Probabilistic Methods in Genomic Data Analysis

Thomas Breslin

Department of Theoretical Physics
Lund University, Sweden

Thesis for the Degree of Doctor of Philosophy

Thesis Advisor: Carsten Peterson

Faculty Opponent: Steen Knudsen
Technical University of Denmark

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 theoretical physics on friday, the 17th of december 2004, at 10.30 a.m.

Organization LUND UNIVERSITY Department of Theoretical Physics Sölvegatan 14A SE-223 62 LUND	**Document Name** DOCTORAL DISSERTATION **Date of issue** December 2004 **CODEN:**
Author(s) Thomas Breslin	**Sponsoring organization**

Title and subtitle
Probabilistic Methods in Genomic Data Analysis

Abstract
In this thesis, three aspects of gene expression data analysis are discussed: Differential gene expression is addressed by a probabilistic method. Gene annotation enrichment analysis is discussed in the context of multiple hypothesis testing and the choice of null hypothesis. The possibility of inferring the activity of cellular signaling pathways from microarray data is explored. The methods developed are applied to various data sets. The method for differential gene expression is applied to aspects of B cell differentiation. The methods for annotation analysis and pathway activity inference are applied to data sets of breast cancer, colon cancer and leukemia.

Summary in Swedish
Den nyligen utvecklade microarray-tekniken gör det möjligt att mäta mängden av mRNA för tusentals gener samtidigt. För att tolka dessa data krävs statistiska metoder och denna avhandling behandlar tre sådana. Den första metoden är utvecklad för att skilja ut de gener vars mRNA-mängd är olika stor i olika vävnadstyper. Den andra metoden är avsedd för att analysera och beräkna statistisk signifikans för annoteringar av gener i samband med analys av microarray-data. Den tredje metoden har för ändamål att relatera microarray-data till tidigare biologisk kunskap om signalvägar i cellen. Metoderna appliceras på ett antal olika dataset. Med den första metoden studeras B cellsdifferentiering och de övriga två tillämpas på data för bröstcancer, coloncancer och leukemi.

Key words
Probabilistic methods, microarray, differential gene expression, annotation analysis, pathway profiling.

Classification system and/or index terms (if any)

Supplementary bibliographical information		**Language** English
ISSN and key title		**ISBN** 91-628-6336-3
Recipient's notes	**Number of pages** 100	**Price**
	Security classification	

Distribution by (name and address)
Thomas Breslin, Dept. of Theoretical Physics,
Sölveg. 14A, SE-223 62 Lund

To Anna

This thesis is based on the following papers:

I C.M. Högerkorp, S. Bilke, T. Breslin, S. Ingvarsson and C. Borrebaeck,
CD44 stimulated human B cells express transcripts specifically involved in immunomodulation and inflammation as analyzed by DNA microarrays
Blood, **101**, 2307–2313 (2003).

II P. Tsapogas, T. Breslin, S. Bilke, A. Lagergren, R. Månsson, D. Liberg, C. Peterson and M. Sigvardsson,
RNA analysis of B-cell lines arrested at defined stages of differentiation allows for an approximation of gene expression patterns during B-cell development
Journal of Leukocyte Biology, **74**, 102-110 (2003).

III S. Bilke, T. Breslin and M. Sigvardsson
Probabilistic estimation of microarray data reliability and underlying gene expression
BMC Bioinformatics, **4**, (2003)

IV T. Breslin, P. Edén and M. Krogh
Comparing functional annotation analyses with Catmap
To appear in *BMC Bioinformatics*

V T. Breslin, M. Krogh, C. Peterson, C. Troein
Signal transduction pathway profiling of individual tumor samples
To be submitted

During my years as a PhD student, I have also contributed to the following paper:

M. Sigvardsson, D.R. Clark, D. Fitzsimmons, M. Doyle, P. Åkerblad, T. Breslin, S. Bilke, R. Li, C. Yeamans, G. Zhang and J. Hagman
Early B-cell factor, E2A, and PAX-5 cooperate to activate the early B cell-specific mb-1 promoter
Molecular and Cellular Biology, **22**, 8539–51 (2002)

Contents

Introduction **1**

Differential Gene Expression and Clustering 5

Annotation Enrichment Analysis 7

Pathways and Gene Expression Data 10

The Papers 14

Acknowledgments 16

References 17

Original Papers

I CD44 stimulated human B cells express transcripts specifically involved in immunomodulation and inflammation as analyzed by DNA microarrays

II RNA analysis of B-cell lines arrested at defined stages of differentiation allows for an approximation of gene expression patterns during B-cell development

III Probabilistic estimation of microarray data reliability and underlying gene expression

IV Comparing functional annotation analyses with Catmap

V Signal transduction pathway profiling of individual tumor samples

Introduction

One of the most influential scientific ideas of the twentieth century is the central dogma of molecular biology (Crick, 1970). It states that the transfer of information, in any living cell, is from DNA to RNA and then from RNA to proteins. Hence, the DNA contains the essential information for building and maintaining the cells of an organism. In any given cell, the RNA molecules that are produced indicate which genes are expressed, and thus which proteins that are going to be produced by that cell. Proteins, in turn, are the end point effectors of cellular diversity. The composition of proteins determines the fundamental properties of any cell, and how it responds to its environment; whether it is a neuron of the brain, a hepatocyte of the liver or a leukemic cancer cell of the blood (Alberts *et al.*, 2002).

Underlying the central dogma is the structure of the DNA molecule (Watson and Crick, 1953). It consists of two anti-parallel strands where the basic building blocks; Adenine, Guanine, Cytosine and Thymine, form A-T or C-G pairs with one of the molecules in a pair on each strand of the back bone. The pairs are held together by hydrogen bonds which can be broken if the temperature is increased. When the temperature is lowered, the separated strands rejoin and form a double strand in such a way that the bases are rejoined in the same A-T and C-G pairs. This principle of base pairing along the DNA strand is fundamental for many of the techniques that provide genomic information, especially DNA microarrays.

DNA microarrays, introduced by Schena *et al.* (1995) and Lockhart *et al.* (1996), enable us to intercept the flow of information from DNA to proteins at the RNA level, thus yielding information about which genes that are being used to synthesize proteins in any given tissue, at any moment. The process of obtaining this data is rather involved, for a review see, *e.g.* Cheung *et al.* (1999) or Holloway *et al.* (2002). A schematic outline may be given as follows: First a sample is collected from the tissue of interest. Such a sample typically contains millions of cells, and thus it is important to bear in mind that the obtained data are averages over the cell population. The cells of the sample are then disrupted, RNA is extracted and, if necessary, amplified. By reverse transcription, the RNA is translated into DNA which is labeled with a fluorescent dye. Finally this

DNA (or in some cases RNA copies of it) is hybridized onto a microarray. A microarray is a small solid surface, upon which known genes or gene fragments are deposited or synthesized, at high density, in a spatially ordered manner. The hybridization consists of splitting the double strands of the sample DNA, and then applying the sample to the microarray. After this, the temperature is lowered, and the fluorescently labeled single strands of the sample attach to the immobilized probes of the microarray by base pairing. By measuring the amount of fluorescence at each spot of the array, one obtains a measure of the RNA level of each gene in the sample.

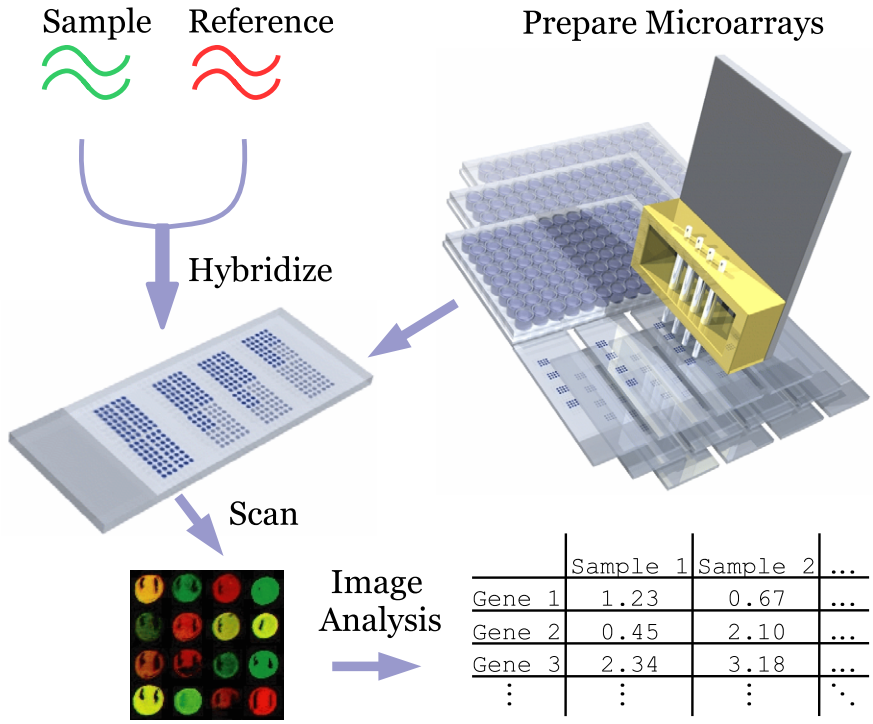


Figure 1: Overview of the cDNA microarray technology. Here, both the sample (labeled with green fluorescent dye) and a reference (labeled with red fluorescent dye) are hybridized onto the same array. The array is scanned, and the amount of green and red fluorescence at each spot is quantified. Finally, a file containing the ratios of green and red fluorescence is created. Adapted in part, with permission, from Hedenfalk (2002).

The introduction of microarrays, with which the expression levels of thousands of genes can be measured in parallel, has extensively improved our knowledge and understanding of how the genome works. The applications of this technique are manifold, and in what follows I shall give a brief account of two of them.

Functional characterization of genes and gene regulation.

One of the first discoveries of gene expression analysis on the whole genome scale was that genes with similar functions show similar expression. This has been confirmed in many studies, and taken as a starting point for inferring the functions of previously uncharacterized genes (Eisen *et al.*, 1998; Chu *et al.*, 1998; Spellman *et al.*, 1998; Brown *et al.*, 2000; Hughes *et al.*, 2000). A plausible biological reason for co-expression of a set of genes is co-regulation. Many studies have thus been devoted to pinpointing the sets of transcription factors, and upstream binding sites, that are responsible for the co-regulation of genes (Holstege *et al.*, 1998; Bussemaker *et al.*, 2000; Jensen and Knudsen, 2000; Keles *et al.*, 2002). More elaborate studies have also taken into account the combinatorial effects of transcription factors (Pilpel *et al.*, 2001), and by combining microarray technology with chromatin immunoprecipitation large parts of the *S. cerevisiae* transcriptional regulatory network have been unraveled (Lee *et al.*, 2002).

Molecular classification of tissues

Molecular classification of tissues is in some respects orthogonal to the functional characterization of genes, even though most studies incorporate both aspects in the analysis of gene expression data. The main assumption is that the state of a tissue, or system of cells, may be characterized by the unique combination of expression levels of the genes harbored by the cells of the tissue. Pioneering work has validated this approach to tissue classification using hierarchical clustering (Eisen *et al.*, 1998; DeRisi *et al.*, 1996). Many different techniques have since been applied and tested. Both unsupervised methods, like hierarchical clustering or self organizing maps (Golub *et al.*, 1999; Tamayo *et al.*, 1999), and supervised methods like support vector machines (Furey *et al.*, 2000) or artificial neural networks (Khan *et al.*, 2001), have been found useful. However, each method has its own advantages and shortcomings, and the choice of method must therefore be determined by the biological problem at hand (Quackenbush, 2001).

Concluding remarks and thesis outline

In the quest of a holistic understanding of cellular function, much remains to be charted. Post-translational modifications of proteins and exact measurements of protein concentrations are beyond what can be inferred from mRNA[1] levels alone (Gygi *et al.*, 1999). The emerging proteomics techniques will thus, doubtlessly, contribute to a better understanding of the cellular machinery (Pandey and Mann, 2000; Mann *et al.*, 2001). However, it has recently been recognized that the functional roles of non-protein-coding RNA may have been gravely overlooked. Thus, a number of amendments to the views held in the light of the central dogma have been suggested (Mattick, 2003). Indeed, the possible applications of DNA microarrays in such research are numerous.

The field of microarray data analysis has grown vast over the past decade. It addresses all issues from image analysis and normalization to biological hypothesis generation by the inclusion of prior knowledge. In what follows, I will discuss my contributions to some of these areas, beginning with methods for purely statistical description of the data, in particular for differential gene expression and clustering. Thereafter, I will discuss statistical considerations of assigning significance to categories of annotation and finally some aspects of how to integrate the data analysis with prior knowledge, in the form of cellular signaling pathways. Each section starts out with a description of the nature of the problem. Then follows a brief review of related research, and finally some aspects of the methods employed and developed in this thesis are discussed. The last section, Papers, gives a brief summary of each of the five papers, with emphasis on findings and conclusions.

[1] As the aim of the introduction is to be general and schematic, I have refrained from using explicit denotations of RNA type until this point. In this section, however, the protein-coding RNA, termed messenger RNA (mRNA) is contrasted to non-protein-coding RNA.

Differential Gene Expression and Clustering

Nature of the problem

A very straightforward question that can be posed with the help of a microarray is which genes differ between two biological *varieties*. Biological varieties are defined by the investigator, and they can be confined at any level of description; neurons versus hepatocytes, cultured *S. cerevisiae* in S versus G0 phase or tumors subjected to one type of treatment versus those subjected to another. To measure gene expression in a variety, a sample must be obtained. Depending on the biological heterogeneity of the variety and the microarray measurement process itself, the obtained values will vary (Schuchhardt *et al.*, 2000). Thus, to separate relevant differences between varieties from irrelevant differences between samples, several samples from each variety are required (Lee *et al.*, 2000). When studying gene expression over many varieties, the problem is extended to grouping genes by the similarity of their expression patterns over the varieties. Such clustering analyses are usually performed using only one sample from each variety but, if many varieties are considered, individual sample deviations are less likely to influence the result.

Related research

A common starting point for many methods for differential gene expression is normalization. Various normalization schemes have been proposed throughout the literature but no method stands out as the final solution to this problem (Quackenbush, 2002). The purpose of normalization is to put every sample on equal footing, and this problem turns out to be even harder when comparing data from different labs or different platforms (Yauk *et al.*, 2004). Some studies indicate that transforming the expression levels of a sample into ranks yields better reproducibility for the samples of a variety (Cheng *et al.*, 2003; Kim *et al.*, 2004). Indeed, this can be expected since any *monotonous* normalization of the data will result in the same ranking. Recently, some methods for detecting differential gene expression, based solely on ranks, have been put forth (Breitling *et al.*, 2004b; Martin *et al.*, 2004). In the method employed in papers I-III, we choose an even cruder discretization of the data by considering only binary expression levels. Working with discrete measures of gene expression lends itself to a very straightforward to approach to assessing differential gene expression and cluster analysis.

Methods employed

In papers I-III we employ and develop a probabilistic method for detecting differential gene expression using discretized gene expression data. The discretization is based on Affymetrix^{TM} Absent/Present calls, and thus our model considers only binary values of expression; 0 or 1. The general framework is however not restricted to binary states, but an extension would require further testing and development. For a variety consisting of n samples, the genes are represented by vectors of ones and zeros $\mathbf{d}_g$, for each gene g. To describe the obtained distribution of expression vectors it is necessary to consider at least two types of genes; genes that are expressed in the variety and genes that are not. If all vectors contain only zeros or ones, then this is sufficient. However, for real data this is rarely the case and to account for vectors with a mixture of ones and zeros we have chosen to consider a third type of gene; one that varies randomly between the samples of a variety. In summary, we consider three types of genes: σ_0 genes which are always zero in any sample, σ_r genes which are randomly zero or one in any sample and σ_1 genes which are always one in any sample. These three types of gene expression will henceforth be referred to as underlying states (of gene expression).

The question to be answered is then: Given the data[2], $\mathbf{d}_g$, for a gene, what underlying state is it in? A simple answer would be to say that if $\mathbf{d}_g$ is only zeros, then it is a σ_0 gene, if it is only ones, then it is a σ_1 gene and if it is a mixture, then it is a σ_r gene. If there are many samples of the variety, one could devise some form of majority vote, for instance $\geq$ 90% zeros or ones makes it a σ_0 or σ_1 gene, respectively, and anything else makes it a σ_r gene. A slightly more sophisticated model would also consider sample specific effects; if one sample consistently contradicts the others, this should be incorporated into the majority voting scheme. In a probabilistic formulation we wish to determine the probability $P(\sigma|\mathbf{d}_g)$ where σ is one of the possible underlying states: σ_0, σ_r, σ_1.

Estimating the sample specific variations, *i.e.*, error probabilities, can be done by modeling the observed distribution of vectors, $\mathbf{d}_g$. In paper III we model this distribution by using parameters for the probability of each underlying state; $P(\sigma_0)$, $P(\sigma_r)$, $P(\sigma_1)$ and sample specific errors $\{P^i_{1\to 0}, P^i_{0\to 1}\}_{i=1}^n$. Once the model is fitted to the observed data the desired probability can be computed for each gene and for each possible underlying state using Bayes' theorem:

$$P(\sigma|\mathbf{d}_g) = \frac{P(\mathbf{d}_g|\sigma)P(\sigma)}{P(\mathbf{d}_g)} \ . \tag{1}$$

Thus we obtain a probabilistic, rather than a static, assignment of a gene to an underlying state.

[2] In paper III this data, $\mathbf{d}_g$, is referred to as S in order to contrast the observed state of a gene to the underlying, *i.e.*, modeled states σ.

To compute the probability a gene being differentially expressed between two varieties A and B, we consider all possible combinations $\sigma^A\sigma^B$ for the gene. The probability of each combination is simply the product of the conditional probabilities of the underlying states in each variety. The probability of the gene being higher in B is thus defined as the sum $P(\sigma_0^A|\mathbf{d}_g^A)P(\sigma_r^B|\mathbf{d}_g^B) + P(\sigma_0^A|\mathbf{d}_g^A)P(\sigma_1^B|\mathbf{d}_g^B) + P(\sigma_r^A|\mathbf{d}_g^A)P(\sigma_1^B|\mathbf{d}_g^B)$. Similarly, we compute the probability that a gene goes down, or remains indifferent, between the two varieties. When considering many varieties, we compute the probability of any expression profile over the varieties, and thus we obtain an inherently probabilistic assignment of genes to expression profiles.

The method as presented so far is based on few assumptions and a very crude, binary, discretization of data. An obvious critique of the model is that transcriptional response is known to be both binary and discretized (Biggar and Crabtree, 2001), and furthermore that expression averages over a heterogeneous cell population will by necessity be continuous. Yet, the method has been proven valid for assessing differential gene expression (papers I to III). Extending it by discretizing the data into more than two levels is rather forthright and deserves further efforts of evaluating and testing.

Annotation Enrichment Analysis

Nature of the problem

The purpose of statistical analysis of gene expression data is often to single out the genes that are relevant to a biological question from those which are not. If the set of relevant genes is small it may be easy, for an experienced investigator, to infer the biological context from which the relevant genes were derived. However, in many analyses the set of relevant genes can be huge. Surveying the literature for all those genes is often too daunting a project. Thus, the investigator is lead to use faster methods such as searching the annotations of the genes for overrepresented keywords. In this section, I will discuss the problem of how to put such annotations enrichment analyses into a solid statistical framework.

The output of gene expression analysis methods often come in two flavors. Some methods output well defined subsets of the genes present on the microarray. A canonical example of this is k-means clustering (Tavazoie *et al.*, 1999). Other methods provide an ordering of the genes, defined by their relevance to the biological question. An example of this is ordering genes based on the p-value of a test of differential expression between two biological varieties. The questions underlying any annotation enrichment analysis are slightly different for the two forms of output. For the former type, one has to ask which annotations that are overrepresented in the set of relevant genes. For the latter, one has

to ask which annotations that are overrepresented towards the top of the ordered gene list and give a precise definition to the meaning of overrepresentation towards the top of the list.

Compilations of contextual gene annotations are steadily growing. Some examples include GenBank annotations, Unigene annotations, SwissProt keywords, enzyme classifications (Bairoch, 2000) and Gene Ontology annotations. The Gene Ontology (Ashburner *et al.*, 2000) is perhaps the most commonly used source of annotation. In addition to providing mere terms of annotation, henceforth referred to as categories[3], it also imposes a tree like structure[4] onto those categories in such a way that less specific categories appear as the mother nodes of more specific categories.

The number of categories applicable to the genes on a microarray can be overwhelming. At the time of writing, the Gene Ontology contains no less than 17,977 distinct categories. This raises the issue of multiple hypothesis testing: If testing a thousand independent hypotheses, fifty of them are expected to be significant, at the 5% level, by chance alone. An excellent review of this matter can be found in Manly *et al.* (2004). For practical reasons though, multiple hypothesis testing may not always be a problem. Often, the investigator is only interested in the ordering of categories, in pursuit of novel hypotheses. Finding that a category such as 'alcohol metabolism' is a highly significant is of no interest in a study separating genes that participate in alcohol metabolism from those that do not. To be useful, the annotation enrichment analysis must provide elucidating or unexpected results that can be tested by other means than the microarray data alone. For this purpose, a ranking of the categories is often sufficient. If however the significant categories seem biologically unreasonable, it may be worthwhile to consider the multiple hypotheses testing issue before starting up a big or expensive experiment in the lab.

Another aspect of hypothesis testing is the choice of null hypothesis. For most data sets there are two obvious choices to consider: A random permutation of sample labels, or a random permutation of gene labels.

Related research

Annotation enrichment analysis goes back to the early days of microarray data analysis (DeRisi *et al.*, 1997; Lashkari *et al.*, 1997; Eisen *et al.*, 1998). In these, and many other studies, the annotations of sets of interesting genes were simply listed in tables. Later methods have brought more statistical rigor to this annotation enrichment analysis. For the case of clearly defined subsets of genes, many authors have employed Fisher's exact test or variants thereof (Zeeberg *et al.*, 2003; Draghici *et al.*, 2003; Beissbarth and

[3] This choice of terminology refers to the annotation as describing a category of genes

[4] In a technical sense, the Gene Ontology is a directed acyclic graph.

Speed, 2004). The null hypothesis for this test is that there is no association between a gene belonging to a certain category, and it being in the subset of relevant genes. The table below illustrates such a situation where the relevant genes are, for example, the set of genes belonging to a certain cluster.

	Metastasis	Non-Metastasis	
Relevant genes	10	90	100
Non-Relevant genes	10	890	900
	20	980	1000

If a subset of 100 relevant genes is selected from a set of 1000 genes and 20 of the 1000 genes belong to the category 'Metastasis', the expected number of genes with this annotation in the relevant subset is $(100/1000) \times 20 = 2$. Thus, in this example, genes of the category 'Metastasis' are clearly overrepresented in the subset. The p-value of this observation is the probability to observe 10 or more 'Metastasis' genes in the relevant subset. This probability is given by

$$p = \sum_{i=10}^{20} \frac{\binom{20}{i}\binom{980}{100-i}}{\binom{1000}{100}}. \tag{2}$$

Similarly, p-values of underrepresentation may be calculated.

For ordered gene lists, where no clear cutoff can be defined, various suggestions have been put forth. Some authors (Berriz *et al.*, 2003; Breitling *et al.*, 2004a) suggest using an optimization procedure to determine the cutoff for each category. This method, however, renders the interpretation of p-values meaningless, and the authors have amended this by using the optimized p-values as scores. Having established the score for a category, they proceed directly to calculate a multiple hypothesis corrected p-value for the score by comparing it to the distribution of scores obtained under the null model of randomly permuted gene lists. A different approach, termed gene set enrichment analysis (GSEA), is presented by Mootha *et al.* (2003). They use a Kolmogorov-Smirnov running sum to score each category, and they then proceed to calculate multiple hypothesis corrected p-values for each category under the null hypothesis of random sample label permutations.

Methods employed

In paper IV, a few novel tools are brought to the table of annotation enrichment analysis methods. For ranked gene lists, each category is assigned a score, based on the Wilcoxon ranksum of the positions of the genes of the category in the ranked list (Wilcoxon, 1945). The p-value of the score is assessed, *for each category individually*, by the fraction of permuted gene lists that yield a better score then the score tested. We have implemented this method in the program Catmap which is freely available for download[5]. The program can take any set of permuted gene lists as input when calculating p-values for the categories. Thus, by generating such lists, a user can employ any null model that he or she sees fit, for example sample label permutation.

We also devise a method for adjusting the p-values for multiple hypothesis testing. This method takes its starting point along the following line of reasoning. Suppose that N independent categories of annotation are tested. Then the probability of obtaining at least one category with a p-value below q is $1-(1-q)^N$. If the categories are not independent, we assume that the probability is given on the same form but with N replaced by an effective number of independent categories, N_{eff}. To estimate this effective number of independent categories we use K permuted gene lists, generated according to the null hypothesis. For each list and all categories, we extract the lowest p-value, yielding the set $\{p_i\}_{i=1}^{K}$. By applying maximum likelihood estimation, we obtain the following relation for N_{eff}:

$$N_{\text{eff}} = \frac{K}{-\sum_{i=1}^{K} \ln(1-p_i)}. \tag{3}$$

The individual category p-values are then adjusted as $p_{adj} = 1-(1-p)^{N_{\text{eff}}}$. For small p-values this is similar to a Bonferroni correction, *i.e.*, $p_{adj} = p \cdot N_{\text{eff}}$ (Bonferroni, 1936). The K permuted gene lists are also used to obtain a false discovery rate. The false discovery rate for the j highest ranked categories is obtained as follows: Given the p-value of the j:th highest ranked category, p_j, and the set of p-values for all categories in the K permuted lists, S_K, the false discovery rate is defined as the number of p-values in S_K that are smaller than p_j, divided by $K \cdot j$.

Pathways and Gene Expression Data

Nature of the problem

It is reasonable to hypothesize that prior contextual knowledge, in the form of pathways, can be used to gain insight into the regulatory mechanisms underlying patterns of gene expression data. The notion of pathways is however rather broad. For example, a differentiation pathway refers to the sequence of events, by which an unspecialized precursor

[5] `http://bioinfo.thep.lu.se/Catmap`

cell transforms into a fully differentiated cell with specialized functions. A metabolic pathway refers to a series of enzyme-catalyzed biochemical reactions in a cell. The series is defined in such a way that the product of one reaction is the substrate of the next reaction. In paper V, we have chosen to study cellular signaling pathways, which are given by extra-cellular signaling molecules (ligands) that activate receptors of the cell. Activated receptors then initiate intracellular signaling events, which eventually regulate the activity of various transcription factors. These transcription factors, in turn, regulate the expression levels of various genes, termed downstream targets of the pathway.

To gauge the activity of cellular signaling pathways, both proteomic and gene expression data would be desirable. For the three cancer data sets studied in paper V (Golub *et al.*, 1999; van 't Veer *et al.*, 2002; Sotiriou *et al.*, 2003), and indeed for many other data sets, only the gene expression data is available. Since post-translational modifications of proteins, such as phosphorylation and methylation, can not directly be inferred from gene expression data and, furthermore, protein concentrations are not perfectly correlated to mRNA levels, we have chosen to rely on the expression level of downstream targets alone when gauging pathway activity. Although straightforward in its interpretation, there are some drawbacks of this crude measurement. Many pathways overlap in terms of which transcription factors they regulate, and many transcription factors overlap in terms of which genes they regulate. These facts have to be accounted for when interpreting the results of pathway activity measurements.

