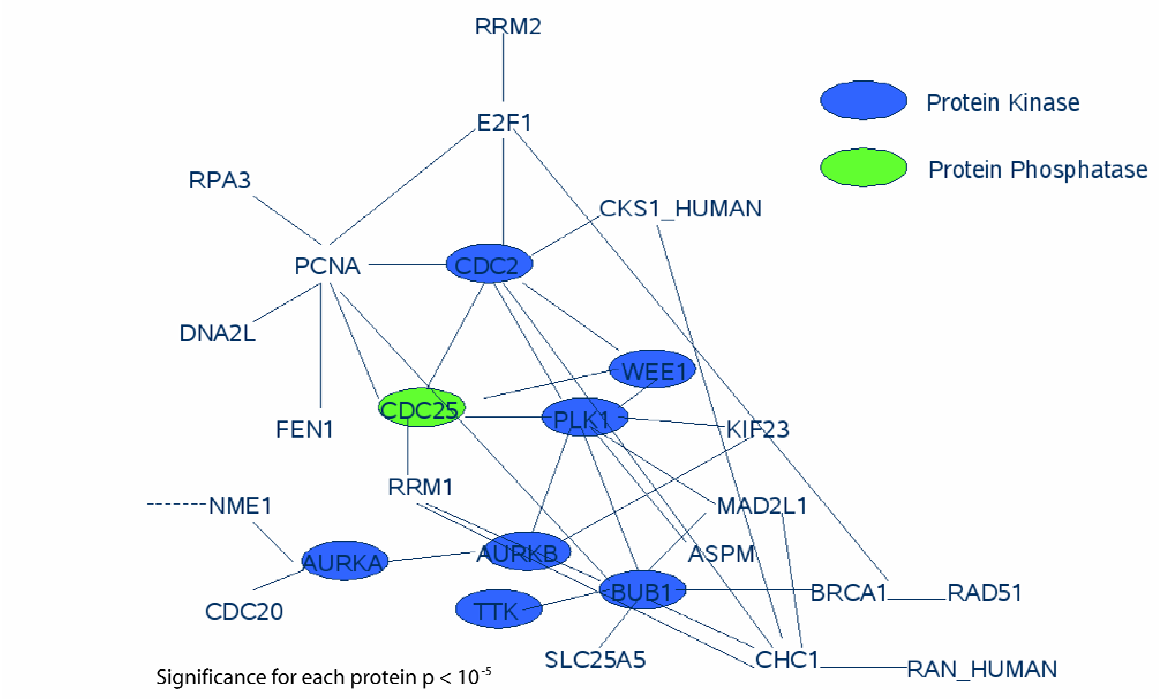


Figure 5: Protein interaction network of significantly different expressed genes. The genes encoding these proteins show a significant expression difference between the "s" and "b" group (moderate t-test). Remarkably CDC2 is involved in a small interaction network of protein kinases and almost all of these interaction partners(CDC25, WEE1, AURKB, AURKA, BUB1) are associated with the cell cycle.

synthesis, DNA methylation, chromatin remodelling and cell cycle regulation [36]. (ii) E2F transcription factor 1 (E2F1), this protein can mediate both cell proliferation and p53-dependent/independent apoptosis [37]. It is lower expressed in group "s". (iii) Nucleolin is an abundant multifunctional phosphoprotein of proliferating and cancerous cells [38, 39, 40, 41] and highly expressed in "b".

Interaction partners of CCND1 are also significantly differently expressed (**Figure 7**): CCND1 and CDK4 are assumed to be involved in cell cycle progression of MCL, MYC is suspected of increasing MCL's proliferation rate. FOS, JUN and MYBL2 are partly known to play a role in cancer, but not explicitly in MCL. FOS ("v-fos FBJ murine osteosarcoma viral oncogene homolog") and JUN ("jun oncogene") are weakly downregulated in "b". Other interaction partners such as MYC ("V-myc myelocytomatosis viral oncogene homolog (avian)"), MYBL2 ("V-myb

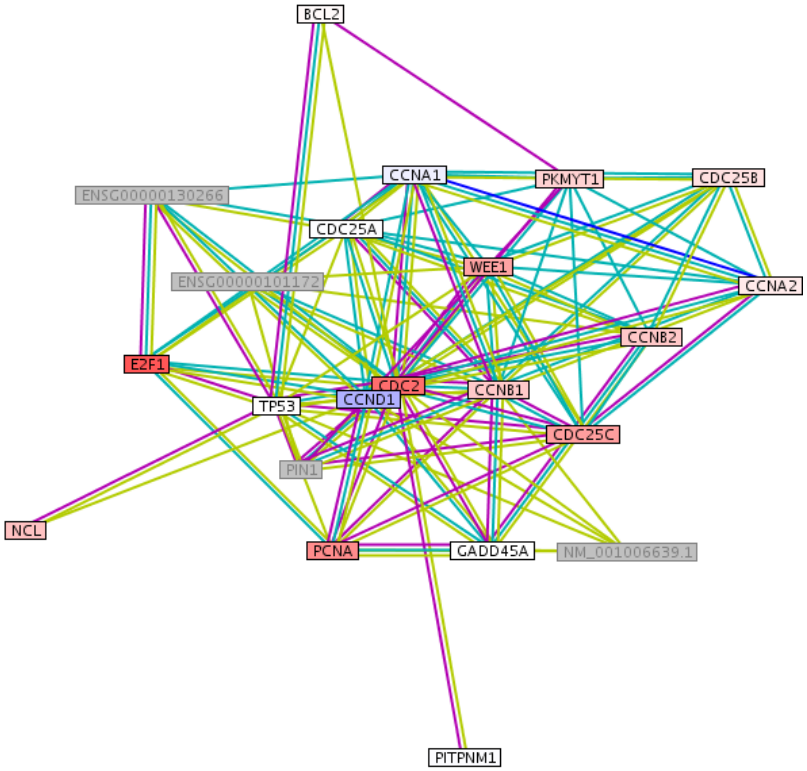


Figure 6: Differences in gene expression of interaction partners of CDC2 in MCL subgroups. In this network figure, red indicates high expression and blue low expression in the subgroup “b” of the proliferation signature. White indicates no gene expression difference and grey the unavailability of the gene in our data set. “Cell division cycle 2” (CDC2) gene interacts in different manners with “cyclin D1” (CCND1), “cell division cycle 25C”(CDC25C), “proliferating cell nuclear antigen”(PCNA), “E2F transcription factor 1”(E2F1) and WEE1. CDC2 and CCND1 are both required for the G1/S transition. The genes WEE1 and CDC25C phosphorylate and dephosphorylate the complexes bound with CDC2 in a cell cycle regulating manner. The "proliferating cell nuclear antigen" (PCNA) is involved in DNA replication whereas "E2F transcription factor 1" (E2F1) controls cell cycle and mediates cell proliferation and apoptosis. A cell cycle regulated transcription activator “Nucleolin” (NCL) shows little difference.

myeloblastosis viral oncogene homolog (avian)-like 2"), CDK4 ("Cyclin-dependent kinase 4") and CDK6 show higher gene expression values in bad prognosis patients below the median of survival. Moreover, there are some genes with similar significance and expression difference, associated with other functions (**Table 3**). Most of them are associated with DNA metabolism. Three of them, "suppressor of cytokine signaling 1" (SOCS1), "tubulin, alpha 1b" (TUBA1B), and

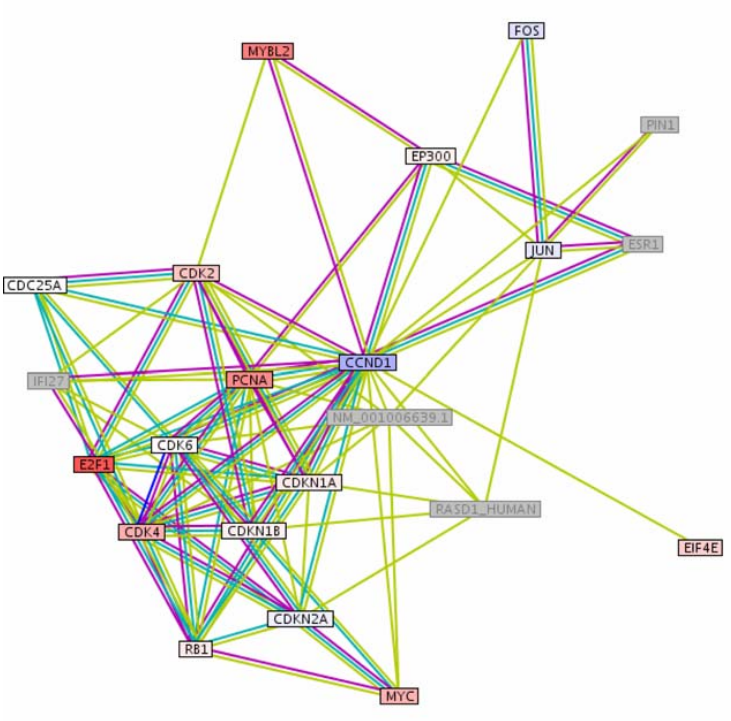


Figure 7: Protein interaction partners of CCND1: Different gene expression in MCL subgroups. The colors red, blue and grey mean “over expressed”, "down regulated” (in “b”) and “not available in the data set”. FOS encodes for a leucine zipper protein and plays a role in regulation of cell proliferation, differentiation, transformation and tumourigenesis [58]. The JUN protein interacts directly with specific target DNA sequences to regulate gene expression [59] and is involved in tumorigenesis by cooperating with oncogenic alleles of Ras, an activator of the mitogen activated protein kinases [60]. MYC and MYBL2 play a role in cell cycle progression and act as transcription factors. MYC is also associated with apoptosis, cellular transformation, cell growth, proliferation, differentiation, and a variety of hematopoietic tumors, leukemias and lymphomas [61, 62, 63], and was part of the original proliferation signature [6]. MYBL2 has been shown to play a role in the G1/S transition [64] and proliferation [65] and is known to be regulated by CCND1 [66, 67]. CDK4 and CDK6 are important regulators of cell cycle transition from G1 to S, phosphorylate, and thus regulate the activity of tumor suppressor protein Rb [68].

“CCAAT/enhancer binding protein (C/EBP), beta” (CEBPB) are mentioned here. CEBPB, is a transcription factor. It plays an important role in immune and inflammatory responses [42]. Additionally it can stimulate the expression of the collagen type I gene. TUBA1B encodes for an important part of the microtubules. SOCS1 is a member of cytokine-inducible inhibitors of signalling [43] and inhibits protein kinase activity.

Table 3: **Genes separating good (s) and bad (b) prognosis not associated with cell cycle and proliferation association**

EnsemblID	gene	p-value	fold change
ENSG00000185338	SOCS1	2.3029981E-10	1.0293059
ENSG00000123416	TBAK_HUMAN	6.1972505E-10	1.0070857
ENSG00000172216	CEBPB	7.545418E-10	0.7460686

CGH data reveals new genes implicated in MCL outcome

We applied the Wilcoxon rank-sum test on the CGH data and compared the patients with good “s” and bad prognosis “b” (over and below median of survival). The null hypothesis corresponds to no differences between the two entities. The resulting p-values for every band of chromosome 9 are compared in **Figure 8**. They show strongly the significance of the first four bands 9p24, 9p23, 9p22 and 9p21. On these bands are MCL related genes such as "cyclin-dependent kinase inhibitor 2B" (CDKN2B) and "cyclin-dependent kinase inhibitor 2A" (CDKN2A). TP53 mutations are associated with the blastoid variant of MCL and with a worse prognosis. The bands 9q33 and 9q34 have a weaker significance. To visualize this result more clearly **Figure 2S** in the supplement plots the densities of the p-values. A peak in the density indicates significant bands of the Wilcoxon test.

The Wilcoxon rank sum test showed similar results for chromosome 7. Here, the bands 7p21, 7p15, 7p14 are potentially related to the classification of “s” and “b” patients. Now the log p-values and their densities are plotted against the bands in **Figure 9** and in **Figure 3S** (supplement). The explorative analyses of chromosome 7 could not show such a clear relation as in chromosome 9.

Specific gene expression differences in patients with good or bad prognosis are well supported by the CGH data of chromosome 9. We checked the location of the signature genes as we wondered if they were on chromosome 7 or 9, however this was not the case. Also the genes of the gene network in **Figure 6** are located elsewhere. No result mentioned before could explain the relationship between the subgroups and the subgroup-separating CGH-data of chromosome 9. We thus investigated the gene expression data of these bands. Again a moderate t-test was applied to

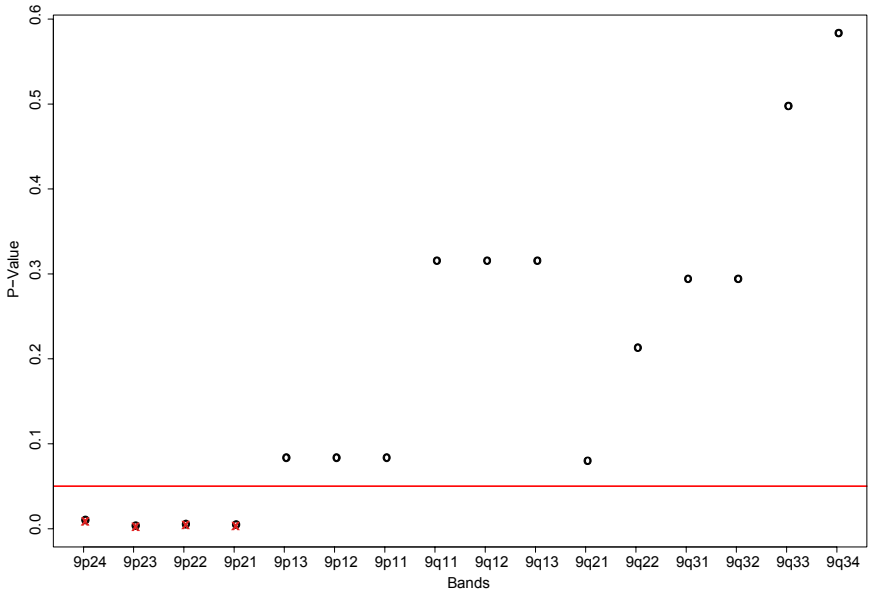


Figure 8: P-values of the Wilcoxon test for the bands of chromosome 9. This figure plots the bands of Chromosome 9 on the x-axis against the p-values of the Wilcoxon test(y-axis), which tested each band between the two groups "s" and "b". The p-values of the first four bands 9p24, 9p23, 9p22, 9p21 are very small, compared to the remaining ones. This affirms the proposed subgroups "s" and "b" and indicates that the first four bands have a relation to this classification.

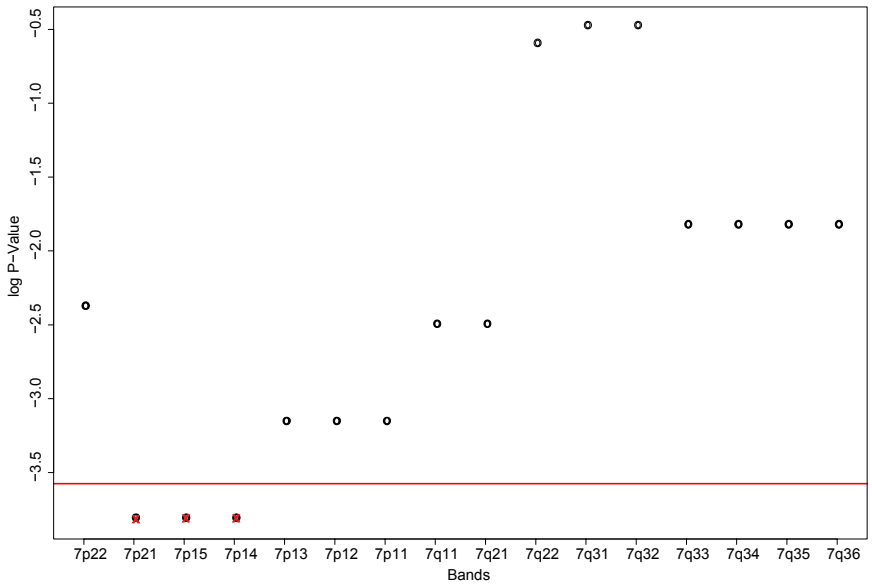


Figure 9: P-values of the Wilcoxon test for the bands of chromosome 7. The Wilcoxon test was applied to all bands of chromosome 7 over the two groups "s" and "b". The bands of chromosome 7 (x-axis) are plotted against the log p-values (y-axis). Three bands show a very low p-value: 7p21, 7p15, 7p14. As the four bands of chromosome 9, they could have a relation to the “s” – “b” classification.

rank genes differentially expressed between “s” and “b”. The top five are listed in **Table 4**, e.g. the "Heat Shock 70kDa protein 5" and a catalytic subunit of "Protein Phosphatase 6". Several of their functions implicate them to be critical in cancer development. Their genomic position revealed a quite remarkable clustering of these genes, shown in **Figure 4S**. Three of the genes seem to be located very closely to each other. The “heat shock 70kDa protein 5” (HSPA5), also referred to as ’immunoglobulin heavy chain-binding protein’ (BiP) targets misfolded proteins for degradation, and has an anti-apoptotic property. It is induced in a wide variety of cancer cells and cancer biopsy tissues and contributes to tumor growth and confers drug resistance to cancer cells [44]. The PPP6C gene encodes for a catalytic subunit of the Ser/Thr phosphatases, the “protein phosphatase 6 catalytic subunit” [45]. The pre-B-cell leukemia transcription factor 3 (PBX3) shows extensive homology to PBX1, a human homeobox gene involved in t(1;19) translocation in acute pre-B-cell leukemias. But in contrast to PBX1 the expression of PBX3 is not restricted to particular states of differentiation or development [46]. It is also known that if HoxB8, a homeobox gene identified as a cause of leukaemia, binds to the Pbx cofactors it blocks differentiation in certain cell types [47]. “Prostaglandin-endoperoxide synthase 1” (PTGS1) is the key enzyme in prostaglandin biosynthesis, and is also known to play a role in the human colon cancer [48, 49]. The expression of the alternative splice variants is differentially regulated by cytokines and growth factors [50, 51, 52]. Very little is known about “quiescin Q6-like 1” (QSCN6L1), except its major role in regulating the sensitization of neuroblastoma cells for IFN-gamma-induced apoptosis [53]. A similar clear clustering as on chromosome 9 could not be detected on chromosome 7.

Table 4: The best "s" and "b" separating genes of chromosome 9 bands 9p24, 9p21, 9q33, and 9q34. A moderate t-test revealed the following ones as the genes with the highest significance. Althogh the significance is weak, it is quite remarkable that these genes here show a distinct clustering on the basis of genomic positions, which can be observed in **Figure 4S**.

gene	start bp.	end bp.	fold change	p-value	official full name
HSPA5	127036953	127043430	0.4364743	0.03039	Heat shock 70kDa protein 5
PPP6C	126948673	126991918	0.2798860	0.03385	Protein phosphatase 6, catalytic subunit
PBX3	127548372	127769477	0.3976210	0.03385	Pre-B-cell leukemia homeobox 3
PTGS1	124173050	124197803	0.4124149	0.03927	Prostaglandin-endoperoxide synthase 1
QSCN6L1	138240395	138277470	-0.3886557	0.03927	Quiescin Q6 sulfhydryl oxidase 2

Discussion

Several different marker genes and events have been proposed for MCL, e.g the translocation t(11;14)(q13;q32) [1], immunohistochemically [54] and Repp86 proteins as a proliferation markers [55] and increased levels of cyclin D1.

The present study consolidates gene expression and CGH-data regarding MCL subgroups with good or bad prognosis to an overall picture. These subgroups are indicated and confirmed by exploratory analyses. This picture shows as yet unknown relations and differences between patients from these groups.

Correspondence analysis is an unsupervised tool to project high dimensional data into lower dimensional subspaces. Surprisingly, its second component separates well the shorter and longer living patients according to the median of survival. This result is in close agreement with the median of the outcome predictor score derived by the proliferation signature [6] as a discriminator.

A new predictor of survival with similar predictive power as the proliferation signature of 20 genes

[6] was developed requiring gene expression values of only seven genes. With the key genes CDC20, HPRT1 and CDC2 the seven-gene-predictor matches with three genes from the 20 genes proliferation signature. Moreover, the four genes CENPE, BIRC5, ASPM and IGF2BP3 add to its predictive power and are associated with chromosome movement, inhibition of apoptosis and tumors. It was shown that a four gene predictor (CDC2, ASPM, tubulin-alpha, CENP-F) [6] is also able to predict length of survival with high statistical significance. Besides the fact, that the proliferation signature is more efficient and powerful than the four gene model, our model meets extensive re-annotation of the genes through the clone IDs.

These CGH data support the association of alterations in chromosomal regions and outcome of MCL patients.

Gene expression analysis comparing long and short surviving patients delivered cell cycle related genes and their protein-protein interactions. A dense interaction network differently regulated in good or bad prognosis includes CDC2 and interaction partners for cell cycle control and proliferation (CCND1, CDK4, MYC and E2F1; CDC25, WEE1, AURKB, AURKA, BUB1, PCNA, FOS, JUN and MYBL2). However, we identified furthermore non proliferation genes differentially implicated in MCL prognosis such as SOCS1 and CEBPB.

The Wilcoxon rank sum test revealed relations between the bands 9p24, 9p23, 9p22 and 9p21 and the difference between the longer and shorter living patients. Investigation of those bands regarding most significant differentially expressed genes revealed a cluster of genes with properties such as “differentiation blocking”, “anti apoptotic” and “apoptosis inducing”. Supporting our finding, the band 9p21 was suggested be implicated in MCL patient outcome [56]. Some bands of chromsome 7 identified further expression differences somewhat weaker associated with the outcome. As the annotation and properties of embedded genes are not completely known, further data are required to better explain the relation between gene functions and survival. CGH-data may improve the power of gene expression based predictors [57]. Besides others, the band 9p21 was associated with a poor clinical outcome, which affirms our finding.

Our study extends these CGH results in two ways: (i) exploratory analysis shows here for the first time, that in fact CGH-data alone can predict prognosis in MCL, (ii) CGH-data point here directly to several genes regulated differently in good or bad prognosis patients.

