Statistical and Functional Analysis of Genomic and Proteomic Data

Yingchun Liu

Department of Theoretical Physics
Lund University, Sweden

Dissertation for the degree of Doctor of Philosophy

Advisor: Markus Ringnér

Faculty Opponent: Sayan Mukherjee
Duke University, USA

To be presented, with the permission of the Faculty of Natural Sciences of Lund University, for public criticism in Lecture Hall F of the Department of Physics on Friday, the 26th of January 2007, at 10.15 a.m.

Organization **LUND UNIVERSITY** Department of Theoretical Physics Sölvegatan 14A SE-223 62 LUND	**Document Name** **DOCTORAL DISSERTATION** **Date of issue** December 2006 **CODEN:**
Author(s) Yingchun Liu	**Sponsoring organization**

Title and subtitle

Statistical and functional analysis of genomic and proteomic data

Abstract

High-throughput technologies have led to an explosion in the availability of data at the genome scale. Such data provide important information about cellular processes and causes of human diseases, as well as for drug discovery. Deciphering the biologically relevant results from these data requires comprehensive analytical methods. In this dissertation, we present methods for gene and protein expression data analysis. Our major contributions include a method for differential in-gel electrophoresis data analysis capable of removing protein-specific dye bias in the data, a method for finding unknown biological groups using expression data, and a method for identifying active and inactive signaling pathways in a gene expression signature based on the enrichment of downstream target genes of pathways.

DOKUMENTDATABLAD enl SIS 61 41 21

Key words

2D-gel, dye bias, expression data, linear mixed model, microarray, regulatory motif, signaling pathway, TGF-beta, unsupervised classification

Classification system and/or index terms (if any)

Supplementary bibliographical information	**Language** English
ISSN and key title	**ISBN** 91-628-6997-3
Recipient's notes	**Number of pages** 80 **Price** **Security classification**

Distribution by (name and address)

Yingchun Liu, Dept. of Theoretical Physics,
Sölvegatan 14 A, SE-223 62 LUND

Signature _____ **Date** 2006-12-05

Information is only as good
as your ability to make use of it

Acknowledgments

I am grateful to a large number of people who have inspired and helped me throughout the years of my studies.

First and foremost I would like to thank my advisor Markus Ringnér for his guidance and support throughout the past four years. I have benefited greatly from his brilliance and excellent combination of knowledges. I have learned so much from him, including how to think analytically, how to do research, and how to write and present research. Except for being a great advisor, Markus also has a fantastic personality. His sense of humor and encouragement for me make it a great pleasure to work with him. Markus, it has been a wonderful and enriching experience to work closely and learn from you.

I would also like to especially thank Morten Krogh for his significant influence on my mathematical and scientific thinking. I am indebted to him for the skills I learned from him and for his insightful comments. Morten is also full of good ideas. I was often inspired by discussing with him and I really enjoyed working with him.

I am very grateful to Carsten Peterson. His concerns and support have been a big encourage to me. Carsten has great insight into problems and discusses ideas openly. I always feel enlightened after talking to him.

All my co-authors have tremendous contributions to the work in this dissertation. In particular, Stefan Karlsson and Göran Karlsson have brought me ideas in biology and have led me to the field of signaling pathways. I am thankful to them for this.

Other wonderful folks in the department include Anders Irbäck, who

always greets me with a warm smile and has been extremely patient in explaining difficult concepts with which I was unfamiliar; Michael Green, who has been a great colleague and friend. Whenever I turn to him for help, he always welcomes me with a big smile. I have learned many computer skills from him; Peter Johansson, who helped me a lot with physics and Swedish; Jari Häkkinen, who assisted in my projects; Liwen You and Carl Troein, who brought motivating discussions; Patrik Edén, who helped me with the pathway project; Henrik Jönsson, who brought helpful ideas; Mattias Ohlsson and Leif Lönnblad, who helped me with computer problems. I feel very grateful to all of you.

I would also like to thank Anders Blomberg for organizing fantastic scientific activities in the research school, and thank Olle Nerman and Ziad Taib for writing a letter of recommendation on my behalf.

Many friends have brought me a lot of joys and made my life colorful over the past years. Especially, I would like to thank Paul for always being my best friend. Aaron, who has inspired me and shown me what it means to be excellent.

Finally, I want to give my deep thanks to my parents, sister, and brother. I am grateful for their love and support for me.

This dissertation is based on the following papers:

I M. Krogh, Y. Liu, S. Bengtsson, B. Valastro and P. James
Analysis of DIGE data with a linear mixed model incorporating protein-specific dye effects
LU TP 06-40 (submitted)

II Y. Liu and M. Ringnér
Multiclass discovery in array data
BMC Bioinformatics **5**:70 (2004)

III G. Karlsson, Y. Liu, J. Larsson, M-J. Goumans, J-S. Lee, S.S. Thorgeirsson, M. Ringnér and S. Karlsson
Gene expression profiling demonstrates that TGF-β1 signals exclusively through receptor complexes involving Alk5 and identifies targets of TGF-β signaling
Physiological Genomics **21**, 396-403 (2005)

IV Y. Liu and M. Ringnér
Revealing signaling pathway deregulation by using gene expression signatures and regulatory motif analysis
LU TP 06-36 (submitted)

Contents

Biological Background

We begin with a brief overview of the concepts of molecular biology that are relevant to this dissertation followed by an introduction to technologies for generating genome-wide data. For further knowledge of molecular biology, the reader can refer to the book [1].

Basic Concepts of Molecular Biology

DNA, gene, mRNA, and protein

Cells are basic units of life. All living organisms are built from cells. At the center of the cell, there is the cell nucleus which contains the genetic code, DNA (Deoxyribonucleic acid), of the cell. The DNA molecule consists of two long sequences of nucleotides, where each nucleotide is composed of one sugar molecule, one phosphate molecule, and one of the four bases: adenine (A), cytosine (C), guanine (G), and thymine (T). These bases can form complementary base pairs in the form of A-T and C-G, joined by hydrogen bonds. The two sequences of the DNA are joined by such base pairs and twisted into a double helix (Figure 1). DNA is organized into separate chromosomes in the cell nucleus, and the whole genetic information encoded in the DNA for an organism is termed genome.

A *gene* is a region on the DNA sequence that codes for proteins. In the human genome, there are over 30,000 genes, and there are even more proteins than genes, because each gene can code for multiple proteins. Genes encode proteins through two main steps. Firstly, the DNA sequence of a gene is transcribed into another molecule called mRNA

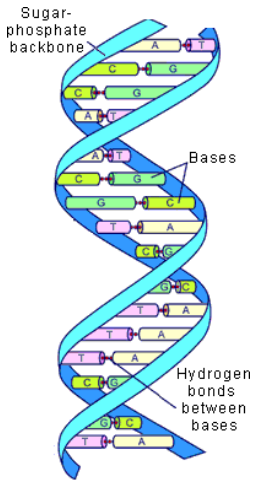


Figure 1: An illustration of the DNA molecule. Image taken from http://web.jjay.cuny.edu/~acarpi/NSC/12-dna.htm

(Ribonucleic acid) by base pairing, thereby the mRNA contains a nucleotide sequence complementary to its template DNA sequence. This process is termed *transcription*. When a gene is transcribed into mRNA, the gene is said to be expressed. The *expression level* of this gene refers to its mRNA abundance. Secondly, the mRNA leaves the cell nucleus and travels to the cellular processing units called ribosomes, where it serves as template for protein synthesis. Every three consecutive nucleotides of the mRNA sequence are converted into one amino acid, and the amino acids are linked together by peptide bonds into a poly-peptide chain. This process is termed *translation*. Finally, the chain folds into a protein with specific three dimensional structure (see Figure 2).

Gene expression regulation

The are often many types of cells in an organism. In the human body, there are brain cells, lung cells, liver cells, skin cells, and so on. Although all the cells contain the same DNA, they appear different and have different functions. This is because different genes are expressed in

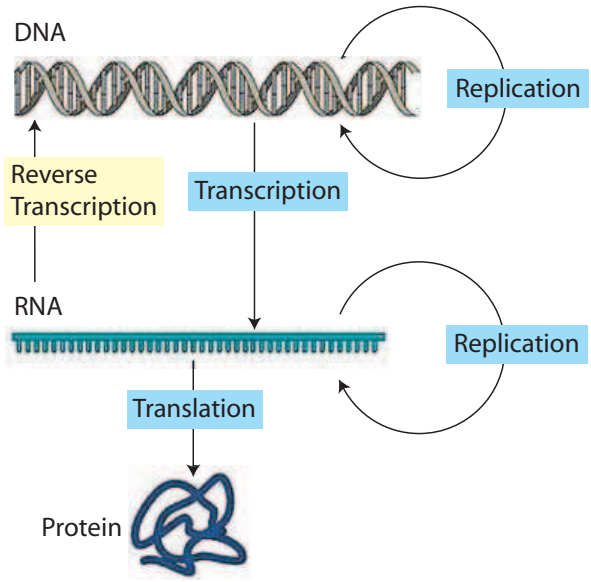


Figure 2: The central dogma of molecular biology. Genetic information encoded in the DNA is passed to RNA through transcription and then to proteins through translation.

different types of cells, which leads to production of different proteins.

Gene expression is largely controlled by regulatory proteins called *transcription factors*. To control gene expression, transcription factors bind to specific short sequences on the DNA in the region upstream of the transcription start site of the gene. This binding recruits or impedes proteins necessary for transcription to enhance (*up-regulate*) or inhibit (*down-regulate*) the transcription of this gene. Such an upstream region of a gene is called a *promoter* region.

In the cell, one gene can be regulated by many transcription factors. One transcription factor can bind to many short sequences in the promoter regions of different genes to regulate their transcription. This forms a complex regulatory network. Similar binding sequences of a transcription factor are represented by a common pattern called the *motif* of the transcription factor. One transcription factor could have multiple motifs.

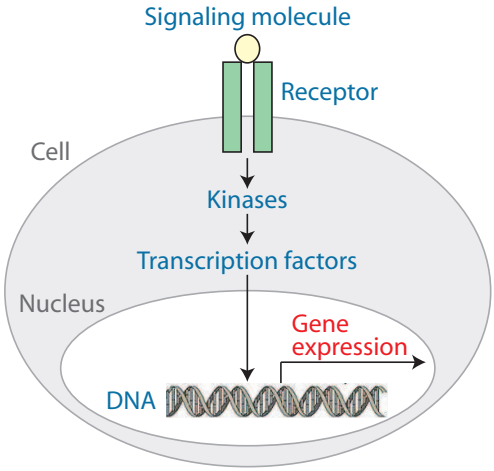


Figure 3: A schematic representation of the signal transduction within a eukaryotic cell. Once the signaling molecule binds to the receptor, the signal is passed through a number of steps into the cell nucleus, where gene expression is affected.

Signaling and metabolic pathways

Cells have the ability to communicate with their internal and external environment, and adjust their functions in response to environmental changes. This ability is achieved through a number of signaling pathways that receive and process signals.

A signaling pathway consists of a set of molecules, such as ligands, receptors, enzymes, and transcription factors. Ligands are signaling molecules from the external or internal environment, and receptors are proteins located on the cell membrane or within the cell that bind to signaling molecules. Once ligands bind to receptors, the signal is propagated within the cell through a cascade of biochemical interactions between receptors and transcription factors (Figure 3). As a result, various transcription factors are activated or inactivated, which in turn alters the expression levels of many genes and eventually alters the activities of biological processes.

Signal transduction is at the core of many biological processes. For example, cell growth, proliferation, differentiation, and apoptosis are

all controlled by signaling pathways. The cell can respond to multiple signals at a time, and the cell's response to one specific signal could activate multiple signaling pathways. Various signaling pathways within the cell form a signaling network.

In addition to signaling pathways, metabolic pathways are also important for the cell's survival. A metabolic pathway is a series of enzyme-catalyzed biochemical reactions that produce energy storage molecules and biomolecules for the cell. In a metabolic pathway, the product of one reaction is the substrate of the next reaction. Metabolic pathways can share common substrates and enzymes forming a metabolic network.

Gene and Protein Expression Profiling

Most processes within the cell are carried out by proteins and their interactions with other molecules. For example, proteins function as enzymes that catalyze biochemical reactions, receptors that receive and propagate signals, and transcription factors that regulate gene expression. Each process is governed by a specific set of active proteins, and the activities of proteins thereby provide important information about the ongoing processes in the cell. In many cases, protein activities are correlated with their abundances. Since proteins are produced from the mRNAs of genes, protein abundances are often correlated with the mRNA abundances of the genes encoding the proteins. Therefore, gene expression studies have the potential to reveal the active processes in the cell.

Cells affected by diseases often have a set of genes differentially expressed with respect to normal cells, because different cellular processes typically are activated as a response to genetic or cellular changes. Such differentially expressed genes may provide insight into the causes of the diseases or be potential drug targets. However, there are hundreds to thousands of genes in an living organism. It is impossible to know which genes to be examined without detailed prior knowledge. This had been a bottleneck for biomedical research using traditional biotechnologies, which can only measure the expression levels of one or few genes at a time. Fortunately, the invention of DNA microarray technology about a decade ago has made genome-wide gene expression studies possible [2].

Microarray-based gene expression profiling

DNA microarrays can measure expression levels of thousands of genes simultaneously on a single slide. Each slide contains thousands of spatially separated spots on the surface. And each spot contains multiple copies of a short DNA sequence that represents one gene. To measure gene expression levels, the mRNA contents of cells are extracted from samples and reversely transcribed into complimentary DNAs (cDNA). The cDNAs are labeled with a fluorescent dye that absorbs and emits light at specific wavelengths. Then, the cDNAs are hybridized on the array where they bind to their complementary DNA sequences in the spots. Finally, the array is scanned to obtain the emitted fluorescent intensities for all spots. The intensity of each spot indicates the expression level of the gene it represents.

There are many types of DNA microarrays, which differ in array fabrication or choice of dyes [3]. In a two-color DNA microarray [4], which is relevant for this dissertation, the mRNA abundances of genes in two samples are compared directly on the same array by labeling them with different fluorescent dyes. After hybridization, the array is scanned at two different wavelengths, which generates two intensities for each spot corresponding to the expression level of this gene in the two samples. I will refer to the two intensities as red and green in the later context. A schematic overview of cDNA microarray technology is shown in Figure 4.

The mRNA abundance is, however, not always correlated with protein abundance. For example, the rates of mRNA decay, translation, and protein decay can influence this correlation. In addition, protein activity is not always correlated with its abundance either. Proteins could be activated or inactivated by post-translational modifications. Hence, large-scale data at protein level, termed *proteomic data*, are useful. Technologies like protein microarray [6], two dimensional poly-acrylamide gel electrophoresis (2D-PAGE) [7,8], and mass spectrometry [9] are used to generate data for proteomic research. These technologies are still immature and under development though, due to complex features of proteins. So far, the most widely applied approach for proteomic research has been protein expression analysis.

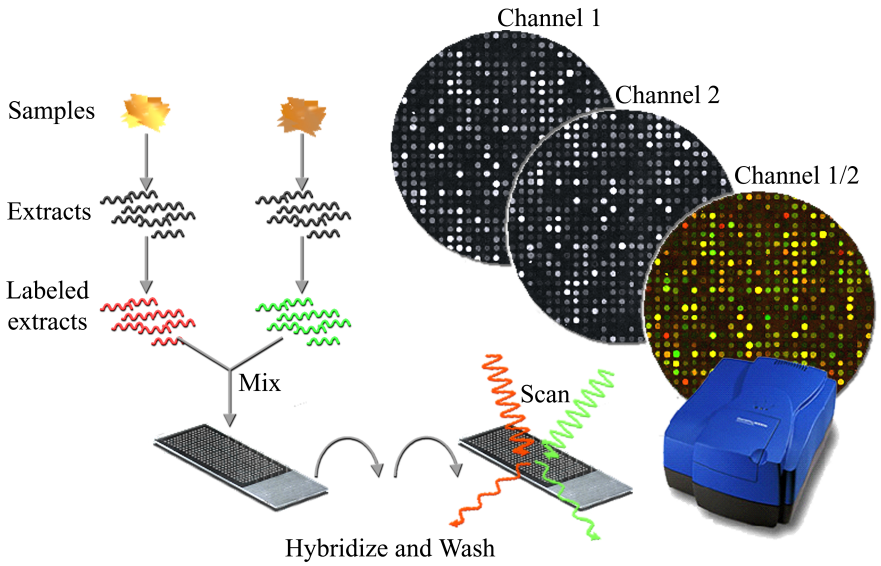


Figure 4: A schematic overview of the two-color DNA microarray technology. Two samples are labeled with a red and a green fluorescent dye respectively and are hybridized on the same array. After hybridization, the array is scanned to obtain two images. The two images are often merged into one image where spots are colored on a scale from red to yellow to green corresponding to the relative gene expression in the two samples. Reprinted with permission from Johan Vallon-Christersson [5].

2D gel-based protein expression profiling

One common way of measuring the abundances of thousands of proteins simultaneously is by means of 2D-PAGE. In 2D gels, proteins extracted from a sample are first separated by isoelectric point using an immobilized pH gradient. Next, proteins are separated by molecular weight, because proteins with different weights move at different speed on the gel. The resulting gel is then stained to visualize the protein spots. Finally, the gel is scanned to obtain the intensities of all spots.

Traditional 2D-PAGE can only deal with one sample on each gel. More recently, differential in-gel electrophoresis (DIGE) [10] was introduced,

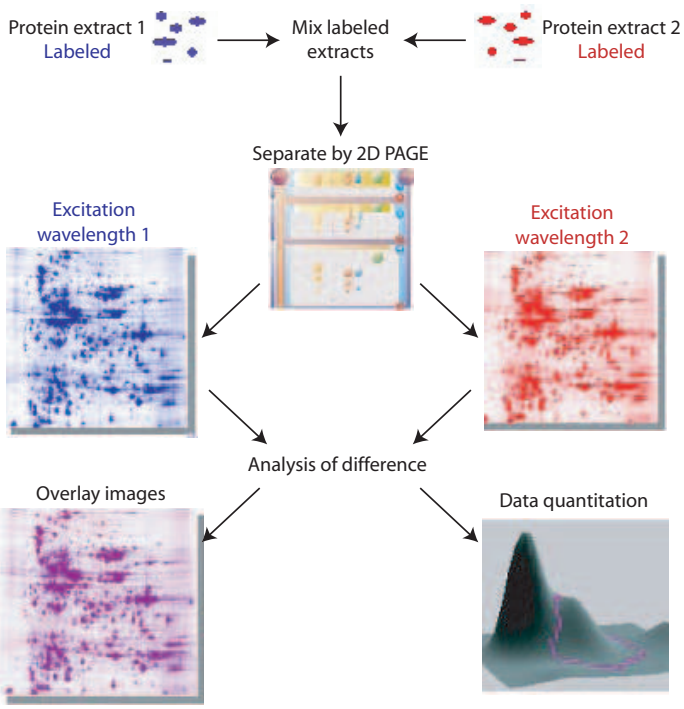


Figure 5: Outline of a DIGE experiment to compare protein abundances in two samples. Protein extractions of two samples are labeled with two different dyes and are resolved on the same gel. Finally, two scanned images are obtained, and the intensities of the spots indicate the abundances of proteins.

which can measure protein abundances of up to three samples on the same gel. In DIGE gels, protein extractions from samples are labeled with different fluorescent dyes and are mixed prior to 2D gel electrophoresis. Finally, the abundances of proteins in these samples are determined by the scanned intensities for the spots. An overview of the DIGE technology is shown in Figure 5.

Genomic and Proteomic Data Analysis

Recent advances in technologies have made a vast amount of data at the genome scale available, including complete genome sequences, genome-wide protein-DNA binding sites, and genome-wide gene and protein expression profiles under various conditions. Such data provide important genetic and cellular information. However, transforming these immense amounts of data into biological information is challenging, especially when there are measurement uncertainties in the data. A successful transformation relies on theoretically-founded methods with deep understanding of the biology. In this chapter, I will discuss some of the challenges: normalization of expression data, identification of differential expression, unsupervised classification, and investigation of functional networks. These are addressed in the appended papers.

Normalization of Expression Data

Expression data obtained by using microarrays are noisy. Differences in the RNA abundances between samples are often mixed with nonbiological variations. Dye bias is the most common nonbiological variation, caused by different labeling or scanning properties of dyes. In particular, the RNA may bind to one dye better than the other, or the same RNA sample labeled with different dyes could have different measured intensities. Except for the dye bias, differences between arrays would exist when the hybridization efficiency on each array is different. Differences between spots would exist when there is different amount of cDNAs printed on each array for the same gene. And different print-tips for different locations on the array would introduce variation as

well. Before applying microarray data for biological studies, the data must be normalized to remove nonbiological variations arising from the technology.

There are many statistical approaches to normalizing two-color DNA microarray data. For instance, normalization can be done separately for each array, using only the red (R) and green (G) intensities for this array, or it can be done using multiple arrays. Overall, many of these approaches aim to have all normalized $\log_2(R/G)$ ratios on each array centered around zero. The underlying assumption is that the numbers of up- or down-regulated genes in each sample are roughly the same, when a random set of genes are printed on each array. Consequently, the mean $\log_2(R/G)$ for each array should be close to zero.

Global normalization is the simplest and most widely used approach, where all the green intensities are multiplied with a constant factor such that the red and green intensities have equal mean or median. This can be done using all the genes on each array, or a selected set of genes, e.g. housekeeping genes [11] or externally spiked genes [12] whose expression levels are constant across multiple conditions. However, dye effects are often dependent on spot intensity and location on the array, so intensity dependent and print-tip based local normalization methods seem more appropriate in this regard [13, 14].

Each of the approaches above is likely to remove only certain nonbiological variations in the data. A more general approach for normalizing DNA microarray data is to use the analysis of variance (ANOVA) models, including fixed-effects ANOVA and mixed-model ANOVA [15–17]. The ANOVA models can be designed to account for variations arising from many sources including arrays, dyes, spots, and their confounding effects, by considering each of them as an unknown parameter of the model. The normalized expression levels of each gene can be obtained by fitting the model using data from all arrays.

In paper I, we introduced a linear mixed model that is able to correct for protein-specific dye effects in DIGE data. DIGE data for protein expression studies have similar properties as microarray data and must be normalized as well.

Identification of Differential Expression

A common task of microarray and DIGE data analysis is to find the genes or proteins differentially expressed between biological groups, for example, disease versus healthy, different cell types, and different conditions. Genes or proteins with expression patterns associated with these groups could provide insight into the causes of diseases, be molecular markers differentiating between cell types, and reflect the active cellular processes under different conditions.

The simplest way of finding differentially expressed genes is the fold change, which considers all genes, whose log ratios between two groups are larger than an arbitrary threshold as differentially expressed [18,19]. The statistical significance of a gene being differentially expressed is, however, unknown for the fold change. More commonly, differentially expressed genes are identified by performing a statistical test gene-by-gene.

For comparison between two groups, the most widely used tests are t-test [20, 21], modified t-tests (significance analysis of microarrays [22]; the regularized t-test [23]), paired t-test [24], Pearson correlation [25], Wilcoxon rank-sum test [26], and permutation test. Each of these tests emphasizes different aspects of the data. The difference between the t-test and its modifications lies in the calculation of the variance of each gene, so that a gene with too small fold change but small variance by chance will not be selected, or vice versa. In the paired t-test, array and spot effects are taken into account, where, for each gene printed on the array, the expression levels from two groups are compared directly and used as a pair in the t-test. The Pearson correlation measures the association between the expression of a gene with the group label, rather than tests the difference in the mean expression of two groups. Compared with all types of t-tests and Pearson correlation, the Wilcoxon rank-sum test and permutation tests do not require normal distribution of data. As microarray data are noisy and often do not form a normal distribution, the nonparametric tests are appealing in this context. The permutation test is also very flexible. It can be used to test the significance of a score constructed in any way, including the scores used in the different tests.

To identify genes differentially expressed between multiple groups, the ANOVA F test [27] and Kruskal-Wallis test [28] are widely used. In addition, the likelihood ratio test and Wald test are also common, when using models for expression analysis. These two tests can be applied to comparison between any number of biological groups.

Each of these tests has its merits. None of them is superior to others for all microarray data sets. The particular test used depends on the data set under study. Statistical methods for identifying differentially expressed proteins are similar.

In paper I, we applied the likelihood ratio test and the Wald test to identify differentially expressed proteins. We employed the Wilcoxon rank-sum test and the Kruskal-Wallis test, in paper II, to find differentially expressed genes.

Unsupervised Classification

Unsupervised classification refers to revealing unknown biological classes in a collection of samples. In medicine, an important usage of unsupervised classification is to find new subtypes of cancers. Cancer is a complex genetic disease having many subtypes. Classification of cancer is primarily based on the histopathological appearance of the tumor. However, tumors with similar histologic appearance could have developed from different genetic aberrations and have different responses to therapy. For example, different subtypes of breast tumors have different responses to chemotherapy. Cancer treatment based on conventional diagnosis is thus difficult. A major challenge of cancer treatment has been to find specific therapies to pathogenetically distinct tumor types.