Cellular signaling pathway activity can be characterized from two principally different points of view. From the point of view of the entire data set, one may measure the degree of co-expression of downstream targets as an indicator of differential pathway activity. From the point of view of each individual sample, one may measure the relative amount of upregulation (or downregulation) of the downstream targets, in relation to the other samples of the data set, and take this measure as indicative of the relative pathway activity level in that sample.

Related research

In a broader perspective, cellular signaling pathways are only one specific instance of gene regulatory systems. Many aspects of gene regulatory systems have been discussed throughout the literature. Modeling the dynamics of gene regulation is reviewed in, *e.g.*, de Jong (2002), and explicitly explored for the yeast transcriptional network (Kauffman *et al.*, 2003). Studies on the topology of the gene regulatory networks are presented, e.g., in Shen-Orr *et al.* (2002) and Lee *et al.* (2002).

A common starting point, when inferring co-regulation of genes, is the expectation that co-regulated genes exhibit similar expression over a set of samples. In the context of metabolic networks, this hypothesis is utilized by Rahnenführer *et al.* (2004), who devise

various scores of average co-regulation for all pairs of non-identical genes in a pathway. One of the earlier studies employing measures of co-regulation to assign statistically significant scores to biologically relevant pathways is presented in Zien *et al.* (2000). The methods of Zien *et al.* (2000) and Rahnenführer *et al.* (2004) both take their starting point in fixed sets of genes representing putatively active pathways. A more flexible approach is given by Vert and Kanehisa (2003), whose starting point is a network of genes. Given the expression profiles of those genes in a data set, they search for regularities in the expression profiles of genes with respect to the network topology and are thus able to extract pathways that are differentially active over the samples of the data set.

Compilations of metabolic and signal transduction pathways in the form of databases are steadily growing. Some examples are the KEGG (Kanehisa and Goto, 2000) and LIGAND (Goto *et al.*, 2000) databases, the STKE database (Gough, 2002), and the TRANSPATH (Krull *et al.*, 2003) and TRANSFAC (Wingender *et al.*, 2001) databases, which are used in paper V.

Methods employed

The three cancer data sets studied in paper V are normalized in the following way: First, each sample, i, of a data set is centered so that its mean expression is zero. Letting $x_{g,i}$ represent the expression level of gene g in sample i, normalized samples will fulfill $\sum_g x_{g,i} = 0$. Second, each gene is centered over the samples of a data set, *i.e.*, $\sum_i x_{g,i} = 0$. This ensures that the expression of a gene is measured in relation to its mean over all the samples.

In assessing differential pathway activity over the samples of a data set, we use the standard Pearson correlation coefficient r. For two genes, g and h, it is defined by:

$$r_{g,h} = \frac{\sum_i (x_{gi} - \bar{x}_g)(x_{hi} - \bar{x}_h)}{\sqrt{\sum_i (x_{gi} - \bar{x}_g)^2} \sqrt{\sum_i (x_{hi} - \bar{x}_h)^2}} \,, \tag{4}$$

where $\bar{x}_g$ refers to the mean expression of gene g over the samples.

We use two slightly different scores when measuring differential pathway activity over the samples of a data set. Both scores are based on summing the values of $r^2_{g,h}$ for pairs of downstream targets of the pathway. The first score uses all such pairs, and the second, more restrictive score, uses only pairs of downstream targets that do not share the same transcription factor. For both scores, the significance is calculated based on random permutation of genes in the data set.

To measure the activity of a pathway in the individual samples of a data set, we devise a score that is simply the sum of normalized expression levels $x_{g,h}$ for all downstream targets of the pathway. A significance is given to the score based on the fraction of cases, under random permutation of genes, for which a higher or lower score is obtained. These

fractions define two p-values, p_+ and p_-, from which the overall p-value is defined as $2 \times \min\{p_+, p_-\}$. The pathway is said to be active if $p_+ < p_-$; and inactive otherwise.

To analyze the association between the individual sample pathway activity, and clinical variables of the data sets, we use contingency tables. For every pathway and data set, we divide the samples into three groups: Samples where the pathway is active at a 5% significance level, samples where it is inactive at a 5% significance level and samples where it is not significant. For each of these contingency tables, we then calculate p-values using a χ^2 test.

The Papers

Papers I and II

In these papers we apply the probabilistic estimation of microarray data reliability and underlying gene expression to different biological problems. In both papers, expression data from Affymetrix GeneChip arrays is discretized into two levels based on the Affymetrix Absent/Present call. Paper I is a study of the effect of CD44 ligation on cultured human B cells. In all, the study comprises six biological varieties; stimulated and non-stimulated cells after 6, 24, and 72 hours in culture. Each variety contains four samples from different human donors. The analysis discerns the temporal effects of CD 44 ligation and eventually novel functions of CD 44 ligation are suggested by the results. Paper II is a study of stage specific gene expression in maturing murine B lymphocytes. Here, the four varieties are four specific stages of B cell development: pro B cells, pre B cells, mature B cells and plasma cells. Each variety again contains four samples which are derived from standard cell lines, arrested at the corresponding stage of development. In this study, the probabilistic analysis is accompanied by a dCHIP analysis and RT-PCR confirmation of some of the results. It may be noted that out of 37 well known control genes, 10 were misclassified by the dCHIP algorithm and 5 by the probabilistic method.

Paper III

Paper III contains a formal description of the method used in papers I and II. The accuracy of parameter estimation is tested with synthetically generated data, and it is found to be valid in the parameter ranges applicable to typical data sets. Using synthetically generated data we are able test the effect of correlations between samples of a variety. Thus, assumption that σ_r genes are represented by *random* vectors of zeros and ones is scrutinized, and we find that for correlations below 0.2, the parameter estimation is still reliable. Furthermore, we compare the probabilistic method to a standard t-test using the same data as in paper II. We find that on a set of known control genes, the probabilistic method is at least comparable to the t-test despite the crude discretization at only two levels.

Paper IV

Paper IV is a study of how the choice of score function, and null hypothesis, influence the significance of annotations for ordered gene lists. Analyzing three publically available data sets (Alon *et al.*, 1999; Golub *et al.*, 1999; van 't Veer *et al.*, 2002), and ordering genes according to their absolute Pearson correlation to sample labels, we find, as expected, that the significance assigned by a cutoff based score function is highly dependent on

the cutoff parameter. Thus, for ordered gene lists, it is preferable to use a parameter free score function such as the Wilcoxon rank sum or the Kolmogorov-Smirnov statistic. Furthermore, we conclude that a random permutation of gene labels often results in lower p-values than random permutation of sample labels. We also find that the highest ranking annotations (based on p-value) are in some respects similar, but in other respects dissimilar between the two different null hypotheses. The reasons for this are further elucidated in the paper. Finally, we present a method for estimating the number of independent annotations, and we show that this estimation is supported by the data.

Paper V

In paper V we study signal transduction pathway activity in three publically available cancer data sets. Using pathways and downstream targets obtained from the TRANSPATH and TRANSFAC databases, augmented with downstream targets of the estrogen–estrogen receptor complex, we verify that many pathways are significantly differentially regulated over the samples of the data sets. The estrogen–estrogen receptor downstream targets are found to be significantly differentially regulated in the breast cancer data sets of van 't Veer *et al.* (2002) and Sotiriou *et al.* (2003), but not in the leukemia data set of Golub *et al.* (1999). We also devise an individual sample pathway activity score. Using this score we find many pathways for which the number of significantly regulated samples is much higher than expected by chance. Finally, we investigate the association between sample specific pathway activity and clinical variables. We find a number of clinical variables that are strongly associated to pathway activity. Furthermore, our results indicate that the sensitivity of the association decreases when lowering the p-value cutoff for pathway activity while the specificity of the association increases.

Acknowledgments

Many people are entitled to their due and proper in this section, and any omissions are caused by lack of time (or sleep), not by lack of appreciation. That having been said, I would like to start out by thanking my supervisor, Carsten Peterson,* for having given me a piece of the action, and guidance along my path, in the field of computational biology. During little over four years of research, I have had the opportunity to work closely with a number of eminent fellows. I will credit them here in name order; Sven Bilke, my first research partner and part time supervisor, Patrik Edén, an authority on scientific writing, Carl-Magnus Högerkorp, the CD44 expert, Morten Krogh, a role model in his attitude to science, Mikael Sigvardsson, the B cell guru and Carl Troein, master of real-time implementation – and Quake.

In the course of my research, I have had the privilege to work in stimulating and scientifically challenging environment. For this, I owe many thanks to the entire computational biology group, and to our head* of complex systems – for constantly driving us to excel. Being the first member of this group to write a thesis on genomic data analysis, the formulation of this introduction has not been without snags. Thus, I am most grateful to all the people who have helped me, especially Carl, Morten and Patrik (mentioned above), and Ewa Breslin, who have carefully proofread the entire manuscript.

On account of the social pleasantries throughout my stay at this department, I would like to thank all members of staff, past and present, for the nice atmosphere and the many opportunities of celebration. On account of sporting events, for which ample opportunities have been provided, both the badminton team and the old golf team (Jönsson, 2001), have made my stay here healthy and pleasant. I would also like to thank my roommates over the years for the many on-hours of socialization and discussion. For the on-, off- and any-hours of socialization, discussion and indeed many other things, I thank my good friends.

Finally, I thank my family for their love and support, my mother and father for endorsing my studies and my grandmother for lighting the flame of my passion for science. Last, but most importantly, I thank you Anna, for bearing with me through thick and thin and helping me onto the path that led to this thesis.

References

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). *Molecular Biology of the Cell*. Garland Publishing, fourth edition.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A*, 96(12):6745–50.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–29.

Bairoch, A. (2000). The enzyme database in 2000. *Nucleic Acids Res*, 28(1):304–305.

Beissbarth, T. and Speed, T. P. (2004). Gostat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465.

Berriz, G. F., King, O. D., Bryant, B., Sander, C., and Roth, F. P. (2003). Characterizing gene sets with funcassociate. *Bioinformatics*, 19(18):2502–2504.

Biggar, S. R. and Crabtree, G. R. (2001). Cell signaling can direct either binary or graded transcriptional responses. *EMBO J*, 20(12):3167–76.

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze.*, 8:3–62.

Breitling, R., Amtmann, A., and Herzyk, P. (2004a). Iterative group analysis (iga): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. *BMC Bioinformatics*, 5(1):34.

Breitling, R., Armengaud, P., Amtmann, A., and Herzyk, P. (2004b). Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett*, 573(1-3):83–92.

Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M., and Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci U S A*, 97(1):262–267.

Bussemaker, H. J., Li, H., and Siggia, E. D. (2000). Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc Natl Acad Sci U S A*, 97(18):10096–100.

Cheng, C., Kimmel, R., Neiman, P., and Zhao, L. P. (2003). Array rank order regression analysis for the detection of gene copy-number changes in human cancer. *Genomics*, 82(2):122–129.

Cheung, V. G., Morley, M., Aguilar, F., Massimi, A., Kucherlapati, R., and Childs, G. (1999). Making and reading microarrays. *Nat Genet*, 21(1 Suppl):15–19.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., and Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast. *Science*, 282(5389):699–705.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(258):561–563.

de Jong, H. (2002). Modeling and simulation of genetic regulatory systems: a literature review. *J Comput Biol*, 9(1):67–103.

DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A., and Trent, J. M. (1996). Use of a cdna microarray to analyse gene expression patterns in human cancer. *Nat Genet*, 14(4):457–60.

DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278(5338):680–686.

Draghici, S., Khatri, P., Martins, R. P., Ostermeier, G. C., and Krawetz, S. A. (2003). Global functional profiling of gene expression. *Genomics*, 81(2):98–104.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, 95(25):14863–14868.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–14.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537.

Goto, S., Nishioka, T., and Kanehisa, M. (2000). Ligand: chemical database of enzyme reactions. *Nucleic Acids Res*, 28(1):380–382.

Gough, N. R. (2002). Science's signal transduction knowledge environment: the connections maps database. *Ann N Y Acad Sci*, 971:585–587.

Gygi, S. P., Rochon, Y., Franza, B. R., and Aebersold, R. (1999). Correlation between protein and mrna abundance in yeast. *Mol Cell Biol*, 19(3):1720–30.

Hedenfalk, I. (2002). *Gene expression profiling of hereditary breast cancer*. PhD thesis, Lund University, Department of oncology, Lund, Sweden.

Holloway, A. J., van Laar, R. K., Tothill, R. W., and Bowtell, D. D. (2002). Options available–from start to finish–for obtaining data from dna microarrays ii. *Nat Genet*, 32 Suppl:481–489.

Holstege, F. C., Jennings, E. G., Wyrick, J. J., Lee, T. I., Hengartner, C. J., Green, M. R., Golub, T. R., Lander, E. S., and Young, R. A. (1998). Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, 95(5):717–28.

Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M., and Friend, S. H. (2000). Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–26.

Jensen, L. J. and Knudsen, S. (2000). Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation. *Bioinformatics*, 16(4):326–33.

Jönsson, H. (2001). *Variational methods in combinatorial optimization and phylogeny reconstruction*. PhD thesis, Lund University, Department of theoretical physics, Lund, Sweden.

Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30.

Kauffman, S., Peterson, C., Samuelsson, B., and Troein, C. (2003). Random boolean network models and the yeast transcriptional network. *Proc Natl Acad Sci U S A*, 100(25):14796–14799.

Keles, S., van der Laan, M., and Eisen, M. B. (2002). Identification of regulatory elements using a feature selection method. *Bioinformatics*, 18(9):1167–75.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., and Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med*, 7(6):673–679.

Kim, B. S., Rha, S. Y., Cho, G. B., and Chung, H. C. (2004). Spearman's footrule as a measure of cdna microarray reproducibility. *Genomics*, 84(2):441–448.

Krull, M., Voss, N., Choi, C., Pistor, S., Potapov, A., and Wingender, E. (2003). Transpath: an integrated database on signal transduction and a tool for array analysis. *Nucleic Acids Res*, 31(1):97–100.

Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., Brown, P. O., and Davis, R. W. (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc Natl Acad Sci U S A*, 94(24):13057–62.

Lee, M. L., Kuo, F. C., Whitmore, G. A., and Sklar, J. (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cdna hybridizations. *Proc Natl Acad Sci U S A*, 97(18):9834–9839.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J. B., Volkert, T. L., Fraenkel, E., Gifford, D. K., and Young, R. A. (2002). Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298(5594):799–804.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, 14(13):1675–80.

Manly, K. F., Nettleton, D., and Hwang, J. T. (2004). Genomics, prior probability, and statistical tests of multiple hypotheses. *Genome Res*, 14(6):997–1001.

Mann, M., Hendrickson, R. C., and Pandey, A. (2001). Analysis of proteins and proteomes by mass spectrometry. *Annu Rev Biochem*, 70:437–73.

Martin, D. E., Demougin, P., Hall, M. N., and Bellis, M. (2004). Rank difference analysis of microarrays (rdam), a novel approach to statistical analysis of microarray expression profiling data. *BMC Bioinformatics*, 5(1):148.

Mattick, J. S. (2003). Challenging the dogma: the hidden layer of non-protein-coding rnas in complex organisms. *Bioessays*, 25(10):930–939.

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D., and Groop, L. C. (2003). Pgc-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, 34(3):267–73.

Pandey, A. and Mann, M. (2000). Proteomics to study genes and genomes. *Nature*, 405(6788):837–46.

Pilpel, Y., Sudarsanam, P., and Church, G. M. (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat Genet*, 29(2):153–159.

Quackenbush, J. (2001). Computational analysis of microarray data. *Nat Rev Genet*, 2(6):418–27.

Quackenbush, J. (2002). Microarray data normalization and transformation. *Nat Genet*, 32 Suppl:496–501.

Rahnenführer, J., Domingues, F. S., Maydt, J., and Lengauer, T. (2004). Calculating the statistical significance of changes in pathway activity from gene expression data. *Statistical Applications in Genetics and Molecular Biology*, 3(1).

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–70.

Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H., and Herzel, H. (2000). Normalization strategies for cdna microarrays. *Nucleic Acids Res*, 28(10):E47.

Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of escherichia coli. *Nat Genet*, 31(1):64–68.

Sotiriou, C., Neo, S. Y., McShane, L. M., Korn, E. L., Long, P. M., Jazaeri, A., Martiat, P., Fox, S. B., Harris, A. L., and Liu, E. T. (2003). Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci U S A*, 100(18):10393–10398.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, 9(12):3273–97.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A*, 96(6):2907–12.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. (1999). Systematic determination of genetic network architecture. *Nat Genet*, 22(3):281–285.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R., and Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536.

Vert, J. P. and Kanehisa, M. (2003). Extracting active pathways from gene expression data. *Bioinformatics*, 19 Suppl 2:II238–II244.

Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1:80–83.

Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhauser, R., Pruss, M., Schacherer, F., Thiele, S., and Urbach, S. (2001). The transfac system on gene expression regulation. *Nucleic Acids Res*, 29(1):281–283.

Yauk, C. L., Berndt, M. L., Williams, A., and Douglas, G. R. (2004). Comprehensive comparison of six microarray technologies. *Nucleic Acids Res*, 32(15):e124.

Zeeberg, B. R., Feng, W., Wang, G., Wang, M. D., Fojo, A. T., Sunshine, M., Narasimhan, S., Kane, D. W., Reinhold, W. C., Lababidi, S., Bussey, K. J., Riss, J., Barrett, J. C., and Weinstein, J. N. (2003). Gominer: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*, 4(4):R28.

Zien, A., Kuffner, R., Zimmer, R., and Lengauer, T. (2000). Analysis of gene expression data with pathway scores. *Proc Int Conf Intell Syst Mol Biol*, 8:407–17.

Paper I

CD44-stimulated human B cells express transcripts specifically involved in immunomodulation and inflammation as analyzed by DNA microarrays

Carl-Magnus Högerkorp, Sven Bilke, Thomas Breslin, Sigurdur Ingvarsson, and Carl A. K. Borrebaeck

A number of studies have implicated a role for the cell surface glycoprotein CD44 in several biologic events, such as lymphopoiesis, homing, lymphocyte activation, and apoptosis. We have earlier reported that signaling via CD44 on naive B cells in addition to B-cell receptor (BCR) and CD40 engagement generated a germinal center–like phenotype. To further characterize the global role of CD44 in B differentiation, we examined the expression profile of human B cells cultured in vitro in the presence or absence of CD44 ligation, together with anti-immunoglobulin (anti-Ig) and anti-CD40 antibodies. The data sets derived from DNA microarrays were analyzed using a novel statistical analysis scheme created to retrieve the most likely expression pattern of CD44 ligation. Our results show that genes such as interleukin-6 (IL-6), IL-1α, and β_2-adrenergic receptor (β_2-AR) were specifically up-regulated by CD44 ligation, suggesting a novel role for CD44 in immunoregulation and inflammation. (Blood. 2003;101:2307-2313)

Introduction

B-cell differentiation is highly regulated by components in the surrounding microenvironment[1] and is a central process in the humoral immune response. This often involves germinal center reactions eventually leading to population of the memory compartment as well as to generation of plasma cells.[2] Important players in this process are, for example, adhesion molecules from the families of integrins, selectins, and immunoglobulins, as well as chemoattractants, such as the chemokines.[3]

The cell surface glycoprotein CD44 is a member of the hyaladherin or link protein superfamily (LPSF) that interacts with the polysaccharide hyaluronan (HA) in the extracellular matrix (ECM). It is widely distributed in the body and mediates cell-cell and cell-matrix interactions. In the hematopoietic system CD44 is expressed on all cell types and has been shown to play a role in lymphopoiesis and lymphocyte homing[4] as well as in lymphocyte activation[5,6] and apoptosis.[6-7] CD44 has furthermore been associated with several different pathologic states where the linkage to cancer and autoimmune diseases is the most notable.[4,8]

All mechanisms supporting B-cell differentiation from a mature naive B cell to an immunoglobulin (Ig)–producing plasma cell are still not entirely understood, although many components have been identified.[2] We and others have previously investigated the role of CD44 in the regulation of T-cell–dependent B-cell activation[9] and subsequent germinal center formation,[10] where ligation of CD44 was shown to contribute to the induction of a phenotype closely resembling a germinal center (GC) B cell.

B cells up-regulate CD44 upon activation,[9,11,12] and HA-CD44 interactions induce activation of mature B cells in vivo.[5] However, CD44 is down-regulated during the germinal center reaction.[13-15] Interestingly, ectopic GC-like structures are found in several autoimmune and inflammatory diseases,[16,17] indicating that a microenvironment promoting formation of GC-like follicular structures is generated in these pathologic states. Apart from T cells and follicular dendritic cells at these GC-like structures,[18,19] an increase in expression levels of CD44 was evident.[20,21]

To further elucidate the functional effects of CD44 ligation on B cells, we assessed the transcriptional profiles from anti-CD44-stimulated naive B cells also costimulated via CD40 and the B-cell receptor (BCR). The transcriptional profile of approximately 6800 genes was evaluated by using a high-density DNA microarray technique. To facilitate the analysis of the complex patterns in the data set, we developed a novel statistical analysis scheme that accounts for inherent errors in the experimental handling and process often seen in these types of studies.[22,23] In summary, the CD44-dependent regulated genes in our analysis were found to mainly pertain to proteins involved in inflammation and immunomodulation. Interestingly, the expression patterns regulated by CD44 fall into 2 groups: (1) genes augmented or repressed by CD44 in a temporal fashion and (2) genes directly regulated—that is, induced—by CD44 alone. Our results suggest a role for CD44 in the modulation of the mature B-cell response and may also have implications in inflammatory and autoimmune diseases, because genes such as the β_2-adrenergic receptor for norepinephrin, platelet-derived endothelial cell growth factor, as well as interleukin-6 (IL-6) and IL-1α were induced by CD44.

Materials and methods

Antibodies

R-phycoerythrin (RPE)–conjugated anti-CD38 antibodies were obtained from Biosciences (BD; San Jose, CA). Anti-IgD-fluorescein isothiocyanate

From the Department of Immunotechnology, Lund University, Sweden; and Department of Complex Systems, Lund University, Sweden.

Submitted June 21, 2002; accepted October 29, 2002. Prepublished online as *Blood* First Edition Paper, October 31, 2002; DOI 10.1182/blood-2002-06-1837.

Supported by Cancerfonden.

Reprints: Carl A. K. Borrebaeck, Department of Immunotechnology, Lund University, PO Box 7031, SE-220 07 Lund, Sweden; e-mail: carl.borrebaeck@immun.lth.se.

The publication costs of this article were defrayed in part by page charge payment. Therefore, and solely to indicate this fact, this article is hereby marked "advertisement" in accordance with 18 U.S.C. section 1734.

(FITC) and anti-CD3–phycoerythrin (PECy5) was obtained from Dako (Glostrup, Denmark), and anti-CD38-PECy5 antibody was purchased from PharMingen (San Diego, CA). Mouse antihuman IgM (AF6) and mouse antihuman CD44 (BU52) antibodies were kindly provided by I. MacLennan (University of Birmingham, United Kingdom). Mouse antihuman CD40 (S2C6) was a gift from S. Pauli (Stockholm University, Sweden).

Cells

Human tonsils were obtained from 4 different pediatric patients undergoing routine tonsillectomy at the University Hospitals of Lund or Malmö. Briefly, tonsils were minced and T cells were removed by rosetting with neuraminidase-treated sheep red blood cells. Mononuclear T-cell–depleted cells were isolated by density centrifugation using Ficoll-Isopaque (Amersham Pharmacia Biotech, Uppsala, Sweden). The interphase fraction, containing predominantly B cells, was washed in phosphate-buffered saline (PBS) containing fetal bovine serum (FBS) (10%) and incubated with precoated anti-CD38 Dynabeads, sheep antimouse IgG, or pan-mouse IgG (Dynal Biotech, Oslo, Norway) for 45 minutes on ice. Cells depleted for CD38 were stained with anti-CD38-PECy5, anti-CD3-PECy5, and anti-IgD-FITC antibodies. Positive selection using flow cytometric cell sorting of $IgD^+/CD38^-$ B cells to a purity exceeding 98% was performed on a FACS Vantage SE cell sorter (BD).

Cell culture condition

$IgD^+/CD38^-$ B cells (2.0×10^6) from the 4 different human donors were separately cultured at 37°C, 5% CO_2, for a total of 6 hours, 24 hours, and 72 hours on 1.5×10^5 CD32-transfected fibroblasts (162CG7) with 0.2 μg/mL anti-IgM (AF6) and 1.0 μg/mL anti-CD40 (S2C6) antibodies, with or without 1.0 μg/mL anti-CD44 (BU52) antibodies. The culture was performed in flat-bottomed 24-well plates (Costar, Corning, NY) using complete medium: RPMI 1640 supplemented with 2 mM L-glutamine, 1% nonessential amino acids, 50 μg/mL gentamicin (Gibco Life Technologies, Gaithersburg, MD), and 10% FBS (Hyclone Laboratories, Logan, UT).

Isolation of mRNA

Freshly isolated cells from each separate culture were lysed in Trizol (Life Technologies). The RNA was extracted from the cell lysate by adding 0.2 vol chloroform. The aqueous phase containing the RNA was separated and subsequently precipitated with isopropanol and washed in 75% ethanol. The RNA pellet was dissolved in diethylene pyrocarbonate (DEPC)-treated water and further purified with the RNeasy Mini Kit (QIAGEN, Hilden, Germany). The total RNA content was assessed spectrophotometrically (GeneQuant II, Amersham, Pharmacia Biotech) within a 260/280 nm OD (optic density) ratio of 1.9:2.1. After a second precipitation step in 2.5 vol ethanol and subsequent wash, the RNA was resuspended in DEPC-treated water. Five micrograms of total RNA was used for the cDNA and cRNA synthesis, as previously described[24]; cRNA is quality controlled by gel electrophoresis.

Hybridization and scanning of the DNA chips

A hybridization cocktail was prepared with the biotinylated and fragmented cRNA at 50 μg/mL, as described previously,[24] and hybridized onto HuGeneFL microarrays (Affymetrix, Santa Clara, CA). The probe array was then stained with a solution of 2 mg/mL acetylated bovine serum albumin (BSA) and 10 μg/mL streptavidin R-phycoerythrin (Molecular Probes, Eugene, OR). A secondary stain was performed with acetylated BSA, normal goat IgG (Sigma Chemical, St Louis, MO), and biotinylated goat antistreptavidin antibody (Vector Laboratories, Burlingame, CA) for amplification. A final staining step with streptavidin R-phycoerythrin was performed before the probe arrays were scanned in the gene array scanner and checked using the Micro Array Suite 4.0 (Affymetrix), as described previously.[24] Several controls assessing the overall processing, the hybridization, and the quality of the material are included on the microarray. A total of 6 arrays were run for every donor, one for every time point for both the anti-CD44–stimulated and nonstimulated cells. In total, 24 arrays were evaluated.

Statistical analysis

A brief review of the mathematical details of the statistical data analysis is given here, and a more detailed description is available elsewhere (S.B., T.B., and M. Sigvardsson, unpublished data, 2002). The data consist of 6 different biologic varieties: anti-CD44-treated cells at 6, 24, and 72 hours after the onset of treatment and untreated cells sampled at the same time points. Each variety consists of 4 samples—that is, from the 4 different human donors. For each variety, data are discretized using the Affymetrix "present/absent" calls so that each gene is represented by a vector $S \in \{0,1\}^4$. The distribution of observed states S in the variety is assumed to be generated by 3 underlying biologic states: σ_0 for expression below the detection level of the chips, σ_1 for expression above detection level, and σ_T for genes with varying expression levels in the 4 samples. The latter state is assumed to give rise to random vectors S with equal probabilities of 0 or 1 at each position. For each measurement, a certain noise characteristic is assumed. This noise characteristic is modeled by misclassification probabilities $\{P^i_{0\to1}P^i_{1\to0}\}^4_{i=1}$.

