TRANSCRIPTIONAL PROFILING OF ADIPOGENESIS

HUBERT HACKL

DOCTORAL THESIS

Graz University of Technology
Institute for Genomics and Bioinformatics
Krenngasse 37, 8010 Graz, Austria

Graz, February 2004

Abstract

Distinct profiles of gene expression mirror the complex molecular mechanisms that regulate development during cellular differentiation and throughout life. To study the key events and processes in adipogenesis the gene expression of 3T3-L1 cell line during differentiation from fibroblast like preadipocytes to mature adipocytes with a 27.648 element focused murine cDNA microarray, comprising adipose specific genes and ESTs from early embryonic stages (NIA 15k clone set) was analyzed. Three independent time series experiments of 3T3-L1 adipocyte differentiation were performed in reference design. RNA from 8 time points (0d, 6h, 12h, 24h, 2d, 3d, 7d, 14d) was hybridized against RNA from the preconfluent stage in duplicate with reversed dye assignment to study the gene expression profile over the whole differentiation process. 780 genes found to be more than twofold up- or downregulated in at least 4 timepoints in comparison to the preconfluent stage were selected for further analysis.

A number of automated methods were proposed to extract biological meaning in the wealth of data. Clustering algorithms were performed (k-means, principal component analysis) to categorize these genes by their expression course. Additionally, an iterative reverse engineering approach based on mutual information and correlation were applied to elucidate gene-gene relations. To study the gene expression profiles in the context of relevant pathways, a database and web portal for genomics of lipid-associated disorders (GOLD.db) was initiated.

Many known and unknown differentially expressed genes in the mitotic clonal expansion phase and the late terminal adipocyte differentiation as well as transcriptional regulators could be identified and confirmed by Real Time PCR. Due to the focused approach and a novel thorough functional annotation process new promising targets were revealed.

Supplementary information is available at http://genome.tugraz.at/adipocyte

Keywords: Adipogenesis, Microarray, Transcriptional Profiling, Functional Annotation

Publications

This thesis was based on the following publications, as well as upon unpublished observations:

Papers

Hackl H, Burkard T, Sanchez Cabo F, Di Camillo B, Sturn A, Fiedler R, Paar C, Rubio R, Quackenbush J, Schleiffer A, Eisenhaber F, Trajanoski Z. Large scale gene expression analysis and functional annotation of adipocyte differentiation. *in preparation*

Hackl H, Maurer M, Mlecnik B, Hartler J, Trost E, Stocker G, Miranda Saavedra D, Trajanoski Z. GOLD.db: Genomics of Lipid-Associated Disorders Database. *submitted*

Hackl H, Sanchez Cabo F, Sturn A, Wolkenhauer O, Trajanoski Z. Analysis of DNA Microarray Data. *Curr Top Med Chem*, in press

Trost E, Hackl H, Maurer M, Trajanoski Z. Java Pathway Editor. *Bioinformatics*, 19:786-787, 2003

Pieler R, Sanchez Cabo F, Hackl H, Thallinger G, Z Trajanoski. ArrayNorm: Comprehensive normalization and analysis of microarray data. *Bioinformatics*, in press

Conference Proceedings and Abstracts

Hackl H, Burkard T, Paar C, Fiedler R, Sturn A, Stocker G, Rubio RM, Quackenbush J, Schleiffer A, Eisenhaber F, Trajanoski Z. Large scale gene expression analysis and functional annotation of adipocyte differentiation. Keystone Symposia: Molecular Control of Adipogenesis and Obesity, Banff, Canada, 210, 2004

Hackl H, Burkard T, Gaspard R, Quackenbush J, Trajanoski Z. Gene Expression analysis during adipocyte differentiation. High Level Scientific Conferences (HLSC): Molecular Mechanisms in Metabolic Diseases: Obesity, Diabetes Type 2, Lipid disorders and Atherosclerosis, Günzburg/Ulm, Germany, 9, 2003

Hackl H, Trost E, Maurer M, Miranda Saavedra D, Hofmann W, Trajanoski Z. Genomics of Lipid-Associated Disorders Database, Keystone Symposia: PPARs: Transcriptional Regulators of Metabolism and Metabolic Diseases, Keystone, CO, USA, 59, 2003

Hackl H, Yu Y, Quackenbush J, Trajanoski Z. AdipoChip: Focused Murine cDNA Microarray, Keystone Symposia: Molecular Control of Adipogenesis and Obesity, Keystone, CO, USA, 76, 2002

Hackl H, Sturn A, Michopoulos V, Quackenbush J, Trajanoski Z. Potential binding sites for PPARs in promoters and upstream sequences. 9th International Conference on Intelligent Systems for Molecular Biology ISMB 2001, Copenhagen, Denmark, 71, 2001

Hackl H, Sturn A, Michopoulos V, Quackenbush J, Trajanoski Z. In silico identification of PPARg target genes, Fourth Annual Conference on Computational Genomics (TIGR), Baltimore, MD, USA, J Comput Biol, 7:639, 2000

Hackl H, Thallinger GG, Wach P, Leberl FW, Trajanoski Z. High resolution CCD scanner for cDNA microarray analysis. Cambridge Healthtech Institutes 2nd Annual Lab-Chips & Microarrays for Biotechnical and Biomedical Applications, Zurich, Switzerland, 2000

Contents

1 Introduction 7
1.1 Background 7
1.2 Objectives 9
2 Methods 11
2.1 DNA microarray technology 11
2.2 Development and design of a focused murine cDNA microarray 15
2.3 3T3-L1 cell line 17
2.4 Experimental design 18
2.5 Microarray experiments 22
2.6 Data analysis 23
2.7 Functional annotation 26
2.8 Real Time RT-PCR 29
2.9 Promoter analysis 30
2.10 Analysis of the gene expression data in context of relevant pathways . . 33
2.11 Strategy for building a regulatory gene interaction 38
3 Results 41
3.1 3T3-L1 cell differentiation 41
3.2 Results and quality of the microarray experiments 42
3.3 Clustering genes according to their expression profiles 44
3.4 Functional annotation, gene ontology, and biological processes 55
3.5 Confirmation of microarray results by real time RT-PCR 57
3.6 Possible associated genes identified by a reverse engineering approach . 58
3.7 3T3-L1 differentiation expression data in context of several pathways . 61
4 Discussion 64
5 Conclusion 70
References 71

Glossary 83
Acknowledgement 87
Publications 88

1 Introduction

1.1 Background

Obesity causes or exacerbates many health problems, both independent from and in association with other diseases. In particular, it is associated with the development of type 2 diabetes mellitus, coronary heart disease, an increased incidence of certain forms of cancer, respiratory complications and osteoarthritis [1]. Obesity is diagnosed when weight normalized for height, or body mass index (BMI) - the weight in kilograms devided by the square of the height in meter - exceeds a defined threshold of 30. There is evidence of a marked increase in the incidence of obesity in most Western societies [1–4]. In the United States, for example, the change of the average BMI from 26.7 to 28.1 between 1991 and 2000 has led to a substantial raise in the number of people with BMI$>$30 [4]. What accounts for this epidemic portions is determined by the combination of genetic, environmental, and psychosocial factors. Physiological studies had previously suggested that body weight and energy stores are homeostatically regulated, with either weight loss or gain producing concerted changes in energy intake and expenditure that resist the initial perturbation [3]. The major function of adipocytes in this context is to store triacylglycerol in periods of energy excess and to mobilize this energy during times of deprivation.

Growth in adipose tissue and long-term changes in fat storage is the result of both hypertrophy (increase in size) and hyperplasia (increase in number) of adipocytes. Hypertrophy is thought to be the initial event that occurs during development of obesity. However, adipocytes cannot grow and accumulate lipids indefinitely. Consequently, increasing number accounts for the adipose tissue expansion observed in obesity and during normal adipose tissue development. The process of adipocyte development, also known as adipogenesis, follows a highly ordered and well characterized temporal sequence $invitro$. Initially, there is growth arrest of proliferating preadipocytes, usually achieved in cultured cell lines after contact inhibition. Following addition of a hormonal cocktail cells undergo one or two additional rounds of cell division known as clonal expansion before they start terminal differentiation and develop the mature adipocyte phenotype. The mitotic clonal expansion phase is a prerequisite for the later differentiation shown by inhibition at different points of the cell cycle during differentiation of 3T3-L1 cells, a widely used cell model to study adipogenesis [5]. The whole differentiation process is regulated by a transcriptional cascade (Figure 1). Members of several transcription factor families have been implicated in this process, including the CCAAT/enhancer-binding proteins C/EBPα,

C/EBPβ, C/EBPδ, the nuclear hormone receptor peroxisome proliferator-activated receptor $\gamma 2$ (PPAR$\gamma 2$), and the adipocyte determination and differentiation-dependent factor-1/sterol regulatory element-binding protein-1c (ADD1/SREBP1c).
C/EBPβ and C/EBPδ are induced very early during differentiation and have been shown to promote adipogenesis, possibly through induction of C/EBPα and PPARγ [6–8], and abrogation of their activity blocks adipose conversion [8, 9]. C/EBPα and PPARγ in turn are able to regulate each other, and to induce their own expression [10]. Although PPARγ is sufficient to induce the expression of many adipocyte genes, C/EBPα is required to confer insulin sensitivity to the adipocyte [10]. The basic helix-loop-helix (bHLH) transcription factor ADD1/SREBP1c could potentially be involved in a mechanism that links lipogenesis and adipogenesis, since ADD1/SREBP1c can activate a broad program of genes involved in fatty acid and triglyceride metabolism in both fat and liver and can also accelerate adipogenesis [11]. The activation of the adipogenesis process by ADD1/SREBP1c could be effected via direct activation of PPARγ [12] or through generation of endogenous ligands for PPARγ [13]. Recently, a number of new molecules and transcription factors involved in adipocyte differentiation were described, either activating or inhibiting adipogenesis. Moreover, it is now appreciated that the white adipose tissue (WAT) is not only a storage for energy (lipids) but plays also an active role in metabolic processes and the adipocyte differentiation by secreting a large number of bioactive molecules including tumor-necrosis factor α (TNFα) [14], interleukin-6 [15], leptin [16, 17], resistin [18, 19], and adiponectin [20]. This illustrates that although well studied the current model of transcriptional regulation gives not a comprehensive picture of the current processes.

Traditional molecular biology has focused on studying the function of individual genes considered in isolation. Although these conventional methods will keep essential the emerging high throughput technologies, allowing to study the expression of thousands of genes in parallel, have changed this view. With large scale analyses it is now possible to monitor the expression of most genes involved in a specific process over time. Subsequently, it is taken into account that data of knock-out experiments and molecular analysis of individual genes always reflects combinatorial regulation and redundancy in the effects of genes.

Some studies have addressed the question how genes are expressed during adipocyte differentiation of 3T3-L1 cells with either membrane based microarrays [21] or high density oligonucleotide microarrays [22–26]. There are also first proteomic approaches to identify secreted proteins during the 3T3-L1 cell differentiation [27]. These studies demonstrate the power of

this type of analysis to a model of differentiation. Thousands of genes were shown differentially expressed - some of them even substantial - during the time course partly show a high agreement with results from previous experiments. However, there is the demand for further studies with optimal experimental design and sound analyses, providing potential associations of differentially expressed genes and subsequently, leading to the identification of new targets. At the end, this will facilitate the generation of new hypotheses.

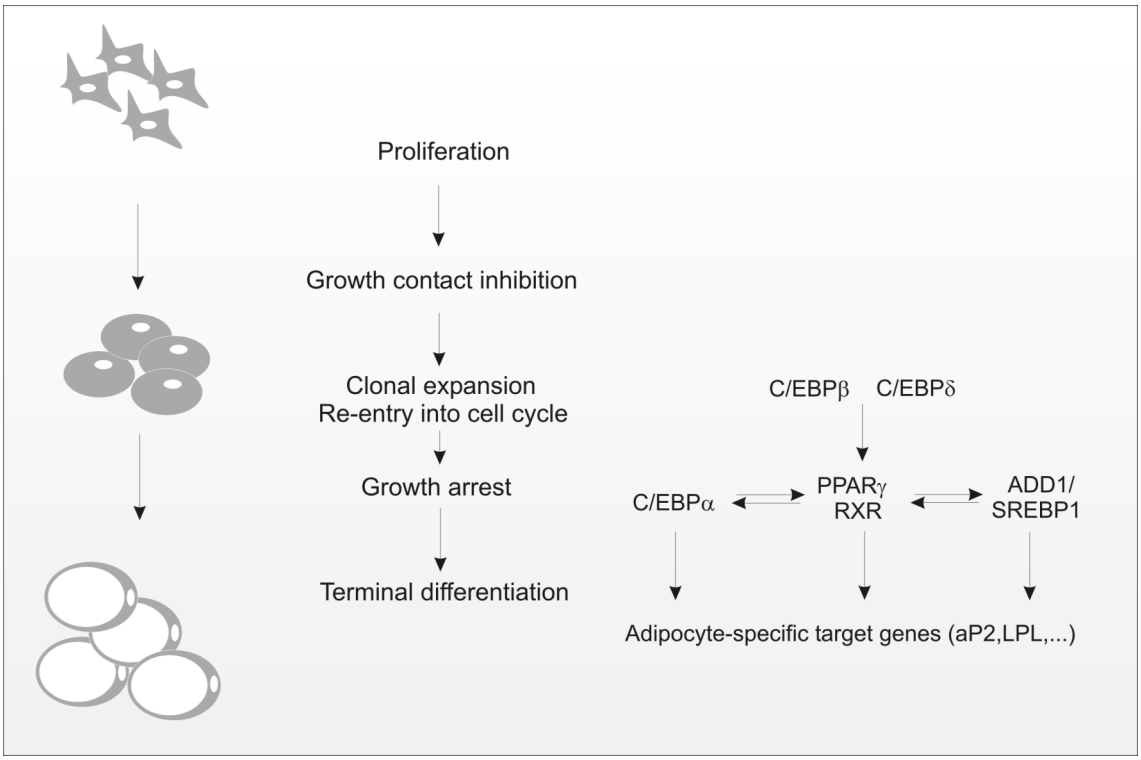

Figure 1: The different stages of adipogenesis (left) and the relation between the key players of the transcriptional regulatory cascade (right) adopted from [28]

1.2 Objectives

The main objective of this thesis was to study transcriptional profiling of 3T3-L1 cells during differentiation to elucidate regulatory processes and to identify new target genes involved in adipogenesis. For this purpose a focused approach with cDNA microarrays, including adipocyte specific genes and many expressed sequence tags (ESTs) from early developmental stages was pursued. To assure high quality of the data appropriate experimental design is fundamental. Commonly, after performing microarray experiments scientists have to face the problem to

study the results gene by gene and put them into a meaningful biological framework. Therefore, a concern of this thesis was to provide methods facilitating biological interpretation and applying here within the current study.
Consequently, the specific aims of this study were:

- Design and development of a focused murine cDNA microarray
- Differentiation of 3T3-L1 preadipocyte cell line as model for adipogenesis
- Large scale gene expression analysis during adipocyte differentiation with microarrays
- Preprocessing of microarray data and identification of differentially expressed genes
- Functional annotation of selected genes (ESTs) and integration of gene ontology
- Identification of clusters of genes with similar expression profile over time
- Building a regulatory network from the known gene-gene interaction derived from literature
- Development of a strategy for reverse engineering methods to build a network based on the microarray data to provide possible associations between genes
- *In silico* identification of target genes and promoter analysis
- Confirmation of the gene expression levels of potential candidates with Real Time RT-PCR
- Initiating a database and web portal to upload pathways and microarray data sets and to analyze the gene expression levels in the context of several pathways

2 Methods

2.1 DNA microarray technology

DNA microarray technology has become an important tool in biomedical research during the last years. All variants of this technology allow simultaneous measurement of expression levels of thousands of genes in a single experiment. The resulting patterns are characteristic for the responses of cells or tissues to their environment, to differentiation into specialized tissues, or to dedifferentiation into neoplastic cells. The great potential of DNA microarrays not only lies in viewing the technology as a collection of individual expression measurements, but also in generating a composite picture of the expression profile of the cell. Therefore microarrays are widely used in basic research as well as in clinical medicine and pharmacogenomics.

The two major platforms for microarrays are spotted arrays [29], where the probes are mechanically deposited on modified glass slides by contact or inkjet printing, and in situ arrays, where oligo probes, usually 20 to 25 nucleotides in length, are synthesized via photo litography and combinatorial chemistry techniques (GeneChip arrays, Affymetrix, Santa Clara, CA). In the latter approach, each gene or expressed sequence tag (EST) is represented on the array by 16-20 probe pairs, consisting of a perfect match (PM) and a mismatch (MM) oligonucleotide that differs from the perfect match by only a single base in the center position. The purpose of the MM sequence is to capture the non-specific binding that distorts the measured intensity level of the PM [30, 31]. Templates for the gene of interest used in cDNA microarray technology are obtained and amplified by polymerase chain reaction (PCR). Following purification and quality control, aliquotes are printed on coated glass microscope slides using a high precision robot. Total RNA from test cells (e.g. treated cells) and reference cells (e.g. untreated cells) is reverse transcribed to cDNA and fluorescently labeled with different dyes, commonly Cy3 and Cy5 are used. This is in contrast to in situ arrays, where only one labeled RNA sample is used. The fluorescent targets are pooled and allowed to hybridize under stringent conditions to the elements on the array. The bound fluorescent dyes are excited and slides are scanned by a laser scanner. The resulting monochrome images must then be analyzed to identify the arrayed spots and to measure the relative fluorescence intensities for each element. The basic principle of the cDNA microarray technology is illustrated in Figure 2.

The typical procedure for a microarray experiment comprises several steps and several issues

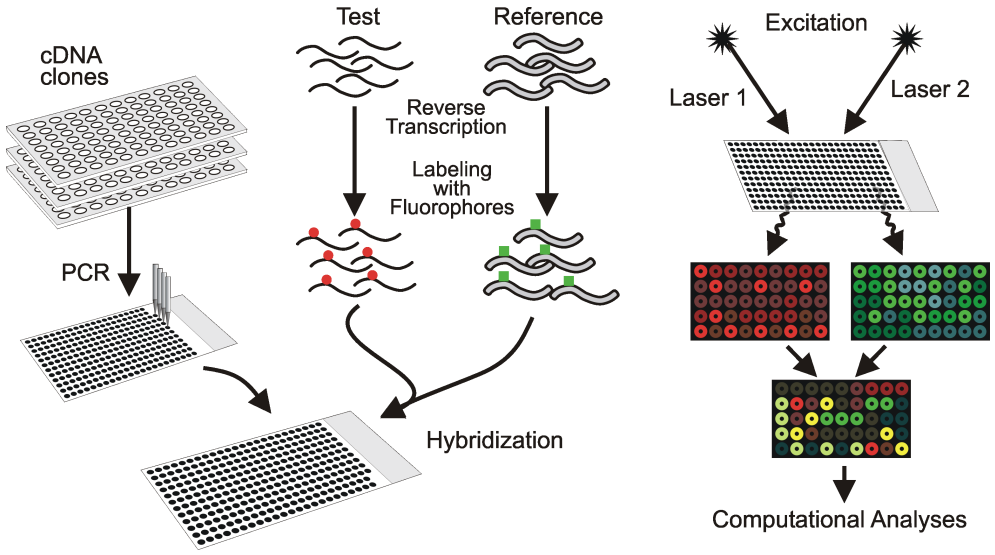

Figure 2: Schematic overview of the cDNA microarray technology

have to be considered (see figure 3). Starting from the biological question the most important point is the experimental design. The objective of the experimental design is to make the analysis of the data and the interpretation of the results as simple and as powerful as possible, given the purpose of the experiment and the constraints of the experimental material. The next question is how to design the array, that is which elements should be included on the array and what type of array to use. For spotted microarrays, one has the option to spot cDNA - in general PCR products (1,000-1,500 bp in length) of clones with an inserted cDNA element representing an expressed sequence tag (EST) or a gene - or oligonucleotides, designed to capture specific genes. In addition, the elements can be designed to represent the most unique part of a given transcript, making the discrimination of closely related genes or splice variants possible. Spotted arrays allow a greater degree of flexibility in the choice of arrayed elements, particularly for the preparation of smaller, customized arrays for specific investigations. It is essential to use optimized hybridization and labeling procedures, to assure a high signal to background ratio in the resulting images. It has turned out, that indirect labeling gives better results than direct labeling [32]. The difference is that the latter uses already labeled nucleotides for first strand cDNA synthesis in contrast to indirect labeling, where cDNA synthesis is performed with chemically modified nucleotides and the fluorescent dyes are coupled in a further step. Moreover it has been shown that prehybridization of the slide with bovine serum albumin (BSA) block non-specific binding of the samples to the surface of the slide and adding Cot1-DNA

and poly(A)-DNA to the hybridization buffer prevents unspecific binding [33] to improve the quality of hybridization. There are many commercial and freely available software packages for image quantitation, which comprise spot recognition, segmentation and intensity extraction. Although there are minor differences between these programs, most give high-quality, reproducible measures of intensities. Before analyses of the data several transformation steps are important: background correction, remove of bad quality spots, log-2 transformation, and normalization. A couple of methods [34–36] have been proposed for the process of normalization, which aims to remove systematic errors by balancing the fluorescence intensities of the two labeling dyes. Some commonly used methods for calculating the normalization factor include 1) global normalization that uses all genes on an array 2) houskeeping genes normalization that uses invariant expressed housekeeping genes, and 3) internal control normalization that uses known amount of exogenous control genes added during hybridization. In cases where the dye bias depends on spot intensity and spatial location on the array non-linear normalization methods, like the locally weighted scatterplot smoothing (LOWESS) normalization method, are preferable. LOWESS detects systematic deviations in the MA-plot, this is the $\log_2$-ratios plotted versus the average spot intensity, and performs a robust linear local fit of the data. After normalization, assuming that most genes are equally expressed, the distribution of log-ratios should be centered around zero, the ratios should be independent of spot intensity, and the fitted line should be parallel to the x-axis. The lowess algorithm can be applied to the whole array (global) or over print tip groups to consider spatial effects. Another possibility to reduce the dye bias problem is to perform dye swap experiments and perform dye swap normalization.

The simplest way to identify genes of potential interest through several related experiments is to search for those that are consistently either up- or downregulated. To that end, a simple statistical analysis of gene-expression levels will suffice. However, identifying patterns of gene expression and grouping genes into expression classes might provide much greater insight into their biological relevance. Although a different number of unsupervised clustering algorithms were previously applied to gene expression data (hierarchical clustering [37], k-means [38], clustering affinity search technique (CAST) [39], self organizing maps (SOM) [40], self-organizing tree algorithm (SOTA) [41], principal component analysis (PCA) [42], singular value decomposition (SVD) [43, 44], correspondence analysis (CA) [45], gene shaving [46]) there is no one-for-all solution and an appropriate choice of data analysis technique depends on both the data and on the goal of the study. The principal problem is that even if random data are analyzed the algorithms will group genes in clusters and different methods produce different

cluster results, considering also the possibility of different similarity measures. The aim is generally to define clusters that minimize distances between the genes in one cluster and maximize intercluster variance [47]. Although there are attempts for automated cluster validation [48], validation can be made testing and comparing several clustering methods on the current data. Supervised methods or class predictors, like support vector machines (SVM) [49], are generally used for finding genes with expression levels that are significantly different between groups of samples as well as finding genes that actually predict a characteristic of samples. Common applications for this type of analysis can be found in the classification of tumor samples [50]. As every high throughput method, microarray studies produce a tremendous amount of data. Therefore, it is inevitable to design relational databases able to store data and information about the experiment in a safe and yet easily retrievable manner (a review of the current approaches can be found in [51]). It is important that this information is archived according to accepted scientific standards, which will then allow scientists to share common information and make valid comparisons among experiments, and allow complex queries of the data using standard language or controlled vocabulary. One such standard, minimum information about a microarray experiment (MIAME) [52] has been formalized by the Microarray Gene Expression Data Consortium (MGED) (http://www.mged.org/Workgroups/MIAME/miame.html). Another standard to emerge is the microarray gene expression markup language (MAGE-ML) [53]. This is a descriptive language to exchange data widely adopted by several microarray database systems and applications. Both ArrayExpress [54] and the Gene-Expression Omnibus (GEO) [55], which are two of the most prominent public microarray data repositories, intend to support the MIAME and MAGE-ML standards, as well as MGED ontology. Standardization attempts can also be found in the gene annotation process. The gene ontology (GO) [56] is providing controlled vocabulary as knowledge of gene and protein roles. Since this information about the molecular function and biological processes is structured in trees, it is possible to group genes according to their common paths in the tree of GO terms for different genes. Moreover it can help to look for similar functions in already clustered genes. The major challenge after computational analysis is to facilitate the search for biological meaning in the data and to generate new hypotheses. The 'list of genes' resulting from microarray analysis should not be viewed as an end itself. Its real value increases only as that list moves through biological validation, ranging from the numerical verification of expression levels with alternative techniques to ascertaining the meaning of the results. Search for common regulatory sequences and transcription factor binding sites, especially for genes with similar expression profile in time series

experiments is a possibility to understand the architecture of genetic regulatory networks. It is optimistic to assume that expression data alone will be sufficient for the inference of complete regulatory pathways. However, several recent reverse engineering approaches, like Boolean or Baysian networks tackle parts of the problem, discovering new gene associations and modeling subnetworks. A different method is to analyze gene expression data in the context of known metabolic pathways, as in KEGG [57] or BioCarta. Tools for this mapping process [58] offer the potential to reveal differentially regulated genes under certain physiological conditions in a specific cellular component. Mapping to genomic sequences or chromosomes, respectively, will reveal the vicinity to genetic markers or correlated expression patterns of adjacent pairs [59]. Subsequently, automated search for interesting literature clusters ranked according to data analysis parameter or creating keyword hierarchies show the relationship of the results of the microarray experiment to published literature [60–62]. The greatest potential to glean new insights, however, will be achieved by computational approaches, which integrate genomic data with disparate information. Probabilistic relational models (PRM) [63] is one approach to allow the inclusion of multiple types of information (e.g. gene ontology definitions, protein motifs, or functional information) in the computational process itself. The challenges associated with clinical applications and at the and with personalized medicine will overcome by extensive bioinformatic solutions for these data integration.

2.2 Development and design of a focused murine cDNA microarray

For the specific application we decided to use cDNA microarrays, because this type of arrays provides the flexibility to include both, first a large number of genes (ESTs), which are uncharacterized and previously not associated with adipogenesis and second genes (ESTs) that are important in adipocyte biology and lipid metabolism (focused approach). A 27.648 element mouse microarray was developed and clones from early developmental stages, clones for adipose specific genes, as well as a brain specific library were utilized. The chip comprises the following libraries (features):

- *627 adipose specific clones:* All expressed sequence tags (ESTs) from a set of 18,376 non-redundant clones spotted on a nylon membrane array shown to be differentially expressed in 3T3-L1 adipocytes, when compared to preadipocytes [21] were included on the microarray. It was recently demonstrated using oligonucleotide microarrays, that the expression of many genes is regulated by leptin [64]. Since leptin is a negative target for PPAR, ESTs shown to be differentially expressed were included on the microarray. And

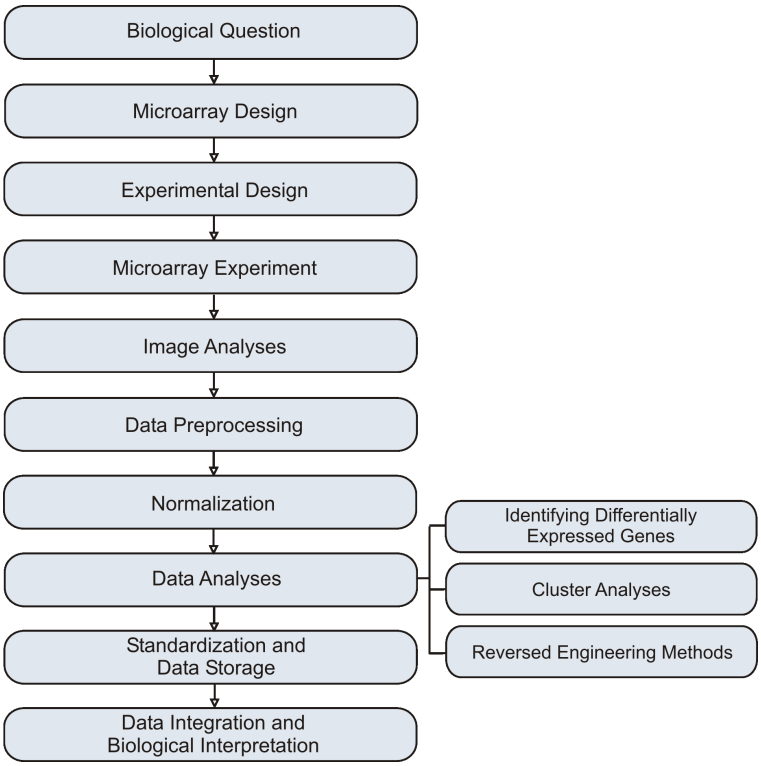

Figure 3: Procedure for microarray experiment and data analyses

third, all relevant genes, known to be important to adipocyte biology and lipid metabolism including transcription factors, coactivators and -repressors, enzymes, and signal transduction molecules were included.

To select eligible clones (ESTs) for the chosen genes, the TIGR Mouse Gene Index Build 5.0 (http://www.tigr.org/tdb/tgi.shtml) was used. Tentative Consensus (TC) sequences within the TIGR Gene Index (TGI) are unique, high-fidelity, virtual transcripts produced by clustering and assembling of ESTs and annotated gene sequences from the GenBank. The applied strategy was to select a TC where the functional annotation matches the gene name in the derived list. One EST within a TC was selected according to the availability and quality of the corresponding cDNA clone. A high sequence similarity is ensured due to the restrictive assembly process of the TCs (only those ESTs are clustered, which are more than 95% similar in an overlap of 40 bp). However, some genes of interest are not represented in the TGI. In this case ESTs were obtained by a high score BLAST search of GenBank entries with a complete coding sequence of the mRNA against the mouse EST

database. The corresponding cDNA clones were ordered through the IMAGE consortium (Research Genetics, Huntsville, AL, USA).

- *15k NIA mouse clone set:* ESTs from early embryonic cDNA library (National Institutes of Aging, National Institutes of Health, Bethesda, MD, USA). 15,000 ”unique” cDNA clones were rearrayed among 52,374 ESTs from pre- and periimplantation embryos, E12.5 female gonad/mesonephros, and newborn ovary. Up to 50% are derived from novel genes.

- *11k BMAP clone set:* 10560 clones from non-normalized and normalized, serially subtracted cDNA libraries from 10 brain regions of adult mouse brain, spinal cord, and retina (mouse strain c57BL6J). These libraries were constructed within the Brain Molecular Anatomy Project (BMAP) contracted by the National Institutes of Health (NIH) and the University of Iowa.

The insert of most of these clones (NIA and BMAP clone set) were sequence verified. The insert size of the clones was about 1-1.5 kbp. All PCR products were purified using size exclusion vacuum filter plates (Millipore) and spotted in singular out of 50% DMSO onto amino-silanated glass slides (UltraGAPS II, Corning) in an 4x12 print tip group pattern. Negative controls (genomic DNA, genes from Arabidopsis thaliana, and DMSO) and positive controls (Cot1-DNA and Salmon sperm DNA) were included in each of the 48 blocks. Samples were bound to the slides by ultraviolet crosslinking at 200 mJ in a Stratalinker (Stratagene).

2.3 3T3-L1 cell line

The understanding of the adipocyte differentiation process - from a fibroblast progenitor cells to mature adipocytes - is derived primarily from cell culture models, i.e. either established cell lines or primary preadipose cells. Primary preadipose cells can be isolated from the stromal vascular fraction of adipose tissue and, when treated in cell culture with a combination of adipogenic effectors, can differentiate into adipocytes. Established preadipocyte lines have advantages over primary preadipocytes in that they provide a homogenous population and can be carried in culture indefinitely. The most widely used culture models are the 3T3-L1 and 3T3-F422A culture lines, which are derived from disaggregated Swiss 3T3 mouse embryos [65, 66]. These cell lines can be induced to differentiate into adipocytes that display the morphological [67] and biochemical [68] characteristics of adipocytes in situ. Moreover, when 3T3-F422A

cells are implanted subcutaneously into athymic mice, they give rise to fat pads that are indistinguishable from adipose tissue [67]. Confluent 3T3-L1 preadipocytes can be differentiated synchronously by a defined hormon cocktail. Maximal differentiation is achieved upon treatment with the combination of insulin, a glucocorticoid, an agent that elevates intracellular cAMP levels, and fetal bovine serum [69]: Insulin is known to act through the insulin-like growth factor 1 (IGF-1) receptor. Dexamethasone (Dex), a synthetic glucocorticoid agonist, which is traditionally used to stimulate the glucocorticoid receptor pathway and methylisobutylxanthine (MIX), which is a cAMP-phosphodiesterase inhibitor used to raise the cAMP level and to stimulate the cAMP dependent protein kinase pathway.

2.4 Experimental design

2.4.1 Differentiation of 3T3-L1 cell line

3T3-L1 cells (ATCC number CL-173) were grown in 100mm diameter dishes in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal bovine serum, 100 units/ml penicillin, 100 μg/ml streptomycin, and 2mM L-Glutamine in an atmosphere of 5% CO_2 at 37°C. Two days after reaching confluence (day 0), cells were induced to differentiate by a two day incubation of a hormone cocktail [69, 70] (100 μM 3-iso-butyl-1-methylxanthine, 0.25 μM dexamethasone, 1 μg/ml insulin, 0.16 μM pantothenic acid, and 3.2 μM biotin) added to the standard medium described above. After 48 h (day 2) cells were cultured in the standard media in the presence of 1 μg/ml insulin, 0.16 μM pantothenic acid, and 3.2 μ M biotin until day 14 (Figure 5). Nutrition media were changed every second day.

2.4.2 Experimental design for microarray analysis

All biological conclusions and predictions resulted from microarray data consequently rely on the quality of the data and the use of appropriate analytical methods and statistical tests. Subsequently, it is important to focus on the design of a microarray experiment. Carefully designed biological experiments are indispensable for the analysis of data and the interpretation of results. The general principles discussed here are for two-color microarray experiments. However, many of these issues will also apply to single-color gene expression assays. Microarray experiments should be treated as general biological experiments and microarrays as measurement of a biological quantity. The following issues should be considered for the design of a microarray experiment:

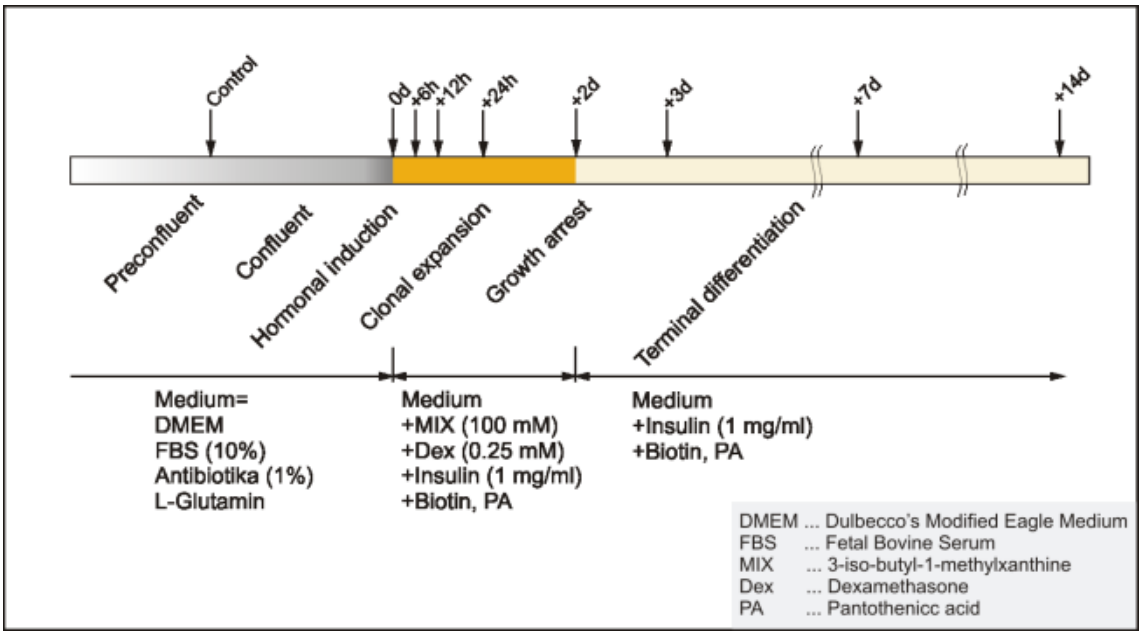

Figure 4: Design of the differentiation experiment. Arrows indicate when RNA is extracted from the cells

- *Biological replicates:* Biological variability is intrinsic to all organisms and can be substantial even for inbred mice [71]. Therefore it is necessary to perform repeated hybridizations with RNA samples from independent sources. The number of the biological replicates depends on the type of the biological question, e.g. whether two types of tissues differ with regard to the expression profiles. In this particular example it is obvious that one RNA sample for each tissue does not answer the biological question, since there may be a substantial biological variability within this tissues [72].

- *Technical replicates:* In microarray experiments there are two possibilities for replicated measurements to reduce variability introduced by measurement errors: replicated features within a slide and replication of hybridization with the same RNA samples. It is useful to have a few technical replicates to ensure that the procedures, reagents and equipment are working properly [72]. However, since technical replicates do not represent independent measurement, it is strongly recommended to use biological replication as principal source of replicated slides [34]. Dye-swap can be also used as a technical replicate and is useful for reducing the systematic bias. In this case replications represent repeated hybridization with the same samples, with interchanged dyes in the second hybridization. This can also

be used for gene-wise normalization to balance the systematic differences in the red and green intensities.

- *Pooling:* In addition to the optimal design there are constraints on the number of slides, the amount of RNA available or cost considerations, all of which will effect the experimental design. In cases where a large number of experimental units may be available and the number of slides is limited, there is the option to pool individual samples. This may also comprise cases with limited amount of RNA. If all available samples are pooled together the biological variance are minimized, but all independent replication would be eliminated. Therefore it is better to use several pools and fewer technical replicates.

- *Control versus reference RNA:* For a pairwise comparison of differential gene expression one of the two RNA samples, which are hybridized to a microarray slide, is used as a control, e.g. RNA from an untreated cell line. In cases of comparison of many different types of samples, it would be desirable to use a more universal reference with a broad coverage of genes, e. g. RNA from pooled cell lines. This approach is commonly used if the focus of the analysis is to determine tumor subtypes [73].