Conclusion

After careful re-annotation of involved genes we found two subgroups of MCL patients which were found and supported by exploratory analysis of gene expression values and CGH-data, network analysis and literature mining. We obtained an improved classification of MCL regarding prognosis. Differentially expressed genes formed a tight protein interaction network of kinases. A seven gene predictor appeared as an easy to measure prognosis indicator for clinical use. The Wilcoxon rank sum test as well as PCA was applied successfully to a CGH data set in this study. Both identify bands on chromosome 9. Following the indicated bands, we found differentially expressed MCL related genes.

Competing interests:

The authors declare that they have neither financial competing interests nor other competing interests.

Authors´ contributions:

SB carried out the essential technical work for the study including data validation, calculations, statistical analysis and result figures and for Ms writing. JCE aided in these tasks including Ms. writing as well as with her expertise in analyzing gene expression data. SP as well as MW provided databank support and results. JS supervised databank support and results and added own observations. AR and HKMH provided the patient data as well as pathology expert advice during the analysis of the data and participated in the critical discussion of the results. TM supervised statistical analysis and gave expert advice including important methodological contributions. TD

led and guided the study, gave supervision, led the Ms writing, and analyzed the different data and results. All authors participated in the writing of the Ms and approved the final version of the Ms.

Acknowledgements:

We thank the State of Bavaria for support (IZKF B-36; ENB Lead Structures of Cell Function) and DFG (SFB688 TP A2).

References

1. Bogner C, Peschel C, Decker T: **Targeting the proteasome in mantle cell lymphoma: A promising therapeutic approach.** *Leuk Lymphoma* 2006, **47**:195-205.
2. Argatoff LH, Connors JM, Klasa RJ, Horsman DE, Gascoyne RD: **Mantle cell lymphoma: a clinicopathologic study of 80 cases.** *Blood* 1997, **89**:2067-2078.
3. Bosch F, Lopez-Guillermo A, Campo E, Ribera JM, Conde E, Piris MA, Vallespi T, Woessner S, Montserrat E: **Mantle cell lymphoma: presenting features, response to therapy, and prognostic factors.** *Cancer* 1998, **82**:567-575.
4. Raty R, Franssila K, Joensuu H, Teerenhovi L, Elonen E: **Ki-67 expression level, histological subtype, and the International Prognostic Index as outcome predictors in mantle cell lymphoma.** *Eur J Haematol* 2002, **69**:11-20.
5. Velders GA, Kluin-Nelemans JC, De Boer CJ, Hermans J, Noordijk EM, Schuuring E, Kramer MH, Van Deijk WA, Rahder JB, Kluin PM, Van Krieken JH: **Mantle-cell lymphoma: a population-based clinical study.** *J Clin Oncol* 1996, **14**:1269-1274.
6. Rosenwald A, Wright G, Wiestner A, Chan WC, Connors JM, Campo E, Gascoyne RD, Grogan TM, Müller-Hermelink HK, Smeland EB, Chiorazzi M, Giltnane JM, Hurt EM, Zhao H, Averett L, Henrickson S, Yang L, Powell J, Wilson WH, Jaffe ES, Simon R, Klausner RD, Montserrat E, Bosch F, Greiner TC, Weisenburger DD, Sanger WG, Dave BJ, Lynch JC, Vose J, Armitage JO, Fisher RI, Miller TP, LeBlanc M, Ott G, Kvaloy S, Holte H, Delabie J, Staudt LM: **The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma.** *Cancer Cell* 2003, **3**:185-197.
7. Alizadeh A, Eisen M, Davis RE, Ma C, Sabet H, Tran T, Powell JI, Yang L, Marti GE, Moore DT, Hudson JR Jr, Chan WC, Greiner T, Weisenburger D, Armitage JO, Lossos I, Levy R, Botstein D, Brown PO, Staudt LM: **The lymphochip: a specialized cDNA**

microarray for the genomic-scale analysis of gene expression in normal and malignant lymphocytes. *Cold Spring Harb Symp Quant Biol* 1999, **64**:71-78.

8. Weniger M, Engelmann JC, Schultz J: **Genome Expression Pathway Analysis Tool--analysis and visualization of microarray gene expression data under genomic, proteomic and metabolic context.** *BMC Bioinformatics* 2007, **8**:179.
9. Gentleman R, Carey V, Bates M, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biology* 2004, **5**:R80.
10. R Development Core Team: R: *A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria; 2007.
11. Smyth GK: **Limma: linear models for microarray data.** In: *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, Edited by Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W. New York: Springer 2005:397-420.
12. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments.** *Statistical applications in genetics and molecular biology* 2004; **3**:Article 3.
13. Y. Benjamini and Y Hochberg: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *Journal of the Royal Statistical Society*, Series B (Methodological) 1995, **57**:125-133.
14. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P: **STRING: known and predicted protein-protein associations, integrated and transferred across organisms.** *Nucleic Acids Research* 2005; **33**(Database issue):D433-437.
15. Venables WN and Ripley BD: *Modern Applied Statistics with S.* Fourth edition. Springer 2002.
16. Ter Braak CJF: **Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis.** *Ecology* 1986, **67**:1167–1179.
17. Oksanen J, Kindt R, Legendre P, O'Hara RB: **vegan: Community Ecology Package**, 2007, [http://cran.r-project.org/. R package version 1.8-4].
18. Wilcoxon, F: **Individual Comparisons by Ranking Methods.** *Biometrics Bulletin* 1945, **1**:80–83.
19. Andersen P, Gill R: **Cox's regression model for counting processes, a large sample study.** *Annals of Statistics* 1982, **10**:1100-1120.

20. Therneau T, Grambsch P, Fleming T: **Martingale based residuals for survival models.** *Biometrika* 1990, **77**:147-160.
21. Norbury C, Nurse P: **Cyclins and cell cycle control.** *Curr Biol* 1991, **1**:23-24.
22. Norbury C, Nurse P: **Animal cell cycles and their control.** *Annu Rev Biochem* 1992, **61**:441-470.
23. Sethi N, Monteagudo MC, Koshland D, Hogan E, Burke DJ: **The CDC20 gene product of Saccharomyces cerevisiae, a beta-transducin homolog, is required for a subset of microtubule-dependent cellular processes.** *Mol Cell Biol* 1991, **11**:5592-5602.
24. Yen TJ, Li G, Schaar BT, Szilak I, Cleveland DW: **CENP-E is a putative kinetochore motor that accumulates just before mitosis.** *Nature* 1992, **359**:536-539.
25. Ambrosini G, Adida C, Altieri DC: **A novel anti-apoptosis gene, survivin, expressed in cancer and lymphoma.** *Nat Med* 1997, **3**:917-921.
26. Beardmore VA, Ahonen LJ, Gorbsky GJ, Kallio MJ: **Survivin dynamics increases at centromeres during G2/M phase transition and is regulated by microtubule-attachment and Aurora B kinase activity.** *J Cell Sci* 2004, **117**:4033-4042.
27. Bond J, Roberts E, Mochida GH, Hampshire DJ, Scott S, Askham JM, Springell K, Mahadevan M, Crow YJ, Markham AF, Walsh CA, Woods CG: **ASPM is a major determinant of cerebral cortical size.** *Nat Genet* 2002, **32**:316-320.
28. Mueller-Pillasch F, Lacher U, Wallrapp C, Micha A, Zimmerhackl F, Hameister H, Varga G, Friess H, Buchler M, Beger HG, Vila MR, Adler G, Gress TM: **Cloning of a gene highly overexpressed in cancer coding for a novel KH-domain containing protein.** *Oncogene* 1997, **14**:2729-2733.
29. Monk D, Bentley L, Beechey C, Hitchins M, Peters J, Preece MA, Stanier P, Moore GE: **Characterisation of the growth regulating gene IMP3, a candidate for Silver-Russell syndrome.** *J Med Genet* 2002, **39**:575-581.
30. Nielsen J, Christiansen J, Lykke-Andersen J, Johnsen AH, Wewer UM, Nielsen FC: **A family of insulin-like growth factor II mRNA-binding proteins represses translation in late development.** *Mol Cell Biol* 1999, **19**:1262-1270.
31. Aleem E, Kiyokawa H, Kaldis P: **Cdc2-cyclin E complexes regulate the G1/S phase transition.** *Nat Cell Biol* 2005, **7**:831-836.
32. Malumbres M, Barbacid M: **Cell cycle kinases in cancer.** *Curr Opin Genet Dev* 2007, **17**:60-65.
33. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TK, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC,

Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK, Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang L, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobe GC, Dang CV, Garcia JG, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, Pandey A: **Development of human protein reference database as an initial platform for approaching systems biology in humans.** *Genome Res* 2003, **13**:2363-2371.

34. Lampson MA, Renduchitala K, Khodjakov A, Kapoor TM: **Correcting improper chromosome-spindle attachments during cell division.** *Nat Cell Biol* 2004, **6**:232-237.
35. Tang Z, Shu H, Oncel D, Chen S, Yu H: **Phosphorylation of Cdc20 by Bub1 provides a catalytic mechanism for APC/C inhibition by the spindle checkpoint.** *Mol Cell* 2004, **16**:387-397.
36. Maga G, Hubscher U: **Proliferating cell nuclear antigen (PCNA): a dancer with many partners.** *J Cell Sci* 2003, **116**:3051-3060.
37. Crosby ME, Almasan A: **Opposing roles of E2Fs in cell proliferation and death.** *Cancer Biol Ther* 2004, **3**:1208-1211.
38. Lapeyre B, Bourbon H, Amalric F: **Nucleolin, the major nucleolar protein of growing eukaryotic cells: an unusual protein structure revealed by the nucleotide sequence.** *Proc Natl Acad Sci U S A* 1987, **84**:1472-1476.
39. Derenzini M, Sirri V, Trere D, Ochs RL: **The quantity of nucleolar proteins nucleolin and protein B23 is related to cell doubling time in human cancer cells.** *Lab Invest* 1995, **73**:497-502.
40. Srivastava M, Pollard HB: **Molecular dissection of nucleolin's role in growth and cell proliferation: new insights.** *FASEB J* 1999, **13**:1911-1922.
41. Grinstein E, Shan Y, Karawajew L, Snijders PJ, Meijer CJ, Royer HD, Wernet P: **Cell cycle-controlled interaction of nucleolin with the retinoblastoma protein and cancerous cell transformation.** *J Biol Chem* 2006, **281**:22223-22235.
42. Akira S, Isshiki H, Sugita T, Tanabe O, Kinoshita S, Nishio Y, Nakajima T, Hirano T, Kishimoto T: **A nuclear factor for IL-6 expression (NF-IL6) is a member of a C/EBP family.** *EMBO J* 1990, **9**:1897-1906.
43. Starr R, Willson TA, Viney EM, Murray LJ, Rayner JR, Jenkins BJ, Gonda TJ, Alexander WS, Metcalf D, Nicola NA, Hilton DJ: **A family of cytokine-inducible inhibitors of signalling.** *Nature* 1997, **387**:917-921.
44. Li J, Lee AS: **Stress induction of GRP78/BiP and its role in cancer.** *Curr Mol Med* 2006, **6**:45-54.

45. Stefansson B, Brautigan DL: **Protein phosphatase 6 subunit with conserved Sit4-associated protein domain targets IkappaBepsilon.** *J Biol Chem* 2006, **281**:22624-22634.
46. Monica K, Galili N, Nourse J, Saltman D, Cleary ML: **PBX2 and PBX3, new homeobox genes with extensive homology to the human proto-oncogene PBX1.** *Mol Cell Biol* 1991, **11**:6149-6157.
47. Knoepfler PS, Sykes DB, Pasillas M, Kamps MP: **HoxB8 requires its Pbx-interaction motif to block differentiation of primary myeloid progenitors and of most cell line models of myeloid differentiation.** *Oncogene* 2001, **20**:5440-5448.
48. Garavito RM, Mulichak AM: **The structure of mammalian cyclooxygenases.** *Annu Rev Biophys Biomol Struct* 2003, **32**:183-206.
49. Wiese FW, Thompson PA, Warneke J, Einspahr J, Alberts DS, Kadlubar FF: **Variation in cyclooxygenase expression levels within the colorectum.** *Mol Carcinog* 2003, **37**:25-31.
50. DeWitt DL: **Prostaglandin endoperoxide synthase: regulation of enzyme expression.** *Biochim Biophys Acta* 1991, **1083**:121-134.
51. Hla T, Ristimaki A, Appleby S, Barriocanal JG: **Cyclooxygenase gene expression in inflammation and angiogenesis.** *Ann N Y Acad Sci* 1993, **696**:197-204.
52. Herschman HR: **Regulation of prostaglandin synthase-1 and prostaglandin synthase-2.** *Cancer Metastasis Rev* 1994, **13**:241-256.
53. Wittke I, Wiedemeyer R, Pillmann A, Savelyeva L, Westermann F, Schwab M: **Neuroblastoma-derived sulfhydryl oxidase, a new member of the sulfhydryl oxidase/Quiescin6 family, regulates sensitization to interferon gamma-induced cell death in human neuroblastoma cells.** *Cancer Res* 2003, **63**:7742-7752.
54. Katzenberger T, Petzoldt C, Holler S, Mader U, Kalla J, Adam P, Ott MM, Müller-Hermelink HK, Rosenwald A, Ott G: **The Ki67 proliferation index is a quantitative indicator of clinical risk in mantle cell lymphoma.** *Blood* 2006, **107**:3407.
55. Schrader C, Janssen D, Meusers P, Brittinger G, Siebmann JU, Parwaresch R, Tiemann M: **Repp86: a new prognostic marker in mantle cell lymphoma.** *Eur J Haematol* 2005, **75**:498-504.
56. Rubio-Moscardo F, Climent J, Siebert R, Piris MA, Martin-Subero JI, Nielander I, Garcia-Conde J, Dyer MJ, Terol MJ, Pinkel D, Martinez-Climent JA: **Mantle-cell lymphoma genotypes identified with CGH to BAC microarrays define a leukemic subgroup of disease and predict patient outcome.** *Blood* 2005, **105**:4445-4454.
57. Salaverria I, Zettl A, Bea S, Moreno V, Valls J, Hartmann E, Ott G, Wright G, Lopez-Guillermo A, Chan WC, Weisenburger DD, Gascoyne RD, Grogan TM, Delabie J, Jaffe ES, Montserrat E, Müller-Hermelink HK, Staudt LM, Rosenwald A, Campo E: **Specific**

secondary genetic alterations in mantle cell lymphoma provide prognostic information independent of the gene expression-based proliferation signature. *J Clin Oncol* 2007, **25**:1216-1222.

58. Milde-Langosch K: **The Fos family of transcription factors and their role in tumourigenesis.** *Eur J Cancer* 2005, **41**:2449-2461.
59. Hartl M, Bader AG, Bister K: **Molecular targets of the oncogenic transcription factor jun.** *Curr Cancer Drug Targets* 2003, **3**:41-55.
60. Weiss C, Bohmann D: **Deregulated repression of c-Jun provides a potential link to its role in tumorigenesis.** *Cell Cycle* 2004, **3**:111-113.
61. Eisenman RN: **Deconstructing myc.** *Genes Dev* 2001, **15**:2023-2030.
62. Marcu KB, Bossone SA, Patel AJ: **myc function and regulation.** *Annu Rev Biochem* 1992, **61**:809-860.
63. Pelengaris S, Khan M, Evan G: **c-MYC: more than just a matter of life and death.** *Nat Rev Cancer* 2002, **2**:764-776.
64. Golay J, Cusmano G, Introna M: **Independent regulation of c-myc, B-myb, and c-myb gene expression by inducers and inhibitors of proliferation in human B lymphocytes.** *J Immunol* 1992, **149**:300-308.
65. Sala A, Watson R: **B-Myb protein in cellular proliferation, transcription control, and cancer: latest developments.** *J Cell Physiol* 1999, **179**:245-250.
66. Horstmann S, Ferrari S, Klempnauer KH: **Regulation of B-Myb activity by cyclin D1.** *Oncogene* 2000, **19**:298-306.
67. Cesi V, Tanno B, Vitali R, Mancini C, Giuffrida ML, Calabretta B, Raschella G: **Cyclin D1-dependent regulation of B-myb activity in early stages of neuroblastoma differentiation.** *Cell Death Differ* 2002, **9**:1232-1239.
68. Schafer KA: **The cell cycle: a review.** *Vet Pathol* 1998, **35**:461-478.

-

Additional Files:

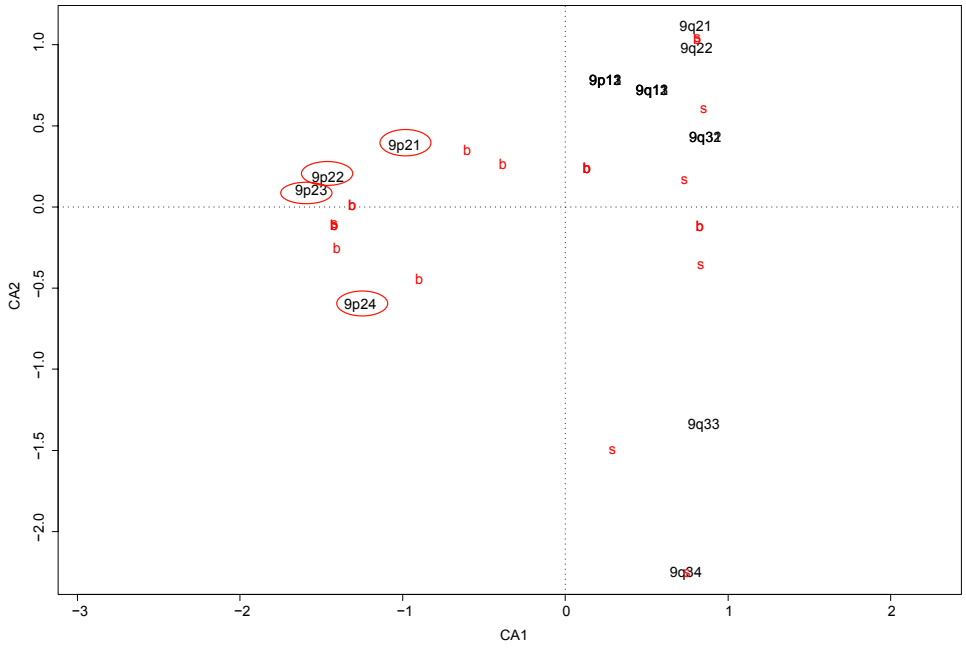
Additional File 1.

File Name: Figure 1S.pdf

File format: PDF

Title: Correspondence analysis of chromosome 9 over the "s" and "b" group.

Description: The first order factor axis separates almost completely these two groups. It is also obvious that the first four bands 9p24, 9p23, 9p22, 9p21 attract most of all b-patients. This leads to the assumption, that these four bands are responsible for the difference of the longer living “s” and the shorter living “b” patients. The second order factor axis separates at first glance strongly the last two bands 9q33 9q34 from the rest.



Additional File 2.

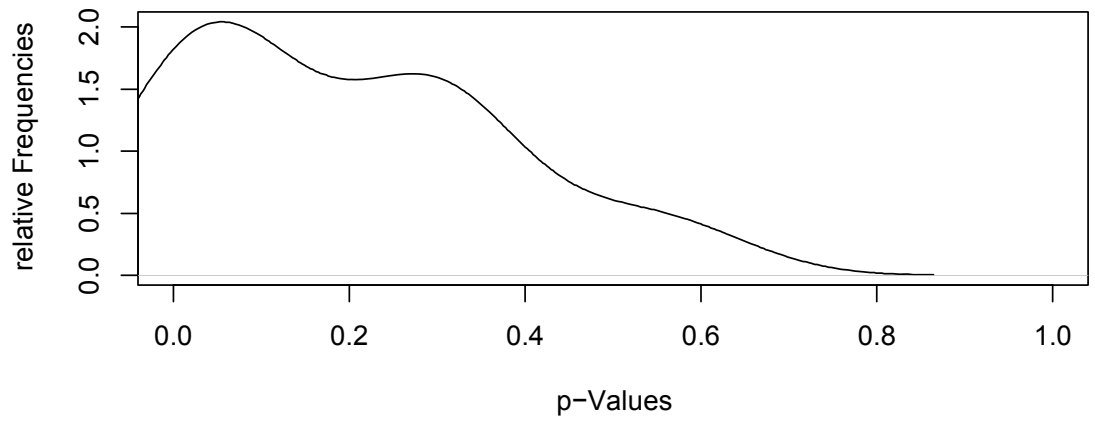
File Name: Figure 2S.pdf

File format: PDF

Title: Density plot of p-values of the Wilcoxon test for the bands of chromosome 9.

Description: The p-values of Wilcoxon test for the bands (x-axis) of chromosome 9 over the subgroups “s” and “b” are represented in their relative frequencies (y-axis). The peak of the first bands indicates that signal of the test ranges from p-value 0 to 0.1. The p-values of the first four bands 9p24, 9p23, 9p22, 9p21 vary between these limits.

This affirms the proposed subgroups "s" and "b" and indicates that the first four bands have a relation to this classification.



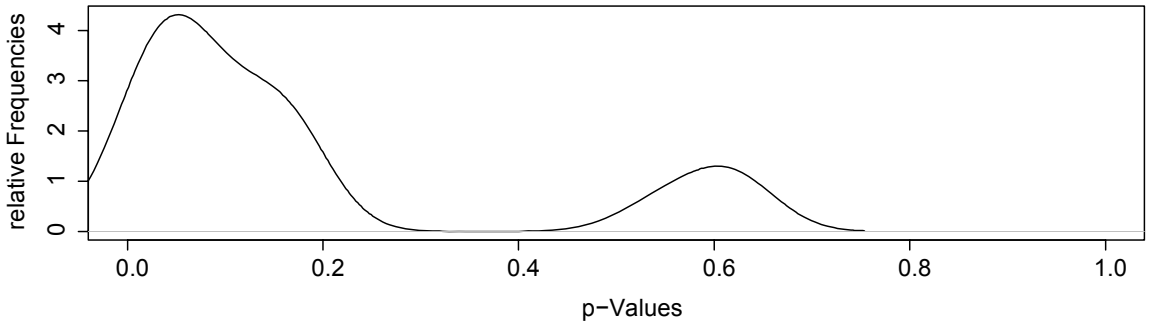
Additional File 3.

File Name: Figure 3S.pdf

File format: PDF

Title: Density plot of p-values of the Wilcoxon test for the bands of chromosome 7.