To simplify the notation, we introduce the S dependent variables:

$$\rho^i_{1\to0} \equiv P^i_{1\to0}\delta_{S_i,0} + (1 - P^i_{1\to0})\delta_{S_i,1}$$

$$\rho^i_{0\to1} \equiv P^i_{0\to1}\delta_{S_i,1} + (1 - P^i_{0\to1})\delta_{S_i,0}$$

The distribution of observed states S is then modeled by

$$P(S) = P(\sigma_1)\Pi^4_{i=1}\rho^i_{1\to0} + P(\sigma_0)\Pi^4_{i=1}\rho^i_{0\to1} + P(\sigma_T)\Pi^4_{i=1}\frac{1}{2}(\rho^i_{1\to0} + \rho^i_{1\to0})$$

This distribution is fitted to the observed data by χ^2 minimization of the unweighted errors. With the parameter estimates generated from the fit, we may now calculate the belief in terms of probability for an underlying state given the observed one. $P(\sigma|S) = [P(S|\sigma)P(\sigma)]/P(S)$ where $\sigma \in \{\sigma_0,\sigma_T,\sigma_1\}$. For the $3^6 = 729$ possible expression profiles over the whole set of varieties $v \in \{6h+, 24h+, 72h+, 6h-, 24h-, 72h-\}$, we calculate the probability of each one of them by $\Pi\, P(\sigma^v|S^v)$.

For simplicity, the underlying states σ_0, σ_T, and σ_1 are throughout the paper referred to by their index symbols 0, T, and 1. Further, the expression profile of a gene over the 6 varieties will be denoted, for example, 0T1000, where the position of 0, T, or 1 in this pattern refers to the expressional state in the variety 6h+, 24h+, 72h+, 6h−, 24h−, 72h−, respectively.

Flow cytometry and ELISA

Freshly isolated cells from each separate culture were assayed for IL-1α surface expression by flow cytometry using a FACScan (BD) and for IL-6 expression in the supernatant using enzyme-linked immunosorbent assay (ELISA). Anti-IL-1α-PE was purchased from PharMingen, while the human IL-6 ELISA kit was purchased from R&D Systems (Minneapolis, MN).

Results

To study the transcriptional changes associated with a CD44 stimuli on naive B cells, the entire cell population from ectomized pediatric tonsils was fractionated and a naive B-cell subset, defined by immunoglobulin (Ig) D cell surface expression and a lack of CD38, was collected. These cells were propagated in the presence of anti-CD40 and anti-IgM antibodies and immobilized on CD32-transfected fibroblasts.[25] The phenotypic and functional changes then attributed to a CD44 stimulation were evaluated at 6, 24, and 72 hours over 4 separate samples using high-density DNA microarrays displaying probes for approximately 6800 genes and several hundred expressed sequence tags (ESTs).

To investigate the sample coherence, we constructed a Hamming distance matrix of all samples (Figure 1). This validation clearly shows how samples labeled 1 in both treated and untreated cells at 72 hours deviate from the other samples. The

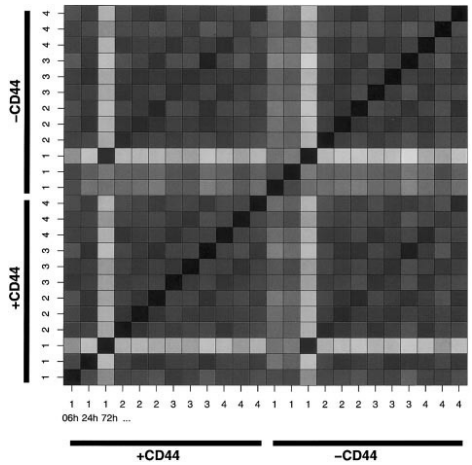


Figure 1. Hamming distance matrix visualizing the number of genes differing between any 2 samples. The color scale ranges from black, indicating no difference, to white, indicating a 2828-gene difference. The ordering is as follows (from the lower left corner up and right): CD44-stimulated (indicated as +CD44) sample no. 1 at 6, 24, and 72 hours; CD44-stimulated sample no. 2, at 6, 24, and 72 hours; sample no. 3 at 6, 24, and 72 hours; and sample no. 4 at 6, 24, and 72 hours. In a corresponding series is the control receiving no CD44 stimuli (indicated -CD44).

deviation is also noted as raised misclassification probabilities for these samples (data not shown). Because the algorithm is designed to handle deviations even of this magnitude, the samples were retained.

To assess the difference in gene expression induced by CD44 ligation, the genes were clustered according to their most probable expression profile over the 6 varieties. These expression profiles can be viewed as 6-letter strings composed by 0, T, or 1 in the following order: 6h+, 24h+, 72h+, 6h−, 24h−, 72h−. For example, a gene with most probable expression profile 01T000 is below the detection level of the chip at 6 hours in the treated group, above the detection level at 24 hours in the treated group, and transient (varying) in the treated group at 72 hours. However, the expression level is below the detection level at all times in the untreated group.

To select genes directly regulated by CD44 (ie, induced or repressed by CD44) at all times, we extracted those having an expression pattern matching (1/T, 1/T, 1/T, 0, 0, 0) and (0, 0, 0, 1/T, 1/T, 1/T). To select temporally regulated genes (ie, genes having an earlier or delayed induction by CD44), we extracted those having an expression pattern matching (1/T, 1/T/0, 1/T/0, 0, 1/T/0, 1/T/0) and (0, 1/T/0, 1/T/0, 1/T, 1/T/0, 1/T/0). The selected clusters were further refined by applying a ranking of the genes having the highest appearance frequency within each specific pattern. In this way the most significant members were assessed. Tables 1 and 2 show genes directly regulated by CD44, and Tables 3 and 4 show genes temporally regulated by CD44.

The ligation of CD44 on a naive B-cell population was recently shown by us to partially induce a germinal center phenotype,[10] as defined by an up-regulation of CD10, CD77, and CD95. The present analysis confirmed the induced presence of CD10 (MME) and CD95 (TNFRSF6) on the transcriptional level (Table 2). The neutral glycosphingolipid CD77 is not represented on the microarray. In the present setting the difference in CD10 and CD95 expression was attributed to a temporal induction, because the CD95 transcript was found in the 11101T cluster and the CD10 is

Table 1. Genes directly regulated and induced by anti-CD44

HGNC*	Annotation	Accession no.	Pattern
IL1A	Interleukin 1, alpha	M28983	TTT000
NA	Human clone 23908 mRNA sequence	U79290	TTT000
NA	Human (BSF-2/IL6) gene for B cell stimulatory factor-2	Y00081	TTT000
PDE6A	Phosphodiesterase 6A, alpha subunit	M26061	TT1000
NA	*H sapiens* mRNA sequence (15q11-13)	X69636	TT0000
C18B11	C18B11 homolog	U67934	TT0000
MLLT3	Myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, *Drosophila*); translocated to, 3	L13744	T1T000
SOX5	SRY (sex-determining region Y)-box 5	S83308	T10000
ATP1A2	ATPase, Na^+K^+ transporting, alpha-2 polypeptide	J05096	T10000
AQP1	Aquaporin 1	U41518	T0T000
STX1A	Syntaxin 1A	L37792	T0T000
PLG	Plasminogen	M34276	T01000
IF	I factor	Y00318	T00000
AFAP	Actin filament associated protein	D25248	1TT000
SH3BP2	SH3-domain binding protein 2	AB000462	11T000
GCNT2	Glucosaminyl (*N*-acetyl) transferase 2, I-branching enzyme	L41607	0TT000
PYGL	Phosphorylase, glycogen; liver	M14636	0TT000
ADRB2	Adrenergic, beta-2-, receptor, surface	M15169	0TT000
PIGA	Phosphatidylinositol glycan, class A	S78467	0TT000
ECGF1	Endothelial cell growth factor 1	S72487	0T1000
NBL1	Neuroblastoma, suppression of tumorigenicity 1	D28124	0T1000
MITF	Microphthalmia-associated transcription factor	Z29678	0T0000
STK9	Serine/threonine kinase 9	X89059	00T000
ALB	Albumin	U22961	00T000
CAMK2D	Calcium/calmodulin-dependent protein kinase (CaM kinase) II delta	U50361	00T000
MAPK14	Mitogen-activated protein kinase 14	L35253	00T000
RGS3	Regulator of G-protein signaling 3	U27655	00T000
SERPINC1	Serine (or cysteine) proteinase inhibitor, clade C (antithrombin), member 1	M21642	00T000

*Annotation based on HUGO Gene Nomenclature Committee (HGNC): http://www.gene.ucl.ac.uk/nomenclature/.
NA indicates not applicable.

Table 2. Genes directly regulated and repressed by anti-CD44

HGNC*	Annotation	Accession no.	Pattern
SSTR4	Somatostatin receptor 4	L14856	000TTT
NA	*H sapiens* CREB gene	X68994	000TTT
PALM	Paralemmin	D87460	000TT0
NA	*H sapiens* mRNA for NK receptor	X97230	000TT0
SLC14A1	Solute carrier family 14 member 1	U35735	000T0T
PRDM2	PR domain containing 2	U17838	000T0T
HLCS	Holocarboxylase synthetase	D87328	000T01
APC	Adenomatosis polyposis coli	M73548	000T00
MLLT10	Myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, *Drosophila*); translocated to, 10	U13948	0001T0
PLS1	Plastin 1	L20826	0000TT
LU	Lutheran blood group	X80026	0000TT
SERPINA1	Serine (or cysteine) proteinase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1	K01396	0000T0
LAMP2	Lysosomal-associated membrane protein 2	L09717	0000T0
GUCY2F	Guanylate cyclase 2F, Retinal	L37378	00000T
EPHB3	EphB3	X75208	00000T

*Annotation based on HUGO Gene Nomenclature Committee (HGNC): http://www.gene.ucl.ac.uk/nomenclature/.
NA indicates not applicable.

found in the 11T0TT cluster. This indicated that CD44-dependent induction of these molecules occurred at the earlier time points.

The list of directly regulated genes represents several novel members. In the cluster designated 00T000, representing genes differentially expressed at 72 hours by CD44 ligation, is the SERPINC1 transcript for antithrombin III found. This cluster also contains the RGS3 transcript for the G-protein signaling regulator 3 protein, which is involved in the regulation of chemoattractant-mediated migration in lymphocytes.[26,27] The cluster 0TT000, representing genes that are differentially up-regulated after 6 hours by the CD44 ligation, contains the transcript for the β_2-adrenergic receptor (ADRB2). This receptor binds the sympathetic nervous system (SNS) mediator norepinephrine. This is of interest, because norepinephrine innervation of lymphoid organs plays a central regulatory role in the immune system.[28]

In the cluster 0T1000 we found the platelet-derived endothelial cell growth factor (ECGF1). This cytokine is known for its involvement in cancer and rheumatoid arthritis and has, for instance, been shown to induce inflammation and hyperplasia in synovial cells.[29] In the cluster described by TTT000, the immunomodulating interleukins IL-6 and IL-1α (IL1A) were found. IL-6 is secreted upon inflammation and has a wide array of functions both centrally and peripherally.[30] IL-1α is another cytokine that has been implicated in functions both centrally and peripherally. In this context, it is a cytokine that has been considered to have overlapping functions with IL-1β,[31] although IL-1α has been found not to compensate for IL-1β loss.[32] However, several studies have reported a role for IL-1α in T helper cell regulation,[33-34] suggesting a more confined functionality.

The up-regulation of IL-1α and IL-6 mRNA led us to investigate their presence at the protein level. Flow cytometry analysis was used to detect surface expression of IL-1α at the various time points after CD44 ligation. The population of IL-1α-expressing cells increased by a factor of 3.9 (SEM 0.44) at 24 hours. In total, 10% (SEM 0.16) of the total CD44-stimulated B-cell population was IL-1α^+ at 24 hours (data not shown). The IL-6 production at the various time points after CD44 ligation was assessed by ELISA. The average net increase of CD44-induced IL-6, observed at 24 and 72 hours, was 1.6 times (SEM 0.26), reaching a total concentration of 300 pg/mL (SEM 30.0). These findings validate the CD44 involvement in the induction of these 2 immunomodulatory cytokines.

Discussion

The hyaluronan (HA) receptor CD44 has been implicated in the regulation of the immune system, and as a member of the hyaladherin family of proteins its role in adhesion and extracellular

Table 3. Genes temporally regulated by induction by CD44

HGNC*	Annotation	Accession no.	Pattern
PRPF18	PRP18 pre-mRNA processing factor 18	U51990	1TT00T
NFKB2	Nuclear factor of kappa light polypeptide gene enhancer in B-cells 2	U20816	1T10TT
KDELR1	KDEL (Lys-Asp-Glu-Leu) endoplasmic reticulum protein retention receptor 1	X55885	1T10TT
PNMT	Phenylethanolamine *N*-methyltransferase	X52730	1T10T1
DDX7	DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide 7 (RNA helicase, 52 kDa)	D26528	1T100T
MME	Membrane metallo-endopeptidase (neutral endopeptidase, enkephalinase, CALLA, CD10)	J03779	11T0TT
MPP3	Membrane protein, palmitoylated 3	U37707	11T0T1
KIAA0121	KIAA0121 gene product	D50911	1110T1
TNFRSF6	Tumor necrosis factor receptor superfamily, member 6	X63717	11101T
NA	Human XIST	X56199	111011
DIPA	Hepatitis delta antigen-interacting protein A	U63825	111011
PRIM2	Primase, polypeptide 2A	X74331	111011
GPR31	G protein-coupled receptor 31	U65402	10T0T0
HR44	Hr44 antigen	X91103	10T0T0

*Annotation based on HUGO Gene Nomenclature Committee (HGNC): http://www.gene.ucl.ac.uk/nomenclature/.
NA indicates not applicable.

Table 4. Genes temporally regulated by repression by CD44

HGNC*	Annotation	Accession no.	Pattern
KIAA0196	KIAA0196 gene product	D83780	00T1TT
CRYGEP1	Crystallin, gamma E pseudogene 1	K03008	00T1TT
LYZ	Lysozyme	J03801	00T1TT
KIAA0141	KIAA0141 gene product	D50931	00TT0T
TERF1	Telomeric repeat binding factor	U40705	00TT1T
EFNA3	Ephrin-A3	U14187	00TTT0
RNTRE	Related to the N terminus of tre	D13644	00TTT1
NUCB2	Nucleobindin 2	X76732	00TTT1
PAFAH2	Platelet-activating factor acetylhydrolase 2	D87845	00TTTT
GCG	Glucagon	J04040	00TTTT
ELL	ELL gene	U16282	00TTTT
MNAT1	Menage a trois 1 (CAK assembly factor)	X87843	00TTTT
NR4A3	Nuclear receptor subfamily 4, group A, member 3	X89894	01T1TT
SLC18A1	Solute carrier family 18, member 1	U39905	0T0TTT
NA	Human clone 23907 mRNA sequence	U90907	0T111T
PLK	Pololike kinase	U01038	0T1T11
CCNG2	Cyclin G2	U47414	0T1T11
MX2	Myxovirus (influenza virus) resistance 2	M30818	0T1T1T
HTR1A	5-hydroxytryptamine (serotonin) receptor 1A	M83181	0TT11T
FRDA	Friedreich ataxia	U43747	0TT11T
MEF2A	MADS box transcription enhancer factor 2, polypeptide A	X68505	0TT1T1
SNCA	Synuclein, alpha	U46901	0TTT11
APOC3	Apolipoprotein C-III	X01388	0TTT1T
PRKACA	Protein kinase, cAMP-dependent, catalytic, alpha	M80335	0TTTT1
AUH	AU RNA binding protein/enoyl-coenzyme A hydratase	X79888	0TTTT1

*Annotation based on HUGO Gene Nomenclature Committee (HGNC): http://www.gene.ucl.ac.uk/nomenclature/.
NA indicates not applicable.

interactions has been extensively investigated.[4] Furthermore, a crucial role of CD44 in inflammation has also lately been demonstrated.[35] In the present study we assessed the transcriptional outcome of ligation of CD44 on mature naive tonsillar B cells. The dataset was analyzed employing a novel statistical analysis scheme created to retrieve the most likely expression pattern of an observed distribution. We were thus able to extract the patterns for genes classified as directly or temporally regulated by the anti-CD44 stimuli (Tables 1-2).

Our previous observation that CD44 ligation in naive B cells in fact induced a GC B-cell–like phenotype was corroborated in the present transcriptional analysis, in that CD10 (MME) and CD95 (TNFRSF6) were shown to be temporally regulated by CD44. The neutral endopeptidase CD10 has a strong association to activated germinal center B-cell populations,[36] and the only other stage in B-cell ontogeny where CD10 is expressed is at the pro-B-cell stage in bone marrow. The functional role of CD10 is associated to inflammation, and CD10 provides protection to several neuropeptides and other mediators of inflammation in tissue. Some of the substrates include endothelin, bradykinin, substance P, calcitonin gene-related peptide, neuropeptide Y, and IL-1β.[37] The CD95 expression is also highly associated with activation, especially in lymphocytes.[38] This receptor is a prerequisite for a physiological control of the activated immune system, enabling the possibility to clear malfunctioning cells by apoptosis. If this control is perturbed, the imbalance in the immune system could lead to autoimmunity. Interestingly, an increased susceptibility to CD95-mediated activation-induced cell death has been linked to CD44 functionality on thymocytes.[6,7]

Among the components that were found to be directly regulated by CD44 were the cytokines IL-6 and IL-1α. IL-6 is a pleiotropic molecule involved as a regulator of inflammation both centrally and peripherally.[30] A role of IL-6 in the induction and progression of arthritis has also been demonstrated in IL-6–deficient mice.[39] IL-6–deficient mice, furthermore, show impaired antibody production and reduced numbers of both B and T cells in lymph nodes as well as reduced germinal center formation.[40] The induction of IL-6 is not unique to CD44, because IL-6 induction has been associated also with CD40-mediated signaling.[41] However, CD44 has been shown to augment IL-6 production in human rheumatoid synovial cells[42] and polymorphonuclear cells,[43] which is in agreement with our findings where the CD44 ligation generated an increase in IL-6 protein production. Interestingly, it has also recently been reported that IL-6 production can be specifically attributed to CD44 v6 at least in macrophages.[44]

Direct regulation by the CD44 stimuli was evident also for the IL-1α transcript (IL1A). This induction seems to be present relatively early in time and is sustained throughout the 72 hours as it follows the TTT000 pattern. Several studies have demonstrated a CD44-dependent induction of IL-1.[45,46] However, in contrast to previous studies, the CD44 ligation on the BCR and CD40-activated B cells induced IL-1α. The immunomodulatory function of IL-1α is mostly associated with T helper cell proliferation. It induces T helper-2 (T_H2) cell proliferation independently of IL-4 and initiates autocrine IL-1α production,[34] which furthermore has been shown to increase the T_H2 responsiveness to IL-4.[47] Our confirmation that IL-1α was up-regulated at the protein level also demonstrated that a CD44 ligation augmented this surface expression, and a nearly 4-fold increase in the number of IL-1α–producing cells was evident. Taken together, these results suggest an attractive role of CD44-activated B cells in the regulation of T-cell proliferation.

CD44 has also been implicated in lymphocyte migration. In this study SERPINC1 and RSG3, both potent regulators of migration, were shown to be regulated by CD44. The serpin, antithrombin III (SERPINC1), is an important regulator of the coagulation cascade as well as being involved in inflammation.[48] Antithrombin III has also been found to inhibit chemokine-mediated migration through

the heparan sulfate proteoglycans syndecan-4 [49] and by a similar interaction also to interfere with nuclear factor-κB (NF-κB) activation,[50] which explains the anti-inflammatory effects of the enzyme. Another regulator of chemokine-mediated migration with a differential expression pattern is the G-protein signaling regulator RGS3. This molecule is a modulator of chemotaxis and cell adhesion mediated by G protein-coupled chemokine receptors[26] and has been found to efficiently inhibit B-cell chemotaxis mediated by the CXCR4, CXCR5, and CCR7.[27] Thus, the regulatory properties of SERPINC1 and RSG3 may be of critical importance, because a balance in the chemokine-mediated migration of lymphocytes has shown to be instructive in the formation of the secondary lymphoid organ structures.[3]

Another up-regulated immunoregulatory gene dependent on CD44 stimulation is the β_2-adrenergic receptor (β_2-AR). The adaptive response to inflammation involves components of the central nervous system (CNS). The systemic release of proinflammatory cytokines triggers the enrollment by CNS of the hypothalamic-pituitary-adrenal (HPA) axis and the sympathetic nervous system (SNS). These systems in concert modulate the immune response by a subsequent release of glucocorticoids from the adrenal glands and norepinephrine (NE) by postganglionic noradrenergic nerve endings, respectively. The peripheral discharge of glucocorticoids and NE will inhibit T_H1 responses and promote a T_H2 response thereby counteracting the inflammatory reaction.[43,51,52] The sympathetic regulation of the immune system is mediated mainly by the β_2-AR, which has been found on all lymphocytes except for T_H2 cells. In B cells it has been reported that the β_2-AR facilitates T_H2-dependent IgG1 and IgE production quantitatively after antigen stimulation in mice.[53,54] It is also proposed that the β_2-AR plays a role in germinal center formation, because NE depletion attenuates GC formation.[53] The β_2-AR transcript ADRB2 is contained in a cluster characterizing genes induced after 24 hours (0TT000). This indicates that in line with proinflammatory cytokines[55] the CD44 stimuli seem to be a regulator of the β_2-AR expression of B cells. The β_2-adrenergic receptor uses the cyclic adenosine monophosphate (cAMP) and the protein kinase A (PKA) pathway. This pathway has been reported to regulate T_H1 and T_H2 responses and T_H2-dependent IgE and IgG production in B cells by a number of effectors.[56] Thus, the induced presence of the β_2-AR suggests a B cell involved in a T_H2-dependent response.

Interestingly, a recent study demonstrated rapid down-regulation of CD44 on NK cells by NE treatment,[57] and it is possible to speculate on a similar mechanism in the heavily innervated marginal zone. Such a mechanism presents a possible scenario where the CD44 participates in the decision making of the B-cell faith. For example, antigen-induced CD44 [9,11,12] will as a costimultory molecule[9] participate in a T cell-B cell cognate interaction with a sequential up-regulation of the β_2-AR. This β_2-AR up-regulation will mount the pivot point in this regulatory process at which the neuroendocrine system decides whether a germinal center formation should be supported or not. The β_2-AR signaling will consequently induce cellular events facilitating GC formation[53,54] characterized by a CD44 down-regulation.

In summary, our results suggest a novel role for CD44 in immunoregulation and inflammation. More specifically, possible mechanisms for CD44 involvement in the humoral response and the formation of germinal center reactions based on the up-regulation of transcript for CD10, CD95, IL-6, IL-1α, and β_2-AR are proposed.

References

1. Butcher EC, Picker LJ. Lymphocyte homing and homeostasis. Science. 1996;272:60-66.
2. MacLennan IC. Germinal centers. Annu Rev Immunol. 1994;12:117-139.
3. Moser B, Loetscher P. Lymphocyte traffic control by chemokines. Nat Immunol. 2001;2:123-128.
4. Lesley J, Hyman R, Kincade PW. CD44 and its interaction with extracellular matrix. Adv Immunol. 1993;54:271-335.
5. Rafi A, Nagarkatti M, Nagarkatti PS. Hyaluronate-CD44 interactions can induce murine B-cell activation. Blood. 1997;89:2901-2908.
6. Foger N, Marhaba R, Zoller M. CD44 supports T cell proliferation and apoptosis by apposition of protein kinases. Eur J Immunol. 2000;30:2888-2899.
7. Chen D, McKallip RJ, Zeytun A, et al. CD44-deficient mice exhibit enhanced hepatitis after concanavalin A injection: evidence for involvement of CD44 in activation-induced cell death. J Immunol. 2001;166:5889-5897.
8. Naor D, Sionov RV, Ish-Shalom D. CD44: structure, function, and association with the malignant process. Adv Cancer Res. 1997;71:241-319.
9. Guo Y, Wu Y, Shinde S, Sy MS, Aruffo A, Liu Y. Identification of a costimulatory molecule rapidly induced by CD40L as CD44H. J Exp Med. 1996; 184:955-961.
10. Ingvarsson S, Dahlenborg K, Carlsson R, Borrebaeck CA. Co-ligation of CD44 on naive human tonsillar B cells induces progression towards a germinal center phenotype. Int Immunol. 1999;11: 739-744.
11. Camp RL, Kraus TA, Birkeland ML, Pure E. High levels of CD44 expression distinguish virgin from antigen-primed B cells. J Exp Med. 1991;173: 763-766.
12. Arch R, Wirth K, Hofmann M, et al. Participation in normal immune responses of a metastasis-inducing splice variant of CD44. Science. 1992; 257:682-685.
13. Kremmidiotis G, Zola H. Changes in CD44 expression during B cell differentiation in the human tonsil. Cell Immunol. 1995;161:147-157.
14. Feuillard J, Taylor D, Casamayor-Palleja M, Johnson GD, MacLennan IC. Isolation and characteristics of tonsil centroblasts with reference to Ig class switching. Int Immunol. 1995;7:121-130.
15. Dahlenborg K, Pound JD, Gordon J, Borrebaeck CA, Carlsson R. Terminal differentiation of human germinal center B cells in vitro. Cell Immunol. 1997;175:141-149.
16. Stott DI, Hiepe F, Hummel M, Steinhauser G, Berek C. Antigen-driven clonal proliferation of B cells within the target tissue of an autoimmune disease. The salivary glands of patients with Sjogren's syndrome. J Clin Invest. 1998;102:938-946.
17. Kim HJ, Berek C. B cells in rheumatoid arthritis. Arthritis Res. 2000;2:126-131.
18. Lindhout E, van Eijk M, van Pel M, Lindeman J, Dinant HJ, de Groot C. Fibroblast-like synoviocytes from rheumatoid arthritis patients have intrinsic properties of follicular dendritic cells. J Immunol. 1999;162:5949-5956.
19. Aarvak T, Natvig JB. Cell-cell interactions in synovitis: antigen presenting cells and T cell interaction in rheumatoid arthritis. Arthritis Res. 2001;3: 13-17.
20. Haynes BF, Hale LP, Patton KL, Martin ME, McCallum RM. Measurement of an adhesion molecule as an indicator of inflammatory disease activity. Up-regulation of the receptor for hyaluronate (CD44) in rheumatoid arthritis. Arthritis Rheum. 1991; 34:1434-1443.
21. Mikecz K, Dennis K, Shi M, Kim JH. Modulation of hyaluronan receptor (CD44) function in vivo in a murine model of rheumatoid arthritis. Arthritis Rheum. 1999;42:659-668.
22. Schadt EE, Li C, Su C, Wong WH. Analyzing high-density oligonucleotide gene expression array data. J Cell Biochem. 2000;80:192-202.
23. Li C, Wong WH. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci U S A. 2001;98:31-36.
24. Ek S, Hogerkorp CM, Dictor M, Ehinger M, Borrebaeck CA. Mantle cell lymphomas express a distinct genetic signature affecting lymphocyte trafficking and growth regulation as compared with subpopulations of normal human B cells. Cancer Res. 2002;62:4398-4405.
25. Banchereau J, de Paoli P, Valle A, Garcia E, Rousset F. Long-term human B cell lines dependent on interleukin-4 and antibody to CD40. Science. 1991;251:70-72.
26. Bowman EP, Campbell JJ, Druey KM, Scheschonka A, Kehrl JH, Butcher EC. Regulation of chemotactic and proadhesive responses to chemoattractant receptors by RGS (regulator of G-protein signaling) family members. J Biol Chem. 1998;273: 28040-28048.
27. Reif K, Cyster JG. RGS molecule expression in murine B lymphocytes and ability to down-regulate chemotaxis to lymphoid chemokines. J Immunol. 2000;164:4720-4729.
28. Elenkov IJ, Wilder RL, Chrousos GP, Vizi ES. The sympathetic nerve—an integrative interface between two supersystems: the brain and the immune system. Pharmacol Rev. 2000;52:595-638.
29. Waguri-Nagaya Y, Otsuka T, Sugimura I, et al. Synovial inflammation and hyperplasia induced

by gliostatin/platelet-derived endothelial cell growth factor in rabbit knees. Rheumatol Int. 2000;20:13-19.