The optimal arrangement of the microarray experiment is driven by the objectives and conditioned by several constrains. The effects of the different parameters are discussed in [71] and can be summarized by the following formulae: Suppose that n pools of k individuals each were created, and each pool will be measured using m technical replicates on microarray slides with r repetitions of each clone, the mean squared error (MSE), which quantifies the precision of estimates, can be calculated by

$$MSE = \sqrt{(\frac{\sigma_B^2}{k^a} + (\sigma_A^2 + \sigma_e^2/r)/m)/n} \tag{1}$$

where σ_B^2 is the intrinsic variation of the biological units within a experimental class, σ_A^2 represents the variation between technical replicates, and σ_e^2 represents the measurement error within a single array. For pooling a denotes a constant with $0 < a < 1$. In the case of $a = 0$ pooling will have no effect, and in case of $a = 1$ the variance is reduced in direct proportion to the pool size. Let C_1 represent the cost of an experimental unit and C_M be the cost for a single technical replicate the overall cost for the experiment would be:

$$cost = n(kC_1 + mC_M) \tag{2}$$

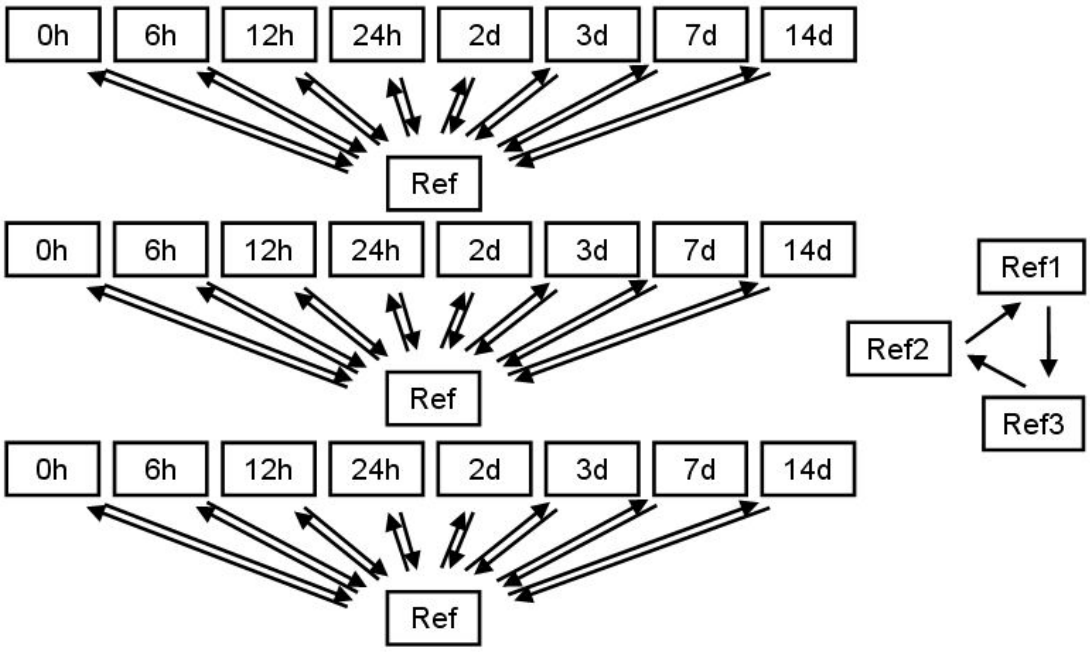

Figure 5: Experimental design of the adipogenesis study. One array represents one microarray slide. The tips of the arrows point to the RNA labeled with Cy3 and the opposite side to the RNA labeled with Cy5. Due to the dye swapping there are two arrows per time point. Additionally to the three independent experiments, the reference samples in each experiment were hybridized to the microarrays (a total of 51 microarrays).

Although a detailed knowledge of the variance components is hardly accessible, some general guidelines can be derived. E.g., when measurement is expensive it is preferable to add experimental units rather than technical replicates, or if the variability between individual samples is large and the units are not too costly, it may be worthwhile to pool samples.

The discussed issues were considered in the current study. Three independent time series experiments of 3T3-L1 adipocyte differentiation were performed in reference design to reduce biological variation. As a reference RNA from the preconfluent stage of the 3T3-L1 cells (80 % confluent) was pooled from 36 dishes for each of the three experiments. At each time point RNA was pooled from 3 culture dishes. As technical replicate each hybridization was repeated with reversed dye assignment (Figure 5). The variation in the gene expression of the references between the independent experiments was checked by three hybridizations of the reference RNA samples to the microarrays.

2.5 Microarray experiments

Cells were harvested and total RNA was isolated at the preconfluent stage and at 8 time points with TRIzol reagent (Invitrogen-Life Technologies) based on the method described previously [74]. The quality of the RNA was checked by Agilent 2100 Bioanalyzer RNA assays through inspection of the 28S and 18S ribosomal RNA intensity peaks.

The used labeling and hybridization procedures were based on those developed at The Institute for Genomic Research [33] and can be viewed at http://gold.tugraz.at. Briefly, 20 μg of total RNA was indirectly labeled with Cy 3 and Cy5, respectively. The Random Hexamer (Invitrogen) primed first strand cDNA synthesis was carried out using Superscript Reverse Transcriptase II (Invitrogen) in the presence of amino allyl dUTP (Sigma), dATP, dGTP, dCTP, dTTP (Invitrogen), DTT, and 1X first strand buffer overnight at 42 °C. cDNA was purified with QIAquick columns (Qiagen) according manufacturer's directions, but using potassium phosphate wash and elution buffer instead of supplied buffers PE and EB. N-hydroxi succinimde (NHS) esters of Cy3 and Cy5 (Amersham) were coupled to the amino allyl dUTPs incorporated in the cDNA. Coupling reactions were quenched by 0.1 M sodium acetate (pH=5.2) and unincorporated dyes were removed using QIAquick columns (Qiagen). Slides were prehybridized in 1% BSA, 5xSSC, 0.1 %SDS for 45 min at 42°C and washed in MilliQ water and 2-Propanol and dried in a centrifuge. Fluorescent cDNA samples were dried in a SpeedVac, resuspended in 12 μl hybridization buffer (50 % formamide, 5XSSC, 0.1 % SDS) and combined. 20 μg mouse Cot1 DNA and 20 μg poly(A) DNA were added, denatured at 95°C for 3 min. and snap cooled on ice for 1 min. Sample with a final sample volume of 26 μl was applied to the prehybridized slide, covered with a glass cover slip (Roth) and hybridized in a humidified chamber for 20 hours at 42°C in the dark. Slides were washed 4 min in a 1xSSC, 0.2 % SDS solution (42 °C), 4 min in 0.1xSSC, 0.2% SDS, 4 min and 2.5 min in a 0.1xSSC, dipped twice in MilliQ water and dried in a centrifuge.

Slides were scanned with a GenePix 4000B microarray scanner (Axon Instruments) at 10 μm resolution. Photo multiplier voltages (PMT) were selected in order that the histogram of the red channel (635nm)and the green channel (532nm) were overlapping to a large extend and few spots were saturated. Identical settings were used for the scanning of the corresponding dye-swapped hybridized slides. The resulting TIFF images for each of the two fluorophors were analyzed with GenePix Pro 4.1 (Axon Instruments) to get relative gene expression levels for each gene.

2.6 Data analysis

2.6.1 Data preprocessing

Data were filtered for low intensity, inhomogeneity, and saturated spots by following criteria. If the median of the pixel intensities within a spot are differing more than 20% from the mean of the pixel intensities spots were considered as inhomogeneous and were filtered out. Spots with more than 10% of saturated pixels (fluorescence intensity $>$65535) were excluded from further analysis. To get expression values for the saturated spots, slides were scanned a second time with lower photo multiplier tube settings and analyzed again. Genes with a very low expression value are often removed in order not to confound their signal with the background intensity. Low intensity spots were defined as those where the sum of the medians/means of the pixel intensities in both channels was lower than 1000 or not more than 55% of the pixels within a spot had intensities higher than the intensity of the surrounding background plus one standard deviation of the background pixels. All spots of both channels were background corrected, by estimation and subtraction of the local background.

2.6.2 Normalization

There are different sources of systematic (sample effect, array effect, dye effect and gene effect) and random errors associated with microarray experiments [75]. Due to the different physical properties of the fluorescent dyes, the major portion of this bias is introduced by the dye effect. Therefore it is indispensable to normalize the data, which is known as removing of all non-biological variation introduced in the measurement and minimizing the random error to get reliable results [34, 35]. As method of choice dye-swap normalization was applied. The expression ratio T for gene i at each time point in relation to the reference was calculated by

$$Ti = \sqrt{\frac{R_{i1}}{G_{i1}} . \frac{G_{i2}}{R_{i2}}} \tag{3}$$

where R_i refers to the red signal and G_i for the green signal for gene i. At the first hybridization, indicated by index 1, the green dye was assigned to the reference and the red dye to the sample. In the second hybridization, indicated by index 2, the assignment of the dyes was reversed. Moreover, it was checked that the course of the ratios over the signal intensities is equal in both hybridizations, which is a prerequisite for applying this normalization method. Since this process is applied gene-wise it is advantageous over global methods, where it has to be assumed that most of the genes are expressed equally in both of the channels. Subsequently,

genes showing substantial differences in the intensity ratios between technical replicates were excluded from further analysis based on a two-standard-deviation cut-off. The resulting ratios were $\log_2$ transformed and for each time point in most cases averaged over 3 independent experiments. Only in a few cases due to missing values, ratios where averaged over 2 independent experiments.

2.6.3 Identification of differentially expressed genes

Regardless of the experiment performed it is invariably of interest to identify genes that are differentially expressed between one or more pairs of samples in the data set. It is useful to reduce the number of genes to those that are most variable between samples. This is often accomplished applying a fixed fold-change cut-off (generally 2-fold) to the mean of the ratios over the biological replicates. To find out the genes mostly regulated in adipocyte differentiation, genes without missing values in all time points and more than 2-fold up- or down-regulated in at least 4 time points were selected for further analysis. The drawback of ranking genes according to their mean is that the variability of the ratios over replicates is not constant across genes. A solution is to rank the genes according to the absolute value of the t-statistic:

$$t = \frac{\overline{T}}{s/\sqrt{n}} \tag{4}$$

where $\overline{T}$ is the mean of the ratios and s the standard deviation of the ratios for n biological replicates. Some studies considered that unrealistically small values for s can lead to a large t-statistic by penalizing s [76–78]. After ranking the genes a cut-off value has to be found above which genes are assigned significantly differentially expressed. For each t-statistic a p-value, that is the probability that the null-hypothesis is rejected although it is true - can be calculated based on the student-t-distribution. In general, the cut-off for the significance is p=0.05. However, the problem in microarrays is that due to multiple testing, for example in testing the null-hypothesis for 10.000 genes 500 genes are probably false discovered as significant differentially expressed. To avoid this the p-values should be corrected to control the family-wise type I error rate [79]. The most stringent method is the Bonferroni correction, where the p-value is divided by the number of comparisons [80]. Other possibilities are the step-down method [81] or bounding of the false discovery rate [82]. For more experimental groups the Analysis of Variance (ANOVA) is commonly used to identify differentially expressed genes [75]. ANOVA provides a F-value, which refers to an estimator of the variance based on the variance within groups devided by an estimator of the variance based on the variance among groups. If $F \gg 1$

the null hypothesis - that the means of all groups are equal - is rejected. Selection of an adequate cut-off p-value results in a corresponding F-value above the gene is considered significantly differentially expressed. The expression data from the adipocyte differentiation were filtered for genes with at least 2 biological replicates in all time points and one-way ANOVA was applied to detect significant differentially expressed genes based on a p-value of 0.05. The application of these parametrical tests or further data analysis, e.g. cluster analysis, presumes normal distributed data. It was shown previously, that the estimated histogram of gene expression data in logarithmic scale is unimodal and symmetric and presents a roughly normal distribution [83]. However, this has to be tested for the results of each microarray hybridization in an experiment by estimating the histogram and analyzing the Q-Q-plots. There are also several other approaches and none parametrical tests to identify differentially expressed genes: mixed models [84, 85], empirical Bayes analysis [78], regularized t-test [86], significance analysis of microarrays (SAM) [77], regression analysis [87], and maximal likelihood approach [88].

2.6.4 Cluster analysis

Selected genes were grouped according to their similarity of the expression profiles over time with different unsupervized clustering methods. Each gene expression profile was mathematically expressed as vector of the log ratios at each time point $\{\mathbf{x} = (x_1, x_2, .., x_n)$;$\mathbf{y} = (y_1, y_2, .., y_n)\}$. As distance measurement the Euclidean distance:

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{5}$$

was used to calculate similarity between the expression profile of two genes, where n is the number of time points. K-means clustering [38, 89] was applied to the selected genes with different number of preselected clusters (k). Finally, $k = 12$ was chosen as number for cluster partitioning, because it provided the best compromise between number of clusters and separation between them. In k-means clustering k reference vectors are initialized randomly and genes are partitioned to their most similar reference vector. Each reference vector is recalculated as the average of the genes that mapped to it. These steps are repeated until convergence, that is, all genes map to the same partition. Consistency of the first clustering approach was checked using principal component analysis (PCA) [42]. The central idea of principal component analysis is to reduce the dimensionality of the data set while retaining as much as possible variation in the data set. Principal components (PCs) are linear transformations of the original

set of variables. Moreover, PCs are uncorrelated and ordered so that the first few PCs contain most of the variations in the original data set. The first 3 PCs were projected in 3 dimensional space, so that maximum variability is visible and clusters can be separated. In the 3D view of the PCs genes were colored according their assignment to the k-means clusters. When the principal components are calculated from the covariance matrix there is a direct relation between PCA and singular value decomposition (SVD) [43, 90], and therefore the same mathematical background is applicable. All cluster analyses were carried out within the clustering software Genesis [91].

2.7 Functional annotation

2.7.1 Annotation process

Each of the ESTs corresponding to the spotted clones on the microarray were annotated according to the TIGR mouse gene index (the provided tentative annotation was used as gene name). The affiliation of these ESTs to UniGene Clusters and the assigned cluster titles were also specified. Based on this annotation and accession numbers the following databases were tracked and unique identifiers were provided to glean further related information for each gene(EST) over the web:

- TIGR Mouse Gene Index [92]
- UniGene Cluster [93]
- RefSeq [94]
- LocusLink [94]
- Swiss-Prot [95]
- SOURCE [96]
- MGI (Mouse Genome Informatics) [97]
- ENSEMBL transcripts [98]
- GeneLynx [99]

Although, a large variety of gene related information is accessible through these databases, the annotation of not well characterized genes is moderate and information is limited. Hence for

the selected and analyzed genes (ESTs) further annotation processes were accomplished with a novel annotation and sequence analysis system (Annotator, Research Institute of Molecular Pathology, Vienna). At the first step for each EST sequence a corresponding protein sequence has to be found; the applied strategy is shown in Figure 6 in detail. For this purpose a Megablast search with parameters word length (w=70) and percent identity (p=95%) against different nucleotide databases (RefSeq, FANTOM, UniGene, NT, TIGR MGI) was carried out and the corresponding protein was extracted from the same or a related database entry. If the accession number was not available in the entry a Blastx search against the mouse proteome set from the International Protein Index (IPI) was conducted to extract amino acid sequences. For the remaining ESTs searches were repeated first with another compilation of the RefSeq (another build of RefSeq including the provisional and automated generated RefSeq records [94], and second using Blastn [100] instead of Megablast to search against the databases again. The most suitable hit (not always the best hit) was used to select the corresponding protein. Finally a Blastn search against the ENSEMBL mouse genome was performed and ESTs with long stretches ($>$100 bp) of not specified nucleotides (N) were excluded. Starting from the protein sequence within the annotator system several global domains were predicted: Low complexity regions with SEG, Pfam, Prosite and Prosite pattern with HMMER, Transmembrane domains (TOPPRED, DAS TM Filter, SAPS), coiled coil regions, GPI anchor, SignalP region, and a series of relevant short signals. Based on this predictions and the annotated hits of sequence homology searches in databases using Blastp, the selected ESTs were assigned new annotations.

2.7.2 Gene ontology

In the annotation process one has to consider that the nomenclature of genes is not consequent, that means often there exists several aliases and symbols for the same gene and there are no standardized names (except e.g. for yeast). This is not astonishing since historically genes were discovered in different processes or isoforms of the genes in different tissues were studied in many different laboratories all over the world and each time a new name had been assigned. To overcome this lack in consistent nomenclature the Gene Ontology^{TM}(GO) Consortium [56] is dedicated to produce a controlled vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing. The goal is not to define new names for a gene but providing three structured networks of defined terms to describe gene product attributes: molecular function, biological process, and cellular component. GO terms and the corresponding GO numbers are organized in structures called directed acyclic

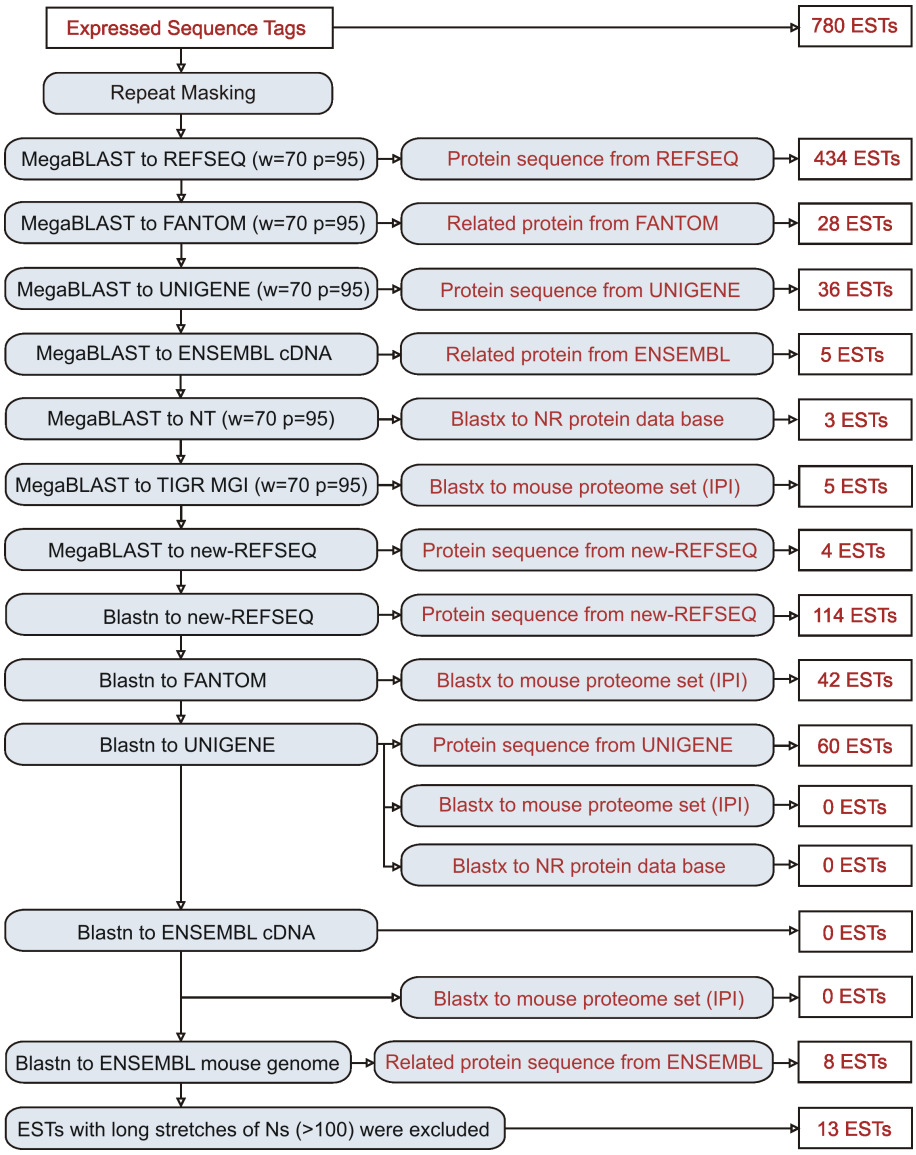

Figure 6: Strategy to search for protein sequences corresponding to selected ESTs. On the left, searches against nucleotide databases are displayed, in the middle selection of protein sequence, and on the right numbers of ESTs with assigned protein sequences

graphs (DAGs), which differ from hierarchies in that a 'child' (more specialized term) can have many 'parents' (less specialized terms). Use of GO in analysis of experimental data from high throughput methods enables integration of biological background data in a controlled manner. For each of the ESTs spotted on the focused mouse microarray the GO terms were derived from the SOURCE database so far they were available. Gene ontology for the selected ESTs was integrated and further analyzed.

2.8 Real Time RT-PCR

To confirm microarray results samples from the second time course of the microarray experiment were analyzed with Real Time RT-PCR. cDNA was synthesized from 2,5 μg of total RNA in 20 μl using random hexamers and SuperScript III reverse transcriptase (Invitrogen). The design of LUX^{TM}primers was done using the web service of Invitrogen (Table 1). Real time quantitative RT-PCR analyses for the genes described in Table 1 were performed starting with 50 ng of reverse transcribed total RNA, with 0.5X Platinum Quantitative PCR SuperMix-UDG®(Invitrogen), with a ROX reference dye, and with a 200 nM concentration of both LUX^{TM}labeled sense and antisense primers (Invitrogen) in a 25 μl reaction on an ABI PRISM 7000 Sequence Detection System instrument (Applied Biosystems). It is difficult to completely eliminate genomic DNA from RNA preparations. Therefore, a no amplification control (NAC) was included, which is a sample where the reverse transcriptase was omitted. Also included in the assays were no template controls (NTC). To measure PCR efficiency, serial dilutions of reverse transcribed RNA (0.24 pg to 23,8 ng) were amplified, and a line was obtained by plotting cycle threshold (C_T) values as a function of starting reverse transcribed RNA, the slope of which was used for efficiency calculation using the formula

$$E = 10^{|1/slope|} - 1. \tag{6}$$

Ribosomal 18 S RNA amplifications were used to account for variability in the initial quantities of cDNA. As calibrator the control sample (preconfluent stage) were used. The relative quantitation for any given gene (R) with respect to the calibrator was determined as follows:

$$R = (1 + E)^{-(\Delta C_{T,sample} - \Delta C_{T,calibrator})} \tag{7}$$

with

$$\Delta C_T = C_T - C_{T,18SrRNA}. \tag{8}$$

It was tested if the PCR efficiency for the genes of interest was comparable to the one of the internal reference (18S rRNA). Finally, the resulting ratios were compared to the normalized ratios from the second time series analysis.

Gene	Forward primer	Reverse primer
PPAR	caccaTGCGGAAGCCCTTTGGtG	GGGCGGTCTCCACTGAGAAT
LPL	AGCAGACGCGGGAAGAGATT	caaccaAGGTCTTGCTGCTGTGGTtG
c-myc	CCCTAGTGCTGCATGAGGAGA	cagcgTTGCTCTTCTTCAGAGTCGCtG
Cyclin A2	CAGAGCTGGCCTGAGTCATTG	gacctaGTGGCGCTTTGAGGTAGGtC
Decorin	TCATAGAACTGGGCGGCAAC	caccaaAGGGATCGCAGTTATGTTGGtG
Nur77	GGCTTCTTCAAGCGCACAGT	gtacatctCTTGTTTGCCAGGCAGATGtAC
BTEB1	TGGCTGTGGGAAAGTCTATGG	atacagAAGGGCCGTTCACCTGTAtG

Table 1: Sequences of forward and reverse primer for Real Time RT-PCR analysis to confirm microarray experiments for the given genes. Lower case letters (bases) on the 5' end indicate complement to the 3' end building a hairpin structure. Lower case t on 3' end indicates where the fluorophore is coupled.

2.9 Promoter analysis

Clustering methods of high-throughput expression profiling data provide us with the information about co-expression of genes in different conditions. The underlying assumption is that co-expressed genes are co-regulated or can share the same function (guilt by association).
A set of co-regulated genes usually share a similar set of regulatory motifs. This implies that in the regulatory regions of the genes, most notably within the promoter sequence, the same transcription factors or modules of transcription factors bind to the DNA and mediate the rate of transcription leading to similar expression profiles. With the advent of sequencing of whole genomes, computational methods arose to predict regulatory events and target genes for transcription factors. This none trivial task consists of identification of the promoter, transcription factor binding sites, combinatorial regulatory elements and transcription factor target genes.

2.9.1 Identification of promoters

The first step towards the identification of a transcription factor binding site is to find the promoter, which is defined here to be the upstream region proximal to a genes' transcription start site (TSS). Several approaches were based on the recognition of relatively conserved signals like TATA box, CCAAT box, other known transcription factor binding sites, or differences in the content like triplet base pairs around TSS, hexamer frequencies within a closed region or presence of CpG islands [101–103] for the identification of promoters. However, the prediction is limited and also a number of false positive promoters were predicted. Based on this

observations we follow the approach of aligning full-length mRNA(cDNA) with their counterparts genomic sequences and extract the upstream genomic sequence of the most 5' predicted exon. This approach can achieve over 80% accuracy in promoter localization [104–107]. Promoters with a length of 5000 bp upstream of human and mouse transcripts were derived from the PromoSer web service [108], that facilitates extraction of user specified regions around the transcription start site. As input the RefSeq and mRNA numbers, or if they were not available the EST accession numbers were used. Through mapping of full length cDNA clones (60.770 full-length cDNA mouse clones from the RIKEN project and 12.340 full-length human cDNA sequences from IMS) as well as entries from the eukaryotic promoter database [109] with BLAT [110] it is assured that there is an extension for mapping of RefSeq and mRNA GenBank sequences to assign the TSS correctly in the genomic sequence. Starting from the mRNA or RefSeq entries of selcted mouse transcripts homologs in human with RefSeq accession number were identifed using the HomoloGene database [93], which is based on reciprocal best Blast hits. Only those regions in the homolog promoters were considered, where the similarity was over 70%, identified using a sliding window of 50bp.

2.9.2 Analysis of regulatory elements in promoters

After promoter regions are identified, it is then possible to search for *cis*-regulatory elements computationally by screening genomic sequences for the presence of transcription factor binding site motifs that have already been identified. Because most transcription factors bind to short (5-25 bp) degenerate sequence motifs that occur very frequently in the genome a position weight matrix (PWM) is used to represent the binding specificity of these factors quantitatively. A collection of PWM for different transcription factors derived from binding sites in different organisms can be found in the TRANSFAC [111] and JASPAR database [112]. Only vertebrate PWMs, which are derived from more than 10 sequences were considered in the current analysis (116 PWMs in total). In order to find potential transcription factor binding sites weights $w_i(i)$ were calculated at each position i based on the information content [113]:

$$w_i(i) = \frac{100}{\ln 4}\Big(\sum_{b \in A,C,G,T} p_i(b) * \ln p_i(b) + \ln 4\Big) \tag{9}$$

The similarity score sim for a found motif was determined according to:

$$sim = \frac{\sum_{j=1}^{n} w_i(j) * score(b,j)}{\sum_{j=1}^{n} w_i(j) * \max(score(j))} \tag{10}$$

where j is the position $b \in A, C, G, T$, n the number of positions in the PWM, and $score$ is the frequency at a position for the nucleotide A,C,G, or T, respectively. Above a threshold of sim$>$0.85 the found motif in the promoter sequence was considered as putative binding site.

Gene	Organism	Binding sequence
CYP4A6/P450 IV	rabbit	CTGAACT AGGGCA A AGTTGA
CYP4A1/P450 IV	rat	TGAAACT AGGGTA A AGTTCA
L-fatty acid binding protein	rat	CAAATAT AGGCCA T AGGTCA*
3-hydroxy-3-methyl-glutaryl-CoA-synthase	rat	AAAAACT GGGCCA A AGGTCT*
Enoyl-CoA-hydratase	rat	CAAATGT AGGTAA T AGTTCA*
Malic enzyme	rat	GCATTCT GGGTCA A AGTTGA
PEPCK	rat	CACAACT GGGATA A AGGTCT
PEPCK	rat	ACTCCCA CGGCCA A AGGTCA*
Acyl-CoA oxidase	rat	GGGGACC AGGACA A AGGTCA*
Liver specific type 1 sugar transporter	rat	GCGTTAC AGGACA A AGGCCA
Malic enzyme	rat	GTGTTAG AGGGCA C AGGTCC*
Acyl-CoA oxidase	rat	GAGAGCA AGGTAG A AGGTCA*
Acyl-CoA synthetase	rat	GTCTTTC AGGGCA T CAGTCA*
Palmitoyl transferase fatty acid transport	mouse	AGGAAGT GGGGCA A AGGGCA
aP2 adipocyte lipid binding protein	mouse	TCTCTCT GGGTGA A ATGTGC*
aP2 adipocyte lipid binding protein	mouse	TCTTACT GGATCA G AGTTCA
c-Cbl-associating protein	mouse	TTGACAC AGGCTA A AGGTCA
Uncoupling protein 1	mouse	TTCAGTG TGGTCA A GGGTGA*
Apolipoprotein C-III	human	AGGGCGC TGGGCA A AGGTCA*
Acyl-CoA oxidase	human	AACTAGA AGGTCA G CTGTCA
Lipoprotein lipase	human	TCCGTCT GCCCTT T CCCCCT
Muscle-type carnitine palmitoyltransferase I (CPT I)	human	TGACCTT TTCCCT A CATTTG

Table 2: Experimental verified binding sites for PPARγ in genes from different organisms and different functions

2.9.3 *In silico* identification of potential PPAR target genes

Since PPARγ plays a predominant role in the process of adipogenesis there is great interest in predicting downstream target genes of PPARγ, which could be involved in the regulation of essential processes during adipocyte differentiation.

The concept was to extract experimentally verified binding sites for PPARγ in genes from different organisms and different functions (Figure 2) and build a model for the sequence specifity of the binding process of PPARγ promoters to the upstream activator sequence (UAS). PPARs bind neither as homodimer nor as monomer but strictly depend on the retinoid X receptor (RXR) as DNA-binding protein. The PPAR:RXR binding is characterized by the following properties: an extended 5'-half-site flanking region and an adenine as the spacing nucleotide between the two hexamers, giving the consensus sequence 5'-AWCT AGGNCA A AGGTCA-3'. A position weight matrix was derived by ClustalW alignment of the verified binding sites and counting the bases at each position of the binding sites including a flanking region.

2.10 Analysis of the gene expression data in context of relevant pathways

2.10.1 Network model for adipogenesis

In the last years many factors were identified, which either inhibit or activate adipocyte differentiation [114, 115]. However, the underlying mechanistic models of how some of these factors are regulating the differentiation process are only partly understood. It was shown that effects could be mediated via signaling pathways (Wnt-signaling [116]), by binding of transcription factors to the promoter of major regulators (binding of GATA-2 and GATA-3 to the PPARγ promoter [117]), post translational modifications (phosphorylation of Rb by cyclin-dependent kinase inhibitor p21 [118]), formation of inactive heterodimers (C/EBPβ and CHOP-10 [119]), or direct protein-protein interaction (transactivation of C/EBP by Rb [120]). The major question is how effects of these identified factors can be integrated into the well established transcriptional regulatory network consisting of members of the C/EBP family of transcription factors and PPARγ as described in the introduction (Chapter 1.1). Quite a number of review articles have summarized the current understanding of the adipogenesis process and how involved factors are related [28, 114, 115, 118, 121–139]. Based on this information the attemption was made to sketch a regulatory network model for adipogenesis (see Figure 7), well considering the inexactness of mapping complex relations to one inhibition or activation connection.

Whereas much effort has been directed toward understanding of the terminal stages of adipocyte

differentiation, the molecular mechanisms underlying the transition between cell proliferation and differentiation of preadipocyte remain largely elusive. It is likely that factors controlling the cell cycle could play a role in adipocyte differentiation. E2F, a cell cycle specific transcription factor when bound to DNA exists either as free E2F heterodimer with dimerization partner DP or associated in larger complexes containing members of the retinoblastoma family (pRb, p107, p130) and members of the cyclin/cyclin dependent kinase (cdk) protein families [28]. Upon re-entry in the cell cycle Rb and other retinoblastoma proteins are phosphorylated by cdk releasing the E2F complex, resulting in the activation of E2F targets and in promotion of cell cycle progression to S-phase [140]. After several rounds of DNA synthesis cyclin dependent kinase inhibitors (p21, p27, p18) are induced, which mediates the cell cycle exit [118]. It was shown that E2F1 induces PPARγ transcription during clonal expansion whereas E2F4, when associated to p107 or p130 represses PPARγ expression during terminal adipocyte differentiation [141]. Moreover, cyclin D1 is able to repress PPARγ through a pRb- and cdk-independent mechanism [142] and cdk inhibitor p21 expression can be stimulated by binding of C/EBPα [143]. Immediately within (2-4h) after induction C/EBPβ and -δ are expressed. At this stage C/EBPβ is unable to bind DNA and thus cannot activate the regulatory genes responsible for terminal differentiation. Only after a long lag period (10-12h) C/EBPβ aquires DNA binding activity. This occurs as the cells synchronously reenter the cell cycle, traverse the G1/S checkpoint and begin mitotic clonal expansion.

Since cAMP level enhancing agents, glucocorticoids and insulin are used for optimal differentiation important issues were raised how these agents influence this differentiation process or rather what are the mechanisms to regulate C/EBPβ and C/EBPδ. Glucocorticoids do not directly affect the appearance of C/EBPβ, while stimulating early adipogenic effects, but dexamethasone, a synthetic glucocorticoid, has been shown to induce C/EBPδ [7]. Furthermore, dexamethasone is down-regulating pref-1 which in turn inhibits adipocyte differentiation and is expressed highly in preadipocyte and is absent in mature adipocytes [144, 145]. Insulin, or the insulin like growth factor1 (IGF1) has been identified as the most potent inducer of clonal expansion [146] and the adipogenic action could be mediated via a down stream effector of insulin action (protein kinase B (PKB/Akt)) [129]. The phosphodiesterase inhibitor methylisobutylxanthine (MIX) increases the cAMP levels and leads to a rise of C/EBPβ [147]. The cAMP responsive transcription factor CREB has also been found to play a role in adipogenesis [148].

Whereas the role for PPARγ in adipocyte differentiation of preadipocytes is well documented, the function of PPARδ is still unclear. Activation of endogenous PPARδ in 3T3-L1 cells en-

hances the expression of PPARγ, but terminal differentiation is only modestly promoted [149], although it was previously shown that both PPARγ and adipose conversion were promoted by fatty acid treatment of 3T3-C2, with forced expression of PPARδ [150].

Although TGF-β promotes proliferation it has inhibitory effects on adipogenesis. This might be mediated by SMAD3, because overexpression of this transcription factor potently blocks differentiation. Although SMAD3 does not greatly influence the expression of C/EBPβ or C/EBPδ, SMAD3 inhibits induction of CEBPα and PPARγ [151]. Interestingly, FGF10 an important signaling molecule is required for the development of adipose tissue [152]. Prevention of FGF10 signaling inhibited adipocyte differentiation and in the absence of FGF10, expression of C/EBPβ was prevented.

A number of inflammatory cytokines, including TNFα interleukin 1 (IL-1), IL-6, IL-11, and interferon γ (IFN-γ) are also inhibitors of preadipocyte differentiation in vitro. Some of them act through their respective receptors on active members of the signal transducers and activators of transcription (STAT) family. STAT1, -5A and -5B are induced during differentiation and STAT3 and STAT6 genes are constitutively expressed [153,154]. In contrast, it was shown that STAT5A is adipogenic in nonprecursor cells [155]. Growth hormone (GH), which effects *in vitro* effects on cellular models of adipocyte differentiation have demonstrated an overall inhibitory effect by increasing cell proliferation and lipolysis, has been shown to decrease adiposity in vivo and inhibits the differentiation of primary preadipocytes in culture, although the in vivo effects might be more complex [156, 157]. The action of GH could be mediated via STAT signaling [115]. TNFα in turn acts through TNFα receptor 1 to inhibit adipogenesis, in part through sustained activation of the ERK pathway [158]. Differentiation of 3T3-L1 preadipocytes grown in serum free media has been reported to depend on epidermal growth factor (EGF) and platelet-derived growth factor (PDGF) [146]. However, it was shown that they inhibit differentiation of mouse, rat, and human preadipocytes, suggesting that the inhibitory effects of these growth factors depend on the origin, the state of development of the target preadipocytes, and culture conditions [114]. Macrophage colony-stimulating factor (MCSF) stimulates adipogenesis *in vitro* and *in vivo* and might play a physiological role in the induction of adipocyte hyperplasia, since it is rapidly down-regulated $invitro$ upon exposure to TNF-α [159].

Inhibitors of DNA binding 2 and 3 (Id2, Id3) are expressed in 3T3-F442A preadipocytes but suppressed in differentiated adipocytes. Ectopic expression of Id3 can inhibit adipogenesis,

not by binding directly to DNA but formatting of nonfunctional heterodimers with other bHLH proteins such as ADD1/SREBP1 [160, 161].

There are several transcriptional activators with critical roles in adipocyte cell growth and differentiation including HMGI(Y) [162], which is involved in the regulation of chromatin structure and function, FOXC2 [163], FOXO1 [164], Krüppel-like factor 2 (KLF2) [165], Olf-1/early B-cell marker (O/E-1) [166], and nuclear factor of activated T cells (NFAT) [167], which was shown to physically interact with PPARγ.

Polyunsaturated fatty acids, which are less efficient in increasing the number of adipocytes in vivo, are more effective than saturated fatty acids in stimulating preadipocyte differentiation in culture, presumably mediated through the ability to act as ligands or precursor of ligands for PPARγ. In contrast to the general effect of polyunsaturated fatty acids prostaglandins have highly specific effects on preadipocyte differentiation. Prostacyclin (PGI_2) is a major metabolite of arachidonic acid in adipose tissue and stimulates adipogenesis [168]. On the other hand, prostaglandin $F_{2\alpha}$ ($PGF_{2\alpha}$) inhibits the differentiation of preadipocytes in culture [169]. Further it was shown that prostaglandin D_2 and its 15-deoxy-J_2 derivative may be endogenous ligands for PPARγ and therefore activate adipogenesis [170].

2.10.2 Genomics of lipid associated disorder database (GOLD.db)

It is apparent that large scale analysis of adipogenesis will not give comprehensive answers for the development of obesity and related disorders in general and dysregulation of neutral lipid hoemeostasis in particular although some regulators seem to play an overall important role. The analyses of the current study were embedded in a broader context. In order to provide a reference for pathways and information of the relevant genes and proteins in an efficiently organized way, the Genomics Of Lipid-associated Disorder database (GOLD.db) was created. The GOLD.db integrates disparate information of the function and properties of genes and their protein products that are particularly relevant to biology, diagnosis, treatment, and the prevention of lipid-associated disorders. The main focus was to provide biological pathway image maps and visual pathway information. For each element in the pathway, specific information exists including structured information about a gene, protein and its 3D-structure, gene regulation, function, literature, and links. For this purpose a pathway editor was developed [171]. This tool provides the possibility to draw elements - typically representing a gene as a part of a pathway - and the connection between those elements. The benefit of this tool is that information can be

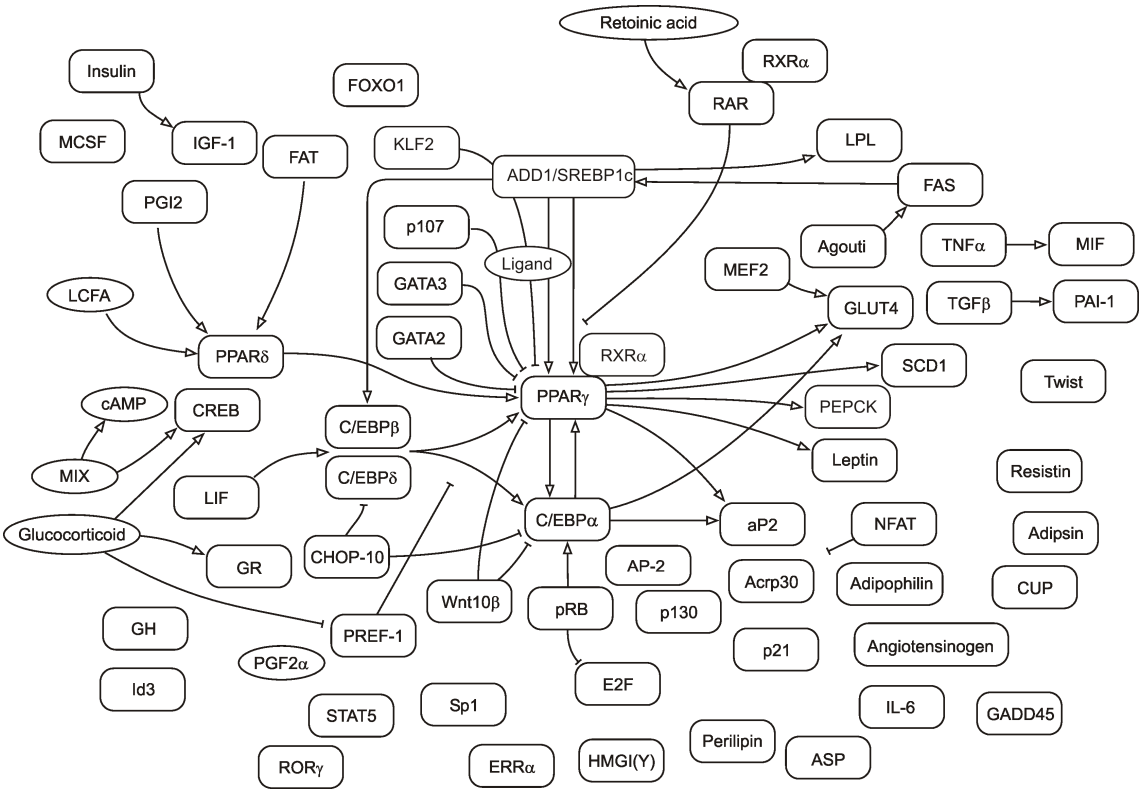

Figure 7: Regulatory network for adipogenesis

appended to each element via an input mask. This information can be accessed by clicking on the corresponding element in the image map, which was saved and uploaded to the web page. Currently annotated pathways are the insulin signaling pathway, the IGF-I pathway, and the adipogenesis regulatory network. Other pathways of lipid metabolism will follow. Available KEGG pathways can also be adapted with the pathway editor based on provided XML files [57] and uploaded in the same way. Several relevant KEGG pathways for different organisms are already provided. The pathway editor is executable as stand alone program and available at http://genome.tugraz.at. Besides including protocols, videos, references, and links to relevant genomic resources, a relational database for the clone resources used for the microarray experiments were implemented. Information about the vector, the sequence and length of the insert, primers for PCR amplification, tissue, organism, accession number, library, container, storage information, date and experimentator is stored.