Description: The p-values from the Wilcoxon test applied on the bands of chromosome 7 are plotted against their relative frequencies. A peak occurs between the limits of 0 and 0.1. The p-values of some bands vary between these limits. These bands are the significant signal of the performed test, affirm the proposed subgroups "s" and "b" and could have a relation to this classification.



Additional File 4.

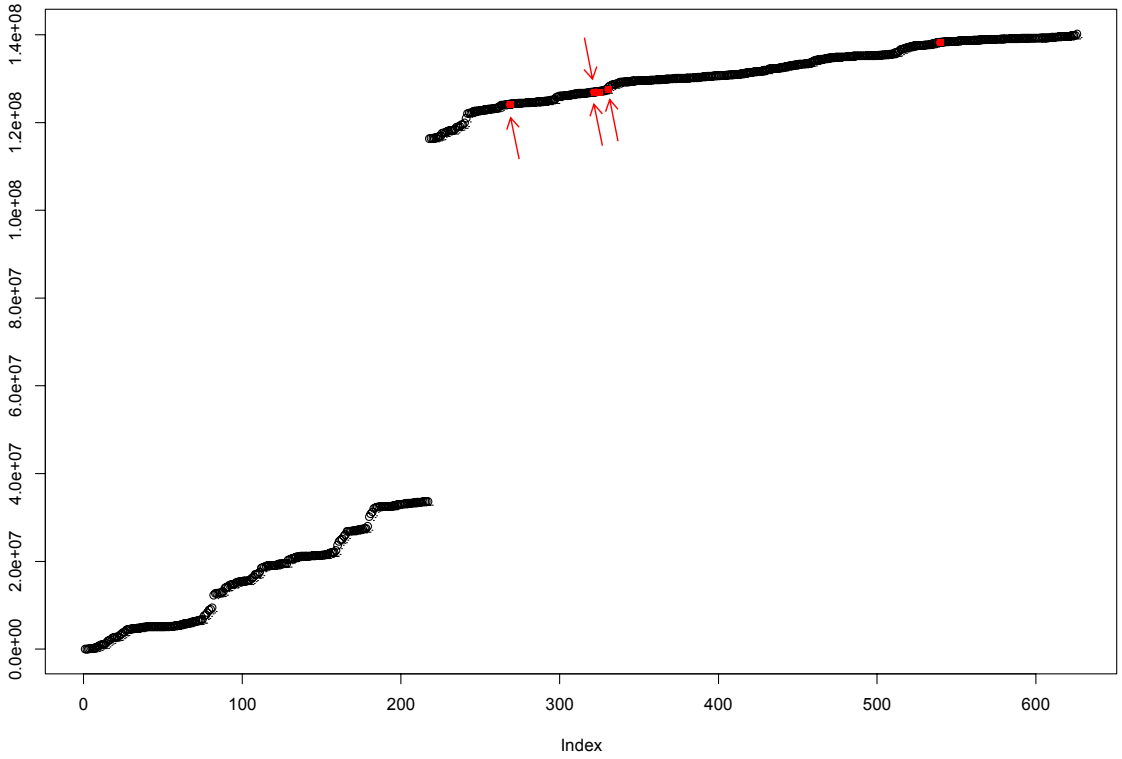
File Name: Figure 4S.pdf

File format: PDF

Title: Plotted base pair positions of genes on Chromosome 9.

Description: Here all genes, which are located on the bands 9p24, 9p21, 9q33, and 9q34 of chromosome 9 are sorted and plotted according to their starting genomic position. The positions are plotted on the y axis. The x-axis represents the genes.

A moderate t-test revealed the best “s” and “b” separating genes in our dataset in these bands. Their starting points are drawn in red. Remarkably three are close to each other.



Additional File 5.

File Name: RatiosBlenk

File format: Text

Title: Gene expression ratios used in this study.

Description: The text file contains all the data (Patients, Ensembl.ID etc.) used for the study after normalization. For the raw intensities please refer to the GEO accession number.

(Text file too large to be printed here, please refer to the supplement of the manuscript at BMC Cancer.)

Additional File 6.

File Name: Prognosis List

File format: Text

Title: Different prognosis assigned to patients

Description: The text file contains how different prognosis can be assigned to patients (over / below median of survival). Please refer to the paper for detailed explanation.

(Text file too large to be printed here, please refer to the supplement of the manuscript at BMC Cancer.)

Chapter 7

Germinal Center B cell-like (GCB) and Activated B cell-like (ABC) type of diffuse large B cell lymphoma (DLBCL): Analysis of molecular predictors, signatures, cell cycle state and patient survival

ORIGINAL RESEARCH

Germinal Center B Cell-Like (GCB) and Activated B Cell-Like (ABC) Type of Diffuse Large B Cell Lymphoma (DLBCL): Analysis of Molecular Predictors, Signatures, Cell Cycle State and Patient Survival

S. Blenk[1], J. Engelmann[1], M. Weniger[1], J. Schultz[1], M. Dittrich[1], A. Rosenwald[2], H.K. Müller-Hermelink[2], T. Müller[1] and T. Dandekar[1]

[1]Department of Bioinformatics, University of Würzburg, Biozentrum, Am Hubland D-97074 Universität Würzburg, Germany. [2]Institute for Pathology, Josef-Schneider-Str. 2, 97080 Würzburg, Germany.

Abstract: Aiming to find key genes and events, we analyze a large data set on diffuse large B-cell lymphoma (DLBCL) gene-expression (248 patients, 12196 spots). Applying the *loess* normalization method on these raw data yields improved survival predictions, in particular for the clinical important group of patients with medium survival time. Furthermore, we identify a simplified prognosis predictor, which stratifies different risk groups similarly well as complex signatures.

We identify specific, activated B cell-like (ABC) and germinal center B cell-like (GCB) distinguishing genes. These include early (e.g. CDKN3) and late (e.g. CDKN2C) cell cycle genes.

Independently from previous classification by marker genes we confirm a clear binary class distinction between the ABC and GCB subgroups. An earlier suggested third entity is not supported. A key regulatory network, distinguishing marked over-expression in ABC from that in GCB, is built by: ASB13, BCL2, BCL6, BCL7A, CCND2, COL3A1, CTGF, FN1, FOXP1, IGHM, IRF4, LMO2, LRMP, MAPK10, MME, MYBL1, NEIL1 and SH3BP5. It predicts and supports the aggressive behaviour of the ABC subgroup. These results help to understand target interactions, improve subgroup diagnosis, risk prognosis as well as therapy in the ABC and GCB DLBCL subgroups.

Keywords: regulation, gene expression, cancer, immunity, prognosis

Introduction

Diffuse large B-cell lymphomas (DLBCL) are the most frequent B cell Non-Hodgkin's lymphomas. Diagnosis relies at present on morphological, immune-phenotypic and laboratory parameters. Clinically, the International Prognostic Index (IPI; age, tumor stage, serum lactate dehydrogenase concentration, performance status, and the number of extranodal disease sites) (The International NHL Prognostic Factors Project, 1993) is often used to predict outcome in DLBCL. On the molecular level, gene expression signatures have been defined that predict outcome in DLBCL independent of the IPI (Rosenwald et al. 2002).

Alizadeh et al. (2000) investigated the gene expression patterns of "diffuse large DLBCL, follicular lymphoma and chronic lymphatic leukemia. They identified two novel distinct types of the DLBCL by gene expression profiling. The "activated B cell-like DLBCL"(ABC) group has a lower overall survival rate than the "germinal centre B cell-like DLBCL" (GCB) group. Von Heydebreck et al. (2001) applied their class discovery method ISIS on a subset of 62 samples and 4026 clones of the data by Alizadeh et al. (2000) and confirmed for these data the two entities ABC and GCB. The survival analysis of Rosenwald et al. (2002), assigned several genes to gene expression signatures and based on this an outcome predictor of survival. The constituents are the "Germinal-center B-cell signature", "MHC class II signature", "Lymph-node signature", "Proliferation signature" and the gene "BMP6". The predictor has a greater prognostic power in classifying patients into risk groups than the IPI (The International Non-Hodgkin's Lymphoma Prognostic Factors Project 1993). Starting with 36 well known DLBCL prognosis genes from the literature, Lossos et al. (2004) found a six gene based outcome predictor and applied it to the data sets of Alizadeh et al. (2000) and Rosenwald et al. (2002). The latter one is an

Correspondence: T. Dandekar, Department of Bioinformatics, University of Würzburg, Biozentrum, Am Hubland D-97074 Universität Würzburg, Germany. Tel: +49-(0)931-8884558, 888-4551; Fax: +49-(0)931-8884552; Email: steffen.blenk@pta.de, dandekar@biozentrum.uni-wuerzburg.de

ongoing study and thus an extension and revision of the old data from Rosenwald et al. (2002) was possible for us (see Material and Methods).

In this study we investigate first the robustness of the data (Rosenwald et al. 2002) with respect to advanced and more appropriate normalization methods. For that, "loess" and "scale" are performed on the data set, as we are aware, for the first time and the results are discussed. Next, unbiased statistical classification analysis confirms for this enlarged data set the classical subgroups ABC DLBCL and GCB DLBCL independent from hierarchical clustering. Furthermore it supports those subgroups being homogeneous entities in the data.

Our analysis includes the expression values for the above 36 DLBCL prognosis genes and we apply more adequate tools from the Bioconductor library (Gentleman et al. 2004) to derive better predictors than e.g. the six-spot predictor found by (Lossos et al. 2004). Moreover, we identify and demonstrate that expression of early and late cell cycle genes distinguishes well the pathological entities ABC and GCB DLBCL.

Finally, we show that the most significant gene expression differences found including cell cycle genes, classical marker genes and all best separating genes are integrated into a compact key regulatory network with clear expression differences between both diffuse large B-cell-lymphoma subgroups. This finding is confirmed comparing the average distribution of genes on the Lymphochip and the connection distances between them in the human interactome as well as by confirming key gene expression differences found in our main data set from new analysis of further gene expression data by Shipp et al. 2002. A picture emerges where a central regulatory circuit tunes immune signatures, apoptotic and proliferation pathways in different ways between ABC and GCB DLBCL. The introduced methods can also be applied to other studies of gene expression analysis in cancer to establish improved prognosis predictors, identify regulatory circuits and for proper group classification.

Materials and Methods

Gene expression data and materials

Patient samples were obtained after informed consent and were treated anonymously during microarray analysis. DLBCL lymph-node biopsies were either snap frozen, frozen in OCT or disaggregated and frozen as a viable cell suspension. DLBCL gene expression was measured with cDNA arrays containing genes preferentially expressed in lymphoid cells or genes known or presumed to be part of cancer development or immune function ("Lymphochip" microarrays (Alizadeh et al. 1999)). Our array includes spots to measure individual exons of the same gene which may be expressed differently in both lymphoma subgroups.

Microarray procedures

Fluorescent images of hybridized microarrays were obtained using a GenePix 4000 microarray scanner (Axon Instruments). Images were analysed with ScanAlyze (M. Eisen; http://www.microarrays.org/software), and fluorescence ratios (along with numerous quality control parameters; see ScanAlyze manual) were stored in a custom database. Single spots or areas of the array with obvious blemishes were flagged and excluded from subsequent analyses. Messenger RNA was extracted according to standard procedures (Sambrook and Russel, 2001) from tumor biopsy specimens of DLBCL patients. All cDNA microarray analyses were performed using poly-(A)+ mRNA (Fast Track, Invitrogen). For each hybridization, fluorescent cDNA probes were prepared from an experimental mRNA sample (Cy5-labelled) and a reference mRNA sample (Cy3-labelled) consisting of a pool of nine lymphoma cell lines (Raji, Jurkat, L428, OCI-Ly3, OCI-Ly8, OCI-Ly1, SUDHL5, SUDHL6 and WSU1). The use of a common reference cDNA probe allows the relative expression of each gene to be compared across all samples.

The original data generated by Rosenwald et al. (2002), in which the subgroups were defined by hierarchical clustering was provided to us by the authors. In our study we analyse an enlarged data set as follows: more patients (a total of 248 patients, each patient array included 12196 gene spots corresponding to 3717 genes), including a more recent classification. The outcome of this are 12.3% more ABC and 5.2% less GCB patients. 19 patients have been removed from the ABC and GCB groups. In detail, five ABC patients were removed from the earlier ABC classification, however, 14 other ones are now associated with it. From the earlier GCB group, 14 patients were assigned to other entities and 11 other patients were newly classified as GCB.

Altogether, 25 patients were thus newly recruited into these two groups. Moreover, each spot is now analyzed in the new study individually. There was no pooling of data on datapoints (spots) as done in older analyses (Rosenwald et al. 2002). We further fully account for the changes in patients analysed (described above) by such an individual spot analysis. In summary this yielded about 3.3 times more data points per patient.

Statistical analyses were performed using the statistical software package R (R Development Core Team 2005) and Bioconductor (Gentleman et al. 2004). For normalization of gene expression data, methods such as vsn, loess and scaling methods were used. To detect differentially expressed genes, functions from the Bioconductor package "limma" were applied. Its special strength is the robust statistics based on linear models and a moderated t-test statistics including multiple testing correction methods (Smyth, 2005, pp 397–420; Smyth, 2004). Based on diagnostic plots we chose gene expression normalization using within-array and between-array normalization methods. The within-array normalization "loess" (Yang et al. 2001, pp 141–152; Yang et al. 2002) adjusts expression log-ratios in the way that they average to zero within each array to make genes on one array comparable to each other. We applied the "scale" method (Yang et al. 2001, pp 141–152; Yang et al. 2002; Smyth and Speed, 2003) for between-array normalization. It scales log-ratios to have the same median-absolute-deviation (MAD) across arrays. By this, log-ratios are normalized to show similar variance across a batch of arrays.

Unbiased class discovery was performed using the ISIS method (*i*dentifying *s*plits w*i*th clear *s*eparation; von Heydebreck et al. 2001). It searches for binary class distinctions in the gene expression levels in an unsupervised fashion. The diagonal linear discriminat score (DLD) quantifies for every found bipartition how strongly the two classes are separated. A maximum sample size of 150 patients for each ISIS run considered 3000 measurements and delivered 50 best separating genes.

Cox regression hazard models were done applying the R package "survival" (Andersen et Gill 1982; Therneau et al. 1990), to calculate the influence of gene expression values on the survival time and Kaplan Meier estimates. The outcome predictor score is calculated with the coefficients of the Cox model and the gene expression values.

Supervised class analyses were performed using "Prediction Analysis of Microarrays" (PAM) (Tibshirani et al. 2002). PAM performs a nearest shrunken centroid method to identify a subset of genes that best characterizes samples as ABC or GCB DLBCL. It computes a standardized centroid for each class and shrinks the prototypes for a given classification error threshold. In the resulting list the obtained optimal (for the given error) shrunken centroid identifier is followed by the number of genes it contains. The chosen classifier is validated by ten-fold cross-validation.

Smaller gene sets typically show larger error rates. However, if almost equally good performing classifiers existed, we parsimoniously chose the one containing the smallest number of genes. The proposed best gene set used for our analysis had 31 spots (labelled by an 'x' character in Fig. 2).

Protein association networks were identified by the STRING database, version 6.3 (von Mering et al. 2005), of known and predicted protein-protein interactions. It combines information from genomic context, experiments, other databases, co-expression and text-mining. Homology predictions transfer and extend these data further. We used the STRING database with a Bayesian confidence level of 0.400 (medium confidence) and a custom limit of 0 (only direct interactions of proteins are considered).

Results

Improving prognosis prediction and separation of DLBCL subtypes

Statistical validation of the DLBCL subgroups ABC DLBCL and GCB DLBCL

Both subgroups were originally introduced on the basis of gene expression profiling. There has been some suggestion that certain diffuse large B-cell lymphomas form a third group (Hans et al. 2004). Furthermore, it was interesting to see whether this classification is also valid for this data set by an unsupervised classification method. To decide independently of any pre-clustering of specific marker genes whether there are two, three or even more lymphoma subgroups and whether they overlap with groups according to other group definitions (e.g. pathology). ISIS (see Materials and Methods) systematically investigates unsupervised

all possible bipartitions of the gene expression data (excluding mediastinal lymphomas; see Materials and Methods) without prior knowledge of marker genes or signature pre-classification (Fig. 1). Nevertheless the bipartitions with the three highest separation scores support and identify the two pathological entities ABC and GCB. Distinct subgroups (splits) within the ABC or GCB entities are not validated by ISIS. In particular, no appropriate bipartition could be observed using previously putatively classified Type 3 patients and the ABC or GCB samples (data not shown). The precise separation into exactly these two subgroups is thus well supported even by an unbiased statistical method independent of predefined expression signatures.

Survival prognosis detection on the updated data and after advanced normalization

The signatures by Rosenwald et al. (2002) are independent from the clinical IPI score (see Introduction) and useful predictors within the low, medium and high IPI risk groups on their data set (Rosenwald et al. 2002). We now tested the performance of advanced normalization methods, namely the methods "loess" (Yang et al. 2001; Yang et al. 2002) and "scale" (Smyth and Speed, 2003; Yang et al. 2001; Yang et al. 2002) on our data set. The IPI score is considered here only as an independent and established clinical prognosis marker. On a normalized data set of 240 patients and considering all individual spots we utilised Kaplan Meier plots (Fig. S1) and reveal the good performance of the gene expression profiles (Rosenwald et al. 2002) also for this data set using the improved normalization procedure. The low risk IPI group in the renormalized data is not as well separated between the best and worst quartile as in Rosenwald et al. (2002). The separation of the high risk group is virtually unchanged. However, in the medium risk group a better separation was achieved by the renormalization and single spot analysis of the enlarged patient data. For the medium risk patients a better separation into high

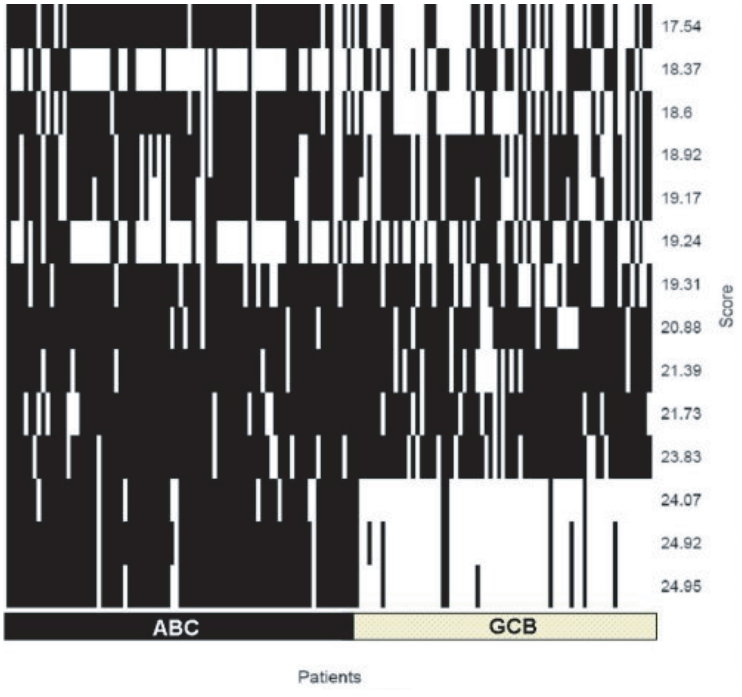


Figure 1. DLBCL splits into sub-groups independent of signatures. Optimal bipartitions of patients are calculated by ISIS based on optimal bipartition subsets of genes (50). Every column of the x-axis represents a patient. On the bottom, the DLBCL-type of the patient is labelled. On the y-axis every row shows the bipartitions ranked in increasing score of separation quality. The three best bipartitions show a very consistent and clear signal separating the ABC- from the GCB-patients. The unsupervised method ISIS reveals the ABC-GCB classification independent of proliferation signatures. No evidence for a previously suggested third group "Type 3" was found. Only a few patients are falsely assigned if compared to the DLBCL gene signature assignment.

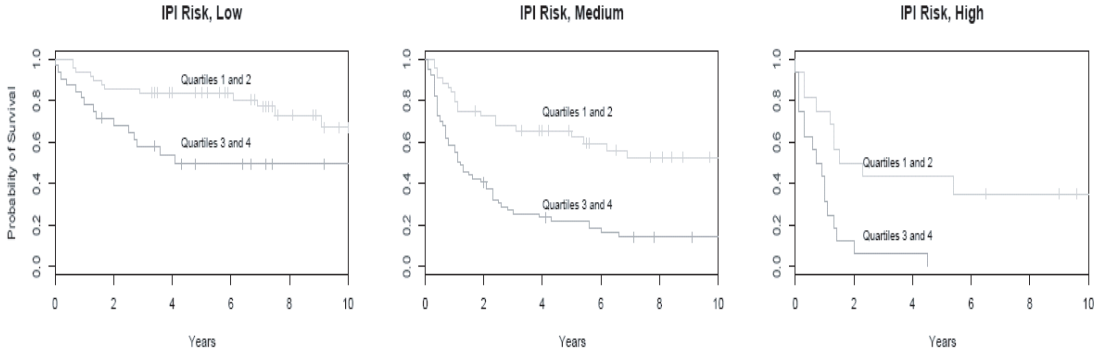


Figure 2. Prognosis prediction applying a molecular predictor of 6 gene spots after improved normalization. Kaplan-Meier plots show large differences in the survival rate for all risk groups. They are estimated by a Cox-Regression Hazard model of the genes listed in Table 1. Normalization was improved applying the "loess" method. x-axis: time (years); y-axis: probability of survival, predicted for the risk groups "low", "medium" and "high".

and low risk is particularly important for prognosis prediction. This method including the advanced normalization can also be applied to any other microarray data set.