30. Tilg H, Dinarello CA, Mier JW. IL-6 and APPs: anti-inflammatory and immunosuppressive mediators. Immunol Today. 1997;18:428-432.
31. Labow M, Shuster D, Zetterstrom M, et al. Absence of IL-1 signaling and reduced inflammatory response in IL-1 type I receptor-deficient mice. J Immunol. 1997;159:2452-2461.
32. Horai R, Asano M, Sudo K, et al. Production of mice deficient in genes for interleukin (IL)-1α, IL-1β, IL-1α/β, and IL-1 receptor antagonist shows that IL-1β is crucial in turpentine-induced fever development and glucocorticoid secretion. J Exp Med. 1998;187:1463-1475.
33. Zubiaga AM, Munoz E, Huber BT. Production of IL-1 α by activated Th type 2 cells. Its role as an autocrine growth factor. J Immunol. 1991;146: 3849-3856.
34. Huber M, Beuscher HU, Rohwer P, Kurrle R, Rollinghoff M, Lohoff M. Costimulation via TCR and IL-1 receptor reveals a novel IL-1α-mediated autocrine pathway of Th2 cell proliferation. J Immunol. 1998;160:4242-4247.
35. Teder P, Vandivier RW, Jiang D, et al. Resolution of lung inflammation by CD44. Science. 2002; 296:155-158.
36. Liu YJ, Arpin C, de Bouteiller O, et al. Sequential triggering of apoptosis, somatic mutation and isotype switch during germinal center development. Semin Immunol. 1996;8:169-177.
37. Koehne P, Schaper C, Graf K, Kunkel G. Neutral endopeptidase 24.11: its physiologic and possibly pathophysiologic role in inflammation with special effect on respiratory inflammation. Allergy. 1998; 53:1023-1042.
38. Onel KB, Tucek-Szabo CL, Ashany D, Lacy E, Nikolic-Zugic J, Elkon KB. Expression and function of the murine CD95/FasR/APO-1 receptor in relation to B cell ontogeny. Eur J Immunol. 1995; 25:2940-2947.
39. Ohshima S, Saeki Y, Mima T, et al. Interleukin 6 plays a key role in the development of antigen-induced arthritis. Proc Natl Acad Sci U S A. 1998; 95:8222-8226.
40. La Flamme AC, Pearce EJ. The absence of IL-6 does not affect Th2 cell development in vivo, but does lead to impaired proliferation, IL-2 receptor expression, and B cell responses. J Immunol. 1999;162:5829-5837.
41. Baccam M, Bishop GA. Membrane-bound CD154, but not CD40-specific antibody, mediates NF-κB-independent IL-6 production in B cells. Eur J Immunol. 1999;29:3855-3866.
42. Fujii K, Tanaka Y, Hubscher S, Saito K, Ota T, Eto S. Crosslinking of CD44 on rheumatoid synovial cells augment interleukin 6 production. Lab Invest. 1999;79:1439-1446.
43. Sconocchia G, Campagnano L, Adorno D, et al. CD44 ligation on peripheral blood polymorphonuclear cells induces interleukin-6 production. Blood. 2001;97:3621-3627.
44. Khaldoyanidi S, Karakhanova S, Sleeman J, Herrlich P, Ponta H. CD44 variant-specific antibodies trigger hemopoiesis by selective release of cytokines from bone marrow macrophages. Blood. 2002;99:3955-3961.
45. Webb DS, Shimizu Y, Van Seventer GA, Shaw S, Gerrard TL. LFA-3, CD44, and CD45: physiologic triggers of human monocyte TNF and IL-1 release. Science. 1990;249:1295-1297.
46. Noble PW, Lake FR, Henson PM, Riches DW. Hyaluronate activation of CD44 induces insulin-like growth factor-1 expression by a tumor necrosis factor-α-dependent mechanism in murine macrophages. J Clin Invest. 1993;91:2368-2377.
47. McArthur JG, Raulet DH. CD28-induced costimulation of T helper type 2 cells mediated by induction of responsiveness to interleukin 4. J Exp Med. 1993;178:1645-1653.
48. Esmon CT. Role of coagulation inhibitors in inflammation. Thromb Haemost. 2001;86:51-56.
49. Kaneider NC, Reinisch CM, Dunzendorfer S, Romisch J, Wiederman CJ. Syndecan-4 mediates antithrombin-induced chemotaxis of human peripheral blood lymphocytes and monocytes. J Cell Sci. 2002;115:227-236.
50. Oelschlager C, Romisch J, Staubitz A, et al. Antithrombin III inhibits nuclear factor κB activation in human monocytes and vascular endothelial cells. Blood. 2002;99:4015-4020.
51. Sternberg EM. Neuroendocrine regulation of autoimmune/inflammatory disease. J Endocrinol. 2001;169:429-435.
52. Kohm AP, Sanders VM. Norepinephrine: a messenger from the brain to the immune system. Immunol Today. 2000;21:539-542.
53. Kohm AP, Sanders VM. Suppression of antigen-specific Th2 cell-dependent IgM and IgG1 production following norepinephrine depletion in vivo. J Immunol. 1999;162:5299-5308.
54. Kasprowicz DJ, Kohm AP, Berton MT, Chruscinski AJ, Sharpe A, Sanders VM. Stimulation of the B cell receptor, CD86 (B7–2), and the β 2-adrenergic receptor intrinsically modulates the level of IgG1 and IgE produced per B cell. J Immunol. 2000;165:680-690.
55. Baerwald CG, Burmester GR, Krause A. Interactions of autonomic nervous, neuroendocrine, and immune systems in rheumatoid arthritis. Rheum Dis Clin North Am. 2000;26:841-857.
56. Fedyk ER, Adawi A, Looney RJ, Phipps RP. Regulation of IgE and cytokine production by cAMP: implications for extrinsic asthma. Clin Immunol Immunopathol. 1996;81:101-113.
57. Nagao F, Suzui M, Takeda K, Yagita H, Okumura K. Mobilization of NK cells by exercise: down-modulation of adhesion molecules on NK cells by catecholamines. Am J Physiol Regul Integr Comp Physiol. 2000;279:R1251-R1256.

Paper II

RNA analysis of B cell lines arrested at defined stages of differentiation allows for an approximation of gene expression patterns during B cell development

Panagiotis Tsapogas,* Thomas Breslin,† Sven Bilke,† Anna Lagergren,* Robert Månsson,* David Liberg,* Carsten Peterson,† and Mikael Sigvardsson*

Laboratory for Cellular Differentiation, Department for Stemcell Biology, and †Department of Theoretical Physics, Lund University, Sweden

Abstract: The development of a mature B lymphocyte from a bone marrow stem cell is a highly ordered process involving stages with defined features and gene expression patterns. To obtain a deeper understanding of the molecular genetics of this process, we have performed RNA expression analysis of a set of mouse B lineage cell lines representing defined stages of B cell development using Affymetrix™ microarrays. The cells were grouped based on their previously defined phenotypic features, and a gene expression pattern for each group of cell lines was established. The data indicated that the cell lines representing a defined stage generally presented a high similarity in overall expression profiles. Numerous genes could be identified as expressed with a restricted pattern using dCHIP-based, quantitative comparisons or presence/absence-based, probabilistic state analysis. These experiments provide a model for gene expression during B cell development, and the correctly identified expression patterns of a number of control genes suggest that a series of cell lines can be useful tools in the elucidation of the molecular genetics of a complex differentiation process. *J. Leukoc. Biol.* 74: 102–110; 2003.

Key Words: *gene expression · immunoglobulin · progenitor B cells*

INTRODUCTION

B lymphocyte development is a highly ordered process proceeding from the progenitor cells in the bone marrow (BM) to the immunoglobulin (Ig)-secreting plasma cell in the spleen, gut, or BM [1, 2]. The early steps of this developmental pathway can be divided into distinct stages based on the recombination status of the Ig genes and on the expression pattern of surface markers and the presence of intracellular proteins [1–6]. The early progenitor B (pro-B) cells in the mouse have their Ig genes in a germ-line configuration and express the surface molecules B220 and AA4.1, the signal-transducing molecule Igβ (B29), and the α-subunit of the interleukin (IL)-7 receptor [5, 6]. Subsequent differentiation results in the expression of the recombination-activating genes *Rag-1* and *Rag-2* and initiation of Ig recombination events [7]. This generates a functional Ig heavy-chain (IgH) gene that is transcribed, translated, and displayed on the cell surface in complex with the surrogate light-chain components λ5 and VpreB, as well as the signal-transduction molecules Igα (*mb-1*) and Igβ (*B29*) [8]. Subsequent differentiation allows for rearrangements of the Ig light-chain (IgL) genes that replace the surrogate light-chain genes on the surface of the B cell [8]. This immature cell is then subjected to negative selection to delete self-reactive cells before it leaves the BM to enter peripheral lymphoid organs, where it becomes a mature B cell [9]. If the cells are activated by interaction with antigens and obtain T cell help, they mature into terminally differentiated plasma cells secreting large amounts of antibodies [10–12]. The extensively studied biology of B cell development, in combination with the defined stages of differentiation, makes it a useful system for investigations of complex molecular events that might provide clues to general features of cellular development.

To obtain a deeper understanding of the molecular processes involved in B cell development and to create a map over stage-restricted gene expression, we wanted to establish a model system that allowed for a reasonable approximation of the gene expression profile during B lymphoid differentiation. Features such as varying proliferation status, heterogeneous populations, and difficulties to obtain sufficient amounts of material limit the use of primary sorted cells. The existence of numerous B cell lines arrested at defined differentiation stages would then pose a possibility to overcome some of these problems. They also provide a highly reproducible source of material that allows for the performance of large-scale gene expression analysis without the use of intermediate amplification steps. The use of cell lines does, however, introduce a risk of obtaining cell line-specific features as a result of the transformation process. To reduce the risk of analyzing cell line-specific features, we used several representative cell lines for each of four major stages in B cell development: pro-B, pre-B,

Correspondence: Mikael Sigvardsson, Laboratory for Cellular Differentiation, Department for Stemcell Biology, BMC B12, Lund, 22184, Sweden. E-mail: Mikael.Sigvardsson@stemcell.lu.se

Received January 9, 2003; revised March 6, 2003; accepted March 24, 2003; doi: 10.1189/jlb.0103008.

B, and plasma cells, and investigated the gene expression pattern in these cell lines by Affymetrix™ microarrays containing ~12,000 gene tags. This allowed for the correct classification of a large number of control genes using dCHIP-based, relative expression level analysis [13] or presence/absence (P/A)-based, probabilistic state analysis (S. Bilke et al., submitted). We also identified a large number of additional genes that now can be considered as candidates to display stage-restricted expression patterns during B cell development.

MATERIALS AND METHODS

Tissue-culture conditions

All cells were grown at 37°C and 5% CO_2 in RPMI supplemented with 7.5% fetal calf serum, 10 mM HEPES, 2 mM pyruvate, 50 μM 2-mercaptoethanol, and 50 μg/ml gentamycin (all purchased from Life Technologies AB, Täby, Sweden). The pro-B cell lines were grown in RPMI as above, supplemented with 10% of IL-3 containing WEHI3-conditioned media. The Ba/F3 subclones were kind gifts from Drs. Rudolf Grosschedl (Gene Center, Munich), Ramiro Gisler (Department for Stemcell Biology, Lund University), and Johan Forssel (Department of Cell and Molecular Biology, Umeå University), and the Ly9D cells were a gift from Dr. Meinrad Busslinger (IMP, Wienna). The pre-B cell line 18–81 was a gift from Dr. Inge-Lill Mårtensson (The Babraham Institute, Cambridge), and all the other cell lines were gifts from Dr. Thomas Leanderson (Department of Immunology, Lund University).

Gene expression analysis

RNA was prepared using Trizol (Gibco, Grand Island, NY), and 7.5 μg of total RNA was annealed to a T7-oligo T primer by denaturation at 70°C for 10 min followed by 10 min of incubation of the samples on ice. First-strand synthesis was performed for 2 h at 42°C using 20 U Superscript reverse transcriptase (RT; Gibco) in buffers and nucleotide mixes according to the manufacturer's instructions. This was followed by a second-strand synthesis for 2 h at 16°C, using RNaseH, *Escherichia coli* DNA polymerase I, and *E. coli* DNA ligase (all from Gibco), according to the manufacturer's instructions. The obtained, double-stranded cDNA was then blunted by the addition of 20 U T4 DNA polymerase and incubated for 5 min at 16°C. The material was then purified by phenol:cloroform:isoamyl alcohol extraction followed by precipitation with NH_4Ac and ethanol. The cDNA was then used in an in vitro transcription reaction for 6 h at 37°C using a T7 IVT kit and biotin-labeled ribonucleotides. The obtained cRNA was purified from unincorporated nucleotides on an RNAeasy column (Qiagen, Valencia, CA). The eluted cRNA was then fragmented by incubation of the products for 2 h in fragmentation buffer (40 mM Tris-acetate, pH 8.1, 100 mM KOAc, 150 mM MgOAc). The final, fragmented cRNA (20 μg) was hybridized to Affymetrix™ chip U74Av2 (Affymetrix, Santa Clara, CA) in 200 μl hybridization buffer [100 mM 2-(N-morpholino)-ethanesulfonic buffer, pH 6.6, 1 M NaCl, 20 mM EDTA, 0.01% Tween 20], supplemented with Herring sperm DNA (100 μg/ml) and acetylated bovine serum albumin (500 μg/ml) in an Affymetrix Gene Chip Hybridization Oven 320. The chip was then developed by the addition of fluorescein isothiocyanate (FITC)-streptavidin followed by washing using an Affymetrix Gene Chip Fluidics Station 400. Scanning was performed using a Hewlett Packard Gene Array scanner.

Data analysis

Probabilistic estimation of gene expression pattern was performed using the Breslin/Bilke method (S. Bilke et al., submitted). Hierarchical tree clusters were generated using the dCHIP program (Li and Wong [13], <http://biosun1.harvard.edu/complab/dchip/>). The initial analysis (see Fig. 2) was performed using the perfect match (PM)-only model, and genes were filtered according to 0.50 <standard deviation/mean between <10.00, and P call in the array used ≥20%. The pro-B, pre-B, B, and plasma cell lines (see Fig. 4 and Supplemental Figs. 1–4 available at http://www.jleukbio.org/) were treated as replicates using the same model as above but a P call above 10%. Classification of genes with an apparent restricted expression pattern in the P/A analysis into functional groups was performed manually using the National Center for Biotechnology Information (NCBI; National Institutes of Health, Bethesda, MD) database (see supplementary tables). Nondefined genes in the data set were resubmitted in a Blast search into Genebank allowing for the identification of most of these entries.

RT and polymerase chain reactions (PCRs)

RNA was prepared from cells using Trizol (Life Technology), and cDNA was generated by annealing 1 μg total RNA to 0.5 μg random hexamers in 10 μl diethylpyrocarbonate-treated water. RT reactions were performed with 200 U Superscript RT (Life Technologies) in the manufacturer's buffer supplemented with 0.5 mM dNTP, 10 mM dithiothreitol, and 20 U RNase inhibitor (Boehringer Mannheim, Bromma, Sweden) in a total volume of 20 μl at 37°C for 1 h. One-twentieth of the RT reactions were used in the PCR assays. PCR reactions were performed with 1 U Taq-polymerase (Life Technologies) in the manufacturer's buffer, supplemented with 0.2 mM dNTP in a total volume of 25 μl. Primers were added to a final concentration of 1 mM.

For all PCRs, the program was the same except for the number of cycles (Y) and the annealing temperature (XX). The common parts of the program were (Y cycles) 95°C for 2 min, 95°C for 45 s, XX°C for 45 s, 72°C for 1 min, and 72°C for 2 min. Annealing temperatures–cycles were: actin, 57ºC, 25 cycles; hypoxanthine guanine phosphoribosyl transferase (HPRT), 52ºC, 25 cycles; λ5, 60ºC, 30 cycles; Pax-5, 60ºC, 30 cycles; J-chain, 60ºC, 28 cycles; Bach1, 55°C, 30 cycles; RhoB, 56°C, 30 cycles; Yes, 55°C, 30 cycles; Sel1, 55°C, 25 cycles; Arhgef, 55°C, 27 cycles; protein kinase C (PKC)-β, 55°C, 28 cycles; and Pftaire-1, 55°C, 30 cycles.

Oligonucleotides used for RT-PCR were: actin, sense 5′GTTTGAGACCTTCAACACC, antisense 5′GTGGCCATCTCCTGCTCGAAGTC; B29, sense 5′GGTGAGCCGGTACCAGCAATG, antisense 5′AGTTCCGTGCCACAGCTGTCG; λ5, sense 5′TGTGAAGTTCTCCTCCTGCTC, antisense 5′ACCACCAAAGTACCTGGGTAG; Pax-5, sense 5′CTACAGGCTCCGTGACGCAG, antisense 5′GTCTCGGCCTGTGAAATAGG; Bach-1, sense 5′ACTCTCAGTTCCGTCAACTGC, antisense 5′TTCCTCTTGCGACAGCGTTGC; Arhgef3 (EST-1), sense 5′AAACATCCGTCCACTCTCTTCC, antisense 5′TACTGTACACATGGGTCATGTGC; Pftaire-1, sense 5′TGCTCTAGCATACATTGAACC, antisense 5′CTCCCCACTTAAAGAACTCC; PKC-β-II, sense 5′ATCCACCAGTCCTAACACC, antisense 5′AAGCAAGCATTTTCTCTCC; RhoB, sense 5′CTGATCGTGTTCAGTAAAGACGAATTCC, antisense 5′TTGTTGGCCACCAGGATGATGG; Yes-associated protein, sense 5′GCAGTTACAGATGGAGAAGGAG, antisense 5′TTGCATCTCCTTCCAGTGTGC; HPRT, sense 5′GCTGGTGAAAAGGACCTCT, antisense 5′CACAGGACTAGAACACCTGC; J-chain, sense 5′GTAGGTGGTACCTATACAATAACA, antisense 5′AGGGTAGCAAGAATCGGGGGTCAA.

Isolation and purification of BM progenitors and mature peripheral B cells

BM cells were sorted on a FACSVantage Cell Sorter (Becton Dickinson, San Jose, CA), equipped with a 488-nm argon ion (Coherent Enterprise II, Santa Clara, CA) and a 633-nm He-Ne (Model 127, Spectra-Physics, Mountain View, CA) laser. Antibodies used were B220 antigen-presenting cell, CD43 phycoerythrin (PE), IgM biotin (Streptavidin TRI), CD19 FITC, and CD138 (Syndecan–1) PE (all from PharMingen, San Diego, CA). The purity of all sorted cell populations is reproducible over 95%. To obtain activated and mature B cells, magnetic cell sorter-purified $B220^+$ spleen cells were incubated in 50 ng/ml lipopolysaccharide (LPS; Sigma Chemical Co., St. Louis, MO) at 37°C for 72 h.

RESULTS

B cell lines arrested at defined stages of development generally display gene expression profiles typical for the differentiation stage

To identify cell lines for our RNA analysis, we searched the literature and selected representatives based on previously published results. As representatives of the pro-B cell stage, we used Ly9D cells and three different subclones of the cell

line Ba/F3 [14, 15]. These are IL-3-dependent lines without Ig rearrangements but with expression of sterile Ig transcripts [14–16]. Ba/F3 cells also express low levels of the B lineage-restricted *mb-1* and *B29* genes [16] and have been differentiated into B cells in vivo [14]. The nitrous–urea-induced cell line 70Z/3 as well as the Abelson virus-transformed lines 230–238, 40EI, and 18–81 represented the pre-B cells. As representatives for the B cell stage, we used the cell lines WEHI231, M12, K46, and A20. The mouse myeloma cell lines J558, S194, MPCII, and SP2.0 represented the Ig-secreting plasma cell stage. To obtain some information concerning gene expression pattern in the cell lines as compared with primary sorted B lineage cells, we extracted RNA from $B220^+$, $CD43^+$, IgM^- (pro-B) cells, $B220^+$, $CD43^-$, IgM^-, $CD19^+$ (pre-B) BM cells, and $B220^+$ spleen cells (B cells). Plasma cells were obtained by sorting Syndecan-1^+, $B220^-$ BM cells or by 72 h LPS stimulation of $B220^+$ splenocytes. The RNA expression patterns in the different cell populations were then analyzed by RT-PCR experiments (**Fig. 1A**) using amplification of actin and HPRT message as a control for the quality of the cDNA. This suggested that the *B29* gene was expressed at all stages of development [17], and the surrogate light-chain gene λ*5* [18] was expressed mainly in pre-B cells. Some λ5 expression was also detected in the Syndecan-1^+ $B220^-$ BM cells representing primary plasma cells. This probably reflects contamination with pre-B cells rather than true expression of this gene at this late developmental stage, as no λ5 expression could be detected in the LPS-stimulated splenocytes. The expression of the transcription factor Pax-5 [19] was high in the pre-B and B cells, and it appeared to be lower in the plasma-cell populations. In contrast, the Ig-associated J-chain was expressed at a higher level in the plasma cells than in the other populations [20]. Analysis of the expression pattern of the same genes in a selection of cell lines (Fig. 1B) indicated that all these expressed the B lineage-restricted *B29* (Igβ) gene, and only the pre-B cell lines expressed λ*5* message. *Pax-5* was not expressed in the pro-B cell lines or in the plasma cell lines, and the pre-B and B cell lines expressed this transcription factor. Message encoding the J-chain was present in one of the pre-B cell lines (230–238) and in cell lines representing the mature or the plasma cell stage. This suggests that the cell lines display expression of B cell markers in patterns comparable with those observed in primary sorted cells.

To expand the analysis of the gene expression patterns in the different B lineage cell lines, we analyzed RNA from these on Affymetrix™ gene chip microarrays containing ~12,000 sequence tags. The data were then analyzed using the dCHIP program [13], allowing for the identification of differentially expressed genes in the cell line samples (**Fig. 2**). This analysis indicated that although there were differences in expression levels, the cell lines previously defined as belonging to a specific stage of development generally displayed a similar cluster of differentially expressed genes. The exception was the B cell line WEHI231, which appeared to group with the pre-B cells rather than with the B cells, possibly reflecting that this cell represents an immature BM-derived B cell rather than a

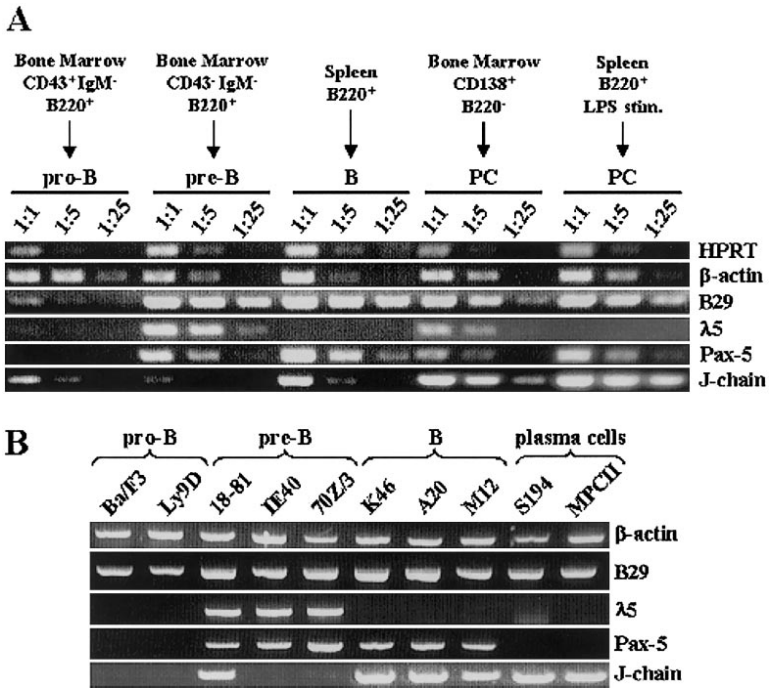
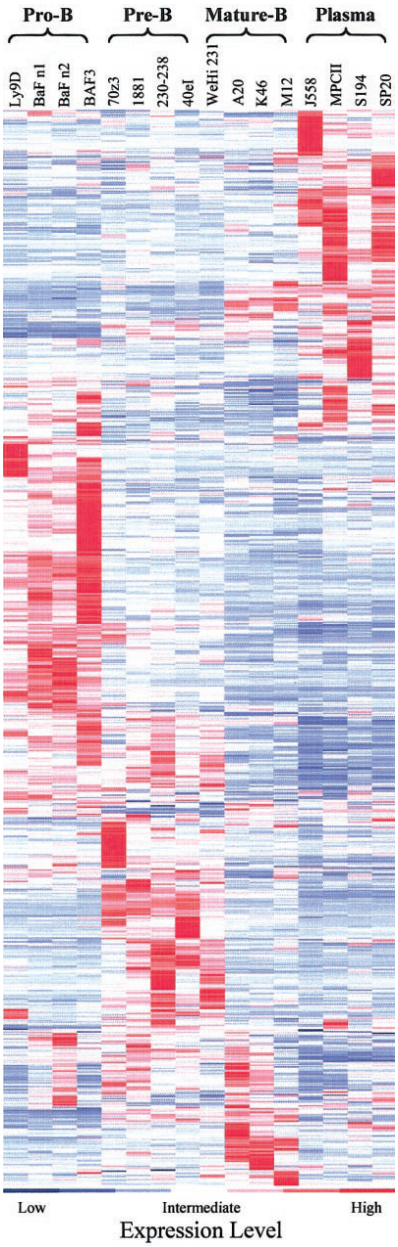


Fig. 1. Cell lines arrested in development express stage-specific genes in patterns resembling primary-sorted B lineage cells. (A) Ethidium bromide-stained agarose gels with PCR products obtained by RT-PCR analysis of primary-sorted cells as indicated. The cDNA was diluted in three steps to allow for a degree of quantification of the expressed transcripts. (B) A gel with a RT-PCR experiment displaying the expression of the same genes in a panel of cell lines as indicated.

mature peripheral cell [21, 22]. The homogeneity in expression levels within the cell line groups representing the different stages was also analyzed by the extraction of normalized expression levels for a set of genes linked to B cell development [1, 3–6] (**Fig. 3**). The expression of the transcription factor Id-1 was reduced upon progression into the pre-B cell stage, and the levels of the signal transducer B29 was increased and maintained in all the cell line groups. mRNA encoding λ5, V-preB1, VpreB-3, and the IL-7 receptor α subunit were all transiently up-regulated in the pre-B cell lines. Mb-1, EBF, and CD19 expression was high in the pre-B cells, but the mRNA could also be detected in the B cell lines. CD24a [human serum albumin (HSA)] was expressed in the pre-B and mature B cells as was PKC-β. The germinal center-specific transcription factor BCL-6 [23, 24] was expressed specifically in the B cell lines, and the Ig-associated J-chain was expressed in the B and the plasma cells. Only the latter cells expressed the plasma-cell transcription factor BLIMP [25]. Thus, although the standard deviations in some cases were substantial, these data indicate that the cell lines representing a specific developmental stage present rather homogenous gene expression patterns and support the idea that stage-specific genes can be identified using a series of B cell lines. To reduce the impact of cell line-specific features, we then treated all cell lines belonging to a certain differentiation stage as replicates and performed another dCHIP analysis. This resulted in the identification of a large number of genes, including several genes with previously defined expression patterns, as expressed in a stage-specific manner (**Fig. 4** and **Attachment Figs. 1–4**).