The GOLD.db was implemented in Java (http://java.sun.com/) technology. Hence, the pathway editor as well as the web application are platform independent. The web application of

GOLD.db was build in Java Servlets and JavaServer Pages technology based on the Model-View-Controller Architecture. For the implementation, the freely available struts framework (http://jakarta.apache.org/struts/) was used. This code can be easily deployed in any Servlet Container. As Servlet Container Tomcat and as database management system Oracle 9i was used. The interface between the Java and the database management system was established using Java database connectivity (JDBC) 2.0. For additional storage and communication between the pathway editor components, the markup language XML containing structured, human readable information, was used.

2.10.3 Mapping of gene expression data to pathways

Through the integration of several types of information deeper insights into the molecular mechanisms and biological processes can be gained than just by the analysis of one type of experimental results. GOLD.db provides the possibility to map gene expression data (for instance results of microarray studies) to the corresponding elements of the available pathways similar to previous efforts [58]. Either an individual or a provided gene expression data set can be used to visualize the gene expression at different experimental conditions sequentially in the context of the pathways. If an element (gene) of the pathway is included in the data set, the related symbol in the image map is color coded according to the relative gene expression or the log ratio in two color microarray experiments, respectively. As key for the mapped relation the RefSeq numbers [94] are used. Hence, only those elements in the data set are mapped, where the RefSeq number in the data set is specified. For the KEGG pathways each element classified by the enzyme classification number (EC) is virtually subdivided into different corresponding RefSeq entries, since one EC is represented by one or more RefSeq entries. Large scale expression data from previous studies related to obesity and diabetes are provided together with the results from the current study. The interrelation of genes in the context of the pathways were studied.

2.11 Strategy for building a regulatory gene interaction

The global gene expression pattern is the result of the collective behavior of individual regulatory pathways. In such highly interconnected cellular signaling networks, gene function depends on its cellular signaling, thus understanding the network as a whole is essential [172]. However, dynamic systems with large numbers of variables present a difficult mathematical problem. One way to make progress in understanding the principles of network behavior is to radically simplify the individual molecular interactions and focus on the collective outcome.

There are several approaches for the reverse engineering problem, that is given an amount of data what can be deduced about the underlying regulatory network: Boolean networks [173], Bayesian networks [174], and Differential equations [175]. However, there is limitation in applying these methods to common microarray data sets. The problem lies in the dimensionality of the data, the few numbers of experimental conditions in comparison to the number of genes or the needed data for a given connectivity, which refers to the number of regulatory inputs per gene. Therefore, a novel iterative approach based on mutual information and correlation is proposed to evaluate gene-gene associations. Although this model will imply little about actual molecular mechanism involved, much helpful information will be gained on genes critical for the biological process. New associations can be derived in comparison to other analysis techniques such as cluster analysis since the information content of the gene expression profiles is used instead of similarity measurements of the expression profiles.

2.11.1 Discretization

The current approach was using mutual information, which in turn is based on the entropy (information content within a individual gene expression profile). The entropy was computed using discrete probabilities and since gene expressions were measured on a continuous scale, expression profiles had to be discretized. The idea was to use an iterative approach, where thresholds for the discretization are increased at each step. Log ratios were assigned 1 for values greater than 1, 0 for values between -1 and 1, and -1 for values lower than -1. Genes showing flat profiles over all time points were filtered out, since it was assumed that they do not participate in regulatory processes. Genes with the identical profile after discretization were merged to clusters. At the next iteration thresholds for the log ratios were increased from +1/-1 to +1.5/-1.5 and for the following iterations accordingly.

2.11.2 Iterative approach based on mutual information and correlation

The Shannon Entropy provides us with the information content of the expression profile for gene A

$$H(A) = -\sum_{i=1}^{n} p(x_i) log_2(p(x_i)) \tag{11}$$

where x_i is the gene expression level at time point i and $p(x_i)$ the probability of the occurrence of the discrete gene expression level x_i at one of the n time points in the profile. Higher entropy for a gene means that its expression levels are more randomly distributed. The mutual informa-

tion MI is a measure of the additional information known about one expression pattern A when given another B and was applied previously for the construction of relevance networks [176]:

$$MI(A,B) = H(A) - H(A|B) = H(A) + H(B) - H(A,B) \tag{12}$$

Mutual information can also be calculated by subtracting the entropy of the joint expression profiles from the individual gene entropies. A mutual information of zero means that the joint distribution of expression values holds no more information than the genes considered separately. A higher mutual information between two genes means that one gene is non-randomly associated with the other. In this way, mutual information can be used as a metric between two genes related to their degree of independence. We hypothesize that the higher mutual information is between two genes, the more likely they have a biological relationship.
The REVEAL algorithm [177] was used to look for any profile X for a given profile Y where

$$MI(Y,X) = H(Y). \tag{13}$$

In these cases X determines Y exactly since $H(Y|X) = 0$ and the genes were regarded as related. Additionally, the correlation matrix was calculated for the centroids of the clusters based on the Pearson correlation coefficient

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{14}$$

where x_i refers to the log ratio in the expression profile X at the time point i and y_i accordingly in the expression profile Y. A threshold of r $>$0.63 was defined based on the type II error and the significance level of 0.05. Only those related genes (clusters) with a correlation coefficient above the threshold were combined to a subnetwork. Subnetworks with more than 20 non-flat expression profiles were considered for further iterations. The algorithm was implemented in the R statistical computing language; subnetworks were visualized using the network visualization system Osprey [178].

3 Results

3.1 3T3-L1 cell differentiation

The phenotypic changes of 3T3-L1 cells during the differentiation from preadipocytes to mature adipocytes become apparent by microscopic images taken at different timepoints in one differentiation experiment (figure 8). During the growth phase cells of preadipocyte lines as well as primary preadipocytes are morphologically similar to fibroblasts. At confluence (represented by day 0 in Figure 8) induction of differentiation by appropriate treatment leads to drastic cell shape changes. The preadipocyte converts to a spherical shape, accumulates lipid droplets, and progressively acquires the morphological and biochemical characteristics of the mature white adipocyte.

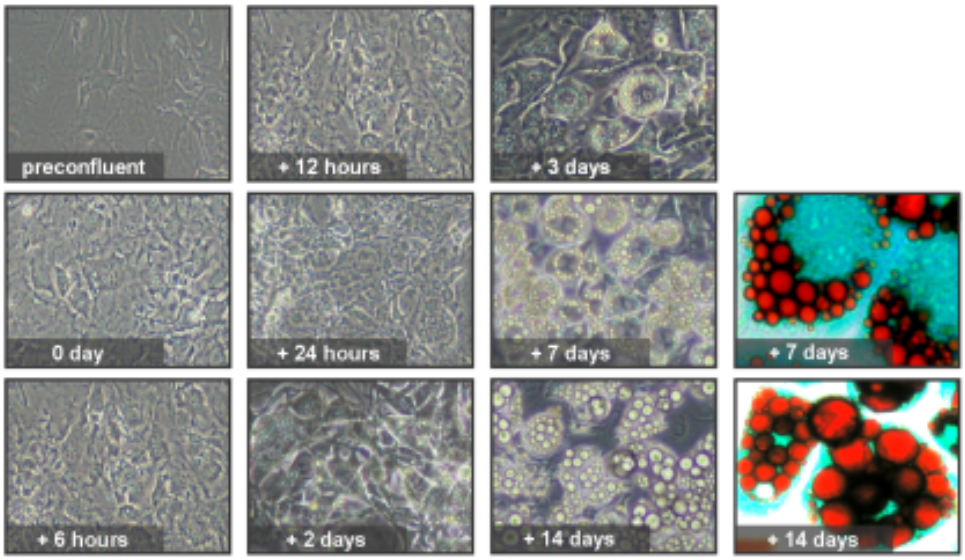

Figure 8: Images of the cells during the differentiation at different time points, where RNA was extracted. On the right Oil-red-O staining of lipid droplets and light green counter staining of cytoplasm are shown for day 7 and day 14 in the differentiation course

Although cells from preadipocyte cell lines like the 3T3-F442A are developing adipose tissue after being implanted subcutanously into athymic mice, the inclosed lipid droplets as indicated by Oil-red-O staining are not merged to one large drop as known for white adipose tissue. This indicates that not the same mature status of the fat cells is achieved as in vivo. First inclusion of lipid droplets appeared 2-3 days after treatment. Remarkable was that fat cells tended to

differentiate starting from initial clusters before they cover most of the bottom of the culture dish, suggesting the occurrence of cell-to-cell communications.

3.2 Results and quality of the microarray experiments

After the experiments to study the differentiation process of 3T3-L1 cells using a 27.648 element focused microarrays were performed, several steps in the analysis process were considered. A summary of the relevant steps is given in Figure 9. After filtering, normalization, and averaging over 3 experiments data were screened for genes which are more than 2 fold up- or downregulated in at least 4 time points. 780 genes (ESTs) out of the 14.368 elements, which show a complete profile over all timepoints were considered for further analysis. Since thousands of elements are analyzed in parallel it is very important to check the overall quality of the microarray data.

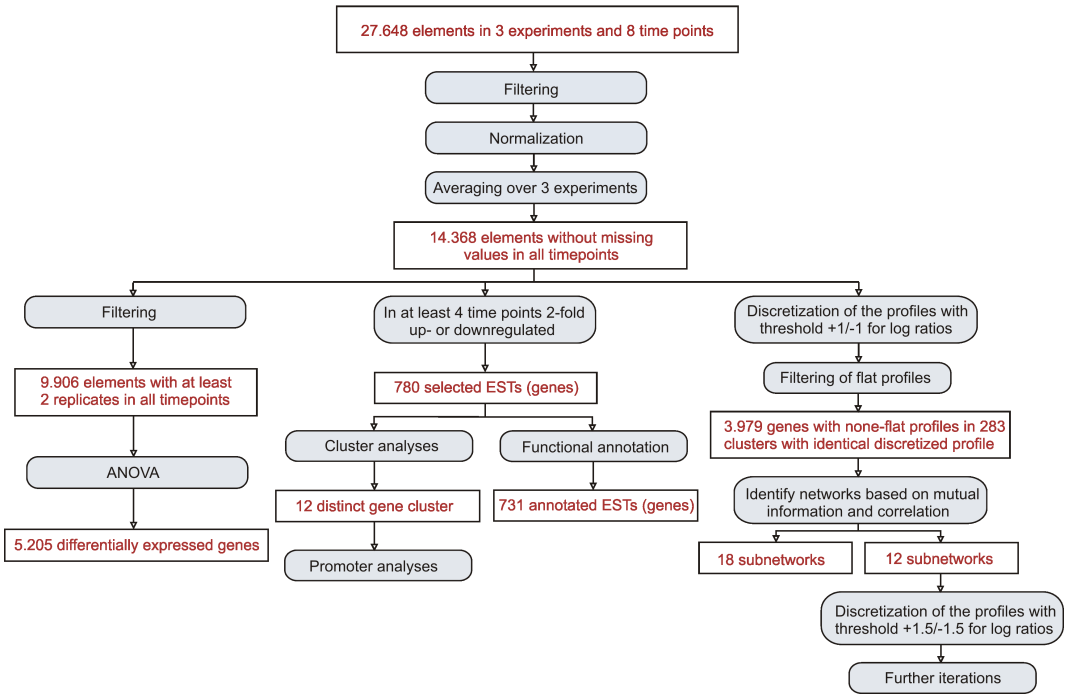

Figure 9: Computational analysis

As a representative example the hybridization data from the third experiment at day 7 versus the reference from preconfluent stage were used to show the consistency and quality of the data in several ways (figure 10). The consistency between technical replicates (dye swap) became evident by analyzing the scatter plot for the ratios of the dye swapped pair ($r^2 = 0.987$). It

turned out that after normalization there was no intensity dependency of the log ratios, the distribution of the log ratios was centered around 0 and in a certain range related to a normal distribution.

Three unrelated adipocyte differentiation experiments were performed. In order to keep the experiments totally independent for each experiment the cells were grown separately till 80% confluency and reference RNA was isolated for each experiment. To study differences in gene expression levels between this reference RNA from the 3 different time points microarray experiments in loop design were carried out. The consistency in gene expression levels is apparent in the scatter plots in Figure 11.

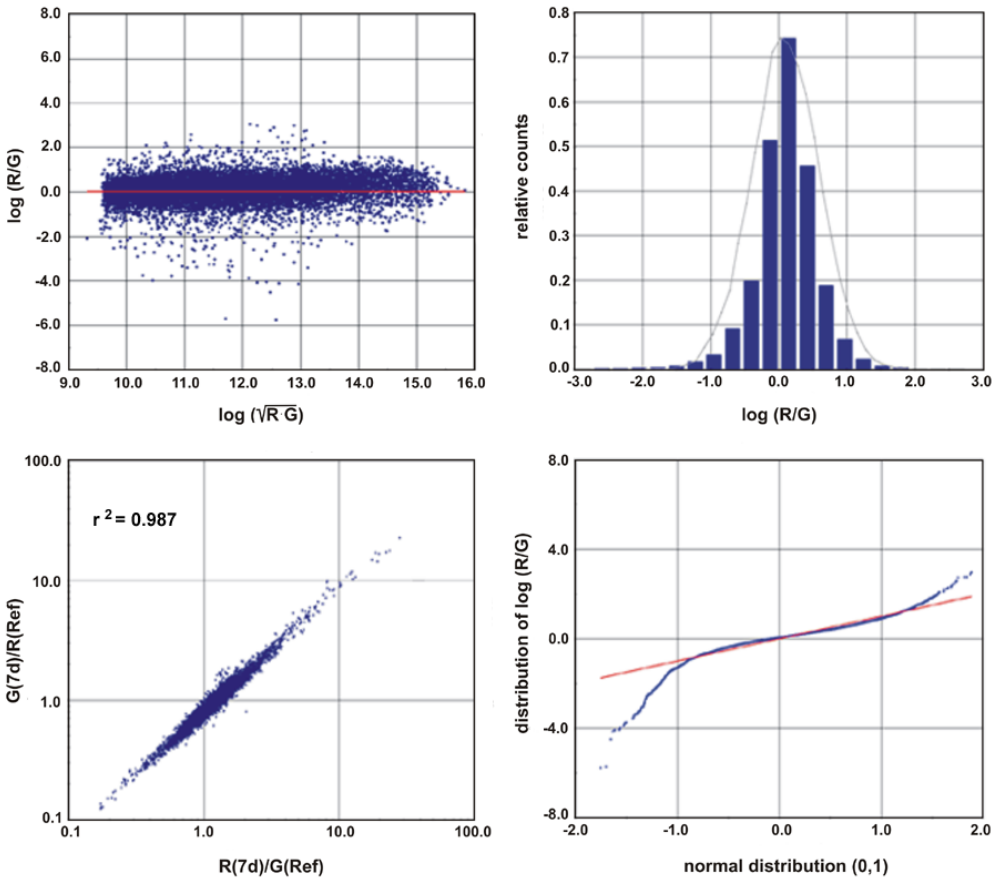

Figure 10: Visualization of the distribution and the quality of the microarray data after normalization. As representative example the results of 3rd experiment at time point 7d are shown: MA-plot (top left), histogram for log ratios (top right), comparison of technical replicates (dye swap) (bottom left), and QQ-plot to check normality of the data (bottom right)

For the one-way analysis of variance (ANOVA) only those 9.902 genes (ESTs) were considered which had no missing values in at least 2 out of the 3 experiments. 5.205 genes (ESTs) were identified by ANOVA as significantly differentially expressed. The identified candidates and their relative expression levels ($\log_2$ ratios) can be visualized at the supplementary webpage (http://genome.tugraz.at/adipocyte). Although this selection is based on statistical significance, the number of remaining genes is still very high and analyses would get complex. Therefore a more stringent criteria were applied and clustering, promoter analyses, and functional annotation were performed,which resulted in 780 genes only. For the reverse engineering approach the profiles of all 14.368 genes with a complete profile were discretized and initially used in the iterative approach.

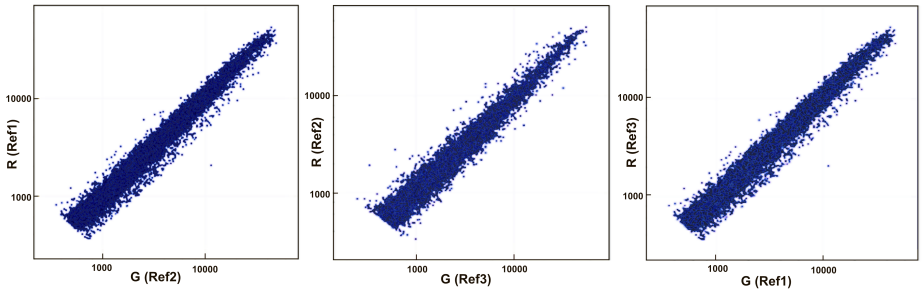

Figure 11: Consistency of gene expression levels from the reference of the 3 independent experiments. Scatterplot of intensities from the reference of the 1^{st} experiment versus the 2^{nd} experiment (left) and reference from 2^{nd} experiment versus 3^{rd} experiment (middle) and reference from 3^{rd} experiment versus 1^{st} experiment (right)

3.3 Clustering genes according to their expression profiles

The data were analyzed by k-means clustering, which groups genes based on the similarity of their patterns of gene expression. This cluster analysis indicated that the selected 780 genes can be grouped most parsimoniously into 12 temporally distinct patterns, each containing between 23 and 143 genes (Figure 13), suggesting that the regulation of adipogenesis may be more complex than previously assumed. Genes in 4 clusters are mostly upregulated, and genes in 8 clusters are mostly downregulated during adipogenesis. More than half of the genes were not described before to be involved in adipocyte differentiation, and also a number of new transcriptional regulators could be identified. To focus on some aspects of the results, only a subset of genes is shown and discussed here. All the data are available online

(http://genome.tugraz.at/adipocyte). To validate the k-means clustering, principal component analysis was performed and the first 3 principal components were visualized in the three dimensional space. Visual separation of the clusters in this view was achieved by coloring of the genes according to the k-means clustering. As apparent in Figure 12 the genes are clustered together and clusters are clearly separated in space implicating reliability of k-means clustering.

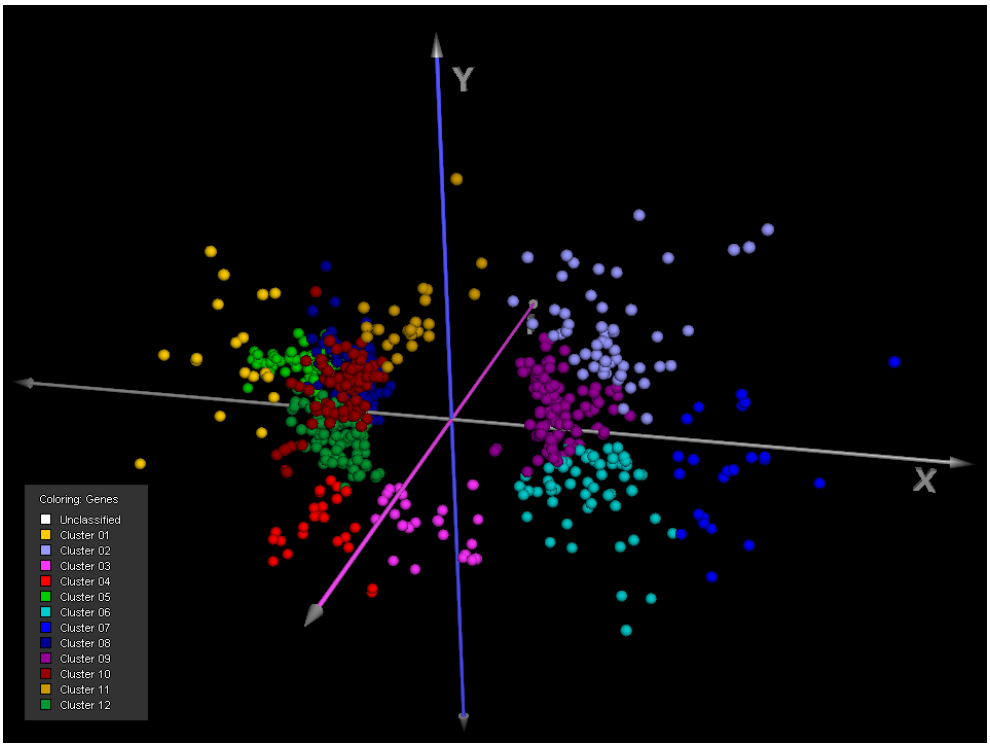

Figure 12: 3D-visualization of principal component analysis (genes are colored according to previous k-means clustering)

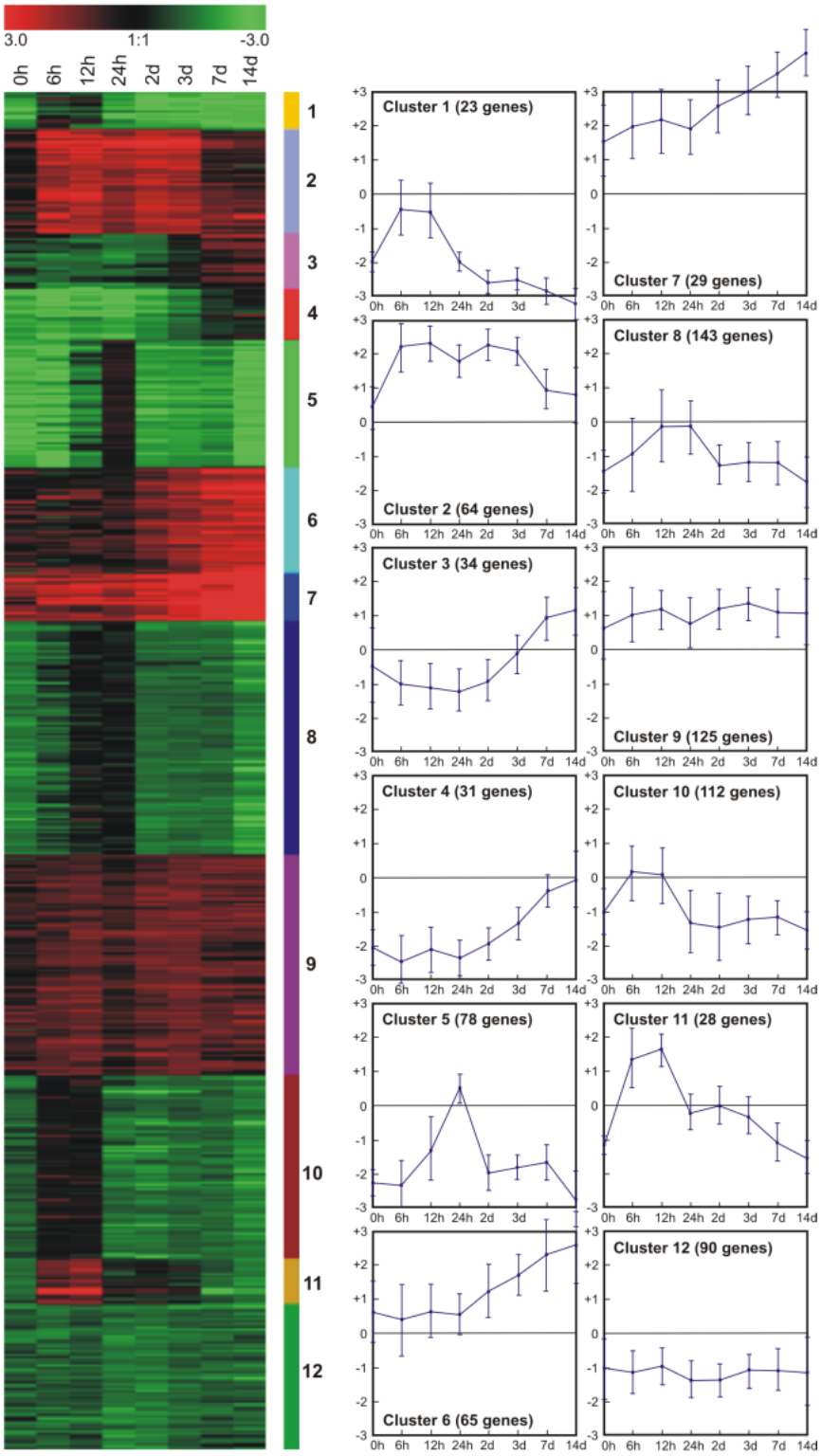

Figure 13: Results of k-means clustering (k=12)

3.3.1 Genes defining the adipocyte phenotype

The adipocyte phenotype can be defined by the induction of genes in the late phase of differentiation. In the current experiment clusters comprising genes, which are more abundant in the late time points (3d, 7d, 14d) than in the early phases, are relevant. Whereas genes of cluster 3 and cluster 4 are downregulated during the earlier time points of the differentiation course, in cluster 6 genes are in general slightly and in cluster 7 highly upregulated. However, genes of all 4 clusters were induced in the later phase. Microarray analysis indicated that the key transcription factors PPARγ2, C/EBPα and SREBP-1c were highly upregulated. Many genes that are markers of the differentiated adipocyte increased in parallel with these factors. These included known gene targets of either of these factors including lipoprotein lipase (LPL), c-Cbl-associated protein(CAP), Stearoyl-CoA desaturase 1 (SCD1), Carnitine palmitoyltransferase II (CPT II), which are grouped in cluster 6 (Figure 14). Since PPARγ2 plays a central role in the adipocyte differentiation process, genes with similar regulatory profile were identified. For this purpose the 780 selected genes were normalized (each gene expression profile was mean centered and devided by the standard deviation) and sorted according to similarity of the PPARγ2 profile (Figure 15). A number of genes involved in lipid and fatty acid metabolism showed similar regulatory profiles, most of them could also be found in cluster 6: acetyl-Coenzyme A dehydrogenase, pyruvate carboxylase, insulin-induced gene 1 (INSIG-1), scavenger receptor class B member I (SR-BI), low density lipoprotein receptor. Some genes of cluster 3 and 4 also show similarity to the regulatory profile of PPARγ2 and many of them are involved in steroid metabolism (like lanosterol synthase, 3-hydroxy-3-methylglutaryl-Coenzyme A synthase, hydroxysteroid (17-beta) dehydrogenase 7, farnesyl diphosphate synthase, URB protein). Malic Enzyme, an enzyme in fatty acid synthesis, which has a PPAR response element in the promoter was also present in cluster 3. HMG-CoA synthase, which is a known target gene of PPARγ and has been proposed as control site of ketogenesis, is highly upregulated in the late phase of differentiation. Interestingly, the signaling molecule fibroblast growth factor 10 (FGF-10) is highly upregulated (see cluster 6), which confirms previous studies [152], where it was shown that FGF-10 is required in the development of white adipose tissue and necessary for maintaining the abundance of C/EBPβ and consequently contributes to the progress of the adipocyte differentiation program in 3T3-L1 cells. In contrast to previous observations the signal transducer and activator of transcription 6 (STAT6) was not constitutively expressed, but induced during differentiation like it was shown for STAT1 and STAT5. A functional relation between genes in cluster 7, which are highly expressed in every time point and even more at the end can not eas-

ily deduced, although many of them are located extracellular. For example, osteoblast specific factor 2 and the proteoglycan decorin, which seems to play a role in the extracellular matrix, are bone related factors whereas angiotensinogen, which is known to be secreted by adipocytes, has a different function and is involved in regulation of blood pressure.

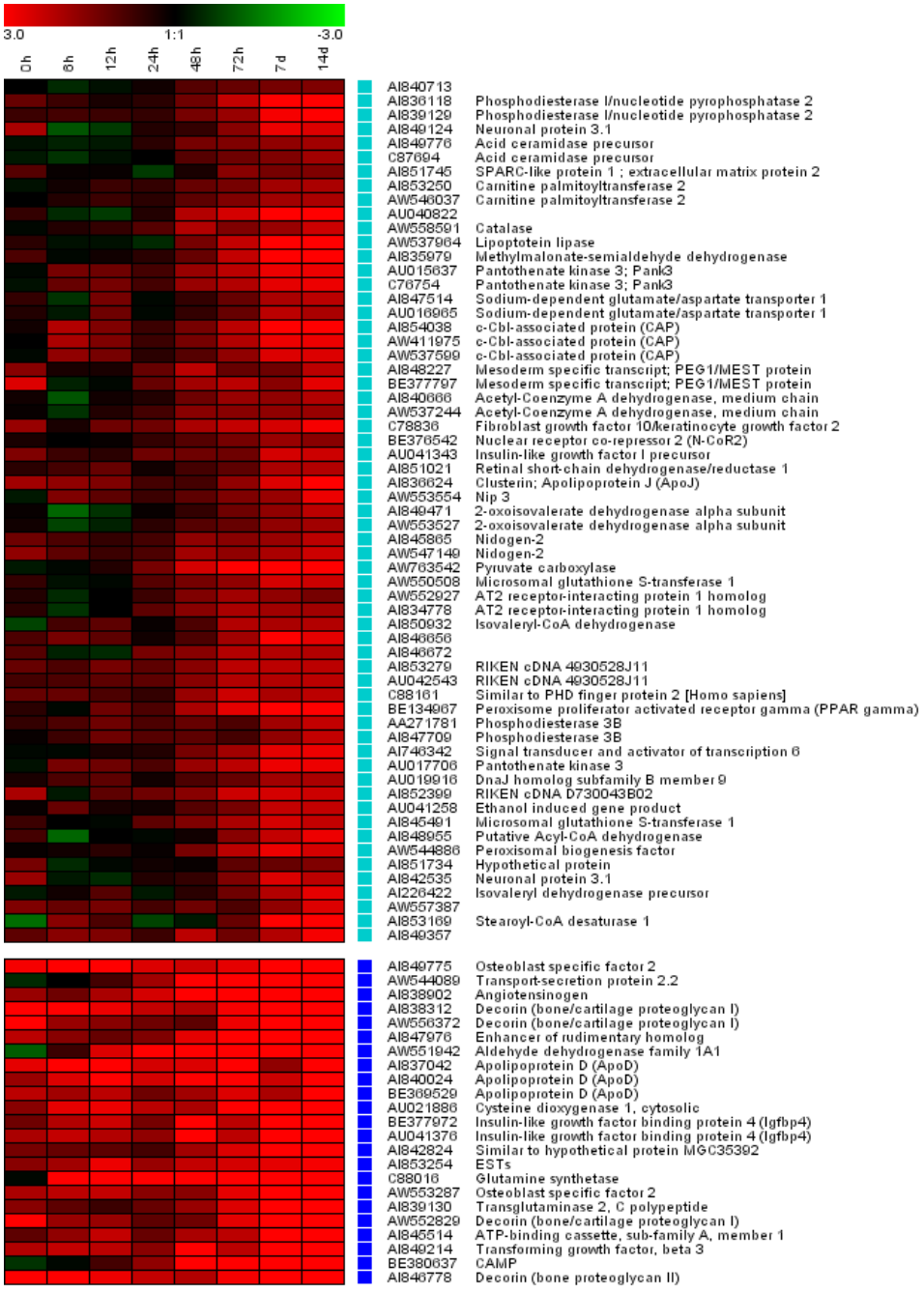

Figure 14: Genes and their relative expression levels of cluster 6 (top) and cluster 7 (bottom)

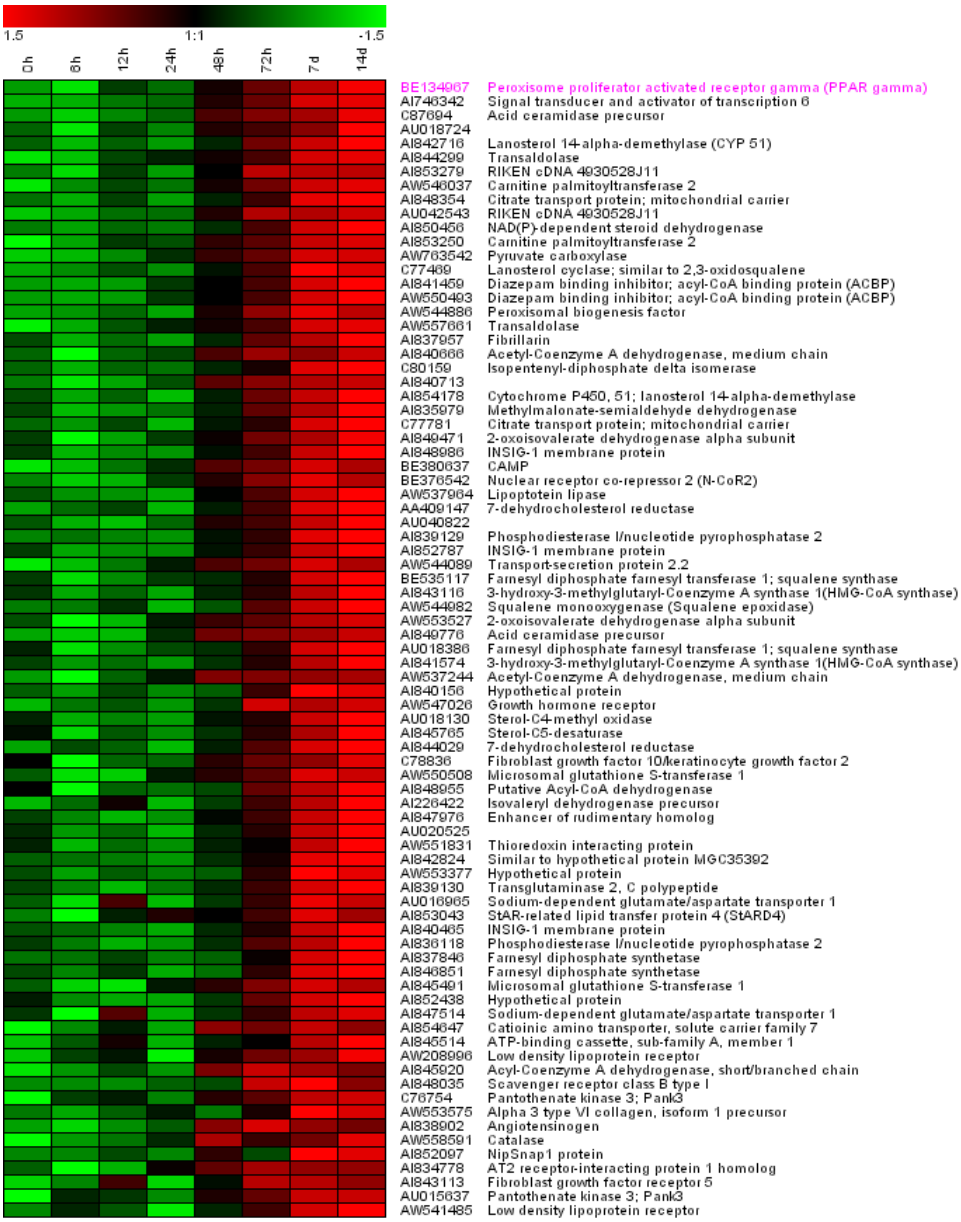

Figure 15: Normalized genes sorted by similarity to the profile of PPARγ

Controversially, the transforming growth factor β 1 (TGFβ1) is not secreted and activated in mature adipocyte [151], current results indicate, however, that TGFβ3 is highly induced. TGFβ3 seems therefore not to have the same inhibitory effects on adipogenesis as TGFβ1 and the proper function in this context has to be elucidated. Moreover, the ATP-binding cassette, subfamily A, member 1 (ABCA1) and apolipoprotein D (apoD), a member of the lipocalin superfamily of carrier proteins that transport small hydrophobic molecules [179], are highly induced

over all time points during the differentiation course. Whereas ABCA1 is an established target for the liver X receptor (LXR), apoD was identified as LXR responsive gene $invitro$ and $invivo$ effected by direct binding of LXR/RXR heterodimer to the promoter just recently [180]. LXR in turn is known to play a role in the modulation of lipid metabolism in adipocytes but is not a regulator of adipocyte differentiation.

3.3.2 Cell cycle related genes during mitotic clonal expansion

Cell proliferation and differentiation are mutually exclusive events. However, a close relationship has been established between both cell processes early during the adipocyte differentiation program. Reentry into the cell cycle of growth arrested preadipocytes is known as the clonal expansion phase. Growth arrested preadipocytes undergo several rounds of cell cycle before terminal differentiation into adipocytes.

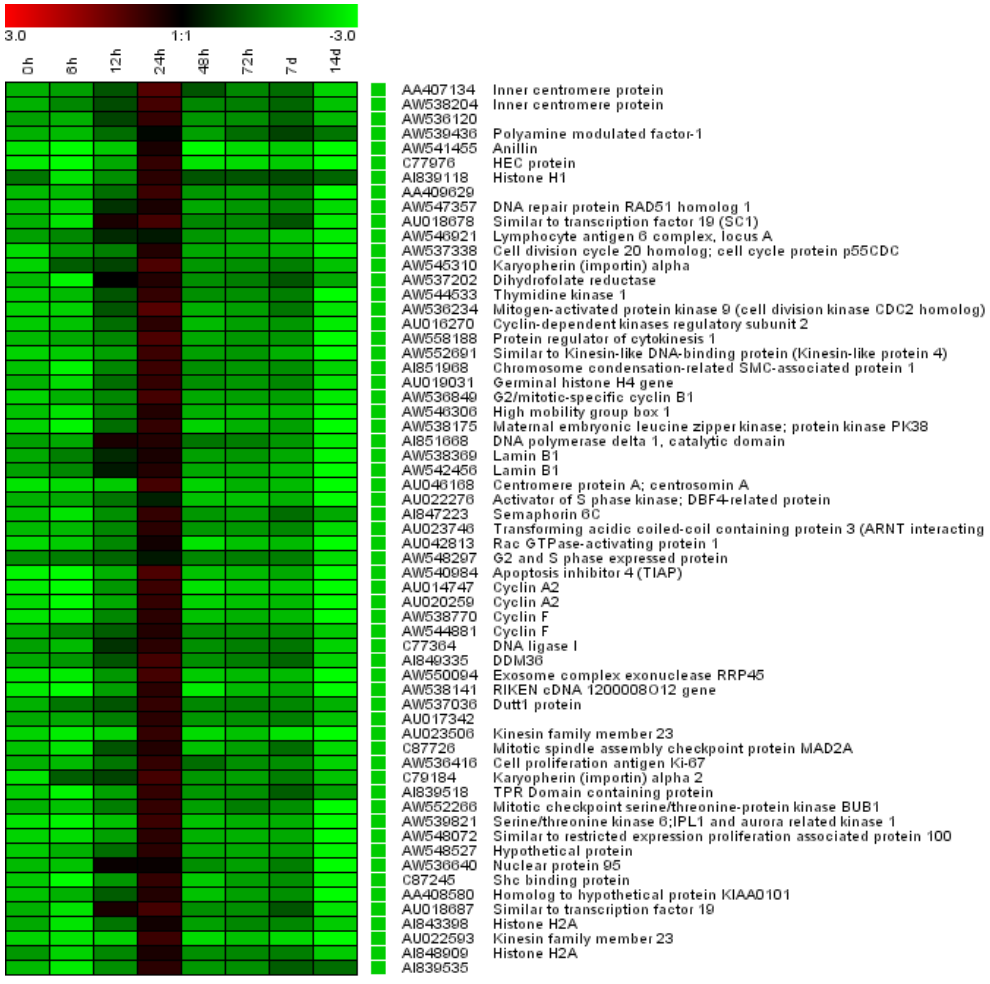

Figure 16: Genes and their relative expression levels of cluster 5

Most genes of cluster 5 and some of cluster 8 mirror the cell cycle events in the early phases. The expression profiles show steady downregulation during the whole process and a sharp upregulation of the relevant genes at time point 24h after hormonal induction.