An improved six-spot predictor for survival prognosis comparing multi- and univariate analysis

The immune signature requires the measurement of gene expression for many genes. We investigated whether a combination of array spots achieves similar good classification. Multivariate analysis (4 spots results in Table S1 and Table S2, they include immune genes) was computationally prohibitive for more than 4 spots. However, by univariate analysis we could systematically test the capability of gene expression values from individual spots to separate patients with good or bad prognosis in Kaplan-Meier plots. We considered for all three IPI classes the separation of best patient quartile with good prognosis from the worst patient quartile with poor prognosis. Using all genes and the 160 patients from the training-set we identified the spots predicting outcome best. Together, in a multivariate model, they form a predictor separating best and worst quartiles for all three IPI categories including the 80 patients from the validation-set. The five-spot-predictor considers different splicing forms in HLA-DRB5. Five spots (HLA-DPa, Brca, HLA-DQa, and two clones of HLA-DRB5; details in Suppl. Material) are about equal to the six gene predictor of Lossos et al. (2004). However, six genes and spots (HLA-DPa, HLA-DQa, HLA-DRb5, SEPT1, EIF2S2 and IDH3A genes, Fig. 2) show even an improvement for this classification task. The separation of the best and worst quartiles in the three IPI classes is comparable (Fig. 3) to the prediction success of the complete signature of Rosenwald et al. (2002) and classifies different patient quartiles better than the set proposed by Lossos et al. (2004; using LMO2, BCL6, FN1, CCND2, SCYA3 and BCL2 for overall survival in DLBCL). Our predictor is delivered by bioinformatical analysis of gene expression measurements, whereas Lossos et al. used real time PCR. However, our method can also be applied to real time PCR data.

Moreover, we tested the influence of the high correlation between the genes HLA-DPa, HLA-DQa and HLA-DRB5 on the quality of the predictor.

Table 1. Optimal molecular survival predictor applying six genes.

Gene name	Gene description
HLA-DPa	Major histocompatibility complex, class II, DP alpha 1
HLA-DQa	Major histocompatibility complex, class II, DQ alpha 1
HLA-DRb5	Major histocompatibility complex, class II, DR beta 1
SEPT1	Serologically defined breast cancer antigen NY-BR-24=Similar to DIFF6
EIF2S2	Eukaryotic translation initiation factor 2 subunit 2
IDH3A	Isocitrate dehydrogenase 3 (NAD+) alpha

The gene symbol (left side) is followed by the gene description. Three of these genes are HLA major histocompatibility complex genes (HLA).

Table 2. Regulatory network of genes best distinguishing ABC and GCB.

Functional categories	Gene	Description
Proliferation	CCND2	cyclin D2, regulates G1 to S transition of CDK4/CDK6; CTGF, fibroblast growth factor
	MAPK10	map kinase 10
	MYBL1	transcriptional activator in the proliferation of neurons, spermatogenic and B-lymphoid cells (recognition sequence: 5´YAAC(GT)G-3´)
	ASB13	ankyrin repeat and sox box-containing protein 13, mediates protein-protein interactions, sox box couples suppressors of cytokine signalling and binding partners with elongin B and C complex to target them for degradation
	SH3BP5	SH3 domain binding protein, targets protein-protein interaction
Block of proliferation	MME	synonyms CALLA, common acute lymphocytic leukemia antigen, the synonym CD10 stresses its properties as a tumor suppressor gene
	BCL7A	putative tumor suppressor gene in T-cell lymphoma
Apoptosis	BCL2	integral outer mitochondrial protein to block apoptosis
	BCL6	transcriptional repressor, necessary for germinal center formation in lymph nodes
Differentiation	CTGF	fibroblast differentiation
	FOXP1	forkhead box P1
	LMO2	LIM domain only 2 transcription factor for hematopoetic development
	LAMP	expressed in lymphoid cells during development
	COL3A1	collagen type III
	FN1	fibronectin 1, cell adhesion
	NEIL1	base excision repair
Immune cell specific	IGHM	immunoglobulin heavy chain gene
	IRF4	interferon regulatory factor 4

The genes of the network in Figure 4 (suppl.) are associated to the functional categories “Proliferation”, “Block of proliferation”, “Apoptosis”, “Differentiation” and “Immune cell specific”, by their annotation. Most of them are part of the antagonists “Proliferation” and “Block of proliferation”. This indicates the complex regulation and importance of proliferation in the determination of ABC and GCB lymphomas. Classical lymphoma genes (see Table S4) known previously are given in *italics*.

The survival prediction with predictors of non correlated genes from the univariate analysis yields no improvement in the results (data not shown).

Genes best distinguishing DLBCL subgroups

Nearest shrunken centroid analysis using the R-package PAM (“Prediction Analysis of Microarrays”) identifies best separating genes for the two subgroups (ABC and GCB DLBCL) with smallest cross-validation error (Fig. S2). Gene numbers of classifiers are plotted versus the resulting error rates. The optimal classifier (Table S3) requires only 18 genes (31 spots) with an overall cross validation error of 6.2% (5 out of 82 ABC DLBCL samples were falsely predicted as GCB (6.1%); 7 out of 112 GCB DLBCL as ABC (6.25%)). Larger gene sets show similar error rates (see Materials and Methods), smaller gene sets result in inferior classification (Fig. S2). GCB DLBCL is correctly predicted even with fewer genes, however, the error for ABC DLBCL samples increases strongly (Fig. S2 lower plot). For clinical application both entities have to be well separated.

Functional relationship of the genes differently expressed in ABC and GCB

Classical lymphoma gene-markers compared to the identified best separating genes

We tested whether 35 classical lymphoma genes (listed in Table S4; as described in Monti et al. 2005;

Lee et al. 2003; Willis et al. 1999; Polo et al. 2004; Rosenwald et al. 2002) separate well the two major subtypes of DLBCL. Three metabolic enzyme genes for LDH (IPI score prognosis marker), IDH and PDH were added. Altogether these 38 genes correspond to 180 spots. PAM analysis identified a set of 9 well classfying genes (21 spots) (Table S5 and S6), with an overall error rate of 14% (10% training set; 15% for the validation group). However, the classical genes require more spots and their separation is not as good as the optimal prediction set above (Fig. S2). After this we merged these classical lymphoma marker genes with the best separating gene set found above for classification. We found, however, that here the best separating genes achieve all top ranks in this task (Table S7). Only mitogen-activated protein kinase 10 (MAPK10), the best classical lymphoma marker, reaches top ranks. BCL6 as the next best classical marker reaches only rank 31. Below we show that classical lymphoma genes are close to but not identical to the central regulatory network and genes best separating GCB and ABC DLBCL.

Cell cycle genes are differently expressed in ABC and GCB

Cell cycle is critical for cancer cell proliferation and we next investigated by PAM analysis (see Material and Methods) whether the functional group of cell cycle genes alone could separate the two B-cell lymphoma groups. We identified 473 spots, which correspond and are homologous to the cell cycle genes found by de Lichtenberg et al. (de Lichtenberg et al. 2005). These genes are annotated according to expression in the cell cycle state (100 steps between 0 and 99 for a full cell cycle).

The separation between the lymphoma subgroups improves as more genes are used. 77 cell cycle genes (Table S8, Table S9; error rate of 15.4%) yield low error rates using a medium sized gene set (classification optimum, see materials and methods). These include genes such as Butyrophilin-like protein 9 (BTNL9), early B-cell factor (EBF), TSC22 domain family member 1, Cyclin-G2 (CCNG2), Interleukin-6 (IL6), immediate early response protein 5 (IER5) and further homologues of typical cell cycle stage-specific genes (de Lichtenberg et al. 2005) such as TIMP metallopeptidase inhibitor 1(TIMP1) and v-maf musculoaponeurotic fibrosarcoma oncogene homolog (MAF), which mainly reflect the late cell cycle states. Figure 3 compares the complete cell cycle genes in our data set with the subset of 77 genes in a density plot. The black line indicates all cell cycle states of the whole chip and the blue line the subset of 77 genes. The densities of these gene sets clearly differ in the early (steps 0–18) and in the late steps (75–85) of cell cycle ($p = 6.65 \cdot 10^{-10}$; Wilcoxon one sided test).

Cell cycle spots, which show the biggest difference in gene expression values between ABC and GCB DLBCL, are in the late steps 72, 80, 84 and 85 (Fig. S3; M/A plot, ie,middle intensity of the genes against difference in expression of both lymphoma subgroups). Moreover, these cell cycle states form a compact cluster in the plot. This data indicate a clear difference in cell cycle states regarding the two DLBCL subgroups.

Cell cycle genes, classical lymphoma genes and best separating genes form a compact network important for DLBCL subtype distinction between ABC and GCB

Are the genes differentially expressed in ABC and GCB DLBCL specially connected, and in particular, if so, how do their respective gene products interact with each other? To analyze this systematically, different large scale protein interaction databases were investigated such as the hand curated HPRD database (Peri et al. 2003). The large protein-protein interaction database STRING (von Mering et al. 2005) allowed us to establish an interaction network (Fig. S4, Fig. S5). Note that this analysis focuses on the clearly differentially expressed genes in ABC and GCB (Table S7). Classical lymphoma gene markers (dark grey boxes) as listed in Table S5 combine and interact with the compact cluster of the most powerful differentiating genes (white boxes) for the whole data set (Table S3) as delivered by PAM. The connections are mainly found by text-mining; however, the two interactions between BCL6—IRF4 and between SH3BP5—MAPK10 are available from the HPRD data set (experimental/biochemical data) as a direct physical interaction (blue). The different article sources re-examine the interaction predictions for different cancer entities: “DLBCL”, “no cancer disease” and “other cancer”. Note that these categories support the interactions from three different view points (Fig. S5). We find that 11 of the 18 best separating genes and 8 of the 9 separating classical lymphoma genes are members

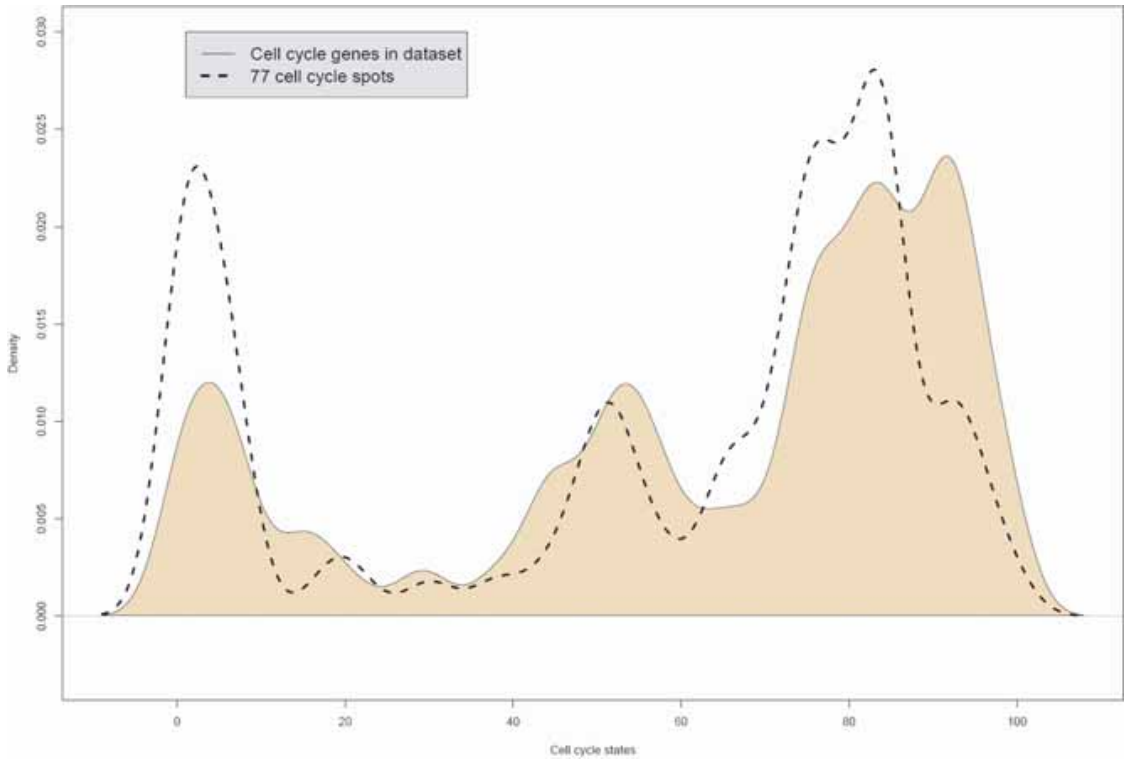


Figure 3. Early and late cell cycle genes are overrepresented in the best separating cell cycle gene set. The density plot compares the distribution of different cell cycle gene sets. x-axis: cell cycle states (from 0 to 99; complete cell cycle). y-axis: relative frequencies. Black line: density of all mapped cell cycle genes of de Lichtenberg et al (de Lichtenberg et al. 2005) in the data set. The area under this line is coloured for easier comparison. Blue line: Optimal separating subset of cell cycle genes (77 spots). Two peaks in the early and late cell cycle states show cell cycle gene expression differences between the subgroups ABC and GCB.

of this dense interaction network. This is supported by the interaction data, the HPRD database and various specific interaction evidence types collated by the STRING database.

The remaining 8 genes, 7 from the first mentioned set and 1 from the latter one, are not part of the databases. Cyclin D2 (CCND2) occurs in both subsets and we obtain a protein association network of 18 nodes. Regarding network regulation the underlined genes are higher expressed in ABC, all others are higher expressed in GCB subtype: ASB13, BCL2, BCL6, BCL7A, CCND2, COL3A1, CTGF, FN1, FOXP1, IGHM, IRF4, LMO2, LRMP, MAPK10, MME, MYBL1, NEIL1 and SH3BP5 (Table S10). The characteristics of the network are described in Table 2: Protein functions involved in the network include stimulation of proliferation, block of proliferation, apoptosis, differentiation and immune cell specific functions. Both DLBCL subgroups show clear differences in these specific pathways and sub-networks. Furthermore, the large collection of protein associations from the STRING database shows that all these different proteins separating the two subgroups are connected by first order interactions. As a control for this finding of a compact regulatory network separating both entities regarding gene expression, we tested that all Lymphochip genes are equally distributed with regard to the human interactome and not pre-clustered (Fig. S6). Moreover, the characteristic path length for randomly picked genes from the Lymphochip is 3.985 (Fig. S7) and clearly longer than the direct interactions (path lengths one or two) found for the differentially regulated network (Fig. S4).

Moreover, 5 of the 8 cell cycle genes, identified in Figure S3 above, to be regulated differently are directly interacting with this regulatory network (Fig. S5). The genes with a significantly higher expression in the ABC group are marked by a red

rectangle, whereas green ellipses mark higher expression in GCB. These differences are an interesting pointer for a more specific anti-cancer treatment.

Gene functions for well separating genes

The shorter survival of patients with ABC DLBCL is connected to pathways expressed differently from GCB DLBCL; thus the well known BCL2, as a central apoptosis blocker is higher expressed and allows cancer cell survival in ABC DLBCL. BCL6, a transcriptional repressor important for B-cell differentiation, is down-regulated in ABC DLBCL. Altogether, apoptosis genes are lower expressed in the ABC DLBCL subtype.

Furthermore, the low gene expression values of the gene MME, a proliferation blocker, CCND2 and BCL7A, both genes which promote proliferation, and high values of SH3BP5 in the ABC DLBCL patients stimulate proliferation.

Both the immune cell specific genes IGHM and IRF4 are higher expressed in ABC DLBCL; however, all genes which are associated with differentiation are down-regulated.

In conclusion, this network indicates down-regulation of apoptosis and differentiation for the ABC DLBCL patients whereas the proliferation and immune cell stimulating genes are up-regulated.

From the cell cycle genes which are connected to the network, IL6 and IER5 show higher values in the ABC group whereas BTNL9 and CCNG2 show an up-regulation in the GCB group. For the latter it is known that CCNG2 and IL6 block the proliferation.

In order to further validate the found gene expression differences, we show that several of these are found again after analyzing further data from Shipp et al. (Shipp et al. 2002; Wright et al. 2003; Table S12).

Do the clear gene expression differences between both subgroups reflect only differences in B-cell specific regulation? In order to gain a first impression regarding T-cell regulatory pathways from our data we tested whether notch genes, trans-membrane receptors important in T cell differentiation and repressed in many cancers (Reizis and Leder, 2002), regulate differently the target genes in the two groups. Target genes are regulated by GY-box-, Brd-box-, and K-box-class microRNAs in the 3'-UTRs e.g. in Drosophila (Lai et al. 2005). We mapped all genes of the Lymphochip to the transcripts annotated in ensembl. We screened these and found candidate notch target genes, whose transcripts bear the mentioned target sequences. All three boxes were found in the genes given in supplementary Table S11. From these transcripts the "Deoxycytidine kinase" gene (ENSG00000156136, DCK) and the "Translocation associated membrane protein 2" (ENSG00000065308, TRAM2) show clear gene expression differences between the ABC and GCB subgroups.

Discussion

Marker genes for DLBCL subtypes

This study improves marker gene detection for prognosis and subtype diagnosis of diffuse large B-cell lymphomas (DLBCL) applying a wide range of methods useful also for other gene expression measurements in cancer. A special patient group are primary mediastinal B-cell lymphomas. Patients recognized with this disease (6 cases) were excluded from the data set and hence are neither visible nor contained in the further analysis. This is in accordance with previous studies (Rosenwald et al. 2002) and other data sets (Alizadeh et al. 2000; Shipp et al. 2002; Wright et al. 2003).

The classification of all other diffuse large B-cell lymphoma into two pathological entities has been established by marker genes and their expression (Alizadeh et al. 2000). A third entity has been discussed (Hans et al. 2004) but was disputed again in the light of recent data. Our statistical analysis by ISIS method (von Heydebreck et al. 2001) provides an independent method and validates and supports only these two subgroups. In addition to previous work (Rosenwald et al. 2002), ISIS analysis here clearly indicates for a large data set the bipartition of all patient data into the two subgroups ABC and GCB through an unbiased and independent statistical method. An adequate normalization of the gene expression intensities applying the loess method (Yang et al. 2001; Yang et al. 2002) allowed a better separation for best and worse outcome quartiles of survival, in particular for patients with medium IPI score where a better separation is important for accurate prognosis. We found a simplified (6 instead of 17 gene spots) survival predictor useful for clinical monitoring e.g. applying RT-PCR (Lossos et al. 2003).

Multivariate analysis showed that a four-spot predictor does not perform well. However, univariate analysis found a six spot prognosis predictor which is superior to a previous six-spot predictor (Lossos et al. 2004) and to an alternative five spot predictor, in particular regarding high risk patients.

Integrated picture of all gene regulation differences

Following this, the statistical analysis identified all genes which well distinguish the ABC and GCB DLBCL subgroups including differences in early and late cell cycle which could be exploited for a differential cytostatic therapy in the two subgroups.

We considered all the identified gene expression differences in order to obtain a detailed description of the differences between both DLBCL subgroups regarding regulation of the cellular network. We show that immune signatures, apoptotic and proliferation pathways are tuned in different ways between ABC and GCB DLBCL. A central circuit of genes is formed by genes that distinguish both lymphoma subgroups and are regulated differently. We also verified this for other data after completion of the first analysis. For the data in Shipp et al. (2002) and Wright et al. (2003) once again key genes from the central network shown in Figure S4 are confirmed as having a significant different regulation in this totally different data and patient set (Table S12). Classical lymphoma genes are either directly or indirectly interacting with it. Besides this central network other pathways are also implicated, we showed that two Notch pathway targets are specifically up-regulated. PAM has been shown previously to be a powerful method for gene selection (Tibshirani et al. 2002).

The different predictors shown in this study were the best predictors according to PAM curves and statistical analysis and gave clear improvements for prognosis prediction compared to previous studies (Rosenwald et al. 2002; Lossos et al. 2004) including a six spot predictor for clinical application. Furthermore, our results are based on experimental gene expression data on 248 patients and individual analysis of 12196 array spots whereas pooled data and fewer patients were used in older studies (Rosenwald et al. 2002; Lossos et al. 2004). Interesting marker genes were found in this study by different statistical methods (PAM, ISIS, LIMMA). Clearly, using other methods (e.g. support vector machines) different gene sets can be obtained. In our study, the ISIS method is applied for explorative analysis and unbiased classification without prior knowledge or gene signatures. It supports independently the two distinct B-cell lymphoma subgroups. The different gene sets were further validated against each other by including classical marker genes. Moreover, we validate in our study key marker genes we found by analysis of additional and further data (Shipp et al. 2002; Wright et al. 2003). A new perspective from this study is that genes found differently expressed in the two B-cell lymphoma types form a compact interaction network including cell cycle genes. This is obtained by another independent analysis method (protein-protein interaction database STRING). Furthermore, the delineated regulatory network adds biological data and data from large-scale interaction databases to show that the identified marker genes are in fact members of a closely interacting regulatory network, with molecular functions that mirror the differences in pathology of the two subgroups GCB and ABC DLBCL.

The identification of cell cycle genes expressed differently indicates here new possible targets for therapy. Differences between the ABC and GCB DLBCL subgroups are at the beginning and the end of the M-phase and the early part of the G1 phase. Inhibiting early cell cycle genes, overexpressed in ABC and adding known cytostatic drugs such as mitosis inhibitors and early G1 blocker may be particularly useful for ABC DLBCL patients. A more detailed therapy profile would take the further differences in regulation into account.

Conclusion

The present analysis reveals through the use of an array of methods a detailed picture of molecular markers differentiating cancer subtypes. We apply it to GCB and ABC DLBCL for clinical use in determining prognosis and diagnosis. This included efficient six spot predictors for prognosis and clinical application. The entities ABC and GCB DLBCL have been confirmed by statistical analysis independent of gene expression signatures, a third entity could not be supported. The resulting genes with altered expression were found to form a tightly connected regulatory network including cell cycle genes, apoptosis and immune

differentiation implicated in the aggressive behaviour of ABC DLBCL compared to the GCB DLBCL subtype.

Acknowledgments

We thank the State of Bavaria for support (IZKF B-36; ENB Lead Structures of Cell Function; SFB688 TP A2).

References

Alizadeh, A.A., Eisen, M.B., Davis, R.E. et al. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–11.

Alizadeh, A., Eisen, M., Davis, R.E. et al. 1999 The lymphochip: a specialized cDNA microarray for the genomic-scale analysis of gene expression in normal and malignant lymphocytes. *Cold Spring Harbor Symposia on Quantitative Biology*, 64:71–8.

Andersen, P. and Gill, R. 1982. Cox's regression model for counting processes, a large sample study. *Annals of Statistics*, 10:1100–1120.

de Lichtenberg, U., Jensen, L.J., Brunak, S. et al. 2005. Dynamic complex formation during the yeast cell cycle. *Science*, 307(5710):724–7.

Gentleman, R., Carey, V., Bates, M. et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80.