Interestingly, recent studies have found that some morphologically similar tumors can be molecularly divided into subclasses with distinct pathogenese. For example, microarray-based gene expression studies have identified subtypes of cutaneous melanoma [29] and four subtypes of breast tumors [30]. These and similar findings have triggered the enthusiasm for unsupervised classification of samples using gene expression data. Many methods have been developed for this purpose.

The most widely used method for unsupervised classification is perhaps agglomerative hierarchical clustering [31, 32]. This method begins by considering each sample as a cluster, and then merges the two closest clusters based on a similarity metric. This process of merging is repeated until there is a single cluster left. Finally, the samples are organized into a tree structure. Classes can be obtained by cutting the tree at a particular height. The k-means method is another commonly used method [33]. Starting from k randomly or carefully chosen data points, called 'centroids', the k-means method iteratively assigns samples to the nearest centroid's cluster and adjusts the centroids to represent the center of the new clusters, optimizing some objective function. Eventually, samples are divided into k clusters. In addition to these two methods, there are a few model-based methods that assume data are sampled from a model distribution, for example a mixture of Gaussian distributions, and seek for parameters that best fit the data [34].

In paper II, we took a strategy different from those described above by using information about differentially expressed genes. Samples from known different classes usually have an overabundance of genes differentially expressed, compared with random classes. These differentially expressed genes are often up-regulated in one class and down-regulated in the other, which forms nice expression patterns characterizing the distinction between compared biological groups. These expression patterns are intrinsic features of the data. Even if we did not know the class labels of the samples, there would still exist such expression patterns. Therefore, in unsupervised classification, we are likely to find the biologically relevant classes in the data, if we could find a partition that exhibits such expression patterns. We applied simulated annealing to find the best partitions.

Investigation of Functional Networks

Finding genes or proteins differentially expressed between biological groups and identifying unknown groups using expression profiles are approaches capable of characterizing biological groups and diagnosing diseases. Furthermore, mapping the differentially expressed genes and proteins to

biological categories, including chromosome location and biological processes, help account for the observed differences between biological groups.

It is important to realize that, in living cells, many genes and proteins function coordinately for complex functions. It is crucial to understand the relationships between them. This understanding has significant impact on drug discovery, because complex diseases often depend on altered interactions between a few genes, rather than changes in a single gene. For example, p53 is a tumor suppressor responsible for DNA repair. It inhibits cell growth in response to DNA damage. But p53 function is controlled by the Mdm2 protein interacting with it. Mdm2 enhances degradation of p53 [35]. If a person has lung cancer, and also has an abnormal overabundance of MDM2 protein in his lung cells, this person can not be cured by simply increasing p53 transcription.

Ideally, we would like to learn the relationships of all genes and proteins in an organism, but biological complexity increases exponentially with the number of genes and the interactions between them. Such large-scale studies are only feasible for simple organisms. In yeast, studies have shown that it is possible to infer molecular pathways [36] and regulatory networks [24,37,38] from gene expression data, using probabilistic models, bayesian or boolean networks. For higher organisms, efforts have been focusing on learning the structure and dynamics of small systems, such as the cell cycle [39] and specific signaling pathways [40]. Such small systems can be studied by perturbation (for example, knocking out interesting genes or overexpressing specific proteins) followed by monitoring the response of each element over time. Finally, the relationships of elements of the system and its response to individual perturbations can be described by mathematical models.

Alternatively, unlike the studies above attempting to learn the detailed relationships, some other studies aim to uncover the coordinate behavior of many genes and proteins in terms of known systems, including metabolic pathways [41, 42] and signaling pathways [43]. Since each pathway usually involves many genes and proteins, and different pathways often share common genes, the set of all known pathways form a complex network of genes and proteins. Active and inactive pathways provide insight into the regularities in the observed gene expression pro-

files with respect to the topology of this gene network.

In addition, gene expression data alone might not be enough for learning functional relationships. More recently, two studies have tried to identify regulatory programs in large-scale transcriptional signatures in cancer, by integrating microarray data with regulatory motifs [44] and DNA copy number [45].

In paper IV, we presented a strategy to study the regulatory mechanisms responsible for the observed gene expression profiles in the context of signaling pathways. We integrated pathway information with regulatory motif data for pathway analysis.

Summary of the Papers

Paper I

In Paper I, we study the dye effects in protein expression data generated by DIGE experiments, where abundances of thousands of proteins in three samples are measured simultaneously on one gel. Each of the samples is labeled with a distinct fluorescent dye. Prior to comparison of protein abundances, differences between dye intensities must be removed. This is usually done by a global normalization within each dye channel. However, we find that dye effects are in fact protein-specific and cannot be removed by any global normalization methods. To address this problem, we introduce an algorithm, a linear mixed model, which incorporates protein-specific dye effects and is applicable to most experimental designs. The algorithm is implemented in a JAVA program called DIGEanalyzer that automatically corrects for protein-specific dye effects and identifies differentially expressed proteins between any linear combination of groups. DIGEanalyzer is available at http://bioinfo.thep.lu.se/digeanalyzer.html.

Conclusion: Dye effects in DIGE data are protein-specific which cannot be corrected for by global normalization methods. We present a program that corrects for protein-specific dye effects and identifies differentially expressed proteins.

Paper II

In Paper II, we present a method to find biologically relevant groups in a set of samples using microarray data. Unlike many methods that cluster experiments based on their distances in gene expression space, we look for partitions of samples that have an overabundance of differentially expressed genes, starting from a predefined number of groups and randomly labeled samples. We evaluate this method using two published microarray data sets: small round blue cell tumors (SRBCT) and breast tumors. The SRBCT data set contains samples belonging to four different SRBCT types and the breast tumors data set contains non-BRCA1/2 familial breast tumors. When applying the method on these data sets, interestingly, we find that the SRBCT data set can be separated perfectly into two groups: tumors and cell lines or into three groups reflecting print batches of microarrays. Our method is able to detect such groups and remove genes discriminating them from analysis, which enables us to find the biologically relevant groups in these data with high success rates. This method is available as a PERL program at http://bioinfo.thep.lu.se/classdiscoverer.

Conclusion: Unknown biological groups in a set of samples could be identified by looking for partitions of samples with an overabundance of differentially expressed genes.

Paper III

In paper III, we study the TGF-β signaling system. Transforming growth factor-β1 (TGF-β) regulates cellular functions, such as proliferation, differentiation, and apoptosis through the TGF-β signaling pathway. It is well-known that the TGF-β signal is transduced through receptor complexes composed of TGF-β receptor type II (TβRII) and activin-like kinase receptor-5 (Alk5) on the cell surface. In this study, we screen for alternative receptors for TGF-β in murine embryonic fibroblast (MEF) cells using gene expression profiling and functional assays. We also identify gene targets of TGF-β signaling in MEF cells.

Conclusion: TGF-β signals exclusively through receptor complexes involving Alk5 in MEF cells.

Paper IV

In Paper IV, we present a method to study the regulatory mechanisms underlying diseases and other biological observations in terms of signaling pathways. We look for active and inactive signaling pathways in the gene expression signatures characteristic of these observations. The method takes a gene signature as input and outputs the signaling pathways whose activation or inactivation might have resulted in the observed expression patterns of these genes. In the analysis, all pathways in the TRANSPATH database are extracted and each is characterized by a set of transcription factors mediating it. The activity of each pathway in the gene signature is inferred based on the enrichment of the downstream target genes of the pathway. Since there are few known target genes, we search for putative target genes by looking for the binding motifs of the transcription factors in the promoter regions of genes. This method is different from many methods in two aspects: First, the activities of pathways are determined by the enrichment of target genes of pathways rather than that of molecular components of pathways. Second, putative target genes that contain the motifs of the transcription factors are used, instead of the few known target genes. We evaluate this method using six human and mouse gene expression signatures.

Conclusion: Regulatory motif analysis of gene expression signatures reveals signaling pathway activation or inactivation.

References

[1] Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD: *Molecular biology of the cell*. Garland Publishing 1994.

[2] Schena M, Shalon D, Davis RW, Brown P: **Quantitative monitoring of gene expression patterns with complementary DNA microarray**. *Science* 1995, **270**(5235):467–470.

[3] Heller MJ: **DNA microarray technology: devices, systems, and applications**. *Annu Rev Biomed Eng* 2002, **4**:129–153.

[4] Xiang CC, Chen Y: **cDNA microarray technology and its applications**. *Biotechnol Adv* 2000, **18**:35–46.

[5] Vallon-Christersson J: **Functional and molecular characterization of BRCA1 and BRCA2 associated breast cancer**. *PhD thesis*, Lund University 2005.

[6] Templin MF, Stoll D, Schwenk JM, Potz O, Kramer S, Joos TO: **Protein microarrays: promising tools for proteomic research**. *Proteomics* 2003, **3**(11):2155–2166.

[7] O'Farrell PH: **High resolution two-dimensional electrophoresis of proteins**. *J Biol Chem* 1975, **250**(10):4007–4021.

[8] Scheele GA: **Two-dimensional gel analysis of soluble proteins. Charaterization of guinea pig exocrine pancreatic proteins**. *J Biol Chem* 1975, **250**(14):5375–5385.

[9] Aebersold R, Mann M: **Mass spectrometry-based proteomics**. *Nature* 2003, **422**(6928):198–207.

[10] Alban A, David SO, Bjorkesten L, Andersson C, Sloge E, Lewis S, Currie I: **A novel experimental design for comparative two-dimensional gel analysis: two-dimensional difference gel electrophoresis incorporating a pooled internal standard**. *Proteomics* 2003, **3**:36–44.

[11] Lee PD, Sladek R, Greenwood CMT, Hudson TJ: **Control genes and variability: absence of ubiquitous reference transcripts in diverse mammalian expression studies**. *Genome Res* 2002, **12**(2):292–297.

[12] van de Peppel J, Kemmeren P, van Bakel H, Radonjic M, van Leenen D, Holstege FCP: **Monitoring global messenger RNA changes in externally controlled microarray experiments**. *EMBO Rep* 2003, **4**(4):387–393.

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

[14] Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: **Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation**. *Nucleic Acids Res* 2002, **30**(4):e15.

[15] Kerr MK: **Analysis of variance for gene expression microarray data**. *J Comput Biol* 2000, **7**(6):819–837.

[16] Lee MLT, Lu W, Whitmore GA, Beier D: **Models for microarray gene expression data**. *J Biopharm Stat* 2002, **12**:1–19.

[17] Carter MG, Hamatani T, Sharov AA, Carmack CE, Qian Y, Aiba K, Ko NT, Dudekula DB, Brzoska PM, Hwang SS, Ko MSH: **In situ-synthesized novel microarray optimized for mouse stem cell and early developmental expression profiling**. *Genome Res* 2003, **13**(5):1011–1021.

[18] DeRisi JL, Iyer VR, Brown PO: **Exploring the metabolic and genetic control of gene expression on a genomic scale**. *Science* 1997, **278**(5338):680–686.

[19] Draghici S: **Statistical intelligence: effective analysis of high-density microarray data**. *Drug Discov Today* 2002, **7**(11 Suppl):55–63.

[20] Salmon KA, Hung Sp, Steffen NR, Krupp R, Baldi P, Hatfield GW, Gunsalus RP: **Global gene expression profiling in Escherichia**

coli K12: effects of oxygen availability and ArcA. *J Biol Chem* 2005, **280**(15):15084–15096.

[21] Callow MJ, Dudoit S, Gong EL, Speed TP, Rubin EM: **Microarray expression profiling identifies genes with altered expression in HDL-deficient mice**. *Genome Res* 2000, **10**(12):2022–2029.

[22] Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response**. *Proc Natl Acad Sci U S A* 2001, **98**(9):5116–5121.

[23] Baldi P, Long AD: **A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes**. *Bioinformatics* 2001, **17**(6):509–519.

[24] Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data**. *Nat Genet* 2003, **34**(2):166–176.

[25] Mansson R, Tsapogas P, Akerlund M, Lagergren A, Gisler R, Sigvardsson M: **Pearson correlation analysis of microarray data allows for the identification of genetic targets for early B-cell factor**. *J Biol Chem* 2004, **279**(17):17905–17913.

[26] Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB: **Nonparametric methods for identifying differentially expressed genes in microarray data**. *Bioinformatics* 2002, **18**(11):1454–1461.

[27] Li H, Wood CL, Liu Y, Getchell TV, Getchell ML, Stromberg AJ: **Identification of gene expression patterns using planned linear contrasts**. *BMC Bioinformatics* 2006, **7**:245.

[28] Croonquist PA, Linden MA, Zhao F, Van Ness BG: **Gene profiling of a myeloma cell line reveals similarities and unique signatures among IL-6 response, N-ras-activating mutations, and coculture with bone marrow stromal cells**. *Blood* 2003, **102**(7):2581–2592.

[29] 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**(6795):536–540.

[30] Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D: **Molecular portraits of human breast tumours**. *Nature* 2000, **406**(6797):747–752.

[31] Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns**. *Proc Natl Acad Sci U S A* 1998, **95**(25):14863–14868.

[32] Khan J, Simon R, Bittner M, Chen Y, Leighton SB, Pohida T, Smith PD, Jiang Y, Gooden GC, Trent JM, Meltzer PS: **Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays**. *Cancer Res* 1998, **58**(22):5009–5013.

[33] Ben-Dor A, Shamir R, Yakhini Z: **Clustering gene expression patterns**. *J Comput Biol* 1999, **6**(3-4):281–297.

[34] Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL: **Model-based clustering and data transformations for gene expression data**. *Bioinformatics* 2001, **17**(10):977–987.

[35] Piette J, Neel H, Marechal V: **Mdm2: keeping p53 under control**. *Oncogene* 1997, **15**(9):1001–1010.

[36] Segal E, Koller D: **Discovering molecular pathways from protein interaction and gene expression data**. *Bioinformatics* 2003, **19**:i264–i272.

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

[38] Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N: **Revealing modular organization in the yeast transcriptional network**. *Nat Genet* 2002, **31**(4):370–377.

[39] Qu Z, Weiss JN, MacLellan WR: **Regulation of the mammalian cell cycle: a model of the G1-to-S transition**. *Am J Physiol Cell Physiol* 2003, **284**(2):349–364.

[40] Wiley HS, Shvartsman SY, Lauffenburger DA: **Computational modeling of the EGF-receptor system: a paradigm for systems biology**. *Trends Cell Biol* 2003, **13**:43–50.

[41] Vert JP, Kanehisa M: **Extracting active pathways from gene expression data**. *Bioinformatics* 2003, **19 Suppl 2**:II238–II244.

[42] Rahnenfuhrer J, Domingues FS, Maydt J, Lengauer T: **Calculating the statistical significance of changes in pathway activity from gene expression data**. *Stat Appl Genet Mol Biol* 2004, **3**:Article16.

[43] Breslin T, Krogh M, Peterson C, Troein C: **Signal transduction pathway profiling of individual tumor samples**. *BMC Bioinformatics* 2005, **6**:163.

[44] Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Barrette TR, Ghosh D, Chinnaiyan AM: **Mining for regulatory programs in the cancer transcriptome**. *Nat Genet* 2005, **37**(6):579–583.

[45] Adler AS, Lin M, Horlings H, Nuyten DSA, van de Vijver MJ, Chang HY: **Genetic regulators of large-scale transcriptional signatures in cancer**. *Nat Genet* 2006, **38**(4):421–430.

Paper I

Analysis of DIGE data using a linear mixed model allowing for protein-specific dye effects

Morten Krogh[1,4], Yingchun Liu[1], Sofia Bengtsson[4], Barbara Valastro[2,3] and Peter James[4]

[1]Computational Biology and Biological Physics, Department of Theoretical Physics, Lund University, Lund, Sweden
[2]Basal Ganglia Pathophysiology Unit, Department of Experimental Medical Science, Lund University, Lund, Sweden
[3]In Vivo Pharmacology, Merz Pharmaceuticals GmbH, Frankfurt am Main, Germany
[4]Department of Protein Technology, Biomedical Centre, Lund University, Sweden

Email: Morten Krogh*- mkrogh@thep.lu.se;

*Corresponding author

Abstract

Differential in-gel electrophoresis (DIGE) experiments allow three protein samples to be run per gel. The three samples are labeled with the spectrally resolvable fluorescent dyes, Cy2, Cy3, and Cy5, respectively. Usually, protein abundances across dyes are compared directly after a global normalization within each dye channel. Here, we show that protein-specific dye effects exist causing over- or underrepresentation of a protein according to the dye used for labeling. This cannot be corrected for by a global normalization, and an effective data analysis should take the dye effects into account. We present a linear mixed model for analysis of DIGE data which takes dye effects into account, and have implemented the algorithm in a freely available Java program, called DIGEanalyzer. Three DIGE experiments from our laboratory, with 173, 64, and 24 gels, respectively, were used to quantify and verify the dye effects. The fraction of proteins with a statistically significant (0.001 level) dye effect were 19%, 34%, and 23%, respectively. The median magnitude of the dye effect is 1.07 fold change for Cy5 versus Cy3 and 1.16 fold change for Cy3 versus Cy2. Importantly, a number of proteins had very large dye effects showing up to a 6 fold change according to which dye was used for labeling.

Introduction

Proteomics is the field of quantifying and identifying numerous proteins simultaneously [1–4]. Knowledge about proteins is crucial for both basic cellular biology and for medicine. A typical task is to find the proteins that are up- or downregulated between two or more biological groups, where each group is represented by one or more biological and technical replicates. The most common ways of measuring relative

protein abundances are either by means of 2 dimensional poly-acrylamide gel electrophoresis (2D-PAGE) [5–7] followed by mass spectrometry for identification of the proteins, or by means of multi-dimensional chromatography coupled to mass spectrometry using differential labeling of the samples, e.g. ICAT [8], cICAT [9] and iTRAQ [10]. The traditional 2D gel experiments, which are still widely used, employ one sample per gel, and require comparisons between spot volumes measured on different gels. More recently, the differential in-gel electrophoresis (DIGE) technique was introduced [11]. Here three samples, labeled with the fluorescent dyes Cy2, Cy3, and Cy5, respectively, are resolved on the same gel. Samples could then either be directly compared on the same gel, or a common reference sample, typically a pool of all biological samples, could be run on all gels [12]. The standardized abundance of a spot from a biological sample is defined as the ratio between the volume of that spot and the volume of the corresponding spot from the reference sample. It has been shown that, in a comparison between biological samples from different gels, standardized abundances give more precise protein quantification than raw volumes [12]. The common reference sample also serves the purpose of facilitating the spot matching between gels by having the same sample on all gels. Since its inception, the DIGE method has been used to quantify protein changes between distinct biological groups including cancer cells [13–15]. Several studies on statistical aspects of the DIGE technique have been published as well [16–19]. So far, analyses of DIGE experiments have assumed that there is no protein-specific dye effects, i.e., that after a global normalization for each dye, protein abundances can be directly compared across dyes. Karp and Lilley [16] noted that a certain spot was much brighter in the Cy5 image than in the Cy3 image. However, to our knowledge, no systematic analysis of dye effects in DIGE has been performed, and no algorithms incorporating protein-specific dye effects have been developed.

In this study, we first show that differential dye effects are indeed prevalent when using Cy dyes. We show that certain proteins incorporate Cy3 much more efficiently than Cy5 whereas other proteins behave in the opposite fashion. A similar effect is seen between Cy2 and Cy3. Then we introduce an algorithm - a linear mixed model - which finds differentially expressed spots between two or more biological groups from a DIGE experiment. The algorithm is applicable to most experimental designs and automatically corrects for dye effects. The output from the algorithm is, for each protein, an estimate of the mean protein abundance in each group, the dye effects, the biological and technical variances, standard deviations on all quantities, and p-values for differentially expressed proteins and dye effects. By applying the algorithm to three DIGE data sets from our laboratory we have quantified typical magnitudes of the dye effects. A java program, called DIGEanalyzer, implementing the linear mixed model is freely available on the web page http://bioinfo.thep.lu.se/DIGEanalyzer.html

Experimental section

Gel experiments

Three DIGE data sets were used, one from breast tumors, one from ovarian tumors and one from rat brains. All DIGE gels were run with three samples labeled with Cy2, Cy3, and Cy5, respectively. One of the 3 samples was a pooled reference sample and two others were biological samples.

The breast tumor data set consists of 173 samples divided into four clinical groups: Group A of size 43, group B of size 15, group C of size 105 and group D of size 10. Protein extracts were prepared and labeled with DIGE minimal labeling according to manufacturer's instructions (GE Healthcare) using 400 pmol dye per 50 μg protein. All samples were pooled

and the pool was labeled with Cy2. Of the 173 samples, 171 were labeled with both Cy3 and Cy5 and the duplicates were run on different gels. One additional sample was run three times and the last sample was run once. A total of 173 DIGE gels were used. Isoelectric focusing was carried out on Serva IPG blue strips (24 cm, pH 4-7, Serva). The second dimension was run using the Ettan DALT II system (GE Healthcare) on 12.5% SDS-PAGE gels. The gels were fixed for 25 minutes in 10% HAc and 30% EtOH and equilibrated in water. The gels were scanned using an Amersham Biosciences Typhoon 9400 variable imager. Image analysis was performed using DeCyder 6.5 (GE Healthcare).

The ovarian tumor data set consists of 64 samples divided into four clinical groups: Group A of size 27, group B of size 12, group C of size 17 and group D of size 8. Protein extracts were prepared and labeled with DIGE minimal labeling according to manufacturer's instructions (GE Healthcare) using 600 pmol dye per 50 μg protein. All samples were pooled and the pool was labeled with Cy5. Every individual sample was labeled with both Cy2 and Cy3 and the duplicates were run on different gels together with the pool. A total of 64 DIGE gels were used. Isoelectric focusing was carried out on Immobiline DryStrips (24 cm, pH 4-7, GE Healthcare). The second dimension was run using the Ettan DALT II system (GE Healthcare) on 12.5% SDS-PAGE gels. The gels were fixed for 25 minutes in 10% HAc and 30% EtOH and equilibrated in water. The gels were scanned using an Amersham Biosciences Typhoon 9400 variable imager. Image analysis was performed using DeCyder 5.0 (GE Healthcare).

The rat brain data set consists of 24 samples divided into four clinical groups, A,B,C and D, each of size 6. Protein extracts were prepared and labeled with DIGE minimal labeling according to manufacturer's instructions (GE Healthcare) using 400 pmol dye per 50 μg protein. A fraction from all samples were pooled and the pool was labeled with Cy2. Every individual sample was labeled with both Cy3 and Cy5 and the duplicates were run on different gels. A total of 24 DIGE gels were used. Isoelectric focusing was carried out on Immobiline DryStrips (24 cm, pH 3-10, GE Healthcare) using the Ettan IPGphor Manifold (GE Healthcare). The second dimension was run using the Ettan DALT II system (GE Healthcare) on 12.5% SDS-PAGE gels. The gels were fixed for 25 minutes in 10% HAc and 30% EtOH and equilibrated in water. The gels were scanned using an Amersham Biosciences Typhoon 9400 variable imager. Image analysis was performed using DeCyder 5.0 (GE Healthcare).

These data sets are available as supplementary material.

Linear mixed model

The linear mixed model is described in the results section. Parameters are estimated with maximum likelihood estimation. P-values are calculated with the Wald test and the likelihood ratio test, which agree quite well. A full account of the algorithm is given as supplementary material.