While regulating cell cycle, progression through S-phase is partially regulated by cdk2/cyclinA complexes and subsequently, cdc2/cyclin B1 complexes modulate S/M transitition [181]. In fact cyclin A2, cyclin B1, and cyclin F was shown to be upregulated around 24h (Figure 16). Further evidence of these cell cycle events were given by the observation that histone H2A, histone H1, histone H4 are also induced at the same time frame. The inner centromere protein (INCENP), which builds a complex together with Aurora-B, is required for a number of mitotic events, such as accurate chromosome segregation and completion of cytokinesis [182], and showed a similar expression course. Interestingly, the cell proliferation antigen Ki67 was also shown (Figure 16) that it shared expression characteristics with the other cell cycle specific molecules. The Ki67 protein (pKi67), a diagnostic marker for several types of cancer, shares structural similarities with other proteins to be involved in cell cycle regulation, however, little progress was made in the last years in understanding its actual function [183]. Metallothionein 1 (MT1) as well as MT2 are induced after induction and downregulated again after 24h. This is consistent with previous findings in 3T3-L1 cells: that the initial burst of proliferation is controlled by the translocation of zinc metallothionein into the cell nucleus where zinc is donated to proteins that are involved in the G0/G1 to S transition [184]. The resulting expression profiles are similar to the total cellular content of the MT protein, as was previously shown [184]. There are also a number of molecules, like kinesin family member 23, nuclear protein 95, dutt1 protein, karyopherin (importin) alpha 2, activator of S phase kinase, mitotic checkpoint serine/threonine-protein kinase BUB1, mitogen-activated protein kinase 9, which are grouped together in cluster 5 and whose function is related to the cell cycle processes.

3.3.3 Transcriptional factors regulated during differentiation

A number of transcription factors, transcriptional regulators, and signaling molecules were differentially expressed during adipocyte differentiation. Whereas many of the terminal highly upregulated transcription factors are previously studied and identified as essential during adipocyte differentiation, among them the key players PPARγ2, C/EBPα and SREBP-1c and other factors like Twist homolog [114], X-box binding protein [22] or FOXO-1 [164] some factors with different expression profiles are not known to play a role in adipogenesis. The kruppel-like factors, also known as basic transcription element binding proteins (BTEB), seem to be relevant in the context of adipocyte differentiation, albeit with different functions. Kruppel-like factor 2 (KLF2) was shown to inhibit PPARγ expression and to be a negative regulator of adipocyte differentiation, whereas KLF15 to induce adipocyte maturation [165]. In the current study KLF9

(BTEB1), KLF5 (BTEB2) were upregulated in the intermediate phases of adipocyte differentiation, however, KLF4 was downregulated. The glucocorticoid-induced leucine zipper functions as transcriptional repressor (of PPARγ2) and antagonizes glucocorticoid induced adipogenesis [185, 186]. This is consistent with the observation that GILZ is highly upregulated during the first two days, during that time span dexamethasone is added to the differentiation cocktail, and downregulated at the end of differentiation, when PPARγ is highly induced.

Gene expression analyses indicated also that a large group of DNA binding inhibitors (Id genes) and high mobility group proteins (HMG genes) as well as the C/EBP homologous protein 10 (CHOP10), another type of transcriptional inactivator by forming of none functional heterodimers, were downregulated with distinct kinetic profiles during differentiation (Fig. 17 lower panel). The reciprocal regulation of adipogenesis by c-myc and C/EBPα is known for a long time: it was shown that expression of c-myc prohibited the normal induction of C/EBPα and prevented adipogenesis [187]. This was confirmed by the expression profile of C/EBPα and c-myc, respectively. C-myc is upregulated very early in the differentiation, when C/EBPα is not differentially expressed and during the end the prevented adipogenesis C/EBPα is highly induced, whereas c-myc is downregulated. Interestingly, many of the regulatory factors are down regulated during the mitotic clonal expansion around 24h after induction in the time series, indicating that they are not involved in cell cycle processes.

The orphan nuclear receptor NUR/77 was highly upregulated with an distinct expression profile. Its role in the process of adipocyte differentiation was not elucidated so far. NUR/77 turned out to be an example for a promising candidate, since nuclear receptors per se are very promising candidates for follow up studies and as potential drug targets. They combine DNA binding activity and receptor activity for specific ligands. Moreover, it was shown that NURR1, a member of the same family of nuclear receptors, exhibits similar expression profile in previous studies [24]. There are also a number of trancription factors (see Figure 17), which also may play a potential role in activating or inhibiting the adipogenesis process, predestined for further investigations.

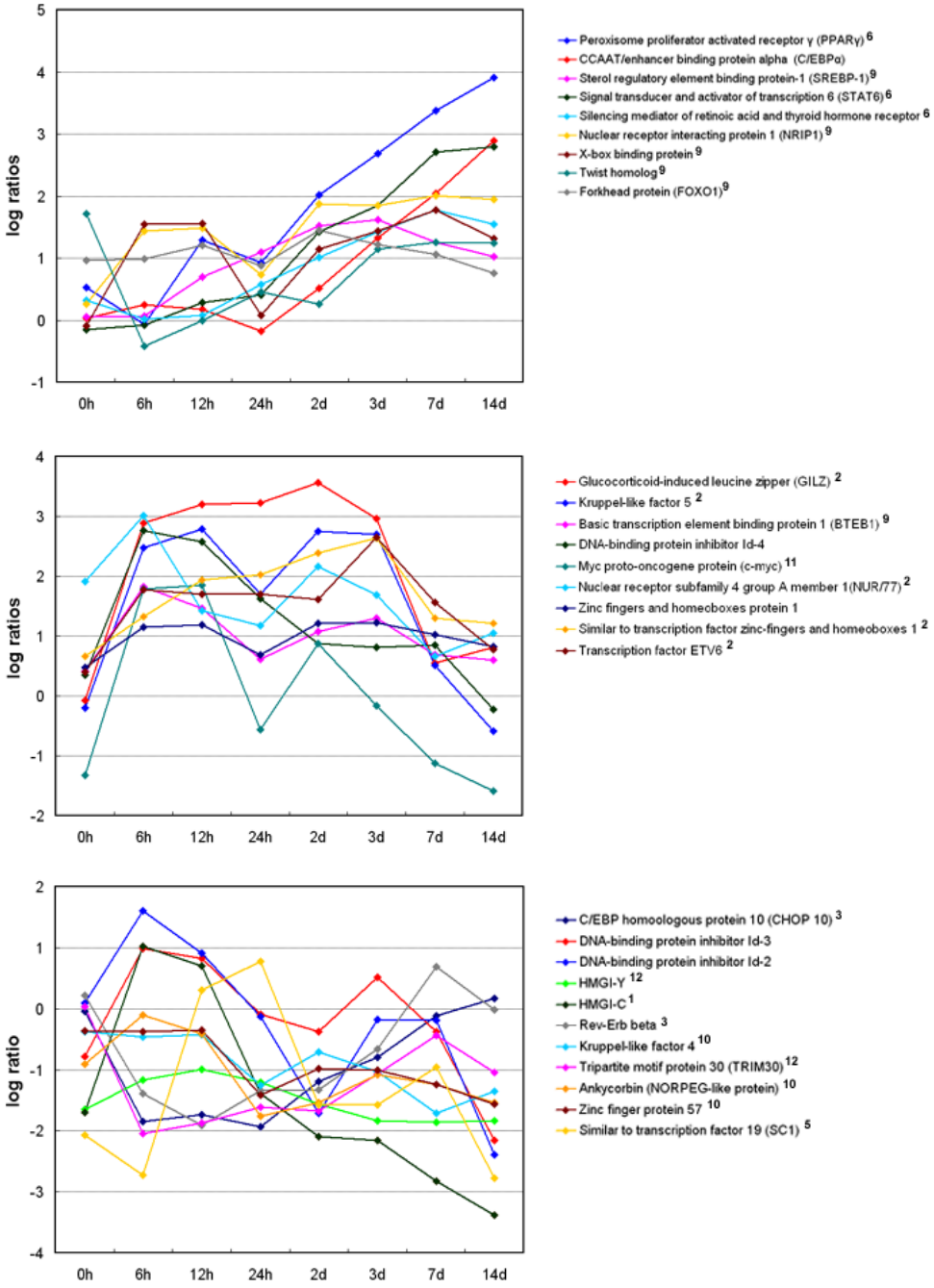

Figure 17: Selected transcriptional regulators and their relative gene expression profile (log ratios) during 3T3-L1 adipocyte differentiation. The affiliation of the factors to their respective clusters from previous k-means clustering is indicated in superscript. A large group of transcription factors and regulatory molecules showing a sustained increase in expression (upper panel), others are highly upregulated only in intermediate phases of differentiation (middle pannel), or mostly downregulated or rather downregulated in the late phase of adipocyte differentiation (lower panel).

3.4 Functional annotation, gene ontology, and biological processes

For each selected gene several steps were undertaken to assign the correct function and derive the involvement in specific biological processes. For 422 genes out of the 780 selected genes a gene ontology assignment for biological processes could be found. They could be divided in physiological process (53%), development (8%), and cellular processes (35%). In Figure 18 the subdivision into the next underlying level of the gene ontology hierarchy is shown. About 3% of the genes, which are assigned to development processes are involved in gene expression regulatory and epigenetic processes. 250 genes are assigned to gene ontology terms related to metabolism. Further subdivisions for the biological process metabolism are also shown in Figure 18. Since more GO terms can be assigned to one gene the percentage indicating not directly the number of genes, however, it gives information about current biological processes in the studied context. The assigned gene ontology terms mirror the already identified course of events. From the 422 genes remain 66 genes assigned to the GO term ”cell cycle”. Almost a third of them (21 genes) are within cluster 5. Only about 9% of the 250 metabolism genes are related to lipid metabolism.

Although the structured tree based gene ontology is very helpful in defining the function of a gene it is limited to well characterized genes (within the used microarray only 10.823 out of the 26.356 elements (without control features) have at least one GO entry). Therefore further efforts were pursued to annotate the remaining genes and assign function. All predicted functional motifs and protein domains in the corresponding protein sequence and the prediction method, the assigned name, the localization within the cell, a summarized function and available gene ontology terms for cellular compartment, biological process, and molecular function are accessible through the supplementary webpage (http://genome.tugraz.at/adipocyte). For 731 distinct ESTs out of the 780 selected ESTs 695 protein sequences were annotated with the annotator system (IMP). Since for some of the EST sequences more than one protein entry was found (1 to n relation) and different EST sequences were related to the same protein sequence (n to 1 relation) the number of entries in the result file is different. From the resulting 833 ESTs to 722 were assigned a name and to 637 a summarized function.

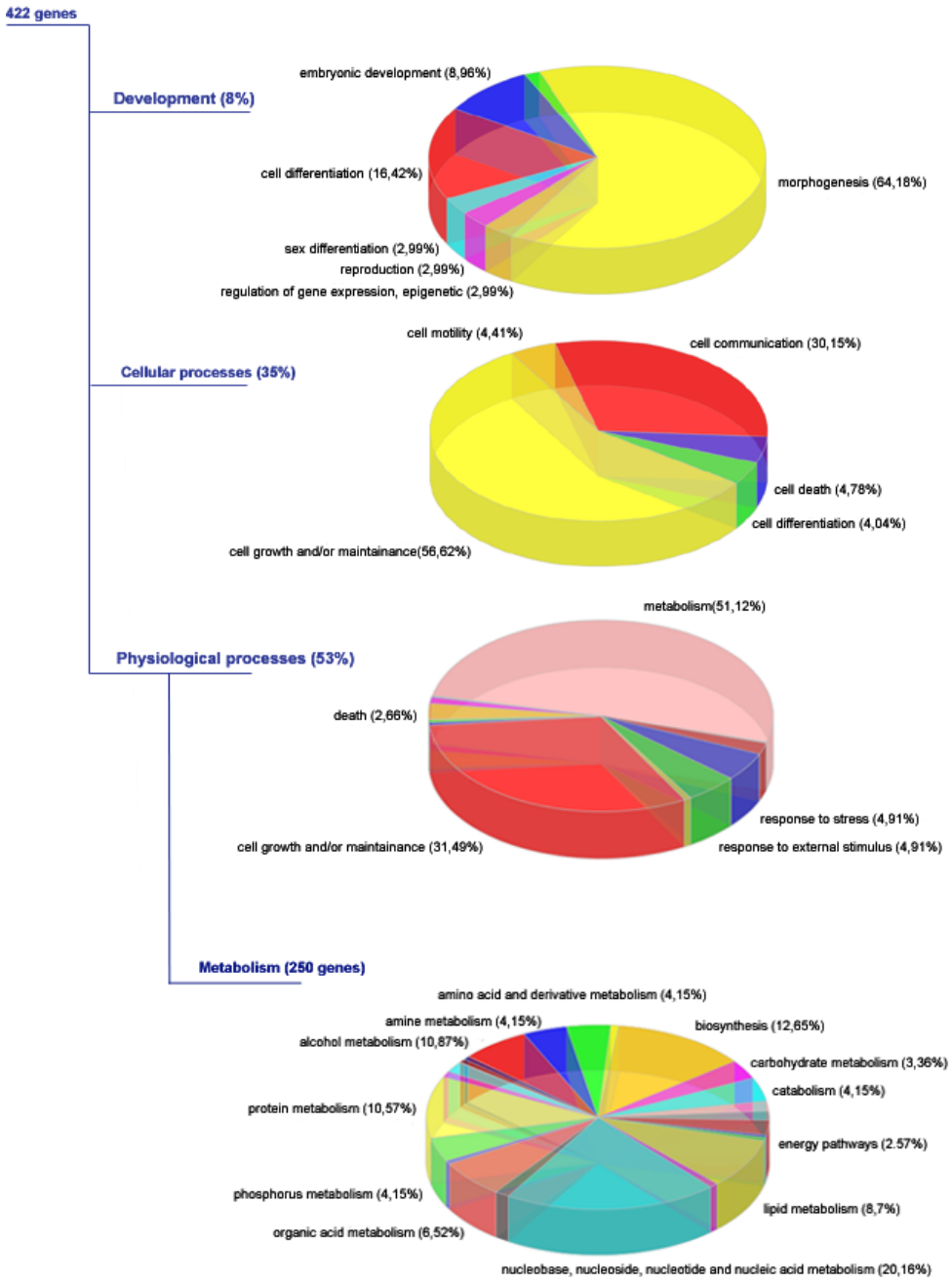

Figure 18: Distribution of the gene ontology assignments for 422 genes out of the 780 selected genes with assigned gene ontology terms for biological process

3.5 Confirmation of microarray results by real time RT-PCR

The microarray results were confirmed by real time RT-PCR measurements. Several candidates with different function and expression profiles were selected for this analysis: a complete profile of the early marker gene c-myc, the transcription factor BTEB1, and the highly expressed gene decorin were determined and also the relative expression levels at one time point (0h) of the cell cycle gene cyclin A, the major player PPARγ2, and the lipoprotein lipase were measured using Real Time RT-PCR. There was a high correlation (r^2=0.87) between microarray results and acquired relative gene expression levels from Real Time RT-PCR measurements of samples of the second microarray experiment as charted in Figure 19. It was shown that the profiles of the selected genes are identified correctly over all time points and over the whole range. This is also supported by further comparison to the averaged log ratios over all three experiments (Figure 19). Two measurements were controverse: microarray analyses indicating upregulation of the gene and RT-PCR measurements indicating downregulation. However, based on the low relative gene expression levels at these time points the genes wouldn't be considered differentially expressed.

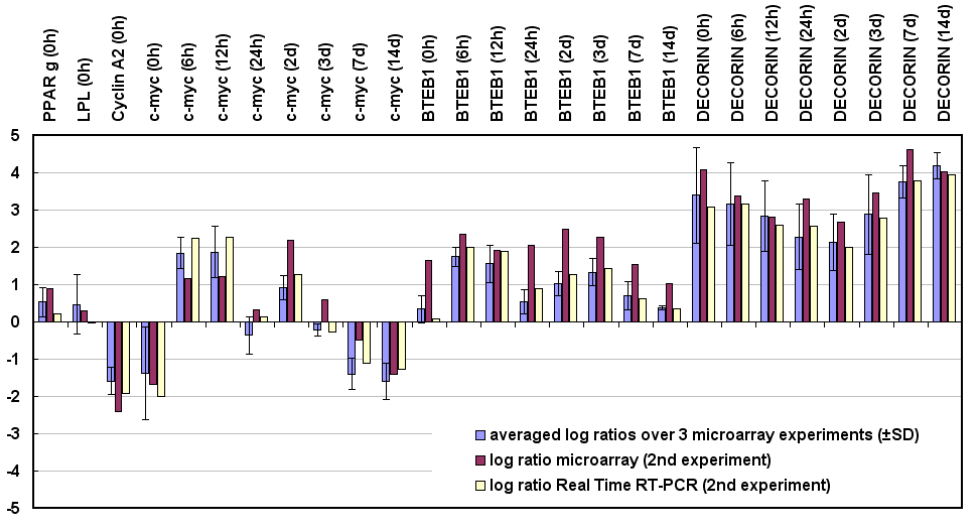

Figure 19: Comparison of RT -PCR results and microarray results from the second experiment

3.6 Possible associated genes identified by a reverse engineering approach

Applying of a novel reverse engineering approach based on mutual information and correlation to the data of the current study revealed in total 47 subnetworks of possibly associated genes in an iterative procedure. 24 subnetworks were identified in the first iteration, 2 of them were further analyzed in the second iteration setting a different threshold for the discretization. This process resulted in 17 new subnetworks, 2 of them were analyzed in the third iteration and 1 of the 9 resulting subnetworks in the forth iteration. The subnetworks are visualized in Figure 20, where the 283 grouped genes with identical discretized profile are represented by nodes and the relations between those groups (cluster) are shown by edges. As indicated, subnetworks with more than 20 none flat expression profiles (nodes) were submitted to the next iteration step. Each subnetwork, the respective genes and their continuous logarithmic expression profiles can be accessed by the supplementary web page. This method showed interesting regulatory combination of genes, which could not be detected by previous analysis steps. Some possibly related genes of specified subnetworks are shown in Table 3.

HMG-CoA synthase and lipoprotein lipase (LPL), both known targets for PPARγ are together in subnetwork 1-12, although in previous clustering procedures those genes were not in the same cluster due to the dissimilarity in their expression profiles. Results of subnetwork 24 suggests a relation between PPARγ and Acetyl-CoA synthetase. Although further evidence is required, this could be a causal relation, since Acyl-CoA synthase another enzyme of the fatty acid biosynthesis pathway, was previously identified as target gene for PPARγ. The great advantage of the proposed method is that genes that are oppositely regulated - which could indicate some regulatory mechanism - can be detected and grouped together as demonstrated by selected genes of subnetwork 1-1-5 and 1-5-1. The selected genes high mobility group protein HMGI-C and fibroblast growth factor 10 (FGF10) form subnetwork 1-1-5. DNA binding protein inhibitor Id-2 and fatty acid synthase from subnetwork 1-5-1, respectively, are known to play a role in adipocyte differentiation, but the regulatory mechanisms between those genes were not described. In addition there are many genes, partial uncharacterized, for which associations were suggested and if supported by other analytic methods - either computational or molecular biological - would be a good starting point for further studies.

Name	Cl.	0h	6h	12h	24h	2d	3d	7d	14d
Subnetwork 8									
Myo-inositol 1-phosphate synthase A1	44	-1.13	-1.02	0.17	-0.49	-0.2	-0.36	-1.85	-2.26
Matrilin-2	191	0.99	0.95	1.09	1.19	1.37	1.3	0.78	0.74
Non-muscle alpha-actinin 4	249	1.07	1.01	0.77	0.38	0.56	0.86	1.15	1.78
Subnetwork 24									
Acetyl-CoA synthetase	140	-1.12	-1.53	-0.82	-1.05	-0.81	0.2	0.45	0.99
Collagen alpha 2(VI)	274	0.08	0.7	1.06	0.92	1.59	1.39	1.13	1.19
PPAR-gamma	274	0.53	-0.07	1.29	0.93	2.02	2.68	3.37	3.91
Subnetwork 1-12									
HMG-CoA synthase	55	-1.84	-2.38	-2.15	-2.4	-1.69	-1.14	-0.12	0.79
Lipoprotein lipase (LPL)	119	0.47	-0.21	-0.16	-0.51	1.41	2.31	3.78	4.4
Subnetwork 1-1-5									
DNA-binding protein inhibitor Id-2	16	0.1	1.56	0.99	-0.1	-1.37	-0.2	-0.22	-2.7
Fatty acid synthase	35	-1.06	0.52	0.75	-0.1	0.39	0.72	1.77	2.05
Subnetwork 1-5-1									
High mobility group protein HMGI-C	1	-1.7	1.02	0.7	-1.39	-2.1	-2.16	-2.82	-3.38
Fibroblast growth factor-10 (FGF-10)	8	1.84	0.27	1.26	1.34	2.02	2.35	2.53	2.89
Subnetwork 1-1-1-4									
deoxycytidylate deaminase	4	-1.4	0.8	1.47	-1.25	-1.06	-1.42	-2.44	-2.86
C/EBP alpha	10	0.03	0.25	0.18	-0.17	0.52	1.33	2.04	2.89

Table 3: Selected genes from different subnetworks identified by a novel reverse engineering process based on mutual information and correlation

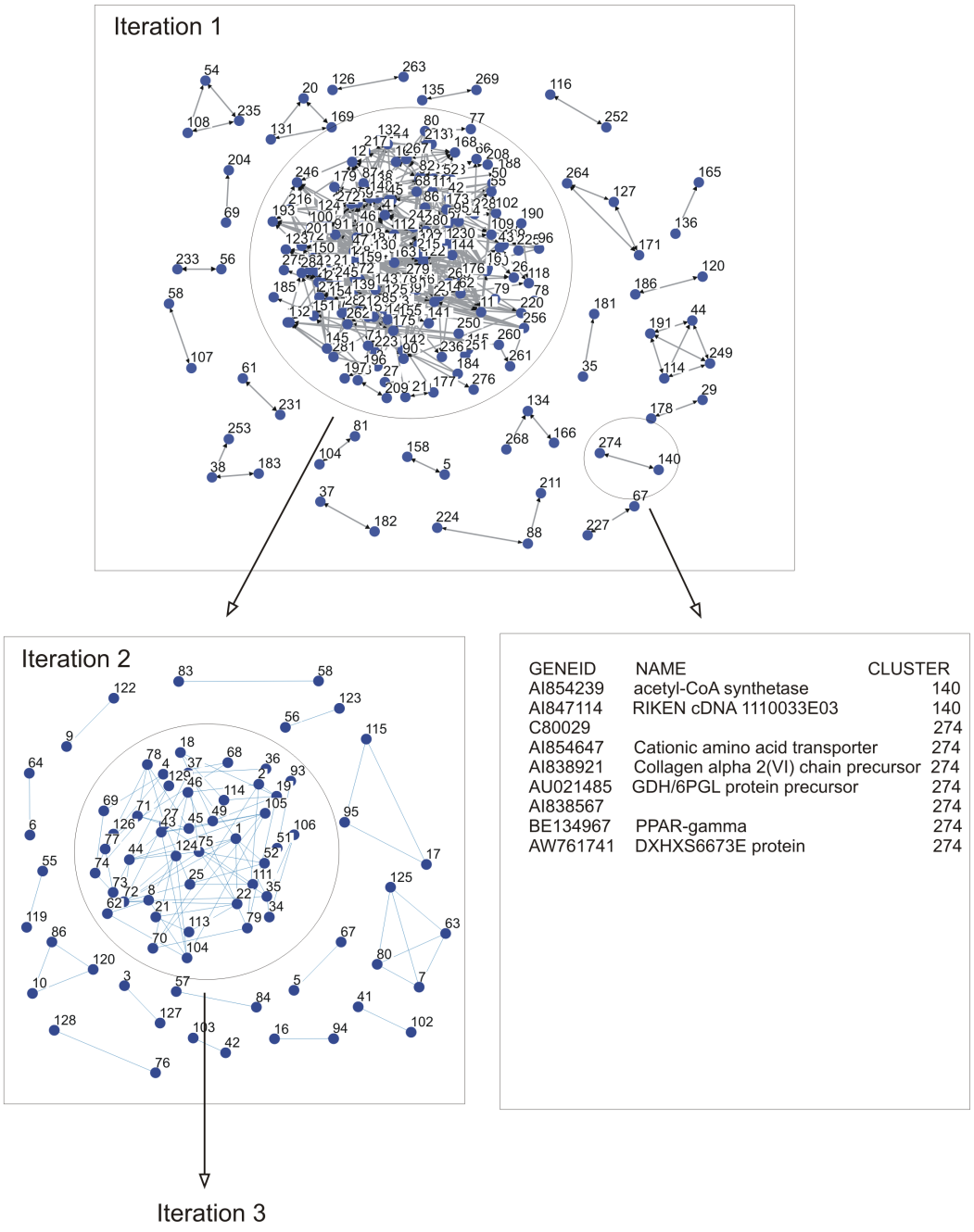

Figure 20: Identification of subnetworks by a reverse engineering approach based on mutual information and correlation. Subnetworks for the first and second iteration are shown and genes of the subnetwork which includes PPARγ are listed

3.7 3T3-L1 differentiation expression data in context of several pathways

Since in late adipocyte differentiation fat droplets are emerging and inclosed into the fat cells, lipogenesis and in more general lipid metabolism might play a dominant role. Microarray results should reflect according changes in gene expression of enzymes, those controlling respective reactions in several pathways. Microarray data were analyzed in the context of some representative pathways including lipid metabolism pathways like fatty acid metabolism, fatty acid biosynthesis, sterol synthesis and carbohydrate pathways like glycolysis/gluconeogenesis or citrate cycle (TCA cycle). For this purpose relative gene expression levels were mapped to corresponding elements (enzymes) in the pathway diagrams using the developed web portal and database GOLD.db (genomics of lipid associated disorders database), accessible through http://gold.tugraz.at. The used pathway diagrams were derived from the KEGG database [57]. It is possible to map expression levels from each of the 8 time points during 3T3-L1 differentiation. The provided data set comprises significantly differentially expressed genes identified by ANOVA those genes, which are more than two fold up- or downregulated in at least 4 time points.

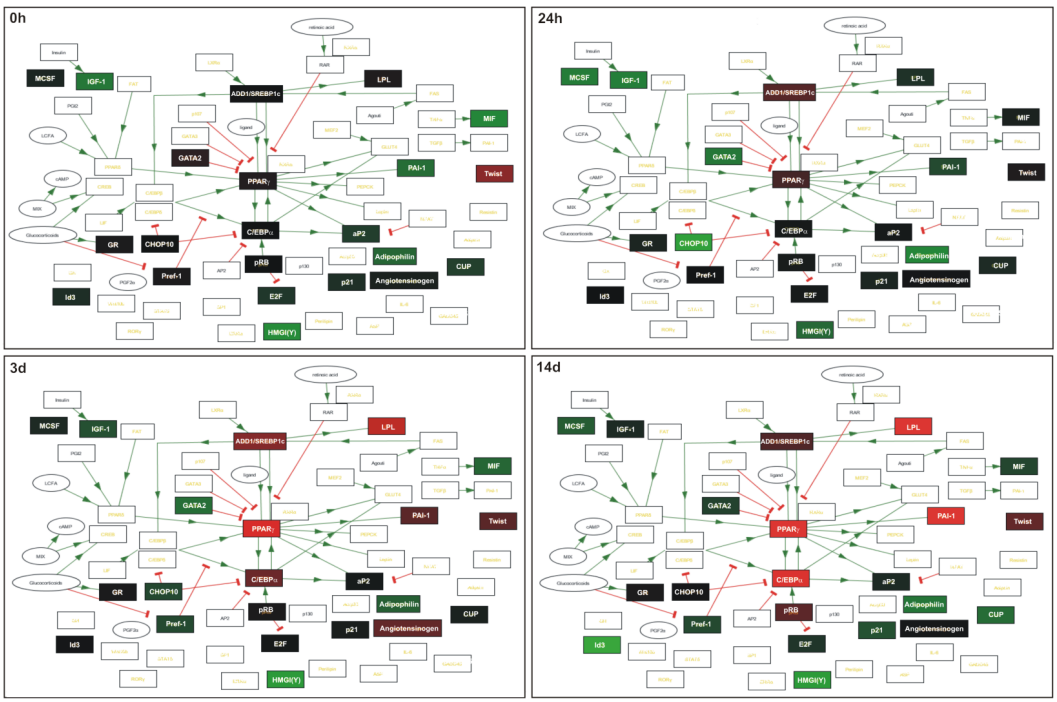

Figure 21: Mapping of 3T3-L1 dataset to adipogenesis regulatory network

For 2414 genes out of the selected genes there was a RefSeq number available, and therefore could be potentially mapped to the pathways. In Figure 22 the mapping of relative expression levels were illustrated schematically. The elements in the pathways were color coded according to the log ratios at different time points. For example in the fatty acid metabolism a concerted induction of the expression of a number of enzymes reactions in this pathway including the acyl-CoA dehydrogenase, long-chain-acyl-CoA dehydrogenase, acyl-CoA oxidase, glutaryl-CoA-dehydrogenase, carnithine palmitoyltransferase, 3 hydroxyacyl dehydrogenase was observed. In contrast, the acetyl-CoA acyltransferase in the fatty acid metabolism is downregulated in the late phase of adipocyte differentiation.

Interestingly, the citrate cycle is switched on in the mature status of the 3T3-L1 adipocytes indicated by the slight raise of several enzymes in this pathway (malate dehydrogenase, citrate synthase, isocitrate dehydrogenase (NAD), isocitrate dehydrogenase (NADP), succinate-CoA ligase, succinate dehydrogenase, dihydrolipoyl dehydrogenanse). The function of the citrate cycle is to provide energy in form of reduced electron carriers, which are utilized in the mitrochondrial respiratory chain for ATP synthesis, starting from acetyl-CoA in a series of enzymatic reactions. The cycle also serves as important source of biosynthetic intermediates. The pyruvate carboxylase, which was highly upregulated especial during the late phase of differentiation, catalyzes the conversion of pyruvate to oxaloacetate, an intermediate in the citrate cycle. Malic enzyme was also upregulated and catalyzes a reaction from pyruvate to malate, another example for an intermediate in the citrate cycle. These anaplerotic processes are required, since bionsynthetic reactions tend to deplete citrate cycle intermediates.
In the glycolysis/gluconeogenesis pathway as indicated in Figure 22 for time point 14d expression of genes for specialized enzymes are differentially regulated in one time point and also changing during the time course implicating not a clear picture of the expression results in context of this pathway. Although some of the enzymes are upregulated, pyruvate kinase and phosphopyruvate hydratase are downregulated at the end of differentiation. Expression profiles of further genes could be studied in the context of other pathways like those involved in the steroid metabolism as previous already found by clustering analyses. The relation between regulatory factors involved in the adipocyte differentiation process can be followed by mapping of the relative expression levels in the time course to the provided scheme of the regulatory network for adipogenesis, as can be done on http://gold.tugraz.at and is also available on the supplementary web page. The synchronous upregulation of the key players and some of target

genes like SCD-1 of LPL are apparent, also the early downregulation of inhibitory factors as CHOP-10 or GATA-2 (see Figure 21). Altogether, gene expression results support the suggested regulatory relations in adipocyte differentiation.

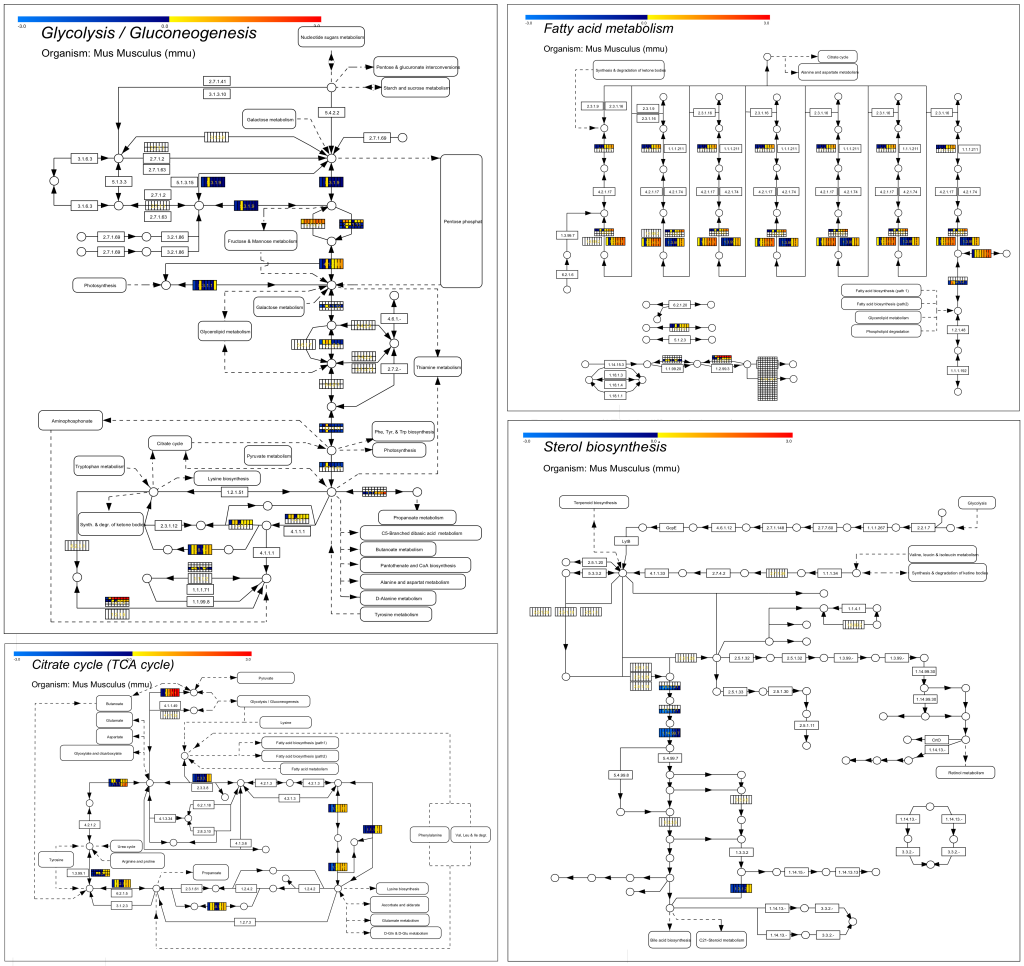

Figure 22: Schematic illustration for mapping of gene expression levels from 8 time points during 3T3-L1 adipocyte differentiation to the Glycolysis/Gluconeogenesis pathway (top left), fatty acid metabolism pathway(top, right), citrate cycle (bottom, left), and sterol biosynthesis pathway (bottom, right)

4 Discussion

Understanding the regulatory processes that control the adipocyte differentiation and the development of adipose tissue is the basis for developing therapeutic strategies for the treatment and prevention of obesity and related diseases. To a great extent, the current knowledge of the underlying molecular mechanisms of adipogenesis in physiological and pathophysiological states are derived from several monogenic mouse models of obesity, cell lines and primary cell cultures. In vitro systems have been extensively studied for more than 20 years. This has led to a dissection of the molecular and cellular events taking place during the transition from undifferentiated fibroblast-like preadipocytes into mature round fat cells. The gene expression profiles during this differentiation of 3T3-L1 cells were studied by large scale analysis with 27.648 element focused murine cDNA microarrays. Although additional noncell autonomous factors may be necessary to achieve a maximal level of the expression of genes *in vivo* [22], identified changes in expression levels in the 3T3-L1 cell line should mirror relevant regulatory processes. Previously it was shown that subcutanously injected preadipose cells in athymic mice led to mature fat pads indistinguishable from white adipose tissue (WAT). This indicates that *in vitro* differentiated adipocytes have many characteristics of adipose cells *in vivo* and that adipose cell acquisition occurs by a similar mechanism. Therefore, the 3T3-L1 cell line, which was used in the accomplished study, should be a suitable model to monitor the adipocyte differentiation processes considering also the acquired knowledge about some of the occuring mechanisms.

Stimulation of growth-arrested preadipocytes with a hormone cocktail, composed of insulin, dexamethasone and cAMP elevating agents, results in a rapid induction of the cell cycle. A first indication that adipocyte requires proliferative stimuli is the observation that both insulin and cAMP elevating agents are considered as mitogenic substances. In particular insulin, or the insulin like growth factor 1 (IGF-1) has been identified as the most potent inducers of clonal expansion. Hence, it is not surprising that the 3T3-L1 cells undergo mitotic clonal expansion before terminal differentiation, as also observed within the current analysis (see cluster 5 in Figure 16). The question if this process is a required step for adipocyte differentiation is more controversial. Blocking cell cycle re-entry with the DNA polymerase alpha inhibitor aphidicolin, resulted in the absence of lipid droplets [188]. Another anti proliferation reagent, rapamycin, was also shown to inhibit the clonal expansion phase and consequently preadipocytes fail to

differentiate [189]. Moreover, the MEK inhibitor UO126 and the cyclin-dependent kinase inhibitor roscovitine block mitotic clonal expansion in 3T3-L1 cells [5] indicating that this is prerequisite for terminal differentiation. However, it could not reproduced as reported in [190] that mitotic clonal expansion is not a required process using the inhibitor of mitogen-activated protein kinase PD98059. If this findings can be translated to other models is rather uncertain, since it was shown that precursor cells from human adipose tissue do not require cell division to enter the differentiation process *in vitro* [191]. These cells may have already undergone possibly critical cell divisions *in vivo* and may be in a late stage of adipocyte development. Recently it was shown that adipocyte differentiation of C3H10T1/2 cells, an immortalized mouse embryo fibroblasts cell line, depend on cycle progression through S-phase but do not require mitosis [192].

A number of genes were upregulated around 24h after induction and downregulated at all other timepoints. These genes were found mostly in cluster 5 and some in cluster 8. Many of this genes are already known to play a role in cell cycle process and partially known as marker for several types of cancer like Ki67. Moreover there are also uncharacterized genes or genes not known to be related to the cell cycle sharing similar expression profiles. Considering the short time frame of the cell cycle and mitotic clonal expansion of several hours [118] compared to the phase of terminal differentiation, which takes almost two weeks, it can be assumed, that the expression pattern of the sharp upregulation around 24h is characteristic for the cell cycle process. Cluster analysis are intended to elucidate genes that show similar expression profile, that share the same function or regulators (guilt by association). This seems especially fulfilled in this case, since previous microarray analysis during early adipocyte differentiation [24] showed the same peak in the expression profile in this time frame. Some of the identified genes are identical to those of the current analysis including cyclin A and cyclin B. Consequently it is reasonable to assume that not well characterized genes with the corresponding expression profile are related to cell cycle processes. Since it turned out that mitotic clonal expansion should be required for terminal adipocyte differentiation in 3T3-L1 cells, questions arise if the identified candidates are essential for the clonal expansion phase and subsequently for terminal adipocyte differentiation. In [193] was shown that intervention in the cell cycle process by prevention of degradation of the cyclin-dependent kinase inhibitor p27 leads to a block in the cell cycle entry and thus differentiation of preadipocyte inhibited. This demonstrates that manipulating essential genes could be detected by the failed terminal differentiation. Intervention in this early process by systematic transient down-knocking or silencing of special genes by RNA interfer-

ence (RNAi) technologies (for a review see [194]) would be an eligible method to assign gene function.