Hans, C.P., Weisenburger, D.D., Greiner, T.C. et al. 2004. Confirmation of the molecular classification of diffuse large B-cell lymphoma by immunohistochemistry using a tissue microarray. *Blood*, 103(1):275–82.

Lai, E.C., Tam, B. and Rubin, G.M. 2005. Pervasive regulation of Drosophila Notch target genes by GY-box-, Brd-box-, and K-box-class microRNAs. *Genes and Development*, 19(9):1067–80.

Lee, J.W., Yoo, N.J., Soung, Y.H. et al. 2003. BRAF mutations in non-Hodgkin's lymphoma. *British Journal of Cancer*, 89 (10):1958–60.

Lossos, I.S., Czerwinski, D.K., Alizadeh, A.A. et al. Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes. *New England Journal of Medicine*, 350(18):1828–37.

Lossos, I.S., Czerwinski, D.K., Wechser, M.A. et al. 2003. Optimization of quantitative real-time RT-PCR parameters for the study of lymphoid malignancies. *Leukemia*, 17(4):789–95.

Monti, S., Savage, K.J., Kutok, J.L. et al. 2005. Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. *Blood*, 105 (5):1851–1861.

Peri, S., Navarro, J.D., Amanchy, R. et al. 2003. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Research*, 13:2363–2371.

Polo, J.M., Dell'Oso, T., Ranuncolo, S.M. et al. 2004. Specific peptide interference reveals BCL6 transcriptional and oncogenic mechanisms in B-cell lymphoma cells. *Nature Medicine*, 10(12):1329–35.

Development Core Team, R. 2005. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Reizis, B. and Leder, P. 2002. Direct induction of T lymphocyte-specific gene expression by the mammalian Notch signaling pathway. *Genes and Development*, 16(3):295–300.

Rosenwald, A., Wright, G., Chan, W.C. et al. 2002. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *New England Journal of Medicine*, 346(25):1937–47.

Sambrook, J. and Russell, D.W. 2001. Molecular Cloning. A laboratory Manual. 3rd Edition. Cold Sprin Harbor Laboratory Press. Cold Spring Harbor, New York.

Shipp, M.A., Ross, K.N., Tamayo, P. et al. 2002. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.*, 8(1):68–74.

Smyth, G.K. 2005. Limma: linear models for microarray data. In: Gentleman R, Carey V, Dudoit S, et al(eds). Bioinformatics and Computational Biology Solutions using R and Bioconductor. New York: Springer. 397–420.

Smyth, G.K. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, 3:Article 3.

Smyth, G.K. and Speed, T.P. 2003. Normalization of cDNA microarray data. *Methods*, 31:265–273.

The International Non-Hodgkin's Lymphoma Prognostic Factors Project, . 1993. A predictive model for aggressive non-Hodgkin's lymphoma. *New England Journal of Medicine*, 329(14):987–94.

Therneau, T., Grambsch, P. and Fleming, T. 1990. Martingale based residuals for survival models. *Biometrika.*

Tibshirani, R., Hastie, T., Narasimhan, B. et al. 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 99(10):6567–72.

von Heydebreck, A., Huber, W., Poustka, A. et al. 2001. Identifying splits with clear separation: a new class discovery method for gene expression data. *Bioinformatics*, 17 (Suppl 1):S107–14.

von Mering, C., Jensen, L.J., Snel, B. et al. 2005. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, 33(Database issue): D433–7.

Willis, T.G., Jadayel, D.M., Du, M.Q. et al. 1999. Bcl10 is involved in t(1; 14)(p22; q32) of MALT B cell lymphoma and mutated in multiple tumor types. *Cell,* 96(1):35–45.

Wright, G., Tan, B., Rosenwald, A. et al. 2003. A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. *Proc. Natl. Acad. Sci. U.S.A.*, 100(17):9991–6. Epub 2003 Aug 4.

Yang, Y.H., Dudoit, S., Luu, P. and Speed, T.P. 2001. Normalization for cDNA microarray data. In: M. L. Bittner, Y. Chen, A. N. Dorsel, and E. R. Dougherty (eds). Microarrays: Optical Technologies and Informatics. *Proceedings of SPIE*, 4266:141–152.

Yang, Y.H., Dudoit, S., Luu, P. et al. 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.*, 30(4):e15.

Germinal Center B Cell-Like (GCB) and Activated B Cell-Like (ABC) Type of Diffuse Large B Cell Lymphoma (DLBCL): Analysis of Molecular Predictors, Signatures, Cell Cycle State and Patient Survival

S. Blenk[1], J. Engelmann[1], M. Weniger[1], J. Schultz[1], M. Dittrich[1], A. Rosenwald[2], H.K. Müller-Hermelink[2], T. Müller[1] and T. Dandekar[1]

Supplemental Methods

To systematically identify spots which describe the outcome and cooperate well with each other in the Cox regression hazard model a multivariate analysis is desirable. However, this requires a huge search space of combinations to be tested. To reduce this we considered only four spot combinations of (i) the gene spots suggest by Rosenwald et al. (Rosenwald et al. 2002), (ii) the 36 important genes for diffuse large B-cell lymphoma chosen by Lossos et al. (Lossos et al. 2004) or (iii) the LDH-, IDH-, and PDH gene spots (the latter to better reflect IPI-scores). Cox Regression Hazard analysis was performed on all possible four tuples of these 153 indicator spots testing 160 patients (several days of calculation time on a LINUX cluster with 20 nodes of Pentium IV CPUs). Table S1 shows the gene content of the ten best multivariate four-spot-predictors (the next best combinations after removing these spots is found in Table S2). The best multivariate four-spot combination is compact and small, but neither as good as the five spot predictor in results nor as the signatures from Rosenwald et al. (Rosenwald et al 2002). The analysis further shows that there is a correlation with survival prediction for the clinical parameter LDH (Table S2), but the prediction based on this well known parameter (part of the IPI score) is even worse then the results shown in Table S2.

In contrast (see below), the new five-spot and six-spot predictors identified by univariate analysis will be useful heuristics for diagnosis and clinic, e.g. to identify risk quartiles and subgroups (Fig. S1).

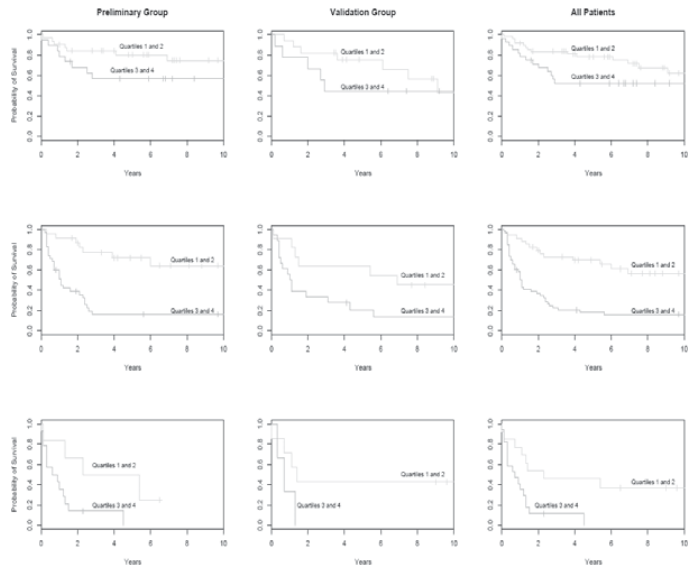


Figure S1. Kaplan Meier plots of the IPI groups. The Kaplan Meier plots estimated by the molecular predictor of Rosenwald et al. (Rosenwald et al. 2002) applied on the new normalized gene expression data of the 240 diffuse large B-cell lymphoma patients. The plots show different groups according to their IPI risk and the training set as Training, Validation and all patients. The left column represents the training-group, the middle one the validation group and the right one all patients. The rows show the IPI risk groups. The first line shows low risk, the second one the medium risk and the last line the high risk patients. The x-axis is the time in years and the y-axis the probability of survival.

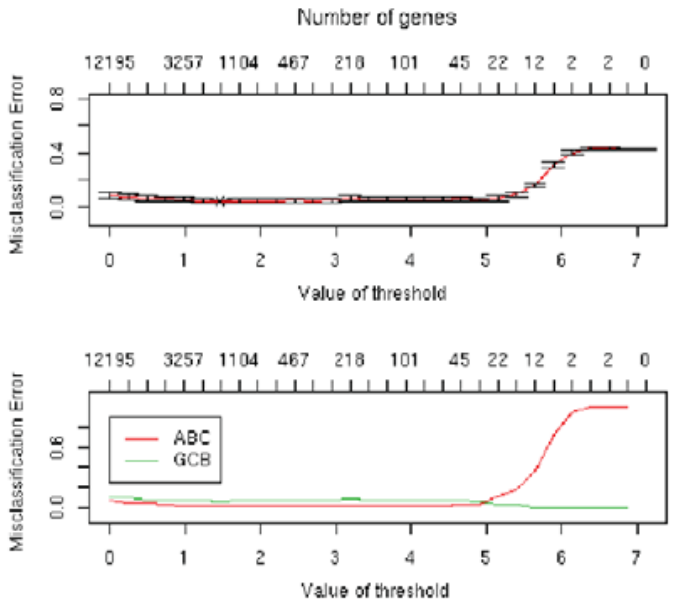


Figure S2. PAM misclassification error of the ABC and GCB subgroups over all genes. The upper plot shows the overall error while the lower one shows the subgroup specific errors. In both, the various thresholds on the lower x-axis correspond to different numbers of genes, labelled on the upper x-axis. The y-axis represents the error and ranges from 0 to 1. The good overall performance of PAM requires only few genes to decrease the error dramatically. The error rate decreases strongly between the thresholds of 6 and 5, which represent the amount of shrinkage. Hence we chose a threshold below 5 with the corresponding set of best separating genes (an optimal choice with few errors and a low number of genes). The performance for the single subgroups shows a big difference between ABC and GCB. Whereas GCB shows a good performance even with few genes, the prediction quality of ABC decreases dramatically in the case of ABC patients. This indicates a complex pattern of gene expression in ABC patients which is defined in more than 15 genes.

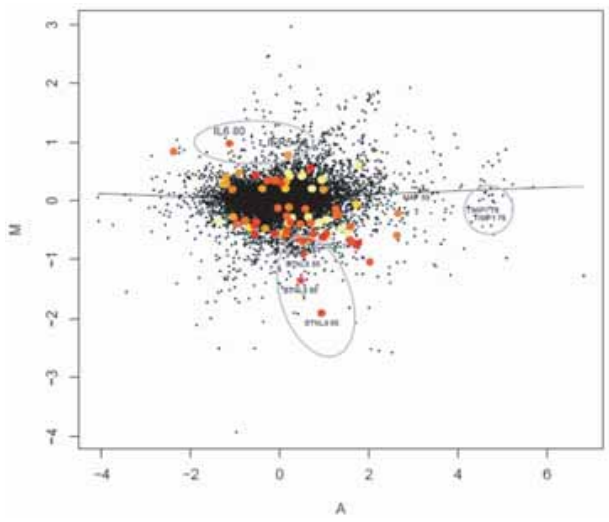


Figure S3. Cell cycle genes with extreme expression differences shown by a MA-plot of normalized gene expression values. The M values on the y-axis correspond to the gene expression difference between the ABC and GCB patient medians and the A values on the x-axis correspond to the average expression of all genes in both groups. The colored points represent the 77 cell cycle spots chosen by PAM analysis. The color scale ranges from yellow to red, whereas yellow is annotated to cellcycle state 0 and red to state 99. Additionally some cell cycle genes show more extreme A values(circle). They are labeled with their names and their cell cycle state. Remarkably, some genes associated with a late high cell cycle state cluster together regarding their gene expression values in both dimensions (ellipse). Again, late cell cycle states indicate a high difference in the M-value (difference in gene expression) between the two subgroups. A locally weighted regression smoothing line (lowess) shows that systematic and random variations are well controlled by the normalization procedure: Its shape fits almost perfectly the horizontal line.

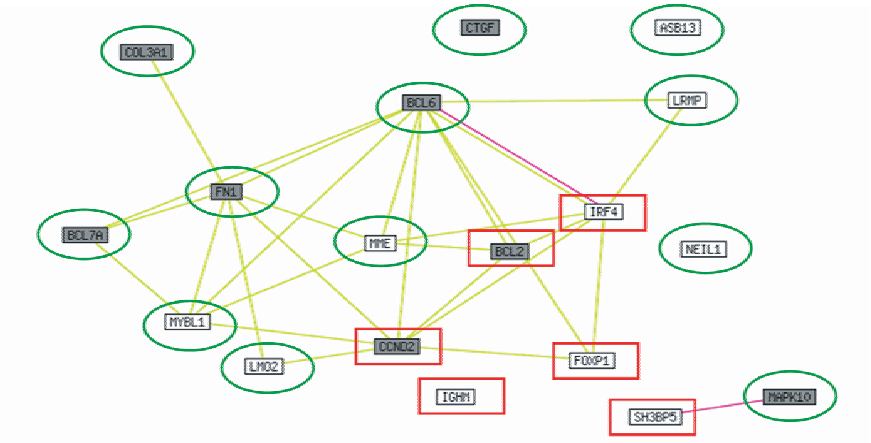


Figure S4. Regulatory network differently regulated in ABC and GCB B-cell lymphomas. This figure shows the resulting network and interaction pattern with each other for the best separating genes applying data from the STRING meta-database of protein interactions. Classical lymphoma genes and best separating gene set form a tight network with the best separating genes in the centre. Shown are the strongly connected network members. They consist of (i) classical lymphoma marker genes (grey boxes), and (ii) the most powerful predictive genes in the PAM analysis (white boxes). Genes which show a significant higher expression in the ABC subgroup are marked by a red rectangle. They are associated to proliferation, block of proliferation, apoptosis, differentiation and specific for immune cells, as most of the remaining ones. Green ellipses mark higher expression in GCB. The almost fully connected gene network demonstrates that both classes of genes are well participating in the interaction network according to the STRING meta-database. Furthermore, the STRING analysis shows that almost all connections between both classes – the yellow colored edges - are based on literature (mainly Medline reports). Only the interaction of "interferon regulatory factor 4" (IRF4) and "B-cell CLL/lymphoma 6" (BCL6) is confirmed by large-scale interaction screen experiments.

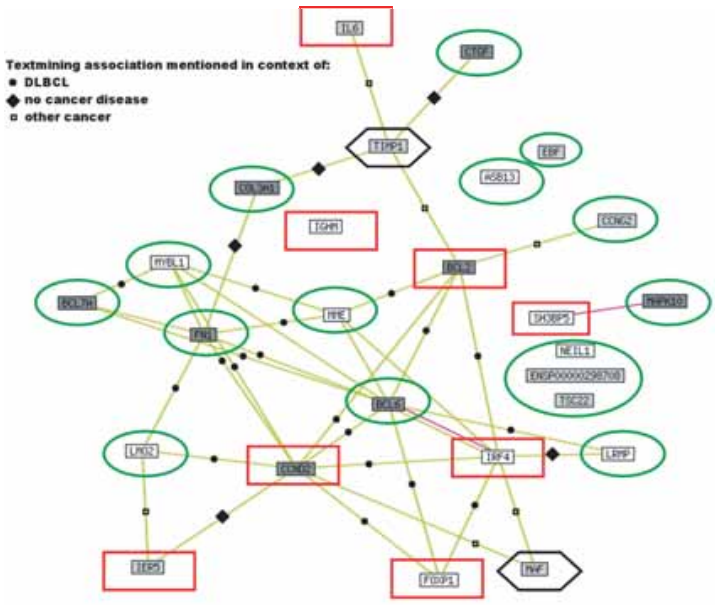


Figure S5. Regulatory network differently regulated in ABC and GCB B-cell lymphomas. Functional protein association network using interactions predicted by the STRING database: the most powerful predictive genes in the PAM analysis (white boxes; see Figure 4S), classical textbook lymphoma genes (dark grey boxes), additional the cell cycle genes (light grey boxes; see Figure 3S: 5 of these 8 cell cycle genes are connected directly with the network. TIMP1 even connects the so far uninvolved classical lymphoma gene CTGF with the network. This indicates how well the cell cycle genes fit to the existing graph). The new connections are confirmed by text mining of PubMed abstracts(circles: DLBCL, diamonds: "no cancer disease", empty square: "other cancer"); these different data complement each other. The genes with a significantly higher expression in the ABC group are marked by a red rectangle. Green ellipses mark higher expression in GCB. Black hexagons mark genes which have a very high average gene expression value in both entities and are an important part for the network.

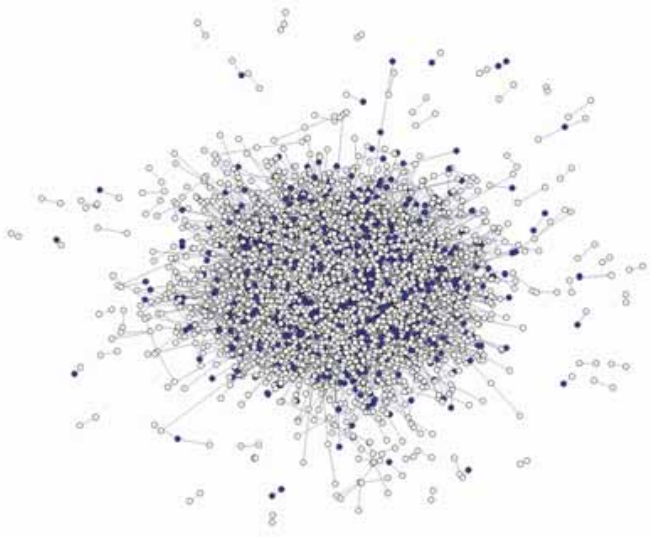


Figure S6. The Lymphochip genes in the human interactome. This plot shows the human interactome as a protein interaction network. The proteins(circles) of the lymphochip are filled out. Interactions are drawn as a line. Characteristic path length and the longest path are 4.642 and 15, respectively.

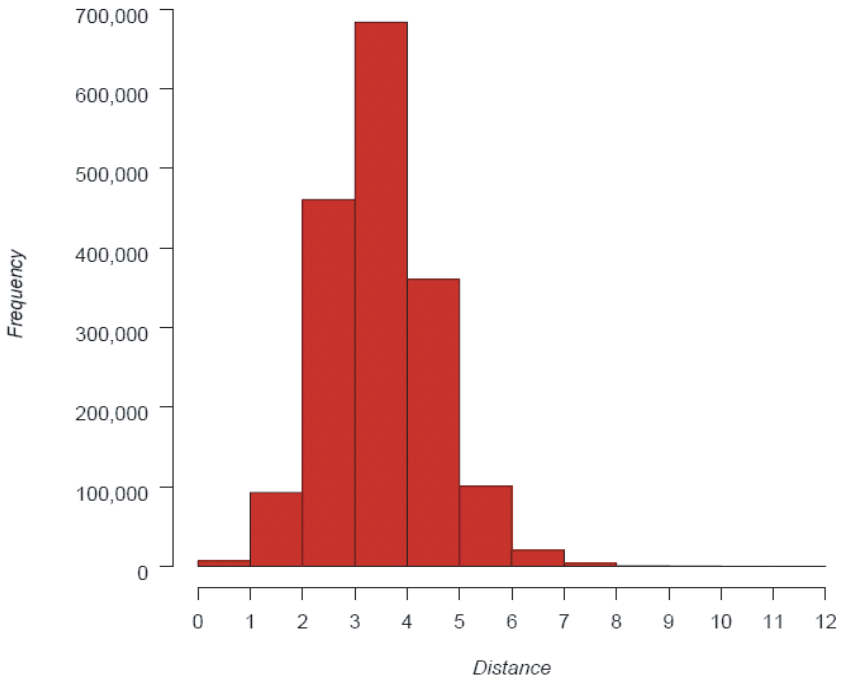


Figure S7. Histogram of the protein interaction distances. The genes of the Lymphochip were mapped to the protein interaction graph in the human interactom. The histogram shows the occurring distances of these genes in the interactome. The longest distance is 11 whereas the characteristic path length is 3.985.

Table S1. Multivariate Cox regression hazard models.

Nr	Multivariate Cox regression hazard model			
1	HGAL	Germ-S	ACTa1	HLA-DRA
2	HGAL	CD54(2)	ACTa1	HLA-DRA
3	HGAL	CD54(2)	HLA-DRA(2)	ACTa1
4	HGAL	CD54(2)	HLA-DRA(3)	ACTa1
5	HGAL	ACTa1	HLA-DRA	CD54
6	HGAL	MHCIIDQa1	CD54(2)	ACTa1
7	HGAL	CD54(2)	MHCIIDRb	ACTa1
8	HGAL	Germ-S	MHCIIDRb	ACTa1
9	HGAL	Germ-S	HLA-DRA(2)	ACTa1
10	HGAL	Germ-S	HLA-DRA(3)	ACTa1

A heuristic search of multivariate Cox regression hazard models revealed this 10 best fitting models. All possible multivariate Cox regression hazard models of four 4 genes from 36 important genes for diffuse large B-cell lymphoma and the metabolic genes LDH, IDH and PDH were calculated and these ten gene sets fit best. Genes are abbreviated according to GenBank nomenclature.

Table S2. Next best multivariate Cox regression hazard models.

Nr.	Multivariate Cox regression hazard model			
1	CD10	IRF4	HLA-DRb5	LDH(2)
2	IRF4(2)	BCL7A	HLA-DRb5	LDH(2)
3	MYC	IRF4(2)	HLA-DRb5	LDH
4	MYC	IRF4(2)	HLA-DQa1	LDH
5	PLAU	IRF4	BCL7A	HLA-DRb5
6	IRF4	BCL7A	HLA-DRb5	LDH(2)
7	PLAU	IRF4(2)	BCL7A	HLA-DRb5
8	IRF4	BCL6	BCL7A	HLA-DRb5
9	CD10	IRF4(2)	HLA-DRb5	LDH(2)
10	MYC	IRF4(2)	HLA-DRb5	LDH(2)

If the genes appearing in Table S1 are removed, and the heuristic search of multivariate Cox regression hazard models is redone, these ten models are the next best fitting. The genes are represented by their GenBank abbreviation. The metabolic marker LDH from the IPI score occurs in the four best fitting models as well as in the the majority of the models.

Table S3. Genes which distinguish best between ABC and GCB according PAM analysis.