Results and Discussion

Three DIGE experiments were performed, one with breast tumors (173 gels), one with ovarian tumors (64 gels) and one with rat brains (24 gels). Details about the experiments are given in the experimental section. Almost all samples were run in duplicate with two different dyes. Spot detection, spot volume quantification and normalization, and matching across gels were performed in DeCyder. All expression values in this study are base 2 logarithms of standardized abundance. Standardized abundance (SA) is defined as the ratio between the spot volume of the sample and the volume of the corresponding reference spot on the same gel. A set of matched spots across all gel images, from hereon called a matched spot set, would typically represent an

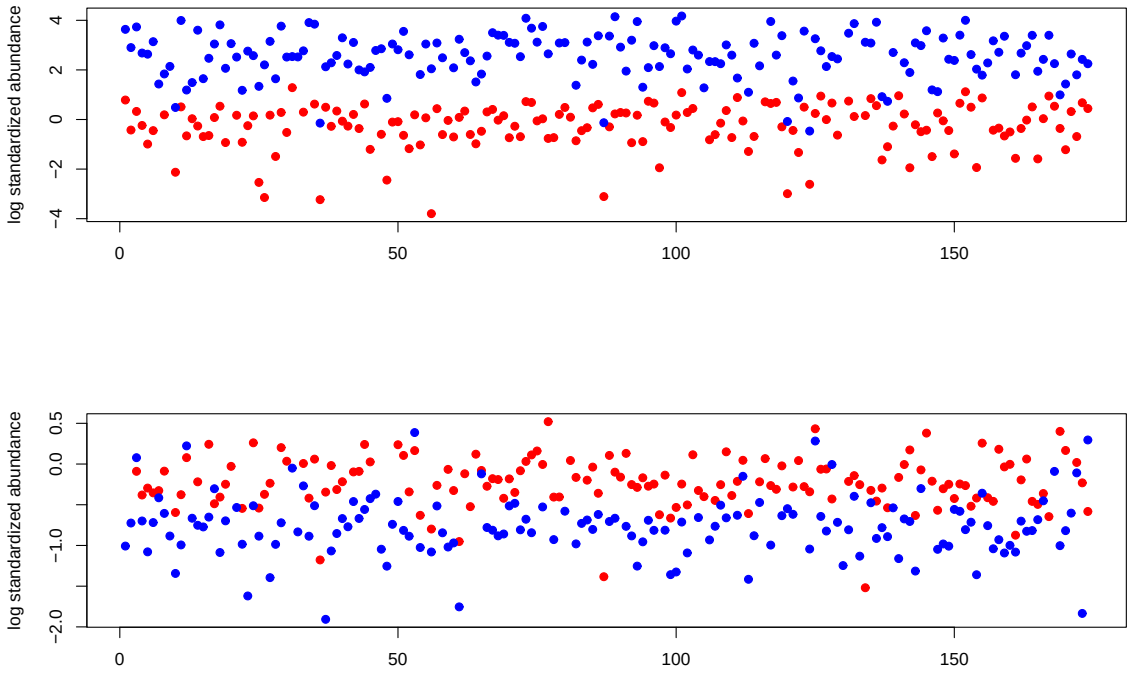


Figure 1: Protein-specific dye effect. For each of the 173 samples in the breast tumor data set, the log2 standardized abundance is shown in the Cy5 channel (blue) and Cy3 channel (red). Samples are shown along the x-axis in arbitrary order. For each x-axis location, i.e., sample, there are two points except for missing values. The two points would have been on top of each other if there had been no dye effect. The upper panel shows a matched spot set with higher overall standardized abundance in the Cy5 channel (binomial p-value $= 2 \times 10^{-46}$), whereas the lower show another matched spot set with higher overall standardized abundance in the Cy3 channel (binomial p-value $= 2 \times 10^{-28}$).

isoform of a protein, but in rarer cases it represents two or more unseparated proteins. Some of the spots in a matched spot set could be missing of course. We filtered out all matched spot sets with more than half of all spots missing, leaving 2027 matched spot sets in the breast tumor study, 757 in the ovarian tumor study and 2339 in the rat study.

First, we performed a statistical test for protein-specific dye effects. For each matched spot set in the breast tumor study, the number of samples, $N_{3>5}$, with a higher expression in the Cy3 measurement than in the Cy5 duplicate was calculated. The total number of samples with a non-missing measurement in both Cy3 and Cy5 is called N. If there were no systematic difference between Cy3 and Cy5 expression levels, $N_{3>5}$ should, for each matched spot set, follow a binomial distribution with probability 0.5 and N draws. The two-sided p-value of rejecting probability 0.5 is calculated as the sum of the lower and upper tail probabilities. So for each matched spot set, a p-value for rejecting equal expression in the Cy3 and Cy5 channels was obtained. Of the 2027 matched spot sets in the breast tumor study, 289 have a p-value

below 0.001. Of those 289 matched spot sets, 85 have higher expression in the Cy3 channel and 204 in the Cy5 channel. 55 matched spot sets have a p-value below 10^{-10} and 640 a p-value below 0.05. The most significant matched spot sets with higher expression in Cy3 and Cy5, respectively, are shown in Figure 1. Similar calculations were performed for the two other data sets. Of the 757 matched spot sets in the ovarian tumor study, 221 have a p-value below 0.001 and 438 have a p-value below 0.05. Of the 2339 matched spot sets in the rat brain study, 100 have a p-value below 0.001 and 430 have a p-value below 0.05. In all three studies, the number of matched spot sets with a significant dye effect is much higher than expected by chance. The effect is clearly protein-specific, and no global normalization in the Cy2, Cy3 or Cy5 channel would remove these effects because they occur with positive and negative sign. Later, we show the quantification of the dye effects in more detail.

The results above imply that dye corrections are necessary to fully take advantage of DIGE experiments. The simplest way to eliminate dye effects is to carry out a dye swap and use the average of the two measurements as the resulting expression value. However, there are two problems with this approach. Firstly, the dye swap doubles the number of gels needed, or, equivalently, halves the number of biological samples in the experiment. Hence, depending on cost issues and availability of biological material, the dye swap could be a suboptimal experimental design. Secondly, missing values occur as a result of spot detection and matching, which could lead to some data points not having a dye swap partner even if the original experimental design had dye swaps of all samples.

Here, we present an DIGE analysis algorithm which does not require dye swapping. The algorithm works for the following quite general experimental designs: One or more biological groups are compared, one or more biological replicas within each group are used, and one or more technical replicas of these are used. Each sample is labeled with one dye out of any number of possible dyes. Almost any experiment follows this design, and it would be straightforward to expand our algorithm to even more levels of nested groups if necessary.

In DIGE, there usually is two dyes in total (the reference sample is used to define the standardized abundance), but that is not a requirement of the algorithm. Let x_{ijdk} be the base 2 logarithm of the standardized abundance for the technical replica k of the biological replica j in group i measured with dye d. We model x_{ijdk} as

$$x_{ijdk} = \mu_i + m_d + \varepsilon_{ij} + \varepsilon_{ijdk}$$

where μ_i is the mean expression of group i and m_d is the effect of dye d. ε_{ij} and ε_{ijdk} represent the biological and technical random effects, respectively, with $\varepsilon_{ij} \sim N(0, \sigma_b^2)$ and $\varepsilon_{ijdk} \sim N(0, \sigma_t^2)$. This is a linear mixed model with the group mean and the dye effect as fixed effects and the biological and technical variation as random effects [20]. In the case where there are solely biological or technical replicates by experimental design or missing measurements, the model is adjusted to only have the corresponding random effect. The model is over-parametrized, so we set $m_1 = 0$. It is only the relative difference between two or more dyes that is well defined. The parameters μ, m, σ_b and σ_t are estimated using maximum likelihood estimation. The p-value that any linear combination of parameters is equal to zero is calculated with both the Wald test and the likelihood ratio test. It is known theoretically that these two tests more or less agree, which is also what we see in our data sets. A full account of all details is given in supplementary material.

The linear mixed model was applied to all three data sets. Figure 2 shows a histogram of the dye effects for proteins with a dye effect p-value below 0.001. The dye effects are measured in log2 of standardized abundance. It is seen that, for all three data sets, there are both

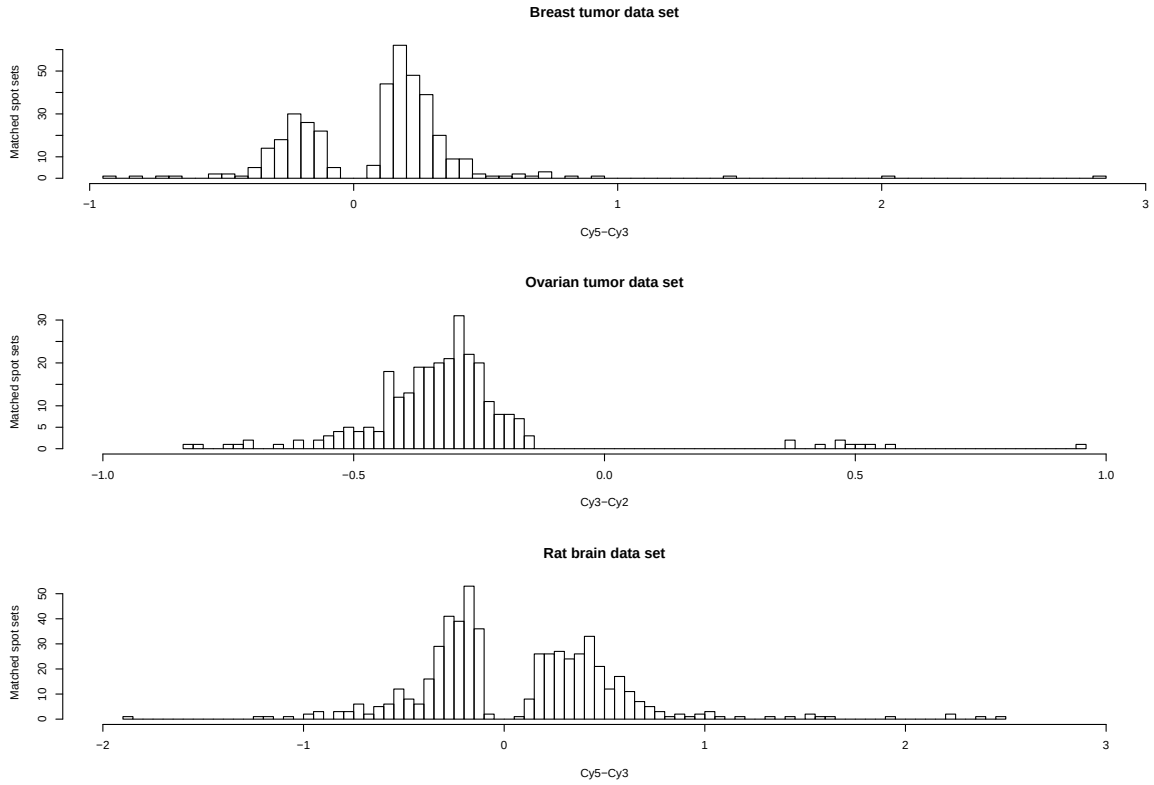


Figure 2: Histogram of dye effects. Histograms of the magnitude of the dye effects for all matched spot sets with a dye effect Wald test p-value below 0.001. The x-axis shows the magnitude of the dye effect using log2 of standardized abundance. The y-axis shows the number of matched spot sets. The matched spot sets were grouped into 100 bins along the x-axis.

positive and negative dye effects. This rules out that a single global normalization could remove the dye effects. The largest single dye effect is found in the breast tumor data set and has a log2 value of 2.8 for Cy5-Cy3. This corresponds to a 6 fold change. Table 1 lists various characteristics of the distributions of dye effects in the three data sets. The p-values in table 1 are calculated with the linear mixed model, and hence deviates a bit from the binomial test employed above. It is reassuring that both the linear mixed model and the binomial test find significant dye effects for all three data sets.

Conclusions

We have shown that protein-specific dye effects occur in DIGE experiments. No global normalization can remove these effects. The only reason we can think of for this protein-specificity is that the labeling process is amino acid sequence and dye specific with an interaction between the two. Suppose we normalize Cy2, Cy3, and Cy5 intensities such that some protein has equal intensities in the three channels, then some other protein would give a stronger signal in Cy5 than Cy3, another protein a stronger signal in Cy2 etc. Since the labeling is a chemical binding, such an effect is not surprising; some amino acid sequences might preferentially bind to Cy5, others to Cy3 etc.

Table 1 - Magnitude of dye effects.

	Breast tumor Cy5-Cy3	Ovarian tumor Cy3-Cy2	Rat brain Cy5-Cy3
Number of proteins	2027	757	2339
Median absolute value of dye effect	0.08	0.22	0.11
Number of proteins with p-value $\leq$0.001	381	258	545
Number of proteins with p-value $\leq$0.05	809	466	1014
Number of proteins with absolute dye effect larger than 0.5 (1.4 fold change)	19	28	141
Number of proteins with positive dye effect	960	104	1097
Number of proteins with negative dye effect	1066	653	1242

For each of the three data sets, six characteristics of the dye effect distribution is given. The dye effect is the difference of log2 standardized abundance between the two dyes indicated below each of the data sets. All values are calculated with the linear mixed model. P-values are calculated with the Wald test which gives slightly different results from the binomial test.

An algorithm has been presented that, for each matched spot set, estimates the dye effect and corrects for it. The algorithm is quite generic, works for most experimental designs, and can also be used in the case of just one dye, where dye effects do not exist. Also, the algorithm is not restricted to DIGE experiments or even protein experiments, even though that has been the focus of this study. The experimental design need not include dye swaps. Actually, with this algorithm in hand, it would be beneficial, given a fixed number of gels, to run twice the number of biological samples instead of making dye swaps.

The magnitude of dye effects is summarized in Table 1 and Figure 2. The breast tumor and rat brain studies both use Cy5 and Cy3. It is reassuring that those two data sets show similar behavior for the dye effects. The median absolute value of the dye effects are 0.08 and 0.11 respectively, both data sets have a quite symmetrical distribution around zero, and both have some proteins with very large dye effects. The ovarian tumor data set, which uses Cy3 and Cy2, has a larger median absolute value of the dye effect of 0.22 and a distribution which is skewed. It should be noted that a global normalization could change the center of the distribution and hence the median absolute value of the dye effect. Using this algorithm and mass spectrometry for protein identification, it might be possible to gain insight into the chemical reasons for protein-specific dye effects.

A java program, called DIGEanalyzer, implementing the linear mixed model is made publicly available on the web page http://bioinfo.thep.lu.se/DIGEanalyzer.html. A screen shot of DIGEanalyzer is shown in Figure 3.

Acknowledgements

MK and PJ are supported by the Swedish Foundation for Strategic Research, the Knut and Alice Wallenberg Foundation through the Swegene consortium and the Strategic Science Foundation (SSF) CREATE Health centre. SB is supported by Cancerfonden (Sweden). YL is supported by the research school in genomics and bioinformatics. BV is supported by Merz pharmaceutical Gmbh (Germany) and by the NSERC (Canada). We thank Markus Ringner for proofreading the manuscript.

Supporting Information Available:

The algorithm is made available in a java program at http://bioinfo.thep.lu.se/DIGEanalyzer.html.

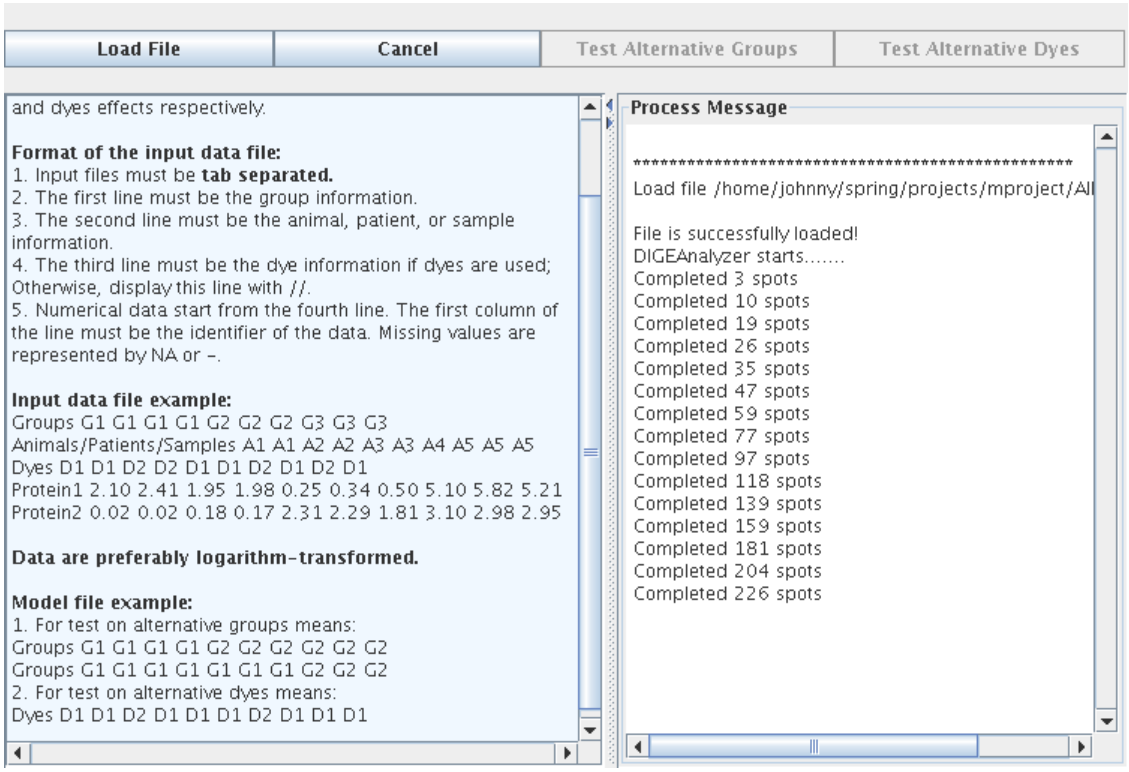


Figure 3: DIGEanalyzer. Screen shot of DIGEanalyzer during the analysis of a dataset. The program is written in Java and the user runs the program through a clickable window.

References

1. Stults, J. T; Arnott, D. *Methods in Enzymology* **2005,** 402, 245-289.
2. Govorun, V. M.; Archakov, A. I. *Biochemistry(Moscow)* **2002,** 67, 1109-1123.
3. Pandey, A.; Mann, M. *Nature* **2000,** 405, 837-846.
4. James, P. Proteome Rsearch: Mass Spectrometry *Springer-Verlag, Berlin* **2001,**
5. O'Farrell, P. H. *J. Biol. Chem.***1975,** 250, 4007-4021.
6. Klose, J. *Humangenetik* **1975,** 26, 231-243.
7. Scheele, G. A. *J. Biol. Chem.* **1975,** 250, 5375-5385.
8. Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R. *Nat. Biotechnol.* **1999,** 17, 994-999
9. Hansen, K. C.; Schmitt-Ulms, G.; Chalkley, R. J.; Hirsch, J.; Baldwin, M. A.; Burlingame, A. L. *Mol. Cell. Proteomics* **2003,** 2, 299-314
10. Ross, P. L.; Huang, Y. N.; Marchese, J. N.; Williamson, B.; Parker, K.; Hattan, S.; Khainovski, N.; Pillai, S.; Dey, S.; Daniels, S.; Purkayastha, S.; Juhasz, P.; Martin, S.; Bartlet-Jones, M.; He, F.; Jacobson, A.; Pappin, D. J. *Mol. Cell. Proteomics* **2004,** 3 (12), 1154-1169.
11. Ünlü, M.; Morgan, M. E.; Minden J. S. *Electrophoresis* **1997** 18, 2071-2077.
12. Alban, A.; David, S. O.; Bjorkesten, L.; Andersson, C.; Sloge, E.; Lewis, S.; Currie, I. *Proteomics* **2003** 3(1), 36-44.
13. Gharbi, S.; Gaffney, P.; Yang, A.; Zvelebil, M. J.; Cramer, R.; Waterfield, M. D.; Timms, J. *Mol. Cell. Proteomics* **2002,** 1(2), 91-98.

14. Morita, A.; Miyagi, E.; Yasumitsu, H.; Kawasaki, H.; Hirano, H.; Hirahara, F. *Proteomics* **2006** 6(21), 5880-5890.

15. Rantalainen, M.; Cloarec, O.; Beckonert, O.; Wilson, I. D.; Jackson, D.; Tonge, R.; Rowlinson, R.; Rayner, S.; Nickson, J.; Wilkinson, R. W.; Mills, J. D.; Trygg, J.; Nicholson, J. K.; Holmes, E. *Journal of Proteome Research* **2006** 5(10), 2642-2655.

16. Karp, N. A.; Lilley , K. S. *Proteomics* **2005** 5, 3105-3115.

17. Karp, N. A.; Kreil, D. P.; Lilley , K. S. *Proteomics* **2004** 4, 1421-1432.

18. Karp, N. A.; Griffin, J. L.; Lilley , K. S. *Proteomics* **2005** 5, 81-90.

19. Karp, N. A.; Spencer, M.; Lindsay, H.; O'Dell, K.; Lilley , K. S. *Journal of Proteome Research* **2005** 4, 1867-1871.

20. McCulloch, C. E.; Shayle R. S. Generalized, Linear, and Mixed Models *John Wiley & Sons, USA* **2001**

Supplemental Methods

The Linear Mixed Model

Let x_{ijdk} be the base 2 logarithm of the abundance for the k^{th} technical replica of the j^{th} biological replica in group i measured with dye d. We model x_{ijdk} as

$$x_{ijdk} = \mu_i + m_d + \varepsilon_{ij} + \varepsilon_{ijdk} \qquad (1)$$

where μ_i is the mean expression of the i^{th} group, m_d is the effect of the d^{th} dye. ε_{ij} and ε_{ijdk} represent the biological and technical random effect respectively, with $\varepsilon_{ij} \sim N(0, \sigma_b^2)$ and $\varepsilon_{ijdk} \sim N(0, \sigma_t^2)$. The model is overparameterized and the constraint $m_1 = 0$ is used. In the case where there are solely biological or technical replicates by experimental design or missing measurements, the model is adjusted to only have the corresponding random effect.

Parameter Estimation

Suppose that there are n groups, n_i biological replica for group i, n_{ij} measurements for the j^{th} biological replica in group i and p dyes used, the likelihood function of this model is

$$\begin{aligned} L(\boldsymbol{\theta}|X) &= \Pi_{i=1}^{n}\Pi_{j=1}^{n_i} \int d\varepsilon_{ij} \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\left(-\frac{\varepsilon_{ij}^2}{2\sigma_b^2}\right) \\ &\quad \times \Pi_{dk}^{n_{ij}} \int d\varepsilon_{ijdk} \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left(-\frac{\varepsilon_{ijdk}^2}{2\sigma_t^2}\right) \\ &\quad \times \delta(x_{ijdk} = \mu_i + m_d + \varepsilon_{ij} + \varepsilon_{ijdk}) \qquad (2) \end{aligned}$$

where $\boldsymbol{\theta} = (\mu_1, \ldots, \mu_n, m_1, \ldots, m_p, \sigma_b, \sigma_t)$. m_1 is set to be zero. The rest of the parameters are estimated by maximum likelihood estimation (mle). Thus, only the relative difference between two or more dyes is well defined in the model. Parameter values of spots with $n_{\text{measurements}} \leq n_{\text{groups}} + n_{\text{dyes}} + 2$ are set to be "NaN" (unavailable), as spots with too few measurements are not interesting and the mle becomes unreliable.We impose the constraint $\sigma_b^2 \geq \frac{\sigma_t^2}{n_r}$ [1], where $n_r = \frac{n_{\text{measurements}}}{n_{\text{individuals}}}$ to avoid solutions where σ_b runs to zero. It is unreasonable to claim that σ_b^2 is smaller than $\frac{\sigma_t^2}{n_r}$, since the latter quantity approximately sets the precision with which each biological sample is measured.

The maximization of the likelihood was processed in a way similar to the EM (Expectation Maximization) approach. Initializing σ_b and σ_t with certain values, we calculated the mle estimated values of the other parameters. These values were then substituted into the likelihood function and σ_b and σ_t were found by numerical maximization using the conjugate gradient descent method. The new estimated σ_b and σ_t were again used to calculated new values of the other parameters. This procedure was repeated until it converged. We used two sets of initial values for σ_b and σ_t. For one of them, σ_t was initialized with the standard deviation of the measurements of all spots. σ_b was twice of σ_t. For the other set, we initialized σ_b and σ_t with values corresponding to a point on the line $\sigma_b^2 = \frac{\sigma_t^2}{n_r}$ that had maximal likelihood. We took the estimated parameter values from the set having larger likelihood. In the case where the model has only one random effect, of course no numerical optimization is needed.

The variance of the estimated parameter values was calculated from the information matrix. For N parameters, the information matrix takes the form

of an NxN matrix [2], with elements:

$$I(\boldsymbol{\theta})_{ij} = -E\left[\frac{\partial}{\partial \boldsymbol{\theta}_i} \ln L(\boldsymbol{\theta}|X) \frac{\partial}{\partial \boldsymbol{\theta}_j} \ln L(\boldsymbol{\theta}|X)\right] \quad (3)$$

The variances of the partameters are given by the diagonal elements of the inverse of the information matrix.

Likelihood Ratio Test

We used likelihood ratio test to investigate the equivalence of group means and dye means. Let the original model for the data be model_1, and the constrained model (for instance equal dye means) be model_0. The likelihood ratio test statistic is given by

$$LR = 2 \ln \left(\frac{L_1(\boldsymbol{\theta}_1|X)}{L_0(\boldsymbol{\theta}_0|X)} \right) \quad (4)$$

where L_0 and L_1 are the maximal likelihood of model_0 and model_1 respectively. LR is known to asymptotically follow a χ^2 distribution. The P-value of rejecting the hypothesis that model_0 describes the data as well as model_1 is calculated as the probability of a value of LR or larger from a χ^2 distribution with degrees of freedom equal to the difference in dimensionality of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_0$. The likelihood ratio test is not performed for spots with $n_{\text{measurements}} < 3n_{\text{groups}} + n_{\text{dyes}}$ since we found that this approach became unreliable when there were too few measurements.

Wald Test

We also calculated the P-values for equal group means and dye means using the Wald test [3]. Let $\hat{\boldsymbol{\theta}}$ be the mle estimate of $\boldsymbol{\theta}$ and $\text{I}(\boldsymbol{\theta})$ be the information matrix for $\hat{\boldsymbol{\theta}}$, the Wald test statistic for the hypothesis $H : \boldsymbol{\theta} = \boldsymbol{\theta}_*$ is given by

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*)'[\text{I}(\boldsymbol{\theta})_*]^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*) \quad (5)$$

This test statistic is approximately χ^2 distributed with order equal to the dimension of $\boldsymbol{\theta}$, similar to that of the likelihood ratio test.