During adipocyte differentiation, acquisition of the adipocyte phenotype is characterized by chronological changes in the expression of numerous genes. This is reflected by the appearance of early, intermediate and late mRNA/protein markers and triglyceride accumulation. Whereas the early marker gene c-myc was detected throughout this study, C/EBPβ and C/EBPδ, which drive the adipogenesis process by binding to responsive elements in the promoter of C/EBPα and PPARγ, were not. Besides the expression profiles of the key players in adipocyte differentiation (PPARγ, C/EBPα, SREBP1) the expression profile of downstream targets of this key factors could be also confirmed. A number of transcriptional regulators showed distinct expression profiles indicating that they may play potential and different roles in the adipocyte differentiation process. Some of them were also identified in previous studies like the X-box binding protein or the kruppel like factor 4 [22]. Not only transcription factors, but also many genes related to several processes important in adipocyte differentiation like triglyceride metabolism, sterol metabolism, fatty acid transport were discovered during this study. Many genes involved in these processes were also identified by studies using in situ arrays, however, the great advantage of the current approach in order to identify novel targets is founded using a focused cDNA microarray, which utilizes also clones from early development stages, in combination with a novel functional annotation system. At a first step microarray results have to be confirmed by other methods like it is done in the current analysis with real time PCR.

Large scale gene expression analyses with microarrays differ in several aspects from conventional approaches like Northern blot analysis. The main difference is the high number of genes, which can be analyzed even in a single experiment with microarrays. This implicates that a part of the features can comprise uncharacterized ESTs or genes which expression profiles could help to identify novel targets and wouldn't be analyzed otherwise based on the current knowledge (discovery driven approach). The advantage of such approaches is that it is not necessary to make some a priori guesses of the outcome. There are some limitations on pursuing specific questions associated with microarrays, since not every element in a microarray can be analyzed and sometimes this could be a missing link in supporting the hypothesis (hypothesis driven approach). Some aspects could be overcome by adequate experimental design and optimized experimental procedures. Consequently, microarrays are often used to screen for novel targets and potential candidates which are further analyzed by other methods. For the specific applica-

tion we decided to use cDNA microarrays, because this type of arrays provides the flexibility to include both: first, a large number of genes(ESTs), which are uncharacterized and previously not associated with adipogenesis and second, genes (ESTs) that are important in adipocyte biology and lipid metabolism (focused approach).

An important question to consider in the analysis of microarray data is that one is trying to derive conclusions regarding protein function from RNA expression measurements. The activity of proteins, however, is not only regulated on the transcriptional level. A number of other regulatory events take place and influence the activity of the protein like binding of proteins to AU rich sequences in the untranslated regions of the transcript will influence RNA stability, posttranlational modifications (e.g. phosphorylation) or allosteric regulation of the protein, which are not detectable with cDNA microarrays. There are proteomics studies, which tried to compare mRNA abundance and protein levels [195, 196]. The reported correlation was very low ($r < 0.4$) and partly uncorrelated. Just in few cases there was a good agreement detectable. The complexity of the proteome is a dimension higher than that in the genome, since at every timepoint at a specific cellular location in the cell there is the possibility of several modifications of each protein. Hence, it is apparent that such approaches are very limited. The adipocyte differentiation process was shown to be regulated by a transcriptional cascade [129], and therefore should be mostly regulated on the transcriptional level. Previous studies showed a high agreement on protein levels detected by Western blots with the expression levels of several genes during adipocyte differentiation detected with microarrays including CHOP10, Id-2, cyclin A [24]. Importly, high throughput analyses of metabolites, protein content and post-translational modifications have become available and the techniques are rapidly improving in sensitivity. In the future, integrative analysis of data from microarrays and other sources will provide a more complete picture of the cellular processes. Not only these experimental methods will improve our understanding of occurring processes in the fat cell development, but also computational based methods. Once a protein sequence was derived for a specific EST (gene) sequence interpretation and prediction of domains can be used to derive the molecular function. Nonglobular regions like transmembrane proteins, low complexity regions or the occurance of known domains as retrievable from several databases like PFAM and PROSITE can be detected and together with BLAST search hits, literature, cell localization, and microarray data can be functional annotated. Although only for known gene available, gene ontology annotations has supplemented this procedure to identify involved processes.

Experimental design in microarray studies is very important, since thousands of measurements are conducted at once and if not well designed the significance of each measurement is poor. Hence, a major focus of this study was the experimental design of the microarray analysis. To consider the biological variability within one experiment RNA was pooled from 3 dishes. Since pooling is not independent, as biological replicates, 3 independent time series experiment were performed. High quality of hybridization results as shown in Figure 10 was achieved through optimized protocols for spotting and hybridization procedures, improved signal intensity by indirect labeling and pre-hybridization with BSA and stringent filtering of bad quality spots. Each hybridization was repeated with dye swap assignment and as apparent in the MA-plot there was almost no intensity dependency detected. Gene wise dye swap normalization was applied, which had the advantage not to rely on the assumption that most genes are not differentially expressed. Two criteria were introduced to cut down the number of genes and to select relevant differentially expressed genes for further analysis. One-way ANOVA led based on a significance level of p=0.05 to 5.205 differentially expressed genes. However, genes which show similar even high expression levels at all time points could not be detected, since difference between time points then is marginal. To consider also those genes and further decrease the number of genes a more stringent criteria (genes which are more than 2 fold up- or down-regulated in at least 4 time points) were applied. 392 from the 780 selected genes were also detected with ANOVA.

Several methods were introduced to help in searching for biological meaning in the wealth of data. These methods tried to extract different biological aspects according to the expression profiles. K-means clustering was applied to group genes based on the similarity in their expression profile. Since the parameter k has to be chosen, several parameters k were tested and k=12 showed the best cluster results in terms of maximal intercluster and minimal intracluster variance. Principal component analysis confirmed that clusters were well separated by k-means clustering. The reverse engineering problem can be formulated as what can be deduced about the underlying regulatory network given an amount of data. Although network models will little imply about the actual molecular mechanisms involved, much helpful information will be gained about genes critical for a biological process. This is also applicable to the proposed novel reverse engineering approach. Since this iterative method is based on mutual information, data have to be discretized. Gene expression profiles were discretized in just 3 levels, because higher number of discretization levels would lead to random patterns due to the noise in the expression

profiles. As threshold for the correlation $r = 0.63$ was chosen, based on the type II error at the significance level of 0.05. New associations can be derived in comparison to other analysis techniques such as cluster analysis because the information content of the gene expression profiles is used instead of similarity measurements of the expression profiles. A different approach was pursued within the GOLD.db, which was developed to facilitate genomic research providing a system for storing, integrating, and analyzing relevant data needed to decipher the molecular anatomy of lipid associated disorders. As part of this database, the possibility to map gene expression data to relevant pathways is provided. Recent advances allow to map expression levels at different conditions (e.g. time points) of a data set to pathways. It is possible either to map gene expression levels from all conditions at once to the pathways or consecutively. This is particularly helpful while deciding if a pathway is turned on in a certain time frame or under certain conditions. A number of KEGG pathways for different organisms, as well as annotated pathways are available.

5 Conclusion

The aim of this study was to elucidate regulatory processes and to identify new target genes involved in adipogenesis. For this purpose a 27.648 element focused murine cDNA microarray comprising a clone set from early embryonal stages (15k NIA clone set) and genes (ESTs) that are important in adipocyte biology and lipid metabolism was developed. Gene expression profiles from 3 independent 3T3-L1 preadipocyte differentiation experiments were acquired.
The expression profiles of the key players in adipocyte differentiation (PPARγ, C/EBPα, SREBP1) and a number of downstream target genes could be confirmed. Further it was shown that 24h after hormonal induction many cell cycle related genes are sharply upregulated, which is in consistency with the required clonal expansion phase before terminal differentiation in 3T3-L1 cells. Many genes that were upregulated in the late phase of differentiation are associated with sterol and lipid metabolism. Furthermore, a number of transcriptional regulators previously not associated with adipocyte differentiation showed distinct expression profiles. Miroarray results of some targets could be validated with real time RT-PCR.
Several methods were proposed to derive meaningful biological information. Cluster results from k-means clustering were most suitable to find out functionally related genes. A reverse engineering approach based on mutual information and correlation provided a number of gene subnetworks showing high-likely associated genes. Additional information is necessary to extract direct regulatory associations. Moreover, mapping of expression data in context of relevant pathways provides the possibility to find out, which pathways are turned on or off, respectively in a specific time frame or condition.
In summary, this was the first time 3T3-L1 differentiation was studied using cDNA microarrays. Due to the focused approach and a novel thorough functional annotation process new promising targets were revealed. Further experiments with the selected targets will provide novel insights into the regulatory mechanisms of adipogenesis.

References

[1] Kopelman PG. Obesity as a medical problem. *Nature*, 404:635–643, 2000.

[2] Must A, Spadano J, Coakley EH, Field AE, Colditz G, Dietz WH. The disease burden associated with overweight and obesity. *JAMA*, 282:1523–1529, 1999.

[3] Spiegelman BM, Flier JS. Obesity and the regulation of energy balance. *Cell*, 104:531–543, 2001.

[4] Friedman JM. A war on obesity, not the obese. *Science*, 299:856–858, 2003.

[5] Tang QQ, Otto TC, Lane MD. Mitotic clonal expansion: A synchronous process required for adipogenesis. *Proc Natl Acad Sci U S A*, 100:44–49, 2003.

[6] Wu Z, Xie Y, Bucher NLR, Farmer SR. Conditional ectopic expression of C/EBPb in NIH-3T3 cells induces PPARg and stimulates adipogenesis. *Genes Dev*, 9:2350–2363, 1995.

[7] Wu Z, Bucher NLR, Farmer SR. Induction of peroxisome proliferator-activated receptor g during the conversion of 3T3 fibroblasts into adipocytes is mediated by C/EBPb, C/EBPd, and glucocorticoids. *Mol Cell Biol*, 16:4128–4136, 1996.

[8] Yeh WC, Cao Z, Classon M, McKnight SL. Cascade regulation of terminal adipocyte differentiation by three members of the C/EBP family of leucine zipper proteins. *Genes Dev*, 9:168–181, 1995.

[9] Tanaka T, Yoshida N, Kishimoto T, Akira S. Defective adipocyte differentiation in mice lacking the C/EBPbeta and/or C/EBPdelta gene. *EMBO J*, 14:365–371, 1997.

[10] Wu Z, Rosen ED, Brun R, Hauser S, Adelmant G, Troy AE, McKeon C, Darlington GJ, Spiegelman BM. Cross-regulation of C/EBP alpha and PPAR gamma controls the transcriptional pathway of adipogenesis and insulin sensitivity. *Mol Cell*, 3:151–158, 1999.

[11] Kim JB, Spiegelman BM. ADD1/SREBP1 promotes adipocyte differentation and gene expression linked to fatty acid metabolism. *Genes Dev*, 10:1096–1107, 1996.

[12] Fajas L, Schoonjans K, Gelman L, Kim JB, Najib J, Martin G, Fruchart JC, Briggs M, Spiegelman BM, Auwerx J. Regulation of peroxisome proliferator-activated receptor gamma expression by adipocyte differentiation and determination factor 1/sterol regulatory element binding protein 1: implications for adipocyte differentiation and metabolism. *Mol Cell Biol*, 19:5495–5503, 1999.

[13] Kim JB, Wright HM, Wright M, Spiegelman BM. ADD1/SREBP1 activates PPARgamma through the production of endogenous ligand. *Proc Natl Acad Sci U S A*, 95:4333–4337, 1998.

[14] Hotamisligil GS, Shargill NS, Spiegelman BM. Adipose expression of tumor necrosis factor-alpha: direct role in obesity-linked insulin resistance. *Science*, 259:87–91, 1993.

[15] Mohamed-Ali V, Goodrick S, Rawesh A. Subcutaneous adipose tissue releases interleukin-6, but not tumor necrosis factor-alpha, in vivo. *J Clin Endocrinol Metab*, 82:4196–4200, 1997.

[16] Zhang Y, Proenca R, Maffei M, Barone M, Leopold L, Friedman JM. Positional cloning of the mouse obese gene and its human homologue. *Nature*, 372:425–432, 1994.

[17] Friedman JM, Halaas JL. Leptin and the regulation of body weight in mammals. *Nature*, 395:763–770, 1998.

[18] Steppan CM, Bailey ST, Bhat S, Brown EJ, Banerjee RR, Wright CM, Patel HR, Ahima RS, Lazar MA. The hormone resistin links obesity to diabetes. *Nature*, 409:307–312, 2001.

[19] Kim KH, Lee K, Moon YS, Sul HS. A cysteine-rich adipose tissue-specific secretory factor inhibits adipocyte differentiation. *J Biol Chem*, 276:11252–11256, 2001.

[20] Scherer A, Krause A, Walker JR, Sutton SE, Seron D, Raulf F, Cooke MP. A novel serum protein similar to C1q, produced exclusively in adipocytes. *J Biol Chem*, 270:26746–26749, 1995.

[21] Guo X, Liao K. Analysis of gene expression profile during 3T3-L1 preadipocyte differentiation. *Gene*, 251:45–53, 2000.

[22] Soukas A, Socci ND, Saatkamp BD, Novelli S, Friedman JM. Distinct transcriptional profiles of adipogenesis in vivo and in vitro. *J Biol Chem*, 276:34167–34174, 2001.

[23] Ross SE, Erickson RL, Gerin I, DeRose PM, Bajnok L, Longo KA, Misek DE, Kuick R, Hanash SM, Atkins KB, Andresen SM, Nebb HI, Madsen L, Kristiansen K, MacDougald OA. Microarray analyses during adipogenesis: understanding the effects of Wnt signaling on adipogenesis and the roles of liver X receptor alpha in adipocyte metabolism. *Mol Cell Biol*, 22:5989–5999, 2002.

[24] Burton GR, Guan Y, Nagarajan R, McGehee RE. Microarray analysis of gene expression during early adipocyte differentiation. *Gene*, 293:21–31, 2002.

[25] Gerhold DL, Liu F, Jiang G, Li Z, Xu J, Lu M, Sachs JR, Bagchi A, Fridman A, Holder DJ, Doebber TW, Berger J, Elbrecht A, Moller DE, Zhang BB. Gene expression profile of adipocyte differentiation and its regulation by peroxisome proliferator-activated receptor-gamma agonists. *Endocrinology*, 143:2106–2118, 2002.

[26] Jessen BA, Stevens GJ. Expression profiling during adipocyte differentiation of 3T3-L1 fibroblasts. *Gene*, 299:95–100, 2002.

[27] Kratchmarova I, Kalume DE, Blagoev B, Scherer PE, Podtelejnikov AV, Molina H, Bickel PE, Andersen JS, Fernandez MM, Bunkenborg J, Roepstorff P, Kristiansen K, Lodish HF, Mann M, Pandey A. A proteomic approach for identification of secreted proteins during the differentiation of 3T3-L1 preadipocytes to adipocytes. *Mol Cell Proteomics*, 1:213–222, 2002.

[28] Fajas L. Adipogenesis: a cross-talk between cell proliferation and cell differentiation. *Ann Med*, 35:79–85, 2003.

[29] Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470, 1995.

[30] Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ. High density synthetic oligonucleotide arrays. *Nat Genet*, 21:20–24, 1999.

[31] Lockhart DJ, Winzeler EA. Genomics, gene expression and DNA arrays. *Nature*, 405:827–836, 2000.

[32] 't Hoen PA, de Kort F, van Ommen GJ, den Dunnen JT. Fluorescent labelling of cRNA for microarray applications. *Nucleic Acids Res*, 31:e20.1–e20.8, 2003.

[33] Hegde P, Qi R, Abernathy K, Gay C, Dharap S, Gaspard R, Hughes JE, Snesrud E, Lee N, Quackenbush J. A concise guide to cDNA microarray analysis. *Biotechniques*, 29:548–556, 2000.

[34] 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*, 30:e15.1–e15.11, 2002.

[35] Quackenbush J. Microarray data normalization and transformation. *Nat Genet*, 32 Suppl:496–501, 2002.

[36] Cavalieri D Leung YF. Fundamentals of cDNA microarray data analysis. *Trends Genet*, 19:649–659, 2003.

[37] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95:14863–14868, 1998.

[38] Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. *Nat Genet*, 22:281–285, 1999.

[39] Ben Dor A, Shamir R, Yakhini Z. Clustering gene expression patterns. *J Comput Biol*, 6:281–297, 1999.

[40] Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR. 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, 1999.

[41] Herrero J, Valencia A, Dopazo J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, 17:126–136, 2001.

[42] Raychaudhuri S, Stuart JM, Altman RB. Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac Symp Biocomput*, 2000:455–466, 2000.

[43] Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A*, 97:10101–10106, 2000.

[44] Alter O, Brown PO, Botstein D. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proc Natl Acad Sci U S A*, 100:3351–3356, 2003.

[45] Fellenberg K, Hauser NC, Brors B, Neutzner A, Hoheisel JD, Vingron M. Correspondence analysis applied to microarray data. *Proc Natl Acad Sci U S A*, 98:10781–10786, 2001.

[46] Hastie T, Tibshirani R, Eisen MB, Alizadeh A, Levy R, Staudt L, Chan WC, Botstein D, Brown P. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol*, 1:RESEARCH0003.1–RESEARCH0003.21, 2000.

[47] Bolshakova N, Azuaje F. Cluster validation techniques for genome expression data. *Signal Processing*, 83:825–833, 2003.

[48] Bolshakova N, Azuaje F. Machaon CVE: cluster validation for gene expression data. *Bioinformatics*, 19:2494–2495, 2003.

[49] Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M, Jr., Haussler D. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A*, 97:262–267, 2000.

[50] Furey TS, Cristianini N, Duffy N, BeDNArski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16:906–914, 2000.

[51] Anderle P, Duval M, Draghici S, Kuklin A, Littlejohn TG, Medrano JF, Vilanova D, Roberts MA. Gene expression databases and data mining. *Biotechniques*, Suppl:36–44, 2003.

[52] Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet*, 29:365–371, 2001.

[53] Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, Bernhart D, Sherlock G, Ball C, Lepage M, Swiatek M, Marks WL, Goncalves J, Markel S, Iordan D, Shojatalab M, Pizarro A, White J, Hubley R, Deutsch E, Senger M, Aronow BJ, Robinson A, Bassett D, Stoeckert CJ, Jr., Brazma A. Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol*, 3:RESEARCH0046.1–RESEARCH0046.9, 2002.

[54] Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A, Rocca-Serra P, Sansone SA. ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res*, 31:68–71, 2003.

[55] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, 30:207–210, 2002.

[56] Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick

L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R, Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res. 2004 Jan 1;32(1):D258-61.*, 32:D258–D261, 2004.

[57] Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GenomeNet. *Nucleic Acids Res*, 30:42–46, 2002.

[58] Dahlquist KD, Salomonis N, Vranizan K, Lawler SC, Conklin BR. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat Genet*, 31:19–20, 2002.

[59] Cohen BA, Mitra RD, Hughes JD, Church GM. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat Genet*, 26:183–186, 2000.

[60] Jenssen TK, Laegreid A, Komorowski J, Hovig E. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet*, 28:21–28, 2001.

[61] Tanabe L, Scherf U, Smith LH, Lee JK, Hunter L, Weinstein JN. MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques*, 27:1210–1217, 1999.

[62] Masys DR, Welsh JB, Lynn FJ, Gribskov M, Klacansky I, Corbeil J. Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics*, 17:319–326, 2001.

[63] Segal E, Taskar B, Gasch A, Friedman N, Koller D. Rich probabilistic models for gene expression. *Bioinformatics*, 17 Suppl 1:S243–S252, 2001.

[64] Soukas A, Cohen P, Socci ND, Friedman JM. Leptin-specific patterns of gene expression in white adipose tissue. *Genes Dev*, 14:963–980, 2000.

[65] Green H, Meuth M. An established pre-adipose cell line and its differentiation in culture. *Cell*, 3:127–133, 1974.

[66] Green H, Kehinde O. An established pre-adipose cell line and its differentiation in culture. II. Factors affecting the adipose conversion. *Cell*, 5:19–27, 1975.

[67] Green H, Kehinde O. Formation of normally differentiated subcutaneous fat pads by an established preadipose cell line. *J Cell Phys*, 101:169–172, 1979.

[68] Mandrup S, Loftus TM, MacDougald OA, Kuhajda FP, Lane MD. Obese gene expression at in vivo levels by fat pads derived from s.c. implanted 3T3-F442A preadipocytes. *Proc Natl Acad Sci U S A*, 94:4300–4305, 1997.

[69] Student AK, Hsu RY, Lane MD. Induction of fatty acid synthetase synthesis in differentiating 3T3-L1 preadipocytes. *J Biol Chem*, 255:4745–4750, 1980.

[70] Le Lay S, Lefrere I, Trautwein C, Dugail I, Krief S. Insulin and sterol-regulatory element-binding protein-1c (SREBP-1C) regulation of gene expression in 3T3-L1 adipocytes Identification of CCAAT/enhancer-binding protein beta as an SREBP-1C target. *J Biol Chem*, 277:35625–35634, 2002.

[71] Churchill GA. Fundamentals of experimental design for cDNA microarrays. *Nat Genet*, 32 Suppl:490–495, 2002.

[72] Simon RM, Dobbin K. Experimental design of DNA microarray experiments. *Biotechniques*, Suppl:16–21, 2003.

[73] 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 J, Jr., Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Staudt LM. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.

[74] Chomczynski P, Sacchi N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal Biochem*, 162:156 –159, 1987.

[75] Kerr MK, Martin M, Churchill GA. Analysis of variance for gene expression microarray data. *J Comput Biol*, 7:819–837, 2000.

[76] Loennstedt I, Speed TP. Replicated microarray data. *Statistica Sinica*, 12:31–46, 2002.

[77] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, 98:5116–5121, 2001.

[78] Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. *Proc Natl Acad Sci U S A*, 96:1151–1160, 2001.

[79] Dudoit S, Yang YH, Callow MJ Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiements. *Statistica Sinica*, 12:111–139, 2002.

[80] Sachs L. Applied Statistics - A Handbook of Techniques. *Springer*, 1982.

[81] Westfall PH, Young SS. Resampling-based multiple testing: examples and methods for p-value adjustment. *Wiley*, 1993.

[82] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc*, 57:289–300, 1995.

[83] Cui X, Churchill GA. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*, 4:210.1 –210.10, 2003.

[84] Pan W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 18:546–554, 2002.

[85] Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS. Assessing gene significance from cDNA microarray expression data via mixed models. *J Comput Biol*, 8:625–637, 2001.

[86] Baldi P, Long AD. Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17:509–519, 2001.

[87] Thomas JG, Olson JM, Tapscott SJ, Zhao LP. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Res*, 11:1227–1236, 2001.

[88] Ideker T, Thorsson V, Siehel AF, Hood LE. Testing for differentially-expressed genes by maximum likelihood analysis of microarray data. *J Comput Biol*, 7:805–817, 2000.

[89] Everitt. Cluster Analysis 122. *Heinemann, London*, 1974.

[90] Holter NS, Mitra M, Maritan A, Cieplak M, Banavar JR, Fedoroff NV. Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc Natl Acad Sci U S A*, 97:8409–8414, 2000.

[91] Sturn A, Quackenbush J, Trajanoski Z. Genesis: cluster analysis of microarray data. *Bioinformatics*, 18:207–208, 2002.

[92] Quackenbush J, Liang F, Holt I, Pertea G, Upton J. The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res*, 28:141–145, 2000.

[93] Wheeler DL, Church DM, Edgar R, Federhen S, Helmberg W, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Suzek TO, Tatusova TA, Wagner L. Database resources of the National Center for Biotechnology Information: update. *Trends Cell Biol*, 32:D35–D40, 2004.

[94] Pruitt KD, Katz KS, Sicotte H, Maglott DR. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet*, 16:44–47, 2000.

[95] Apweiler R, Martin MJ, O'Donovan C, Pruess M. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res*, 32:D115–D119, 2004.

[96] Diehn M, Sherlock G, Binkley G, Jin H, Matese JC, Hernandez-Boussard T, Rees CA, Cherry JM, Botstein D, Brown PO, Alizadeh AA. SOURCE: a unified genomic resource of functional annotations, ontologies, andgene expression data. *Nucleic Acids Res*, 31:219–223, 2003.

[97] Blake JA, Richardson JE snd Davisson MT, Eppig JT. The Mouse Genome Database (MGD). A comprehensive public resource of genetic, phenotypic and genomic data. *Nucleic Acids Res*, 25:85–91, 1997.

[98] Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T, Durbin R, Eyras E, Gilbert J, Hammond M, Huminiecki L, Kasprzyk A, Lehvaslaiho H, Lijnzaad P, Melsopp C, Mongin E, Pettett R, Pocock M, Potter S, Rust A, Schmidt E, Searle S, Slater G, Smith J, Spooner W, Stabenau A, Stalker J, Stupka E, Ureta-Vidal A, Vastrik I, Clamp M. The Ensembl genome database project. *Nucl Acids Res*, 30:38–41, 2002.

[99] Lenhard B, Wahlestedt C, Wasserman WW. GeneLynx mouse: integrated portal to the mouse genome. *Genome Res*, 13:1501–1504, 2003.

[100] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*, 215:403–410, 1990.

[101] Zhang MQ. Identification of human gene core promoters in silico. *Genome Res*, 8:319–326, 1998.

[102] Knudsen S. Promoter2.0: for the recognition of PolIII promoter sequences. *Bioinformatics*, 15:356–361, 1999.

[103] Scherf M, Klingenhoff A, Werner T. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J Mol Biol*, 297:599–606, 2000.

[104] Qin L, Qiu P, Wang L, Li X, Swarthout JT, Soteropoulos P, Tolias P, Partridge NC. Gene expression profiles and transcription factors involved in parathyroid hormone signaling in osteoblasts revealed by microarray and bioinformatics. *J Biol Chem*, 278:19723–19731, 2003.

[105] Qiu P, Ding W, Jiang Y, Greene JR, Wang L. Computational analysis of composite regulatory elements. *Mamm Genome*, 13:327–332, 2002.

[106] Trinklein ND, Aldred SJ, Saldanha AJ, Myers RM. Identification and functional analysis of human transcriptional promoters. *Genome Res*, 13:308–312, 2003.

[107] Lenhard B, Sandelin A, Mendoza L, Engstrom P, Jareborg N, Wasserman WW. Identification of conserved regulatory elements by comparative genome analysis. *J Biol*, 2:13 –13, 2003.

[108] Halees AS, Leyfer D, Weng Z. PromoSer: a large-scale mammalian promoter and transcription start site identification service. *Nucl Acids Res*, 31(13):3554–3559, 2003.

[109] Cavin R, Junier T, Bucher P. The Eukaryotic Promoter Database EPD. *Nucleic Acids Res*, 26:353–357, 1998.

[110] Kent WJ. BLAT–the BLAST-like alignment tool. *Genome Res*, 12:656–664, 2002.

[111] Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Pruss M, Reuter I, Schacherer F. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res*, 28:316–319, 2000.

[112] Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucl Acids Res*, 32:D91–D94, 2004.

[113] Quandt K, Frech K, Karas H, Wingender E, Werner T. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res*, 23:4878–4884, 1995.

[114] Gregoire FM, Smas CM, Sul HS. Understanding adipocyte differentiation. *Physiol Rev*, 78:783–809, 1998.

[115] MacDougald OA, Mandrup S. Adipogenesis: forces that tip the scales. *Trends Endocrinol Metab*, 13:5–11, 2002.

[116] Ross SE, Hemati N, Longo KA, Bennett CN, Lucas PC, Erickson RL, MacDougald OA. Inhibition of adipogenesis by Wnt signaling. *Science*, 289:950–953, 2000.

[117] Tong Q, Dalgin G, Xu H, Ting CN, Leiden JM, Hotamisligil GS. Function of GATA transcription factors in preadipocyte-adipocyte transition. *Science*, 290:134–138, 2000.

[118] Morrison RF, Farmer SR. Role of PPARgamma in regulating a cascade expression of cyclin-dependent kinase inhibitors, p18(INK4c) and p21(Waf1/Cip1), during adipogenesis. *J Biol Chem*, 274:17088–17097, 1999.

[119] Tang QQ, Lane MD. Role of C/EBP homologous protein (CHOP-10) in the programmed activation of CCAAT/enhancer-binding protein-beta during adipogenesis. *Proc Natl Acad Sci U S A*, 97:12446–12450, 2000.

[120] Chen PL, Riley DJ, Chen Y, Lee WH. Retinoblastoma protein positively regulates terminal adipocyte differentiation through direct interaction with C/EBPs. *Genes Dev*, 10:2794–2804, 1996.

[121] Brun RP, Tontonoz P, Forman BM, Ellis R, Chen J, Evans RM, Spiegelman BM. Differential activation of adipogenesis by multiple PPAR isoforms. *Genes Dev*, 10:974–984, 1996.

[122] Mandrup S, Lane MD. Regulating adipogenesis. *J Biol Chem*, 272:5367–5370, 1997.

[123] Loftus TM, Lane MD. Modulating the transcriptional control of adipogenesis. *Curr Opin Genet Dev*, 7:603–608, 1997.

[124] Hwang CS, Loftus TM, Mandrup S, Lane MD. Adipocyte differentiation and leptin expression. *Annu Rev Cell Dev Biol*, 13:231–259, 1997.

[125] Darlington GJ, Ross SE, MacDougald OA. The role of C/EBP genes in adipocyte differentiation. *J Biol Chem*, 273:30057–30060, 1998.

[126] Cowherd RM, Lyle RE, McGehee RE, Jr. Molecular regulation of adipocyte differentiation. *Semin Cell Dev Biol*, 10:3–10, 1999.

[127] Auwerx J. PPARgamma, the ultimate thrifty gene. *Diabetologia*, 42:1033–1049, 1999.

[128] Ntambi JM, Young-Cheul K. Adipocyte differentiation and gene expression. *J Nutr*, 130:3122S–3126S, 2000.

[129] Rosen ED, Spiegelman BM. Molecular regulation of adipogenesis. *Annu Rev Cell Dev Biol*, 16:145–171, 2000.

[130] Rosen ED, Spiegelman BM. Peroxisome proliferator-activated receptor gamma ligands and atherosclerosis: ending the heartache. *J Clin Invest*, 106:629–631, 2000.

[131] Rangwala SM, Lazar MA. Transcriptional control of adipogenesis. *Annu Rev Nutr*, 20:535–559, 2000.

[132] Fajas L, Debril MB, Auwerx J. Peroxisome proliferator-activated receptor-gamma: from adipogenesis to carcinogenesis. *J Mol Endocrinol*, 27:1–9, 2001.

[133] Grimaldi PA. The roles of PPARs in adipocyte differentiation. *Prog Lipid Res*, 40:269–281, 2001.

[134] Gregoire FM. Adipocyte differentiation: from fibroblast to endocrine cell. *Exp Biol Med (Maywood)*, 226:997–1002, 2001.

[135] Rosen ED, Spiegelman BM. PPARgamma : a nuclear regulator of metabolism, differentiation, and cell growth. *J Biol Chem*, 276:37731–37734, 2001.

[136] Tong Q, Hotamisligil GS. Molecular mechanisms of adipocyte differentiation. *Rev Endocr Metab Disord*, 2:349–355, 2001.

[137] Holst D, Grimaldi PA. New factors in the regulation of adipose differentiation and metabolism. *Curr Opin Lipidol*, 13:241–245, 2002.

[138] Camp HS, Ren D, Leff T. Adipogenesis and fat-cell function in obesity and diabetes. *Trends Mol Med*, 8:442–447, 2002.

[139] Valet P, Tavernier G, Castan-Laurell I, Saulnier-Blache JS, Langin D. Understanding adipose tissue development from transgenic animal models. *J Lipid Res*, 43:835–860, 2002.

[140] Richon VM, Lyle RE, McGehee RE, Jr. Regulation and expression of retinoblastoma proteins p107 and p130 during 3T3-L1 adipocyte differentiation. *J Biol Chem*, 272:10117–10124, 1997.

[141] Fajas L, Egler V, Reiter R, Hansen J, Kristiansen K, Debril MB, Miard S, Auwerx J. The retinoblastoma-histone deacetylase 3 complex inhibits PPARgamma and adipocyte differentiation. *Dev Cell*, 3:903–910, 2002.

[142] Wang C, Pattabiraman N, Zhou JN, Sakamaki T Fu M, Albanese C, Li Z, Wu K, Hulit J, Neumeister P, Novikoff PM, Brownlee M, Scherer PE, Jones JG, Whitney KD, Donehower LA, Harris EL, Rohan T, Johns DC, Pestell RG. Cyclin D1 repression of peroxisome proliferator-activated receptor gamma expression and transactivation. *Mol Cell Biol*, 23:6159–6173, 2003.

[143] Cram EJ, Ramos RA, Wang EC, Cha HH, Nishio Y, Firestone GL. Role of the CCAAT/enhancer binding protein-alpha transcription factor in the glucocorticoid stimulation of p21waf1/cip1 gene promoter activity in growth-arrested rat hepatoma cells. *J Biol Chem*, 273:2008–2014, 1998.

[144] Smas CM, Kachinskas D, Liu CM, Xie X, Dircks LK, Sul HS. Transcriptional control of the pref-1 gene in 3T3-L1 adipocyte differentiation. Sequence requirement for differentiation-dependent suppression. *J Biol Chem*, 273:31751–31758, 1998.

[145] Smas CM, Chen L, Zhao L, Latasa MJ, Sul HS. Transcriptional repression of pref-1 by glucocorticoids promotes 3T3-L1 adipocyte differentiation. *J Biol Chem*, 274:12632–12641, 1999.

[146] Schmidt W, Poll-Jordan G, Loffler G. Adipose conversion of 3T3-L1 cells in a serum-free culture system depends on epidermal growth factor, insulin-like growth factor I, corticosterone, and cyclic AMP. *J Biol Chem*, 265:15489–15495, 1990.

[147] Cao Z, Umek RM, McKnight SL. Regulated expression of three C/EBP isoforms during adipose conversion of 3T3-L1 cells. *Genes Dev*, 5:1538–1552, 1991.

[148] Reusch JE, Colton LA, Klemm DJ. CREB activation induces adipogenesis in 3T3-L1 cells. *Mol Cell Biol*, 20:1008–1020, 2000.

[149] Hansen JB, Zhang H, Rasmussen TH, Petersen RK, Flindt EN, Kristiansen K. Peroxisome proliferator-activated receptor delta (PPARdelta)-mediated regulation of preadipocyte proliferation and gene expression is dependent on cAMP signaling. *J Biol Chem*, 276:3175–3182, 2001.

[150] Bastie C, Holst D, Gaillard D, Jehl-Pietri C, Grimaldi PA. Expression of peroxisome proliferator-activated receptor PPARdelta promotes induction of PPARgamma and adipocyte differentiation in 3T3C2 fibroblasts. *J Biol Chem*, 274:21920–21925, 1999.

[151] Choy L, Skillington J, Derynck R. Roles of autocrine TGF-beta receptor and Smad signaling in adipocyte differentiation. *J Cell Biol*, 149:667–682, 2000.

[152] Sakaue H, Konishi M, Ogawa W, Asaki T, Mori T, Yamasaki M, Takata M, Ueno H, Kato S, Kasuga M, Itoh N. Requirement of fibroblast growth factor 10 in development of white adipose tissue. *Genes Dev*, 16:908–912, 2002.

[153] Stephens JM, Morrison RF, Pilch PF. The expression and regulation of STATs during 3T3-L1 adipocyte differentiation. *J Biol Chem*, 271:10441–10444, 1996.

[154] Stephens JM, Morrison RF, Pilch PF. PPARγ ligand-dependent induction of STAT1, STAT5A, and STAT5B during adipogenesis. *Biochem Bioph Res Commun*, 262:216–222, 1999.

[155] Floyd ZE, Stephens JM. STAT5A Promotes Adipogenesis in Nonprecursor Cells and Associates With the Glucocorticoid Receptor During Adipocyte Differentiation. *Diabetes*, 52:308–314, 2003.

[156] Nam SY, Lobie PE. The mechanism of effect of growth hormone on preadipocyet and adipocyte function. *Obes Rev*, 1:73–86, 2000.

[157] Wabitsch M, Braun S, Hauner H, Heinze E, Ilondo MM, Shymo R, De Meyts P, Teller WM. Mitogenic and antiadipogenic properties of human growth hormone in differentiating human adipocyet precursor cells in primary culture. *Pediatr Res*, 40:450–456, 1996.

[158] Xu H, Sethi JK, Hotamisligil GS. Transmembrane tumor necrosis factor (TNF)-alpha inhibits adipocyte differentiation by selectively activating TNF receptor 1. *J Biol Chem*, 274:26287–26295, 1999.

[159] Levine JA. Adipocyte macrophage colony-stimulated factor is a mediator of adipose tissue growth. *J Clin Invest*, 101:1557–1564, 1998.

[160] Moldes M, Lasnier F, Feve B, Pairault J, Djian P. Id3 prevents differentiation of preadipose cells. *Mol Cell Biol*, 17:1796–1804, 1997.

[161] Moldes M, Boizard M, Liepvre XL, Feve B, Dugail I, Pairault J. Functional antagonism between inhibitor of DNA binding (Id) and adipocyte determination and differentiation factor 1/sterol regulatory element-binding protein-1c (ADD1/SREBP-1c) trans-factors for the regulation of fatty acid synthase promoter in adipocytes. *Biochem J*, 344 Pt 3:873–880, 1999.

[162] Melillo RM, Pierantoni GM, Scala S, Battista S, Fedele M, Stella A, De BiasioM, Chiappetta G, Fidanza V, Condorelli G, Santoro M, Croce CM, Viglietto G, Fusco A. Critical role of the HMGI(Y) proteins in adipocytic cell growth and differentiation. *Mol Cell Biol*, 21:2485–2495, 2001.

[163] Wolfrum C, Shih DQ, Kuwajima S, Norris AW, Kahn CR, Stoffel M. Role of Foxa-2 in adipocyte metabolism and differentiation. *J Clin Invest*, 112:345–356, 2003.

[164] Nakae J, Kitamura T, Kitamura Y, Biggs WH, Arden KC, Accili D. The forkhead transcription factor foxo1 regulates adipocyte differentiation. *Dev Cell*, 4:119–129, 2003.

[165] Banerjee SS, Feinberg MW, Watanabe M, Gray S, Haspel RL, Denkinger DJ, Kawahara R, Hauner H, Jain MK. The Kruppel-like factor KLF2 inhibits peroxisome proliferator-activated receptor-gamma expression and adipogenesis. *J Biol Chem*, 278:2581–2584, 2003.

[166] Åkerblad P, Lind U, Liberg D, Bamberg K, Sigvardsson M. Early B-cell factor (O/E-1) is a promoter of adipogenesis and involved in control of genes important for terminal adipocyte differentiation. *Mol Cell Biol*, 22:8015–1773, 2002.

[167] Ho IC, Kim JH, Rooney JW, Spiegelman BM, Glimcher LH. A potential role for the nuclear factor of activated T cells family of transcriptional regulatory proteins in adipogenesis. *Proc Natl Acad Sci U S A*, 95:15537–15541, 1998.

[168] Aubert J, Darimont C, Safonova I, Ailhaud G, Negrel R. Prostacyclin IP receptor up-regulates the early expression of C/EBPβ and C/EBPδ. *Mol Cell Endocrinol*, 160:149–156, 2000.

[169] Serrero G. Paracrine regulation of adipose differentiation by arachidonate metabolites prostaglandin F2 alpha inhibits early and late markers of differentiation in the adipogenic cell line 1246. *Endocrinology*, 131:2545–2551, 1992.

[170] Kliewer SA, Lenhard JM, Willson TM, Patel I, Morris DC, Lehmann JM. A prostaglandin J_2 metabolite binds peroxisome proliferator-activated receptor gamma and promotes adipocyte differentiation. *Cell*, 83:813–819, 1995.

[171] Trost E, Hackl H, Maurer M, Trajanoski Z. Java editor for biological pathways. *Bioinformatics*, 19:786–787, 2003.

[172] D'haeseleer P, Liang S, Somogyi R. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16:707–726, 2000.

[173] Akutsu T, Miyano S, Kuhara S. Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics*, 16:727–734, 2000.