Affymetrix™ P/A analysis allows for the characterization of stage-restricted gene expression

The design of the Affymetrix™ microarrays, with one set of matching and one set of mismatching oligonucleotides for each gene, allows for a comparison of the obtained signals from the two probe sets. The data are then evaluated by Affymetrix™ array analysis software, allowing for the classification of all the studied genes as present (P) or absent (A) in each of the samples. This transforms the data set into binary values creating novel possibilities for mathematical analysis of the obtained data. This does, however, delete all information about relative expression levels; so to investigate if binary values could be used for the identification of stage-specific genes, we constructed a Hamming distance matrix based on P/A analysis. This generates a measure of similarity between any two samples based on the number of genes for which the Affymetrix™ P/A calls differ (**Fig. 5**). The pro-B and plasma cell groups were the most homogenous, and the B cell group displayed a poorer similarity. M12, K46, and A20 appeared similar, while the WEHI231 cells rather resembled the pre-B cells. This indicates that P/A analysis allows for the correct stage classi-

Fig. 2. dCHIP analysis of gene expression data suggests that B cell lines belonging to a certain differentiation stage generally display similar expression patterns of stage-specific genes. The figure displays a dCHIP analysis of the RNA expression patterns in the different cell lines selected for our investigations after filtering and hierarchical clustering. The name of the cell line and the earlier classification are displayed above the data panel. The criteria selected for the definition of differentially expressed genes were a 20% present count and a maximum standard deviation of 0.5 (0.50<standard deviation/mean between <10.00). Expression scales ranging from –3 (blue) to +3 (red) are indicated below the data display.

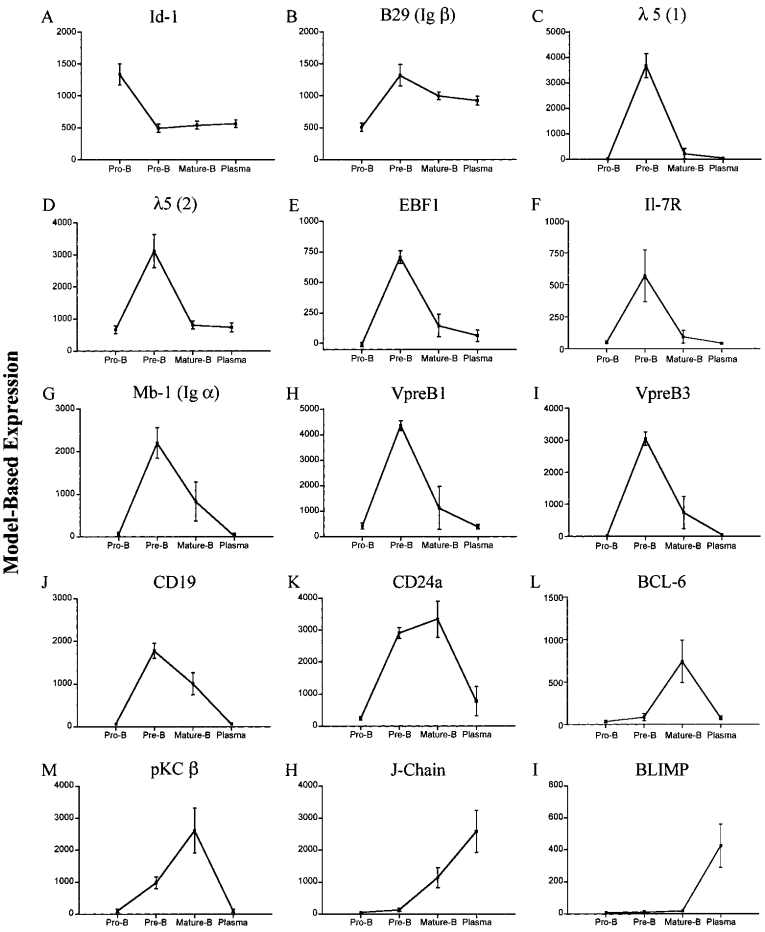


Fig. 3. Investigations of expression pattern of individual control genes indicate that RNA expression data from a group of cell lines can be used for the delineation of stage-specific genes. The figure displays diagrams over normalized expression levels of genes associated with B cell development as indicated on top of each display. The λ5 gene is represented by two probe sets. The standard deviations have been calculated by treatment of the cell lines belonging to a defined stage as replicates. Id-1, Inhibitor of DNA binding-1; EBF1, early B cell factor; BCL-6, B cell lymphoma-6; BLIMP, B lymphocyte-induced maturation protein-1.

fication of the cell lines and that this analysis method could be used to obtain information about stage-specific gene expression.

To extract the collected information from the P/A analysis, we applied the probabilistic estimation method (S. Bilke et al., submitted). Based on the expression of a given gene in the four samples representing each stage, this method yields the conditional probability as to whether the gene should be regarded as absent (A) mixed (M), i.e., expressed in some but not all the cell lines within the group, or present (P) at that specific stage. The analysis scheme further enables us to compute the probability of each possible expression profile over the four developmental stages, resulting in the fact that the most and second-to-most likely expression profile can be extracted. To investigate the feasibility of this analysis model, we extracted expression information from a number of genes with established expression patterns in B cell development (**Table 1**) and compared the result with that obtained with dCHIP analysis. Out of 37 genes, we found five that were not correctly classified as stage-specific from the P/A analysis and 10 that we did not detect as stage-restricted in the dCHIP analysis. The full P/A analysis is shown in **Supplementary Tables 1–7**, available online at http://www.jleukbio.org/. Thus, P/A-based probabilistic state analysis allowed for correct classification of several control genes, some of which could not be defined from the dCHIP analysis and vice versa, suggesting the two methods of data analysis to be complementary.

Predicted stage-restricted gene expression patterns can be verified by RT-PCR analysis

To validate the result of our data analysis, we randomly selected a number of genes suggested to be stage-specifically expressed and investigated their expression by RT-PCR analysis of a set of cell lines (**Fig. 6**). The transcription factor BACH-1, suggested from the P/A analysis to be specifically repressed at the B cell stage (supplementary tables), was

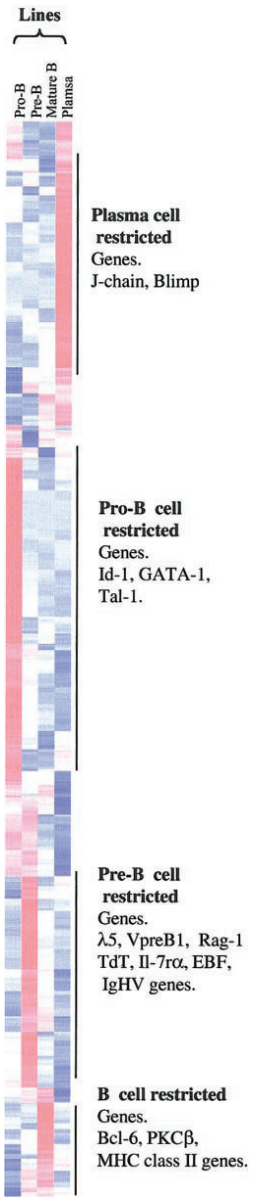


Fig. 4. Treatment of the data from different cell lines representing the same differentiation stage as replicates allows for the identification of stage-specific genes. The figure displays a dCHIP-generated cluster analysis of differential gene expression in the groups of cell lines after filtering and hierarchical clustering. The identification of control genes within the groups is indicated to the right of the color scheme. The full analysis with gene names of differentially expressed genes can be found as Attachment Figures 1–4. Expression scales ranging from –3 (blue) to +3 (red) are indicated below the data display. The figure only displays genes classified as present on more than one chip and to 0.50 < standard deviation/mean between <10.00. TdT, Terminal deoxynucleotidyl transferase; IgHV, Ig heavy-chain variable region; MCH, major histocompatibility complex.

expressed in pro-, pre-, and plasma cells but not in the B cell lines. The same data analysis indicated that the signal-transduction molecule Arhgef-3 was expressed specifically in pre-B cells, and although low levels of message could be detected in cell lines not belonging to the pre-B cell group, the expression appeared to be largely stage-restricted. A similar observation was made for the progenitor cell-restricted protein kinase PFTAIRE, which appeared to be expressed at high levels in the pro-B cells and only at a low level in one of the B cell lines. The P/A and the dCHIP analysis suggested another protein kinase, PKC-β, to be expressed mainly in the pre-B and B cells. mRNA encoding this protein could be detected in the pre-B and B cell lines and also in the Ly9D pro-B cells but not in BaF/3 or plasma cell lines. The dCHIP analysis suggested plasma cell-restricted expression of the signal-transduction molecules RhoB and Yes-associated protein, a finding confirmed by the RT-PCR analysis of the cell lines. This indicated that although specific discrepancies can be found, the overall picture of gene expression patterns using P/A or dCHIP analysis was well supported by the RT-PCR analysis.

DISCUSSION

A set of B cell lines can be used to define stage-specific genes

Here, we report a large-scale expression analysis of genes expressed in cell lines arrested at specific stages of B cell development. The calculated expression patterns of a large number of control genes, previously defined as expressed in the predicted pattern based on earlier experiments using cell line data and analysis of primary cells, indicate that analyzing B

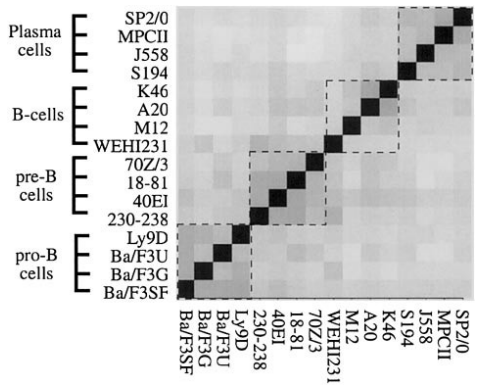


Fig. 5. P/A analysis allows for stage determination of B cell lines. The figure shows a distance matrix based on P/A analysis of the cell lines used for the generation of the data set. Ba/F3 1, 2, and 3 are differentially obtained subclones of the pro-B cell line Ba/F3, and Ly9D is an independently generated pro-B cell clone. 230–238, 40E1, and 18–81 are pre-B cell lines generated by Abelson virus transformation. 70/Z3 is a nitros–urea-induced pre-B cell line. WEHI231, A20, K46, and M12 are all defined as B cell lines, and S194, J558, SP2.0, and MPCII represent plasma cells. The scale ranges from highest similarity (black) to lowest (white).

TABLE 1. dCHIP and Probabilistic State Analysis Can Be Used As Complementary Methods for Gene Expression Analysis

Expression pattern	Probability	Secondary expression pattern	Probability 2	Gene name/reference	dCHIP classification
PAAA	0.934	MAAA	0.019	GATA-1 [26]	Pro-B
PAAA	0.934	MAAA	0.019	GATA-2 [26]	—
PAAA	0.934	MAAA	0.019	Id-1 [5]	Pro-B
APAA	0.790	APMA	0.157	Lef-1 [27]	—
APAA	0.934	AMAA	0.021	Sox-4 [28]	—
APAA	0.712	AMAA	0.243	Rag-1 [4]	Pre-B
APAA	0.790	APMA	0.157	λ5 [4]	Pre-B
APAA	0.790	APMA	0.157	VpreB [4]	Pre-B
AAPA	0.709	AAMA	0.246	SpiB [29]	—
APPA	0.924	APMA	0.024	CD19 [5, 30]	Pre-B-B
APPP	0.910	APPM	0.029	BOB-1 [31–33]	—
APPP	0.910	APPM	0.029	BLNK [34]	—
AMAA	0.955	AMAM	0.016	TdT [4]	Pre-B
APMM	0.922	APAM	0.024	EBF [35]	Pre-B
AAPM	0.702	AAMM	0.244	CD20 [36]	—
—	—	—	—	HSA [3]	Pre-B-B
APMA	0.922	APPA	0.027	Mb-1 [4]	Pre-B
—	—	—	—	Bcl-6 [23, 24]	B
—	—	—	—	Blimp-1 [34]	Plasma
—	—	—	—	J-chain [20]	Plasma
APMA	0.864	APAA	0.082	Pax-5 [19]	—
AMAA	0.931	AAAA	0.025	Rag-2 [4]	—
AAAM	0.539	AAAP	0.418	Syndecan [37]	—
—	—	—	—	Il-7rα [6]	Pre-B
APAA	0.934	AMAA	0.021	Clone BPS3.23 germline Ig variable region heavy chain precursor gene	Pre-B
APAA	0.934	AMAA	0.021	Immunoglobulin H-chain V-region pseudogene	Pre-B
APAA	0.934	AMAA	0.021	Clone BPS3.19 immunoglobulin heavy chain variable region precursor gene	Pre-B
APAA	0.934	AMAA	0.021	Germline immunoglobulin V(H)II gene H17	Pre-B
APAA	0.934	AMAA	0.021	Germline immunoglobulin V(H)II gene H8	Pre-B
APAA	0.934	AMAA	0.021	Clone BPS5.16 immunoglobulin heavy chain variable region precursor gene	Pre-B
APAA	0.934	AMAA	0.021	Ig B cell antigen receptor gene	Pre-B
APAA	0.790	APMA	0.157	Immunoglobulin heavy chain V DSP2.7-JH2 region (Igh) gene	Pre-B
APAA	0.790	APMA	0.157	Clone BHS2.19 immunoglobulin heavy chain variable region precursor gene	Pre-B
APAA	0.790	APMA	0.157	Immunoglobulin heavy and light chain variable region mRNA	Pre-B
APAA	0.790	APMA	0.157	Recombinant antineuraminidase single chain Ig VH and VL domains mRNA	Pre-B
APAA	0.790	APMA	0.157	Clone N1.1.b immunoglobulin heavy chain VDJ region gene	Pre-B
APAA	0.790	APMA	0.157	Immunoglobulin heavy chain gene, CDR3 region	Pre-B
APAA	0.712	AMAA	0.243	Germline immunoglobulin V(H)II gene H18	Pre-B
APAA	0.712	AMAA	0.243	Clone BPS3.26 immunoglobulin heavy chain variable region precursor, gene	Pre-B

Table 1 shows the calculated expression pattern of a set of control genes using dCHIP or P/A analysis. Genes were considered as present (P), absent (A), or mixed (M; S. Bilke et al., submitted) within the samples from each specific cell line group. The first classification goes for pro-B, the second to pre-B, the third to B, and the last for plasma-cell expression. For every gene, the second-highest probability to have another expression pattern is indicated as Probability 2. The gene name as well as reference to expression patterns are indicated in a separate column. The obtained information was compared with that resulting from a dCHIP analysis of the same data set using the same criteria as in Figure 4 (minimum present count 10% and 0.50<standard deviation/mean between <10.00).

cell development by the use of microarrays and cell lines arrested in defined stages results in reasonable approximation of gene expression patterns in B cell development. The analysis also suggests that cell lines defined as belonging to a specific developmental stage display a rather homogenous gene expression pattern. The group displaying the largest differences was the B cell group, possibly because these cell lines can arise at different anatomical sites and at different substages of differentiation. WEHI231 cells displayed an RNA expression profile closer related to the pre-B cells than to the B cells (Fig. 2). As a second, independently generated sample from this cell line gave the same result, we believe this might reflect that WEHI231 cells represent an immature BM-derived B cell. This idea is supported by reports suggesting that WEHI231 cells display defined features of immature B cells [21, 22]. The potential to obtain a degree of dynamic informa-

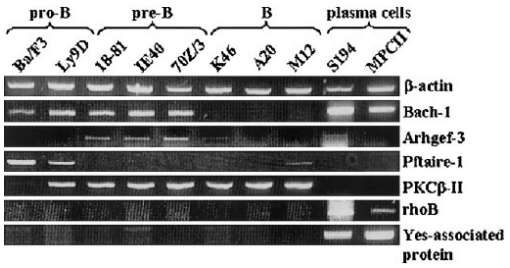


Fig. 6. Calculated gene expression patterns can be verified by RT-PCR analysis. The figure displays agarose gels, with the PCR products obtained using primers amplifying genes predicted to display restricted expression patterns (Attachment Figs. 1–4 and supplementary tables). The identities of the amplified mRNAs are indicated to the right, and the cell line used to generate cDNA is indicated on top of the panel. The PCR product has been visualized by ethidium bromide staining.

tion using stage-specifically arrested cell lines is also supported by analysis of the genes defined as expressed at more than one stage in the probabilistic state analysis. This is because the groups representing continuous expression patterns contained, on average, 17 genes (supplementary tables), and those representing discontinuous patterns contained, on average, six genes, a finding that indicates the existence of a flow of genetic information from the progenitor cell to the plasma cell. Our limited comparison of gene expression patterns in primary cells and cell lines indicates that the general feature of gene regulation is reasonably conserved in the transformed cells. This also suggests that the use of cell-surface markers on primary cells allows for a reasonable degree of enrichment of specific cell stages. This is also supported from microarray analysis of primary BM pre-B and B cells, where several of the same genes could be defined as stage-restricted [38, 39]. One major difference is lack of identification of cell-cycle genes that constitute a prominent group of genes when primary cells are used in expression analysis [38, 39]. This is probably explained by the fact that although only some of the primary cell stages are in cycle, all the cell lines are constitutively cycling, reducing the complexity of this part of the analysis.

Expression analysis of pre-B cell lines suggests simultaneous stage-restricted expression of several nonrearranged IgH genes

Another aspect of B cell development that does not become apparent when using sorted primary cell populations for gene expression analysis is reflected in the detection of RNA encoded by several V-region, heavy-chain (VH) genes, including pseudo-genes, specifically in the pre-B cell lines (Table 1). As the cell lines are of clonal origin, this could indicate that one and the same pre-B cell has the ability to express several VH genes simultaneously. These transcripts were in most cases not detected at the later stages of differentiation, and the mature cells, to a larger extent, expressed IgVL (V-region, light chain) genes. This is likely to reflect an ongoing rearrangement process of the heavy-chain gene in pre-B cell lines [40], with sterile expression of VH genes making them accessible for the recombination machinery [7, 41]. The expression of these V genes appears to be silenced at the later stages of development, possibly to ensure that no additional rearrangements of the heavy-chain genes occur during the assembly of the light-chain genes [8]. It may also be a mechanism contributing to allelic exclusion of the heavy chain to ensure that each single B cell only expresses one type of surface-bound Ig to avoid cross-reactive immune responses [8]. Thus, there may be a biological necessity in this rather complicated expression pattern, possibly demanding differential regulation of IgH promoters during B cell development. This type of information would not be extracted from the use of primary, sorted cells, as a broad expression of Ig genes could be explained by the heterogeneity of the sorted cell populations.

The Affymetrix™ P/A analysis is useful for the identification of stage-specific genes

Although the general picture of gene expression patterns was the same, independently of which method we used for the analysis of our data, the results from the (dCHIP) analysis differed to some extent from that obtained by P/A-based, probabilistic state analysis. As the dCHIP analysis takes into regard the relative transcription levels, there will be one group of genes that is expressed at all developmental stages but at different relative levels, which will be detected using dCHIP but not P/A analysis. These genes will be classified as present in all the groups and therefore not be detected as stage-specifically regulated in a P/A analysis. A bit more surprising was the detection of genes by the P/A-based method that we could not get classified in the dCHIP analysis. This is probably a result of that fact that rather small changes in relative expression values could change the classification from absent to present. Such an alteration might be classified as insignificant in the dCHIP analysis. This means that the P/A method gives a higher sensitivity of the data analysis, but at the same time, it will also increase the probability of detecting nonregulated genes. However, it appears that we in some cases detect different control genes using different analysis methods, indicating that the two approaches are complementary to each other.

We have not performed any extended analysis of the data we obtained as a result of the large amounts of information and the validity of the expression profile, as any individual gene needs further investigation. The analysis does, however, provide information that can be used to create a preliminary map of gene expression patterns that can be used to formulate working hypotheses for complex molecular events in B cell development.

ACKNOWLEDGMENTS

The Swedish Research Council, The Knut and Alice Wallenberg Foundation through the Swegene consortium, and The Swedish Cancer, Barncancer, Magnus Bergwall, Åke Wibergs, Österlunds, and The Crafoord foundations funded this work. P. T. and T. B. contributed equally to this work. We personally

thank the Swegen micro array facility in Lund as well as Drs. S. Cardell and Ramiro Gisler for critical reading of the manuscript.

REFERENCES

1. Ghia, P., ten Boekel, E., Rolink, A. G., Melchers, F. (1998) B-cell development: a comparison between mouse and man. *Immunol. Today* **19,** 480–485.
2. Jelinek, D. F. (2000) Regulation of B lymphocyte differentiation. *Ann. Allergy Asthma Immunol.* **84,** 375–385.
3. Hardy, R. R., Carmack, C. E., Shinton, S. A., Kemp, J. D., Hayakawa, K. (1991) Resolution and characterization of pro-B and pre-pro-B cell stages in normal mouse bone marrow. *J. Exp. Med.* **173,** 1213–1225.
4. Li, Y. S., Hayakawa, K., Hardy, R. R. (1993) The regulated expression of B lineage associated genes during B cell differentiation in bone marrow and fetal liver. *J. Exp. Med.* **178,** 951–960.
5. Li, Y. S., Wasserman, R., Hayakawa, K., Hardy, R. R. (1996) Identification of the earliest B lineage stage in mouse bone marrow. *Immunity* **5,** 527–535.
6. Tudor, K. S., Payne, K. J., Yamashita, Y., Kincade, P. W. (2000) Functional assessment of precursors from murine bone marrow suggests a sequence of early B lineage differentiation events. *Immunity* **12,** 335–345.
7. Grawunder, U., West, R. B., Lieber, M. R. (1998) Antigen receptor gene rearrangement. *Curr. Opin. Immunol.* **10,** 172–180.
8. Melchers, F., ten Boekel, E., Yamagami, T., Andersson, J., Rolink, A. (1999) The roles of preB and B cell receptors in the stepwise allelic exclusion of mouse IgH and L chain gene loci. *Semin. Immunol.* **11,** 307–317.
9. Nussenzweig, M. C. (1998) Immune receptor editing: revise and select. *Cell* **95,** 875–878.
10. Liu, Y. J., Arpin, C. (1997) Germinal center development. *Immunol. Rev.* **156,** 111–126.
11. Liu, Y. J., Banchereau, J. (1997) Regulation of B-cell commitment to plasma cells or to memory B cells. *Semin. Immunol.* **9,** 235–240.
12. MacLennan, I. C. M., Gulbranson-Judge, A., Toellner, K-M., Casamayor-Palleja, M., Chan, E., Sze, D. M-Y., Luther, S. A., Orbea, H. A. (1997) The changing preference of T and B cell partners as T-dependent antibody responses develop. *Immunol. Rev.* **156,** 53–66.
13. Li, C., Wong, W. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA* **98,** 31–36.
14. Palacios, R., Steinmetz, M. (1985) Il-3-dependent mouse clones that express B-220 surface antigen, contain Ig genes in germ-line configuration, and generate B lymphocytes in vivo. *Cell* **41,** 727–734.
15. Palacios, R., Karasuyama, H., Rolink, A. (1987) Ly1+ PRO-B lymphocyte clones. Phenotype, growth requirements and differentiation in vitro and in vivo. *EMBO J.* **6,** 3687–3693.
16. Sigvardsson, M., O'Riordan, M., Grosschedl, R. (1997) EBF and E47 collaborate to induce expression of the endogenous immunoglobulin surrogate light chain genes. *Immunity* **7,** 25–36.
17. Hermanson, G. G., Eisenberg, D., Kincade, P. W., Wall, R. (1988) B29: a member of the immunoglobulin gene superfamily exclusively expressed on beta-lineage cells. *Proc. Natl. Acad. Sci. USA* **85,** 6890–6894.
18. Sakaguchi, N., Melchers, F. (1986) Lambda 5, a new light-chain-related locus selectively expressed in pre-B lymphocytes. *Nature* **324,** 579–582.
19. Adams, B., Dorfler, P., Aguzzi, A., Kozmik, Z., Urbanek, P., Maurer-Fogy, I., Busslinger, M. (1992) Pax-5 encodes the transcription factor BSAP and is expressed in B lymphocytes, the developing CNS, and adult testis. *Genes Dev.* **6,** 1589–1607.
20. Lamson, G., Koshland, M. E. (1984) Changes in J chain and mu chain RNA expression as a function of B cell differentiation. *J. Exp. Med.* **160,** 877–892.
21. Monroe, J. G., Seyfert, V. L., Owen, C. S., Sykes, N. (1989) Isolation and characterization of a B lymphocyte mutant with altered signal transduction through its antigen receptor. *J. Exp. Med.* **169,** 1059–1070.
22. Yellen, A. J., Glenn, W., Sukhatme, V. P., Cao, X. M., Monroe, J. G. (1991) Signaling through surface IgM in tolerance-susceptible immature murine B lymphocytes. Developmentally regulated differences in transmembrane signaling in splenic B cells from adult and neonatal mice. *J. Immunol.* **146,** 1446–1454.
23. Cattoretti, G., Chang, C. C., Cechova, K., Zhang, J., Ye, B. H., Falini, B., Louie, D. C., Offit, K., Chaganti, R. S., Dalla-Favera, R. (1995) BCL-6 protein is expressed in germinal-center B cells. *Blood* **86,** 45–53.
24. Falini, B., Bigerna, B., Pasqualucci, L., Fizzotti, M., Martelli, M. F., Pileri, S., Pinto, A., Carbone, A., Venturi, S., Pacini, R., Cattoretti, G., Pescarmona, E., Lo Coco, F., Pelicci, P. G., Anagnastopoulos, I., Dalla-Favera, R., Flenghi, L. (1996) Distinctive expression pattern of the BCL-6 protein in nodular lymphocyte predominance Hodgkin's disease. *Blood* **87,** 465–471.
25. Turner, C. A., Mack, D. H., Davis, M. M. (1994) Blimp-1, a novel zinc finger-containing protein that can drive the maturation of B lymphocytes into immunoglobulin-secreting cells. *Cell* **77,** 297–306.
26. Weiss, M. J., Orkin, S. H. (1995) GATA transcription factors: key regulators of hematopoiesis. *Exp. Hematol.* **23,** 99–107.
27. Travis, A., Amsterdam, A., Belanger, C., Grosschedl, R. (1991) LEF-1, a gene encoding a lymphoid-specific protein with an HMG domain, regulates T-cell receptor α enhancer function. *Genes Dev.* **5,** 880–894.
28. van de Wetering, M., Oosterwegel, M., van Norren, K., Clevers, H. (1993) Sox-4, an Sry-like HMG box protein, is a transcriptional activator in lymphocytes. *EMBO J.* **12,** 3847–3854.
29. Su, G. H., Ip, H. S., Cobb, B. S., Lu, M-M., Chen, H-M., Simon, M. C. (1996) The Ets protein Spi-B is expressed exclusively in B cells and T cells during development. *J. Exp. Med.* **184,** 203–214.
30. Kozmik, Z., Wang, S., Dorfler, P., Adams, B., Busslinger, M. (1992) The promoter of the CD19 gene is a target for the B-cell-specific transcription factor BSAP. *Mol. Cell. Biol.* **12,** 2662–2672.
31. Luo, Y., Roeder, R. G. (1995) Cloning, functional characterization, and mechanism of action of the B-cell-specific transcriptional coactivator OCA-B. *Mol. Cell. Biol.* **15,** 4115–4124.
32. Strubin, M., Newell, J. W., Matthias, P. (1995) OBF-1, a novel B cell-specific coactivator that stimulates immunoglobulin promoter activity through association with octamer-binding proteins. *Cell* **80,** 497–506.
33. Gstaiger, M., Knoepfel, L., Georgiev, O., Schaffner, W., Hovens, C. M. (1995) A B-cell coactivator of octamer-binding transcription factors. *Nature* **373,** 360–362.
34. Fu, C., Turck, C. W., Kurosaki, T., Chan, A. C. (1998) BLNK: a central linker protein in B cell activation. *Immunity* **9,** 93–103.
35. Hagman, J., Belanger, C., Travis, A., Turck, C. W., Grosschedl, R. (1993) Cloning and functional characterization of early B-cell factor, a regulator of lymphocyte-specific gene expression. *Genes Dev.* **7,** 760–773.
36. Tedder, T. F., Klejman, G., Disteche, C. M., Adler, D. A., Schlossman, S. F., Saito, H. (1988) Cloning of a complementary DNA encoding a new mouse B lymphocyte differentiation antigen, homologous to the human B1 (CD20) antigen, and localization of the gene to chromosome 19. *J. Immunol.* **141,** 4388–4394.
37. Sanderson, R., Lalor, P., Bernfield, M. (1989) B lymphocytes express and lose syndecan at specific stages of differentiation. *Cell Regul.* **1,** 27–35.
38. Hoffman, R., Seidl, T., Neeb, M., Rolink, A., Melchers, F. (2002) Changes in gene expression profiles in developing B cells of murine bone marrow. *Genome Res.* **12,** 98–111.
39. Hoffman, R., Bruno, L., Siedl, T., Rolink, A., Melchers, F. (2003) Rules for gene usage inferred from a comparison of large-scale gene expression profiles of T and B lymphocyte development. *J. Immunol.* **170,** 1339–1353.
40. Schlissel, M. S., Corcoran, L. M., Baltimore, D. (1991) Virus-transformed pre-B cells show ordered activation but not inactivation of immunoglobulin gene rearrangement and transcription. *J. Exp. Med.* **173,** 711–720.
41. Yancopoulos, G. D., Alt, F. W. (1985) Developmentally controlled and tissue-specific expression of unrearranged VH gene segments. *Cell* **40,** 271–281.