References

1. W.H. Press, S.A. Teukolsky, W.T., Vetterling and B.P. Flannery: Numerical recipes in C. *Cambridge University Press* 1999.

2. J.A. Rice: Mathematical Statistics and Data Analysis. *Duxbury*, 1995

3. A. Wald: Asymptotically most powerful tests of statistical hypotheses. Annals of Mathematical Statistics, 12:1-19.

Paper II

BMC Bioinformatics

BioMed Central

Software

Multiclass discovery in array data

Yingchun Liu and Markus Ringnér*

Address: Complex Systems Division, Department of Theoretical Physics, Lund University, Sölvegatan 14A, SE-223 62 Lund, Sweden

Email: Yingchun Liu - spring@thep.lu.se; Markus Ringnér* - markus@thep.lu.se

* Corresponding author

Published: 04 June 2004

BMC Bioinformatics 2004, **5**:70

Received: 21 January 2004

Accepted: 04 June 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/70

Abstract

Background: A routine goal in the analysis of microarray data is to identify genes with expression levels that correlate with known classes of experiments. In a growing number of array data sets, it has been shown that there is an over-abundance of genes that discriminate between known classes as compared to expectations for random classes. Therefore, one can search for novel classes in array data by looking for partitions of experiments for which there are an over-abundance of discriminatory genes. We have previously used such an approach in a breast cancer study.

Results: We describe the implementation of an unsupervised classification method for class discovery in microarray data. The method allows for discovery of more than two classes. We applied our method on two published microarray data sets: small round blue cell tumors and breast tumors. The method predicts relevant classes in the data sets with high success rates.

Conclusions: We conclude that the proposed method is accurate and efficient in finding biologically relevant classes in microarray data. Additionally, the method is useful for quality control of microarray experiments. We have made the method available as a computer program.

Background

A common application in microarray data analysis is to identify genes that, based on their expression levels, discriminate between known classes of experiments. This identification is often achieved by using various statistical measures to, gene-by-gene, correlate the expression levels with the classes of interest. In this way a discriminatory weight is calculated for each gene. For example, Golub *et al*. used a signal-to-noise statistic to find genes with expression patterns that discriminate between samples obtained from patients with acute myeloid leukemia and patients with acute lymphoblastic leukemia [1]. Other examples include using a standard *t*-test to discriminate between breast tumors from carriers of *BRCA1* mutations and carriers of *BRCA2* mutations [2]. For an overview of applications see [3]. In most studies, the number of genes is much larger than the number of experiments. For such a large number of genes, it is crucial to estimate how many genes would correlate with the classes of interest by chance. Often, a *P* value corresponding to the probability of obtaining a given weight by chance is calculated for each weight. One can then investigate if there is an over-abundance of discriminatory genes for classes of interest as compared to randomly selected classes. Indeed, such an over-abundance has been found for many microarray-based classification applications (see *e.g.* [1,2,4]).

Often clustering methods, such as hierarchical clustering [5], *k*-means clustering [6], or self-organizing maps (SOM) [7] are used for unsupervised classification of array data (see [8] for an overview). For example, hierarchical clustering has been used to discover two subtypes of diffuse B-cell lymphoma [9], three subtypes of breast tumors [10], and two subtypes of cutaneous melanoma [4].

Dugas *et al*. have developed an iterative *k*-means clustering method for class discovery in array data [11]. Examples of SOM-based methods for discovery of cancer subtypes include applying SOMs to automatically discover the distinction between acute myeloid leukemia and acute lymphoblastic leukemia [1], and to separate 14 different tumor subtypes [12]. SOMs and *k*-means clustering require the user to predefine the number of clusters to be found. Hsu *et al*. proposed an unsupervised hierarchical self-organizing map approach that automatically identifies a suitable number of clusters, and applied it to a couple of publicly available array data sets [13].

In these clustering methods, experiments are clustered based on the distance between them in gene expression space. An alternative unsupervised classification approach to discover classes in gene expression data, which exploits the fact that there typically is an over-abundance of genes separating known classes was proposed by Ben-Dor *et al*. [14]. In their method, classes are discovered by seeking partitions of experiments with an over-abundance of discriminatory genes. In contrast to many clustering methods, no metric to define distances between experiments is required. Furthermore, classes are discovered based only on the subset of genes that are differentially expressed between the classes, whereas in unsupervised clustering the distances are often based on all the genes. A similar classification method, which also searches for binary class distinctions in a set of samples that show separation in the expression of subsets of genes, has been developed by von Heydebreck *et al*. [15]. These classification methods are well-suited to discover several significant partitions of experiments, each based on a different subset of genes.

Inspired by the method by Ben-Dor *et al*., we have previously used a similar approach to sub-classify familial breast cancer into two classes [16]. Briefly, the approach was as follows. For a given partition of the experiments into two classes (with n_1 and n_2 experiments, respectively), a discriminative weight was calculated for each gene using the signal-to-noise statistic [1]. To assign *P* values to the weights one has to perform random permutation tests. Such a test was used to generate a weight distribution that could be expected for two classes with n_1 and n_2 experiments under the assumption of random gene expression. Using this weight distribution, each weight was assigned a *P* value corresponding to the probability to obtain the weight or larger for a random partitioning into n_1 and n_2 experiments. Candidate partitions of the data were scored with the number of statistically significant weights, that is the number of genes that were significantly different in expression between the classes. A simulated annealing [17] scheme was used, in which partitions were updated by changing the class of a randomly selected experiment, to find the partition of experiments into the classes with the highest score. Our approach is different from the Ben-Dor *et al*. method in two respects. First, they use the total number of misclassification (TNoM) score to find discriminatory genes [18]. Second, we use a fixed *P* value cutoff to find the number of discriminatory genes, whereas they instead use surprise scores [14].

In this work, we have extended our unsupervised classification method for discovery of more than two classes and to allow for missing values in gene expression data. Furthermore, we have made the method publicly available as a computer program. For the breast cancer study [16], we performed random permutation tests for all possible n_1 and n_2 that add up to the total number of experiments. Extending the method to find a preset but arbitrary number of classes and to allow for missing values would result in performing random permutation tests for many more combinations of class sizes. For nonparametric rank-based statistics, analytically calculated *P* values correspond to what would be obtained by random permutation tests. Therefore, we decided to use such statistics instead of a parametric test. Moreover, nonparametric methods have been shown to be robust conservative (low numbers of false positives) in its application to the identification of discriminatory genes in gene expression data [19].

Here, we describe the unsupervised classification method used in our class discovery program in detail, and results from applying it on two publicly available data sets.

Implementation

Identification of differentially expressed genes

The Wilcoxon rank sum test (WT) is used to identify genes differentially expressed in two classes [21]. The nonparametric WT tests for equality of medians of two samples of data, but unlike the *t*-test it makes no assumption that the data is normally distributed. It operates on rank-transformed data rather than the raw values. In our method, the expression values of each gene are ranked across experiments from low to high, disregarding to which class each experiment belongs. For a given partition of experiments into two classes, a discriminatory weight u_g is calculated for each gene (g),

$$\text{class 1: } n_1 \text{ samples} \quad \text{class 2: } n_2 \text{ samples} \quad n_2 > n_1$$

$$w_g = \sum_{e \in \text{class 1}} \text{rank}_g(e)$$

$$u_g = w_g - n_1(n_1 + 1)/2$$

where e denotes an experiment with $\text{rank}_g(e)$ for g. Next, we want to calculate a *P* value of u_g for the null hypothesis that expression values for all experiments are drawn from the same probability distribution. If the *P* value is near

zero, it casts doubt on the null hypothesis and suggests that the medians of expression values are significantly different in the two classes. For $n_1 > 8$ (and thus $n_2 > 8$ also), the *P* value can be calculated by using a normal approximation [22],

$$\begin{aligned} \text{mean}_{u_g} &= n_1 n_2 / 2 \\ \text{var}_{u_g} &= n_1 n_2 (n_1 + n_2 + 1)/12 \\ z &= \left(u_g - \text{mean}_{u_g} \right) / \sqrt{\text{var}_{u_g}} \\ z &\in N(0,1). \end{aligned}$$

For partitions with n_1 being at least 9, we use this normal approximation. For partitions with n_1 smaller than 9, we assign *P* values to u_g using a random permutation test. For each n_1 and n_2, the test is based on 50,000 random permutations of class labels. For the WT, we use two-sided tests.

For partitions of experiments into three or more classes, the Kruskal-Wallis test (KWT) is used to identify discriminatory genes [23]. The KWT is a nonparametric version of the one-way analysis of variance (ANOVA), uses the ranks of the data, and is an extension of the WT to more than two groups. It tests for equality of medians of *k* samples of data. For a given partition of experiments into *k* classes, a discriminatory weight H_g is calculated for each gene,

$$\begin{aligned} &\text{class } i : n_i \text{ samples} \\ &n = \sum_{i=1}^{k} n_i \\ &R_g(i) = \sum_{e \in \text{class i}} \text{rank}_g(e) \\ &H_g = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_g(i)^2}{n_i} - 3(n+1). \end{aligned}$$

For partitions where all classes have at least 5 samples, the sampling distribution of H_g can be approximated very well by a chi-square distribution with *k* - 1 degrees of freedom [22]. Hence, the *P* value for the null hypothesis that expression values for all experiments are drawn from the same probability distribution can be calculated from $\chi^2(k - 1)$. If the *P* value is close to zero it suggests that the median of expression levels for at least one class is significantly different from the others. This does not necessarily mean that every group differs from every other group. For partitions with all classes having at least 5 experiments, we use the chi-square approximation. *P* values are not calculated for partitions into more than 2 classes for which a class have less than 5 experiments.

Missing values is handled by using, for each gene, only those experiments for which there exists measurements in the calculation of the statistical test (WT or KWT).

Scoring partitions of experiments

Each partition of experiments is assigned a score corresponding to the number of genes with *P* values smaller than or equal to a user specified cut-off. Thus, the score is the number of genes with significantly different expression in the classes. Because the total number of genes for a data set is identical for all partitions, we do not correct *P* values for multiple testing. For random gene expressions, we would, for a typical partition, expect a score of *P* multiplied by the total number of genes. We denote this expected score by *E*.

Finding the partition with the highest score

We use simulated annealing (a global optimization method) to find the partition of the experiments into a given number of classes with the highest score [17,24]. This procedure is described in Fig. 1. In addition to the best partition, we extract all partitions investigated in the search that have a score larger than a user specified cut-off. Table 1 contains the values of the parameters used in the analysis of the data sets.

Peeling discriminatory genes

For a data set there may be several biologically relevant partitions. However, one of them is often supported by many more discriminatory genes than the others. In such a scenario, the discovered partitions with the highest scores will mostly be similar to the best one, because shifting one or two experiments will still result in a higher score as compared to other relevant partitions. Hence, it may be difficult to discover important but not dominating partitions. One way to address this issue is to remove genes that contribute to the score for the best partition and run the class discovery program again using this smaller set of genes [14]. Using such a procedure, one can iteratively peel discriminatory genes from the data set to systematically investigate the presence of further partitions with significant scores.

Evaluation

To further evaluate discovered classes we used hierarchical clustering. Clustering was performed using EPCLUST http://ep.ebi.ac.uk/EP/EPCLUST/ with the distance measure parameter set to linear correlation based distance (Pearson).

To investigate discovery of two classes, we used the non-*BRCA1/2* familial breast cancer (termed *BRCAx*) data set by Hedenfalk *et al*. [16]. This data set consists of 16 *BRCAx* samples, for which intensity ratios of 4,795 clones considered to be well-measured are provided. Following

1. Initialization

   (a) Initialize the labeling of the experiments by randomly assigning each experiment to one of k classes. For $k > 2$ require each class to have at least 5 experiments.

   (b) Initialize the 'temperature' $T = T_{start}$.

2. Procedure to optimize the partition score.

   (a) Calculate the partition score (S) for the labeling.

   (b) Randomly pick an experiment and change its label to a different label.
   - if $k = 2$, the experiment is randomly selected from all experiments.
   - if $k > 2$, the experiment is randomly selected from classes having more than 5 experiments.

   (c) Calculate the score for the changed labeling (S_{new}).
   - Accept the changed labeling, if $S_{new} > S$; otherwise, accept it with probability $e^{(S_{new} - S)/T}$.
   - If the changed labeling is accepted set $S = S_{new}$. If not, keep the original labeling and S.

   (d) Decrease T by a factor η ($T \leftarrow \eta T$), if $N_{success}$ changed labelings have been accepted , or N_{total} changed labelings have been proposed at the current T.

   (e) Repeat steps 2(b)-2(d) until T in step 2(d) becomes smaller than T_{end}.

3. Extract all partitions with scores larger than a threshold (S_c) and the discriminatory genes associated with each of them.

Figure 1
The essential algorithmic steps in the class discovery procedure. For actual values of the parameters used in the analysis see Table 1.

Hedenfalk *et al.*, we performed pre-processing of the data such that the log intensity ratios were mean-centered and these values were used as a measure of the expression levels.

To investigate discovery of more than two classes, we used the data for small round blue cell tumors (SRBCTs) of childhood by Khan *et al.* [20]. This data set consists of 88 samples, separated into a training and a test set, for which

Table 1: Parameters in the class discovery procedure and the values used for the SRBCT and the *BRCAx* data.

Parameter	Value
T_{start}	3.0
T_{end}	0.1
η	0.9
$N_{success}$	10
N_{total}	150 (SRBCT) or 50 (*BRCAx*)

Table 2: Batches of production for the 88 SRBCT microarrays.

Batch[a]	Category	Experiments[b]
104	EWS	T1, T2, T3, T4, C1, C2, C3, C4
	BL	C1, C2, C3, C4
	NB	C1, C2, C3
	RMS	T1, T2, T3, T4, C8, C11
	TEST	5, 24
118	EWS	T6, T7, T9, T11, T12, T13, T14, T15, T19
	RMS	T5, T6, T7, T8, C3, C4
	TEST	6, 9, 11, 20, 21
119	EWS	C6, C7, C8, C9, C10, C11
	BL	C5, C6, C7, C8
	NB	C4, C5, C6, C7, C8, C9, C10, C11, C12
	RMS	C2, C5, C6, C7, C9, C10
	TEST	3
143	RMS	T11
	TEST	1, 2, 4, 7, 12, 17
163	RMS	T10
	TEST	8, 10, 13, 14, 15, 16, 18, 19, 22, 23, 25

[a]Identifier of batch of production [b]T: tumor samples; C: cell lines

relative intensities of 2,308 filtered genes are provided. In the training set, there are 63 samples belonging to four different SRBCT types, neuroblastoma (NB), rhabdomyosarcoma (RMS), Burkitt's lymphoma (BL) and Ewing's sarcoma (EWS). The test set consists of 20 SRBCTs (belonging to the four types) and 5 non-SRBCTs. Following Khan *et al.*, we used the logarithm of the relative intensities as a measure of the expression levels. The 88 experiments were performed on microarrays from 5 different batches of production, and included samples from both tumor biopsy material and cell lines (Table 2).

Results and Discussion

Discovery of two classes in the breast cancer data set

To validate our class discovery program, we first applied it to the Hedenfalk *et al*. *BRCAx* data set and looked for two classes, as in our original analysis [16]. In that work, a signal-to-noise statistic [1] and a *P* value cut-off of 0.001 was used to discover two classes supported by 60 discriminatory genes. For a fixed *P* value cut-off, we now expect less discriminatory genes, because of the change to a rank-based statistic. Therefore, we decided to use a somewhat larger cut-off to facilitate a comparison of our set of discriminatory genes with the original set. Importantly, we found the classes discovered to be rather insensitive to changes in *P* value cut-off. Our best scoring partition was supported by 133 discriminatory genes ($P <= 0.005$), whereas we would expect 24 discriminatory genes for a random partition ($E = 24$). Our partition was similar to the one in the original study, except that three samples had shifted class. Moreover, the Hedenfalk *et al*. partition was also highly significant (score 117), and the top scoring partitions were dominated by partitions very similar to it. Of our 117 discriminatory genes 57 overlapped with the 60 found in ref. [16]. We conclude that the partition found in the original study is robust to changes in the details of the class discovery algorithm.

Setting the values of the parameters in the annealing schedule requires experimentation. For our data sets, we selected parameter values for which the program runs relatively fast (5–10 minutes on a standard personal computer), and for which running the program again, with a new random initialization of the classes, often resulted in again finding the best partition found previously. For more conservative parameter values ($T_{start} = 25$, $N_{success} = 100$, $N_{total} = 500$) for which the program takes 10–15 times longer to run, we found that the best partition was essentially found every time the program was run (9 out of 10 times for the breast cancer data).

Discovery of two classes in the SRBCT data set

We applied our class discovery method to the 63 experiments in the SRBCT data training set, leaving the 25 experiments in the test set for verification. For two classes, our best scoring partition was supported by 602 genes ($P <= 0.001$; $E = 2.3$ genes). We investigated the score (number of discriminatory genes) for our best scoring partition as a function of *P* value cut-off. The significance of the partition is relatively insensitive to the *P* value cut-off (Fig. 2). The two classes essentially separated cell lines from tumors, with 4 cell lines (EWS-C4, RMS-C3, C8 and C11) in the tumor class. Thus, we found, in agreement with principal component analysis of the expression data [25], that the dominant separation of the experiments is into cell lines and tumors (94% correct for our two classes). The score for perfect partitioning into cell lines and tumors was 513.

Discovery of three classes in the SRBCT data set

Next, we investigated discovery of more than two classes. For three classes, our best scoring partition was supported by 934 discriminatory genes ($P <= 0.001$; $E = 2.3$). The three classes separated experiments according to batches of microarray production (Table 2). Experiments on print

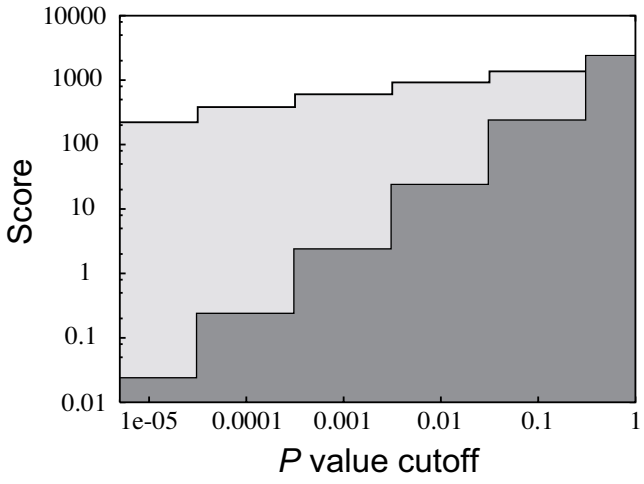


Figure 2
The number of discriminatory genes (score) as a function of the cut-off in *P* value. The data shown is for discovery of two classes in the SRBCT data set. The two curves are for the best partition found (light gray) and for random partitions (dark gray). For the *P* value cut-off 0.001, the best partition is supported by 602 genes, whereas the expectation for a random partition is 2.3 genes.

Table 3: The four classes of experiments identified by the class discovery program (score = 470; *P* < 0.001; *E* = 1.4) after removing 923 genes discriminatory for batches of array production.

Category[a]	Class 1	Class 2	Class 3	Class 4
BL	8	0	0	0
EWS-C	0	8	1	1
EWS-T & RMS-T	0	0	22	1
NB-C & RMS-C	0	2	2	18

[a]T: tumor samples; C: cell lines

Table 4: The four classes of experiments identified by the class discovery program (score = 353; *P* < 0.001; *E* = 1.2) after removing 1076 genes discriminatory for batches of array production or cell lines versus tumors.

Category	Class 1	Class 2	Class 3	Class 4
BL	8	0	0	0
EWS	0	12	10	1
RMS	0	4	13	3
NB	0	0	0	12

batches 104, 118 and 119 were in one class each. There was only one error: RMS-T4 (print batch 104) was in the class with experiments on print batch 118, as were the two experiments from batches 143 and 163. When only using the 61 training experiments belonging to batches 104, 118 and 119, the score for a perfect partitioning into these three batches was 923.

Discovery of four classes in the SRBCT data set

For four classes, our best scoring partition was supported by 1051 discriminatory genes ($P <= 0.001$; $E = 2.3$). The four classes were identical to our result for three classes (according to print batches), except that the BL experiments (on batches 104 and 119) were in a class of their own. Thus, apart from finding the BL class, the dominant separation of the experiments into four classes is according to print batches. In agreement with our result, BL was in the original supervised analysis found to be the category with the most distinct expression profile [20]. The score for perfect partitioning into the four SRBCT categories was 544, which though not the best score is highly significant compared to random expectations.

Peeling genes discriminatory for batches of array production

To reveal partitions into four biologically relevant classes, we proceeded by removing genes discriminatory for print batches from the dataset. We removed the 923 genes found to be discriminatory between the 3 major print batches. The best partition found was supported by 470 discriminatory genes ($P <= 0.001$; $E = 1.4$). The four classes corresponded to BL, EWS cell lines, EWS and RMS tumors, and NB and RMS cell lines, respectively (see Table 3), with 89% of the experiments correctly assigned to these categories. Using the four SRBCT categories instead, the four classes corresponded to correct assignment of 48% of the experiments. For this reduced data set, the score for perfect partitioning into the four SRBCT categories was 390. Hence, the removal of genes discriminatory for print batches had the desired effect: the dominant partition no longer reflected print batches, but instead biologically relevant categories. However, there was a separation of tumors from cell lines.

Peeling genes discriminatory between cell lines and tumors

To further reveal relevant partitions, we also removed the 513 genes discriminatory between cell lines and tumors. There was an overlap between these genes and the 923 previously removed print batch discriminatory genes, resulting in the removal of a total of 1076 genes. Using neural networks, Khan *et al*. identified a set of 96 genes with which they were capable of classifying the four categories. Of these 96 genes, 62 remained in our dataset after peeling. For the peeled data set, the best partition found was supported by 353 discriminatory genes ($P <= 0.001$; $E = 1.2$), including 46 of the 62 genes identified by Khan *et*

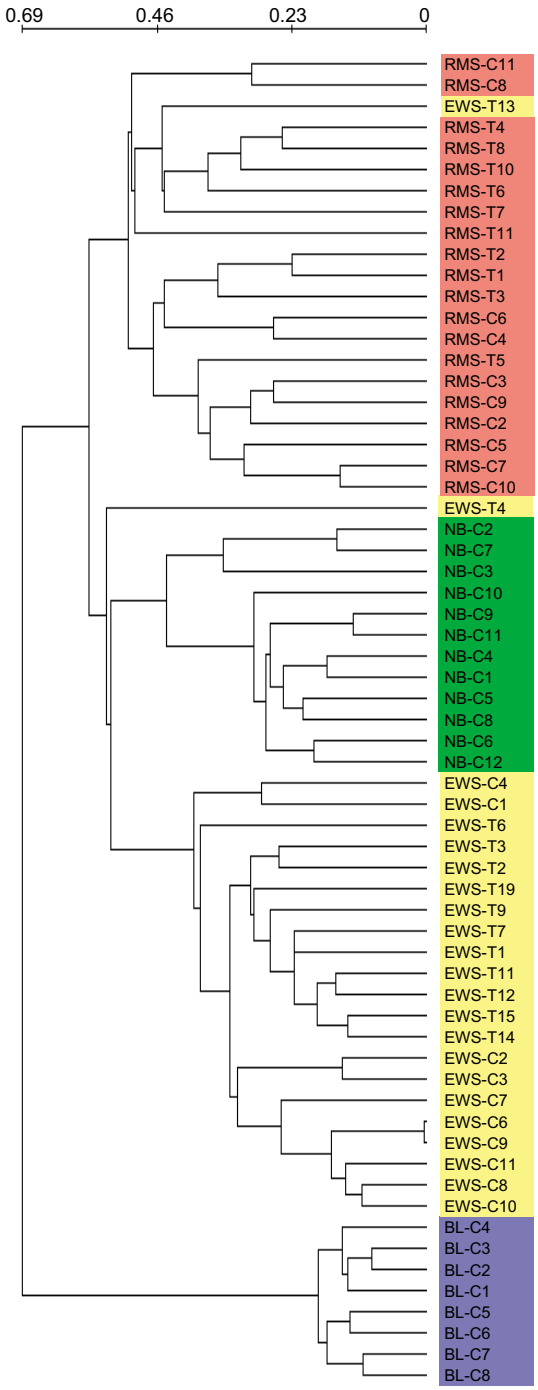


Figure 3
Hierarchical clustering of the 63 SRBCT training experiments. The clustering was performed using the 353 genes discriminatory for the best partition found in the data set reduced for genes discriminating cell lines versus tumors or between print batches. Using the discriminatory genes found by our unsupervised method results in clusters that correspond to the disease categories. The scale shows the linear correlation based distance used to construct the dendrogram.

al. The separation of the SRBCT categories was improved, and with 71% of the experiments correctly assigned, the four classes corresponded to the SRBCT categories (see Table 4). Most of the mis-classifications were mistakes between RMS and EWS experiments, and is due to differences between cell lines and tumors that still remain in the reduced data set. The score for perfect partitioning into the four SRBCT categories was 319, including all 62 of the genes identified by Khan *et al*. Furthermore, hierarchical clustering using the 353 genes significant for the best partition, clearly clustered the SRBCT categories into distinct clusters (Fig. 3). Only two samples were misplaced: EWS-T13 were in the RMS cluster and EWS-T4 was an outlier. We conclude that by removing genes discriminatory between cell lines and tumors or between print batches, we can discover the four SRBCT categories with a high success rate.