[174] Friedman N, Linial M, Nachman I, Pe'er D. Using bayesian networks to analyze expression data. *J Comp Biol*, 7:601–620, 2000.

[175] Chen T, He HL, Church GM. Modeling gene expression with differential equations. *Pac Symp Biocomput*, pages 29–40, 1999.

[176] Butte AJ, Kohane IS. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput*, pages 418–429, 2000.

[177] Liang S, Fuhrman S, Somogyi R. REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. *Pac Symp Biocomput*, pages 18–29, 1998.

[178] Breitkreutz BJ, Stark C, Tyers M. Osprey: a network visualization system. *Genome Biol*, 4:R22.1–R22.4, 2003.

[179] Rassart E, Bediran A, Do Carmo S, Guinard O, Sirois J, Terrisse SL, Milne R. Apolipoprotein D. *Biochim Biophys Acta*, 1482:185–198, 2000.

[180] Hummasti S, Laffitte BA, Watson MA, Galardi C, Chao LC, Ramamurthy L, Moore JT, Tontonoz P. Liver X receptors are regulators of adipocyte gene expression but not differentiation. Identification of apoD as direct target. *J Lipid Res*, M300312:1–35, 2004.

[181] Dynlacht BD. Regulation of transcription by proteins that control the cell cycle. *Nature*, 389:149–152, 1997.

[182] Tanaka TU. Bi-orienting chromosomes on the mitotic spindle. *Curr Opin Cell Biol*, 97:9127–9132, 2002.

[183] Brown DC, Gatter KC. Ki67 protein: the immaculate deception? *Histopathology*, 40:2–11, 2002.

[184] Schmidt C, Beyersmann D. Transient peaks in zinc and metallothionein levels during differentiation of 3T3L1 cells. *Arch Biochem Biophys*, 364:91–98, 1999.

[185] Shi XM, Blair HC, Yang X, McDonald JM, Cao X. Tandem repeat of C/EBP binding sites mediates PPARgamma2 gene transcription in glucocorticoid-induced adipocyte differentiation. *J Cell Biochem*, 76:518–527, 2000.

[186] Shi X, Shi W, Li Q, Song B, Wan M, Bai S, Cao X. A glucocorticoid-induced leucine-zipper protein, GILZ, inhibits adipogenesis of mesenchymal cells. *EMBO Rep*, 4:374–380, 2003.

[187] Freytag SO, Geddes TJ. Reciprocal regulation of adipogenesis by Myc and C/EBP alpha. *Science*, 256:379–382, 1992.

[188] Reichert M, Eick D. Analysis of cell cycle arrest in adipocyte differentiation. *Oncogene*, 18:459–466, 1999.

[189] Yeh WC, Bierer BE, McKnight SL. Rapamycin inhibits clonal expansion and adipogenic differentiation of 3T3-L1 cells. *Proc Natl Acad Sci U S A*, 92:11086–11090, 1995.

[190] Qiu Z, Wei Y, Chen N, Jiang M, Wu J, Liao K. DNA synthesis and mitotic clonal expansion is not a required step for 3T3-L1 preadipocyte differentiation into adipocytes. *J Biol Chem*, 276:11988–11995, 2001.

[191] Entenmann G, Hauner H. Relationship between replication and differentiation in cultured human adipocyte precursor cells. *Am J Physiol*, 270:C1011–C1016, 1996.

[192] Cho YC, Jefcoate CR. PPARgamma1 synthesis and adipogenesis in C3H10T1/2 cells depends on S-phase progression, but does not require mitotic clonal expansion. *J Cell Biochem*, 91:336–353, 2004.

[193] Patel YM, Lane MD. Mitotic clonal expansion during preadipocyte differentiation: calpain- mediated turnover of p27. *J Biol Chem*, 275:17653–17660, 2000.

[194] Carpenter AE, Sabatini DM. Systematic genome-wide screens of gene function. *Nat Rev Genet*, 5:11–22, 2004.

[195] Gygi SP, Rochon Y, Franza BR, Aebersold R. Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol*, 19:1720–1730, 1999.

[196] Chen G, Gharib TG, Huang CC, Taylor JM, Misek DE, Kardia SL, Giordano TJ, Iannettoni MD, Orringer MB, Hanash SM, Beer DG. Discordant protein and mRNA expression in lung adenocarcinomas. *Mol Cell Proteomics*, 1:304–313, 2002.

Glossary

ABCA1	ATP-binding cassette, sub-family A, member 1
ANOVA	Analysis of variance
apoD	Apolipoprotein D
bHLH	Basic helix-loop-helix
BLAST	Basic local sequence alignment tool
BMI	Body Mass Index
BLAST	Basic local alignment search tool
BLAT	Blast like alignment tool
BSA	Bovine serum albumin
BTEB	Basic transcription element binding protein
C/EBP	CCAAT/Enhancer binding protein
cAMP	Cyclic adenosin monophosphate
CA	Correspondence analysis
CAP	c-Cbl-associated protein
CDK	Cyclin dependent kinase
C_T	Cycle threshold
CHOP	C/EBP homologous protein
CPT	Carnitine palmitoyltransferase
CREB	cAMP responsive element binding protein
SAM	Significance analysis of microarrays
DAG	Directed acyclic graphs
DMSO	Dimethylsulfoxide
DTT	Dithiothreitol
EGF	Epidermal growth factor
IGF	Insulin like growth factor
LXR	Liver X receptor
EC	Enzyme classification
EST	Expressed sequence tag
FANTOM	Functional annotation of mouse cDNA
FOX	Forkhead box
FGF	Fibroblast growth factor
GEO	Gene expression omnibus
GH	Growth hormone
GILZ	Glucocorticoid-induced leucine zipper

GO	Gene ontology
GOLD.db	Genomics of lipid-associated disorders database
H	Entropy
HMGI	High mobility group AT-hook 1
HMG-CoA	3-hydroxy-3-methylglutaryl-Coenzyme A
BSA	Bovine serum albumin
BUB1	Budding uninhibited by benzimidazoles 1 homolog
Id	Inhibitor of DNA binding
IFN	Interferon
IL	Interleukin
INCENP	Inner centromere protein
INSIG	Insulin-induced gene
IPI	International protein index
JDBC	Java database connectivity
KEGG	Kyoto encyclopedia of genes and genomes
KLF	Kruppel like factor
LOWESS	Locally weighted scatterplot smoothing
LPL	Lipoprotein lipase
MAGE-ML	Microarray gene expression markup language
MCSF	Macrophage colony-stimulating factor
MGED	Microarray gene expression data consortium
MI	Mutual information
MIX	Methylisobutylxanthine
MSE	Mean squared error
MT	Metallathionein
NFAT	Nuclear factor of activated T cells
NHS	N-hydroxysuccinimide
O/E1	Olf-1/early B-cell marker
NAC	None amplification control
NTC	None template control
NURR	Nuclear receptor related 1
NUR77	Nuclear receptor subfamily 4, group A, member 1
PCA	Principal component analysis
PCR	Polymerase chain reaction
PDGF	Platlet-derived growth factor

PGF	Prostaglandin F
PGI_2	Prostacyclin
PKB	Protein kinase B
PMT	Photomultiplier voltage
PPAR	Peroxisome proliferator activated receptor
PRM	Probabilistic relational models
Rb	Retinoblastoma protein
RNA	Ribonuclein acid
RXR	Retinoid X receptor
PWM	Position weight matrix
RT-PCR	Reverse transcription polymerase chain reaction
SAM	Significance analysis of microarrays
SCD	Stearoyl-CoA desaturase
SREBP1	Sterol regulatory element binding protein 1
SR-BI	Scavenger receptor class B member I
SAM	Significance analysis of microarrays
STAT	Signal transducers and activators of transcription
NTC	None template control
SOM	Self organizing maps
SOTA	Self organizing tree algorithm
SSC	Saline sodium citrate
SDS	Sodium dodecyl sulfate
SVD	Singular value decomposition
SVM	Support vector machines
TC	Tentative consensus sequence
TGI	TIGR Gene Indices
TGF	Transforming growth factor
TIFF	Tagged image file format
TNF	Tumor necrosis factor
TOPPRED	Topology prediction
TSS	Transcription start site
WAT	White adipose tissue
XML	Extensible markup language

Acknowledgement

Major parts of this work was supported by the Austrian Science Foundation, SFB project Biomembranes (F718). I would like express my deepest gratitude to my mentor Zlatko Trajanoski, who make this work possible, for his encouragment, visions, and believing on me. Further thank go to all previous members of the Bioinformatics group and people at the Institute of Genomics and Bioinformatics for fruitful discussions, support and their friendship. A special acknowledgment is dedicated to the people, who has contributed to this work: Fatima Sanchez Cabo, Barbara Di Camillo, Thomas Burkard, Alexander Sturn, Michael Maurer, Christine Paar, Roman Fiedler, Elmar Trost, Wolfgang Hofmann, Bernhard Mlecnik, Gernot Stocker, Andreas Prokesch, Renee Rubio, Jeremy Hasseman, Susanne Prattes, Elke Wagner, Roland Pieler, Alexander Schleiffer, Frank Eisenhaber Also I want to acknowledge John Quackenbush and the colleagues from TIGR, giving me the opportunity to get insights in the exciting world of genomic research. I'm indebted to my parents and family to accompanying me and for their support.

Publications

Journals

Hackl H, Burkard T, Sanchez Cabo F, Di Camillo B, Sturn A, Fiedler R, Paar C, Rubio R, Quackenbush J, Schleiffer A, Eisenhaber F, Trajanoski Z. Large scale gene expression analysis and functional annotation of adipocyte differentiation. *in preparation*

Hackl H, Maurer M, Mlecnik B, Hartler J, Trost E, Stocker G, Miranda Saavedra D, Trajanoski Z. GOLD.db: Genomics of Lipid-Associated Disorders Database. *submitted*

Hackl H, Sanchez Cabo F, Sturn A, Wolkenhauer O, Trajanoski Z. Analysis of DNA Microarray Data. *Curr Top Med Chem*, in press

Trost E, Hackl H, Maurer M, Trajanoski Z. Java Pathway Editor. *Bioinformatics*, 19:786-787, 2003

Pieler R, Sanchez Cabo F, Hackl H, Thallinger G, Z Trajanoski. ArrayNorm: Comprehensive normalization and analysis of microarray data. *Bioinformatics*, in press

Analysis of DNA Microarray Data

HUBERT HACKL[1], FATIMA SANCHEZ CABO[1,2], ALEXANDER STURN[1], OLAF WOLKENHAUER[3], AND ZLATKO TRAJANOSKI[1*]

[1]*Institute of Biomedical Engineering and Christian Doppler Laboratory for Genomics and Bioinformatics, Graz University of Technology, 8010 Graz, Austria*

[2]*Department of Biomolecular Sciences, UMIST, Manchester M60 1QD, U.K.*

[3]*Department of Computer Science, University of Rostock, 18059 Rostock, Germany*

ABSTRACT

Recent advances in DNA microarray technology have great impact on many areas of biomedical research and pharmacogenomics: discovering novel targets and genes, elucidating signatures of complex diseases, transcriptional profiling of models for diseases, and the development of individually optimized drugs based on differential gene expression patterns. Consequently; there is demand for robust methods for data analysis and the choice of adequate statistical tests. This review guides through all steps in the cDNA microarray data analysis pipeline and gives a basic understanding of the challenges in interpreting large microarray datasets.

INTRODUCTION

DNA microarray technology has become an important tool in biomedical research during the last years. All variants of this technology allow simultaneous profiling of expression levels of thousands of genes in a single experiment. The resulting profiles are patterns that are characteristic for the responses of cells or tissues to their environment, to differentiation into specialized tissues, or to dedifferentiation into neoplastic cells. The great potential of DNA microarrays lies not only in viewing the technology as a collection of individual expression measurements, but also in generating a composite picture of the expression profile of the cell. Therefore, microarrays are widely used in basis research as well as in clinical medicine and pharmacogenomics.

*To whom correspondence should be addressed:

Zlatko Trajanoski, PhD.
Institute of Biomedical Engineering, Graz University of Technology
Krenngasse 37
A-8010 Graz
Austria
Phone: +43-316-873-5332
Fax: +43-316-873-5340
Email: zlatko.trajanoski@tugraz.at

Running Title: microarray analysis
Key words: microarray, analysis, statistics, normalization, clustering

Application of microarrays in basic research

Functional genomics, the study of gene function through parallel expression measurements of a genome, can give information about the function of uncharacterized genes. Examining gene expression patterns of biological processes and molecular pathways as well as transcriptional profiling in development and differentiation gives insights into molecular mechanisms and can lead to the generation of new hypothesis for further investigations [1]. Coregulated genes show not only possible functional similarities, but can also share the same regulators [2,3]. During the drug discovery process microarrays are used for target discovery, examining large numbers of genes under relative few conditions using diseased tissues or animal models of disease [4]. Subsequently, targets are selected by certain criteria and the role of each gene in disease is investigated by examining time courses and dose-response curves. To validate potential candidates and identify phenotypic changes and possible side effects, knockout, knock-in, and gene silencing strategies in cells and model organisms are advantageously applied in combination with microarrays [5,6,7,8]. Profiling of the liver – the dominant site of drug metabolism – of drug-treated rats or primary human hepatocytes, gives a rounded picture about the activity of genes, regulated by a network of nuclear receptors that recognize both, xenobiotics and endogenous compounds and gives conclusion about the drug metabolism [4]. Microarray profiling of gene expression can also be used to elucidate mechanisms underlying the adverse events of drugs, to identify biomarkers for physiological events, and potentially even to predict adverse events before human exposure or in more general toxicity.

Application of microarrays in clinical medicine and pharmacogenomics

In pharmacogenomics, interindividual differences in gene expression profiles in the healthy state, the disease state, and after drug administration are studied. Recording of allelic variants in different individuals, evident by the growing number of detected single nucleotide polymorphisms (SNPs), embody also inherent characteristics of this evolving field. Main directions are: (1) identification of specific disease-associated genes, allelic variants or gene products, (2)

identification of gene or allelic variants that affect individual response to current drugs (allowing patient-specific drug stratification), (3) identification of disease-associated molecular expression profiles to pinpoint targets in various pathways, and (4) design and development of individual-optimized drugs based on differential genetic and expression characteristics [9,10,11,12]. Common drug treatments are effective in only 50-75% of all patients and, in some cases, such as specific cancer chemotherapies, the efficacy can be as low as 25% [13]. Consequently, the use of microarray technology in pharmacogenomic studies is proliferating.

The greatest potential for microarrays in clinical settings can be found in research and diagnosis of cancer. Tumor-based gene expression data from DNA microarrays add immense detail and complexity to the information available from traditional clinical and pathological sources; it is a snapshot of the total gene activity of a tumor, providing complex and detailed data on both, the inherent state of the patient and on the current characteristics of the tumor and disease state [14]. A variety of cancer studies and profiling of cancer cell lines were performed using DNA microarrays [14,15,16,17,18,19,20,21,22,23,24].

One can imagine, that the variety of applications requires different types of experiments, different kinds of microarray design, and analysis techniques. Despite the efforts from biologists, computer scientists, and statisticians there is no one-for-all solution. All biological conclusions and predictions resulted from microarray data rely on the quality of the data and the use of appropriate analytical methods and statistical tests. Subsequently, it is important to focus on the design of a microarray experiment. Carefully designed biological experiments are indispensable for the analysis of data and the interpretation of results. However, in microarray experiments analyzing thousands to millions of data points and deriving a meaningful interpretation requires efficient collecting of data and the use of computational methods to store the results and the experimental parameters in an organized way.
Despite the increasing number of papers on expression profiling using microarrays, there is still a lack of standardized procedures for analyzing the datasets. The challenge lies more in the choice of the most suitable method at every stage than in the proposal of new techniques. Furthermore, an understanding of both, the biology and the computational methods is essential for tackling the associated data mining tasks. The purpose of this review is therefore to summarize all important steps during the analysis process of microarray data and to give a basic understanding of the challenges in interpreting large microarray datasets. Several approaches for each basic step, from the microarray design to the biological interpretation, during the microarray analysis are discussed (Fig. 1).

MICROARRAY DESIGN

The two major platforms for microarrays are spotted arrays, where the probes are mechanically deposited on modified glass slides by contact or inkjet printing, and in situ arrays, where oligo probes (usually 20 to 25 nucleotides in length) are synthesized via photolithography and combinatorial chemistry techniques (GeneChip arrays, Affymetrix, Santa Clara, CA). In the latter approach, each gene or EST is represented on the array by probe pairs, consisting of a perfect match (PM) and a mismatch (MM) oligonucleotide that differs from the perfect match by only a single base in the center position. The purpose of the MM sequence is to capture the non-specific binding that distortion the measured intensity level of the PM. Note, that only one dye is required [27] for in situ arrays whereas for spotted microarrays a competitive hybridization of two RNA samples, labeled with two different fluorescent dyes is used.

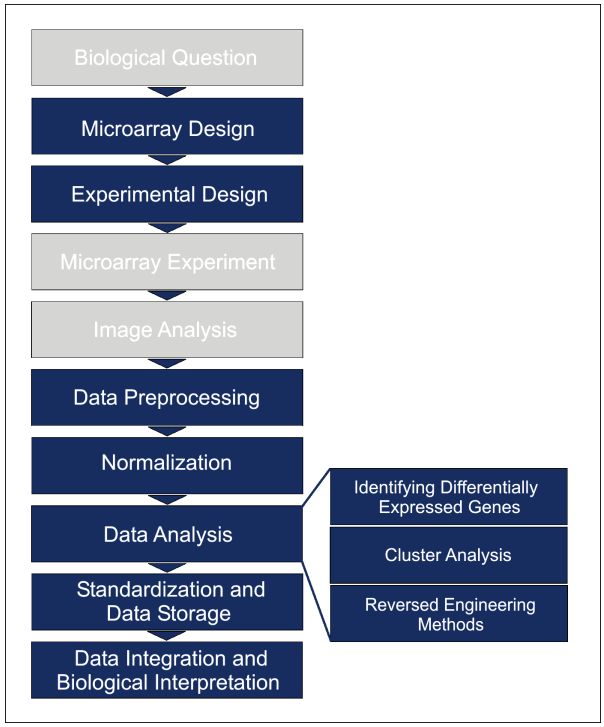

Figure 1: Procedure for microarray experiment and data analysis (boxes shaded in blue are covered by this review)

For spotted microarrays, one have the option to spot cDNA - in general PCR products (1,000-1,500 bp in length) of clones with an inserted cDNA element representing an expressed sequence tag (EST) or a gene - or oligonucleotides, designed for specific genes. Methods based on synthetic oligonucleotides require no time-consuming handling of cDNA resources [28]. In addition, the elements can be designed to represent the most unique part of a given transcript, making the detection of closely related genes or splice variants possible. Although short

oligonucleotides may result in less specific hybridization and reduced sensitivity using of presynthesized, longer oligonucleotides (50-100mers) counteracts these disadvantages [29]. Spotted arrays allow a greater degree of flexibility in the choice of arrayed elements, particularly for the preparation of smaller, customized arrays for specific investigations. In addition, arraying of unsequenced clones from cDNA libraries or clones for ESTs not similar to one characterized gene can be useful for gene discovery and functional annotation. The arrangement and the number of features on microarray slides raises design issues that have impact on the normalization and the microarray data analysis [30]. Repeated positioning of the same element on the array increases precision of the measurement by averaging the intensities and can minimize problems caused by contamination of the surface. The use of internal control features (e.g. features for genes of other organisms) also help to ensure the quality of the data.

EXPERIMENTAL DESIGN

The general principles discussed here are for two-color microarray experiments. However, many of these issues apply also to single-color gene expression assays. Microarray experiments should be treated as general biological experiments and microarrays as measurement of a biological quantity. The following issues should be considered for the design of a microarray experiment:

(1) *Biological replicates:* Biological variability is intrinsic to all organisms and can be substantial even for inbred mice [31]. Therefore it is necessary to perform repeated hybridizations with RNA samples from independent sources. The number of the biological replicates depends on the type of the biological question, e.g. whether two types of tissues differ with regard to the expression profiles. In this particular example it is obvious that one RNA sample for each tissue does not answer the biological question, since there may be a substantial biological variability within this tissues [26].

(2) *Technical replicates*: In microarray experiments there are two possibilities for replicated measurements to reduce variability introduced by measurement errors: replicated features within a slide and replication of hybridization with the same RNA samples. It is useful to have a few technical replicates to ensure that the procedures, reagents and equipment are working properly [26]. However, since technical replicates do not represent independent measurement, it is strongly recommended to use biological replication as principal source of replicated slides [32]. *Dye-swap* can be also used as a technical replicate and is useful for reducing the systematic bias. In this case replications represent repeated hybridization with the same samples, with swapped dyes in the second hybridization. This can also be used for gene-wise normalization to balance the systematic differences in the red and green intensities.

(3) *Pooling:* In addition to the optimal design there are constraints on the number of slides, the amount of RNA available or cost considerations, all of which will effect the experimental design. In cases where a large number of experimental units may be available and the number of slides is limited, there is the option to pool individual samples. This may also comprise cases with limited available amount of RNA. If all available samples are pooled together the biological variance are minimized, but all independent replication would be eliminated. Therefore it is better to use several pools and fewer technical replicates.

(4) Control versus reference RNA: For a pairwise comparison of different gene expression one of the two RNA samples, which are hybridized to a microarray slide, is used as control, e.g. RNA from an untreated cell line. In cases of comparison of many different types of samples, it would be desirable to use a more universal reference with a broad coverage of genes, e. g. RNA from pooled cell lines. This approach is commonly used if the focus of the analysis is to determine tumor subtypes[22].

In summary, DNA microarray experiments require careful planning of the experimental and the microarray design, driven by the objectives and conditioned by several constrains.

PREPROCESSING OF MICROARRAY DATA

A good data analysis should start with an exhaustive look at the quality of the data. For data generated from microarray experiments, the quality of the measurements is strongly dependent on the technology used, since issues as the characteristics of the surface and the material arrayed are determinant in the intensity measures obtained. Hence, the quality problems that arise for two color and high density oligonucleotide chips must be faced separately.

Preprocessing of cDNA microarray data

Testing the quality of the data

Prior to further analysis poor quality spots must be removed from the data set. Most of the image analysis software provides some criteria to quantify the quality of a spot. For example, in GenePix (Axon Instruments Inc., Union City, CA) a spot can be manually

classified as good, bad, absent or not found. Genes with a very low expression value are often removed in order not to confound their signal with the background intensity. Saturated spots are also excluded from the analysis because the scanner cannot provide a reliable measurement of their intensity value. In addition, the quality of a spot can be related to its shape, filtering out those genes for which the standard deviation of its pixels is too big or those for which mean and median values are very different. All thresholds should be calculated based on statistical analysis.

Background correction

The background intensity quantifies the amount of material attached to a slide, even where no spotted material is available. This fluorescence will be added to the true intensities of all spots in the slide. In consequence, background correction must be performed in order to improve the accuracy of the expression values. There are several methods to estimate local and global background [24]. The most common approaches to remove the background effect are:

(1) To filter out those genes for which the estimator of the background intensity is higher than the estimator of the foreground intensity. For the rest of the genes, the background intensity will be overall much lower than the foreground intensity so the subtraction of the background intensity from the foreground intensity will not be necessary.

(2) To subtract the background intensity from the foreground intensity. A problem may arise because negative values could be obtained if the very low intensity spots were not previously removed [35].

(3) If background intensity is bigger than foreground intensity the reason can be that the estimators provided by the image analysis software for either the foreground (usually the median of the pixels within the spot) or the background were not correct. Kooperberg et al. [36] present a novel algorithm based on Bayesian Statistics to estimate the foreground and background from the pixel intensities.

Preprocessing of high density oligonucleotide chip data

Since Affymetrix (Affymetrix Inc., Santa Clara, CA) is the main manufacturer for high density oligonucleotide arrays we will hereafter refer to them as Affymetrix chips [27]. As previously described, several probes from a given gene are arrayed at different locations. Because we are interested in the expression level of the gene, a measurement summarizing the intensity of all probes of this particular gene should be provided. Irizarry et al. [37] review the main expression measures proposed. From them, the Average Difference (AvDiff), the Li and Wong's model [38] and the MAS 5.0 [39] are based on the difference $d_{ij}=(PM_{ij}-MM_{ij})$, where $j=1,\ldots,n_{probes}$ per gene *i*. AvDiff is a simple linear combination of the consistent d_{ij}'s while the Li and Wong's model is already more sophisticated. Affymetrix also realized the necessity of a non-linear method and provided it to its users with the MAS 5.0 based on the Tukey Biweight statistic. Irizarry et al. [37] showed how in most of the cases, the MM_{ij} is capturing not just non-specific binding but also signal. Their novel expression measure is the Robust Multichip Average (RMA) and it is shown to perform better than traditional measurements. This new method is based on the estimation of the background, containing the non-specific signal together with the optical noise [37] and has some similarities with the Bayesian estimator of the background of cDNA microarrays proposed in [36].

NORMALIZATION OF MICROARRAY DATA

Microarray experiments generate a substantial amount of experimental variation, making it difficult to identify the biological variation of interest. The term normalization generally refers to the reduction of a given data set to a standard or normal state. For microarray data, this will be achieved by removing all the non-biological variation introduced in the measurements [40,41] and minimizing the random errors. Only after normalization microarray data will be reliable and comparable.

Sources of systematic error (sample effect, array effect, dye effect and gene effect [42,43]) and also random errors will have different effects depending on the technology used, i.e. cDNA arrays or high density oligonucleotide arrays. Some methods traditionally applied to cDNA microarrays could be also used to normalize data generated by high density oligonucleotide chips. These methods are Analysis of Variance (ANOVA) and Singular Value Decomposition (SVD). The general statistical tool ANOVA is often applied to normalize and analyze microarray data [42,43,44]. Based on the experimental design of the particular experiment to analyze, this method fits a model and provides estimators for the different sources of non-biological variation that contribute into the measured intensity value. Subtracting these estimators from the measured values, a reliable estimator of the intensity measurement can be obtained. This estimator can be used, without further calculations, to estimate the change in the expression level of a particular gene across different conditions.

The use of Singular Value Decomposition (SVD) to normalize microarray data was recently proposed [45]. In this case, the hypothesis is that all non-biological

variability will be summarized in one of the Principal Components (PC). The data set would be normalized removing the noisy PC.

Methods to normalize cDNA microarray data

The different properties of the two dyes used to label the RNA samples to be compared, make the dye effect the most important among the systematic effects to be corrected in cDNA microarrays. The properties that are different for the two dyes are the lower incorporation rate of Cy5, the quantum yield, the photobleaching, and the quenching properties [46]. In consequence, the measures obtained from two samples labeled with different dyes are not comparable and normalization must be applied to the data in order to bring them to the "same scale". The correction of the dye effect is also regarded as "within-slide normalization". The most popular methods are based on the assumption that the majority of the genes are equally expressed in both channels. However, this is not the case for all experiments. According to [47], there are two main approaches to correct the dye effect: (1) to use the whole data set to normalize the data, also known as self-consistency, or (2) to use the quality elements provided in the experiment. This includes the self-normalization [40], the use of spotted controls [48,49] and the use of a reference channel to normalize the data [31,32,50,51].

Normalization by self-consistency

Assuming that most of the genes should be equally expressed in both channels, an expression ζ_i is estimated to force the overall intensity of both channels (the two samples of RNA hybridized onto the same slide and labelled with the two dyes) to be the same. Both channels intensities would be then related according to the expression $R = \zeta_i G$, where R stands for the RNA sample labelled with red (Cy5) and G represents the RNA sample labelled with green (Cy3). Depending on the method used to estimate this expression ζ_i the data will be normalized in different ways. Visualization of the data using e.g. boxplots and scatterplots is essential at this stage in order to choose the normalization method that fits best to the data. These visual methods are also useful to test the effect of normalization in the data. The fit of a *LOWESS* function [52] to the data scattered in a M-A plot (where $M=\log_2(R/G)$ and $A=1/2\cdot(\log_2 R + \log_2 G)$) is becoming a common practice to normalize microarray data [40]. Since it is known that the dye effect is intensity dependent [40], this method is more appropriate than the simple global normalization method, that estimates ζ=median (R/G), which is constant across the whole intensity range. Another common normalization method estimates ζ_i for every gene but based on linear regression methods [53]. In most of the cases, the relationship between both channels is not linear so the use of non-linear techniques such as *LOWESS* is more suitable. In addition, the regressive approach is very sensitive to outliers and *LOWESS* offers a more robust estimation. Although *LOWESS* performs well, some new non-linear fittings have been proposed lately [48, 54]. In addition, different overall expression level can be observed in particular regions of the slide. If the difference corresponds to groups of genes printed with different print-tips, this effect can be corrected estimating a different *LOWESS* function for every subgrid instead of to the whole data set [40]. The correction of spatial systematic effects other than print tip effects is winning increasing importance in the normalization of microarray data [55, 56].

Normalization using quality elements

In general, there are many experiments for which the assumption that most of the genes are equally expressed cannot be known "a priori" or for which a very different number of genes is expected to be differentially expressed in both channels. For example, "boutique arrays" [48] are becoming an extended practice in medical research: After performing a high density spotted array where even the whole genome of an organism under study could have been arrayed, the genes found to be differentially expressed are spotted in another "low-density" array. In those cases the normalization of the data cannot be done using methods previously described and for this purpose quality control elements available from the experimental design should be used.
If enough material is available to replicate the experiment at every one of the biological conditions of interest, self-normalization [40] can be useful. Dobbin et al. [50] showed that to perform dye swap just for half of the biological conditions to be tested should give enough information to remove the dye effect from the data. Self-normalization appears as a natural method to remove the dye effect because the expression level of every gene in the two RNA pools of interest is estimated with the average value of every pool labelled with the two dyes. Given two arrays for which the same samples were labelled with a different dye each time, for every spotted gene *i* it can be proved that under the assumptions estated in [57] then $1/2\cdot(M_i-M_i')\sim$*(true log expression ratio)*, where $M_i=\log_2 (R_i/G_i)$ and M_i' is the same for the slide for which the dyes were swapped. The main advantage of the self-normalization is that it transforms the data while preserving the characteristics of every particular gene. In addition, the computational cost for the calculations is very low. If controls covering the whole intensity range are available, the data can be normalized according to them [40, 48, 49]. For controls for which both channels expression level is expected to be the same (e.g., Microarray Sample Pool (MSP) [40]), a non-linear function can be fitted to the M-A plot of the controls and used to correct the entire data set. However, because the number of controls

available per slide is usually not very large, the fit of a quadratic function using the Levenberg-Marquardt method [58] can be more appropriate than to use the *LOWESS* function.

After within-array normalization, across arrays normalization might also be necessary to make comparable values across different biological conditions. Besides the systematic error introduced in every measurement, there is a random error that cannot be estimated. The only way to reduce the intrinsic variability of a given measurement is replicating measurements. In microarrays, replicated spots and replicated slides (both referred to as technical replicates) are merged after normalization of the data to obtain a final value more reliable than the original one.

Normalization of high density oligonucleotide chip data

As for cDNA microarrays, the measurements obtained with Affymetrix chips contain non-biological variation derived from experimental errors such as differences in sample preparation, manufacture of the arrays, or problems during labeling, hybridization, and scanning [37]. According to [59] two main approaches can be used to correct systematic errors in the Affymetrix data.

The *baseline methods* correct the data based on the expression measure (i.e. AvDiff, Li-Wong's model [38], MAS 5.0 [39] or RMA [37]) and strongly depend on the selection of a baseline array against which all slides must be normalized. Affymetrix [39] propose a global normalization forcing all the arrays in ane experiment to have the same mean value. An improvement to this method is to fit non-linear functions to model the relationship between the baseline array and any of the other slides to be normalized, but just for the subset of rank-invariant genes [38, 60].

Table 1. Overview of normalization methods for cDNA microarrays and for high density oligonucleotide chips

		cDNA microarrays		**High-density oligonucleotide chips**
Non-specific hybridization	Background correction	1. Filtering 2. Subtraction 3. Bayesian method	d=(PM-MM) RMA	1. AvDiff 2. Li&Wong 3. MAS 5.0
Systematic errors	Self consistency	1. Global 2. Linear regression 3. LOWESS 4. LOWESS print-tip 5. Local mean norm.	Baseline methods	1. Global 2. Splines smooth
			Complete data methods	1. Cyclic LOESS 2. Quantile
	Quality elements	1. Self-normalization 2. Controls 3. Reference channel	Controls	

On the other hand, the *complete data methods* correct the data at the probe level. Every probe will be normalized and the expression level of a gene will then be obtained summarizing the normalized intensity level of its probes. The main two methods described in [59] are the cyclic *LOESS* and the quantile normalization. The first is an extension of the *LOWESS* method that applies to cDNA arrays, but the samples are now hybridized onto different slides so all possible pair-wise comparisons must be normalized. The quantile normalization is based on the idea that two data sets have the same distribution if the data in a QQ-plot scatters around a slope line of 45˚. This simple principle can be easily extended to a set of *n* arrays from a microarray experiment, without the need to choose a baseline or reference array.

DETECTION OF DIFFERENTIALLY EXPRESSED GENES

The discussion about the suitability of the normality assumption for microarray data has been one of the core tasks for the microarray data analysts in recent years. Several data transformations have been proposed in order to obtain normally distributed data [35,61]. Gene expression data in logarithmic scale presents a roughly normal distribution, since the

estimated histograms are unimodal and symmetric (Figure 2). Therefore it seems quite reasonable to accept the normality assumption. The consequences of taking this hypothesis as true are several and of great impact in the analysis of microarray data: (1) Simple error models can be formulated to explain, e.g., the variability of the data using ANOVA techniques [42,43]; (2) clustering algorithms (supervised [62] and unsupervised [18,63]) can be applied to the data, (3) reverse engineering methods [64,65] can be used to model gene-gene interactions, etc. In contrast, the subtle but essential task of detecting differentially expressed genes often requires the use of non-parametric tests [66, 67, 68, 69, 70, 71, 72] because the tails of the distribution of gene intensity values are heavier than for normally distributed data (Figure 1). One of the simplest approaches to discover differentially expressed genes consists in removing one by one all those values that make the tails of the distribution of expression levels longer than for normal distributed data. The process is repeated until the data can be considered as normally distributed [48]. Another simple method presented in [67] detects potentially differentially expressed genes looking at the QQ-plot. The values that deviate from a straight line of slope 45˚ are potential differentially expressed genes. Moreover, a simple fold-change detector can be used to detect differentially expressed genes.

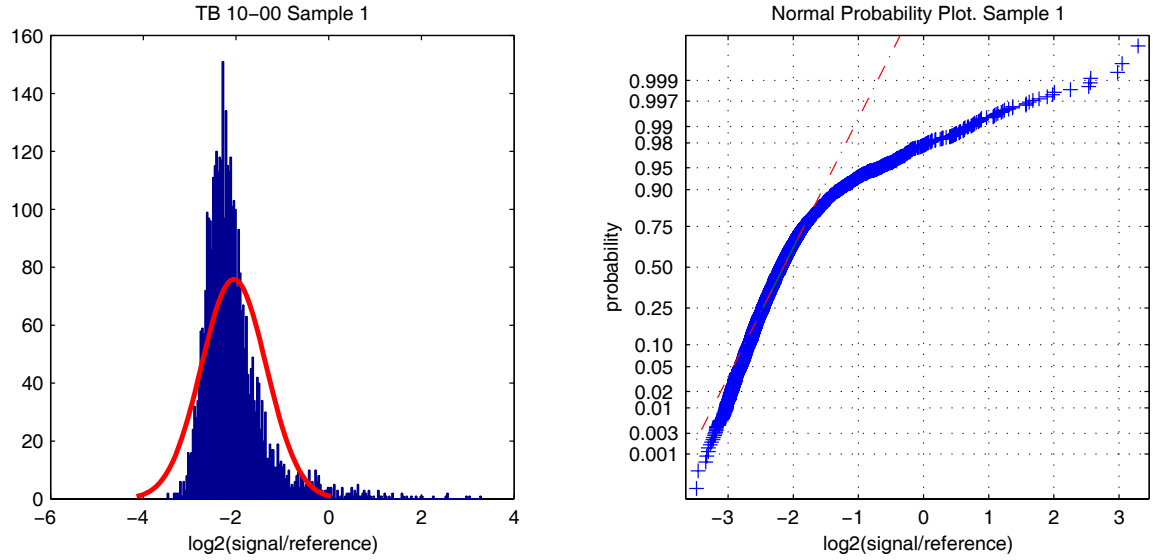

Figure 2. Histogram and QQ-plot of the gene expression data of one slide from a cDNA microarray experiment before normalization. The distribution is unimodal but the right-hand tail is much heavier than the tail of normally distributed data (grey line). In the QQ-plot an agreement between the two distributions up to the higher 10% percentile can be observed.

For quantitative detection, the calculation of a t-statistic is the most applied practice, which can also be generalized to multiple groups via ANOVA F statistic. However, the p-value corresponding to each t-statistic doesn't have necessarily to be calculated based on the student-t distribution [67] and must be adjusted for two reasons: First, the use of parametric methods to detect differentially expressed genes might not be suitable in many situations for which the data is not normal. In addition, due to the great number of null hypothesis tested at the same time (number of spotted genes > 4000), the number of false discoveries must be controlled. For example, if the null hypothesis "gene *i* doesn't change significantly its expression level from condition 1 to condition 2", at a level α=0.05 has to be tested for a microarray study with 4000 genes, the null hypothesis will be rejected being true 200 times (0.05*4000). In consequence, two hundred genes are wrongly selected as differentially expressed. Hence, the p-values should be corrected in order to control the family-wise type I error rate (FWE) [67]. Dudoit et al. [67] describe several methods to adjust the p-values. The single step methods (Bonferroni and Sidák [73]) perform equivalent multiplicity adjustment for all hypotheses. Step-down methods order p-values varying the adjustments needed. The step-down algorithm proposed in [67] adjusts p-values using the permutation based method proposed by Westfall and Young [74]. Another possibility is to bound the false discovery rate (FDR) as proposed by Benjamini and Hochberg [75].A review of the most popular tests to detect differentially expressed genes can be found in [76]. The most used tests to detect differentially expressed genes are (1) the linear regression model [77], (2) the regularized t-test [78], (3) Significance Analysis of Microarray (SAM) [79], and (4) the mixture model approach [80]. Prior to conducting further analyses, the data from genes exhibiting little variation across the different conditions should be excluded. The rational of this filtering process is that genes exhibiting little or no variation across samples

do not contribute valuable information for distinguishing among specimens. Selecting genes or samples with high intensities, fold changes and little missing values is leading to a cleaner dataset with less noise for further analysis. Because the discovery of new targets is one of the main aims of microarray data, many are already the applications of the methods previously described to detect differentially expressed or "active" genes [18,81,82,83].

CLUSTER ANALYSES

With the advent of novel experimental techniques for large-scale, genome-wide transcriptional profiling via microarrays or gene chips, a new field of gene expression data analysis emerged [21,53,64,84-99]. This new momentum to the bioinformatics community has fueled the hope of getting more insight into the processes conducted in a cell, tissue or organism. However, as more and more researchers adopted the microarray technology it soon became increasingly clear that simple data generation is not satisfactory and the challenges lie in storage, normalization, analysis, visualization of results, and most importantly in extracting meaningful biological information about the investigated cellular processes. Therefore, considerable progress has been made in the last couple of years to handle and analyze the millions of data points accumulated by state of the art microarray studies with tens of thousands of sequences per slide and maybe hundreds of slides. Nowadays so many options are available that choosing among them is challenging. Not only bioinformaticians themselves but also bioscientists and physicians using the computational tools need profound skills in bio- and computer sciences. To create and interpret results in an efficient, meaningful and responsible way, at least a fundamental understanding of the used technologies, algorithms, and methods is indispensable. In the following section we will give a brief insight into some common themes in the field of gene expression data analysis and existing computational approaches without claiming to be comprehensive, as new, more sophisticated techniques are being developed perpetually.