Paper III

LU TP 02-14

Probabilistic estimation of microarray data reliability and underlying gene expression

Sven Bilke[*1], **Thomas Breslin**[*2] **and Mikael Sigvardsson**[†3]

[*]Complex Systems Division, Department of Theoretical Physics
University of Lund, Sölvegatan 14A, SE-223 62 Lund, Sweden

[†]The Laboratory for Cell Differentiation Studies
Department for Stem Cell Biology, BMC B12, SE-22185 Lund, Sweden

BMC Bioinformatics **4**, 2003

Abstract

Background: The availability of high throughput methods for measurement of mRNA concentrations makes the reliability of conclusions drawn from the data and global quality control of samples and hybridization important issues. We address these issues by an information theoretic approach, applied to discretized expression values in replicated gene expression data.

Results: Our approach yields a quantitative measure of two important parameter classes: First, the probability $P(\sigma|S)$ that a gene is in the biological state σ in a certain variety, given its observed expression S in the samples of that variety. Second, sample specific error probabilities which serve as consistency indicators of the measured samples of each variety. The method and its limitations are tested on gene expression data for developing murine B-cells and a t-test is used as reference. On a set of known genes it performs better than the t-test despite the crude discretization into only two expression levels. The consistency indicators, i.e. the error probabilities, correlate well with variations in the biological material and thus prove efficient.

Conclusions: The proposed method is effective in determining differential gene expression and sample reliability in replicated microarray data. Already at two discrete expression levels in each sample, it gives a good explanation of the data and is comparable to standard techniques.

[1]sven@thep.lu.se
[2]thomas@thep.lu.se
[3]Mikael.Sigvardsson@stemcell.lu.se

Background

A broad variety of algorithms has been developed and used to extract biologically relevant information from gene expression data. Among others commonly used are visual inspection [1], hierarchical and k-means clustering [2], self organizing maps [3, 4] and singular value decomposition [5, 6]. These methods aim mainly at identifying predominant patterns and thus groups of "cooperating" genes based on the assumption that related genes have similar expression patterns.

Compared to the amount of work devoted to efficient methods to extract information from the data, somewhat less attention has been paid to the question of the reliability of the generated results. The ANOVA analysis [7] allows estimation, and thus elimination, of some systematic error sources. Bootstrapping cluster analysis estimates the stability of cluster assignments [8] based on artificial data-sets generated with ANOVA coefficients. Some authors also considered the question of how well a certain oligo [10] is suited to measure the mRNA expression level of the related gene.

Some work has gone towards the ambitious task of learning topological properties or qualitative features of the genetic regulatory network from expression profiles, see e.g. [11]. A major limiting factor in these attempts is the comparative sparseness of available data. It is therefore reasonable to consider reduced models, for example a Boolean representation of the gene activity. It is known that many biological properties, for instance stability and hysteresis, can be modeled by the dynamics of such reduced models [12, 13, 14].

In this work we investigate the possibility of reducing complexity of gene expression data by discretizing the expression levels. The approach we present enables a new way of extracting biologically relevant information from the data in the following way: A biological variety, i.e. a biological system defined by the investigator, is represented by several samples which are subjected to gene expression analysis. If gene expression levels are discretized into n values, and the variety is represented by m samples, the number of observable expression states for a gene are limited to n^m. These observed states S are modeled as being derived from a smaller number of underlying, biological states σ, through a measurement process. Rather than making static assignments $S \rightarrow \sigma$ we calculate conditional probabilities $P(\sigma|S)$. The number of possible expression profiles for a gene over a set of varieties is limited and the probability of each expression profile is easily calculated. Since the model we use considers both the underlying biology and the measurement process it also generates a measure of sample coherence in each biological variety.

We demonstrate the feasibility of this approach for a *binary* discretization of gene expression. For the discretization step we use the absent/present classification provided by the Affymetrix software [9]. The outcome of our method on a data set covering gene expression in developing murine B-cells is compared to the results of a standard analysis. We show that even with the crude discretization into only two expression levels the method is competitive to statistical methods based on continuous expression levels.

Methods

The Model

A major step in the analysis of gene expression data is to separate the biological content of the data from measurement and sample specific errors. In other words given an observation, i.e. the expression values of a gene in several samples representing the same biological variety, one wants to conclude on the biological state σ, which generated the observation. This can be expressed as a conditional probability,

$$P(\sigma|S) \tag{1}$$

that a gene is in a certain biological state σ given the corresponding observed state S. In the application on which we demonstrate the method we consider three different biological varieties: pro, pre, and mature B-cells. The samples in each variety are different cell lines arrested at the corresponding stage of development.

In this work we take an information theoretic point of view to estimate this probability: The information of interest, the state σ, is "transmitted" in a noisy measurement process and potentially distorted (Figure 1). Using Bayes' theorem, the desired conditional probability Eq. (1) can be expressed as:

$$P(\sigma|S) = \frac{P(S|\sigma)P_\sigma}{P_S} \ . \tag{2}$$

On the right hand side of this equation, $P(S|\sigma)$ is the probability to observe state S if the underlying biological state is σ. In a sense, $P(S|\sigma)$ describes the noise characteristic of the measurement process. In the following we will show how this conditional probability, and the other probabilities on the r.h.s. of Eq. (2) can be estimated.

Given a set of m samples representing the same biological variety, differences in the expression level of a gene between the samples can arise from two independent sources:

1. *Random variation* within the variety. This may be caused by temporal differences in response to the stimuli, slightly different environmental conditions, genotypic differences between samples, etc.
2. *Sample specific errors*. These are mainly caused by the measurement process, e.g. differences in the treatment of the mRNA, scratched arrays, and so on. However, outlier samples, cultured under considerably different conditions, also contribute to sample specific errors.

A separation of these two contributions is possible only with an appropriate model for the variation of gene expression between the samples. In the choice of model, one has

considerable freedom within the bounds set by biological plausibility. A limiting factor on the biological model comes from the type and amount of available data. The data used in this work contains only four samples for each variety. For the model we propose this is the *minimum* number of samples required to estimate the model parameters.

In the discretization of gene expression levels, we use only two discrete values, 0 and 1, for the expression of a gene in a sample. This means that the number of observable states, S, in a variety consisting of m samples is 2^m. With no measurement errors we could immediately conclude on the underlying biological state σ: the two cases, where all observations agree $S = (1, \ldots, 1)$ and $S = (0, \ldots, 0)$ can be mapped to the biological states $\sigma_{\mathbf{1}}$ and $\sigma_{\mathbf{0}}$ respectively, which describe "pure" states without variation. The remaining $N-2$ observable states S, where the individual measurements disagree, correspond to biological states σ with random variation. For the application in our biological study with supposedly *identical* biological systems contributing to the observable states S, the exact pattern leading to contradicting observations does not carry any information, as long as we assume that there are no sample specific errors. Therefore, we subsummize all $N-2$ possible observations as one biological state $\sigma_{\mathbf{r}}$ with a random variation.

The biological rationale for this model is given by the following example: If one considers a biological variety such as cells in the retina of the eye, then a certain number of crucial genes ought to be expressed in all samples. Such genes might include rhodopsin, a molecule that responds to light. In contrast, genes such as the hemoglobin family, which are typical of erythrocytes, ought not to be expressed in the retina. A third class of genes could be considered as independent of the system in the sense that their expression is not directly related to the biological system. Such genes may vary in expression both due to environmental and genetic differences between the samples.

The model discussed so far is depicted graphically in the left part of Figure 1, where a possible distribution of the relative frequencies of the three biological states is depicted, for the case of $m = 4$ samples. The distribution can be described by three numbers: the probabilities $P(\sigma_{\mathbf{1}})$ and $P(\sigma_{\mathbf{0}})$, which contribute to the frequencies of the states $S = (1, 1, 1, 1)$ and $S = (0, 0, 0, 0)$, and $P(\sigma_{\mathbf{r}})$ which contributes to both the frequency of mixed states and the two states above. Describing the mixed states with only one parameter $P(\sigma_{\mathbf{r}})$ implies that the biological variation is modeled evenly and identically distributed independently for each sample. In a second step, the measurement process with possible sample specific errors is modeled as statistically independent between samples. For each sample i, we define two parameters, $P^i_{0\to 1}$ and $P^i_{1\to 0}$, denoted sample specific *error probabilities*.

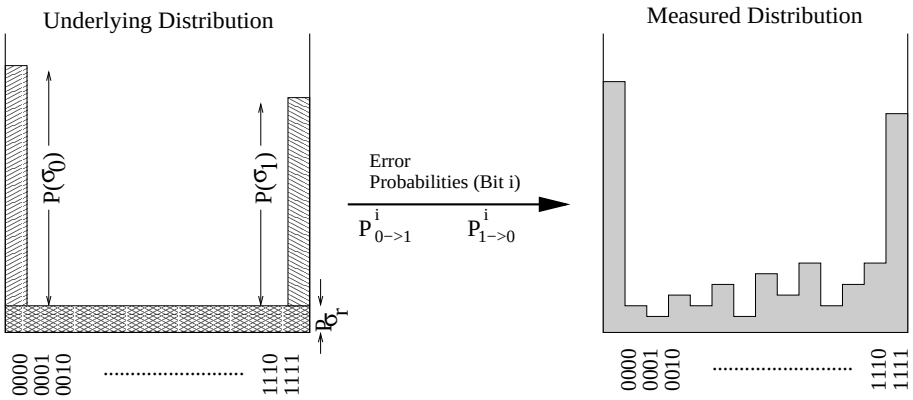


Figure 1: Schematic diagram illustrating the transition from underlying to observed distributions of states, in the case of $m = 4$ samples. The underlying distribution on the left hand side can be described by the probabilities for each underlying state, $P(\sigma_{\mathbf{1}})$, $P(\sigma_{\mathbf{0}})$, and $P(\sigma_{\mathbf{r}})$ (see text). This distribution is then distorted by sample specific errors, $P^i_{0\to 1}$ and $P^i_{1\to 0}$, resulting in an experimentally observed distribution, depicted on the right hand side.

To introduce the full formalism of our current model we start by considering a simple example, again for $m = 4$ samples. An observed state S, $S \equiv (S_1, S_2, S_3, S_4) = (1, 0, 1, 0)$, may be generated by the gene being in state $\sigma_{\mathbf{1}}$ with the probability:

$$P(\sigma_{\mathbf{1}})(1 - P^1_{1\to 0})(P^2_{1\to 0})(1 - P^3_{1\to 0})(P^4_{1\to 0}),$$

or it may be generated by the gene being in state $\sigma_{\mathbf{0}}$ with the probability:

$$P(\sigma_{\mathbf{0}})(P^1_{0\to 1})(1 - P^2_{0\to 1})(P^3_{0\to 1})(1 - P^4_{0\to 1}),$$

or it may be generated by the gene being in state $\sigma_{\mathbf{r}}$ with the probability:

$$\begin{array}{l} P(\sigma_{\mathbf{r}})\times \\ \frac{1}{2}[(P^1_{0\to 1}) + (1 - P^1_{1\to 0})]\frac{1}{2}[(1 - P^2_{0\to 1}) + (P^2_{1\to 0})]\times \\ \frac{1}{2}[(P^3_{0\to 1}) + (1 - P^3_{1\to 0})]\frac{1}{2}[(1 - P^4_{0\to 1}) + (P^4_{1\to 0})] \ . \end{array}$$

With the briefer notation,

$$\begin{array}{rcl} \rho^i_{1\to 0} & \equiv & P^i_{1\to 0}\delta_{S_i,0} + (1 - P^i_{1\to 0})\delta_{S_i,1} \\ \rho^i_{0\to 1} & \equiv & P^i_{0\to 1}\delta_{S_i,1} + (1 - P^i_{0\to 1})\delta_{S_i,0} \end{array}$$

where δ refers to the Kronecker delta (i.e. $\delta_{j,k} = 1$ if $j = k$ and 0 otherwise), we may express the distribution of observed states, in the general case of binary discretization with m samples, as:

$$P(S) = P(\sigma_{\mathbf{1}})\prod_{i=1}^{m}\rho^i_{1\to 0} + P(\sigma_{\mathbf{0}})\prod_{i=1}^{m}\rho^i_{0\to 1} + P(\sigma_{\mathbf{r}})\prod_{i=1}^{m}\frac{1}{2}(\rho^i_{1\to 0} + \rho^i_{0\to 1}) \ . \qquad (3)$$

Altogether the model uses $3 + 2 * m$ variables. These parameters $P(\sigma_{\mathbf{1}})$, $P(\sigma_{\mathbf{0}})$, $P(\sigma_{\mathbf{r}})$ and $\{P^i_{1\to 0}, P^i_{0\to 1}\}_{i=1}^m$ are estimated from the observed distribution of states (right side of Figure 1) by Levenberg-Marquardt [15] chi-square minimization of the unweighted error to the theoretical distribution Eq. (3). Using Eq. (2), and the parameters estimated as above, our belief that a gene belongs to the underlying states $\sigma_{\mathbf{0}}$, $\sigma_{\mathbf{1}}$, $\sigma_{\mathbf{r}}$, given the $2^4 = 16$ observable states S, can now be expressed as:

$$\begin{aligned} P(\sigma_{\mathbf{0}}|S) &= \tfrac{P(S|\sigma_{\mathbf{0}})P(\sigma_{\mathbf{0}})}{P(S)} \\ P(\sigma_{\mathbf{1}}|S) &= \tfrac{P(S|\sigma_{\mathbf{1}})P(\sigma_{\mathbf{1}})}{P(S)} \\ P(\sigma_{\mathbf{r}}|S) &= \tfrac{P(S|\sigma_{\mathbf{r}})P(\sigma_{\mathbf{r}})}{P(S)} \end{aligned}$$

Once the probability that a gene is in a certain biological state $\Sigma^i \in \sigma_{\mathbf{1}}, \sigma_{\mathbf{0}}, \sigma_{\mathbf{r}}$ has been calculated for all varieties $i = 1 \ldots v$, one can calculate the probability that a gene exhibits a certain expression profile over a set of different varieties by taking the product

$$P(\Sigma^1, \ldots, \Sigma^v | S^1, \ldots, S^v) = \prod_{i=1}^{v} P(\Sigma^i | S^i) \; . \tag{4}$$

In this way, the probabilistic state analysis also generates a clustering: For a given expression profile over the varieties, e.g. $\sigma^1_{\mathbf{0}}\sigma^2_{\mathbf{r}} \ldots \sigma^i_{\mathbf{1}}$, we may extract those genes for which this expression profile is the most probable. In fact this is a "soft" clustering, in that an expression profile can belong to several clusters simultaneously with different probabilities. Moreover the genes clustered to a biologically interesting expression profile can be ranked by the probability of Eq. (4).

Experimental data preparation

All cells were grown in RPMI medium supplemented with 7.5% fetal calf serum, 10 mM HEPES, 2 mM pyruvate, 50 mM 2-mercaptoethanol and 50 mg gentamicin per ml (complete RPMI media) (all purchased from Life Technologies AB, T=E4by, Sweden) at 37=B0C and 5% CO2. RNA was prepared using Trizol (GIBCO) and 7.5 =B5g of total RNA was annealed to a T7-oligo T primer by denaturation at 70=B0C for 10 minutes followed by 10 minutes of incubation of the samples on ice. First strand synthesis was performed for 2 hours at 42=B0C using 20 U of Superscript Reverse Transcriptase (GIBCO) in buffers and nucleotide mixes according to the manufacturers instructions. This was followed by a second strand synthesis for 2 hours at 16=B0C, using RNAseH, E coli DNA polymerase I and E coli DNA ligase (all from GIBCO), according to the manufacturers instructions. The obtained double stranded cDNA was then blunted by the addition of 20 U of T4 DNA polymerase and incubation for 5 minutes at 16=B0C. The material

was then purified by Phenol:Cloroform:Isoamyl alcohol extraction followed by precipitation with NH4Ac and Ethanol. The cDNA was then used in an in vitro transcription reaction for 6 h at 37 =B0C using a T7 IVT kit and biotin labeled ribonucloetides. The obtained cRNA was purified from unincorporated nucleotides on a RNAeasy column (Qiagen). The eluted cRNA was then fragmented by incubation of the products for two hours in fragmentation buffer (40 mM Tris-acetate, pH 8.1, 100 mM KOAc, 150 mM MgOAc). 20 =B5g of the final fragmented cRNA was then hybridized to affymetrix chip U74Av2 (Affymetrix) in 200 =B5l hybridization buffer (100 mM MES-buffer, pH 6.6, 1 M NaCl, 20 mM EDTA, 0.01Herring sperm DNA (100 =B5g/ml) and Acetylated BSA (500 =B5g/ml) in an Affymetrix Gene Chip Hybridization oven 320. The chip was then developed by the addition of FITC-streptavidin followed by washing using an Affymetrix Gene Chip Fluidics Station 400. Scanning was performed using a Hewlett Packard Gene Array Scanner.

Results

To evaluate the method we used both real and synthetic data. The experimental data was generated with Affymetrix microarrays for the study of differentiating murine B-cells at different stages in the differentiation process. In this publication the data is only used to demonstrate the feasibility of the proposed method. The biological implications of this study are published elsewhere [20].

Synthetic data and the effect of correlation

For synthetic data, generated with the model parameters $P(\sigma_{\mathbf{0}}) = 0.45$, $P(\sigma_{\mathbf{1}}) = 0.35$, $P(\sigma_{\mathbf{r}}) = 0.2$ and $P^i_{1\to 0} = P^i_{0\to 1} = 0.02$ for all samples i, parameter estimates are, as expected, given with low errors. The parameter values were chosen as typical values from the estimates on real data, see next section, and the result was verified for sample sizes $m = 4$, $m = 5$, and $m = 6$ (data not shown).

An assumption of simple model used to derive Eq. (3) is that randomly varying genes vary *independently* in the samples of a variety. Hence we investigated how severely this assumption influences the estimation of the model parameters.

To assess the influence of correlations between randomly varying genes we generated a data set consisting of four bits, i.e. samples, with the same parameters as above. In the random patterns a correlation was introduced between the third and fourth bit by changing the value of of the fourth to that of the third with a certain probability. We define this probability as the correlation factor. The correlation was introduced before distorting the patterns with error probabilities. We then plotted the mean error in the

estimation of parameters over 500 runs of synthetically generated data for correlation factors in the range $\{0, 0.02, \ldots, 0.98\}$.

Figure 2 shows the error in the estimation of the parameters describing the underlying distribution. We notice that even for fully correlated patterns the estimation error is less than 20% of the correct values. The estimation of the probability for biologically varying genes is somewhat worse, for fully correlated patterns the error is almost 50%. For real data one can, however, expect a much smaller correlation. The average error in the estimates of the error probabilities, as seen in Figure 3, shows the expected behavior: The average error grows with the correlation for the uncorrelated samples, while the estimate for the correlated observations is almost unaffected. Intuitively, the model compensates for the correlation by increasing $P(\sigma_{\mathbf{1}})$ and $P(\sigma_{\mathbf{0}})$ as well as the error probabilities and lowering $P(\sigma_{\mathbf{r}})$. For correlation factors above 0.50, due to the compensation effect, the model deteriorates in explaining the data. This can be seen in the sum $P(\sigma_{\mathbf{0}}) + P(\sigma_{\mathbf{r}}) + P(\sigma_{\mathbf{1}})$ which initially drops from almost 1 to 0.99 as the correlation factor rises from 0 to 0.50 and then from 0.99 to 0.96 for correlation factors in the range 0.50 to 0.98 (data not shown). We hence conclude that it is reasonable not to impose the condition $P(\sigma_{\mathbf{0}}) + P(\sigma_{\mathbf{1}}) + P(\sigma_{\mathbf{r}}) = 1$ in the model, as this sum indicates if samples are strongly correlated in genes whose expression vary around the threshold of discretization.

In summary, for not too large correlations in the biological variance the algorithm gives a good quantitative estimate of the model parameters. In the case of large correlations the qualitative picture given by the estimated parameters is still reliable.

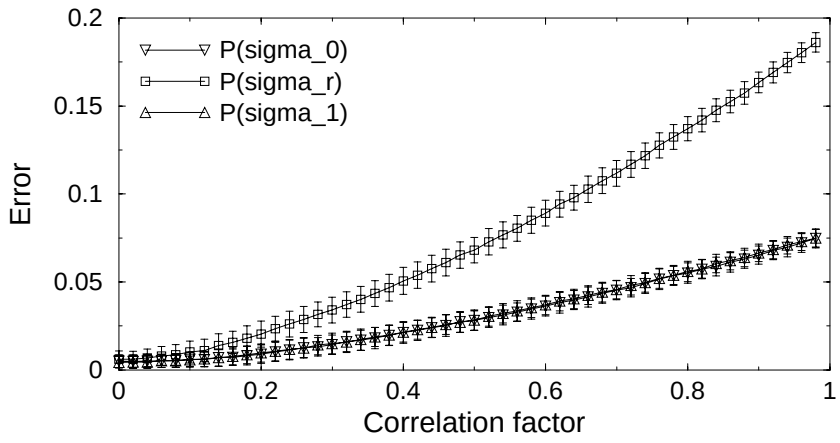


Figure 2: The average error in the estimation of the parameters $P(\sigma_{\mathbf{1}})$, $P(\sigma_{\mathbf{0}})$, $P(\sigma_{\mathbf{r}})$ are given as a function of correlation factor between the third and fourth bit. For correlation factors above 0.2 the error in $P(\sigma_{\mathbf{r}})$ rises considerably.

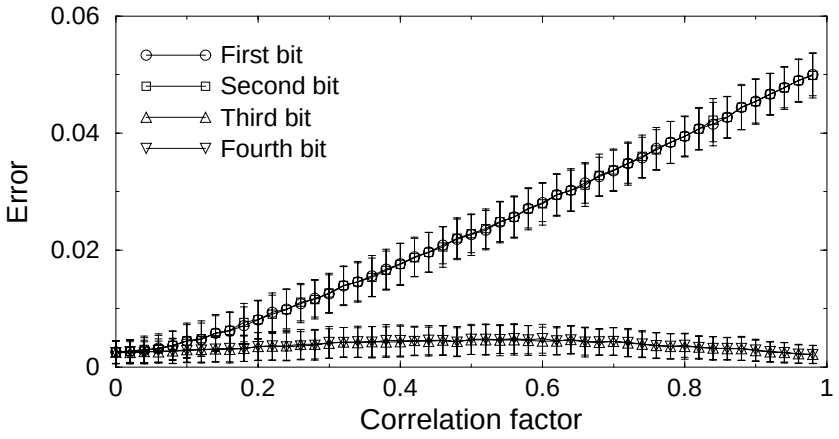


Figure 3: The average error in the estimation of the error probabilities $\{P^i_{0\to 1}\}^4_{i=1}$. For correlation factors above 0.2 $P^1_{1\to 0}$ and $P^2_{1\to 0}$ are notably raised. Patterns were these bits deviate from the other two are then not considered as random but rather caused by an error. This effect could only be avoided by introducing extra parameters for correlation between bits.

Real data

Differentiating B-cells are characterized by phenotypic markers into different stages of development. Here we chose to study the expressional differences between three such stages; pro, pre and mature B-cells. For each of these three varieties we used four different cell lines arrested at the corresponding stage of development. Measurements we performed with Affymetrix array containing probesets for 12488 genes and ESTs on each sample. The discretization of expression levels was given by the Affymetrix GeneChip absent present calls [9].

Our algorithm was used to estimate the parameters $P(\sigma_{\mathbf{1}})$, $P(\sigma_{\mathbf{0}})$ and $P(\sigma_{\mathbf{r}})$, describing the biological distribution and the error probabilities (see Table 1). Theoretically, one expects the three biological probabilities to sum up to one. In our model, Eq. (3), we do not explicitly impose this condition. Nevertheless, the sum of the independently estimated parameters is close to one. This indicates that our model is a reasonable approximation of the biological system and the measurement process.

The error probabilities from Eq. (3) can be used as a consistency index for the samples in a given variety. In the last variety (mature B-cells) the maximum error probability is notably higher. This effect is likely to be explained by the different anatomical origins of the cell lines representing this group. No such differences exist in the other groups since they all

originate in the bone marrow which is the only anatomical site for B cell development in the adult animal [16]. In contrast, the mature B cell can reside in several other sites such as spleen, lymph-nodes and intestine which may affect the gene expression profile in these cells [17, 18]. With only four samples, it is not unlikely that these effects show up in the error probabilities and not only in the random variation parameter $P(\sigma_{\mathbf{r}})$.

	$P(\sigma_{\mathbf{1}})$	$P(\sigma_{\mathbf{0}})$	$P(\sigma_{\mathbf{r}})$	Max P_e	Min P_e	Median P_e
Pro B-cells	0.405	0.460	0.135	0.035	0.0002	0.020
Pre B-cells	0.395	0.450	0.155	0.047	0.003	0.028
Full B-cells	0.343	0.471	0.186	0.073	0.0007	0.022

Table 1: Typical paramter values. Summary of the estimated parameter values for the B-cell data. P_e refers to the set of error probabilities, i.e., $[P^i_{1\to 0}, P^i_{0\to 1}]_{i=1}^4$.