Results for additional SRBCT test samples

Finally, we wanted to investigate the robustness of the discovered classes using our unsupervised method on independent test data. Therefore, we included the 25 test experiments and performed hierarchical clustering of all the 88 experiments in the SRBCT data set using the 353 genes (Fig. 4). Again the SRBCT categories clustered into fairly distinct clusters. 15 of the 20 test experiments belonging to one of the four SRBCT categories clustered into clusters dominated by their category. The 5 misplaced experiments (4 EWS and one NB) were in clusters dominated by RMS. Some of these mistakes reflect that many of the test experiments were performed on print batches not corrected for in our class discovery analysis because they were rarely used for training experiments. We conclude that the discriminatory genes found in an unsupervised class discovery analysis can be used to successfully cluster additional experiments. Nonetheless, one should keep in mind that by incorporating which SRBCT category each experiment belongs to in a supervised analysis, one can separate the experiments with 100% success rate [20].

Conclusions

We have developed an unsupervised classification method for the discovery of two or more classes of experiments in microarray data. The method has been implemented as a publicly available computer program. We have tested the method on two published gene expression data sets and conclude that the proposed method is effective in finding relevant classes. We are planning to make the program available as a plugin for BASE [26], the open-source database for array data maintained by our group.

When applying our method on these data sets, we have found that the best partitions discovered are relatively insensitive to the cut-off in *P* value. Moreover, even

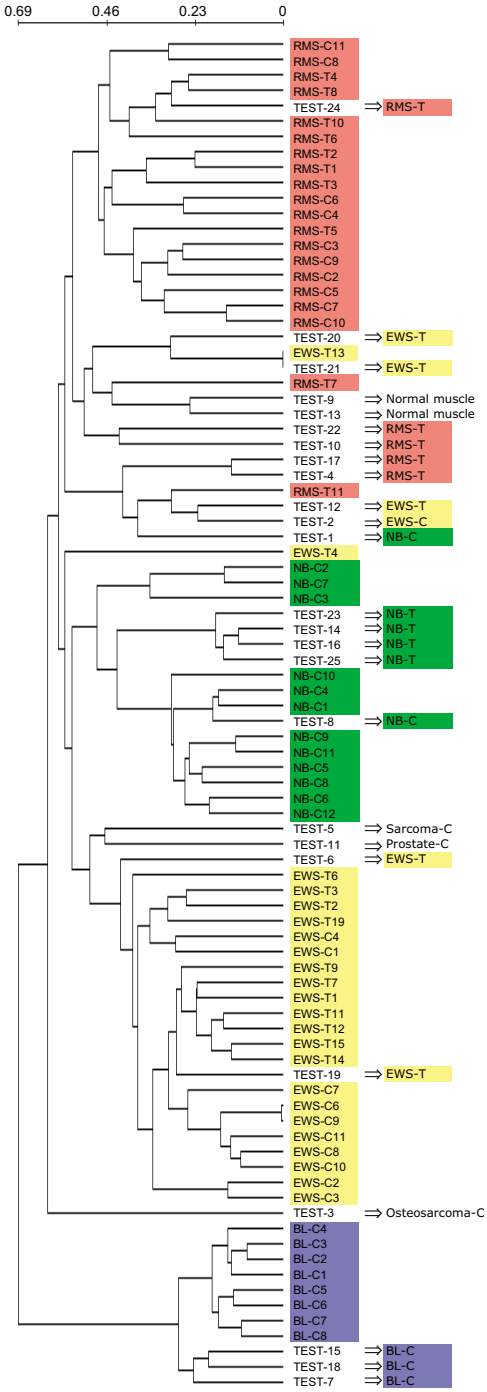


Figure 4
Hierarchical clustering of all 88 SRBCT experiments. The clustering was performed using the 353 genes discriminatory for the best partition found using our unsupervised method applied to the training data set reduced for genes discriminating cell lines versus tumors or between print batches. Using these genes, the test samples cluster in clusters dominated by the correct disease category. The scale shows the linear correlation based distance used to construct the dendrogram.

though there is no guarantee that the simulated annealing algorithm finds the optimal score, we found that high-quality local minimas are discovered. Nevertheless, any user of the program will benefit from experimenting with the values of the parameters in the algorithm to explore the particular details of each data set.

The method was designed for unsupervised classification, but it can also be very useful for quality control when classes of experiments are known. It is common to look for an over-abundance of discriminatory genes separating known classes. In such a scenario, it may be useful to seek the partitions having the largest overabundance of discriminatory genes. Thereby, one can rule out potential problems, as highlighted by our example of the print batches for the SRBCT experiments. The SRBCT experiments were performed on arrays produced by the first generation of cDNA microarray printers. Using our class discovery program on more recent data sets our experience is that differences due to print batches are much smaller (data not shown). Nevertheless, we think that our results using the SRBCT data set illustrates how experimental artifacts can be found and corrected for when using our class discovery program. Here, one should keep in mind the crucial importance of random experimental design: if each SRBCT category had been investigated using a unique print batch, there would be no way to disentangle the relevant biology from artifacts. Moreover, these findings illustrate that it is important to know the procedural steps underlying an experiment to be able to interpret discovered classes.

Availability and requirements

Project name: MCD – Multiclass discoverer

Project homepage: http://bioinfo.thep.lu.se/classdiscoverer

Operating systems: Linux, Windows, Mac OS X

Programming language: Perl

Other requirements: The Perl modules: Algorithm::Numerical::Shuffle, POSIX, Statistics::Distributions, Storable, and Tie::RefHash

License: GNU general public license

Any restrictions to use by non-academics: none

Authors' contributions

YL developed and tested this software under the supervision of MR. Both authors wrote the manuscript.

Acknowledgments

We thank Amir Ben-Dor and Zohar Yakhini for valuable discussions. This work was in part supported by the National Research School in Genomics and Bioinformatics and the Knut and Alice Wallenberg Foundation through the Swegene consortium.

References

1. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ED: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286:**531-537.
2. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi OP, Wilfond B, Borg Å, Trent J: **Gene-expression profiles in hereditary breast cancer.** *N Engl J Med* 2001, **344:**539-548.
3. Ringnér M, Peterson C, Khan J: **Analyzing array data using supervised methods.** *Pharmacogenomics* 2002, **3:**403-415.
4. 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, Hayward N, Trent J: **Molecular classification of cutaneous malignant melanoma by gene expression profiling.** *Nature* 2000, **406:**536-540.
5. Sokal RR, Michener CD: **A statistical method for evaluating systematic relationships.** *Univ Kans Sci Bull* 1958, **38:**1409-1438.
6. Hartigan JA, Wong MA: **A K-means clustering algorithm.** *Applied Statistics* 1979, **28:**100-108.
7. Kohonen T: *Self-Organizing Maps* 3rd edition. Berlin: Springer; 2001.
8. Quackenbush J: **Computational analysis of microarray data.** *Nat Rev Genet* 2001, **2:**418-427.
9. 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, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.** *Nature* 2000, **403:**503-511.
10. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO, Botstein D: **Molecular portraits of human breast tumours.** *Nature* 2000, **406:**747-752.
11. Dugas M, Merk S, Breit S, Dirschedl P: **Mdclust – Exploratory microarray analysis by multidimensional clustering.** *Bioinformatics* 2004, **20:**931-936.
12. Covell DG, Wallqvist A, Rabow AA, Thanki N: **Molecular classification of cancer: unsupervised self-organizing map analysis of gene expression microarray data.** *Mol Cancer Ther* 2003, **2:**317-332.
13. Hsu AL, Tang SL, Halgamuge SK: **An unsupervised hierarchical dynamic self-organizing approach to cancer class discovery and marker gene identification in microarray data.** *Bioinformatics* 2003, **19:**2131-2140.
14. Ben-Dor A, Friedman N, Yakhini Z: **Class discovery in gene expression data.** In *Proceedings of the Fifth Annual Conference on Computational Biology (RECOMB): 2001; Montreal* Edited by: Lengauer T, Sankoff D, Istrail S, Pevzner P, Waterman M. ACM Press; 2001:31-38.
15. von Heydebreck A, Huber W, Poustka A, Vingron M: **Identifying splits with clear separation: a new class discovery method for gene expression data.** *Bioinformatics* 2001, **17(Suppl 1):**S107-114.
16. Hedenfalk I, Ringnér M, Ben-Dor A, Yakhini Z, Chen Y, Chebil G, Ach R, Loman N, Olsson H, Meltzer P, Borg Å, Trent J: **Molecular classification of familial non-BRCA1/BRCA2 breast cancer.** *Proc Natl Acad Sci USA* 2003, **100:**2532-2537.
17. Kirkpatrick S, Gelatt C, Vecchi M: **Optimization by simulated annealing.** *Science* 1983, **220:**671-680.
18. Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z: **Tissue classification with gene expression profiles.** *J Comput Biol* 2000, **7:**559-583.
19. Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB: **Nonparametric methods for identifying differentially expressed genes in microarray data.** *Bioinformatics* 2002, **18:**1454-1461.
20. Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Atonescu CR, Peterson C, Meltzer PS: **Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.** *Nat Med* 2001, **7:**673-679.
21. Wilcoxon F: **Individual comparisons by ranking methods.** *Biometrics* 1945, **1:**80-83.
22. Walpole RE, Myers RH: *Probability and Statistics for Engineers and Scientists* 3rd edition. New York: Macmillan; 1985.
23. Kruskal WH, Wallis WA: **Use of ranks in one-criterion variance analysis.** *J Amer Statist Assoc* 1952, **47:**583-621.
24. Press WH, Teukolsky SA, Vetterling WT, Flannery BP: *Numerical Recipes in C* 2nd edition. Cambridge, UK: Cambridge University Press; 1992.
25. Ringnér M, Edén , Johansson P: **Classification of expression patterns using artificial neural networks.** In *A Practical Approach to Microarray Data Analysis* Edited by: Berrar DP, Dubitzky W, Granzow M. Boston: Kluwer Academic Publishers; 2002:201-215.
26. Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg Å, Peterson C: **Bioarray software environment: a platform for comprehensive management and analysis of microarray data.** *Genome Biol* 2002, **3:**software0003.1-0003.6.

Paper III

Physiol Genomics 21: 396–403, 2005.
First published March 15, 2005; doi:10.1152/physiolgenomics.00303.2004.

Gene expression profiling demonstrates that TGF-β1 signals exclusively through receptor complexes involving Alk5 and identifies targets of TGF-β signaling

Göran Karlsson,[1,2] Yingchun Liu,[2,3] Jonas Larsson,[1,2] Marie-José Goumans,[4] Ju-Seog Lee,[5] Snorri S. Thorgeirsson,[5] Markus Ringnér,[2,3] and Stefan Karlsson[1,2]

[1]*Department of Molecular Medicine and Gene Therapy, Institute of Laboratory Medicine, and* [2]*The Lund Strategic Research Center for Stem Cell Biology and Cell Therapy, Lund University Hospital, Lund, Sweden;* [3]*Complex Systems Division, Department of Theoretical Physics, Lund University, Lund, Sweden;* [4]*Department of Cardiology, Heart Lung Center, Utrecht, the Netherlands; and* [5]*Laboratory of Experimental Carcinogenesis, National Cancer Institute, Bethesda, Maryland*

Submitted 20 December 2004; accepted in final form 9 March 2005

Karlsson, Göran, Yingchun Liu, Jonas Larsson, Marie-José Goumans, Ju-Seog Lee, Snorri S. Thorgeirsson, Markus Ringnér, and Stefan Karlsson. Gene expression profiling demonstrates that TGF-β1 signals exclusively through receptor complexes involving Alk5 and identifies targets of TGF-β signaling. *Physiol Genomics* 21: 396–403, 2005. First published March 15, 2005; doi:10.1152/physiolgenomics.00303.2004.—Transforming growth factor-β1 (TGF-β) regulates cellular functions like proliferation, differentiation, and apoptosis. On the cell surface, TGF-β binds to receptor complexes consisting of TGF-β receptor type II (TβRII) and activin-like kinase receptor-5 (Alk5), and the downstream signaling is transduced by Smad and MAPK proteins. Recent data have shown that alternative receptor combinations aside from the classical pairing of TβRII/Alk5 can be relevant for TGF-β signaling. We have screened for alternative receptors for TGF-β and also for gene targets of TGF-β signaling, by performing functional assays and microarray analysis in murine embryonic fibroblast (MEF) cell lines lacking Alk5. Data from TGF-β-stimulated $Alk5^{-/-}$ cells show them to be completely unaffected by TGF-β. Additionally, 465 downstream targets of Alk5 signaling were identified when comparing $Alk5^{-/-}$ or TGF-β-stimulated $Alk5^{+/+}$ MEFs with unstimulated $Alk5^{+/+}$ cells. Our results demonstrate that, in MEFs, TGF-β signals exclusively through complexes involving Alk5, and give insight to its downstream effector genes.

signal transduction; Smad; microarrays

TRANSFORMING GROWTH FACTOR-β1 (TGF-β) is ubiquitously expressed during embryonic development and in most adult tissues, regulating proliferation, differentiation, and apoptosis (18). Although much is known about the mechanisms of TGF-β signaling, many target genes remain elusive, and the responses to TGF-β are highly context dependent, varying with cell type and physiological state (15). In recent years, a large number of studies have attempted to elucidate the mechanisms of TGF-β function via knockout mouse models either for TGF-β receptors, the ligands, or downstream signaling proteins. As expected, these mice have various abnormalities in development, regulation of immune response, vessel formation, wound healing, and hematopoiesis (2).

Because of its importance in several different contexts, thorough investigations have led to considerable knowledge about the mechanisms of TGF-β signal transduction (13). TGF-β belongs to a superfamily of cytokines that also includes activins/inhibins, bone morphogenetic proteins (BMPs), growth differentiation factors, and a few other related proteins. The TGF-β family receptors consist of type I and type II transmembrane serine/threonine kinase receptors. On the cell surface, TGF-β binds to the constitutively active TβRII. Upon binding of the ligand, two TβRIIs and two TGF-β type I receptors (Alk5s) are brought together, forming a heterotetrameric complex. The kinase domain of Alk5 is phosphorylated by TβRII, and subsequently Alk5 phosphorylates and activates the intracellular receptor-activated (R)-Smads (Smad2 and -3), which heteroligomerize with the co-mediator, Smad4. This complex translocates into the nucleus, where it recruits transcriptional co-activators and co-repressors to control gene expression. In addition, it has been reported that phosphorylation of Alk5 initiates other pathways of signaling, including the mitogen-activated protein kinase (MAPK) family pathways, comprised of the extracellular signal-regulated kinase (ERK) pathway, c-Jun NH_2-terminal kinase (JNK) pathway, and the p38 pathway. The only known function of TβRII is the activation of Alk5, and signaling from solitary TβRIIs has not been reported.

It was earlier thought that Alk5 was the only type I receptor conducting TGF-β signaling. However, there is evidence that TGF-β can bind to Alk1 in endothelial cells and that the subsequent signal transduction through Smad1/5/8 inhibits Alk5 signaling (5, 16). Additionally, in epithelial cells Alk2 is involved in TGF-β-induced epithelial-to-mesenchymal transdifferentiation (9). In this study, we have investigated the possibility of alternative receptors for TGF-β signaling in fibroblasts, by stimulating Alk5-deficient murine embryonic fibroblasts (MEFs) with TGF-β. We conclusively demonstrate that Alk5 is necessary for TGF-β signal transduction in these cells.

Additionally, we present transcriptional profiles of MEF cell lines that are either deficient in Alk5 signaling, have normal signaling, or are stimulated with TGF-β and identify genes that are differentially expressed as a response to TGF-β signaling. Our results create a data set of 465 targets of TGF-β/Alk5 signaling, in which earlier unknown categories of gene targets are revealed.

METHODS

Generation and culture of MEF cell lines. Cre-loxP gene targeting of the *Alk5* exon 3 and isolation of murine embryonic fibroblasts were performed as described (10). Cells were cultured in Dulbecco's modified Eagles medium (DMEM; GIBCO, Stockholm, Sweden) supplemented with 10% fetal bovine serum (FBS; GIBCO) and 1%

Article published online before print. See web site for date of publication (http://physiolgenomics.physiology.org).

Address for reprint requests and other correspondence: M. Ringnér, Complex Systems Division, Dept. of Theoretical Physics, Lund Univ., Sölvegatan 14A, 223 62 Lund, Sweden (e-mail: markus@thep.lu.se).

Physiological Genomics

penicillin-streptomycin (GIBCO). Four different cell lines were generated, two that were *Alk5*$^{-/-}$ [knockout (KO)1 and KO2] and two that were *Alk5*$^{+/+}$ [wild type (WT)1 and WT2]. In the case of TGF-β stimulation experiments, huTGF-β1 (R&D systems, Abingdon, UK) was added in a concentration of 10 ng/ml at 1, 5, 8, or 16 h before cell confluency. Cells were trypsinized and harvested when confluent.

In the cell expansion experiment, 2×10^5 cells from the different cell lines were plated in 60-mm plates and cultured with or without the presence of TGF-β1 (10 ng/ml). Cells were counted and medium was changed at 24, 48, and 72 h postseeding. Three independent experiments were performed.

Western blot analysis. Western blot analysis was performed as described (10).

Transcriptional reporter assays. Assays were performed as described (10).

RNA isolation and microarray hybridization. RNA from harvested MEFs, stored as a dry pellet at −80°C, was isolated using TRIzol reagent (Invitrogen, Stockholm, Sweden) according to manufacturer's protocol, and reverse transcription and hybridization to microarray slides were performed as described (14). Microarray glass slides containing >37,000 mouse cDNA clones (mouse, 36K; National Institute of Mental Health, Bethesda, MD) were used. Slides were scanned using an Axon 4000 scanner (Axon Instruments, Leusden, the Netherlands) and analyzed with GenePix pro v.4.0 software (Axon Instruments).

Ten microarrays were performed to identify downstream targets of TGF-β signaling. These were made using one RNA extraction from the KO1 MEF line and two extractions from the KO2 line (KO2 and KO2_2), with one of the *Alk5*$^{+/+}$ cell lines (WT1) as reference, or the two *Alk5*$^{+/+}$ cell lines (WT1 and WT2) stimulated with TGF-β for 1, 5, and 16 h preharvest, again with the WT1 cell line as reference. *Alk5*$^{-/-}$ and TGF-β-stimulated samples were labeled with Cy5 and reference sample with Cy3 dye in all experiments, except for one dye-swap experiment (KO1_DS) performed to exclude technical errors. The KO2_2 extraction functioned as a technical repeat and was performed 6 mo after the KO2 experiment. The subsequent microarray analysis was carried out using a completely new batch of reagents.

In addition, three arrays were performed to investigate the possibility of alternative receptors for TGF-β signaling. They include experiments done using cDNA from the two *Alk5*$^{-/-}$ cell lines stimulated with TGF-β for 8 h, with cDNA from unstimulated cells of the same *Alk5*$^{-/-}$ cell line as reference, and one experiment using cDNA from RNA extractions of two independently harvested WT1 cultures.

Microarray data, including information about the mouse cDNA clones, are available at the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) website (**http://www.ncbi.nlm.nih.gov/geo/**) with GEO accession no. GSE1742.

Analyses of microarray data. The data extracted with the use of GenePix were stored in BioArray Software Environment (BASE; Ref. 19), and BASE was used for quality control and normalization. Briefly, on the basis of the parameters extracted by the GenePix software, our quality filter required each spot to have a median-based signal-to-noise ratio of at least 3 for each channel and at least 50 foreground pixels. Spots that did not fulfill these requirements were treated as missing values in the subsequent analysis. In addition, for the 10 arrays used to identify targets of TGF-β signaling, spots with more than two missing values across arrays were excluded from further analysis. Each experiment was subsequently normalized using LOWESS regression as implemented in BASE.

Next, we looked for clones showing a large variation in expression across arrays, and selected spots for which the standard deviation of log (base 2) expression ratios was >0.5. To identify clones differentially expressed between TGF-β-stimulated *Alk5*$^{+/+}$ and *Alk5*$^{-/-}$ cells, we employed the rank-based Wilcoxon test as implemented in the MCD software (12). To investigate whether differentially expressed clones were significantly associated with specific Gene Ontology (GO) terms (6), we used the GoMiner software (24). GoMiner accepts gene symbols as input, and we used ACID (17) to map our clones to gene symbols (based on UniGene mouse build 141). In the GoMiner analysis, differentially expressed clones were compared with all clones that passed the quality filter. GO terms with *P*-under ≤ 0.05 and at least five downregulated clones or *P*-over ≤ 0.05 and at least five upregulated clones were selected for discussion. To investigate genes in common between our data and previous studies, public data sets were mapped to gene symbols according to UniGene mouse build 141, using ACID.

Quantitative RT-PCR. Quantitative (q)RT-PCR (TaqMan; Applied Biosystems, Stockholm, Sweden) was performed in an ABI Prism 7700 (Applied Biosystems), according to the manufacturer's protocol, on cDNA from *Alk5*$^{-/-}$, *Alk5*$^{+/+}$, 5-h TGF-β-stimulated *Alk5*$^{+/+}$, and TGF-β-stimulated *Alk5*$^{-/-}$ MEFs. RNA was reverse transcribed (SuperScript II; Invitrogen, Stockholm, Sweden), according to manufacturer's protocol, in the presence of random hexamers, and gene-specific primers (Applied Biosystems) were used to analyze the expression of nine differentially expressed genes from the microarrays [inhibitor of DNA binding (*Id*)-1, *Id2*, *Id3*, growth arrest and DNA damage-inducible-45γ *(Gadd45g)*, eukaryotic translation initiation factor-5A *(Eif5α)*, myelocytomatosis oncogene *(c-myc)*, growth arrest specific *(Gas5)*, activated leukocyte cell adhesion molecule *(Alcam)*, and decorin *(Dcn)*] together with the housekeeping gene hypoxanthine guanine phosphoribosyl transferase (*Hprt)*. Each assay was performed in triplicate, the results were normalized to *Hprt* levels, and relative gene expressions between *Alk5*$^{-/-}$ or TGF-β1-stimulated *Alk5*$^{+/+}$ and unstimulated *Alk5*$^{+/+}$ cells were examined. Additionally, qRT-PCR analysis was performed on all MEF lines, using primers for *Alk1–6* (Applied Biosystems), and on murine embryonic endothelial cells, using primers for *Alk1*. Experiments were performed in triplicate and normalized to *Hprt* levels.

RESULTS

Alk5 is essential for TGF-β signaling in MEFs. To investigate whether Alk5 is crucial for TGF-β signal transduction, we measured expression of other members of the TGF-β receptor family, the most likely candidates to be alternative receptors for TGF-β signaling. Screening for these, using qRT-PCR analysis, revealed that endogenous *Alk1* and *Alk6* expression was absent in MEFs (Fig. 1*A*). The other *Alk*s were highly expressed compared with housekeeping gene expression, with *Alk2* and *-3* having much higher expression than *Alk4* and *-5* in these cells. *Alk5*$^{-/-}$ MEFs did not exhibit any increase in expression of mRNA for Alk receptors, indicating that compensatory upregulation of other TGF-β family receptors (e.g., the BMP family) does not occur as a consequence of Alk5 deficiency. Because qRT-PCR showed the presence of *Alk5* mRNA in *Alk5*$^{-/-}$ MEFs, we performed Western blot analysis and transcriptional activity assays to exclude a functional Alk5 protein. Western analysis clearly illustrated that Smad2 was phosphorylated in the *Alk5*$^{+/+}$ MEFs upon TGF-β stimulation, whereas phosphorylation was absent in *Alk5*$^{-/-}$ cells (Fig. 1*B*). Thus, although *Alk5* mRNA is stable in the *Alk5*$^{-/-}$ MEFs, the deletion of exon 3 makes it unable to produce a functional protein. Additionally, this analysis shows that Smad2 phosphorylation by TGF-β is conducted exclusively through Alk5 in MEFs.

Furthermore, we investigated the transcriptional activity of Smads using a reporter gene approach where Smad-binding elements (SBEs) were coupled to the luciferase gene and transfected into the different MEF cell lines. Figure 1*C* shows that TGF-β induced luciferase activity in WT MEFs trans-

fected with the Smad3/Smad4-binding elements (CAGA) and the Smad4-binding SBE elements coupled to the luciferase reporter gene. Only background activity was detected when the cells, transfected with the Smad1/Smad4-inducible BMP-responsive element (BRE) reporter, were stimulated with TGF-β. No changes in luciferase activities were observed when stimulating *Alk5*$^{-/-}$ cells with TGF-β. The transcriptional assay demonstrates a complete absence of TGF-β-induced Smad signaling as a result of Alk5 deficiency and that all TGF-β Smad signals are transduced through the Smad2/3/4 branch in MEFs. Thus signaling from alternative receptor complexes in the TGF-β family, like the ones seen in endothelial and epithelial cells, cannot be detected in this context.