Data representation

The true power of microarray analysis does not come from the analysis of single hybridization, but rather, from the analysis of many hybridizations under different experimental conditions to identify common patterns of gene expression. A collection of gene expression data can be viewed abstractly as a table or matrix with rows representing genes, columns representing various samples, and each position in the table describing the measurement for a particular gene in a particular sample. This table is commonly referred as gene expression matrix. After normalization, the data for each gene are typically reported as a fluorescence or expression ratio from the two samples compared, hybridized to the same or to a number of arrays. Since almost all results from cDNA microarray experiments are ratios, overexpressed genes are represented by a range between 1 and ∞, whereas underexpressed genes are squashed between 0 and 1. To overcome this discrepancy, the data is usually transformed into logarithmic space, where overexpressed genes are assigned to positive values, underexpressed genes are assigned to negative values and a gene expression at a constant level (with a ratio of 1) is represented by zero. The advantage of this transformation is simply that the data is represented in a more "natural" way. In addition, microarray data in a logarithmic scale is easing further analysis by following an approximately normal distribution. In most cases the *logarithmus dualis* ($\log_2$) is used instead of $\log_{10}$ or *logarithmus naturalis* (ln), because of the better scaling and the more natural understanding of differences in terms of double and half.

Graphical data representation

Since the massive collection of numbers is difficult to assimilate, the primary data is combined with a graphical representation by displaying each data point with a color that quantitatively and qualitatively reflects the original experimental observation, i.e. each data point of the matrix is colored on the basis of the measured fluorescence ratio. This yields to a representation of complex gene expression data that allows assimilation and exploration of the data in a more intuitive manner. The most commonly used color scale ranges usually from saturated green (max. neg. $\log_2$ value) to saturated red (max. pos. $\log_2$ value). Values with $\log_2$ (ratio) close to zero (genes unchanged) are colored black, increasingly positive $\log_2$ (ratio) values with reds of increasing intensity, and increasing negative $\log_2$ (ratio) values with greens of increasing intensity. This means that for each element in the matrix, the relative intensity represents the relative expression, with brighter elements being more highly differentially expressed. Due to various effects during spotting, hybridization, and data analysis, not each spot can be assigned a meaningful ratio. This results in missing values in the data matrix. Missing values usually appear gray.

Data adjustment

In addition to filtering and normalization steps, several data adjustment procedures to enhance certain relationships are available and often used prior to analysis of a given data set. One commonly used procedure is to rescale each vector by subtracting the basal expression level from each al measurement, so that the average expression (mean or median) of each gene or sample is zero. This process referred to as mean or median centering can be used for instance to

remove certain types of biases. Another possibility is to adjust the data so that the minimum and maximum values are ±1, or so that the standard deviation or length of each expression vector is one. It is essential to know not only the mathematical background of these procedures, but also the effects such procedures can have on datasets based on biological observations. Adjusting data is changing the data and the effects of such manipulations can be very adjuvant or precarious, leading to inaccurate or misleading results and conclusions in respect to the investigated issue.

Vector Comparison

A gene expression matrix can be studied by either comparing its rows or columns and by looking for similarities or differences. If two rows are similar, it can be hypothesized that the respective genes are co-expressed or even co-regulated and possibly functionally related (Guilt By Association) [64, 93]. The comparison of experiments can provide information about which genes are differentially expressed in two or more conditions and this enables to study for instance effects that various compounds have on an investigated condition. The rows or columns of the gene expression matrix can be regarded as points in *n*-dimensional space or as *n*-dimensional vectors, where *n* is the number of genes or samples, respectively. Before any comparison of patterns of expression can be performed, a way to measure the similarity (or distance) between the vectors to compare need to be declared. Usually an analog value representing the distance between two vectors is computed by summing the distances between their respective vector elements. How this value is normalized and how the distance is computed depends on the distance measurement used. There is an extensive variety of algorithms available, ranging from simple Euclidian Distance or Pearson Correlation Coefficient to more sophisticated approaches like nonparametric or rank correlation coefficients (e.g. Spearman Rank-Order correlation or Kendall's Tau) or Mutual Information [100-102]. Possibly one "right" distance measure in the expression profile space does not exist, and the choice should depend on the question that is being addressed.
It has to be mentioned, that the only relationship that all of the aforementioned distance measuring approaches can detect are simple, one-to-one relationships. This restriction drastically reduces the amount of calculation required to find relationships between expression patterns, as only pair-wise, linear comparisons are made. The price of this speed is a lack of consideration of what are surely two of the hallmark characteristics of biological systems: the ability to make decisions based on multiple inputs, and the non-linearity of response. Additionally missing values are difficult to handle. A few are usually no problem, but if there are too many in comparison to the number of vector-elements *n*, an arbitrary result may be expected. However, the results obtained by clustering methods using these distance measurements demonstrate that such procedures are a useful way to extract and visualize one-to-one correlations from large datasets [103].

Clustering methods

The appearance of data sets that measure the relative abundance of mRNA of thousands of genes across tens or hundreds of samples has underscored the paucity of quantitative analytical tools available to examine such complex data. The scrutiny of expression profiles is currently in the phase of analysis that mathematicians term "data exploration" and that biologists often call "data mining". The underlying object of this level of analysis is the recognition of any non-random patterns or structures inherent in the data requiring further exploration.
Enabled by the increasing amount of public available microarray studies, comparative analysis of the transcriptome of different cell types, treatments, tissues or even among two or more model organisms promise to significantly enhance the fundamental understanding of the universality as well as the specialization of molecular biological mechanisms. The objective is to develop mathematical tools that are able to distinguish the similar from the dissimilar among two or more large-scale data sets.
Current methodologies to analyze gene expression data sets can be divided into two categories: supervised approaches, or analysis to determine genes or samples that fit a predetermined pattern; and unsupervised approaches, or analysis to characterize the components of a data set without the a priori input or knowledge of a training signal.
Supervised methods or class predictors are generally used for finding genes with expression levels that are significantly different between groups of samples and finding genes that actually predict a characteristic of samples. There are several published supervised methods that find genes or sets of genes that actually predict sample characteristics, such as distinguishing one type of cancer from another. These methods might find individual genes such as the Nearest Neighbor approach [18], and/or a set of genes, such as Decision Trees, Neural Networks and Support Vector Machines [104-106-3]. The latter is the most commonly used method and provides promising results in the area of gene classification [89].
Unsupervised Clustering methods attempt to identify relatively homogeneous groups or clusters of genes that behave similar across a range of conditions or samples. Clustering can be defined as the process of separating a set of objects into several subsets on the basis of their similarity [90]. The aim is generally to define clusters that minimize intracluster variability while maximizing intercluster distances, i.e. finding clusters, which members are similar to each other, but distant to members of other clusters in terms of gene

expression based on the used similarity measurement. The motivation to find such clusters is driven by the assumption that genes that demonstrate similar patterns (coexpressed genes) share common characteristics, such as common function, common regulatory elements, or common cellular origin. A wide range of approaches is available for gleaning insight from the data. However, despite the combined effort of biologists, computer scientists, statisticians and software engineers, there is no one-size-fits-all solution for the analysis and interpretation of genome-wide expression data. An appropriate choice of data-analysis technique depends both on the data and on the goals of the study. Still one of the most popular choices for the analysis of patterns of gene expression is assuredly the Hierarchical Clustering [62] in different variants. Standard agglomerative Hierarchical Clustering produces a representation of the data in the shape of a binary tree often referred to as dendrogram, where the most similar patterns are clustered in a hierarchy of nested subsets.

As an alternative to hierarchical clustering procedures, non hierarchical methods often start with a pre-defined number of clusters and, by iterative reallocation of cluster members, minister the overall intra-cluster dispersion and at the same time try to maximize the distance between clusters. Most noted representatives of non-hierarchical clustering are the k-means algorithms [107], Clustering Affinity Search Technique (CAST) [108], and neural network approaches like Self Organizing Maps (SOM) [109], the Self-Organizing Tree Algorithm (SOTA) [110], Relevance Networks [111] or Gene Expression Terrain Maps [112]. The complicacy using these techniques is that often considerable user input is required in form of difficult starting parameter specification. However, fortunately there are some approaches to circumvent this issue for instance by finding the number of clusters from the data itself in a self-acting, sophisticated way or simply by brute-force (examining all possibilities within a certain range) [113]. All theses clustering techniques can be used in combination with other exploratory techniques, such as Principal Component Analysis (PCA) [114] or related methods like Singular Value Decomposition (SVD) [45, 115] or Correspondence Analysis [116], that help the user to visualize the complexity of the data in a two- or three-dimensional space, allowing groups of related genes to be identified in a more intuitive way. Also related to PCA is a clustering method called Gene Shaving [117].

Post clustering analysis

Several topics of the analytical pipeline, namely image analysis, normalization, and gene expression data clustering and classification have been addressed in numerous publications, e.g. [103]. An extensive record of microarray software can be found on Y.F. Leung´s website (http://ihome.cuhk.edu.hk/~b400559/array.html). Data interpretation, however, proliferated just recently and leaves still a lot of room for new tools to extract knowledge from the increasing amount of microarray data. A key challenge of bioinformatics in the future will be to bridge this considerable gap between data generation and its usability by scientists for incisive biological discovery. Methods of choice may be the mapping of gene expression data onto sequence (e.g. chromosomes) [118] or biological pathways [119], automated literature search [120], as well as extensive promoter analyses [107].

REVERSED ENGINEERING METHODS

Traditionally, Biomolecular Science investigated genes or proteins one at a time allowing the identification of specific pathways. These gene-to-gene interactions can be converted into easy-to-read diagrams as those compiled in the BioCarta (http://www.biocarta.com/support/howto/gene.asp). High-throughput technologies have changed this area of research. New approaches of gene expression pattern analysis are being already applied to try to gain insight into the underlying regulatory processes in a given organism. However, due to the early stage of these technologies it is still difficult to apply these methods. The main problem is often the unbalance between the large number of genes to be modeled and the small number of samples available. To reduce the complexity of the problem the data can be discretized [121] or resampled to obtain more sampling points [122]. Details about the influence on the network inference of the network topology and the way of collecting the data are discussed in [123]. A review of the three main approaches to model the dynamics of a system using expression data can be found in [64, 65]. These three main techniques are Boolean and Bayesian networks, and ordinary differential equations (ODEs). There are differences and similarities among them and some are more appropriate than others depending on the particularities of the problem to be studied. In principle, Bayesian networks [121] and ODEs [124] should enable the formulation of a quantitative model using the continuous measurements available. However, the number of genes that can be modeled using a Bayesian approach is often small due to the computational cost of training a Bayesian network [125]. The same can be applied to ODEs since the number of parameters to be estimated increases when a new gene is aimed to be modeled as a component of the target system. On the other hand, Boolean networks [126] can provide a qualitative model but do not give a measurement of the strength of the relationships. Furthermore, they make use of discrete data. Another controversial point of these techniques is the validation of results. For models proposed to understand small known pathways and how they change when the conditions are perturbed, the best way of validation is performing experiments.

In the case of networks explaining the unknown relationships among genes of a given organism, promoter study can be a tedious but challenging task [127, 128].
In spite of the complexity and drawbacks of the problem, the evolution of microarray data production to ever-larger and more complex data sets will enable to use this huge amount of information for developing innovative approaches to reverse engineer biological networks of molecular interactions, which may unravel the contribution of specific genes and proteins in the cellular context [64].

STANDARDIZATION AND DATA STORAGE

Although new innovative procedures to analyze genomic data are still desirable, one problem during the analysis of gene expression data is not the lack of algorithms and tools, but the multiplicity of practices available to choose from. Moreover, these methods are difficult to compare and each method has its own implementation and frequently a different data format and representation. This diversity of methods makes it difficult and time consuming to compare results from different analyses. Although all of these resources are highly informative individually, their benefit could be augmented if provided in combination with other methods in a unified and centralized environment. Therefore standardized data exchange and calculation platforms, which allow the straightforward and efficient application of different algorithms to the data one is interested in, are and will be highly welcomed by the research community. In the area of microarray databases and analysis tools the MGED (Microarray Gene Expression Data) Society (http://www.mged.org) proposes with MIAME (Minimum Information About a Microarray Experiment) [129] a standard to describe the minimum information required to unambiguously interpret and verify microarray experiments. In addition to MIAME, which is required by several journals for manuscript submission, the Microarray Gene Expression – Markup Language (MAGE-ML) has been designed for microarray data exchange based on the XML standard [130]. Moreover, the ontology working group of the MGED project is developing an ontology for describing samples used in microarray experiments. In addition, the Gene Ontology (http://www.geneontology.org) provides a structured and standardized vocabulary to describe gene products in any organism [131]. The adoption of common standards and onthologies will be of immediate benefit to researchers, scientific journals, and those developing data management systems and tools for data analysis.
Microarray expression analysis typically generates a huge quantity of data. Effective and efficient analyses relies on capturing and recording data on the genes arrayed, the samples used for hybridization, the laboratory conditions and protocols used in the assay, and the parameters associated with obtaining and analyzing the hybridization data. The use of a well designed and comprehensive database in combination with (or including) a laboratory information management system (LIMS) is indispensable to efficiently query the data. Although at present there is no clear standard solution for microarray data storage and analysis software, there are a many open-source, public domain or commercial solutions vying for a share of this evolving market [132-140]. See also Y.F. Leung´s website (http://ihome.cuhk.edu.hk/~b400559/array.html) for an overview. These initiatives will maximize the value of microarray data by permitting greater opportunities for sharing information and thus for discovery, and will ultimately affect the description, analysis, and management of all high-throughput biological data. Standards and the uses of onthologies are essential to manage microarray data, especially in the context of public repositories.

DATA INTEGRATION AND BIOLOGICAL INTERPRETATION

The major challenge after computational analysis is to facilitate the process of looking for biological meaning in the data and to generate new hypotheses. Normally, investigators study the result lists or the gene clusters gene-by-gene. However, it is difficult to put them into a meaningful biological framework. Data interpretation methods proliferated just recently and are still leaving a lot of room for new tools to extract knowledge from the increasing amount of microarray data. A key challenge of bioinformatics in the future will be to bridge this considerable gap between data generation and its usability by scientists for biological discovery. Some methods have been applied successfully for this task, but still there is a lack in automatic tools for biological interpretation. Methods of choice may be the mapping of gene expression data onto sequence (e.g. chromosomes) [2, 118] or extensive promoter analysis [107]. The integration of pathway information with gene expression studies for instance has the potential to reveal differentially regulated genes under certain physiological conditions in a specific cellular component [119,141]. Probabilistic relational models (PRM) allow the inclusion of multiple types of information (e.g. cell type or clinical data) in the computational process itself [94,142]. Furthermore, integration of gene ontology (GO) terms and the automatic search for interesting literature clusters for whole expression sets will be also helpful for the analysis [120,143]. Another issue to be mentioned in this context is data integration. Genes and gene products do not function independently. They contribute to complex and interconnected pathways, networks and molecular systems. The understanding of these systems, their interactions, and their properties will require information from several fields, like genomics,

transcriptomics, proteomics, metabolomics or systematic phenotype profiles at the cell and organism level [144, 145]. Although all of these resources are highly informative individually, the collection of available content would have more efficacies if provided in a unified and centralized, or connected context. Database technologies and computational methods have to be improved to facilitate the integration and visualization of these miscellaneous data types, ranging from genomic data to biological pathways [146]. Sophisticated computational technologies, which establish relation-ships between genotype and the corresponding biological functions may yield to new insights about physiological processes in normal and disease states. Data integration may play also an important role for future trends in the prognosis of cancer and personalized medicine. According to [25,26] a few different generic types of objectives of microarray based cancer studies were pursued so far and successfully applied: comparison of expression profiles obtained from different predefined classes of specimen [17,18,27], determination whether there is a relationship between expression profiles and clinical outcome and development of a prognostic predictor of outcome (e.g. duration of survival)[25], and identification of novel subtypes of specimen within a population [22,23]. The combination of genomic data with some traditional clinical risk factors markedly improved the accuracy of predictions of breast cancer recurrence [14]. The integration of genomic data with heterogeneous patient data like clinical parameters, image data, biomarkers and proteomics data will allow more precise delineations of patients into subgroups that are more homogenous with respect to disease outcomes. Consequently, this leads to future advances in personalized medicine.

CONCLUSION

Microarray technology holds immense potential to improve human health and well-being. Although genome-based analysis methods are permeating biomedical research, the challenges of establishing robust paths from genomic expression information to improved human health remain immense. However, the ongoing investigations in these areas attempt to provide researchers with a markedly improved repertoire of computational tools that facilitate the translation of accumulated information into novel biological insights. This virtual workbench will allow the analyses of the function of genes and gene products in health and disease at an unprecedented level of molecular detail.

ACKNOWLEDGMENTS

The authors express their appreciation to the staff of the Institute of Biomedical Engineering for valuable comments and contributions. This work was supported by the bm:bwk, GEN-AU:BIN, Bioinformatics Integration Network and the Austrian Science Fund, Project SFB Biomembranes F718.
Fatima Sanchez Cabo was supported by an EU Marie Curie Training Site program “Genomics of Lipid Metabolism”.

REFERENCES

[1] Miki, R.; Kadota, K.; Bono, H.; Mizuno, Y.; Tomaru, Y.; Carninci, P.; Itoh, M.; Shibata, K.; Kawai, S.; Kawai, J.; Konno, H.; Watanabe, S.; Sato, K.; Tokusumi, Y.; Kikuchi, N.; Ishii, Y.; Hamaguchi, Y.; Nishizuka, I.; Goto, H.; Nitanda, H.; Satomi, S.; Yoshiki, A.; Kusakabe, M.; DeRisi, J. L.; Eisen, M. B.; Iyer, V.R.; Brown, P. O.; Muramatsu, M.; Shimada, H.; Okazaki, Y; Hayashizaki, Y. ; Delineating developmental and metabolic pathways *in vivo* by expression profiling using the RIKEN set of 18,1816 full-length enriched mouse cDNA arrays. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 2199-2204.

[2] Cohen, B. A.; Mitra, R. D.; Hughes, J. D.; Church, G. M. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat. Genet.* **2000,** *26,* 183-186.

[3] Bussemaker, H. J., Li, H., Siggia, E. D. Regulatory element detection using correlation with expression . *Nat. Genet*. **2001,** *27*, 167-171.

[4] Gerhold, D. L.; Jensen, R. V.; Gullans, S. R. Better therapeutics through microarrays. *Nat. Genet.* **2002**, *32 suppl*, 547-551.

[5] Marton, M. J.; DeRisi, J. L.; Bennet, H. A.; Iyer, V. R.; Meyer, M. R.; Roberts, C. J.; Stoughton, R.; Burchard, J.; Slade, D.; Dai, H.; Basset, D. E., Jr.; Hartwell, L.H.; Brown, P. O.; Friend, S. H. Drug target validation and identification of secondary drug target effects using DNA microarrays. Nature Med. **1998**, *4*, 1293-1301.

[6] Hughes, T. R.; Marton, M. J.; Jones, A. R.; Roberts, C. J.; Stoughton, R.; Armour, C. D.; Bennett, H. A.; Coffey, E.; Dai, H.; He, Y. D.; Kidd, M. D.; King, A. M.; Meyer, M. R.; Slade, D.; Lum, P. Y.; Stepaniants, S. B.; Shoemaker, D. D.; Gachotte, D.; Chakraburtty, K.; Simon, J.; Bard, M.; Friend, S. H. Functional discovery via a compendium of expression profiles. *Cell* **2000**, *102*, 109-126.

[7] Fambrough, D.; McClure, K.; Kazlauskas, A.; Lander, E. S. Diverse signaling pathways activated by growth factor receptors induce broadly overlapping, rather than independent, set of genes. *Cell* **1999**, *97*, 727-741.

[8] Hannon, G. J. RNA interference. *Nature* **2002**, *418*, 244-251.

[9] Pagliarulo, V.; Datar, R.H. Role of genetic and expression profiling in pharmacogenomics: the changing face of patient management. *Curr. Issues Mol. Biol.* **2002**, *4*, 101-110.

[10] Chicurel, M. E.; Dalma-Weiszhaus, D. D.; Microarrays in pharmacogenomics – advances and future promise. *Pharmacogenomics.* **2002**, *3*, 589-601.

[11] Evans, W. E.; Johnson, J. A. Pharmacogenomics: the inherited basis for interindividual differences in drug response. *Ann. Rev. Genomics Hum. Genet.* **2001**, *2*, 9-39.

[12] Evans, W. E.; Relling, M. V. Pharmacogenomics: translating functional genomics into rational therapeutics. *Science* **1999**, *286*, 487-491.

[13] Spear, B. B.; Heath-Chiozzi, M.; Huff, J. Clinical application of pharmacogenetics. *Trends Mol. Med.* **2001**, *7*, 201-204.

[14] Nevins, J. R.; Huang, E. S.; Dressman, H.; Pittman, J.; Huang, A. T.; West, A. M. Towards Integrated Clinico-Genomic Models for Personalized Medicine: Combining gene expression signatures and clinical factors in breast cancer outcomes prediction. *Hum. Mol. Genet.* **2003**, *0*, 2871-2880.

[15] Liotta, L.; Petricoin, E. Molecular profiling of human cancer. *Nat. Rev. Genet.* **2000**, *1*, 48-56.

[16] Cooper, C. S. Applications of microarray technology in breast cancer research. *Breast Cancer Res.* **2001**, *3*, 158-175.

[17] Hedenfalk, I.; Duggan, D.; Chen, Y.; Radmacher, M.; Bittner, M.; Simon, R.; Meltzer, P.; Gusterson, B.; Esteller, M.; Kallionemi, O.-P.; Wilfond, B.; Borg, A.; Trent, J. Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.* **2001**, *344*, 539-548.

[18] Golub, T. R.; Slonim, D. K.; Tamayo, P.; Huard, C.; Gaasenbeek, M.; Mesirov, J. P.; Coller, H.; Loh, M. L.; Downing, J. R.; Caligiuri, M. A.; Bloomfield, C. D.; Lander, E. S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **1999**, *286*, 531-537.

[19] van't Veer, L. J.; Dai, H.; van de Vijer, M. J.; He, Y. D.; Hart A. A. M.; Mao, M.; Peterse, H. L.; van der Kooy, K.; Marton, M. J.; Witteveen, A. T.; Schreiber, G. J.; Kerkhoven, R. M.; Roberts, C.; Linsley, P. S.; Bernards, R.; Friend, S. H. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **2002**, *415*, 530-536.

[20] Scherf, U.; Ross, D. T.; Waltham, M.; Smith, L. H., Lee, J. K.; Tanabe, L.; Kohn, K. W.; Reinhold, W. C.; Myers, T. G.; Andrews, D. T.; Scudiero, D. A.; Eisen, M. B.; Sausville, E.A., Pommier, Y.; Botstein, D., Brown, P. O., Weinstein, J. N. A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.* **2000**, *24*, 236-244.

[21] Clarke, P. A.; te Poele, R.; Wooster, R.; Workman, P. Gene expression microarray analysis in cancer biology, pharmacology, and drug development: progress and potential. *Biochem. Pharmacol.* **2001**, *62*, 1311-1336.

[22] Alizadeh, A. A.; Eisen, M. B.; Davis, R. E.; Ma, C.; Lossos, I. S.; Rosenwald, A.; Boldrick, J. C.; Sabet, H.; Tran, T.; Yu, X.; Powell, J. I.; Yang, L.; Marti, G. E.; Moore, T.; Hudson, J., Jr.; Lu, L.; Lewis, D. B.; Tibshirani, R.; Sherlock, G.; Sherlock, G.; Chan, W. C.; Greiner, T. C.; Weisenburger, D. D.; Armitage, J. O.; Warnke, R.; Levy, R.; Wilson, W.; Grever, M. R.; Byrd, J. C.; Botstein, D.; Brown, P. O.; Staudt, L. M. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **2000**, *403*, 503-511.

[23] Bittner, M.; Meltzer, P.; Chen, Y.; Jiang, Y.; Seftor, E.; Hendrix, M.; Radmacher, M.; Simon, R.; Yakhini Z.; Ben Dor, A.; Sampas, N.; Dougherty, E.; Wang, E.; Marincola, F.; Gooden, C.; Lueders, J.; Glatfelter, A.; Pollock, P.; Carpten, J.; Gillanders, E.; Leja, D.; Dietrich, K.; Beaudry, C.; Berens, M.; Alberts, D.; Sondak, V. Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **2000**, *406,* 536-540.

[24] Perou, C. M.; Sorlie, T.; Eisen, M. B.; van de, R. M.; Jeffrey, S. S.; Rees, C. A.; Pollack, J. R.; Ross, D. T.; Johnsen, H.; Akslen, L. A.; Fluge, O.; Pergamenschikov, A.; Williams, C.; Zhu, S.X.; Lonning, P. E.; Borresen-Dale, A. L.; Brown, P. O.; Botstein, D. Molecular portraits of human breast tumours . *Nature* **2000**, *406*, 747-752.

[25] Simon, R.; Radmacher, M. D.; Dobbin, K. Design of studies using DNA microarrays. *Genetic Epidemiology* **2002**, *23*, 21-36.

[26] Simon, R. M.; Dobbin, K. Experimental design of DNA microarray experiments. *BioTechniques* **2003**, *34*, S16-S21.

[27] Lockhart, D. J.; Winzeler, E. A. Genomics, gene expression and DNA arrays. *Nature* **2000,** *405,* 827-836.

[28] Schulze, A.; Downward, J. Navigating gene expression using microarrays - a technology review. *Nat. Cell. Biol.*. **2001,** *3*: E190-E195.

[29] Kane, M. D.; Jatkoe, T. A.; Stumpf, C. R.; Lu, J.; Thomas, J. D.; Madore, S. J. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res.* **2000,** *28*, 4552-4557.

[30] Yang, Y. H.; Dudoit, S.; Luu, P.; Lin, D. M.; Peng, V.; Ngai, J.; Speed, T. P. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*. **2002,** *30,* e15.

[31] Churchill, G. A. Fundamentals of experimental design for cDNA microarrays. *Nat. Genet.* **2002,** *32 suppl.*, 490-495.

[32] Yang, Y. H.; Speed, T. Design issues for cDNA microarray experiments. *Nat. Rev. Genet.* **2002,** *3*, 579-588.

[33] Kerr, M. K.; Churchill, G. A. Statistical design and the analysis of gene expression microarray data. *Genet. Res.* **2001,** *77*, 123-128.

[34] Yang, Y.H.; Buckley, M.J.; Dudoit, S.; Speed, T.P. Comparison of methods for image analysis on cDNA microarray data. *J. Comp. Graph. Stat.* **2002**, *11*, 108-136.

[35] 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*, 96-104.

[36] Kooperberg, C.; Fazzio, T. G.; Delrow, J. J.; Tsukiyama, T. Improved Background Correction for Spotted DNA Microarrays. *J. Comput. Biol.* **2002**, *9*, 55-66.

[37] Irizarry, R.A.; Hobbs, B.; Collin, F.; Beazer-Barclay, Y.D.; Antonellis, K.J.; Scherf, U.; Speed, T.P. Exploration, Normalization and Summaries of High Density Oligonucleotide Array Probe Level Data. *Biostatistics* **2003**, *2*, 249-264.

[38] Li, C.; Wong, W. Model-Based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci. U S A*. **2001**, *98*, 31-36.

[39] Affymetrix microarray suite user guide version 4 edition. Affymetrix Inc., Santa Clara, CA. **1999.**

[40] Yang, Y.H.; Dudoit, S.; Lin, D.M.; Peng, V.; Ngai, J.; Speed, T.P. Normalization for cDNA microarray data: A robust composite method adressing single and multiple slide systematic variation. *Nucleic Acids Res.* **2002***, 4*, e15.

[41] Quackenbush, J. Microarray data normalization and transformation**.** *Nat. Genet.* **2002**, *32 suppl.,* 496- 501.

[42] Kerr, K.; Martin, M; Churchill, G.A. nalysis of variance for gene expression microarray data. *J. Comput. Biol.* **2000**, *7*, 819-837.

[43] Kerr, K.; Churchill, G.A. Experimental Design for Gene Expression Microarrays. *Biostatistics* **2001,** *2*, 183-201.

[44] Engelen, K.; Coessens, B.; Marchal, K.; De Moor, B.M. MARAN: normalizing micro-array data. *Bioinformatics* **2003**, *19*, 893-894.

[45] Alter, O.; Brown, P.O.; Botstein, D. Singular value decmposition for genome-wide expression data processing and modelling. *Proc. Natl. Acad. Sci. U S A*. **2000**, *18*, 10101-10106.

[46] Tseng, G.C.; Oh, M.; Rohlin, L.; Liao, J.C.; Wong, W.H. Issues in cDNA microarray filtering: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.* **2001**, *29*, 2549-2557.

[47] Kepler, T. B.; Crosby, L.; Morgan, K. T. Normalization and analysis of DNA microarray data by self-consistency and local regression. *Genome Biol.* **2002**, *7*, research0037.1-0037.12.

[48] Wilson, D.L.; Buckley, M.J.; Helliwell, C.A.; Wilson, I.W. New normalization methods for cDNA microarray data. *Bioinformatics* **2003**, *19*, 1325-1332.

[49] Van de Peppel, J.; Kemmeren, P.; van de Bakel, H; Radonjic, M.; van Leenen, D.; Holstege, F. C. P. Monitoring global messenger RNA changes in externally controlled microarray experiments. *EMBO Rep.* **2003**, *4*, 387-393.

[50] Dobbin, K.; Shih, J.H.; Simon, R. Statistical design of reverse dye microarrays. *Bioinformatics* **2003**, *19*, 803-810

[51] Talaat, A.M.; Howard, S.T.; Hale IV, H.; Lyons, R.; Garner, H.; Johnston, S.A. Genomic DNA standards for gene expression profiling in Mycobacterium tuberculosis. *Nucleic Acids Res.* **2002**, *30*, e104.

[52] Cleveland, W.S. Robust Locally Weighted Regression and Smoothing Scatterplots. *J. Am. Stat. Assoc.* **1979**, *74*, 829-836.

[53] Quackenbush, J. Computational analysis of microarray data. *Nat. Rev. Genet.* **2001**, *2*, 418-427.

[54] Workman, C.; Jensen, LJ.; Jarmer, H.; Berka, R.; Gautier, L.; Nielsen, H.B.; Saxild, H.H.; Nielsen, C.; Brunak, S.; Knudsen, S. A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol.* **2002**, *3*, research0048.1-0048.16.

[55] Colantuoni, C.; Henry, G.; Zeger, S.; Pevsner, J. Local Mean Normalization of Microarray Element Signal Intensities across and Array Surface: Quality Control and Correction of Spatially Systematic Artefacts. *BioTechniques* **2002,** *6*, 1316-1320.

[56] Wernisch, L.; Kendall, S.L.; Soneji, S.; Wietzorrek, A.; Parish, T.; Hinds, J.; Butcher, P. D.; Stoker, N.G. Analysis of whole-genome microarray replicates using mixed models. B*ioinformatics* **2003,** *19*, 53-61.

[57] Yang, Y.H.; Dudoit, S.; Luu, P.; Speed, T.P. Normalization for cDNA Microarray Data.In *Microarrays: Optical Technologies and Informatics;*Bittner, M.L.; Chen, Y., Dorsel, A. N.; Dougherty, E. R.; Eds.; SPIE, Society for Optical Engineering, San Jose, CA, 2001**.**

[58] Marquardt, D.W. An Algorithm for Least Squares-Estimation of Nonlinear Parameters. *J. Soc. Industr. Appl. Math.* **1963**, *11*, 431-441.

[59] Bolstad, B.M.; Irizarry, R.A.; Astrand, M.; Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **2003**, *19*, 185-193.

[60] Schadt, E.; Li, C.; Eliss, B.; Wong, W.H. Feature extraction and normalization of algorithms for high-density oligonucleotide gene expression data. *J. Cell. Biochem.* **2001**, *84*, 120-125.

[61] Cui, X.; Kerr, M. K.; Churchill, G.A. Data transformations for cDNA microarray data. *Techn. report* **2001**, The Jackson Laboratory, Bar Habour, Maine.

[62] Eisen, M.B.; Spellman, P. T.; Brown, P. O.; Botstein**,** D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U S A*. **1998**, *95*, 14863-14868.

[63] Brown, M. P. S.; Grundy, W. N.; Lin, D.; Cristianini, N.; Sugnet, C. W.; Furey, T. S.; Ares, M.; Haussler, D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. U S A*. **2000**, *97*, 262-267.

[64] D'haeseleer, P.; Lian, S.; Somogyi. R. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* **2000**, *16,* 707-726.

[65] H. de Jong. Modeling and simulation of genetic regulatory systems: a literature review. *Journal of Computational Biology* **2002**, *9*, 67-103.

[66] Chen Y, Dougherty ER and Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J Biomed. Optics* **1997**, *2*, 364-367.

[67] Dudoit, S.; Yang, Y. H.; Callow, M. J.; Speed, T. P. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiements. *Statistica Sinica* **2002**, *12*, 111-139.

[68] Ideker, T.; Thorsson, V.; Siehel, A.F.; Hood, L.E. Testing for differentially-expressed genes by maximum likelihood analysis of microarray data. *J Comput. Biol.* **2000**, *7*, 805-817.

[69] Lee, M-L. T.; Kuo, F.C.; Whitmore, G. A.; Sklar, J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl. Acad. Sci. U S A*. **2000***, 97*, 9834-9839.

[70] Wolfinger, R.D.; Gibson, G.; Wolfinger, E. D.; Bennett, L.; Hamadeh, H.; Bushel, P.; Afshari, C., Paules, R. S. Assessing gene significance from cDNA microarray expression data via mixed models. *J Comput. Biol.* **2001,** *8* 625-637.

[71] Zhao, Y.; Pan, W. Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments. *Bioinformatics* **2003**, *19*, 1046-1054.

[72] Park, P.J.; Pagano, M.; Bonetti, M. A nonparametric scoring algorithm for identifying informative genes from microarray data. *Pac. Symp. Biocomput*. **2001**, 52-63.

[73] Sachs, L. *Applied Statistics - A Handbook of Techniques*. Springer, 1982.

[74] Westfall, P.H.; Young, S.S. Resampling-based multiple testing: examples and methods for p-value adjustment. In *Wiley series in probability and mathematical statistics*.; Wiley, 1993**;**

[75] Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc.* **1995**, *57*, 289-300.

[76] Pan, W. A Comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* **2002,***18*, 546-554.

[77] Thomas, J.G.; Olson, J.M.; Tapscott, S.J.; Zhao, L.P. An efficient and robust statistical modelling approach to discover differentially expressed genes using genomic expression profiles. *Genome Res.* **2001**, *11*, 1227-1236.

[78] Baldi, P.; Long, A. D. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes.*Bioinformatics* **2001**, *17***,** 509-519.

[79] Tusher, V. G.; Tibshirani, R.; Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U S A*. **2001**, *98*, 5116-5121.

[80] Pan, W.; Lin, J.; Le, C.T. A mixture model approach to detecting differentially expressed genes with microarray data. *Funct. Integr. Genomics* **2003**, *3*, 117-24.

[81] Roberts, C. J.; Nelson, B.; Marton, M. J.; Stoughton, R.; Meyer, M. R.; Bennett, H. A.; He, Y. D.; Dai, H.; Walker, W. L.; Hughes, T. R.; Tyers, M.; Boone, C.; Friend, S. H. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science*. **2000**, *287*, 873-880.

[82] Model, F.; Adorjan, P.; Olek, A.; Piepenbrock, C. Feature selection for DNA methylation based cancer classification. *Bioinformatics*. **2001**, *17 Suppl 1*, S157-164.

[83] Zhan, F.; Hardin, J.; Kordsmeier, B.; Bumm, K.; Zheng, M.; Tian, E.; Sanderson, R.; Yang, Y.; Wilson, C.; Zangari, M.; Anaissie, E.; Morris, C.; Muwalla, F., van Rhee, F.; Fassas, A.; Crowley, J.; Tricot, G.; Barlogie, B.; Shaughnessy, Jr., J. Global gene expression profiling of multiple myeloma, monoclonal gammopathy of undetermined significance, and normal bone marrow plasma cells. *Blood*. **2002**, *99*, 1745-1757.

[84] Brazma, A.; Vilo, J. Gene expression data analysis. *FEBS Lett*. **2000**, *480*, 17-24.

[85] Butte, A. The use and analysis of microarray data. *Nat. Rev. Drug. Discov*. **2002**, *1*, 951-960.

[86] Claverie, J. M. Computational methods for the identification of differential and coordinated gene expression. *Hum. Mol. Genet*. **1999**, *8*, 1821-1832.

[87] Detours, V.; Dumont, J. E.; Bersini, H.; Maenhaut, C. Integration and cross-validation of high-throughput gene expression data: comparing heterogeneous data sets. *FEBS Lett*. **2003**, *546*, 98-102.

[88] Epstein, C. B.; Butow, R. A. Microarray technology - enhanced versatility, persistent challenge. *Curr. Opin. Biotechnol*. **2000**, *11*, 36-41.

[89] Gaasterland, T.; Bekiranov, S. Making the most of microarray data. *Nat. Genet*. **2000**, *24*, 204-206.

[90] Gilbert, D. R.; Schroeder, M.; van Helden, J. Interactive visualization and exploration of relationships between biological objects. *Trends Biotechnol*. **2000**, *18*, 487-494.

[91] Heyer, L. J.; Kruglyak, S.; Yooseph, S. Exploring expression data: identification and analysis of coexpressed genes. *Genome Res*. 1999, *9*, 1106-1115.

[92] Holloway, A. J.; van Laar, R. K.; Tothill, R. W.; Bowtell, D. D. Options available--from start to finish--for obtaining data from DNA microarrays II. *Nat. Genet*. **2002**, *32 Suppl*, 481-489.

[93] Hughes, T. R.; Shoemaker, D. D. DNA microarrays for expression profiling. *Curr. Opin. Chem. Biol*. **2001**, *5*, 21-25.

[94] Kaminski, N.; Friedman, N. Practical approaches to analyzing results of microarray experiments. *Am. J. Respir. Cell Mol. Biol*. **2002**, *27*, 125-32.

[95] Miller, L. D.; Long, P. M.; Wong, L.; Mukherjee, S.; McShane, L.M.; Liu, E.T. Optimal gene expression analysis by microarrays. *Cancer Cell*. **2002,** *2*, 353-61.

[96] Orr, M.S.; Scherf, U. Large-scale gene expression analysis in molecular target discovery. *Leukemia*. **2002,** *16*, 473-7.

[97] Simon, R.; Radmacher, M. D.; Dobbin, K.; McShane, LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J. Natl. Cancer Inst*. **2003**, *95*, 14-18.

[98] Slonim, D. K. From patterns to pathways: gene expression data analysis comes of age. *Nat. Genet*. **2002**, *32 Suppl*, 502-508.

[99] Tamames, J.; Clark, D.; Herrero, J.; Dopazo, J.; Blaschke, C.; Fernandez, J. M.; Oliveros, J.C.; Valencia, A. Bioinformatics methods for the analysis of expression arrays: data clustering and

information extraction. *J. Biotechnol*. **2002**, *98*, 269-83.

[100] Michaels, G.S.; Carr, D.B.; Askenazi, M.; Fuhrman, S.; Wen, X.; Somogyi, R. Cluster analysis and data visualization of large-scale gene expression data. *Pac. Symp. Biocomput*. **1998,** 42-53.

[101] Fuhrman, S.; Cunningham, M.J.; Wen, X.; Zweiger, G.; Seilhamer, J. J.; Somogyi, R. The application of shannon entropy in the identification of putative drug targets. *Biosystems*. **2000**, *55*, 5-14.

[102] Korber, B. T.; Farber, R. M.; Wolpert, D. H.; Lapedes, A. S. Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proc. Natl. Acad. Sci. U S A*. **1993**, *90*, 7176-80.

[103] Dudoit, S.; Gentleman, R. C.; Quackenbush, J. Open source software for the analysis of microarray data. *Biotechniques*. **2003**, *Suppl.,* 45-51.

[104] Brown, M. P.; Grundy, W. N.; Lin, D.; Cristianini, N.; Sugnet, C. W.; Furey, T. S.; Ares, Jr., M.; Haussler, D. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. U S A*. **2000**, *97*, 262-267.

[105] Furey, T. S.; Cristianini, N.; Duffy, N.; Bednarski, D. W.; Schummer, M.; Haussler, D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*. **2000**, *16*, 906-914.