Nr.	Gene
1	MYBL1
2	*Centerin
3	FOXP1
4	LOC96597
5	SH3BP5
6	KIAA0864
7	IRF4
8	ASB13
9	*Similar to human endogenous retrovirus-4 Clone=417048
10	NEIL1
11	MME
12	IGHM
13	LMO2
14	LOC152137
15	KIAA1039
16	LRMP
17	FLJ123633
18	CCND2

From all twelve thousand spots from the lymphoma chip, the listed genes distinguish best between ABC and GCB according to PAM analysis. The best separating genes are written on the top.

Table S4. Classical lymphoma genes.

Nr.	Gene
1	BCL6
2	BRAF
3	ARAF1
4	RAF1
5	RAS
6	MEK
7	MAP
8	HLA-DPα
9	HLA-DQα
10	HLA-DRα
11	HLA-DRβ
12	α-Actinin
13	COL3A1
14	Connective-tissue growth factor
15	FN1
16	KIAA0233
17	PLAUR
18	E2IG3
19	NPM3
20	BMP6
21	CASP10
22	POU2AF1
23	CDKN2A
24	MYC
25	BCL2
26	FCGR2B
27	CyclinD1
28	NFKB2
29	PAX5
30	BCL10
31	CDK6
32	DDX6
33	BCL7A
34	CyclinD2
35	IL-10
36	LDH
37	IDH
38	PDH

Lymphoma associated genes were collected from literature and were also found in the data set. Furthermore we added the metabolic enzymes “lactate dehydrogenase”(LDH), “isocitrate dehydrogenase” (IDH) and “pyruvate dehydrogenase”(PDH). The latter are represented in the data by the genes PDHB, PDHA1, IDH3A, IDH3G, IDH3B, IDH1, IDH3B, IDH3A, LDHB and LDHA.

Table S5. Classical marker genes of lymphoma disease distinguish between ABC and GCB lymphoma subtype (PAM analysis; error rates for this gene set: TR:10% VAL:15.38%; F:CV:14%))

Nr.	Gene
1	FN1
2	BCL6
3	CTGF
4	BCL2
5	MAPK10
6	CCND2
7	COL3A1
8	KIAA0233
9	BCL7A

Table S6. Lymphochip spots of known lymphoma genes.

SpotID	Gene Name
19384	MAPK10
24787	CCND2
15914	MAPK10
24429	BCL6
28472	MAPK10
19268	BCL6
16858	CCND2
17646	BCL2
16789	BCL2
19361	COL3A1
26535	BCL6
28859	BCL2
24367	BCL2
17791	FN1
16016	FN1
16732	FN1
31398	FN1
19379	FN1
27499	KIAA0233
24415	BCL7A
29222	CTGF

180 spots, which are known to deal with lymphoma were tested to distinguish between ABC and GCB subtype by PAM analysis. Successful genes are given in descending order (gene set error rate:TR:10% VAL:15.38%; F:CV:14%)

Table S7. Combined classifier for lymphoma subtypes.

SpotID	Gene Name
24376	*Centerin
17496	MYBL1
28014	MYBL1
19326	IGHM
19254	MME
33991	FOXP1
19384	MAPK10
19375	FOXP1
16049	IGHM
26454	SH3BP5
22118	KIAA0864
24787	CCND2
24787	CCND2
28979	LMO2
15914	MAPK10
19346	SH3BP5
15864	MME
19238	LMO2
30263	ASB13
19291	MYBL1
19312	NEIL1
25036	FLJ12363
26385	MME
19227	LOC96597
22122	IRF4
16886	LRMP
24480	KIAA1039
27378	LRMP
27379	LRMP
24729	IRF4
27673	LRMP
19348	*Similar to
24429	BCL6
28472	MAPK10
26516	*Similar clone=417048
19268	BCL6
	@Homo sapiH08 (LOC152137)
	Sur_clone=232
32529	2321
17646	BCL2

The resulting gene list that distinguishes ABC and GCB if the PAM analysis is performed only on the 31 best spots merged with the well known lymphoma genes. Marked in grey are the 31 best spots from all twelve thousand spots compared. Remarkably, the two classical lymphoma marker genes MAPK10 and CCND2 reach a similar quality in distinguishing ABC and GCB as the best separating ones.

Table S8. Cell cycle gene set that best distinguishes ABC and GCB subgroup. The genes are annotated by their spot ID, ensembl gene-ID and their gene name. Additionally the cell cycle states are given. The latter parameter shows a strong signal in the early and late cell cycle states compared with all available cell cycle states in the data set.

SpotID	Ensembl ID	cell cycle state	Gene
24927	ENSG00000165810	85	BTNL9
33929	ENSG00000165810	85	BTNL9
26913	ENSG00000138764	72	CCNG2
24750	ENSG00000136244	80	IL6
32430	ENSG00000162783	56	IER5
24491	ENSG00000165810	85	BTNL9
30172	ENSG00000138764	72	CCNG2
24930	ENSG00000187837	69	HIST1H1C
24725	ENSG00000011007	59	TCEB3
24908	ENSG00000118515	83	SGK
30355	ENSG00000164330	84	EBF
32096	ENSG00000164330	84	EBF
31931	ENSG00000164543	18	STK17A
26081	ENSG00000180447	80	GAS1
19374	ENSG00000124762	21	CDKN1A
24969	ENSG00000164330	84	EBF
24647	ENSG00000164330	84	EBF
34708	ENSG00000118515	83	SGK
27774	ENSG00000134058	92	CDK7
26401	ENSG00000118515	83	SGK
26725	ENSG00000164330	84	EBF
28881	ENSG00000163918	52	RFC4
17786	ENSG00000102804	1	TSC22D1
24613	ENSG00000102804	1	TSC22D1
33901	ENSG00000100644	2	HIF1A
27538	ENSG00000171656	96	ETV5
27952	ENSG00000179583	76	CIITA
34557	ENSG00000052841	2	TTC17
30021	ENSG00000099953	95	MMP11
27704	ENSG00000164330	84	EBF
26992	ENSG00000102804	1	TSC22D1
26344	ENSG00000138764	72	CCNG2
24832	ENSG00000163918	52	RFC4
26080	ENSG00000163739	76	CXCL1
33329	ENSG00000179583	76	CIITA
17290	ENSG00000134058	92	CDK7
30922	ENSG00000185658	5	BRWD1
26162	ENSG00000135541	91	AHI1
34288	ENSG00000134884	48	NA
33646	ENSG00000185658	5	BRWD1
26951	ENSG00000102804	1	TSC22D1
24977	ENSG00000153936	92	HS2ST1
16661	ENSG00000123080	75	CDKN2C
25942	ENSG00000145050	49	ARMET
22163	ENSG00000169926	6	KLF13
17405	ENSG00000178573	30	MAF
27275	ENSG00000100644	2	HIF1A
30415	ENSG00000164330	84	EBF
34484	ENSG00000151150	50	ANK3
33221	ENSG00000065809	2	FAM107B
32218	ENSG00000179583	76	CIITA
29637	ENSG00000145632	99	PLK2PLK2
27939	ENSG00000179583	76	CIITA
27328	ENSG00000108984	44	MAP2K6

(*Continued*)

Table S8. (*Continued*)

SpotID	Ensembl ID	cell cycle state	Gene
28792	ENSG00000099326	53	ZNF42
30725	ENSG00000175455	65	CCDC14
16736	ENSG00000136244	80	IL6
30874	ENSG00000081320	77	STK17B
28707	ENSG00000123080	75	CDKN2C
33336	ENSG00000175455	65	CCDC14
15871	ENSG00000168310	7	IRF2
28640	ENSG00000100526	0	CDKN3
28748	ENSG00000136244	80	IL6
28430	ENSG00000168310	7	IRF2
26084	ENSG00000128590	38	DNAJB9
30859	ENSG00000117650	93	NEK2
28674	ENSG00000138061	66	CYP1B1
16127	ENSG00000138061	66	CYP1B1
24868	ENSG00000012963	52	C14orf130
30508	ENSG00000081320	77	STK17B
34108	ENSG00000169926	6	KLF13
16053	ENSG00000173757	83	STAT5B
16091	ENSG00000100526	0	CDKN3
33594	ENSG00000179583	76	CIITA
32924	ENSG00000185658	5	BRWD1
32766	ENSG00000135164	74	DMTF1
16597	ENSG00000109971	0	HSPA8

Table S9. The cell cycle genes, which were chosen to distinguish the ABC and the GCB group.

Ensembl gene ID	cell cycle state	Gene symbol
ENSG00000011007	59	TCEB3
ENSG00000012963	52	C14orf130
ENSG00000052841	2	TTC17
ENSG00000065809	2	FAM107B
ENSG00000081320	77	STK17B
ENSG00000099326	53	ZNF42
ENSG00000099953	95	MMP11
ENSG00000100526	0	CDKN3
ENSG00000100644	2	HIF1A
ENSG00000102804	1	TSC22D1
ENSG00000108984	44	MAP2K6
ENSG00000109971	0	HSPA8
ENSG00000117650	93	NEK2
ENSG00000118515	83	SGK
ENSG00000123080	75	CDKN2C
ENSG00000124762	21	CDKN1A
ENSG00000128590	38	DNAJB9
ENSG00000134058	92	CDK7
ENSG00000134884	48	NA
ENSG00000135164	74	DMTF1
ENSG00000135541	91	AHI1
ENSG00000136244	80	IL6
ENSG00000138061	66	CYP1B1
ENSG00000138764	72	CCNG2
ENSG00000145050	49	ARMET
ENSG00000145632	99	PLK2PLK2
ENSG00000151150	50	ANK3

(*Continued*)

Table S9. (Continued)

Ensembl gene ID	cell cycle state	Gene symbol
ENSG00000153936	92	HS2ST1
ENSG00000162783	56	IER5
ENSG00000163739	76	CXCL1
ENSG00000163918	52	RFC4
ENSG00000164330	84	EBF
ENSG00000164543	18	STK17A
ENSG00000165810	85	BTNL9
ENSG00000168310	7	IRF2
ENSG00000169926	6	KLF13
ENSG00000171656	96	ETV5
ENSG00000173757	83	STAT5B
ENSG00000175455	65	CCDC14
ENSG00000178573	30	MAF
ENSG00000179583	76	CIITA
ENSG00000180447	80	GAS1
ENSG00000185658	5	BRWD1
ENSG00000187837	69	HIST1H1C

The cell cycle genes annotated by their ensembl gene-ID and their gene name. Additionally the cell cycle states are annotated. The latter parameter shows a strong signal in the early and late cell cycle states compared with all available cell cycle states in the data set.

Table S10. Gene expression values of the main regulatory network distinguishing ABC and GCB.

Gene	ABC	GCB
ASB13	–	+
MYBL1	–	+
MME	–	+
MAPK10	–	+
LRMP	–	+
LMO2	–	+
FN1	–	+
CTGF	–	+
COL3A1	–	+
BCL6	–	+
BCL7A	–	+
NEIL1	–	+
SH3BP5	+	–
BCL2	+	–
CCND2	+	–
IRF4	+	–
IGHM	+	–
FOXP1	+	–

Genes from Figure 2 and their gene expression values in the subgroups ABC and GCB are shown. The symbol "–" indicates a lower gene expression than "+". In this network, more genes of the more aggressive ABC type have a lower gene expression than the GCB type.

Table S11. List of potential Notch target transcripts.

Gene ID	Transcript ID	Description
ENSG00000156136	ENST00000286648	Deoxycytidine kinase
ENSG00000148158	ENST00000277244	Sorting nexin family member 30
ENSG00000179388	ENST00000317216	Early growth response protein 3
ENSG00000198833	ENST00000361212	Ubiquitin-conjugating enzyme E2 J1
ENSG00000198833	ENST00000361333	Ubiquitin-conjugating enzyme E2 J1
ENSG00000065308	ENST00000182527	Translocation associated membrane protein 2
ENSG00000170584	ENST00000302764	NudC domain containing protein 2
ENSG00000074706	ENST00000265198	phosphoinositide-binding protein PIP3-E
ENSG00000134108	ENST00000256496	ADP-ribosylation factor-like 10C)

For all genes of the Lymphochip, all available transcripts annotated in ensembl were screened for the GY, Brd and K boxes. Only these transcripts bear all three boxes, GY, Brd and K in the 3'-UTRs. They are possible candidates to be regulated by the Notch signalling pathway. Moreover, the Deoxycytidine kinase (ENSG00000156136) and the Translocation associated membrane protein 2 (ENSG00000065308) show different gene expression values between the ABC and GCB subgroups.

Table S12. T-test result of network genes in another data set.

Genes	P-value	T-value
CCND2	6.260705e-06	5.56939706
BCL6	2.490035e-02	–2.34449786
BCL2	1.843571e-03	3.43618678
IRF4	2.082072e-07	6.49044833
LMO2	3.820841e-07	–6.66162303
MAPK10	3.888633e-02	–2.15403094

The genes from the proposed STRING-network in Figure 4 were used to apply a T-test between the ABC and the GCB group in the gene expression data of Shipp et al. The authors Wright et al. found some evidence for these DLBCL groups in there.
The most obvious rejection of the null hypothesis is delivered by IRF4, LMO2, CCND2, BCL2, BCL6 and MAPK10, which are also part of the predictor of Wright et al.

Part III

Concluding Discussion

Microarray experiments can aid in answering scientific questions in very diverse fields, some of them being cancer research and diagnostics, basic research in plant and animal physiology and biodiversity studies. In this work, applications of the microarray technology to fields named above are described. The development of a phylogenetic DNA microarray demonstrated the complete workflow from designing an array, experimental laboratory work to microarray data analysis. In other chapters of this work, the analysis of primary microarray data was performed. In two chapters, the analysis of whole genome expression data of *Arabidopsis thaliana* was described. The first experiment was set up to find out if microwave irradiation had an effect on gene expression on a plant cell culture. In the second experiment, physiological differences between normal and tumor plant cells were analyzed. Secondary data analysis was performed on a dataset of human Diffuse Large B Cell Lymphoma (DLBCL) and a meta-analysis on a large number of datasets from a public database was performed on *Arabidopsis thaliana* datasets. Furthermore, a new software is presented to improve microarray gene expression data analysis from a functional perspective ("GEPAT", chapter 5).

In the chapter about the **development of a phylogenetic microarray** (publication 1, (Engelmann *et al.*, 2008b)), the complete process from selecting appropriate species, selecting capture probes (array design) over hybridization of the microarrays to data analysis is described. On phylogenetic arrays, in contrast to gene expression microarrays, the sequence of one marker gene is spotted for many different species. Evaluating the signal intensities from the individual spots, predictions about which species had been in the sample can be made. DNA microarrays are suitable and cost-effective tools to measure biodiversity when a large number of measurements needs to be performed or when the same habitat needs to be measured regularly.

In chapter 1, I have presented the data analysis of a phylogenetic DNA array based on 12 microarray hybridizations, 10 single alga hybridizations and two mixtures of two algae each. Although we had performed more microarray hybridizations, they could not be integrated in the analysis presented in the paper because of several reasons. The first batch of arrays was spotted with probes made from universal primers located in the 5.8S and 28S rDNA which amplify the complete ITS2 sequence. These probes share about 50 bp of identical sequence at both ends of the probe leading to high levels of cross-hybridization. For the second batch of array hybridizations we used algae-specific primers to amplify shorter probe sequences with less sequence similarity. Cross-hybridization was much less on these arrays and therefore these data were used for the data analysis described in the publication. A third batch of 18 arrays was created with both probes made from universal and from specific primers. The idea was that integrating data from the two probes per alga may yield more robust results for the detection of species. Unfortunately, the spots of the probes with the sequences from universal primers were much brighter than the ones from the specific primers and outshined the spots from the specific probes. Although there was considerable high background, I still analyzed the data of the spots of the specific primers. With the

same approach presented in the publication, I was able to correctly predict for about 50 % of the microarray hybridizations which species were in the sample in a leave-two-out cross-validation. The third batch contained 8 hybridizations with mixtures of two or three algae. Similar to the results described in the publication, the two closest related alga *Scenedesmus acuminatus* and *Scenedesmus obliquus* were the most difficult to distinguish. Besides this, the predictions of the mixtures were very good, indicating that our species microarray is able to detect also complex mixtures of species. Unfortunately, it was not possible to integrate the microarray data from batches two and three into one coherent analysis, which would have increased the number of microarray hybridizations dramatically. This was most likely due to the high background level observed on the arrays of batch three.

One problem in the current analysis procedure is that the case "no species present" cannot be distinguished from the case "all species present in equal amounts". This is due to the variance stabilization preprocessing step where the signal intensities are scaled such that the mean signal intensity is about equal in all arrays. If all spot intensites are very low because no species is present, they are scaled up. This problem could be approached in several ways. One way could be to include a control probe on the array which will bind only to a spiked-in control sequence added to the sample in a known concentration. Then the signal intensities could be calibrated with the control signal intensity. For environmental applications there might be species which are virtually always present in the environment and could serve as controls, but with these "natural" controls, quantitative calibration is of course not possible. Another way would be to change the preprocessing. Once the experimental workflow is more robust, variance stabilization might not be necessary any more and calculations could directly be performed on the median signal intensities.

Compared to light microscopy or sequencing of a barcode region, DNA microarrays require some time and preliminary experiments to design and test the microarray. But once a particular microarray has been established, processing of many samples can be performed rapidly. Nonetheless, DNA microarrays for species detection also face competition from recently developed high throughput sequencing technologies. The advantage of these technologies is that in principle, every species can be identified, not just the ones which would be represented on a particular microarray. But because the initial investment for high throughput sequencing machines is still very high at the moment, it will take several years before they pose serious competition. Until then, phylogenetic DNA microarrays have to compete with barcoding approaches, which are based on classical Sanger sequencing. Their limitation is, however, that they only yield reliable sequence reads for one individual sequence. Mixed environmental samples can therefore only be analyzed when the sequences of the sample are individually cloned and then sequenced. This is more laborious and the sequences of some species might be missed in the cloning procedure. Also, quantification of species is impossible because the cloning step introduces bias.

While for gene expression microarray data, numerous public databases ex-

ist, databases for phylogenetic array data are still missing. This is probably in parts due to the fact that there are no established platforms and data formats for this kind of data yet. It would be desirable to establish such public data repositories and common standards on the design and analysis of phylogenetic arrays, to facilitate future DNA microarray experiments and allow the comparison of existing DNA array datasets.

As stated above, the situation is different for gene expression microarray data. The size of public gene expression array databases has increased rapidly in the past years, but still, the wealth of information stored there could be used more extensively. Most meta-analysis performed so far combined datasets which were set up to answer the same scientific question. They were mainly used to increase the number of replicates which again increases the number of statistically significant differentially expressed genes. **Exploratory meta-analysis** over a wide range of experimental conditions as described here (chapter 2, (Engelmann *et al.*, 2008c)), has the potential to discover novel functions of genes which would have been missed in the analysis of individual datasets. In a meta-analysis of *Arabidopsis thaliana*, for example, a function in pathogen defense could be assigned to a group of serine/threonine kinases with two uncharacterized domains in their receptor part (DUF26). Few web applications like Genevestigator (Zimmermann *et al.*, 2004) which allows browsing in the human, mouse, rat, *Arabidopsis*, and barley transcriptome under different conditions, have been set up to better exploit microarray databases. Also the number of scientifc questions they can answer are usually limited. A general problem in meta-analysis is to achieve comparability. The individual experiments deposited in the database were typically performed in different laboratories, maybe with different chemicals and slightly differing protocols. Despite a standard which defines what information needs to be given by the authors on a microarray experiment (MIAME, Minimal Information About a Microarray Experiment, (Brazma *et al.*, 2001)), standard protocols on how a microarray experiment should be performed are missing. This makes a comparison of datasets from different studies very difficult, especially when only processed data is supplied by the database.

Chances to get reasonable results from a meta-analysis are better when it can be based on unprocessed ("raw") data, e.g. CEL-files from Affymetrix arrays or the scanner outputs from other platforms. Therefore, databases which require unprocessed data to be deposited should be preferred. Still, the problem of normalizing a possibly very heterogenous dataset persist. Great care needs to be taken in the preprocessing steps and possible outlier hybridizations or even complete datasets need to be removed to achieve comparability. Once this more elaborate preprocessing has been done, a meta-analyis can yield additional insights into the function of genes and their regulation.

We have dealt with this issue with a very strict outlier removal, discarding 35 of 76 contrasts (pairwise comparisons of groups of similarly treated microarrays). Although the remaining data was considerably more homogenous than the complete data set, one can argue that too much information was lost when discarding almost 50% of the data. In this trade-off between homogeneity of

the data and using as much data as possible, we favored homogeneity to lower the risk that our analysis would be confused by experimental artifacts.

Another promising meta-analysis approach has been proposed by Hibbs *et al.* (2007). To better exploit the available microarray data of *Saccharomyces cerevisiae* experiments, Hibbs *et al.* (2007) have built a web application named SPELL where the regulation of a gene can be studied over about 2400 conditions. Given a small set of query genes, SPELL determines the most informative conditions for these genes and searches for genes with similar expression profiles in the selected conditions. By this co-expression analysis, hypotheses about gene functions can be proposed which can then be experimentally validated. It would be desirable to have similar methods also for other more complex organisms, although they might require more sophisticated approaches because of their more complex regulation mechanisms compared to the single-cell organism *Saccharomyces cerevisiae*.

Co-expression analysis might also be reasonable for *Arabidopsis thaliana* datasets. Although the regulation mechanisms are more complicated than for yeast, individual variation is smaller than for example between humans or mice. With principal components analysis, an unsupervised cluster method, a simple explorative meta-analysis was performed to find similarities between a microwave treated *Arabidopsis thaliana* gene expression dataset and 75 datasets from a public repository. The initial question in this analysis was to find out **if microwave irradiation had an effect on the physiology of plant cells** (chapter 3, (Engelmann *et al.*, 2008a)). Because the problem was formulated as a question, this was an excellent opportunity to apply explorative analysis methods. The scientific question was unusual: prove or disprove that there is an effect of a treatment. This asked for a non-standard analysis of the gene expression microarray data. Unsupervised clustering methods could show that differences exist between the two sample groups and small changes in gene expression could be confirmed by supervised differential gene expression analysis and quantitative real-time PCR. But are these small changes physiologically relevant to the plant cells? Some of the differentially expressed genes have functions in photosynthesis and a comparison of the microwave dataset to publicly available datasets indicated similarities to datasets which analyzed the effect of different light treatments. From these findings, one might set up the hypothesis that plants perceive electromagnetic irradiation as some kind of energy which might even have an effect on photosynthesis. However, the time of irradiation was only 24 h in the experiment described here, to get a more realistic picture of the influence of UMTS irradiation on plants, cell cultures and whole plants should be treated for weeks or even month. Additionally, it would be interesting to study whether microwaves have an effect on the genomic level. If there was an effect on defined regions on the genome, this could be studies with SNP arrays in resequencing studies, but if the irradiation would induce genomic changes at different positions in different cells, this effect would be very difficult to trace. In any case, further experiments are needed to clarify if the small changes in gene expression observed here on a cell culture could have an effect on the physiology of whole plants.