It has been shown previously that TGF-β has an inhibitory effect on cell proliferation through blockage in the late phase of G1 (8). To examine whether TGF-β-induced proliferation inhibition was altered in our MEF cell lines, we performed a cell expansion experiment. Each cell line was cultured in cell culture medium and serum with or without the addition of TGF-β. The *Alk5*$^{-/-}$ cells proliferated equally well, irrespective of the presence of TGF-β, whereas the cells from the *Alk5*$^{+/+}$ animals showed an ~2.5-fold decrease in expansion after 72 h of stimulation (Fig. 1*D*). These results demonstrate the potent inhibitory effect of TGF-β on cell proliferation and how this effect is dependent on Alk5.

Furthermore, we studied how the absence of Alk5 affected the transcriptional response of the fibroblasts. We examined the expression of the inhibitor of the DNA binding/differentiation family of genes, known to be transcriptionally repressed by TGF-β signaling (11, 20). qRT-PCR reveals that these genes were not affected by TGF-β in *Alk5*$^{-/-}$ MEFs, whereas in *Alk5*$^{+/+}$ cells *Id* expression was strongly downregulated after TGF-β stimulation (Fig. 2*A*).

To exclude receptors outside the TGF-β superfamily and signaling pathways activated through TβRIIs alone, we performed global gene expression profiling. We speculated that microarray analysis on *Alk5*$^{-/-}$ cells stimulated with TGF-β would unravel possible alternative pathways for TGF-β signaling. However, very few (<0.05%) of the >37,000 spots on the microarrays had a more than twofold differential expression in the two Alk5$^{-/-}$ cell lines used in the experiment, none of which overlapped (Fig. 2*B*). The number of differentially expressed clones in these experiments was similar to when hybridizing two separate RNA extractions from the same cell line (WT1 vs. WT1), where any differential expression is due to noise. In contrast, stimulation with epidermal growth factor (EGF) generated transcriptional repression of *Id1* and induction of *c-myc* in the *Alk5*$^{-/-}$ cells and the *Alk5*$^{+/+}$ cells alike (qRT-PCR, data not shown). This serves as a positive

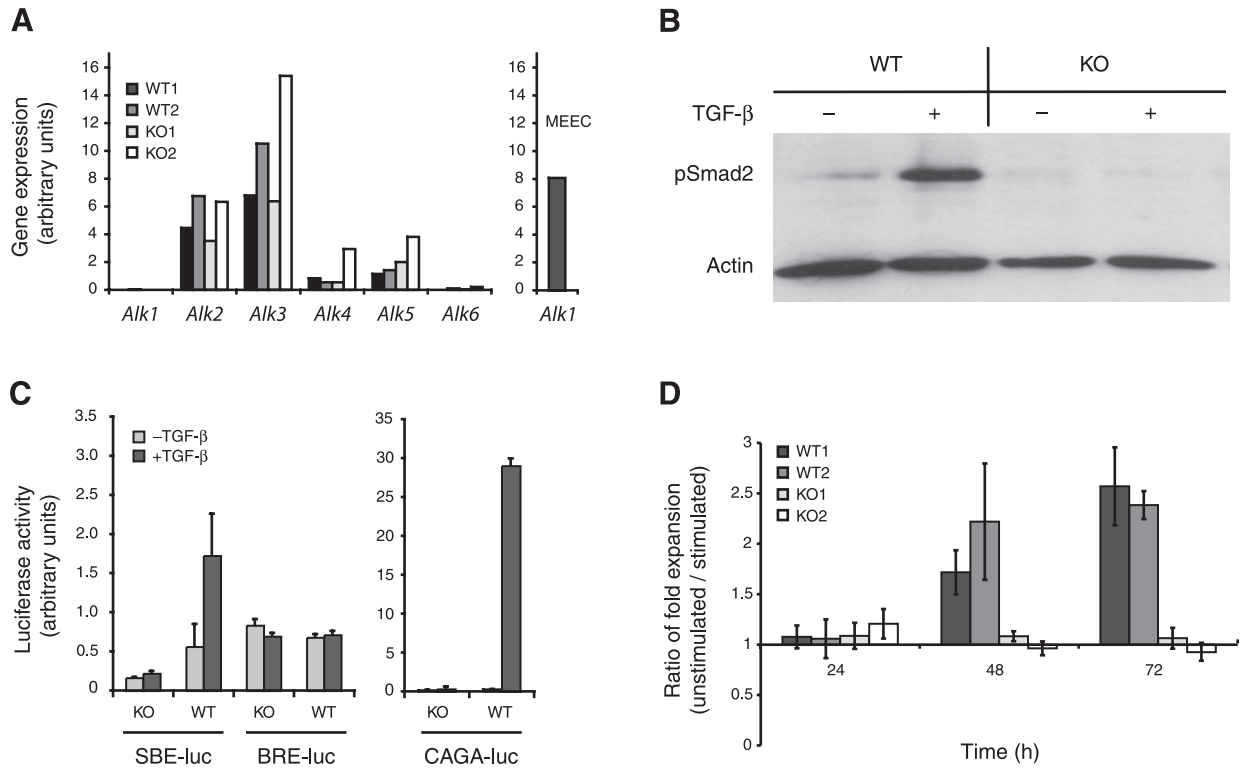


Fig. 1. Alk5 is essential for transforming growth factor (TGF)-β induced Smad signaling. *A*: quantitative (q)RT-PCR screening for *Alk* expression in murine embryonic fibroblasts (MEFs) reveals the absence of *Alk1* and *Alk6* expression. Alk5 deficiency does not alter the expression profile of TGF-β family receptors. Murine embryonic endothelial cells (MEECs) were used as a positive control for *Alk1* expression. Scale is logarithmic, where one unit represents a doubling in expression. *B*: Smad2 is phosphorylated in *Alk5*$^{+/+}$ MEFs upon stimulation, whereas Smad2 phosphorylation is abrogated in *Alk5*$^{-/-}$ cells. *C*: TGF-β induces luciferase activity in *Alk5*$^{+/+}$ cells transfected with the reporter gene coupled to the Smad3/4-binding elements (CAGA) and Smad4-binding elements (SBE). This response is not detected in the Alk5-deficient cells. Low luciferase activity is detected in cells transfected with the Smad1/4-inducible BRE coupled to the reporter gene, and this activity is not changed when cells are stimulated with TGF-β. *D*: unstimulated *Alk5*$^{+/+}$ MEFs proliferate more rapidly compared with cells stimulated with TGF-β, reaching an unstimulated-to-stimulated ratio of expansion of 2–2.5-fold in 72 h. Proliferation of *Alk5*$^{-/-}$ MEFs is not affected by TGF-β stimulation. Error bars represent SE of the mean. pSmad2, phosphorylated Smad2; WT, wild type; KO, knockout.

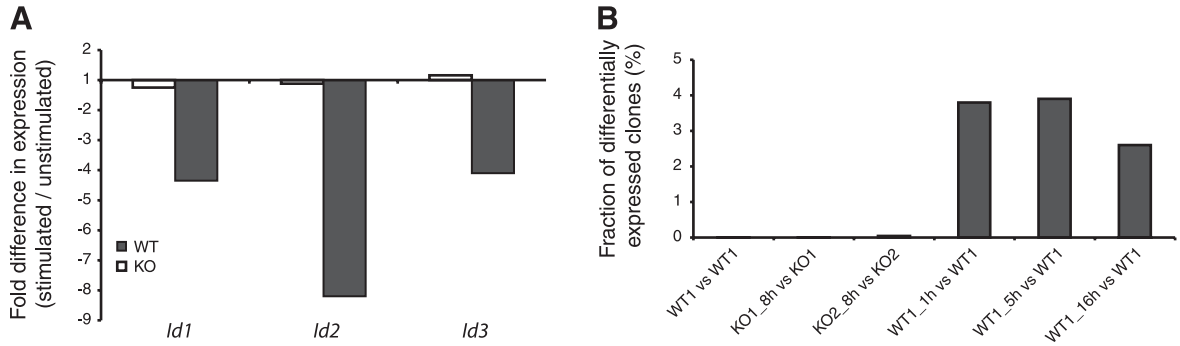


Fig. 2. TGF-β does not affect gene expression in the absence of Alk5. *A*: the expression of three known gene targets of TGF-β signaling (*Id1*, *Id2* and *Id3*) is decreased 4–8.5-fold in response to TGF-β stimulation of $Alk5^{+/+}$ (WT1) MEFs as detected by qRT-PCR. *Id* expression is not altered in stimulated $Alk5^{-/-}$ (KO1) cells. *B*: microarray analysis of $Alk5^{-/-}$ MEFs stimulated with TGF-β for 8 h compared with unstimulated $Alk5^{-/-}$ cells provides a very low fraction of differentially expressed clones (<0.05% with >2-fold differential expression). This fraction is of the same magnitude as the fraction of differentially expressed clones in the self-self control experiment (WT1 vs. WT1). Similar experiments on $Alk5^{+/+}$ MEFs (TGF-β stimulated for 1, 5, and 16 h) demonstrate that TGF-β stimulation alters the expression of up to 4% of the genes in the genome.

control showing that a gene that is a target of TGF-β signaling can be activated in the $Alk5^{-/-}$ cells by another ligand.

Microarray experiments performed on TGF-β-stimulated $Alk5^{+/+}$ cell lines resulted in differential expression of between 2.6 and 3.9% of the clones printed, demonstrating the vast effect of TGF-β on transcriptional regulation in the genome.

Global gene expression analysis reveals 465 targets of Alk5 signaling. The substantial effects of TGF-β made us conduct more extensive global gene expression studies to get a database of gene targets of Alk5 signaling. Total RNA was extracted from unstimulated $Alk5^{-/-}$ and $Alk5^{+/+}$ MEFs stimulated with TGF-β for three different time periods (1, 5, and 16 h). cDNA from each sample was hybridized on the microarray slides together with cDNA from an unstimulated $Alk5^{+/+}$ common reference sample. We speculated that this design would result in four different data sets: early, intermediate, and late differentially expressed clones plus all the clones differentially expressed in the Alk5-deficient fibroblasts. A dye swap control experiment using one of the $Alk5^{-/-}$ lines (KO1) showed that a replicate of one experiment using the same RNA but different dyes for the sample and reference produced highly similar results (R^2 = 0.96 and 0.97 for the KO1 and the reference, respectively; data not shown), thus demonstrating the reliability of our protocols. For the 10 arrays used in this investigation, 15,202 clones passed the quality filtering. Next, we performed hierarchical cluster analysis on all the clones (2,223) that had a standard deviation in expression values over the data set >0.5. This unsupervised analysis divided the microarray experiments in two clusters: one that included all the $Alk5^{-/-}$ hybridizations and another containing all the stimulated $Alk5^{+/+}$ experiments (Fig. 3*A*). This was good evidence that the differences in gene expression among the samples were primarily due to the presence or absence of Alk5 and not because of other biological or technical variances. Because the 1-h TGF-β-stimulated $Alk5^{+/+}$ cell lines seemed to be highly related, whereas the similarities in gene expression patterns in the 5- and 16-h stimulations were more dependent on the cell line, we considered our data to contain differentially expressed clones representing early responsive genes ($Alk5^{+/+}$ MEFs stimulated for 1 h), late responsive genes ($Alk5^{+/+}$ MEFs stimulated for 5 and 16 h), and genes differentially expressed in fibroblasts deficient in TGF-β signaling. Within this set of 2,223 clones with variation among the samples, we isolated the ones that were differentially expressed due to TGF-β stimulation or Alk5 signaling deficiency and hence likely to be the targets of Alk5 signaling. By dividing the samples into the two groups ($ALK5^{-/-}$ and TGF-β-stimulated $Alk5^{+/+}$ samples) defined by the unsupervised hierarchical clustering, and performing a ranked-based Wilcoxon test on the varying clones, we were able to identify 465 clones (of which 445 could be mapped to a UniGene cluster, representing 369 unique genes) that were significantly ($P \leq 0.01$) differentially expressed as a response to TGF-β stimulation or Alk5 signaling deficiency (Supplemental Fig. S1; available at the *Physiological Genomics* web site).[1] We expected 22 clones by chance to be differentially expressed using this P value for the 2,223 clones. Hierarchical clustering of these clones divided them in two groups, either upregulated as a response to TGF-β stimulation and downregulated due to Alk5 deficiency, or the opposite (Fig. 3*B*). To get an overview of the 465 clones, they were categorized according to the biological or molecular function assigned to them in the GO database (Fig. 3*C*). An interesting observation was the finding of a large number of clones in the "protein folding/chaperone activity" category.

With the GoMiner software, one can investigate whether certain GO terms are enriched among the differentially expressed clones compared with their abundance on the microarray. GoMiner confirmed "protein folding" and "chaperone activity" as two of the most enriched categories, with 5.5 and 5.0 times more clones represented among the upregulated clones than one would expect if picking clones randomly from the microarray (Table 1).

Furthermore, we compared the 465 clones identified as downstream targets of TGF-β signaling in our study, with 360 clones identified as targets of TGF-β signaling in a similar study performed on Smad2-, Smad3-, and ERK-deficient MEFs by Yang et al. (23). Among the 360 targets presented in the previous study, 212 were found among our 15,202 clones

[1]The Supplemental Material for this article (Supplemental Fig. S1) is available online at **http://physiolgenomics.physiology.org/cgi/content/full/00303.2004/DC1**.

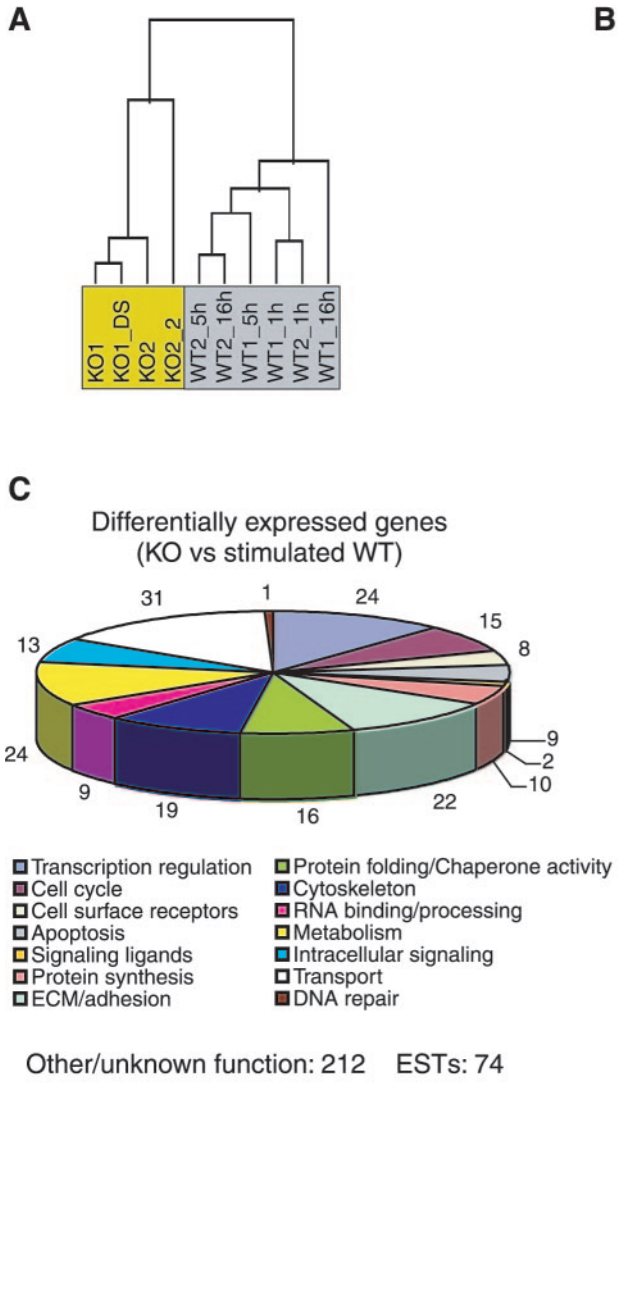


Fig. 3. Global gene expression profiling reveals 465 gene targets for TGF-β signaling. *A*: unsupervised hierarchical clustering of all clones with a standard deviation >0.5 in expression levels divides the data set in two clusters, one containing all the experiments made on *Alk5*$^{-/-}$ MEFs and another containing the ones made on *Alk5*$^{+/+}$ cells stimulated with TGF-β (1, 5, and 16 h), as shown by the resulting dendrogram. *B*: hierarchical clustering of the 465 targets of TGF-β signaling. *C*: 179 of the 465 clones could be assigned to categories according to classification in the Gene Ontology database; 74 were expressed sequence tags, and 212 were either unknown or belonged to a category not presented in the diagram. DS, dye swap; red, upregulated; green, downregulated; KO2_2, repeated RNA extraction from KO2 cell line.

that passed quality filtering. Twenty-nine of these 212 were also present in our list of 465 targets. These 29 clones correspond to 27 unique genes (Table 2).

To validate the results obtained from the microarray experiments, we performed qRT-PCR, measuring gene expression of nine genes with varying magnitude of differential expression. Relative mRNA levels between *Alk5*$^{-/-}$ cells as well as *Alk5*$^{+/+}$ cells stimulated with TGF-β and unstimulated *Alk5*$^{+/+}$ cells were compared (Table 3). The resulting data revealed that in 77% (27 of 35) of the assays performed, an upregulated gene in the microarray experiment was also upregulated when using the qRT-PCR method and vice versa. The validation of the microarray was even stronger when only looking at genes in samples with a differential expression of more than or equal to twofold, where 100% (18/18) of the assays showed the same direction of differential expression in both methods used.

DISCUSSION

The findings of Alk1 as an alternative receptor for TGF-β in endothelial cells (16) raise the question of alternative receptors in other cell types. We reasoned that a model with a complete absence of the classical Alk5 receptor would be ideal to elucidate this question. We have conducted extensive studies, including TGF-β family receptor screening, Smad signaling analysis, and transcriptional and functional assays as well as global gene expression profiling on Alk5-deficient MEFs without finding any signs of alternative receptors for TGF-β signaling in these cells.

Table 1. *GO categories enriched by genes regulated by TGF-β signaling*

GO Category	Total	Up	*P* Value
Physiological process	3,373	74	0.0420
Metabolism	2,379	58	0.0113
Biosynthesis	477	15	0.0424
Protein biosynthesis	273	11	0.0167
Translation	92	6	0.0088
Protein metabolism	995	33	0.0008
Protein folding	65	7	0.0002
Cell motility	78	6	0.0039
Nucleocytoplasmic transport	55	5	0.0041
Nuclear organization and biogenesis	86	5	0.0258
Translation regulator activity	71	9	0.0000
Translation factor activity, nucleic acid binding	68	9	0.0000
Translation initiation factor activity	46	8	0.0000
Binding	2,651	66	0.0020
Nucleotide binding	660	22	0.0073
Purine nucleotide binding	651	22	0.0062
Adenyl nucleotide binding	529	18	0.0128
ATP binding	524	18	0.0116
Chaperone activity	93	9	0.0001
Nuclear membrane	51	5	0.0030
Cell junction	55	5	0.0041
Enzyme inhibitor activity	56	5	0.0045
Nucleolus	56	5	0.0045

GO Category	Total	Down	*P* Value
Vacuole	61	7	0.0032
Lytic vacuole	54	7	0.0016
Lysosome	54	7	0.0016
Extracellular	669	35	0.0021
Extracellular space	604	32	0.0028
Extracellular matrix	113	9	0.0103
Protein binding	989	46	0.0043
Peptidase activity	190	11	0.0425

For Gene Ontology (GO) categories, subcategories are indented below parent categories. Total, total no. of filtered genes matched to category; Up, no. of genes upregulated in transforming growth factor (TGF)-β-stimulated wild types (WTs) matched to category; Down, no. of genes downregulated in TGF-β-stimulated WTs matched to category; *P* value, significance of enrichment of up/downregulated genes as calculated by GOMiner.

Table 2. *Gene targets of TGF-β signaling in common with Yang et al. (23)*

Gene Name	Gene Symbol
RIKEN cDNA 2610312E17 gene	2610312E17
Branched chain aminotransferase 1, cytosolic	Bcat1
Basic helix-loop-helix domain containing, class B2	Bhlhb2
Bcl2-associated athanogene 2	Bag2
Filamin, beta	Flnb
Connective tissue growth factor	Ctgf
RIKEN cDNA B230104P22 gene	B230104P22
Myelocytomatosis oncogene	Myc
RIKEN cDNA 1110017C15 gene	1110017C15
Peroxisome proliferative activated receptor, gamma, coactivator-related 1	Pprc1
Nucleolar and coiled-body phosphoprotein 1	Nolc1
RIKEN cDNA 2210403K04 gene	2210403K04
Synaptic nuclear envelope 2	Syne2
Zinc finger protein 395	Zfp395
High mobility group box transcription factor 1	Hbp1
Sestrin 1	Sesn1
Platelet derived growth factor receptor, alpha polypeptide	Pdgfra
F11 receptor	F11r
Melanoma antigen, family D, 1	Maged1
Zinc finger protein 36, C3H type-like 1	Zfp3611
Laminin, alpha 5	Lama5
RIKEN cDNA 1700065A05 gene	1700065A05
Protein tyrosine phosphatase, receptor type, F	Ptprf
Cyclin-dependent kinase inhibitor 2C (p18, inhibits CDK4)	Cdkn2c
RIKEN cDNA 1110065D03 gene	1110065D03
RIKEN cDNA 5830403L16 gene	Glu1
Shroom	Shrm

Global gene expression profiling only detects transcriptional changes, and therefore posttranslational events could be overseen. For example, recently it has been shown that the TGF-β superfamily type II receptor BMPR-II alone can interact with LIM kinase-1 (LIMK1), and that LIMK1 is released on BMP4 binding to BMPR-II (3). Unbound LIMK1 will then regulate actin dynamics through phosphorylation of cofilin. These kinds of posttranslational events might not be detected by gene expression analysis, although this is unlikely, since such changes in the cell are likely to affect global gene expression profiles to some degree. Thus, taken together, our data strongly suggest that TGF-β signals exclusively through receptor complexes containing Alk5 in fibroblasts.

One microarray study on murine fibroblasts has previously been published (23) where the authors investigated the transcriptome of MEFs lacking Smad2, Smad3, or ERK signaling while having other TGF-β signaling pathways intact. Here we have the advantage of *Alk5*$^{-/-}$ cells that completely lack responses to TGF-β signaling and thus should reveal the majority of its gene targets. While in our study we identified 465 clones as targets of TGF-β signaling, Yang et al. presented 360, and among these two data sets, 29 targets were overlapping. The 27 unique genes that these 29 clones represent include both known targets of TGF-β signaling, like *c-myc* and *Pdgfra*, and some novel targets with interesting functions. In our study, *Pdgfra* is considerably downregulated in all the TGF-β-stimulated WT MEFs and upregulated in three of four *Alk5*$^{-/-}$ experiments. *Pdgfra* is a possible indirect key mediator of the TGF-β-induced proliferation block, since it stimulates proliferation of fibroblasts and has been shown to be

Table 3. *qRT-PCR confirmation of microarray data*

Gene	Method	Fold Differentially Expressed KO1	KO2	WT1_5h	WT2_5h
c-myc	qRT-PCR	−3.1	−3.4	1.1	1.4
	microarray	−1.9	−2.9	1.8	1.6
Gas5	qRT-PCR	4.2	2.7	3.1	12.5
	microarray	−1.2	−1.3	1.9	1.6
elF5a	qRT-PCR	−1.1	−1.4	1.1	1.0
	microarray	−1.1	−1.8	2.0	1.8
Id1	qRT-PCR	2.7	1.5	−4.3	−3.0
	microarray	3.0		−6.3	−1.9
Id2	qRT-PCR	−1.7	1.8	−8.2	−1.4
	microarray	1.4	3.4	−4.9	−3.4
Id3	qRT-PCR	1.5	3.5	−4.1	1.5
	microarray	3.2	2.6	−4.6	−1.5
Gadd45g	qRT-PCR	1.3	−4.3	2.6	3.3
	microarray	−1.1	1.3	3.6	5.2
Alcam	qRT-PCR	4.3	55.2	−1.1	4.8
	microarray	8.2	8.2	−4.5	−1.6
Dcn	qRT-PCR	9.3	15.3	−3.6	2.4
	microarray	7.6	9.3	−10.0	−1.4

KO = ALK5$^{-/-}$. WT = ALK5$^{+/+}$. _5h indicates 5-h stimulation with TGF-β. qRT-PCR, quantitative RT-PCR.

downregulated by overexpression of TβRII in NIH-3T3 fibroblasts (4).