Comparison to conventional t-test on known genes

To determine how well biologically relevant information can be extracted from the discretized data, we compare it with another statistical method based on continuous expression values. We use our method to identify differences in gene expression between two varieties in the following way. A gene that goes up between variety i and variety j is characterized by the states $\sigma^i_{\mathbf{0}}$, $\sigma^j_{\mathbf{1}}$ or $\sigma^i_{\mathbf{0}}$, $\sigma^j_{\mathbf{r}}$ or $\sigma^i_{\mathbf{r}}$, $\sigma^j_{\mathbf{1}}$. Hence the belief that a gene goes up is given by the probability:

$$P(\text{up between variety } i \text{ and } j) = P(\sigma^i_{\mathbf{0}})P(\sigma^j_{\mathbf{1}}) + P(\sigma^i_{\mathbf{0}})P(\sigma^j_{\mathbf{r}}) + P(\sigma^i_{\mathbf{r}})P(\sigma^j_{\mathbf{1}}),$$

suppressing the conditional probabilities, $P(\cdot|S)$, for brevity. Similarly, the belief that a gene goes down is given by the probability:

$$P(\text{down between variety } i \text{ and } j) = P(\sigma^i_{\mathbf{1}})P(\sigma^j_{\mathbf{0}}) + P(\sigma^i_{\mathbf{1}})P(\sigma^j_{\mathbf{r}}) + P(\sigma^i_{\mathbf{r}})P(\sigma^j_{\mathbf{0}})\ .$$

Taking $1 - P(\text{up})$ thus yields the Bayesian p-value of a gene going up. To answer the same question when working on continuous expression data one possibility is to employ a one sided two sample t-test in the Welch approximation of unknown variances in the varieties. This enables testing, for each gene, whether the mean of expression is higher or lower in one variety than in another. For comparison of these two approaches we selected a set of genes based on their well documented expression pattern and biological functions in the developing B lymphocyte [16, 19]. Several of these are functionally linked since they participate directly in somatic DNA rearrangement events occurring specifically at the pre-B cell stage or participate in the regulation of genes involved in this process and thus display restricted expression patterns (pre-B specific). A second set of genes were selected based on their expression in cells that are either committed to the B lineage (B-lineage

specific genes, in pre-B and B-cells) or non committed to this developmental pathway (Not in B-lineage, expressed in pro-B cells) [21].

The result of this comparison is presented in Table 2. For 14 out of the 22 genes the two methods completely agree. Out of these 14 only one (Mb1) does not match the expected target profile. For the other 8 genes, where the two methods yield different results, the probabilistic state analysis gives the expected answer in 5 cases, which should be compared to the two cases, where the *t*-test gives the right answer. In one case (rag-1), neither of the two methods gives the expected result.

For the subset of genes considered here, our method has an advantage of $5 : 2$ in giving the correct (i.e. expected) expression pattern. However, the number of samples is not big enough to draw firm conclusions from this result.

Conclusions

The method we have presented serves several purposes:

1. It gives a measure of the biological variation of the genes' expression in different varieties.
2. It estimates each hybridizations' global error probabilities. These parameters are very useful as they serve as quality/consistency indicators of the samples of each variety.
3. Given the parameters above, it estimates the probability of a gene belonging to each of the three groups $\sigma_{\mathbf{0}}$, $\sigma_{\mathbf{r}}$ and $\sigma_{\mathbf{1}}$. These probabilities in turn indicate weather the gene is likely to be below, fluctuating around or above the threshold of discretization.
4. Clustering of genes to expression profiles over a set of different varieties is achieved with Eq. (4). The probability, i.e. belief, that a gene belongs to a certain cluster is exactly quantified.

This novel approach is proven valuable for quantifying both data reliability and underlying gene expression in microarray experiments. Our method has been successfully applied in two different projects [22] [20].

		Discrete			Continuous			
Gene	Id	I	II	III	I	II	III	Literature
Target profile	-	**S**	**U**	**U**	**S**	**U**	**U**	
CD20	99446_at	S	U	U	S	U	U	B-Cell
Spi	93657_at	S	U	U	S	S	S	specific
Target profile	-	**U**	**D**	**S**	**U**	**D**	**S**	
Sox-4	160109_at	U	D	S	U	D	S	
lef-1	103628_at	U	D	S	U	S	S	
rag-1	93683_at	U	S	S	S	S	U	
VpreB	92972_at	U	D	S	U	D	S	Pre-B
Lambda-5	99429_at	U	D	S	U	D	S	specific
Il-7 receptor	99030_at	U	D	S	U	S	S	
TdT	99030_at	U	D	S	U	S	S	
Target profile	-	**U**	**S**	**U**	**U**	**S**	**U**	
Bob-1	93915_at	U	S	U	U	S	U	
CD 19	99945_at	U	S	U	U	D	U	
Blnk	100771_at	U	S	U	U	S	U	
Pax-5	96993_at	U	S	U	U	S	U	B-lineage
Blk	92359_at	U	S	U	U	S	U	Specific
Mb-1	102778_at	U	D	S	U	D	S	
B29	161012_at	S	S	S	U	S	U	
CD24	100600_at	U	S	U	U	S	U	
Target profile	-	**D**	**S**	**D**	**D**	**S**	**D**	
Id-1	100050_at	D	S	D	D	S	D	
Fag-1	97974_at	S	D	D	D	S	D	Not in
Il-3 receptor	94747_at	D	S	D	D	S	D	B-lineage
CD 63	160493_at	D	S	D	D	S	D	
Gata-2	102789_at	D	S	D	D	D	D	

Table 2: *t*-test vs. probabilistic analysis of gene expression levels. The three groups I, II and III indicate the expressional changes between *Pro-B to Pre-B*, *Pre-B to Mature-B*, and *Pro-B to Mature-B* respectively. U stands for accepting the hypothesis up, D for down, and S (stable) if no hypothesis could be accepted on the 95% confidence level.

Authors' Contributions

SB and TB have contributed to the development of the model and implemented the algorithms. MS has contributed the experimental data and biological expertise.

Acknowledgments

SB and TB are supported by Knut and Alice Wallenberg Foundation through the SWEGENE consortium. MS is supported by the swedish science council and the A. Österlund foundation. The authors also extend their gratitude to Carsten Peterson, Patrik Edén, Jari Häkkinen and Markus Rignér, for fruitful discussions and careful proofreading of the manuscript.

References

[1] RJ Cho, MJ Campbell, EA Winzeler, L Steinmetz, A Conway, L Wodicka, TG Wolfsberg, AE Gabrielian, D Landsman, DJ Lockhart: Genome-Wide transcriptional analysis of the mitotic cell cycle. *Mol.Cel* 1998, 2: 65-73

[2] MB Eisen, PT Spellman, PO Brown, D Botstein: Cluster Analysis of Genome Wide Expression Patterns. *PNAS* 1998, 95: 14863-68

[3] P Tamayo, D Slonim, J Mesirov, Q Zhu, S Kitareewan, E Dimitrovsky, E Lander, TR Golub: Interpreting Gene Expression with Self-Organizing Maps. *PNAS* 1999, 96: 2907-2912

[4] P Törönen, M Kolehmainen, G Wong, E Castren: Analysis and Visualization Of Gene Expression Data Using Self-Organizing Maps. *FEBS Lett* 1999, 451: 142-146

[5] NS Holter, M Mitra, A Maritan, M Cieplak, JR Banavar, N Fedoroff: Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *PNAS* 2000, 97: 8409-8414

[6] O Alter, P Brown, D Botstein, Singular value decomposition for genome-wide expression data processing and modeling. *PNAS* 2000, 97: 10101-10106

[7] MK Kerr, M Martin, GA Churchill: Analysis of Variance for Gene Expression Microarray Data. *J Comp Biol* 2000, 7: 819-837

[8] MK Kerr, GA Churchill: Bootstrapping cluster analysis: Assessing the reliability of conclusions from micro-array experiments. *PNAS* 2001, 98: 9061-8965

[9] Affymetrix technical notes. Available at http://www.affymetrix.com

[10] F Li, GD Stormo: Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics* 2001, 17: 1067-1076

[11] N Friedman, M Linial, I Nachman, D Peér: Using Bayesian Networks to Analyze Expression Data. *J Comp Biol* 2000, 7: 601-620

[12] SA Kauffman: (1969) Metabolic Stability and Epigenesis in Randomly Constructed Genetic Nets. *J Theor Biol* 1969, 22: 437

[13] SA Kauffman: **The Origins of Order** *Oxford University Press* 1993

[14] S Huang: Gene expression profiling, genetic networks and cellular states: an integrating concept of tumor genesis and drug discovery. *J. Mol. Med.* 1999, 77: 469-480

[15] DW Marquardt: *Journal of the Society for Industrial and applied Mathematics* 11: 431-441

[16] P Gia, E ten Boekel, AG Rolink, F Melchers: B-cell development: a comparison between mouse and man. *Immunol Today* 1998, 19: 480-5

[17] AG Rolink, F Melchers, J Andersson: The transition from immature to mature B cells. *Curr Top Microbiol Immunol* 1999, 246: 39-43

[18] AG Rolink, E ten Boekel, T Yamagami, R Ceredig, J Andersson, F Melchers: B cell development in the mouse from early progenitors to mature B cells. *Immunol Lett* 1999, 68: 89-93

[19] D Liberg, M Sigvardsson: Transcriptional regulation in B cell differentiation *Crit Rev Immunol* 1999, 19: 127-53

[20] P Tsapogas, T Breslin, S Bilke, A Lagergren, R Mnsson, D Liberg, C Peterson and M Sigvardsson: RNA analysis of B-cell lines arrested at defined stages of differentiation allows for an approximation of gene expression patterns during B cell development. *J Leukoc Biol*, in press

[21] AG Rolink, C Schaniel, M Busslinger, SL Nutt, F Melchers: Fidelity and infidelity in commitment to B-lymphocyte lineage development. *Immunol Rev* 2000, 175: 104-11

[22] CM Högerkorp, S Bilke, T Breslin, S Ingvarsson, C Borrebaeck: CD44-stimulated human B cells express transcripts specifically involved in immunomodulation and inflammation as analyzed by DNA microarrays. *Blood* 2003 101:2307-13.

Paper IV

LU TP 04-11

Comparing functional annotation analyses with Catmap

Thomas Breslin[1], **Morten Krogh**[2] **and Patrik Edén**[3],

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

To appear in *BMC Bioinformatics*

Abstract

Background: Ranked gene lists from microarray experiments are usually analysed by assigning significance to predefined gene categories, *e.g.*, based on functional annotations. Tools performing such analyses are often restricted to a category score based on a cutoff in the ranked list and a significance calculation based on random gene permutations as null hypothesis.

Results: We analysed three publicly available data sets, in each of which samples were divided in two classes and genes ranked according to their correlation to class labels. We developed a program, `Catmap` (available for download[4]), to compare different scores and null hypotheses in gene category analysis, using Gene Ontology annotations for category definition. When a cutoff-based score was used, results depended strongly on the choice of cutoff, introducing an arbitrariness in the analysis. Comparing results using random gene permutations and random sample permutations, respectively, we found that the assigned significance of a category depended strongly on the choice of null hypothesis. Compared to sample label permutations, gene permutations gave much smaller p-values for large categories with many coexpressed genes.

Conclusions: In gene category analyses of ranked gene lists, a cutoff independent score is preferable. The choice of null hypothesis is very important; random gene permutations does not work well as an approximation to sample label permutations.

[1] thomas@thep.lu.se
[2] mkrogh@thep.lu.se
[3] patrik@thep.lu.se
[4] `http://bioinfo.thep.lu.se/Catmap`

Background

In genome-wide microarray experiments, it is possible to analyse the relevance of many different categories of genes, obtained from prior knowledge in the form of database annotations or from other experiments. These gene annotation analyses can unravel new information about pathways and cellular functions responsible for different phenotypes. Computational tools aiding in this process have recently been developed [1–8], most notably for annotations based on the Gene Ontology (GO) [9]. Generally, category relevance is calculated as the p-value of a score, thus being dependent on both the choice of score and the choice of null hypothesis.

In microarray analyses such as clustering, which provide defined subsets of genes with no internal ranking, it is natural to base the score on the number of category genes in the relevant subset. However, ranking of genes appear in many techniques for microarray analysis, such as correlation of gene expression to target profiles [10] and scoring of genes by their ability to discriminate between experimental conditions [11–13]. A separation of relevant and irrelevant genes can easily be constructed from ranked gene lists by introducing a cutoff, but the choice of cutoff becomes somewhat arbitrary and information in the list is lost. Tools addressing this problem, by using rank-based scores that are independent of a rank cutoff, have adopted the Kolmogorov–Smirnov score [14–17], and a minimized cutoff-based p-value [7, 8], which optimizes the cutoff for each category. The Wilcoxon rank sum [18], investigated here, serves the same purpose.

To calculate a p-value for the assigned score, a set of gene lists, ranked according to a chosen null hypothesis, are needed. The simplest choice of null hypothesis is just random gene permutations, and for some rank-based scores, the p-value can then be calculated analytically, without explicitly performing the permutations. However, the random gene permutations null hypothesis assumes independence of gene expression over biological samples, and the p-value is thus a combination of the p-value of how important the category is and the p-value for the genes of the category being coexpressed. When category genes behave similarly over a wide range of experimental conditions, the coexpression does not indicate relevance of the category for the question under study. In many analyses, a more appropriate null hypothesis is therefore sample label permutations, in which a set of ranked gene lists are generated based on the gene expression correlations to randomly permuted target values of the samples. This approach accounts for correlations between category genes and gives p-values that are bounded from below by the number of possible permutations of the samples in the data set. The latter is particularly important in data sets with few samples. Despite this, publicly available tools for gene annotation analysis are restricted to gene permutations [1–8].

We present a program, `Catmap`, for gene category analysis based on ranked gene lists. The program uses either the number of genes above a cutoff or the Wilcoxon rank sum as score, and the significance of the score can be calculated from a user supplied set of

ranked lists, thus allowing for sample label permutations. Furthermore, the program calculates corrections for multiple category testing, using permutation results to assess an effective number of independent categories, which enables `Catmap` to estimate very small multiple category p-values, that would otherwise have been computationally infeasible. The input to the program is two files and some arguments. The first file contains the biologically relevant ranked list of genes and, if needed, additional ranked gene lists drawn from the null hypothesis. The second file contains the categories and their corresponding genes. The input arguments can either be specified on the command line or in a settings file, and are as follows: 1) a choice between the cutoff score the Wilcoxon rank sum score; 2) a choice of null hypothesis, which can be either the above mentioned user-supplied ranked lists or random gene permutations; 3) the number of permutations used in multiple category testing. If zero, no multiple category testing is performed.

The output of Catmap is two files. The main output file contains all the categories, one on each line ordered according to their significance. The line of a category contains the p-value, the multiple comparison p-value, the false discovery rate, the ROC area (which is a normalized way to represent the Wilcoxon rank sum), the number of genes in the category, and the 25th, 50th, and 75th percentiles of the ranks. The other output file, the companion file, contains all the categories, with all the genes and their ranks listed below. Each line contains a gene and its rank. The program can be downloaded at the web site `http://bioinfo.thep.lu.se/Catmap`, where file format specification and example files are accessible as well.

Results and discussion

Comparing cutoff independent and cutoff-based score functions

We analysed the breast cancer data set of van 't Veer *et al.* [13] with a cutoff-based score function, using different cutoffs. Table 1 presents results for 15 categories with low p-values from cutoff independent scoring, showing that the p-value depends strongly on the choice of cutoff. This is further illustrated by the very different cutoffs at which the minimized cutoff-based p-value was obtained.

Compared to the variations between the cutoff-based alternatives, the results shown in Table 1 are in reasonable agreement for two cutoff independent p-values, using the Wilcoxon rank sum and the minimized cutoff-based p-value, respectively. The p-value based on the Wilcoxon rank sum was most often larger than the minimal cutoff-based p-value. Since the latter is biased by a minimization process, it must be interpreted as a score, rather than a p-value, thus requiring additional analyses to find statistical significance [7, 8].

GoName	GoId	top100	top300	top600	min p	cutoff	WRS p	#g
mitotic cell cycle	0000278	9e-05	5e-10	6e-07	3e-10	290	7e-08	93
M phase	0000279	1e-03	2e-06	1e-04	1e-09	1491	1e-07	41
nuclear division	0000280	1e-03	2e-06	1e-04	4e-09	1491	2e-07	40
M phase of mit. cell cyc..	0000087	6e-04	5e-07	1e-04	2e-08	1491	5e-07	36
mitosis	0007067	5e-04	4e-07	1e-04	6e-08	1491	1e-06	35
cell cycle	0007049	2e-04	3e-07	1e-05	1e-07	1571	6e-06	172
carbon-nitrogen ligase act..	0016884	4e-04	3e-03	1e-02	3e-05	27	2e-05	2
carboxylic acid metab..	0019752	9e-02	1e-04	4e-05	3e-06	711	3e-05	83
organic acid metabolism	0006082	9e-02	1e-04	4e-05	3e-06	711	3e-05	83
intramolecular isomerase ..	0016863	1e+00	3e-02	8e-04	2e-05	609	5e-05	4
cell proliferation	0008283	7e-04	2e-05	1e-03	4e-06	1956	8e-05	264
intramolec. isomerase ..	0016860	1e+00	2e-02	3e-03	3e-05	905	1e-04	9
spindle microtubule	0005876	4e-04	3e-03	1e-02	6e-05	42	1e-04	2
DNA replic. and chro..	0000067	6e-02	3e-04	2e-03	9e-05	852	3e-04	44
regulation of mitosis	0007088	9e-03	8e-03	5e-02	3e-04	1248	5e-04	8

Table 1: Category p-values for cutoff-based and cutoff independent score functions. The 15 GO categories with the lowest Wilcoxon rank sum p-values from the ranked gene list, based on the data set of van't Veer *et al.*, which comprises 5224 genes in total. Three columns show p-values for cutoff based score functions, with cutoffs at position 100, 300 and 600 in the list. The columns "min p" and "cutoff" give the minimal cutoff based p-value and the cutoff where this minimum was attained. The column "WRS" gives the p-value calculated with the Wilcoxon rank sum as score function and random permutation of genes as null hypothesis, and the column marked #g indicates the total number of genes in the category. A full table (sorted by WRS p-value) is given as supplementary information (additional file 2). The supplementary table also contains the ranking of each category using the different methods and the 25th, 50th and 75th percentiles of those genes in the ranked list.

Comparing null hypotheses

Using the Wilcoxon rank sum, we compared the results of different null hypotheses. Three publicly available data sets were examined [11, 13, 19]. As can be seen in Figure 1, p-values based on gene permutations tend to be lower than those based on sample label permutations. For categories with small p-values, there are remarkable differences, in particular for large categories with more than 20 genes.

Since the gene permutation null hypothesis assumes independent genes, we expect a GO category whose genes are uncorrelated to have roughly the same p-value under the two different null hypotheses, whereas a significant category whose genes are highly correlated will get a lower p-value using the gene permutation null hypothesis. To illustrate this coexpression effect, we picked two large categories, "carboxylic acid metabolism" and "M phase", which are encircled in Figure 1. In the data set of van 't Veer *et al.* [13], "carboxylic

acid metabolism" has similar p-values for the two null hypotheses, while "M phase" has a p-value of 10^{-7} using gene permutations but the much higher p-value of $3 \cdot 10^{-2}$ using sample label permutations. As seen in Figure 2, the most highly ranked genes of "M phase" are indeed more coexpressed than the most highly ranked genes of "carboxylic acid metabolism".

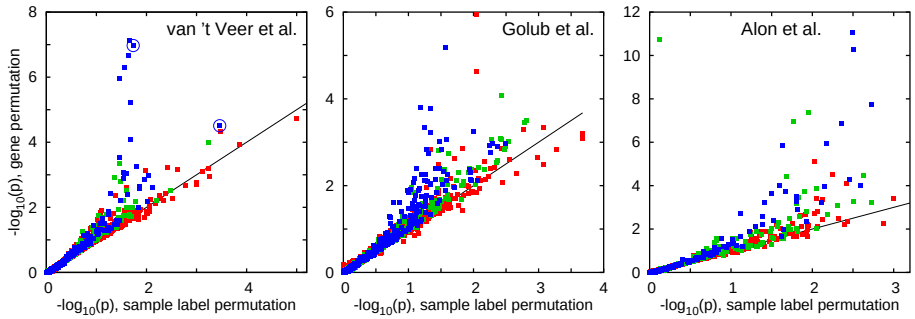


Figure 1: Comparing null hypotheses.Comparison of p-values obtained by sample label permutations and gene permutations, using the data set of van 't Veer *et al.* [13] (left), Golub *et al.* [11] (middle), and Alon *et al.* [19] (right). Sample label permutation results were obtained with 100.000 permutations for the van 't Veer *et al.* data set and with 10.000 permutations for the other data sets. Gene permutation results were calculated as described in Methods. Red, green and blue colours represent categories with 1 to 5, 6 to 20, and over 20 genes, respectively. Encircled boxes in the left figure represent the categories "M phase" and "carboxylic acid metabolism", which are further discussed in the text.

In Table 2, the ranks of categories for the different null hypotheses are compared. There are distinct differences, with only a small overlap among top ten categories. One can clearly see the tendency for the gene permutation null hypothesis to find categories with very many genes, as discussed above.

Table 2 also shows category ranks obtained with two alternative cutoff independent score functions: the Kolmogorov–Smirnov score as used in GSEA [17] and the minimal cutoff-based p-value used in FuncAssociate [7] and iGA [8]. These two alternatives do not calculate *individual* p-values for categories, but ranks categories based on the chosen score. Nevertheless, they give results similar to those obtained with the Wilcoxon rank sum and gene permutation. This is expected, since the minimized p-value is calculated with gene permutations, and the score adopted in GSEA [17] ranks categories similarly to what a Kolmogorov–Smirnov p-value, based on gene permutations, would do. It should be noted that GSEA, FuncAssociate, and iGA calculate multiple hypotheses corrected p-values, but these do not change the ranking of categories.

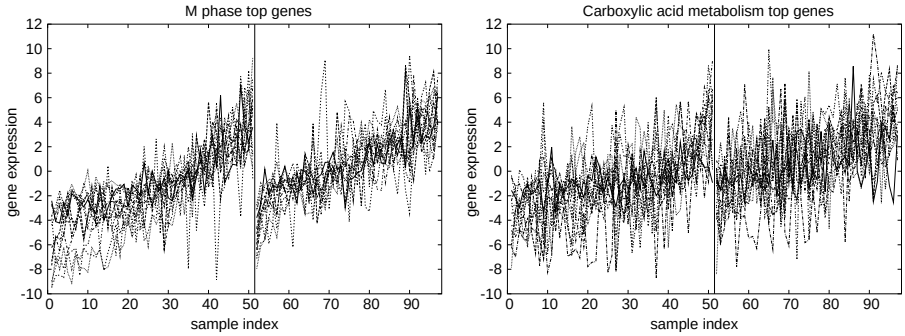


Figure 2: Effects of coexpression on different null hypotheses. Expression profiles, over the 97 samples in van 't Veer *et al.* [13], of the 12 most highly ranked genes in the "M phase" category (left) and 13 most highly ranked genes in the "carboxylic acid metabolism" category (right), respectively. Some genes were inverted since the ranking was based on absolute correlation values to metastasis class. The metastasis free samples are to the left of the vertical line, and within each metastasis class, samples are ordered in increasing average expression of the examined genes. The expressions of each gene was normalized to zero average across samples. The narrower band of expressions in the left figure illustrates the higher Pearson correlation of M phase genes. Average absolute Pearson correlation between gene expressions was 0.74, with standard deviation of 0.16, for the M phase genes, and 0.44, with standard deviation of 0.27, for carboxylic acid metabolism genes.

There is a possible difference (which does not reveal itself in Table 2) between the Kolmogorov–Smirnov score and minimized p-value score on one hand, and the Wilcoxon rank sum on the other, in the treatment of categories for which only a subset of genes have expressions correlating significantly with the question under study. The important genes being in the top of the ranked list will give the category a good score with all three score functions, provided the remaining, seemingly insignificant, genes are distributed in the ranked list as expected by random. However, if these less important genes lie higher in the list than expected by random (though not high enough to affect the Kolmogorov–Smirnov or min-p scores), the category will be considered more important by the Wilcoxon rank sum. Reversely, if the less important category genes prevail in the bottom of the list, the Wilcoxon rank sum score function will deem the category as unimportant, while the other two scores will give the category a high significance, based on the top ranked genes alone. Whether seemingly insignificant genes being ranked better or poorer than explainable by random expectations should be observed or ignored is of course a matter of taste, and a possibility is to use several score functions, that may complement each other. The differences are, however, much smaller than those related to choice of null hypothesis, as revealed in Table 2.

GoName	GoId	WRS		K–S	min-p	25%	50%	75%	#g
		s.l.p.	g.p.						
carbon-nitrogen ligase act..	0016884	1	7	53	16	6	6	27	2
spindle microtubule	0005876	2	13	54	19	38	38	42	2
organic acid metabolism	0006082	3	9	11	8	564	1619	3283	83
carboxylic acid metabolism	0019752	4	8	12	7	564	1619	3283	83
intramolec. isomerase act..	0016863	5	10	10	12	195	412	453	4
deoxynucleoside kinase act..	0019136	6	18	60	51	70	70	117	2
GMP synthase activity	0003921	7	27	317	78	6	6	6	1
intramolec. isomerase act..	0016860	9	12	9	14	195	453	905	9
biotin metabolism	0006768	10	21	62	58	76	76	132	2
nucleus	0005634	71	16	28	10	1100	2353	3797	574
mitotic cell cycle	0000278	121	1	3	1	346	1419	3031	93
M phase	0000279	107	2	1	2	251	848	1862	41
nuclear division	0000280	124	3	2	3	251	867	1862	40
M phase of mit. cell cyc..	0000087	130	4	4	4	238	848	1862	36
mitosis	0007067	152	5	5	5	235	848	1862	35
cell cycle	0007049	142	6	6	6	731	1689	3599	172
cell proliferation	0008283	112	11	7	9	891	1947	3645	264
regulation of cell cycle	0000074	153	29	8	15	731	1689	3604	105

Table 2: Comparison of different cutoff independent approaches. The top ten categories and their corresponding ranks for each of the the four methods: Wilcoxon rank sum (WRS) with sample label permutation null hypothesis (s.l.p.), WRS with gene permutation null hypothesis (g.p.), the Kolmogorov–Smirnov score (K–S) as used in GSEA [17], and the minimal cutoff-based p-value (min-p) [7, 8]. Percentile columns indicate the position of the 25th, 50th and 75th percentile in the ranked gene list which comprises 5224 genes and is based on the data set of van't Veer *et al.*. The last column indicates the number of genes in each category. The full table is available as a supplementary file (additional file 3).

Multiple category testing

The more categories that are being tested, the more likely it is that at least one category gets a very small p-value by chance. To better evaluate the statistical significance of the best scoring categories, we used `Catmap` to calculate false discovery rates and family-wise error rates by permutation tests. This also gave us an effective number of independent categories, N_{eff}, as described in Methods.

The GO contains many small categories which would be reasonable to ignore in a study aiming at biological conclusions, and they were included in Figure 1 mainly to highlight the differences between the null hypotheses. When performing multiple category testing, we restricted the study to large categories, containing more than 20 genes. We tested the 3 sub-ontologies (biological process, molecular function, and cellular component) both separately and together.