Infection with *Agrobacterium tumefaciens*, on the other hand, has a very dramatic effect on the gene expression of plant cells. The bacteria induce changes in gene expression which lead to rapid tumor growth and the production of nutrients used by the bacteria. With the same type of microarray as used in the microwave irradiation experiment (Affymetrix ATH-1 whole genome arrays), the **gene expression and metabolite profiles of *Arabidopsis thaliana* tumors** were analyzed (chapter 4, (Deeken *et al.*, 2006)). With a functional analysis of differentially expressed genes, a shift from aerobic to anerobic energy production necessary for rapid cell growth was observed. The tumor cells reduce photosynthetic energy production which is reflected by the repression of genes involved in photosynthesis and mitochondrial electron transport. These transcriptional changes were accompanied by increased levels of anions, sugars and amino acids. In this analysis, gene expression data was complemented by solute measurements, demonstrating that the integration of different data types has great potential to gain a better understanding of what is going on in a cell or tissue. Future studies could focus on the development of a plant tumor with for example a time-course experiment of samples taken very early, early, after a medium period of time and late after infection to analyze the transcriptional changes during the reorganisation from regular to tumor cells. If early events that finally lead to tumor growth could be identified, these could also help in developing crop plants which are resistant to *Agrobacterium tumefaciens* infection.

Plant tumors show some analogies to animal and human cancers and therefore plant tumor experiments could also yield valuable insights into cancer biology. But in this thesis, also analyses of human cancer patient data were demonstrated. Besides elucidating the molecular changes which lead to cancer development, gene expression data generated with microarrays can aid in the classification of patients into subgroups of the disease or into risk groups with different predicted survival times. These diagnostic applications are clinically very important because they can first help in developing and secondly help in selecting the appropriate therapy for each individual patient. For **Mantle Cell Lymphoma (MCL)** patients, a novel seven gene predictor could be discovered to predict patient survival (chapter 6, (Blenk *et al.*, 2008)). It performs similarly well to former predictors but uses less genes, making it easier to apply in the clinic with low throughput techniques like real-time PCR. Explorative analysis of Comparative Genomic Hybridization (CGH) data of the same patients showed that patients could also be grouped into groups of longer or shorter survival based on this data type. However, for clinical applications, measuring the gene expression levels of the predictor genes will be easier than the analysis of chromosomal aberrations. Integrating the gene expression data with interaction data from STRING (von Mering *et al.*, 2007), an interaction network of proliferation markers and cell cycle genes showed that more aggressive MCL increase the expression of late cell cycle genes.

For **Diffuse Large B Cell Lymphoma (DLBCL)** patients, the two subgroups ABC (Activated B Cell-like) and GCB (Germinal Center B cell-like) described previously could be confirmed by an unsupervised analysis (chap-

ter 7, (Blenk *et al.*, 2007)). Furthermore, a survival predictor with only six spots was derived which yields better predictions for the clinically important patient group with medium survival. A regulatory network with proliferation, anti-proliferation, apoptosis and differentiation genes pointed to differences between the subgroups ABC and GCB which could explain why ABC lymphomas are more aggressive.

All of the projects presented in this work demonstrated that careful selection and application of analysis methods is needed to deduce biological knowledge from high throughput data. As presented in chapter 5, **GEPAT** (Weniger *et al.*, 2007) offers statistical analysis methods and graphical representations for gene expression microarray and comparative genomic hybridization data and connects to several biological databases which allow integration of chromosomal localization, functional annotation and interaction data.

Great potential also lies in the integration and combination of different high throughput data types to get a more holistic picture about differences between two or more sample types. Chromosomal rearrangements are known to play a crucial role in cancer develoment and have an effect on gene expression, therefore, integrating gene expression with CGH data can help to get further insights into cancer biology (Bussey *et al.*, 2006; Nigro *et al.*, 2005). GEPAT currently allows to graphically overlay gene expression and CGH data to find genomic regions of interest, but this approach is of course absolutely model-free. To facilitate and improve this integration, methods need to be developed which are not only valid for a single study, but which are standardized such that they can be applied to all gene expression plus CGH data sets. One possible approach could be to use a hidden Markov model that has all combinations of increased-normal-decreased levels of gene expression and CGH data as states. With this approach, regions where gene expression and CGH data match and others where they disagree could be predicted.

To get further insights into the regulation of gene expression, again with applications to cancer, it is worthwile to integrate microarray gene expression data with transcription factor binding site information. Although many transcription factors are known, their motifs are often very short or degenerated and therefore sophisticated computational methods are needed to spot them in the genome if they are based on the sequence alone (D'haeseleer, 2006; Stormo, 2000). Jeffery *et al.* (2007) presented a more promising approach using unsupervised analysis to identify transcription factor binding motifs associated to gene expression differences between two sample groups. Known transcription factor binding sites could also easily be displayed in GEPAT.

Furthermore, other relatively new high throughput technologies could be combined to yield additional insights. Single nucleotide polymorphism (SNP) data from linkage and association studies could also be combined with gene expression data to analyze the effect of SNPs on gene expression levels when they lie within a coding region but also the effect of SNPs in non-coding, possibly regulatory regions. Small non-coding RNAs (miRNA, siRNA, piRNA) also modulate gene expression and much research and new methods are needed to first discover them in the genome and then fully characterize their functions

(Mattick and Makunin, 2006; Berezikov *et al.*, 2006). miRNA annotation data has also been included in GEPAT. Tiling arrays are used to find non-coding RNAs in the genome. In combination with gene expression data, tiling arrays could be used to find and characterize non-coding RNAs which modulate gene expression in for example different tissues or in cancer. Additional modules to analyze and integrate all the different data types mentioned above could be integrated into GEPAT in the future.

Much research is still required to understand the modulators of gene expression. Besides genomic aberrations, epigenetic changes also play a role in the regulation of transcription. DNA methylation and histone modifications alter the packaging of the DNA which in turn has an effect on gene expression. In general, methylation and histone modifications lead to a tighter packaging of the DNA which makes the genes in these regions less accessible for the transcription machinery (Bock and Lengauer, 2008). It has been shown that methylation patterns change during the life of an individual (Lewin, 2000) and differ between tissues, thereby influencing gene expression patterns (Song *et al.*, 2005). Differential DNA methylation has also been found in human cancers (Weber *et al.*, 2005). In prostate cancer, hypermethylation of CpG islands might be the earliest somatic genome changes which eventually lead to unrestricted cell growth (Yegnasubramanian *et al.*, 2004). ChIP-on-chip experiments using tiling arrays can be used to analyze DNA methylation patterns on a large scale. In these experiments, antibodies are used to pull down methylated regions which are then hybridized to a tiling array and compared to a control sample (Lippman *et al.*, 2005). But because this approach is still rather new, further improvements to the experimental steps and methods for data analysis are needed. The bioinformatic challenge is here to derive a ranked list of overrepresented (methylated) regions of the sample. Methods similar to the ones used in gene expression analysis are applied to the data and have been modified to answer specific needs (see Bock and Lengauer (2008) for a review of available methods). In cancer epigenetics, the challenge is to detect common patterns or functional relationships of specific regions to cancer. This task as well as developing methods to improve diagnosis and therapy can be approached with the modulation of existing bioinformatic methods for the analysis of high throughput data.

Allele-specific transcription is another interesting topic to be addressed in the future. Although today it is known that for most genes, both maternal and paternal alleles are more or less active, for almost 10% of genes, only one allele from the mother or the father is active (Gimelbrant *et al.*, 2007). Before Gimelbrant's work was published, only some immuno globulin genes and genes coding for olfactory receptors had been known to be expressed monoallelicly. More than 1000 genes in the human genome could be monoallelicly expressed and amongst others, they might explain differences in disease susceptibility of monozygotic twins. Dependent on which copy of a gene is expressed, one twin might have a higher chance to develop a disease than the other. This example shows that identical genome information, as in monoallelic twins, does not make identical organisms. Transcription of genes and their regulation play a

crucial part in the characteristics of an individual.

With the advent of massively parallel high throughput sequencing techniques, for the first time, microarrays face serious competition in whole genome expression analysis. Torres *et al.* (2008) report on using the 454 sequencing technology (Margulies *et al.*, 2005) to measure the gene expression profile of *Drosophila melanogaster*. An advantage of the sequencing approach is that also new isoforms and antisense transcripts can be detected and that a quantitative analysis of the transcripts is in principle possible. Even allele-specific gene expression measurements might be feasible, because SNPs in transcripts can be detected by sequencing. While interspecies gene expression studies are difficult to conduct with microarrays because the gene expression values are always relative and usually a reference suitable for several species can only be set up for very closely related species (Oshlack *et al.*, 2007), they are more feasible with the sequencing approach. If a database of high quality is available, the sequence reads can also be mapped to a interspecies database.

Nonetheless, microarrays will remain an invaluable tool for the analysis of whole genome transcription changes, chromosomal rearrangements, single nucleotide polymorphisms and modulators of gene expression. The challenge for the future will be to deduce novel biological knowledge by integrating different data types to get a more systemic picture of the organism under study.

Summary

Microarrays have been used in diverse research fields to answer many biological and medical questions. One important application of DNA microarrays is whole genome gene expression analysis. Because of the large number of genes interrogated in one experiment, statistical analysis methods are needed to handle these huge datasets and receive meaningful and interpretable results. In this thesis, the development of a phylogenetic DNA microarray, the analysis of several gene expression microarray datasets and new approaches for improved data analysis and interpretation are described.

In the **first publication**, the development and analysis of a phylogenetic microarray is presented. I could show that species detection with phylogenetic DNA microarrays can be significantly improved when the microarray data is analyzed with a linear regression modeling approach. Standard methods have so far relied on pure signal intensities of the array spots and a simple cut-off criterion was applied to call a species present or absent. This procedure is not applicable to very closely related species with high sequence similarity because cross-hybridization of non-target DNA renders species detection impossible based on signal intensities alone. By modeling hybridization and cross-hybridization with linear regression, as I have presented in this thesis, even species with a sequence similarity of 97% in the marker gene can be detected and distinguished from related species. Another advantage of the modeling approach over existing methods is that the model also performs well on mixtures of different species. In principle, also quantitative predictions can be made.

To make better use of the large amounts of microarray data stored in public databases, meta-analysis approaches need to be developed. In the **second publication**, an explorative meta-analysis exemplified on *Arabidopsis thaliana* gene expression datasets is presented. Integrating datasets studying effects such as the influence of plant hormones, pathogens and different mutations on gene expression levels, clusters of similarly treated datasets could be found. From the clusters of pathogen-treated and indole-3-acetic acid (IAA) treated datasets, representative genes were selected which pointed to functions which had been associated with pathogen attack or IAA effects previously. Additionally, hypotheses about the functions of so far uncharacterized genes could be set up. Thus, this kind of meta-analysis could be used to propose gene functions and their regulation under different conditions which could then be experimentally validated.

In this work, also primary data analysis of *Arabidopsis thaliana* datasets

is presented. In the **third publication**, an experiment which was conducted to find out if microwave irradiation has an effect on the gene expression of a plant cell culture is described. During the first steps, the data analysis was carried out blinded and exploratory analysis methods were applied to find out if the irradiation had an effect on gene expression of plant cells. Small but statistically significant changes in a few genes were found and could be experimentally confirmed. From the functions of the regulated genes and a meta-analysis with publicly available microarray data, it could be suspected that the plant cell culture somehow perceived the irradiation as energy, similar to perceiving light rays. However, further experiments are needed to analyze whether microwave irradiation has an effect on whole plants.

The **fourth publication** describes the functional analysis of another *Arabidopsis thaliana* gene expression dataset. The gene expression data of the plant tumor dataset pointed to a switch from a mainly aerobic, auxotrophic to an anaerobic and heterotrophic metabolism in the plant tumor. Genes involved in photosynthesis were found to be repressed in tumors; genes of amino acid and lipid metabolism, cell wall and solute transporters were regulated in a way that sustains tumor growth and development.

Furthermore, in the **fifth publication**, GEPAT (Genome Expression Pathway Analysis Tool), a tool for the analysis and integration of microarray data with other data types, is described. It consists of a web application and database which allows comfortable data upload and data analysis. GEPAT also links to biological databases which help in interpreting the results by supplying functional annotation, metabolic or regulatory interaction network membership and chromosomal localization.

In later chapters of this thesis (**publication 6** and **publication 7**), GEPAT is used to analyze human microarray datasets and to integrate results from gene expression analysis with other datatypes. Gene expression and comparative genomic hybridization data from 71 Mantle Cell Lymphoma (MCL) patients was analyzed and allowed proposing a seven gene predictor which facilitates survival predictions for patients compared to existing predictors. In this study, it was also shown that CGH data can be used for survival predictions. For the dataset of Diffuse Large B-cell lymphoma (DLBCL) patients, an improved six-spot survival predictor could also be found based on the gene expression data. From the genes differentially expressed between long and short surviving MCL patients as well as for regulated genes of DLBCL patients, interaction networks could be set up. They point to differences in regulation for cell cycle and proliferation genes between patients with good and bad prognosis for both cancer types.

The results of the different projects described in this thesis have shown that great potential lies in the analysis of microarray data. Novel methods for data analysis open up new perspectives and enable the researcher to draw meaningful conclusions. Challenges for the future lie in finding and characterizing modulators of gene expression and in the integration of different high throughput data types.

Zusammenfassung

Microarrays werden in zahlreichen Forschungsbereichen eingesetzt, um biologische und medizinische Fragestellungen zu beantworten. Eine wichtige Anwendung der DNA-Microarray-Technologie ist die genomweite Analyse der Genexpression. Da die Anzahl der Gene, die in einem Microarray-Experiment untersucht werden, sehr groß ist, werden statistische Auswerteverfahren benötigt um diese Datensätze zu verarbeiten und aussagekräftige und interpretierbare Ergebnisse zu erhalten. In der vorliegenden Dissertation wird die Entwicklung eines phylogenetischen DNA Microarrays, die Analyse von mehreren Microarray-Genexpressionsdatensätzen und neue Ansätze für die Datenanalyse und Interpretation der Ergebnisse vorgestellt.

Die Entwicklung und Analyse der Daten eines phylogenetischen DNA Microarrays wird in der **ersten Publikation** dargestellt. Ich konnte zeigen, dass die Spezies-Detektion mit phylogenetischen Microarrays durch die Datenanalyse mit einem linearen Regressionsansatz signifikant verbessert werden kann. Standard-Methoden haben bislang nur die Signalintensitäten der einzelnen Microarray-Messpunkte betrachtet und eine Spezies als an- oder abwesend bezeichnet, wenn die Signalintensität ihres Messpunktes oberhalb eines willkürlich gesetzten Schwellenwertes lag. Dieses Verfahren ist allerdings aufgrund von Kreuz-Hybridisierungen nicht auf sehr nah verwandte Spezies mit hoher Sequenzidentität anwendbar. Durch die Modellierung des Hybridisierungs- und Kreuz-Hybridisierungsverhaltens mit einem linearen Regressionsmodell konnte ich zeigen, dass Spezies mit einer Sequenzähnlichkeit von 97% im Markergen immer noch unterschieden werden können und ihre Anwesenheit richtig vorhergesagt werden kann. Ein weiterer Vorteil der Modellierung gegenüber herkömmlichen Methoden ist, dass auch Mischungen verschiedener Spezies zuverlässig vorhergesagt werden können. Theoretisch sind auch quantitative Vorhersagen mit diesem Modell möglich.

Um die großen Datenmengen, die in öffentlichen Microarray-Datenbanken abgelegt sind, besser nutzen zu können, bieten sich Meta-Analysen an. In der **zweiten Publikation** wird eine explorative Meta-Analyse auf *Arabidopsis thaliana*-Datensätzen vorgestellt. Mit einer gemeinsamen explorativen Analyse verschiedener Datensätze, die den Einfluss von Pflanzenhormonen, Pathogenen oder verschiedenen Mutationen auf die Genexpression untersucht haben, konnten die Datensätze anhand ihrer Genexpressionsprofile in drei große Gruppen eingeordnet werden: Experimente mit Indol-3-Essigsäure (IAA), mit Pathogenen und andere Experimente. Gene, die charakteristisch für die Gruppe der IAA-Datensätze beziehungsweise für die Gruppe der Pathogen-Datensätze

sind, wurden näher betrachtet. Diese Gene hatten Funktionen, die bereits mit Pathogenbefall bzw. dem Einfluss von IAA in Verbindung gebracht wurden. Außerdem wurden Hypothesen über die Funktionen von bislang nicht annotierten Genen aufgestellt. Daher könnte die hier vorgestellte Meta-Analyse generell dazu dienen, Genfunktionen und ihre Regulation unter verschiedenen Bedingungen vorherzusagen. Diese sollten anschließend experimentell bestätigt werden.

In dieser Arbeit werden auch Primäranalysen von einzelnen *Arabidopsis thaliana* Genexpressions-Datensätzen vorgestellt. In der **dritten Publikation** wird ein Experiment beschrieben, das durchgeführt wurde um herauszufinden, ob Mikrowellen-Strahlung einen Einfluss auf die Genexpression einer Zellkultur hat. Die ersten Schritte der Datenanalyse dieses Datensatzes wurden doppeltblind durchgeführt und explorative Analysemethoden wurden angewendet um herauszufinden, ob die Strahlung einen Effekt auf die Genexpression der pflanzlichen Zellkultur hat. Es wurden geringe aber signifikante Veränderungen in einer sehr kleinen Anzahl von Genen beobachtet, die experimentell bestätigt werden konnten. Die Funktionen der regulierten Gene und eine Meta-Analyse mit öffentlich zugänglichen Datensätzen einer Datenbank deuten darauf hin, dass die pflanzliche Zellkultur die Strahlung als eine Art Energiequelle ähnlich dem Licht wahrnimmt. Allerdings sind weitere Experimente notwendig um dies zu bestätigen und den Einfluss von Mikrowellen-Strahlung auf komplette Pflanzen zu untersuchen.

Des weiteren wird in der **vierten Publikation** die funktionelle Analyse eines *Arabidopsis thaliana* Genexpressionsdatensatzes beschrieben. Die Analyse der Genexpressionsdaten eines pflanzlichen Tumores zeigte, dass der pflanzliche Tumor seinen Stoffwechsel von aerob und auxotroph auf anaerob und heterotroph umstellt. Gene der Photosynthese werden im Tumorgewebe reprimiert, Gene des Aminosäure- und Fettstoffwechsels, der Zellwand und Transportkanäle werden so reguliert, dass Wachstum und Entwicklung des Tumors gefördert werden.

In der **fünften Publikation** in dieser Arbeit wird GEPAT (Genome Expression Pathway Analysis Tool) beschrieben. Es besteht aus einer Internet-Anwendung und einer Datenbank, die das einfache Hochladen von Datensätzen in die Datenbank und viele Möglichkeiten der Datenanalyse und die Integration anderer Datentypen erlaubt. GEPAT ist außerdem mit biologischen Datenbanken verlinkt, die dadurch, dass sie funktionelle Annotationen, Zugehörigkeiten zu metabolischen und regulatorischen Netzwerken und die chromosomale Position der Gene anbieten, bei der Interpretation der Ergebnisse hilfreich sind.

In den folgenden zwei Publikationen (**Publikation 6** und **Publikation 7**) wird GEPAT auf humane Microarray-Datensätze angewendet um Genexpressionsdaten mit weiteren Datentypen zu verknüpfen. Genexpressionsdaten und Daten aus vergleichender Genom-Hybridisierung (CGH) von primären Tumoren von 71 Mantel-Zell-Lymphom (MCL) Patienten ermöglichte die Ermittlung eines Prädiktors aus sieben Genen, der die Vorhersage der Überlebensdauer von Patienten gegenüber herkömmlichen Methoden verbessert. Die Analyse der CGH Daten zeigte außerdem, dass auch dieser Datentyp für die Vorher-

sage der Überlebensdauer geeignet ist. Für den Datensatz von Patienten mit großzellig diffusem B-Zell-Lymphom DLBCL konnte aus den Genexpressionsdaten ebenfalls ein neuer Prädiktor vorgeschlagen werden, der aus sechs Genen besteht. Mit den zwischen lang und kurz überlebenden Patienten differentiell exprimierten Genen der MCL Patienten und mit den Genen, die zwischen den beiden Untergruppen von DLBCL reguliert sind, wurden jeweils Interaktionsnetzwerke gebildet. Diese zeigen, dass bei beiden Krebstypen (MCL und DLBCL) Gene des Zellzyklus und der Proliferation zwischen Patienten mit kurzer und langer Überlebensdauer unterschiedlich reguliert sind.

Die Ergebnisse der in dieser Arbeit vorgestellten Projekte zeigen, dass in der Analyse von Genexpressionsdaten großes Potential steckt. Neue Analysemethoden können neue Einblicke ermöglichen und erlauben den Wissenschaftlern, aussagekräftige Schlussfolgerungen zu ziehen. Die Herausforderungen der Zukunft liegen darin, Regulatoren der Genexpression zu finden und zu charakterisieren, sowie die Daten verschiedener Hochdurchsatz-Technologien miteinander zu verflechten.

Bibliography

Avarre JC, de Lajudie P, and Béna G (2007): Hybridization of genomic DNA to microarrays: a challenge for the analysis of environmental samples. *J Microbiol Methods*, **69** (2): 242–248.

Barrett JC and Kawasaki ES (2003): Microarrays: the use of oligonucleotides and cDNA for the analysis of gene expression. *Drug Discov Today*, **8** (3): 134–141.

Berezikov E, Cuppen E, and Plasterk RHA (2006): Approaches to microRNA discovery. *Nat Genet*, **38 Suppl**: S2–S7.