Another gene overlapping with the study by Yang et al. (23) is Bcl-2-associated athanogene-2 (*Bag2*), previously unknown as a target of TGF-β signaling. *Bag2* is involved in apoptosis and chaperone regulation and was recently reported to be induced by the p38 MAPK pathway (21). In accordance with this, *Bag2* is downregulated in three of four of our $Alk5^{-/-}$ experiments while upregulated in all the TGF-β-stimulated WTs. The finding of *Bag2* is part of the interesting observation that two of the most significant differentially expressed categories, according to GoMiner analysis of our data, are the protein folding and chaperone categories. These include ATP-binding cassette, subfamily E (OABP), member 1 (*Abce1*), *Bag2*, heat shock protein A (*Hspa9a*), *Hsp105*, *Hspe1*, *Hspd1*, *Hspa5bp1,* chaperonin subunit *2* (*Cct2*), *Cct3*, *Cct8,* t-complex protein-1 (*Tcp1*), peptidylprolyl isomerase D (*Ppid*), and DNAJ homolog subfamily A, member 1 (*Dnaja1*). Also interesting is that the vast majority of these genes were upregulated as a response to TGF-β stimulation or downregulated in the Alk5-deficient cells. Analysis after only 1 h of TGF-β stimulation revealed a significant overrepresentation of the chaperone activity and protein folding genes. This significance was lost at later time points, suggesting that TGF-β stimulation rapidly and transiently induces the expression of chaperones. Intriguingly, a large number of the other significantly overrepresented categories of target genes, identified by GoMiner analysis, are also involved in different aspects of protein processing. Among these, the most striking finding is the upregulation of genes involved in translation. Of the 71 genes involved in "translation regulator activity" spotted on the microarray, as many as nine were significantly upregulated in response to TGF-β stimulation. Additionally, eight of these nine genes were subunits of eukaryotic initiating factors (eIFs), responsible for the first step in protein biosynthesis. These include Eif2, subunit 3, structural gene X-linked (*Eif2s3x*), *Eif3s9*, *Eif2s1*, *Eif2s2*, *Eif5a*, and *Eif4g1*. Furthermore, these categories exhibited kinetics similar to the protein folding and chaperone genes with rapid upregulation as a response to TGF-β signaling.

Comparing our differentially expressed genes with identified gene targets from other studies performed on human lung fibroblasts (1) and human dermal fibroblasts (22), we obtained an overlap of ~10%. Because of the differences in experimental design and species, extensive comparisons between our study and these two were not performed. However, *Myc* and connective tissue growth factor (*Ctgf*) were identified in both our study and the one performed by Yang et al. (23) as well as in the study performed on human lung fibroblasts, indicating these genes to be somewhat species-independent targets of TGF-β signaling. Furthermore, it is interesting to note the immense downregulation of the *Id* genes by TGF-β in our study. This is in contrast to findings in the study of human lung fibroblasts, where the *ID* genes were induced by TGF-β (1). On the contrary, a microarray study of epithelial cells reported a repression of *ID1*, *ID2*, and *ID3* expression as a response to TGF-β, in agreement with our results (7). This downregulation of *ID* genes is implicated in the mechanism of proliferation arrest exercised by TGF-β in these cells. Thus our findings showing an *Id* expression in MEFs similar to the expression in adult epithelial cells indicate that the downregulation of these genes is part of the proliferation inhibition by TGF-β in MEFs.

To conclude, the use of $Alk5^{-/-}$ MEFs has given us the unique opportunity to investigate and exclude the use of alternative receptors for TGF-β in fibroblasts. Additionally, the absence of alternative receptors makes this model ideal for the identification of target genes of TGF-β signaling, which will be important for understanding the mechanisms behind the diverse effects of this multifunctional cytokine.

ACKNOWLEDGMENTS

We thank Dr. Mikael Sigvardsson, Dr. Peter ten Dijke, and Johan Vallon-Christersson for valuable discussions and advice.

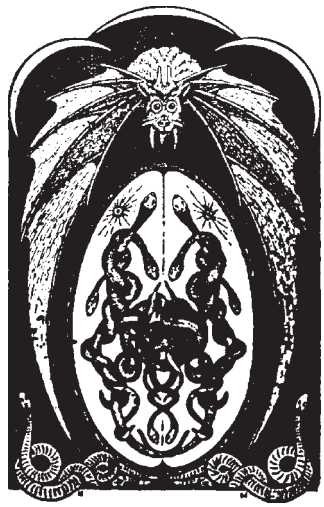
GRANTS

This work was supported by grants to S. Karlsson from Cancerfonden-Sweden, Barncancerfonden-Sweden, The Medical Research Council-Sweden, and The Juvenile Diabetes Research Foundation and a clinical research award from Lund University Hospital and by grants to M. Ringnér from the Swedish Research Council. This work was also supported by a grant to S. Karlsson and M. Ringnér from the Research School in Genomics and Bioinformatics-Sweden. The Lund Stem Cell Center is supported by a Center of Excellence Grant in Life Sciences from the Swedish Foundation for Strategic Research.

REFERENCES

1. **Chambers RC, Leoni P, Kaminski N, Laurent GJ, and Heller RA.** Global expression profiling of fibroblast responses to transforming growth factor-β1 reveals the induction of inhibitor of differentiation-1 and provides evidence of smooth muscle cell phenotypic switching. *Am J Pathol* 162: 533–546, 2003.
2. **Chang H, Brown CW, and Matzuk MM.** Genetic analysis of the mammalian transforming growth factor-β superfamily. *Endocr Rev* 23: 787–823, 2002.
3. **Foletta VC, Lim MA, Soosairajah J, Kelly AP, Stanley EG, Shannon M, He W, Das S, Massague J, Bernard O, and Soosairaiah J.** Direct signaling by the BMP type II receptor via the cytoskeletal regulator LIMK1. *J Cell Biol* 162: 1089–1098, 2003.
4. **Goldberg HJ, Huszar T, Mozes MM, Rosivall L, and Mucsi I.** Overexpression of the type II transforming growth factor-β receptor inhibits fibroblasts proliferation and activates extracellular signal regulated kinase and c-Jun N-terminal kinase. *Cell Biol Int* 26: 165–174, 2002.
5. **Goumans MJ, Valdimarsdottir G, Itoh S, Lebrin F, Larsson J, Mummery C, Karlsson S, and ten Dijke P.** Activin receptor-like kinase (ALK)1 is an antagonistic mediator of lateral TGFβ/ALK5 signaling. *Mol Cell* 12: 817–828, 2003.
6. **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, and White R.** The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 32 *(Database* issue): D258–D261, 2004.
7. **Kang Y, Chen CR, and Massague J.** A self-enabling TGFβ response coupled to stress signaling: Smad engages stress response factor ATF3 for Id1 repression in epithelial cells. *Mol Cell* 11: 915–926, 2003.
8. **Kletsas D, Stathakos D, Sorrentino V, and Philipson L.** The growth-inhibitory block of TGF-β is located close to the G1/S border in the cell cycle. *Exp Cell Res* 217: 477–483, 1995.
9. **Lai YT, Beason KB, Brames GP, Desgrosellier JS, Cleggett MC, Shaw MV, Brown CB, and Barnett JV.** Activin receptor-like kinase 2 can mediate atrioventricular cushion transformation. *Dev Biol* 222: 1–11, 2000.
10. **Larsson J, Goumans MJ, Sjostrand LJ, van Rooijen MA, Ward D, Leveen P, Xu X, ten Dijke P, Mummery CL, and Karlsson S.** Abnor-

mal angiogenesis but intact hematopoietic potential in TGF-β type I receptor-deficient mice. *EMBO J* 20: 1663–1673, 2001.
11. **Ling MT, Wang X, Tsao SW, and Wong YC.** Down-regulation of Id-1 expression is associated with TGFβ1-induced growth arrest in prostate epithelial cells. *Biochim Biophys Acta* 1570: 145–152, 2002.
12. **Liu Y and Ringner M.** Multiclass discovery in array data. *BMC Bioinformatics* 5: 70, 2004.
13. **Massague J, Blain SW, and Lo RS.** TGFβ signaling in growth control, cancer, and heritable disorders. *Cell* 103: 295–309, 2000.
14. **Meyer K, Lee JS, Dyck PA, Cao WQ, Rao MS, Thorgeirsson SS, and Reddy JK.** Molecular profiling of hepatocellular carcinomas developing spontaneously in acyl-CoA oxidase-deficient mice: comparison with liver tumors induced in wild-type mice by a peroxisome proliferator and a genotoxic carcinogen. *Carcinogenesis* 24: 975–984, 2003.
15. **Nathan C and Sporn M.** Cytokines in context. *J Cell Biol* 113: 981–986, 1991.
16. **Oh SP, Seki T, Goss KA, Imamura T, Yi Y, Donahoe PK, Li L, Miyazono K, ten Dijke P, Kim S, and Li E.** Activin receptor-like kinase 1 modulates transforming growth factor-β1 signaling in the regulation of angiogenesis. *Proc Natl Acad Sci USA* 97: 2626–2631, 2000.
17. **Ringner M, Veerla S, Andersson S, Staaf J, and Hakkinen J.** ACID: a database for microarray clone information. *Bioinformatics* 20: 2305–2306, 2004.
18. **Roberts AB.** Molecular and cell biology of TGF-β. *Miner Electrolyte Metab* 24: 111–119, 1998.
19. **Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg A, and Peterson C.** BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. *Genome Biol* 3: SOFTWARE0003, 2002.
20. **Siegel PM, Shu W, and Massague J.** Mad upregulation and Id2 repression accompany transforming growth factor (TGF)-β-mediated epithelial cell growth suppression. *J Biol Chem* 278: 35444–35450, 2003.
21. **Ueda K, Kosako H, Fukui Y, and Hattori S.** Proteomic identification of Bcl2-associated athanogene 2 as a novel MAPK-activated protein kinase 2 substrate. *J Biol Chem* 279: 41815–41821, 2004.
22. **Verrecchia F, Chu ML, and Mauviel A.** Identification of novel TGF-β /Smad gene targets in dermal fibroblasts using a combined cDNA microarray/promoter transactivation approach. *J Biol Chem* 276: 17058–17062, 2001.
23. **Yang YC, Piek E, Zavadil J, Liang D, Xie D, Heyer J, Pavlidis P, Kucherlapati R, Roberts AB, and Bottinger EP.** Hierarchical model of gene regulation by transforming growth factor-β. *Proc Natl Acad Sci USA* 100: 10269–10274, 2003.
24. **Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, Bussey KJ, Riss J, Barrett JC, and Weinstein JN.** GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol* 4: R28, 2003.



Paper IV

Revealing signaling pathway deregulation by using gene expression signatures and regulatory motif analysis

Yingchun Liu[1]and Markus Ringnér*[1]

[1]Computational Biology and Biological Physics, Department of Theoretical Physics, Lund University, Sölvegatan 14A, SE-223 62 Lund, Sweden

Email: Yingchun Liu - spring@thep.lu.se; Markus Ringnér*- markus@thep.lu.se;

*Corresponding author

Abstract

Gene expression signatures consisting of tens to hundreds of genes have been found to be informative for different biological states. Recently, many computational methods have been proposed for biological interpretation of such signatures. However, there is a lack of methods for identifying cell signaling pathways whose deregulation result in an observed expression signature. We present a strategy for identifying such signaling pathways and evaluate the strategy using six human and mouse gene expression signatures.

Background

Genetic aberrations and variations in cellular processes are usually reflected in the expression levels of many genes. Hence, such alterations can potentially be characterized by their gene expression profiles. Gene expression profiling, in particular DNA microarray analysis, has been widely used in attempts to reveal the underlying mechanisms of many diseases, different developmental stages, cellular responses to different conditions, and many other biological phenomena (for example [1–3]). Gene expression signatures consisting of tens to hundreds of genes have been associated with many important aspects of the systems studied. To help realize the full potential of gene expression studies, a variety of methods such as GoMiner [4], EASE [5], Catmap [6], GSEA [7], and ArrayXPath [8], have been developed to relate gene expression profiles or signatures to a broad range of biological categories. Although some of these methods include signaling pathways in their categories, their focus has not been on regulatory mechanisms that control the observed gene expression changes.

Signal transduction is at the core of many

regulatory systems. Cellular functions such as growth, proliferation, differentiation, and apoptosis are regulated by signaling pathways. Appropriate regulation of such pathways is essential for the normal functioning of cells. Cells affected by disease often have one or several signaling pathways abnormally activated or inactivated. For example, cancer is a disease of deregulated cell proliferation and death [9]. To uncover mechanisms underlying cellular phenotypes, it is therefore crucial to systematically analyze gene expression signatures in the context of signaling pathways.

In signal transduction, ligands, usually from outside the cell, interact with receptors on the surface of the cell membrane or with nuclear receptors. These interactions trigger a cascade of biochemical reactions. Proteins, called transcription factors (TFs) and cofactors, are eventually transported to the nucleus of the cell where they turn transcription of target genes on or off. A signaling pathway is composed of a set of molecular components conveying the signal, such as ligands, receptors, enzymes, TFs, and cofactors.

When a pathway is activated, the expression levels of the components of the pathway are not necessarily affected. For example, mutation of a TF can change the expression levels of its target genes, without necessarily affecting the expression levels of the TF itself or other components of the pathway. Also, pathway components might not be regulated at the transcriptional level, instead they are often regulated post-translationally, for example, by phosphorylation. Proteomic data could be used to detect such modifications and be used for pathway analysis, but currently there is a lack of such genome-wide protein data. It has been pointed out that gene expression signatures may be more reliable indicators of pathway activities than protein data for single components in signaling pathways [10]. Taking all these considerations into account, we reason that the activity of a signaling pathway may currently be best characterized by the expression levels of its target genes. In support of this hypothesis, Breslin *et al.* have shown the capacity of expression levels of known target genes to reflect pathway activities [11]. However, knowledge about target genes of transcription factors is far from complete, which hampers accurate prediction of pathway activities. On the other hand, the *cis*-regulatory motifs to which transcription factors bind are often better characterized. For organisms with sequenced genomes, these motifs enable genome-wide identification of putative target genes, by looking for potential transcription factor binding sites in promoter sequences. Therefore, integrating regulatory motif analysis with pathway information would be a potential approach to break this bottleneck for pathway analysis. Recently, the feasibility to use putative binding sites to identify transcription factors responsible for gene expression signatures of human cancer has been demonstrated [12].

Here we present a strategy to discover activated and inactivated signaling pathways from gene expression signatures by using regulatory motif analysis (Figure 1). To achieve this goal, we begin by extracting all signaling pathways in the TRANSPATH database [13], and characterize each pathway by the TFs that mediate it. In all human and mouse promoter sequences, we identify putative binding sites of all the TFs mediating pathways using transcription factor binding site position weight matrices from the TRANSFAC database [14]. Next, we investigate promoters of genes in gene expression signatures for an enrichment of these putative binding sites. Finally, we measure the activity of a pathway in a gene expression signature in terms of the enrichment of binding motifs for the TFs mediating the pathway. Although the use of putative TF binding sites will introduce false-positive target genes for each TF, when the promoters of a set of co-expressed genes are enriched for a putative transcription factor binding site, the gene set is also likely enriched for true target genes. Moreover, our

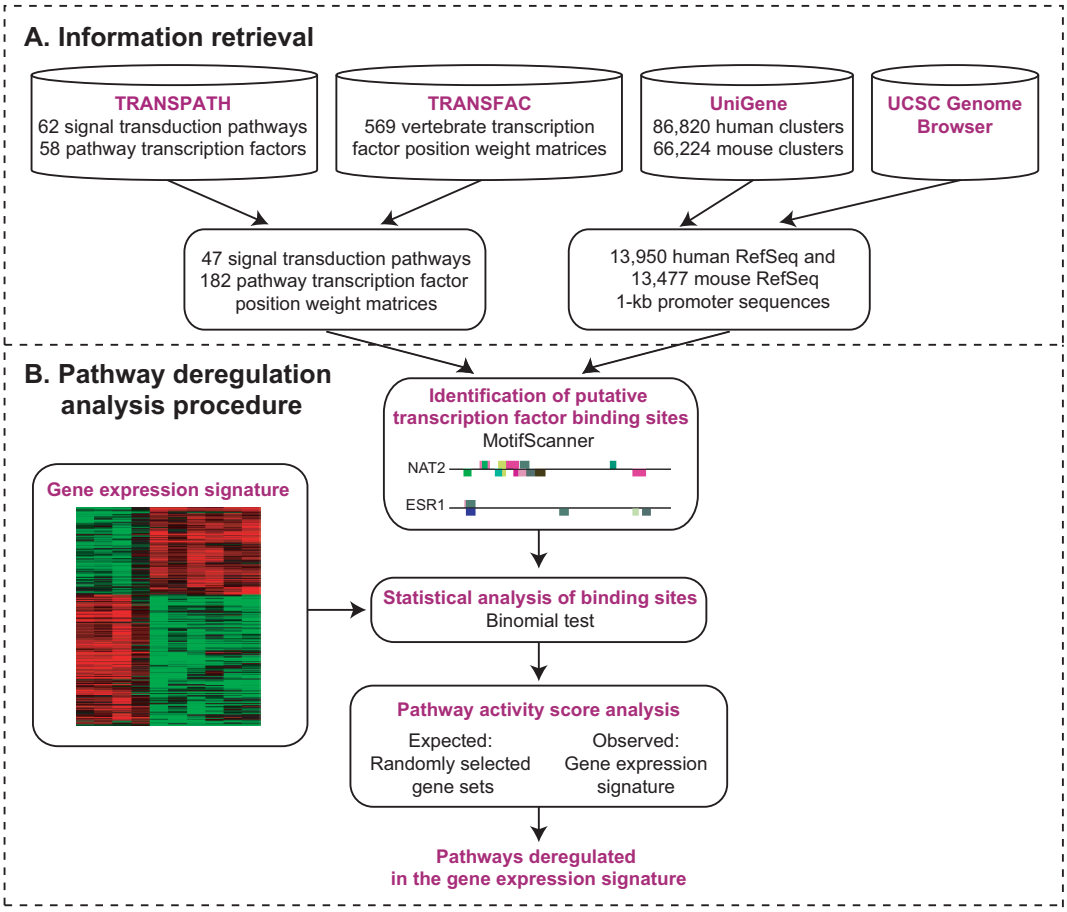


Figure 1: Overview of the method used to reveal pathways deregulated in gene expression signatures. (A) Information was retrieved and integrated from four sources: TRANSPATH, TRANSFAC, UniGene, and the UCSC Genome Browser. (B) Putative transcription factor binding sites in promoter regions were identified using MotifScanner. Enrichment of putative transcription binding sites among genes in a gene signature was assessed using a binomial test. Each pathway was scored in terms of an enrichment for putative binding sites for the transcription factors mediating the pathway. The significance of a pathway's relevance for a gene signature was assessed by using randomly selected gene sets from the genome.

strategy to integrate regulatory motif analysis with knowledge about which TFs act together in pathways further reduces the influence of false-positive targets on the identification of pathways. We have used gene expression signatures containing both up- and down-regulated genes because we think such mixed signatures allow for a more comprehensive identification of pathway deregulation. Obviously, our strategy can also be applied to up- and down-regulated genes separately. Our results for six human and mouse gene expression signatures demonstrate the power of our method to identify relevant pathways. We compared our results with those obtained using the EASE software [5].

Results and Discussion

Gene signatures for oncogenic pathways

To examine the ability of our method to accurately detect the activity of pathways, we obtained gene signatures for three oncogenic pathways produced by Bild *et al.* [15]. These signatures consist of genes for which the expression levels in human mammary epithelial cells were highly correlated with the activation status of the oncogenes E2F3 (268 genes), Myc (218 genes), or Ras (383 genes), respectively. These three oncogenic pathways are often activated in solid tumors, including breast tumors, where they contribute to tumor development or progression. Bild *et al.* verified the activation status of each pathway using various biochemical measurements and demonstrated that the expression patterns in each signature were specific to each pathway. Hence, these signatures are ideal for evaluating our strategy to identify activated pathways. The statistically significant pathways identified by our method for the three gene signatures are shown in Table 1.

The E2F pathway was found extremely significant for the E2F3 gene signature. E2F3 is a member of the E2F transcription factor family (E2Fs). E2Fs can induce cell cycle G1-to-S transition and activate a large number of genes encoding proteins essential for DNA replication [16,17]. E2F-1, another member of E2Fs, can form dimers with DP-1 making this activation more efficient [18]. Our method identified both E2F-1 (P-value$<$0.001) and DP-1 (P-value$<$0.001) as significant TFs for this signature.

TRANSPATH does not contain a strictly defined Myc pathway, but it includes three pathways containing c-Myc as a TF: EGF, Notch, and MAPK. We identified c-Myc as a significant TF for this signature (P-value$<$0.001), and both the EGF and the Notch pathways were found to be significant. MAPK was not found to be significant. The only significant TF found for MAPK was c-Myc, perhaps suggesting that induction of c-Myc is not sufficient to deregulate this pathway. Consistent with this suggestion, it has been shown that elevated c-Myc expression is not sufficient for tumorigenesis in human mammary epithelial cells [19]. Interestingly, we also found the hypoxia-inducible pathway HIF-1 significant. Studies have shown that HIF-1 is activated in many tumors including breast cancer [20], as a consequence of shortage in oxygen supply during sustained tumor growth. Moreover, it has been reported that HIF-1α counteracts Myc to induce cell cycle arrest, and HIF-1α down-regulates Myc-activated genes [21].

In the analysis of the Ras gene signature, we found the MAPK and p38 pathways to be significantly relevant. This finding is consistent with the fact that Ras activates mitogen-activated protein kinases (MAPKs) including ERK and p38. It has been shown in human fibroblasts that a sustained high intensity Ras signal induces MEK and ERK to higher levels, eventually resulting in stimulation of the p38 pathway [22] and that the p38 pathway provides negative feedback for Ras proliferation [23]. Several of the pathways we found to be significant, contained NF-κB as a significant TF (P-value=0.002), including the receptor activator of NF-κB (RANK) and TNFα pathways. It has been shown that NF-κB has an essential role in breast cancer progression, and activation of NF-κB signaling is in particular required for the epithelial-mesenchymal transition in Ras-transformed epithelial cells [24]. We identified the stress pathway as affected, perhaps only because this pathway overlaps the p38 pathway. Also, we identified the TLR3 and TLR4 pathways as responsive to Ras stimulation. A recent study has shown that toll-like receptors (TLRs) are expressed in a variety of tumors and trigger tumor self-protection mechanisms [25] making it plausible that they are induced by Ras activation.

In addition to those pathways affected specifically for an oncogenic activation signature, the caspases pathway was found to be significantly affected for all three signatures. The

Table 1 - Significant pathways for oncogenic gene signatures

Pathway	TFs	Significant TFs	P-value
E2F3 gene signature			
E2F	DP-1, E2F, p53	DP-1, E2F	<0.001
Caspases	CREB, Max, SRF, p53, AP-2α	AP-2α	<0.001
Myc gene signature			
AhR	AhR, ER-α, Sp1, p300, NF-κB, Arnt	AhR, Sp1, NF-κB, Arnt	<0.001
HIF-1	p53, p300, HIF-1α, HNF-4α2, Arnt	HIF-1α, Arnt	<0.001
Notch	Max, LEF-1, p300, c-Myc	Max, c-Myc	<0.001
EGF	c-Fos, Elk-1, Sp1, STAT3, c-Jun, STAT1α, c-Myc	Sp1, c-Myc	0.002
Caspases	CREB, Max, SRF, p53, AP-2α	Max, AP-2α	0.002
c-Kit	MITF, Sp1, Tal-1, p300, GATA-1	MITF, Sp1, Tal-1	0.006
Ras gene signature			
AhR	AhR, ER-α, Sp1, p300, NF-κB, Arnt	Sp1, NF-κB	<0.001
Apoptosis	p53, FOXO3a, NF-κB	p53, NF-κB	0.001
Caspases	CREB, Max, SRF, p53, AP-2α	CREB, p53, AP-2α	0.004
RANK	MITF, PU.1, c-Jun, NF-κB	PU.1, NF-κB	0.008
TNFα	AP-1, NF-κB	AP-1, NF-κB	0.009
TLR4	CREB, CRE-BP2, STAT1, Elk-1, p300, IRF-3, IRF-7, NF-κB	CREB, CRE-BP2, NF-κB	0.015
MAPK	CREB, Elk-1, p53, c-Jun, c-Myc	CREB, p53	0.023
TLR3	CRE-BP2, p300, c-Jun, IRF-3, IRF-7, NF-κB	CRE-BP2, NF-κB	0.034
p38	ELk-1, p53, MITF, PPAR-α, CHOP-10, Max, CREB, PU.1, MRF4, HNF-1α, CRE-BP2, NF-AT2, STAT3	p53, PPAR-α, CHOP-10, CREB, PU.1, CRE-BP2	0.035
Stress	PPAR-γ, c-Ets-1, PPAR-α, Max, NF-AT2, HSF1, c-Jun, Elk-1, p53, CHOP-10, CREB, CRE-BP2, RXR-α, HNF-1α, STAT3, MRF4	PPAR-α, p53, CHOP-10, CREB, CRE-BP2	0.037

caspases pathway triggers cell death. Because evasion of cell death is essential for tumor development [9], it is likely that this pathway is repressed regardless of which of the oncogenes is activated. Indeed, it has been indicated that over-expression of E2F3 or Ras induces tumor invasion through interaction with AP-2α, a characteristic transcription factor in the caspases pathway, in epithelial cells of bladder cancer [26]. It has also been shown that c-Myc represses AP-2α *trans*-activation [27]. Another pathway found to be affected for more than one signature was the AhR pathway, which was found to be significant for both the Myc and Ras gene signatures. It has been demonstrated that the AhR TF is constitutively active at high levels in mammary tumors as compared to in normal mammary glands, and suggested that it contributes to ongoing mammary tumor cell growth [28].