[106] Chow, M. L.; Moler, E. J.; Mian, I. S. Identifying marker genes in transcription profiling data using a mixture of feature relevance experts. *Physiol. Genomics*. **2001**, 5, 99-111.

[107] Tavazoie, S.; Hughes, J. D.; Campbell, M. J.; Cho, R. J.; Church, G. M. Systematic determination of genetic network architecture. *Nat. Genet*. 1999, *22*, 281-285.

[108] Ben-Dor, A.; Shamir, R.; Yakhini, Z. Clustering gene expression patterns. *J. Comput. Biol*. **1999**, *6*, 281-97.

[109] Tamayo, P.; Slonim, D.; Mesirov, J.; Zhu, Q.; Kitareewan, S.; Dmitrovsky, E.; Lander, E. S.; Golub, T. R. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. U S A*. **1999**, *96*, 2907-2912.

[110] Herrero, J.; Valencia, A.; Dopazo, J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*. 2001, *17*, 126-136.

[111] Butte, A. J.; Tamayo, P.; Slonim, D.; Golub, T. R.; Kohane, I. S. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. Proc. *Natl. Acad. Sci. U S A*. **200**0, *97*, 12182-12186.

[112] Kim, S. K.; Lund, J.; Kiraly, M.; Duke, K.; Jiang, M.; Stuart, J. M.; Eizinger, A.; Wylie, B. N.; Davidson, G. S. A gene expression map for Caenorhabditis elegans. *Science*. **2001**, *293*, 2087-2092.

[113] Yeung, K. Y.; Haynor, D. R.; Ruzzo, W.L. Validating clustering for gene expression data. *Bioinformatics*. **2001**, *17*, 309-18.

[114] Raychaudhuri, S.; Stuart, J. M.; Altman, R. B. Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac. Symp. Biocomput*. **2000**, 455-466.

[115] Alter, O.; Brown, P. O.; Botstein, D. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. Proc. *Natl. Acad. Sci. U S A*. **2003**, *100*, 3351-3356.

[116] Fellenberg, K.; Hauser, N. C.; Brors, B.; Neutzner, A.; Hoheisel, J. D.; Vingron, M. Correspondence analysis applied to microarray data. *Proc. Natl. Acad. Sci. U S A*. **200**1, *98*, 10781-10786.

[117] Hastie, T.; Tibshirani, R.; Eisen, M. B.; Alizadeh, A.; Levy, R.; Staudt, L.; Chan, W. C.; Botstein, D.; Brown, P. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol*. **2000**, *1*, RESEARCH0003.

[118] Sturn, A.; Quackenbush, J.; Trajanoski, Z. Genesis: cluster analysis of microarray data. *Bioinformatics*. 2002, *18*, 207-208.

[119] Dahlquist, K. D.; Salomonis, N.; Vranizan, K.; Lawlor, S. C.; Conklin, B. R. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat. Genet*. **2002**, *31*, 19-20.

[120] Jenssen, T. K.; Laegreid, A.; Komorowski, J.; Hovig, E. A literature network of human genes for high-throughput analysis of gene expression. *Nat. Genet*. **2001**, *28*, 21-28.

[121] Friedman, N.; Linial, M.; Nachman, I.; Pe'er, D. Using Bayesian Networks to analyze expression data. *Journal Comput. Biol*. **2000**, *7*, 601-620.

[122] Hastings, W.K. Monte Carlo Sampling methods using Markov chains and their applications. *Biometrika* **1970**, *57*, 97-109.

[123] Smith, V.A.; Jarvis, E.D.; Hartemink, A.J. Influence of network topology and data collection on network inference. *Pacific Symposium in Biocomputation* **2003**, 164-175.

[124] Chen, T.; Hongyu, L.H.; Church, G.M. Modelling Gene Expression with Differential Equations. *Pac. Symp. Biocomput*. **1999**, *4*, 29-40.

[125] Heckerman, D. A Tutorial on Learning with Bayesian Networks. In Learning in Graphical Models, Adaptive Computation and Machine Learning. *MIT Press, Cambridge, Massachusetts* **1999**, 301-354

[126] Akutsu, T.; Miyano, S.; Kuhara, S. Algorithms for inferring qualitative models of biological networks. *Pac. Symp. Biocomput*. **2000**, *5*, 290-301.

[127] Pilpel, Y.; Sudarsanam, P.; Church, G.M. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genetics* **2001**, *29*, 153-159.

[128] Segal, E.; Barash, Y.; Simon, I.; Friedman, N.; Koller, D. From Promoter Sequence to Expression: A Probabilistic Framework. *In Proceedings of the 6th international Conference*

on Research in Computational Moecular Biology (RECOMB) **2002**, 263-272.

[129] Brazma, A.; Hingamp, P.; Quackenbush, J.; Sherlock, G.; Spellman, P.; Stoeckert, C.; Aach, J.; Ansorge, W.; Ball, C. A.; Causton, H. C.; Gaasterland, T.; Glenisson, P.; Holstege, F. C.; Kim, I. F.; Markowitz, V.; Matese, J. C.; Parkinson, H.; Robinson, A.; Sarkans, U.; Schulze-Kremer, S.; Stewart, J.; Taylor, R.; Vilo, J.; Vingron, M. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet*. **2001**, *29*, 365-371.

[130] Spellman, P. T.; Miller, M.; Stewart, J.; Troup, C.; Sarkans, U.; Chervitz, S.; Bernhart, D.; Sherlock, G.; Ball, C.; Lepage, M.; Swiatek, M.; Marks, W. L.; Goncalves, J.; Markel, S.; Iordan, D.; Shojatalab, M.; Pizarro, A.; White, J.; Hubley, R.; Deutsch, E.; Senger, M.; Aronow, B. J.; Robinson, A.; Bassett, D.; Stoeckert, Jr., C. J.; Brazma, A. Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol*. **2002**, *3*, RESEARCH0046.1-RESEARCH0046.9.

[131] Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Res*. **2001**, *11*, 1425-1433.

[132] Gardiner-Garden, M.; Littlejohn, T. G. A comparison of microarray databases. *Brief Bioinform*. **2001**, *2*, 143-158.

[133] Anderle, P.; Duval, M.; Draghici, S.; Kuklin, A.; Littlejohn, T. G.; Medrano, J. F.; Vilanova, D.; Roberts, M. A. Gene expression databases and data mining. *Biotechniques*. **2003**, *Suppl*, 36-44.

[134] Stoeckert, Jr., C. J.; Causton, H. C.; Ball, C. A. Microarray databases: standards and ontologies. *Nat. Genet*. **2002**, *32 Suppl.*, 469-473.

[135] Brazma, A.; Robinson, A.; Cameron, A.; Ashburner, M. One-stop shop for microarray data. *Nature*. **2000**, *403*, 699-700.

[136] Brazma, A.; Parkinson, H.; Sarkans, U.; Shojatalab, M.; Vilo, J.; Abeygunawardena, N.; Holloway, E.; Kapushesky, M.; Kemmeren, P.; Lara, G. G.; Oezcimen, A.; Rocca-Serra, P.; Sansone, S. A. ArrayExpress--a public repository f< or microarray gene expression data at the EBI. *Nucleic Acids Res*. **2003**, *31*, 68-71.

[137] Sherlock, G.; Hernandez-Boussard, T.; Kasarskis, A.; Binkley, G.; Matese, J. C.; Dwight, S. S.; Kaloper, M.; Weng, S.; Jin, H.; Ball, C. A.; Eisen, M. B.; Spellman, P. T.; Brown P. O.; Botstein, D.; Cherry, J. M. The Stanford Microarray Database. *Nucleic Acids Res*. **2001,** *29*, 152-5.

[138] Saal, L. H.; Troein, C.; Vallon-Christersson, J.; Gruvberger, S.; Borg, A.; Peterson, C. BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data . *Genome Biol*. **2002**, *3*, SOFTWARE0003.1-SOFTWARE0003.6.

[139] Edgar, R.; Domrachev, M.; Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002, *30*, 207-210.

[140] Saeed, A. I.; 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.

[141] Forster, J.; Gombert, A. K.; Nielsen, J. A functional genomics approach using metabolomics and in silico pathway analysis. *Biotechnol. Bioeng*. **2002**, *79*, 703-712.

[142] Segal, E., Taskar, B., Gasch, A., Friedman, N., Koller, D.; Rich probabilistic models for gene expression. *Bioinformatics* **2001,** *98,* S243-S252

[143] Doniger, S. W., Salomonis, N., Dahlquist, K. D., Vranizan, K., Lawlor, S. C., Conklin, B. R. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol.* **2003,** *4,* R7.

[144] Collins, F. S.; Green, E. D.; Guttmacher, A. E.; Guyer, M. S. A vision for the future of genomics research. *Nature*. **2003**, *422*, 835-847.

[145] Baxevanis, A. D. The Molecular Biology Database Collection: 2003 update. *Nucleic Acids Res*. **2003**, *31*, 1-12.

[146] Diehn, M.; Sherlock, G.; Binkley, G.; Jin, H.; Matese, J. C.; Hernandez-Boussard, T.; Rees, C. A.; Cherry, J. M.; Botstein, D.; Brown, P. O.; Alizadeh, A. A. SOURCE: a unified genomic resource of functional annotations, ontologies, andgene expression data. *Nucleic Acids Res*. **2003**, *31*, 219-223.

GOLD.db: Genomics of Lipid-Associated Disorders Database

HUBERT HACKL, MICHAEL MAURER, BERNHARD MLECNIK, JÜRGEN HARTLER, ELMAR TROST, GERNOT STOCKER, DIEGO MIRANDA-SAAVEDRA AND ZLATKO TRAJANOSKI*

Institute of Biomedical Engineering and Christian Doppler Laboratory for Genomics and Bioinformatics, Graz University of Technology, Krenngasse 37, 8010 Graz, Austria

ABSTRACT

The GOLD.db (Genomics of Lipid-Associated Disorders Database) was developed to address the need for integrating disparate information on the function and properties of genes and their products that are particularly relevant to the biology, diagnosis management, treatment, and prevention of lipid-associated disorders. The database provides a reference for pathways and information about the relevant genes and proteins in an efficiently organized way. The main focus was to provide biological pathways with image maps and visual pathway information for lipid metabolism and obesity-related research. The GOLD.db provides also the possibility to map gene expression data individually to each pathway. Gene expression at different experimental conditions can be viewed sequentially in context of the pathway. Related large scale gene expression data sets were provided and can be searched for specific genes to integrate information regarding their expression levels in different studies and conditions. Additionally, analytic and data mining tools, reagents, protocols, videos, references, and links to relevant genomic resources were included in the database. GOLD.db is available at http://gold.tugraz.at.

INTRODUCTION

The excessive consumption of high calorie, high fat diets and the adoption of a sedentary life style have made obesity and atherosclerosis major health problems in Western societies. In the USA, over 50% of the population are over-weight (BMI>25) and close to 25% are considered obese (BMI>30) (1, 2). As a consequence, a large fraction of the population is at risk to develop a broad range of common, life-threatening diseases including non-insulin dependent diabetes, various hyper-lipidemias, high blood pressure and atherosclerosis.

*corresponding author:

Zlatko Trajanoski, PhD
Institute of Biomedical Engineering
Graz University of Technology
Krenngasse 37, 8010 Graz, Austria
Tel: +43-316-873-5332
Fax: +43-316-873-5340
Email: zlatko.trajanoski@tugraz.at

Keywords: adipogenesis, insulin signaling, Java, database, pathway, gene expression

Vascular disease including coronary heart disease and stroke is currently the major cause of death in the United States and in other industrialized nations.
At the root of obesity and atherosclerosis is an excessive deposition of neutral lipids. Adipose tissue accumulates predominantly triglycerides,
whereas macrophages along the blood vessel wall mainly accumulate cholesterol and cholesteryl esters.
Accordingly, a detailed understanding of the molecular mechanisms that govern the balance between lipid deposition and mobilization is fundamentally important for the prevention and improved treatment of disease. In addition to the apparent environmental components involved in the pathogenesis of disorders related to lipid and energy metabolism, a large number of studies have provided undisputed evidence that susceptibility genes contribute around 50% of the phenotype. These genes encode products involved in the cellular uptake, synthesis, deposition and/or mobilization of lipids. However, characterization of many if not most of these genes and their products remains rudimentary. Deficiencies in the current level of understanding extend to key enzymes such as important triglyceride hydrolases in adipose tissue (3) or cholesteryl ester hydrolases in macrophages, hormones, signal transduction pathways, and the regulation of the transcription of relevant genes.

While medical molecular biology traditionally associates single genes and gene products with diseases, a growing body of evidence suggests that several common disease phenotypes arise from the delicate interaction of many genes as well as gene-environment interactions. To elucidate the development of obesity and atherosclerosis, it will be necessary to analyze patterns of gene expression and relate them to various metabolic states. To discover novel genes, processes and pathways that regulate lipid deposition and mobilization, a departure from hypothesis-driven research and turn to a discovery-driven approach is necessary. The application of high-throughput technologies and genome-based analysis will provide the tools for the analysis of gene-gene and gene-environment interactions in a systematic and comprehensive manner.

To facilitate genomic research we have initiated the development of a system for storing, integrating, and analyzing relevant data needed to decipher the molecular anatomy of lipid associated disorders. In

order to provide a reference for pathways and information of the relevant genes and proteins in an efficiently organized way, we have created the Genomics Of Lipid-Associated Disorders database (GOLD.db). The GOLD.db integrates disparate information on the function and properties of genes and their protein products that are particularly relevant to the biology, diagnosis management, treatment, and prevention of lipid-associated disorders. The main focus was to provide biological pathways with image maps and visual pathway information. For each element in the pathway, specific information exists including structured information about a gene, protein, 3D-structure, gene regulation, function, literature, and links. The GOLD.db provides also the possibility to map gene expression data individually to each pathway. Additionally, analytic and data mining tools, reagents, protocols, videos, references, and links to relevant genomic resources were included in the database.

DATABASE DESCRIPTION

PATHWAYS

In order to construct the biological pathways of interest, we have developed a pathway editor. This drawing tool provides the possibility to draw elements – typically representing a gene as part of the pathway – and the connection between those elements. The benefit of this tool is that information can be appended to each element via an input mask. This information can be accessed by clicking on the corresponding element in the image map, which was saved and uploaded to the web page. To design this pathway service as flexible as possible, features are provided for the remove, up- and download of relevant pathways (image maps) including the underlying additional information of the elements. However, this service is on a restricted basis to prohibit unauthorized access. Since some pathways tend to become very detailed an option to search for genes or gene accession number, respectively, within the pathway was built in. The pathway editor is executable as a standalone application and is available from http://genome.tugraz.at (4). Currently annotated pathways are the insulin signaling pathway, the IGF-I pathway, and the adipogenesis regulatory network. Other pathways of lipid metabolism will follow in the near future. Available KEGG pathways can also be adapted with the pathway editor based on the provided XML files (5) and uploaded in the same way. Several relevant KEGG pathways for different organisms are already provided.

For each element in the pathway a specific information field exists. The field includes structured information about a gene, protein, 3D-structure, gene regulation, function, literature, and links. The GenBank accession number of the respective gene (typically a RefSeq number) acts as the primary key for the database entries and therefore the declaration of this identity is compulsory. Besides the gene name, symbol name and GenBank accession number for the gene, protein identities for the NCBI, the SWISS-PROT database, and the 3D structures databases can be specified, and the accession numbers displayed and linked to the appropriate databases. The body of the query strings for these links can be changed for all entries of the pathway at once. Since in the case of transcriptional networks, the binding of transcription factors to the DNA is of interest, in the gene regulation field options were implemented to upload and display sequences upstream of the transcription start site (usually the promoter sequence) and transcription factors known to bind to these upstream activator sequences. The description, localization and classification of the factors are entered by the annotator in plain text and are accessed in the same format. The references used to generate the content of the database entries can be appended, including a link to the PubMed entry. There is also the possibility to create a list of all reference entries for the pathway or a list of all upstream sequences in FASTA format, in order to search for transcription factor binding sites. If a clone for a specific gene is available in the clone resources, the clone name will be displayed automatically and a link with optional information about this clone is provided.

MAPPING OF GENE EXPRESSION DATA TO PATHWAYS

Through the integration of several types of biological information deeper insights into the molecular mechanisms and biological processes can be gained than just by the analysis of one type of experimental results. In the GOLD.db it is possible to map gene expression data (for instance results of microarray studies) to the corresponding elements of the available pathways similar to previous efforts (6). Either an individual or a provided gene expression data set can be used to visualize the gene expression at different experimental conditions sequentially in the context of the pathways. If an element (gene) of the pathway is included in the data set, the related symbol in the image map is color coded according to the relative gene expression or the log ratio in two color microarray experiments, respectively. As key for the mapped relation the RefSeq number (7) is used. Hence, only those elements in the data set file are mapped, where the RefSeq number in the data set is specified. For the KEGG pathways each element classified by the enzyme classification number (EC) is virtually subdivided into different corres-ponding RefSeq entries, since one EC is represented by one or more RefSeq entries.

GENE EXPRESSION DATA SETS

Analysis of gene expression patterns in animal models for lipid-associated disorders will help to understand

the fundamental gene relations and regulatory mechanisms responsible for the development of obesity related diseases. The huge amount of data associated with the analysis of large scale gene expression analysis raises the demand of tools for storing, processing and retrieving complex information. Approaches to upload and retrieve gene expression data were pursued within the GOLD.db. Large scale gene expression data sets can be uploaded in form of tab deliminated text files (Stanford file format) as used for cluster analysis programs together with additional information about the experimental conditions and the citation for already published data sets. Within those data sets the search for specific genes is possible to provide integrated visualization of gene expression levels in different studies and experimental conditions. Finally, pathways can be selected where the gene expression data can be mapped.

REAGENTS

We have developed a relational database for tracking the repository of the reagents like clone resources which can be used for microarray studies. Information about the vector, the sequence and length of the clone insert, primers for the PCR amplification, tissue, organism, accession number, library, container, storage information, date and person and access to other clone bases (e.g. IMAGE Consortium) can be stored. Users of the GOLD.db can list these clones and get all the information about each available clone. Clone information or clone lists can be uploaded and selection lists can be created and deleted by users with appropriate access. The input mask is designed in such way that the user can choose one of the elements of the created selection lists.

TOOLS

In order to deal with the huge amount of data associated with large scale studies and to perform sequence based analysis, several bioinformatics tools were integrated. Sequence similarity search against databases can be performed with BLAST (Basic Local Alignment Search Tool) (8), FASTA (9) or HMM (Hiden Markov Models) (10) on a 48-CPU PC cluster. The sequence retrieval system SRS (LION Bioscience AG, Heidelberg, Germany) was included to enable rapid, easy and user friendly access to the large volumes of diverse and heterogeneous data (11). The PathwayEditor can also be downloaded from the GOLD.db to create new pathways.

OUTREACH COMPONENTS

To establish an educational and outreach component heterogeneous sources of information have been made accessible through the GOLD.db. Video presentations of leading scientists in genomics and proteomics research can be streamed and experimental protocols can be uploaded in pdf-format. The included references are not intended to report all citations associated with a gene or its protein products. The goal is to provide a set of citations with background information. Either these citations or those included in the links, can then be used to find related publications in the PubMed. Finally, links are included to a bundle of functional genomics and computational biology resources.

IMPLEMENTATION

The GOLD.db was implemented in Java (http://java.sun.com/) technology. Hence, the pathway editor as well as the web application are platform independent. The web application of GOLD.db is build in Java Servlets and JavaServer Pages technology based on the Model-View-Controller Architecture. For the implementation, the struts framework (http://jakarta.apache.org/struts) was used. This code can be easily deployed in any Servlet Container. We used the Servlet Container Tomcat (http://jakarta.apache.org/tomcat/) which is accessible from all web browsers. Oracle 9i was used as database management system. The interface between the Java and the Database management system was established using Java database connectivity (JDBC) 2.0. Therefore, migration to other freely available DBMSs like mySQL can be easily done. For additional storage and communication between the pathway-editor components, the markup language XML containing structured, human readable information, was used.

CITING AND ACCESSING GOLD.DB

The GOLD.db database should be cited with the present publication as a reference. Access to GOLD.db is possible through the World Wide Web at http://gold.tugraz.at. The pathway editor and the clone tracker are available free of charge to academic, government, and other nonprofit institutions.

FUTURE DIRECTIONS

The vast quantity of gene expression data generated in genomic studies presents a number of challenges for their effective analysis and interpretation. In order to fully understand the changes in expression that will be observed, we must correlate these data with phenotype, genotype, and other information including the tissue distribution and time course expression data gleaned from previous studies. An important goal of our work is the development of tools that allow researchers to efficiently analyze patterns of gene expression and to display them in a variety of useful and informative ways, allowing outside researchers to perform queries pertaining to gene expression results.

We are currently developing a system for visualization of the results of microarray experiments to display relative gene expression for a given gene under specified experimental condition in combination with

other genes at the same or other experimental conditions. This approach will allow addressing further questions by analyzing of these "virtual chip experiments". Connection and integrating to a microarray database and several analysis tools like gene clustering applications (12) will raise new opportunities in understanding mechanisms of different applications and lipid-associated disorders in particular.

ACKNOWLEDGEMENTS

This work was supported by the Austrian Science Fund, Project SFB Biomembranes F718, the GEN-AU projects Bioinformatics Integration Network (BIN) and Genomics of Lipid-Associated Disorders (GOLD). Diego Miranda-Saavedra was supported by an EU Marie Curie Training Site program "Genomics of Lipid Metabolism". Michael Maurer was supported by a grant from the Austrian Academy of Sciences. We would like to thank Alexander Sturn for valuable comments and support for mapping of gene expression data and Dietmar Rieder for help with specifying of enzyme classifications.

REFERENCES

1. Flegal, K.M., Carroll, M.D., Kuczmarski, R.J., Johnson, C.L. (1998) Overweight and obesity in the United States: prevalence and trends, 1960-1994. *Int. J. Obes.*, **22**, 39-47.
2. Must, A., Spadano, J., Coakley, E.H., Field, A.E., Colditz, G., Dietz W.H. (1999) The disease burden associated with overweight and obesity. *JAMA*., **282**, 1523-1529.
3. Zechner, R., Strauss, J., Frank, S., Wagner, E., Hofmann, W., Kratky, D., Hiden, M., Levak-Frank, S. The role of lipoprotein lipase in adipose tissue development and metabolism. *Int. J. Obesity.*, **24**, S53-S56.
4. Trost, E., Hackl, H., Maurer, M., Trajanoski, Z. (2003) Java editor for biological pathways. *Bioinformatics*, **9**, 786-787.
5. Kanehisa, M., Goto, S., Kawashima, S., Nakaya, A. (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res.*, **30**, 42-46.
6. Dahlquist, K.D., Salomonis, N., Vranizan, K., Lawlor, S.C., Conklin B.R. (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat. Genet.*, **31**, 19-20.
7. Pruitt, K.D., Maglott, D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137-140.
8. Altschul, S.F., Gish, W., Miller,W., Myers, E.W., Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403-410.
9. Pearson, W.R. (1991) Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics*, **11**, 635-650.
10. Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755-763.
11. Etzold, T., Ulyanov, A., Argos, P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.*, **266**, 114-128.
12. Sturn, A., Quackenbush, J., Trajanoski Z. (2002) Genesis: Clustering gene expression data. *Bioinformatics*, **18**, 207-208.

BIOINFORMATICS APPLICATIONS NOTE
Vol. 19 no. 0 2003, pages 1–2
DOI: 10.1093/bioinformatics/btg052

Java editor for biological pathways

Elmar Trost, Hubert Hackl, Michael Maurer and Zlatko Trajanoski*

Institute of Biomedical Engineering and Christian Doppler Laboratory for Genomics and Bioinformatics, Graz University of Technology, Krenngasse 37, 8010 Graz, Austria

Received on August 30, 2002; revised on October 23, 2002; accepted on November 13, 2002

ABSTRACT

Summary: A visual Java-based tool for drawing and annotating biological pathways was developed. This tool integrates the possibilities of charting elements with different attributes (size, color, labels), drawing connections between elements in distinct characteristics (color, structure, width, arrows), as well as adding links to molecular biology databases, promoter sequences, information on the function of the genes or gene products, and references. It is easy to use and system independent. The result of the editing process is a PNG (portable network graphics) file for the images and XML (extended markup language) file for the appropriate links.

Availability: http://genome.tugraz.at

Contact: zlatko.trajanoski@tugraz.at

INTRODUCTION

The knowledge about biological pathways, their components, and the interaction between the components is crucial for understanding the function of the cell. With the advance of both, molecular biology technology and information technology, the information about molecular interactions is steadily increasing. Consequently, modeling, editing and annotating biological pathways is becoming an important issue for the organization of knowledge as well as for pathways analysis and computation. The importance of tools for editing pathways including metabolic pathways, signal transduction pathways, or gene regulatory networks was recognized earlier and a set of programs was developed for this purpose. Basically, there are three types of pathway drawing approaches: auto-layout, manual (interactive) drawing, or a hybrid of these two approaches (Kanehisa *et al.*, 2002; Koike and Rzhetsky, 2000; Karp, 2001; Karp *et al.*, 2002; Becker and Rojas, 2001). Of these, interactive drawing tools are useful for the construction of pathway diagrams in a visual way based on available knowledge, and the annotation of the components and interactions between them. However, to the best of our knowledge, there is currently no easy to use and platform independent interactive drawing tool available. Therefore, we have initiated the development of a Java tool to facilitate the representation, visualization and analysis of biological pathways.

*To whom correspondence should be addressed.

PROGRAM OVERVIEW

The pathway editor we have designed represents a novel drawing tool which integrates the possibilities of: (a) charting elements with different attributes (size, colour, labels); (b) drawing connections between elements in distinct characteristics (colour, structure, width, arrows); (c) adding text; and (d) creating a legend and adding literature (Figure 1). The form of each element—typically representing a gene as a part of a pathway—can be edited independently in the drawing plane. The great benefit of this tool is that additional information can be appended to each element via an input mask.

For each element in the pathway a specific information field exists. The field includes structured information about a gene, protein, 3D-structure, gene regulation, function, literature, and links. The GenBank (Benson *et al.*, 2002) accession number of the respective gene (typically an entry of the mRNA, including the feature CDS for the complete coding sequence) acts as the primary key for the database entries and therefore the declaration of this identity is compulsory. Besides the gene name, symbol name and GenBank accession number for the gene, protein identities for the NCBI, the SWISS-PROT (Wu *et al.*, 2002) database, and the 3D structures databases can be specified, and the accession numbers displayed and linked to the appropriate databases. The body of the query strings for these links can be changed for all entries of the pathway at once. Since in the case of transcriptional networks, the binding of transcription factors to the DNA is of interest, in the gene regulation field options were implemented to upload and display sequences upstream of the transcription start site (usually the promoter sequence) and transcription factors known to bind to these upstream activator sequences. The description, localization and

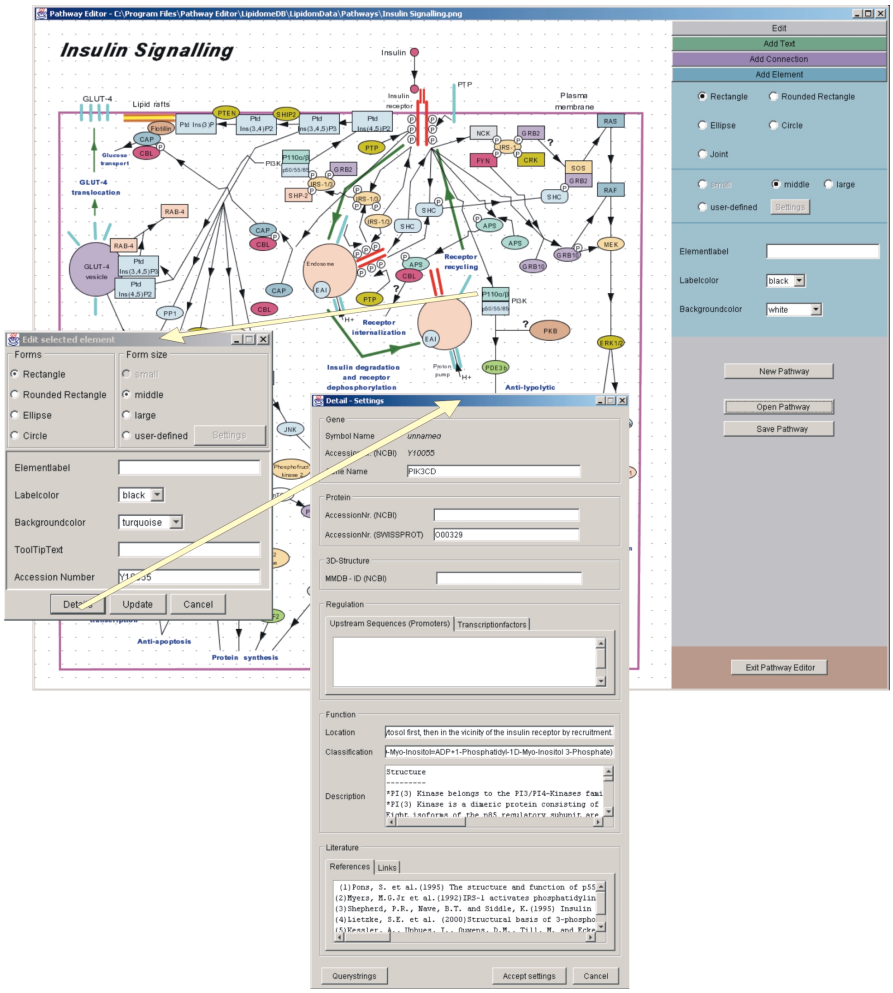

Fig. 1. An example of the use of the Pathway Editor for the construction of the insulin signaling pathway. The information that can be entered for a certain element (p110α/β, represented as a rectangle) is shown in the details-setting window and includes name, location, description and references.

classification of the factors are entered by the annotator in plain text and are accessed in the same format. The result of the editing process is a PNG (portable network graphics) file for the images and XML (extended markup language) file for the appropriate links and annotated information. Image maps can be easily created in a web page by parsing the XML files. An example of an image map constructed using this tool is the annotated pathway for insulin signaling (http://gold.tugraz.at).

The pathway editor was implemented in Java and is freely available.

ACKNOWLEDGEMENTS

This work was supported by the Austrian Science Fund, Project SFB Biomembranes F718. Michael Maurer was supported by a PhD fellowship grant from the Austrian Academy of Sciences.

REFERENCES

Becker,M.Y. and Rojas,I. (2001) A graph layout algorithm for drawing metabolic pathway. *Bionformatics*, **17**, 461–467.

Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2002) GenBank. *Nucleic Acids Res.*, **30**, 17–20.

Kanehisa,M., Goto,S., Kawashima,S. and Nakaya,A. (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res.*, **30**, 42–46.

Karp,P.D. (2001) Pathway databases: a case study in computational symbolic theories. *Science*, **293**, 2040–2044.

Karp,P.D., Paley,S. and Romero,P. (2002) The Pathway Tools software. *Bioinformatics*, **18**, S225–232.

Koike,T. and Rzhetsky,A. (2000) A graphic editor for analyzing signal-transduction pathways. *Gene*, **259**, 235–244.

Wu,C.H., Huang,H., Arminski,L., Castro-Alvear,J., Chen,Y., Hu,Z.Z., Ledley,R.S., Lewis,K.C., Mewes,H.W., Orcutt,B.C. *et al.* (2002) The Protein Information Resource: an integrated public resource of functional annotation of proteins. *Nucleic Acids Res.*, **30**, 35–37.

ArrayNorm: Comprehensive normalization and analysis of microarray data.

R. Pieler[1], F. Sanchez-Cabo[1,2], H. Hackl[1], G. G. Thallinger[1] and Z. Trajanoski[1] *

[1] Institute of Biomedical Engineering and
Christian Doppler Laboratory for Genomics and Bioinformatics,
Graz University of Technology, 8010 Graz, Austria
E-mail: zlatko.trajanoski@tugraz.at
Tel./Fax: +43 (0)316 873 5332/5340

[2] Department of Biomolecular Sciences,
UMIST, Manchester M60 1QD, U.K.

July 9, 2003

* *To whom correspondence should be addressed.*

ABSTRACT

Summary: ArrayNorm is a user-friendly, versatile and platform independent Java application for the visualization, normalization and analysis of two-color microarray data. A variety of normalization options were implemented to remove the systematic and random errors in the data, taking into account the experimental design and the particularities of every slide. In addition, ArrayNorm provides a module for statistical identification of genes with significant changes in expression.
Availability: The package is freely available for academic and non-profit institutions from
http://genome.tugraz.at
Contact: zlatko.trajanoski@tugraz.at

INTRODUCTION

Microarray technology has become an essential tool in functional genomics. However, the huge amount of data generated from a single slide requires the development of powerful computational tools to extract valuable information. While most of the current software focusses on tools to analyze the data (clustering, gene networks, detection of genes differentially expressed, gene annotation) only a few packages have been yet developed for visualization and normalization purposes. These are essential -and not trivial- tasks in order to obtain reliable and comparable data to be used in further analyzes.

The main drawbacks of the available normalization tools are the lack of universality and the poor or inexistent use of the experimental design information (reference or loop design, dye-swap, technical and biological replicates, spotted controls). Among the available packages, those using methods such as ANOVA (Engelen *et al.*, 2003) remove all sources of non-biological variation in one single step, what might appear as a "black-box" to the user. Other packages (Colantuoni *et al.*, 2002), normalize the data in a sequential way but compile just correction options for experiments for which most of the genes are expected to be equally expressed in the two hybridized samples.

Therefore, we have developed ArrayNorm, a platform independent Java suite for the visualization, normalization and analysis of two-color microarray data. The broad selection of normalization options implemented, the way of dealing with the replicated measurements and with the controls, together with the wide range of experiments that can be normalized and analyzed make ArrayNorm one of the most complete freely available tools to normalize microarray data.

PROGRAM OVERVIEW

ArrayNorm guides the user through the normalization process, accounting for the known sources of non-biological variability introduced in microarray data. Starting from the result files of the image analysis software (e.g. GenePix, Axon Instruments, Union City, CA), ArrayNorm organizes the loaded arrays in classes, i.e. the biological conditions to be compared. ArrayNorm enables the user to characterize the experiment according to several features: (1) Replicated slides within a class; (2) slides for which the dyes were swapped; (3) spotted controls; (4) replicated measurements within a slide. ArrayNorm graphically summarizes this experimental design information.

ArrayNorm allows the visualization of the data before and after normalization. To test the quality of the data, *Arrayview* reconstructs the original images representing every spot with a single pixel. The "bad spots" marked by the image analysis software are colored to show potentially contaminated areas. A *Slide Report* shows the percentage of spots excluded from the analysis and background subtraction can be performed if desired. Spatial effects due to different print tips can be detected with the *Boxplots* of the different print-tip groups. *MAplots* (Yang *et al.*, 2002) were implemented to detect intensity dependent effects in the log ratios distribution. The same information can be visualized with the *Scatterplots*. They are essential to decide which normalization method to choose. At last, *Histograms* were implemented to analyze the distribution of the intensity level of both channels and to decide if a t-test is suitable for further detection of differentially expressed genes.

Systematic errors

In (Kerr and Churchill, 2001) four main effects that introduce variability in microarray data are described: *dye, array, gene* and *sample effect.* Some account for biological variability and some others for technical. From all the sources of systematic variation that introduce noise in two-color microarray data, the most important one is the *dye effect.* The different properties of the two cyanine dyes commonly used to label the two hybridized samples make it necessary to balance the intensity level of both compared samples. These are the options implemented in ArrayNorm to remove the *dye effect* from the data:

- Using the overall distribution of the genes in the array. The implemented possibilities are the *global method*, the *linear regression method* and *LOWESS* (Cleveland, 1979) function to account for intensity dependent effects. If spatial effects were detected with the *Boxplots* or the *Arrayview*, all those options can be performed independently for every *subgrid.*

- All the methods described before assume that most of the genes in the slide are expected to be equally expressed in both compared samples. There are many experiments for which this assumption does not hold (low-density microarrays, experiments for

which no *a priori* information about the number of differentially expressed genes is available). If *controls* are available and suitable for the normalization of the data (similar expression level in both samples, covering the whole intensity range) ArrayNorm fits to them a quadratic function, using the Levenberg-Marquardt non-linear fit method (Marquardt, 1963). This function can be used to correct the whole data set. If the dyes were swapped for the replicated slides, another possibility is to use the *self-normalization* (Yang *et al.*, 2002). This method has been proved to improve the correlation among the replicates and to preserve the biological information of the data.

ArrayNorm gives the possibility to reset the data, enabling the comparison of the effect of the different normalization methods on the data.

Random errors

Microarray experiments should provide two different kind of replicates: biological and technical (Churchill, 2002). Whilst biological replicates are independent and account for the variability within the population, technical replicates are used to minimize the technical errors that can arise from the experiment but cannot be removed in a systematic way. ArrayNorm firstly detects and averages replicated spots within an array. Afterwards, the technical replicated slides can be scaled to remove the *array effect* and averaged. The result is a unique value for every spotted gene which will be the estimator of the expression level of this particular gene in a particular class. ArrayNorm allows the user to transform the data into the log scale or to continue the analysis with the intensity ratios. The normalized expression levels of all the genes in the slide can be saved in a text file in the "Stanford flat file" format which can be later used with cluster analysis applications (Sturn *et al.*, 2002). Biological replicates can be used to detect genes differentially expressed.

Differentially expressed genes

After data normalization, ArrayNorm can be employed to detect differentially expressed genes. To approach the problem, a threshold for the minimal fold change can be specified by the user. ArrayNorm also provides a t-test that can be applied to the data if they are normally distributed. The user can define the α value at which the genes can be detected as differentially expressed. To control the Error of Type I it is possible to adjust the p-values using a Bonferroni correction (Dudoit *et al.*, 2002). This t-test can be performed within a class or across different classes for experiments with a reference design. Finally, for those experiments for which the requirement of normality of the data is not met, the Mann-Whitney non-parametric test was implemented (Motulsky, 1995). The selected genes and their significance level can be saved in a text file.

OUTLOOK

ArrayNorm has been tested on Windows 2000, XP, Mac OS X and LINUX platforms. New features to normalize the spatial effects and to test the quality of the replicates are being implemented as well as more sophisticated detectors of differentially expressed genes. Finally, visualization and normalization of Affymetrix data will be included.

ACKNOWLEDGEMENTS

This work was supported by the Austrian GEN-AU project BIN (Bioinformatics Integration Network) and by the EU Marie Curie Training Site grant Genomics of Lipid Metabolism.

REFERENCES

Churchill, G. (2002) Fundamentals of experimental design for cDNA microarrays. *Nature Genetics Supplement,* **32**, 490–495.

Cleveland, W. (1979) Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association,* **74**, 829–836.

Colantuoni, C., Henry, J., Zeger, S. and Pevsner, J. (2002) SNOMAD(standarization and normalization of microarray data: web-accesible gene expression data analysis. *Bioinformatics,* **18** (1), 1540–1541.

Dudoit, S., Yang, Y., Callow, M. and Speed, T. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica,* **12**, 111–139.

Engelen, K., Coessens, B., Marchal, K. and De Moor, B. (2003) MARAN: normalizing micro-array data. *Bioinformatics,* **19** (7), 893–894.

Kerr, K. and Churchill, G. (2001) Experimental design for gene expression microarrays. *Biostatistics,* **2**, 183–201.

Marquardt, D. (1963) An algorithm for least squares-estimation of nonlinear parametrs. *Journal of the Society for Industrial and Applied Mathematics,* **11**, 431–441.

Motulsky, H. (1995) *Intuitive Biostatistics.* Oxford University Press.

Sturn, A., Quackenbush, J. and Trajanoski, Z. (2002) Genesis: cluster analysis of microarray data. *Bioinformatics,* **18** (1), 207–208.

Yang, Y., Dudoit, S., Lin, D., Peng, V., Ngai, J. and Speed, T. (2002) Normalization for cDNA microarray data: a robust composite method adressing single and multiple slide systematic variation. *Nucleic Acids Research,* **30** (4), e15.1–e15.10.