Blenk S, Engelmann JC, Pinkert S, Weniger M, Schultz J, Rosenwald A, Müller-Hermelink HK, *et al.* (2008): Explorative data analysis of MCL reveals gene expression networks implicated in survival and prognosis supported by explorative CGH analysis. *BMC Cancer*, in press.

Blenk S, Engelmann JC, Weniger M, Schultz J, Dittrich M, Rosenwald A, Müller-Hermelink HK, *et al.* (2007): Germinal center B cell-like (GCB) and activated B cell-like (ABC) type of diffuse large B cell lymphoma (DLBCL): Analysis of molecular predictors, signatures, cell cycle state and patient survival. *Cancer Informatics*, **3**: 409–430.

Bock C and Lengauer T (2008): Computational epigenetics. *Bioinformatics*, **24** (1): 1–10.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, *et al.* (2001): Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet*, **29** (4): 365–371.

Buck MJ and Lieb JD (2004): ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, **83** (3): 349–360.

Bullinger L, Rücker FG, Kurz S, Du J, Scholl C, Sander S, Corbacioglu A, *et al.* (2007): Gene-expression profiling identifies distinct subclasses of core binding factor acute myeloid leukemia. *Blood*, **110** (4): 1291–1300.

Bussey KJ, Chin K, Lababidi S, Reimers M, Reinhold WC, Kuo WL, Gwadry F, *et al.* (2006): Integrating data on DNA copy number with gene expression

levels and drug sensitivities in the NCI-60 cell line panel. *Mol Cancer Ther*, **5** (4): 853–867.

Campbell NA (1997): *Biologie*. Spektrum Akademischer Verlag.

Chang F, Steelman LS, Shelton JG, Lee JT, Navolanic PM, Blalock WL, Franklin R, *et al.* (2003): Regulation of cell cycle progression and apoptosis by the Ras/Raf/MEK/ERK pathway (review). *Int J Oncol*, **22** (3): 469–480.

Chiang MK and Melton DA (2003): Single-cell transcript analysis of pancreas development. *Dev Cell*, **4** (3): 383–393.

Clarke PA, te Poele R, Wooster R, and Workman P (2001): Gene expression microarray analysis in cancer biology, pharmacology, and drug development: progress and potential. *Biochem Pharmacol*, **62** (10): 1311–1336.

Cui X and Churchill GA (2003): Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*, **4** (4): 210.

Deeken R, Engelmann JC, Efetova M, Czirjak T, Müller T, Kaiser WM, Tietz O, *et al.* (2006): An integrated view of gene expression and solute profiles of *Arabidopsis* tumors: A genome-wide approach. *Plant Cell*, **18** (12): 3617–3634.

D'haeseleer P (2006): How does DNA sequence motif discovery work? *Nat Biotechnol*, **24** (8): 959–961.

Dudoit S, Shaffer JP, and Boldrick JC (2003): Multiple hypothesis testing in microarray experiments. *Statistical Science*, **18**: 71–103.

Engelmann JC, Deeken R, Müller T, Nimtz G, Roelfsema RG, and Hedrich R (2008a): Is gene activity in plant cells affected by UMTS-irradiation? A whole genome approach. *Computational Biology and Chemistry: Advances and Applications*, submitted.

Engelmann JC, Rahmann S, Wolf M, Schultz J, Fritzilas E, Kneitz S, Dandekar T, *et al.* (2008b): Modeling cross-hybridization on phylogenetic DNA microarrays increases the detection power of closely related species. *Molecular Ecology Resources*, accepted.

Engelmann JC, Schwarz R, Blenk S, Friedrich T, Seibel P, Dandekar T, and Müller T (2008c): Large-scale kernel-based explorative meta-analysis of Affymetrix genome arrays. *Bioinformatics and Biology Insights*, accepted.

Fellenberg K, Hauser NC, Brors B, Neutzner A, Hoheisel JD, and Vingron M (2001): Correspondence analysis applied to microarray data. *Proc Natl Acad Sci U S A*, **98** (19): 10781–10786.

Gene Ontology Consortium (2001): Creating the gene ontology resource: design and implementation. *Genome Res*, **11** (8): 1425–1433.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, *et al.* (2004): Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol*, **5** (10): R80.

Gimelbrant A, Hutchinson JN, Thompson BR, and Chess A (2007): Widespread monoallelic expression on human autosomes. *Science*, **318**: 1136–1140.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, *et al.* (1999): Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286** (5439): 531–537.

Hajibabaei M, Singer GAC, Hebert PDN, and Hickey DA (2007): DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends in Genetics*, **23** (4): 167–172.

Hartmann CH and Klein CA (2006): Gene expression profiling of single cells on large-scale oligonucleotide arrays. *Nucleic Acids Res*, **34** (21): e143.

Hatfield GW, Hung SP, and Baldi P (2003): Differential analysis of DNA microarray gene expression data. *Mol Microbiol*, **47** (4): 871–877.

Hibbs MA, Hess DC, Myers CL, Huttenhower C, Li K, and Troyanskaya OG (2007): Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics*, **23** (20): 2692–2699.

Hubbell E, Liu WM, and Mei R (2004): Supplemental data: robust estimators for expression analysis. Tech. rep., Affymetrix.

Huber W, von Heydebreck A, Sültmann H, Poustka A, and Vingron M (2002): Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18 Suppl 1**: S96–104.

Huber W, Irizarry RA, and Gentleman R (2005): Preprocessing overview. *In: Bioinformatics and computational biology solutions using R and Bioconductor* (edited by R Gentleman, VJ Carey, W Huber, RA Irizarry, and S Dudoit), p. 11, Springer.

Iqbal J, Neppalli VT, Wright G, Dave BJ, Horsman DE, Rosenwald A, Lynch J, *et al.* (2006): BCL2 expression is a prognostic marker for the activated B-cell-like type of diffuse large B-cell lymphoma. *J Clin Oncol*, **24** (6): 961–968.

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, and Speed TP (2003): Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4** (2): 249–264.

Irizarry RA, Wu Z, and Jaffee HA (2006): Comparison of Affymetrix GeneChip expression measures. *Bioinformatics*, **22** (7): 789–794.

Jeffery IB, Madden SF, McGettigan PA, Perriére G, Culhane AC, and Higgins DG (2007): Integrating transcription factor binding site information with gene expression datasets. *Bioinformatics*, **23** (3): 298–305.

Johnson JM, Edwards S, Shoemaker D, and Schadt EE (2005): Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends Genet*, **21** (2): 93–102.

Jolliffe IT (1986): *Principal Component Analysis*. Springer, New York.

Kamme F, Salunga R, Yu J, Tran DT, Zhu J, Luo L, Bittner A, *et al.* (2003): Single-cell microarray analysis in hippocampus CA1: demonstration and validation of cellular heterogeneity. *J Neurosci*, **23** (9): 3607–3615.

Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, *et al.* (2006): From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*, **34** (Database issue): D354–D357.

Lewin B (2000): *Genes VII*. Oxford University Press Inc, New York.

Lichter P, Joos S, Bentz M, and Lampel S (2000): Comparative genomic hybridization: uses and limitations. *Semin Hematol*, **37** (4): 348–357.

Lippman Z, Gendrel AV, Colot V, and Martienssen R (2005): Profiling DNA methylation patterns using genomic tiling microarrays. *Nat Methods*, **2** (3): 219–224.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, *et al.* (2005): Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437** (7057): 376–380.

Marton MJ, DeRisi JL, Bennett HA, Iyer VR, Meyer MR, Roberts CJ, Stoughton R, *et al.* (1998): Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat Med*, **4** (11): 1293–1301.

Mattick JS and Makunin IV (2006): Non-coding RNA. *Hum Mol Genet*, **15 Spec No 1**: R17–R29.

von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Krüger B, Snel B, *et al.* (2007): STRING 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res*, **35** (Database issue): D358–D362.

Mockler TC, Chan S, Sundaresan A, Chen H, Jacobsen SE, and Ecker JR (2005): Applications of DNA tiling arrays for whole-genome analysis. *Genomics*, **85** (1): 1–15.

Moritz C and Cicero C (2004): DNA barcoding: promise and pitfalls. *PLoS Biol*, **2** (10): e354.

Nagata M, Fujita H, Ida H, Hoshina H, Inoue T, Seki Y, Ohnishi M, *et al.* (2003): Identification of potential biomarkers of lymph node metastasis in oral squamous cell carcinoma by cDNA microarray analysis. *Int J Cancer*, **106** (5): 683–689.

Nigro JM, Misra A, Zhang L, Smirnov I, Colman H, Griffin C, Ozburn N, *et al.* (2005): Integrated array-comparative genomic hybridization and expression array profiles identify clinically relevant molecular subtypes of glioblastoma. *Cancer Res*, **65** (5): 1678–1686.

Oren M (2003): Decision making by p53: life, death and cancer. *Cell Death Differ*, **10** (4): 431–442.

Oshlack A, Chabot AE, Smyth GK, and Gilad Y (2007): Using DNA microarrays to study gene expression in closely related species. *Bioinformatics*, **23** (10): 1235–1242.

Quackenbush J (2001): Computational analysis of microarray data. *Nat Rev Genet*, **2** (6): 418–427.

R Development Core Team (2007): *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Scholtens D and von Heydebreck A (2005): Analysis of differential gene expression studies. *In: Bioinformatics and computational biology solutions using R and Bioconductor* (edited by R Gentleman, V Carey, W Huber, R Irizarry, and S Dudoit), Springer.

Smyth GK (2004): Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, **3 (1)**: Article 3.

Smyth GK, Michaud J, and Scott HS (2005): Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, **21** (9): 2067–2075.

Smyth GK and Speed TP (2003): Normalization of cDNA microarray data. *Methods*, **31**: 4.

Smyth GK, Yang YH, and Speed TP (2003): Statistical issues in cDNA microarray data analysis. *Methods in Molecular Biology*, **224**: 111–136.

Song F, Smith JF, Kimura MT, Morrow AD, Matsuyama T, Nagase H, and Held WA (2005): Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression. *Proc Natl Acad Sci U S A*, **102** (9): 3336–3341.

Stormo GD (2000): DNA binding sites: representation and discovery. *Bioinformatics*, **16** (1): 16–23.

Tibshirani R, Hastie T, Narasimhan B, and Chu G (2002): Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A*, **99** (10): 6567–6572.

Tietjen I, Rihel JM, Cao Y, Koentges G, Zakhary L, and Dulac C (2003): Single-cell transcriptional analysis of neuronal progenitors. *Neuron*, **38** (2): 161–175.

Torres TT, Metta M, Ottenwälder B, and Schlötterer C (2008): Gene expression profiling by massively parallel sequencing. *Genome Res*, **18** (1): 172–177.

Tusher VG, Tibshirani R, and Chu G (2001): Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, **98** (9): 5116–5121.

Usadel B, Nagel A, Thimm O, Redestig H, Blaesing OE, Palacios-Rojas N, Selbig J, *et al.* (2005): Extension of the visualization tool MapMan to allow statistical analysis of arrays, display of corresponding genes, and comparison with known responses. *Plant Physiol*, **138** (3): 1195–1204.

Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, and Schübeler D (2005): Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat Genet*, **37** (8): 853–862.

Weniger M, Engelmann JC, and Schultz J (2007): Genome Expression Pathway Analysis Tool–analysis and visualization of microarray gene expression data under genomic, proteomic and metabolic context. *BMC Bioinformatics*, **8**: 179.

Wright G, Tan B, Rosenwald A, Hurt EH, Wiestner A, and Staudt LM (2003): A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. *Proc Natl Acad Sci U S A*, **100** (17): 9991–9996.

Wu Z, Irizarry R, Gentleman R, Martinez-Murillo F, and Spencer F (2004): A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association*, **99**: 909–917.

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, and Speed TP (2002): Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, **30** (4): e15.

Yang YH and Paquet AC (2005): Normalization for cDNA microarray data. *In: Bioinformatics and computational biology solutions using R and Bioconductor* (edited by R Gentleman, VJ Carey, W Huber, RA Irizarry, and S Dudoit), pp. 60–67, Springer.

Yegnasubramanian S, Kowalski J, Gonzalgo ML, Zahurak M, Piantadosi S, Walsh PC, Bova GS, *et al.* (2004): Hypermethylation of CpG islands in primary and metastatic human prostate cancer. *Cancer Res*, **64** (6): 1975–1986.

Zimmermann P, Hirsch-Hoffmann M, Hennig L, and Gruissem W (2004): GENEVESTIGATOR. *Arabidopsis* microarray database and analysis toolbox. *Plant Physiol*, **136** (1): 2621–2632.

Contributions

The work presented in the individual chapters of this thesis has partially been produced in collaborations. Contributions of the people involved in these projects are indicated here:

Chapter 1:
"Modeling cross-hybridization on phylogenetic DNA microarrays increases the detection power of closely related species "

Julia C Engelmann designed the primers, conducted experimental work (DNA isolation, PCR, algae culturing), analyzed the microarray data (data preprocessing, linear modeling, cross-validation, visualization of results) and supervised the project. Sven Rahmann, Tobias Müller and Epameinondas Fritzilas assisted in developing the linear model. Susanne Kneitz provided the microarray facility and expertise. Matthias Wolf selected appropriate algae species. Jörg Schultz and Thomas Dandekar helped revising the manuscript. Julia Engelmann wrote a manuscript which has been accepted by *Molecular Ecology Resources*.

Chapter 2:
"Unsupervised meta-analysis on diverse gene expression datasets allows insight into gene function and regulation "

Julia C Engelmann and Torben Friedrich selected microarray datasets from GEO database and annotated them. Julia C Engelmann normalized the microarray data, set up the contrasts and calculated fold changes and p-values for each contrast. Roland Schwarz performed the clustering, kernel PCA and gene selection in `R`. Julia C Engelmann interpreted the results biologically, discussed the functions of the genes representative for IAA and pathogen related contrasts and performed the analysis in MapMan. Torben Friedrich performed homology modeling of the DUF26 kinases. Steffen Blenk performed literature research. Tobias Müller supervised the project. Julia C Engelmann and Roland Schwarz drafted a manuscript which has been accepted by *Bioinformatics and Biology Insights*.

Chapter 3:
"Is gene activity in plant cells affected by UMTS irradiation? A whole genome approach. "

Julia C Engelmann designed and carried out the microarray data analysis (normalization, quality control, Correspondence Analysis, hierarchical clustering with bootstrapping and differential expression analysis) in `R`. She also conceived and performed the comparison of the microwave irradiation experiment with 74 datasets from a public database and tested the changes in expression measured with qRT-PCR for significance. Rosalia Deeken performed the irradiation experiment on the plant cell culture. Günther Nimtz constructed the irradiation protocol and set up the microwave irradiation facility. Rainer

Hedrich supervised the project. Julia C Engelmann, Rosalia Deeken and Rob Roelfsema drafted a manuscript which has been submitted to *Computational Biology and Chemistry: Advances and Applications*.

Chapter 4:
"An integrated view of gene expression and solute profiles of *Arabidopsis* tumors: A genome-wide approach "

Julia C Engelmann designed and performed the microarray data analysis (normalization, quality control, Correspondence Analysis, differential expression analysis), the functional category analysis, the comparison of the tumor dataset with public datasets studying effects of phytohormones and analyzed the qRT-PCR measurements for significant changes in gene expression. Rosalia Deeken interpreted the biological impact of the results and supervised the experimental work. Both authors wrote the manuscript. Marina Efetova and Tina Czirjak prepared the plant samples. Werner M Kaiser, Olaf Tietz, Markus Krischke and Martin J Mueller performed solute profile measurements. Rainer Hedrich supervised the project.

Chapter 5:
"Genome Expression Pathway Analysis Tool - Analysis and visualization of microarray gene expression data under genomic, proteomic and metabolic context "

Markus Weniger created the software and web interface and wrote the manuscript. Julia C Engelmann proposed the microarray data analysis methods for GEPAT and wrote `R` scripts for these analyses. She also gave general advice in microarray analysis. Jörg Schultz supervised the project and revised the manuscript.

Chapter 6:
"Explorative data analysis of MCL reveals gene expression networks implicated in survival and prognosis supported by explorative CGH analysis "

Steffen Blenk performed the data analysis. Julia C Engelmann aided in data analysis (Correspondence Analysis) and gave advice. Markus Weniger and Stefan Pinkert provided databank support. Andreas Rosenwald and Hans K Müller-Hermelink provided the patient data and pathology expert advice. Tobias Müller gave advice for data analysis. Julia C Engelmann and Tobias Müller revised the manuscript. Thomas Dandekar and Jörg Schultz supervised the project.

Chapter 7:
"Germinal Center B cell-like (GCB) and Activated B cell-like (ABC) type of diffuse large B cell lymphoma: Analysis of molecular predictors, signatures, cell cycle state and patient survival "

Steffen Blenk performed the data analysis. Julia C Engelmann normalized the microarray data and gave advice for data analysis. Markus Weniger assisted in using GEPAT. Andreas Rosenwald and Hans K Müller-Hermelink provided the patient data. Tobias Müller gave advice for data analysis. Thomas Dandekar and Jörg Schultz supervised the project.

Curriculum Vitae

Julia Cathérine Engelmann

Education

07/2004 - 03/2008	**PhD student** in the group of Prof. Thomas Dandekar, Department of Bioinformatics, University of Würzburg.
08/2003 - 06/2004	**Diploma thesis** in the Departments of Plant Physiology and Bioinformatics supervised by Prof. Rainer Hedrich and Prof. Thomas Dandekar, University of Würzburg. Title: "Comparative analysis of gene expression in tumor and reference tissue of *Arabidopsis thaliana*".
10/2001 - 07/2003	**Studies in Biology** at the University of Würzburg.
10/1999 - 09/2001	**Studies in Biology** at the University of Göttingen.
07/1999	**Abitur** (university-entrance diploma) at Goethe-Schule Wetzlar.
07/1997	**High School Diploma** at Olive High School, Mannford, Oklahoma, USA.

Stipends

06/2006	Travel allowance from the Boehringer Ingelheim Fonds to participate in the 12-day course "Integrated Data Analysis for High Throughput Biology" at Cold Spring Harbor Laboratory, USA.

Organizational

2005 - 2007	Organization and supervision of "Girls Day", a future career day for girl students from grades 8-10 funded by the German Federal Ministry of Education and Research.

List of Publications

Publications associated with this thesis

Julia C Engelmann, Sven Rahmann, Matthias Wolf, Jörg Schultz, Epameinondas Fritzilas, Susanne Kneitz, Thomas Dandekar and Tobias Müller: Modeling cross-hybridization on phylogenetic DNA microarrays increases the detection power of closely related species. *Molecular Ecology Resources (accepted).*

Julia C Engelmann, Roland Schwarz, Steffen Blenk, Torben Friedrich, Philipp N Seibel, Thomas Dandekar and Tobias Müller: Unsupervised meta-analysis on diverse gene expression datasets allows insight into gene function and regulation. *Bioinformatics and Biology Insights (accepted).*

Julia C Engelmann, Rosalia Deeken, Tobias Müller, Günther Nimtz, Rob GM Roelfsema and Rainer Hedrich: Is gene activity in plant cells affected by UMTS-irradiation? A whole genome approach. *Computational Biology and Chemistry: Advances and Applications (submitted).*

Steffen Blenk, **Julia C Engelmann**, Stefan Pinkert, Markus Weniger, Jörg Schultz, Andreas Rosenwald, Hans K Müller-Hermelink, Tobias Müller, Thomas Dandekar (2008): Explorative data analysis of MCL reveals gene expression networks implicated in survival and prognosis supported by explorative CGH analysis. *BMC Cancer (in press).*

Markus Weniger, **Julia C Engelmann**, Jörg Schultz (2007): Genome Expression Pathway Analysis Tool - Analysis and Visualization of Microarray Gene Expression Data under Genomic, Proteomic and Metabolic Context. *BMC Bioinformatics* 8: 179 (12 pp.).

Steffen Blenk, **Julia C Engelmann**, Markus Weniger, Jörg Schultz, Markus Dittrich, Andreas Rosenwald, Hans-Konrad Müller-Hermelink, Tobias Müller and Thomas Dandekar (2007): Germinal center B cell-like (GCB) and activated B cell-like (ABC) type of diffuse large B cell lymphoma (DLBCL): Analysis of molecular predictors, signatures, cell cycle state and patient survival. *Cancer Informatics* 3: 409-430.

Rosalia Deeken, **Julia C Engelmann**, Marina Efetova, Tina Czirjak, Tobias Müller, Werner M Kaiser, Olaf Tietz, Markus Krischke, Martin J Mueller, Klaus Palme, Thomas Dandekar, and Rainer Hedrich (2006): An Integrated View of Gene Expression and Solute Profiles of Arabidopsis Tumors: A Genome-Wide Approach. *Plant Cell* 18: 3617-3634.

Conference Contributions

Large-scale kernel-based explorative meta-analysis of Affymetrix genome arrays.
Julia C Engelmann, Roland Schwarz, Steffen Blenk, Torben Friedrich, Philipp N Seibel, Thomas Dandekar and Tobias Müller. ISMB/ECCB, Vienna, Austria, Jul 21-26, 2007 *(Poster)*.

DNA microarrays for species identification.
Julia C. Engelmann and Tobias Müller. Bioinformatics Symposium, University of Würzburg, 27. Juli 2006 *(Oral Presentation)*.

Biodiversity: About fish and chips.
Julia C Engelmann, Susanne Kneitz, Sven Rahmann, Matthias Wolf, Jörg Schultz, Thomas Dandekar and Tobias Müller. 7th International EMBL PhD Student Symposium, Heidelberg, Dec 1-3, 2005 *(Poster)*.

Integrative analysis of gene expression and CGH data using Hidden Markov Models.
Julia C Engelmann, Philipp N Seibel, Bastian Kröckel and Tobias Müller. 8th International Meeting of the Microarray Gene Expression Data Society (MGED8), Bergen, Norway, Sept 11-13, 2005 *(Poster)*.

ERKLÄRUNG

Hiermit erkläre ich ehrenwörtlich, dass ich die vorliegende Dissertation selbständig angefertigt und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe. Die Dissertation wurde bisher weder in gleicher noch ähnlicher Form in einem anderen Prüfungsverfahren vorgelegt. Außer dem Diplom in Biologie von der Universität Würzburg habe ich bisher keine weiteren akademischen Grade erworben oder versucht zu erwerben.

Würzburg, April 2008

Julia C. Engelmann