Taken together, our results for these three oncogenic gene signatures demonstrate the power of our method to accurately identify the active pathways. Moreover, we found additional pathways known to be relevant for each oncogenic pathway. These results highlight the potential of our method to generate hypotheses for connections between pathways.

Gene signatures for the TGF-β pathway

Sets of genes claimed to belong to a gene signature are often sensitive to sample selection and have small overlaps for different studies [29,30]. This issue has raised debate about the credibility of such signatures. A possible explanation for small overlaps is that there may be redundancy in expression profiles; many gene sets are equally good at distinguishing a phenotype of interest. In this case, gene sets with small over-

Table 2 - Significant pathways for TGF-β gene signatures

Pathway	TFs	Significant TFs	P-value
Yang *et al.* gene signature			
AhR	AhR, ER-α, Sp1, p300, NF-κB, Arnt	AhR, Sp1, p300, NF-κB, Arnt	<0.001
EGF	c-Fos, Elk-1, Sp1, STAT3, c-Jun, STAT1α, c-Myc	Sp1, c-Jun, c-Myc	<0.001
c-Kit	MITF, Sp1, Tal-1, p300, GATA-1	Sp1, p300	<0.001
p53	TFIIA, E2F-1, p53, p300, BRCA1, YY1	E2F-1, p53, p300, BRCA1	<0.001
Caspases	CREB, Max, SRF, p53, AP-2α	CREB, Max, p53, AP-2α	<0.001
MAPK	CREB, Elk-1, p53, c-Jun, c-Myc	CREB, p53, c-Jun, c-Myc	<0.001
E2F	DP-1, E2F, p53	DP-1, E2F, p53	<0.001
HIF-1	p53, p300, HIF-1α, HNF-4α2, Arnt	p53, p300, HIF-1α, Arnt	<0.001
Stress	PPAR-γ, c-Ets-1, PPAR-α, Max NF-AT2, HSF1, c-Jun, Elk-1, p53, CHOP-10, CREB, CRE-BP2, RXR-α, HNF-1α, STAT3, MRF4	Max, c-Jun, p53, CREB, CRE-BP2, RXR-α	0.001
TLR3	CRE-BP2, p300, c-Jun, IRF-3, IRF-7, NF-κB	CRE-BP2, p300, c-Jun, NF-κB	0.002
TLR4	CREB, CRE-BP2, STAT1, Elk-1, p300, IRF-3, IRF-7, NF-κB	CREB, CRE-BP2, p300, NF-κB	0.002
p38	ELk-1, p53, MITF, PPAR-α, CHOP-10, Max, CREB, PU.1, HNF-1α, CRE-BP2, NF-AT2, STAT3, MRF4	p53, Max, CREB, CRE-BP2	0.003
JNK	CRE-BP2, p53, HSF1, PPAR-γ, STAT3, c-Jun, c-Ets-1	CRE-BP2, p53, c-Jun	0.004
TGF-β	LEF-1, CRE-BP2, Smad2, Smad3, Smad4	CRE-BP2, Smad4	0.006
EDAR	c-Jun, NF-κB	c-Jun, NF-κB	0.015
IL-1	ELk-1, c-Jun, NF-κB	c-Jun, NF-κB	0.015
TCR2	c-Jun, NF-κB, NF-AT	c-Jun, NF-κB	0.018
RANK	MITF, PU.1, c-Jun, NF-κB	c-Jun, NF-κB	0.020
Hypoxia	ER-α, p53, AP-1, HIF-1α	p53, HIF-1α	0.033
Notch	Max, LEF-1, p300, c-Myc	Max, p300, c-Myc	0.037
Karlsson *et al.* gene signature			
AhR	AhR, ER-α, Sp1, p300, NF-κB, Arnt	AhR, Sp1, Arnt	<0.001
EGF	c-Fos, Elk-1, Sp1, STAT3, c-Jun, STAT1α, c-Myc	Sp1, STAT1α, c-Myc	<0.001
c-Kit	MITF, Sp1, Tal-1, p300, GATA-1	Sp1, Tal-1	<0.001
p53	TFIIA, E2F-1, p53, p300, BRCA1, YY1	E2F-1, BRCA1	<0.001
Caspases	CREB, Max, SRF, p53, AP-2α	Max, AP-2α	<0.001
E2F	DP-1, E2F, p53	DP-1, E2F	0.002
HIF-1	p53, p300, HIF-1α, HNF-4α2, Arnt	HIF-1α, Arnt	0.006
Notch	Max, LEF-1, p300, c-Myc	Max, c-Myc	0.019

laps may still arise from activation or repression of identical pathways.

To validate our method as a guide to pathway analysis in this regard, we analyzed target genes of the transforming growth factor-β (TGF-β) pathway from two independent studies. One data set contains 360 genes identified by comparing expression profiles of murine embryonic fibroblast (MEF) cells deficient in Smad2, Smad3, or MAPK ERK, which are mediators of TGF-β signaling, with wild-type MEFs in response to 1, 2, or 4 hours of TGF-β stimulation [31]. The other data set contains 465 targets differentially expressed between MEF cells with the TGF-β receptor Alk5 knocked out and wild-type MEFs stimulated with TGF-β for 2, 4, or 16 hours [32].

Whereas there are only 29 genes in common for the two data sets, many of the active pathways we found are the same (Table 2). In particular, all five pathways with P<0.001 for the Karlsson *et al.* data set also have P<0.001 for the Yang *et al.* data set. We identified the TGF-β pathway as significant for the Yang *et al.* target genes, but not for the Karlsson *et al.* genes. This discrepancy is possibly due to the different durations of TGF-β stimulation in the two experiments. Yang *at al.* reported that Smad3/Smad4 binding motifs are only present in immediate-early target genes but not in the

intermediate ones [31]. The lack of an overabundance of genes containing Smad binding motifs in the Karlsson *et al.* data set, suggests that it consists of intermediate or late response genes. A target gene of TGF-β signaling is Myc and it is one of the genes in common for both data sets. The repression of Myc by TGF-β stimulation is mediated by the transcription factors E2F4/5 and DP-1 [33]. In agreement with this picture, we found all six pathways that were significant for the Myc gene signature (Table 1) as well as the E2F pathway to be significant for both TGF-β data sets (Table 2).

The fibroblasts used by Yang *et al.* to identify TGF-β responsive genes included MEFs with genetic ablation of MAPK ERK. The oncogene Ras activates ERK and eight of the ten pathways we found to be significant for the Ras gene signature (Table 1) were also found to be significant for the Yang *et al.* gene signature (Table 2). This finding indicates that the Yang *et al* gene signature is a mixture of the transcriptional response to both MAPK- and Smad-signaling. For this data set, four pathways appeared as significant only because they contain the TFs c-Jun and NF-κB. These two TFs also appear in other significant pathways supported by additional significant TFs, including the pathways AhR, EGF, MAPK, and p38. Biochemical investigations are required to reveal if the pathways with only c-Jun and NF-κB are indeed deregulated, or if they are false positives likely to go away as the information in pathway databases improves.

Table 3 - Significant pathways for the breast cancer prognosis gene signature

Pathway	TFs	Significant TFs	P-value
E2F	DP-1, E2F, p53	DP-1, E2F	<0.001
AhR	AhR, ER-α, Sp1, p300, NF-κB, Arnt	AhR, Sp1	0.017
Caspases	CREB, Max, SRF, p53, AP-2α	AP-2α	0.039

This analysis of TGF-β signaling provides an example that pathway analysis can be used to find common pathways underlying gene sets with small overlaps. In addition, we have again verified that our method identifies relevant pathways.

Poor prognosis gene signature for breast cancer

Finally, we tested the ability of our method to identify signaling pathways involved in a disease by using a gene expression signature from breast tumor samples. We used a signature distinguishing patients who developed distant metastases within five years from patients who remained disease free for at least five years [34]. This poor prognosis gene signature contains 70 genes that we investigated for pathway activities. The signature consists of genes annotated as being involved in cell cycle, invasion, metastasis, and angiogenesis [34].

Consistent with the functional annotation of the genes, we found a pathway regulating the cell cycle, E2F, most significantly associated with the poor prognosis signature (Table 3). Activation of the E2F pathway can induce the transition from G1 to S phase in the cell cycle. The percentage of cells in a tumor cell population that are in S phase is known to be associated with shorter disease-free survival [35]. We also found the AhR pathway to be significant (Table 3). The AhR pathway has been suggested to inhibit apoptosis while promoting transition to an invasive, metastatic phenotype for breast tumors [28]. Interestingly, we found the caspases pathway that regulates apoptosis to be significant (Table 3). This finding is consistent with the indication in recent studies that apoptosis is a central mechanism regulating metastasis [36].

Our analysis of the poor prognosis signature highlights the potential of our method to reveal pathways that both are consistent with functional annotations of genes in signatures and

provides a more detailed insight into the molecular mechanisms underlying the annotations.

Pathway component enrichment analysis

We compared our results with results obtained by looking for enrichment of pathway components in the gene signatures. EASE is a widely used software application that is customizable to allow the use of functional information about genes from various sources [5]. Among other things, EASE can search for an enrichment of genes in pathways from the KEGG, GenMAPP, and BBID pathway databases.

We used EASE to search for enrichment of pathway components for the six gene signatures. We could only identify significant pathways (EASE score <0.05) for one data set: the Ras oncogenic gene signature. For this gene signature two pathways, signal transduction - homo sapiens and phosphatidylinositol signaling system, both from KEGG, were found to be significant. It has been indicated that Ras activates the phosphatidylinositol signaling system, although not at levels sufficient for oncogenic transformation of human mammary epithelial cells [19].

This analysis makes it clear that identifying pathways from a gene signature by mapping genes in the signature to pathway components is difficult. As expected, genes in a gene signature are more likely target genes of the deregulated pathways.

Conclusions

We have presented a strategy to identify signaling pathways whose deregulation result in an observed gene expression signature. The strategy is based on combining identification of putative TF binding sites in promoter regions of genes with knowledge about which TFs act in the same pathway. The major conclusions from our results for six human and mouse gene expression signatures are as follows. First, it is feasible to identify pathways deregulated in mammalian gene expression signatures by viewing such signatures as a collection of target genes of the TFs mediating the pathways. Second, while binding site analysis alone can identify key transcription factors, combining such analysis with pathway information improves the potential to direct attention to possible mechanisms driving an observed transcriptional response. Third, mapping gene expression signatures onto pathways can guide the identification of common regulatory programs driving different signatures with small overlaps, as well as the identification of diverse regulatory programs driving a single signature. As pathway databases are steadily growing in size and quality, we expect that methods combining regulatory motif analysis with pathway information will be even more useful in the future.

Methods

Pathway information retrieval

Signal transduction pathways were taken from the TRANSPATH database (Release 7.1). For the 62 pathways defined in the database, 58 components were identified as TFs mediating at least one pathway. We extracted pathway-TF pairs from the map files provided by TRANSPATH and extracted DNA binding motifs of these TFs from TRANSFAC (Release 10.1). The binding motifs used were 6-24bp long and each was represented by a position-weight-matrix (PWM) that indicates the experimentally determined frequency of the four nucleotides at each position. The 58 pathway TFs were associated with 182 PWMs and 47 pathways were represented by at least one PWM. These 47 pathways were used in our subsequent analysis (Figure 1A).

Identification of transcription factor binding sites

Each human and mouse cluster in the UniGene database (Human build 193; Mouse build 155), was associated with RefSeq reference sequences using ACID [37]. Clusters that did not match a RefSeq or matched multiple RefSeqs were excluded from the analysis. This procedure resulted in 13,950 human and 13,477 mouse RefSeqs for which we retrieved 1-kb promoter sequences from the University of California Santa Cruz Genome Browser using human assembly hg18 and mouse assembly mm7 (Figure 1A).

Putative TF binding sites in the promoter sequences were identified by using MotifScanner, a part of the Toucan software [38,39], which can search for the occurrences of a list of known motifs in each query sequence. MotifScanner requires several arguments including: i) a set of query sequences, ii) a background model which scores the frequencies of single nucleotides or oligonucleotides of fixed size, and iii) a set of motifs represented by PWMs. In our analysis, all 1-kb promoter sequences for a species were used both as a query set and to generate a background model for oligonucleotides of size three [40]. All PWMs for the pathway TFs were used when searching for putative binding sites. Default values were used for all other MotifScanner parameters. For each promoter sequence, MotifScanner outputs the number of occurrences for each motif.

Statistical analysis of binding sites

A binomial test was used to assess which putative binding sites were enriched in promoters of genes in a gene signature (Figure 1B). The possible number of start positions (R) in n promoter sequences for a motif was approximated as

$$R(n) = 2 \times \sum_{i=1}^{n} (L_i - w + 1) \ , \qquad (1)$$

where L_i is the length of the i^{th} sequence and w is the length of the motif. To generate a null model, we calculated the overall frequency (f) of each motif (m) by dividing the observed number of occurrences of the motif (K) in all human or mouse promoter sequences (N) with the number of possible start positions

$$f_m = \frac{K}{R(N)} \ . \qquad (2)$$

Based on this overall frequency, we calculated a P-value corresponding to the probability of observing k or more occurrences of the motif m in a set of n ($n \leq N$) promoter sequences by chance as described in [38]

$$\text{P-value}(m) = \sum_{j=k}^{R(n)} \binom{R(n)}{k} \times f_m^k \times (1-f_m)^{R(n)-k} \ . \qquad (3)$$

Thus, a small P-value indicates an enrichment for motif m in the promoters of genes in a gene signature.

Statistical analysis of pathway activities

The activity of a pathway in a gene expression signature was assessed by investigating if the signature was enriched in binding motifs for the TFs that mediate the pathway (Figure 1B). Letting TF(p) denote the set of TFs for a pathway p, and M(t) the set of binding motifs for a TF t, we first defined a score for a TF t as

$$S(t) = - \sum_{m \in \text{M}(t)} \log(\text{P-value}(m)) \ , \qquad (4)$$

and second a score for a pathway p as

$$S(p) = \sum_{t \in \text{TF}(p)} S(t) \ . \qquad (5)$$

We generated gene sets of the same size as the gene signature by randomly selecting genes from the human or mouse genome. Finally, we calculated P-values for each pathway and each TF by calculating the probability of obtaining an equal or larger score for a random gene set as compared to the gene signature. We used 1000 randomly selected sets in each of our analyses.

TFs with P-value<0.1 were considered significant. Pathways were considered significant if they met two criteria: (i) P-value<0.05, and (ii) at least two significant TFs or one significant TF unique for the pathway. An important aspect of our strategy is that we have not introduced a binary decision whether a gene is a target gene of a TF or not, rather the P-value for each motif was used throughout the pathway analysis.

Gene signatures

We obtained six different publicly available human and mouse gene signatures. Gene identifiers were mapped to UniGene clusters using ACID [37]. Gene identifiers that mapped to multiple UniGene clusters were removed from further analysis.

Enrichment for pathway components in gene signatures was evaluated using EASE [5]. In the EASE analysis, we selected the categories BBID pathway, GenMAPP pathway, and KEGG pathway, used the EASE score as the primary score, and used all mouse or human genes for which we identified promoter regions as the general population of genes. For all other EASE settings, we used default values.

Availability

Software for the method was written using the PERL programming language and is freely available upon request.

Acknowledgements

We thank Morten Krogh and Jari Häkkinen for helpful discussions.

References

1. Brandenberger R, Wei H, Zhang S, Lei S, Murage J, Fisk GJ, Li Y, Xu C, Fang R, Guegler K, Rao MS, Mandalam R, Lebkowski J, Stanton LW: **Transcriptome characterization elucidates signaling networks that control human ES cell growth and differentiation**. *Nat Biotechnol* 2004, **22**(6):707–716.
2. Dean SO, Rogers SL, Stuurman N, Vale RD, Spudich JA: **Distinct pathways control recruitment and maintenance of myosin II at the cleavage furrow during cytokinesis**. *Proc Natl Acad Sci U S A* 2005, **102**(38):13473–13478.
3. Bjorklund M, Taipale M, Varjosalo M, Saharinen J, Lahdenpera J, Taipale J: **Identification of pathways regulating cell size and cell-cycle progression by RNAi**. *Nature* 2006, **439**(7079):1009–1013.
4. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, Bussey KJ, Riss J, Barrett JC, Weinstein JN: **GoMiner: a resource for biological interpretation of genomic and proteomic data**. *Genome Biol* 2003, **4**(4):R28.
5. Hosack DA, Dennis GJ, Sherman BT, Lane HC, Lempicki RA: **Identifying biological themes within lists of genes with EASE**. *Genome Biol* 2003, **4**(10):R70.
6. Breslin T, Eden P, Krogh M: **Comparing functional annotation analyses with Catmap**. *BMC Bioinformatics* 2004, **5**:193.
7. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proc Natl Acad Sci U S A* 2005, **102**(43):15545–15550.
8. Chung HJ, Park CH, Han MR, Lee S, Ohn JH, Kim J, Kim J, Kim JH: **ArrayXPath II: mapping and visualizing microarray gene-expression data with biomedical ontologies and integrated biological pathway resources using Scalable Vector Graphics**. *Nucleic Acids Res* 2005, **33**(Web Server issue):621–626.
9. Hanahan D, Weinberg RA: **The hallmarks of cancer**. *Cell* 2000, **100**:57–70.
10. Downward J: **Cancer biology: signatures guide drug choice**. *Nature* 2006, **439**(7074):274–275.
11. Breslin T, Krogh M, Peterson C, Troein C: **Signal transduction pathway profiling of individual tumor samples**. *BMC Bioinformatics* 2005, **6**:163.
12. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Barrette TR, Ghosh D, Chinnaiyan AM: **Mining for regulatory programs in the cancer transcriptome**. *Nat Genet* 2005, **37**(6):579–583.
13. Krull M, Pistor S, Voss N, Kel A, Reuter I, Kronenberg D, Michael H, Schwarzer K, Potapov A, Choi C, Kel-Margoulis O, Wingender E: **TRANSPATH:**

an information resource for storing and visualizing signaling pathways and their pathological aberrations. *Nucleic Acids Res* 2006, **34**(Database issue):546–551.

14. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E: **TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes**. *Nucleic Acids Res* 2006, **34**(Database issue):108–110.
15. Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, Joshi MB, Harpole D, Lancaster JM, Berchuck A, Olson JAJ, Marks JR, Dressman HK, West M, Nevins JR: **Oncogenic pathway signatures in human cancers as a guide to targeted therapies**. *Nature* 2006, **439**(7074):353–357.
16. Johnson DG, Schwarz JK, Cress WD, Nevins JR: **Expression of transcription factor E2F1 induces quiescent cells to enter S phase**. *Nature* 1993, **365**(6444):349–352.
17. Dyson N: **The regulation of E2F by pRB-family proteins**. *Genes Dev* 1998, **12**(15):2245–2262.
18. Helin K, Wu CL, Fattaey AR, Lees JA, Dynlacht BD, Ngwu C, Harlow E: **Heterodimerization of the transcription factors E2F-1 and DP-1 leads to cooperative trans-activation**. *Genes Dev* 1993, **7**(10):1850–1861.
19. Zhao JJ, Gjoerup OV, Subramanian RR, Cheng Y, Chen W, Roberts TM, Hahn WC: **Human mammary epithelial cell transformation through the activation of phosphatidylinositol 3-kinase**. *Cancer Cell* 2003, **3**(5):483–495.
20. Pugh CW, Gleadle J, Maxwell PH: **Hypoxia and oxidative stress in breast cancer. Hypoxia signalling pathways**. *Breast Cancer Res* 2001, **3**(5):313–317.
21. Koshiji M, Kageyama Y, Pete EA, Horikawa I, Barrett JC, Huang LE: **HIF-1alpha induces cell cycle arrest by functionally counteracting Myc**. *EMBO J* 2004, **23**(9):1949–1956.
22. Deng Q, Liao R, Wu BL, Sun P: **High intensity ras signaling induces premature senescence by activating p38 pathway in primary human fibroblasts**. *J Biol Chem* 2004, **279**(2):1050–1059.
23. Chen G, Hitomi M, Han J, Stacey DW: **The p38 pathway provides negative feedback for Ras proliferative signaling**. *J Biol Chem* 2000, **275**(50):38973–38980.
24. Huber MA, Azoitei N, Baumann B, Grunert S, Sommer A, Pehamberger H, Kraut N, Beug H, Wirth T: **NF-kappaB is essential for epithelial-mesenchymal transition and metastasis in a model of breast cancer progression**. *J Clin Invest* 2004, **114**(4):569–581.
25. Huang B, Zhao J, Li H, He KL, Chen Y, Chen SH, Mayer L, Unkeless JC, Xiong H: **Toll-like receptors on tumor cells facilitate evasion of immune surveillance**. *Cancer Res* 2005, **65**(12):5009–5014.
26. Wolff EM, Liang G, Jones PA: **Mechanisms of Disease: genetic and epigenetic alterations that drive bladder cancer**. *Nat Clin Pract Urol* 2005, **2**(10):502–510.
27. Batsche E, Cremisi C: **Opposite transcriptional activity between the wild type c-myc gene coding for c-Myc1 and c-Myc2 proteins and c-Myc1 and c-Myc2 separately**. *Oncogene* 1999, **18**(41):5662–5671.
28. Schlezinger JJ, Liu D, Farago M, Seldin DC, Belguise K, Sonenshein GE, Sherr DH: **A role for the aryl hydrocarbon receptor in mammary gland tumorigenesis**. *Biol Chem* 2006, **387**(9):1175–1187.
29. Ein-Dor L, Kela I, Getz G, Givol D, Domany E: **Outcome signature genes in breast cancer: is there a unique set?** *Bioinformatics* 2005, **21**(2):171–178.
30. Michiels S, Koscielny S, Hill C: **Prediction of cancer outcome with microarrays: a multiple random validation strategy**. *Lancet* 2005, **365**(9458):488–492.
31. Yang YC, Piek E, Zavadil J, Liang D, Xie D, Heyer J, Pavlidis P, Kucherlapati R, Roberts AB, Bottinger EP: **Hierarchical model of gene regulation by transforming growth factor beta**. *Proc Natl Acad Sci U S A* 2003, **100**(18):10269–10274.
32. Karlsson G, Liu Y, Larsson J, Goumans MJ, Lee JS, Thorgeirsson SS, Ringnér M, Karlsson S: **Gene expression profiling demonstrates that TGF-beta1 signals exclusively through receptor complexes involving Alk5 and identifies targets of TGF-beta signaling**. *Physiol Genomics* 2005, **21**(3):396–403.
33. Chen CR, Kang Y, Siegel PM, Massague J: **E2F4/5 and p107 as Smad cofactors linking the TGF-beta receptor to c-myc repression**. *Cell* 2002, **110**:19–32.
34. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH: **Gene expression profiling predicts**

clinical outcome of breast cancer. *Nature* 2002, **415**(6871):530–536.

35. Sigurdsson H, Baldetorp B, Borg A, Dalberg M, Ferno M, Killander D, Olsson H: **Indicators of prognosis in node-negative breast cancer**. *N Engl J Med* 1990, **322**(15):1045–1053.
36. Mehlen P, Puisieux A: **Metastasis: a question of life or death**. *Nat Rev Cancer* 2006, **6**(6):449–458.
37. Ringnér M, Veerla S, Andersson S, Staaf J, Häkkinen J: **ACID: a database for microarray clone information**. *Bioinformatics* 2004, **20**(14):2305–2306.
38. Aerts S, Van Loo P, Thijs G, Mayer H, de Martin R, Moreau Y, De Moor B: **TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis**. *Nucleic Acids Res* 2005, **33**(Web Server issue):393–396.
39. Aerts S, Thijs G, Coessens B, Staes M, Moreau Y, De Moor B: **Toucan: deciphering the cis-regulatory logic of coregulated genes**. *Nucleic Acids Res* 2003, **31**(6):1753–1764.
40. Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouze P, Moreau Y: **A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling**. *Bioinformatics* 2001, **17**(12):1113–1122.

Additional Files

The following additional data are available with the online version of this article. A file in tab-delimited format listing the results for all pathways for each gene signature: E2F3 (Additional data file 1), Myc (Additional data file 2), Ras (Additional data file 3), Yang (Additional data file 4), Karlsson (Additional data file 5), and breast cancer prognosis (Additional data file 6).