As expected from the discussion above, several categories with coexpressed genes got small p_{multiple} and small false discovery rates with random gene permutations. In contrast, when using sample label permutations, the smallest p_{multiple} was obtained in the data set of van 't Veer *et al.* [13] for the biological process category "organic acid metabolism", which contained 83 genes and had $p = 3 \cdot 10^{-4}$ and $p_{\text{multiple}} = 0.02$. Interestingly, organic acid metabolism is known in the literature to be relevant for breast cancer [20, 21]. For this data set and the biological process categories, there was a 38% false discovery rate among the top 15 categories.

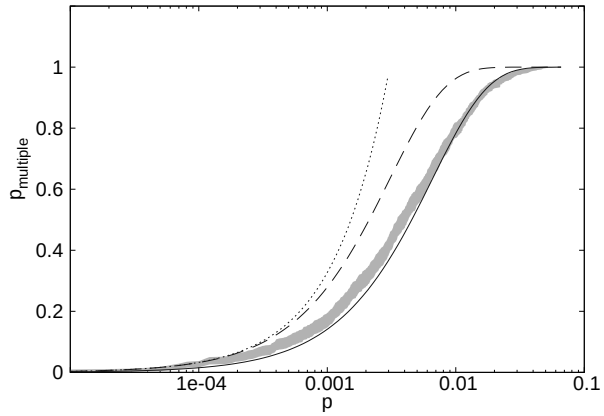


Figure 3: Fitting an effective number of independent categories. The multiple category p-value, p_{multiple}, versus p-value for the data set of van 't Veer *et al.* [13], using 327 large Gene Ontology categories with more than 20 genes. The yellow band shows 95% confidence interval of sample label permutation results, based on 1000 random lists, and the blue curves show the results of Equation (1), with the fitted $N_{\text{eff}} = 152$ (solid line), the total number of categories $N = 327$ (dashed line), and also the Bonferroni correction (dotted line).

For all 3 sub-ontologies, the effective number of categories, N_{eff}, was around half of the full number of categories, N. In the data set of van 't Veer *et al.* [13] the numbers were $N_{\text{eff}} = 83$ versus $N = 166$ for biological process, $N_{\text{eff}} = 69$ versus $N = 119$ for molecular function, and $N_{\text{eff}} = 22$ versus $N = 42$ for cellular component. For all categories together the real number of large categories was $N = 327$ whereas $N_{\text{eff}} = 152$. Using random gene permutations for the same data set and categories, we got $N_{\text{eff}} = 170$.

The fact that N_{eff} for the two null hypotheses are so close is a general phenomena that we see in all our examples (data not shown). Furthermore, for all data sets and ontologies studied, N_{eff} was approximately half of the total number of categories. If this is a general feature for GO categories, the simple Bonferroni correction would not be totally

unreasonable for small p-values.

Figure 3 shows that the fit with an effective number of categories was good; in the range where permutations results were available it did not deviate more than a factor of two. The example in Figure 3 was obtained with 100.000 sample label permutations, and minimal p-values were found for 1000 random gene lists.

It should be noted that whenever several ranked lists are examined as part of a project, this additional source of multiple hypotheses testing should also be corrected for. An example of such a correction, for cutoff-based score functions, is presented by Corà *et al.* [22].

Conclusions

We developed a computer program for calculating the significance of gene categories in a ranked list of genes. Corrections for multiple category testing can be performed by the program. To investigate the properties of different scores and null hypotheses, we analyzed three publicly available data sets [11, 13, 19].

Commonly [1–6], a subset of relevant genes is defined from a ranked gene list by introducing a cutoff in the list. Our results show that the obtained p-values of biologically relevant categories depend strongly on the choice of cutoff. The cutoff independent Wilcoxon rank sum score overcomes the problem, representing an alternative to the Kolmogorov–Smirnov score [14–17] and the minimized cutoff-based p-value [7, 8]. The ranking of categories for the three cutoff independent scores are very similar.

Though sample label permutations in many situations represent a better null hypothesis than gene permutations, available gene annotation analysis tools are restricted to the latter. Our implementation allows for both null hypotheses, and we find that both the p-values and the ranking of categories depend strongly on the choice of null hypothesis. Compared to sample label permutations, gene permutations gave much smaller p-values for large categories with many coexpressed genes.

Methods

Algorithm

The implemented algorithm treats the categories sequentially and independently. As score function for category relevance, the program uses either the Wilcoxon rank sum or the number of genes above a given cutoff in the ranked list. The latter is implemented for

method comparison and for the case of a defined subset of relevant genes, without internal ranking.

For the case of the Wilcoxon rank sum, the user can supply a set of ranked lists distributed according to an appropriate null hypothesis, or request random permutations of genes as the null hypothesis. In the latter case, the significance of the score is calculated analytically by the program, using either an exact calculation by an iterative method, a Gaussian approximation, or a continuous volume approximation. The program chooses method based on a balance between accuracy and computation time. For details, see additional file 1: '*p*-values for the Wilcoxon rank sum score'.

For the case of the cutoff-based score function, the *p*-value of category relevance is determined with Fisher's exact test [23], corresponding to randomly permuted genes as null hypothesis.

When N independent categories are tested simultaneously, family-wise error rate simply means calculating the probability,

$$p_{\text{multiple}}(q) = 1 - (1 - q)^N, \tag{1}$$

that at least one category has a *p*-value below any given number q by chance. For correlated categories, we make the assumption that the same functional form describes $p_{\text{multiple}}(q)$, with N replaced by an effective number of independent categories N_{eff}. We find N_{eff} by generating a number, K, of ordered lists under the null hypothesis and calculating the *p*-values of all categories. We fit N_{eff} using the maximum likelihood estimation

$$\frac{1}{N_{\text{eff}}} = \frac{-\sum_{k=1}^{K} \ln(1 - p_k)}{K}, \tag{2}$$

where p_k is the minimal *p*-value for the k'th ordered list.

The false discovery rate for the j highest ranked categories is found by counting the number of *p*-values from K permuted lists lower than the *p*-value of the j:th category and divide this number with $K \cdot j$.

For the case of sample label permutations, when a user supplied set of ranked gene lists are used to represent the null hypothesis, the first K lists are used to find N_{eff} and false discovery rates, and the remaining lists are used to calculate *p*-values for each of the K lists.

Implementation

The algorithm is implemented in the Perl program `Catmap.pl` and is released under the GNU General Public License (GPL). `Catmap.pl`, together with user instructions, is available for download at `http://bioinfo.thep.lu.se/Catmap`.

Public Data Sets

Using `Catmap`, we analysed three publicly available data sets with gene annotations from the Gene Ontology.

The data set of van 't Veer *et al.* [13] consists of 97 patients with primary sporadic breast cancer, of which 46 had metastases within five years following treatment. Quality filtering was performed as described in [13], and rendered about 5,000 genes which were ranked according to their absolute Pearson correlation to metastasis class. A Gene Ontology analysis of the data set has previously been performed with the 231 top genes as the subset of important genes and random gene permutations [24].

The data set of Golub *et al.* [11] consists of bone marrow samples from leukemia patients, 27 with AML and 11 with ALL. The published data contains expression levels for 5000 genes, which after removal of genes with no variance across samples rendered 4812 genes which were ranked according to their absolute Pearson correlation to leukemia type.

The data set of Alon *et al.* [19] consists of 40 tumour and 22 normal colon tissue samples. The 2000 genes in the published data set were ranked according to their absolute Pearson correlation to tissue type.

Gene Ontology Associations

All genes were first mapped to corresponding UniGene clusters [25]. For the data set of Golub *et al.* [11] this mapping was given from chip annotation files provided by Affymetrix, whereas for the other data sets [13, 19], the mapping was done via GenBank accession numbers. GO annotations for UniGene clusters were extracted with ACID [26], and completed by back propagating all lower level associations on the GO graph.

Authors' contributions

TB and MK implemented the algorithms in `Catmap`. All authors participated in the design of the study, prepared, read, and approved the final manuscript.

Acknowledgments

MK and TB are grateful for financial support from the Swedish Foundation for Strategic Research. PE was supported by the Swedish Foundation for Strategic Research through the Lund Center for Stem Cell Biology and Cell Therapy. The authors also thank Kasper Astrup Eriksen, Peter Johansson, Henrik Jönsson, Carsten Peterson and Markus Ringnér for fruitful discussions.

References

[1] 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**:R28.

[2] Robinson MD, Grigull J, Mohammad N, Hughes TR: **Funspec: a web-based cluster interpreter for yeast.** *BMC Bioinformatics* 2002, **3**:35.

[3] Khatri P, Draghici S, Ostermeier GC, Krawetz SA: **Profiling gene expression using onto-express.** *Genomics* 2002, **79**:266–270.

[4] Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, Lawlor SC, Conklin BR: **Mappfinder: using gene ontology and genmapp to create a global gene-expression profile from microarray data.** *Genome Biol.* 2003, **4**:R7.

[5] Beissbarth T, Speed T: **GOstat: Find statistically overrepresented Gene Ontologies within a group of genes.** *Bioinformatics* 2004, **20**:1464–1465.

[6] Hosack DA, Dennis Jr G, Sherman BT, Lane HC, Lempicki RA: **Identifying biological themes within lists of genes with EASE**. *Genome Biol.* 2003, **4**:R70.

[7] Berriz GF, King OD, Bryant B, Sander C, Roth FP: **Characterizing gene sets with funcassociate.** *Bioinformatics* 2003, **19**:2502–2504.

[8] Breitling R, Amtmann A, Herzyk P: **Iterative Group Analysis (iGA): A simple tool to enhance sensitivity and facilitate interpretation of microarray experiments**. *BMC Bioinformatics* 2004, **5**:34.

[9] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. the gene ontology consortium.** *Nat. Genet.* 2000, **25**:25–29.

[10] Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen B, Brown PO, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization.** *Mol. Biol. Cell.* 1998, **9**:3273–3297.

[11] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring**. *Science* 1999, **286**:531–537.

[12] Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu 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.

[13] van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, 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**:530–536.

[14] Kolmogorov AN: **Sulla determinazione empirica di una legge di distribuzione**. *Giorn. Dell Inst. Ital. Degli Attuari* 1933, **4**:83–91.

[15] Smirnov NV: **On the estimation of the discrepancy between empirical curves of distribution for two independent samples**. *Bull. Moscow Univ.* 1939, **2**:3–16.

[16] Jensen LJ, Knudsen S: **Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation.** *Bioinformatics* 2000, **16**:326–333.

[17] Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstrale M, Laurila E, Houstis N, Daly MJ, Patterson N, Mesirov JP, Golub TR, Tamayo P, Spiegelman B, Lander ES, Hirschhorn JN, Altshuler D, Groop LC: **PGC-1alpha-responsive genes involved in oxidative phosphorylati on are coordinately downregulated in human diabetes.** *Nat. Genet.* 2003, **34**(3):267–73.

[18] Wilcoxon F: **Individual comparisons by ranking methods.** *Biometrics* 1945, **1**:80–83.

[19] Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ: **Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays**. *Proc. Natl. Acad. Sci. USA* 1999, **96**:6745–6750.

[20] Kuhajda FP: **Fatty-acid synthase and human cancer: new perspectives on its role in tumor biology.** *Nutrition* 2000, **16**:202–208.

[21] Kumar-Sinha C, Ignatoski KW, Lippman ME, Ethier SP, Chinnaiyan AM: **Transcriptome analysis of her2 reveals a molecular connection to fatty acid synthesis.** *Cancer Res.* 2003, **63**:132–139.

[22] Cora D, Di Cunto F, Provero P, L Silengo P, Caselle M: **Computational identification of transcription factor binding sites by functional analysis of sets of genes sharing overrepresented upstream motifs.** *BMC Bioinformatics* 2004, **5**:57.

[23] Fisher RA: **The use of multiple measurements in taxonomic problems.** *Ann. Eugen.* 1936, **7**:179–188.

[24] Draghici S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA: **Global functional profiling of gene expression.** *Genomics* 2003, **81**:98–104.

[25] **UniGene**[`http://www.ncbi.nlm.nih.gov/UniGene`].

[26] Ringnér M, Veerla S, Andersson S, Staaf J, Häkkinen J: **ACID: a database for microarray clone information**. *Bioinformatics* 2004, **20**:2305–2306.

Additional Files

Additional file 1 — *p*-values for the Wilcoxon rank sum score.

File name: Catmap_supp.pdf

File format: pdf

Additional file 2 — Supplement to Table 1.

File name: Table1_supp.csv

File format: csv

Additional file 3 — Supplement to Table 2.

File name: Table2_supp.csv

File format: csv

Paper V

LU TP 04-32

Signal transduction pathway profiling of individual tumor samples

Thomas Breslin[1], **Morten Krogh**[2],
Carsten Peterson[3] **and Carl Troein**[4]

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

To be submitted

Abstract
We devise a method for analyzing tumor gene expression data in terms of signal transduction pathway activity. We use pathways compiled from the TRANSPATH/TRANSFAC databases and the literature, and three publicly available cancer microarray data sets. Variation in pathway activity, across the samples, is gauged by the degree of correlation between downstream targets of a pathway. Two correlation scores are applied; one considers of all pairs of downstream targets, and the other considers only pairs without common transcription factors. Furthermore, we devise a score for pathway activity in individual samples, based on the averege expression value of the downstream targets. Statistical significance is assigned to the scores using reshuffling of genes as null model. Hence, for individual samples, the status of a pathway is given as a sign, $+$ or $-$, and a p-value. This approach defines a projection of high-dimensional gene expression data onto low-dimensional pathway activity scores. Finally, we find that several sample-wise pathway activities are significantly associated with clinical classifications of the samples.

[1] thomas@thep.lu.se
[2] mkrogh@thep.lu.se
[3] carsten@thep.lu.se
[4] carl@thep.lu.se

Introduction

The interpretation of microarray data is facilitated by combining the data, or results of data analysis, with prior contextual knowledge, *e.g.*, ontologies [2,3,10,18], pathways [11, 13, 19] and other annotation groups of interest [14]. This study aims at inferring cellular signaling pathway activity from tumor microarray data, on a sample-by-sample basis, using prior knowledge about pathways. Furthermore, we examine whether the pathway activity of individual samples is associated to clinical classifications of the samples.

Signaling pathway activity scoring is a more direct measurement of biological processes than ontology mapping, which aims at finding over-representation of genes in various groups of contextual annotation. A cellular signaling pathway (see Fig. 1) is composed of a series of signaling molecules that convey information, typically from the outside of the cell to the nucleus. The initial step consists of extracellular signaling molecules, ligands, that activate receptors of the cell. These receptors then initiate intracellular signaling events, which eventually regulate the activity of various transcription factors. These transcription factors, in turn, regulate the expression levels of various genes, termed downstream targets of the pathway.

To characterize pathway activity, it would be desirable to have both proteomic and gene expression data. Gene expression data alone is not sufficient for assessing protein concentrations [7] and post-translational modifications of proteins. In the absence of proteomic data, one is thus forced to rely on aspects of the pathway that are detectable at the mRNA level. The foremost candidate for this is the downstream targets, which we will focus on here. It is of course also possible that the mRNA levels of effector proteins in a pathway change due to altered pathway activity. However, such effects are outside the scope of this paper. A further complication is that many pathways overlap, both in terms of having common transcription factors, and in terms of distinct transcription factors having common downstream targets. Our methods will not be able to distinguish very similar pathways, and in some sense this problem can be seen as a result of the ambiguities that follow when the full protein network is partitioned into separate pathways.

Cellular signaling pathways are subject to intense research, and current knowledge is compiled into databases such as STKE [5], TRANSPATH [8] and TRANSFAC [17]. These databases do not yet account for all pathways or transcription factors, but develop over time. Most pathway information utilized in this work is collected from TRANSPATH/TRANSFAC, where information about transcription factors and downstream targets is readily available. The only exception is the estrogen receptor pathway, which is taken from [9]. We analyze three microarray data sets in this study: Two breast cancer data sets [15, 16], and one leukemia data set [4].

For the three data sets, we assess pathway activity from two related, but different, points of view. The first is to examine which pathways behave in a coherent way across the entire

data set, *i.e.*, which pathways have significantly co-expressed downstream targets. This is done both with and without accounting for the fact that downstream targets of a single transcription factor are correlated irrespective of pathway behavior. The second point of view is to assess pathway activity of individual samples, relative to the other samples in the same experiment, yielding an active or inactive status for each pathway in each sample. Finally, we relate the sample-wise pathway activity to clinical classifications of samples by way of contingency tables. Several pathways are found to be highly predictive of the clinical classifications.

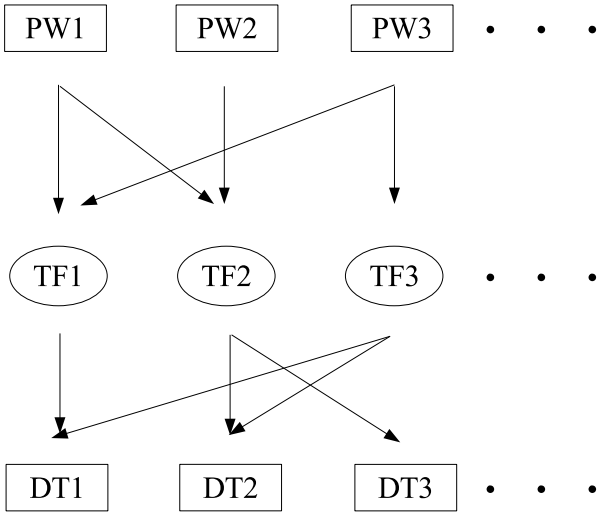


Figure 1: Pathways (PW) → transcription factors (TF) → downstream targets (DT).

Methods

Pathway information and UniGene clusters

Transcription factors for 23 pathways were extracted from TRANSPATH [8]. The downstream target genes of those transcription factors were obtained from TRANSFAC [17]. Since our study contains breast cancer data, we have augmented the pathway information with the Estrogen Receptor (ER) pathway compiled from [9], where 89 direct target genes were identified. 59 of them were induced by the ER complex and 30 were repressed. 8 out of the 89 genes were previously verified. We employ all six combinations of induced/repressed/all and verified/all, yielding 6 versions of the ER pathway. The 29

pathways employed are listed in Table 2. Downstream target genes were represented as UniGene IDs (http://www.ncbi.nlm.nih.gov/UniGene), using UniGene Hs build 171. For the analysis of the data sets, gene identifiers were converted into UniGene IDs and expression values of clones belonging to the same UniGene cluster were averaged.

Data Sets

The following three publicly available data sets were analyzed:

1. The breast cancer data set of van 't Veer *et al.* [16], consisting of samples from 117 patients, of which 46 developed metastases. After UniGene merging the data set contains 20663 genes.
2. The breast cancer data set of Sotiriou *et al.* [15], consisting of 99 samples of different clinical classifications, with 4878 genes after UniGene merging.
3. The leukemia data set of Golub *et al.* [4], derived from bone marrow samples from 38 patients, 27 of which were diagnosed with Acute Myeloid Leukemia (AML) and 11 with Acute Lymphoblastic Leukemia (ALL). After UniGene merging and removal of genes with no variance across the samples, 4701 genes remain.

Normalization of microarray data

The data sets in [15, 16] are given in the form of log ratios of expression values in the samples versus a reference. The data set in [4] is given in the form of Affymetrix average difference values. For the calculation of *Group Correlation Score* and *Exclusive Group Correlation Score* the affymetrix average difference values were logarithm transformed, since the Pearson correlation is very sensitive to single outlier samples. For the *Group Sample Score* the original differences were kept. We denote the expression value for gene g in sample s by x_{gs}, with missing values allowed. We normalized the expression values in two steps. First, for each sample, the mean of all genes was subtracted, in order to ensure that no samples are up- or down-regulated on average. The transformed expression values satisfy:

$$\sum_g x_{gs} = 0 \qquad \text{for all } s.$$

Second, for each gene, the mean of all samples was subtracted from the expression values of that gene, yielding:

$$\sum_s x_{gs} = 0 \qquad \text{for all } g.$$

The second normalization implies that the expression value of a gene is measured relative to the same gene in other samples.

Pearson correlation p-values

To determine if a group of downstream target genes is significantly co-expressed, a total score of the group is needed. Two scores were used here, both based on the Pearson correlation of a pair of genes g and h:

$$r_{g,h} = \frac{\sum_s (x_{gs} - \bar{x}_g)(x_{hs} - \bar{x}_h)}{\sqrt{\sum_s (x_{gs} - \bar{x}_g)^2} \sqrt{\sum_s (x_{hs} - \bar{x}_h)^2}},$$

where the sums exclude missing values and $\bar{x}_g$ is the mean of expression values for gene g.

The *Group Correlation Score* is defined as the sum of squares of Pearson correlations among all pairs of genes in a group of genes:

$$GCS = \sum_{g \neq h} r_{g,h}^2$$

where the sum runs over all genes in the group. The square ensures that both correlations and anti-correlations contribute to the score. We use the *Group Correlation Score* for the downstream target genes of a single transcription factor, as well as for those of an entire pathway.

The *Exclusive Group Correlation Score*, on the other hand, is only applicable for the downstream targets of a pathway. It is defined as

$$EGCS = \sum_{TF(g) \cap TF(h) = \emptyset} r_{g,h}^2$$

where the sum runs exclusively over pairs of genes g and h that do not share any transcription factor.

The p-value of a score is defined as the fraction of random cases, drawn under the null hypothesis, which achieve a higher score than the score tested. For both scores, *GCS* and *EGCS*, our null hypothesis is reshuffling of the genes on the microarray. This null hypothesis keeps the structure and overlap of all pathways fixed, but changes the identity of the genes.

Pathway activity for individual samples

For each sample, s, and pathway,PW, the *Group Sample Score* is defined as follows:

$$GSS_s = \sum_{g \subset PW} x_{gs},$$

where the sum runs over all downstream target genes of the pathway.

The null hypothesis is again reshuffling of the genes from the microarray. We are interested in pathways both with high and low scores. Hence, we consider the p-values for the score being higher (p_+) and lower (p_-) than random, respectively, and the final p-value is given by two times the smaller of these two p-values:

$$p = 2 \cdot \min(p_+, p_-).$$

The pathway is said to be active ($+$) if $p_+ < p_-$, and inactive ($-$) otherwise.

Family-wise p-value

If N independent hypotheses are tested simultaneously, the probability to obtain K or more p-values below q is given by a binomial distribution:

$$p_{fw} = \sum_{i=K}^{N} \binom{i}{N} q^i (1-q)^{(N-i)}.$$

We refer to this probability as the *family-wise p-value*.

Results

In this section we assess variation in transcription factor and pathway activity across the samples. We then proceed to probe pathway activity in individual samples. Finally, we study the association between pathway activity and the clinical classifications of the samples.

Co-expression of the downstream targets of a transcription factor

As a prelude to the study of pathways, we quantified the degree of correlation among downstream targets of single transcription factors. For this purpose we used the *Group Correlation Score* defined in Methods. The p-values were calculated using random reshuffling of the genes. Table 1 shows the most significant transcription factors in the van 't Veer data set. We see, as expected, that several transcription factors have significantly correlated downstream targets. For the data set of Sotiriou *et al.*, 8 out of 42 transcription factors have a p-value below 0.1, and for the Golub *et al.* data set, the corresponding numbers are 6 out of 39. Among these three data sets, the p-values are noticeably better for data sets with more samples and genes. With this in mind we conclude that the downstream targets of transcription factors are co-expressed in the these data sets, albeit not to a high degree. The full lists for all 3 data sets can be found in the Supplement [1].

TF	# of DT	*p*-value
NF-κB	19	3e-4
RelA	11	9e-4
NF-κB1	10	2e-3
ER-induced	50	3e-3
STAT1α	2	5e-3
ER-induced (v)	5	7e-3
STAT6	6	7e-3
C/EBPα	21	7e-3
ER	77	7e-3
ER(v)	7	1e-2
GATA-1	7	2e-2
c-Rel	3	5e-2
Elk-1	4	6e-2
STAT4	3	7e-2
SMAD-3	6	8e-2

Table 1: The 15 most significant transcription factors (TF) in the van 't Veer *et al.* [16] data set, and their number of downstream targets (DT). *p*-values are based on the Group Correlation Score. In all 54 TFs were studied for this data set. ER pathway notation as in Table 2.

Co-expression of the downstream targets of a pathway

Here we use the same *Group Correlation Score* as above, applied to the downstream targets of entire pathways rather than those of individual transcription factors. Table 2 shows the results for the van 't Veer data set, where 21 out of 29 pathways have a *p*-value below 0.05, which is significantly more than expected by chance. However, many of the downstream targets have common transcription factors, which might be the major cause of the co-expression. To eliminate such a contribution we used the *Exclusive Group Correlation Score*, which considers only pairs of downstream targets lacking common transcription factors. The *p*-values for the *Exclusive Group Correlation Score* are also shown in Table 2. Although these *p*-values are larger, they are still significant; out of 20 pathways with more than one transcription factor, 12 have a *Exclusive Group Correlation Score p*-value below 0.05, which is still more than expected by chance. We conclude that the co-expression of downstream targets in a pathway can only in part be explained by the genes having common transcription factors. This co-expression at the pathway level justifies the view of pathways as functional units. Similar tables for the two other data sets are shown in the Supplement [1]. Both data sets have smaller *p*-values than expected by chance, albeit not as convincingly as the van 't Veer data set.

Pathway	# of TF	# of DT	GCS *p*-value	EGCS *p*-value
IL-1	5	21	3e-4	3e-1
fMLP	9	27	3e-4	1e-1
TLR4	9	41	8e-4	1e-3
RANK	6	41	2e-3	5e-2
ER-induced	1	50	2e-3	n/a
EDAR	6	41	2e-3	5e-2
Oncostatin M	1	2	6e-3	n/a
PDGF	8	15	6e-3	2e-2
ER	1	77	7e-3	n/a
ER-induced (v)	1	5	8e-3	n/a
IL-4 - STAT6	1	6	8e-3	n/a
TGF$-\beta$ network	7	23	1e-2	1e-2
EGF network	12	53	1e-2	1e-2
Insulin	7	45	1e-2	4e-2
ER (v)	1	7	2e-2	n/a
TNF-α	8	61	2e-2	5e-2
VEGF	3	8	2e-2	2e-2
TPO	6	10	3e-2	2e-2
IFN	6	10	3e-2	2e-2
PRL	6	10	3e-2	2e-2
IL-10	2	7	6e-2	5e-2
IL-12 - STAT4	1	3	7e-2	n/a
c-Kit	4	87	8e-2	6e-2
ER-repressed	1	27	1e-1	n/a
T-cell antigen receptor	4	10	3e-1	3e-1
B-cell antigen receptor	4	10	3e-1	3e-1
Wnt	2	8	4e-1	3e-1
ER-repressed (v)	1	2	6e-1	n/a
IL-2 - STAT5	2	4	8e-1	7e-1

Table 2: Pathways in the van 't Veer *et al.* [16] data set, ordered by significance and their number of transcription factors (TF) and downstream targets (DT). Also shown are Group Correlation Score (GCS) and Exclusive Group Correlation Score (EGCS) *p*-values. ER means both induced and repressed ER-pathway and (v) means that the pathway has been verified in a second experiment (see [9]).

Pathway assignments for individual samples

After having established that downstream target genes are co-expressed in some pathways, we proceeded to study the status of pathway activity in individual samples. To this end we employed the *Group Sample Score*, which for each pathway designates every sample in a data set as either active or inactive, with an associated *p*-value.

Table 3 shows *p*-values and pathway activity status, for six samples in the van 't Veer data set. For example, in the first sample the RANK pathway is designated as inactive with a *p*-value of 0.004, whereas the inactiveness of the ER-induced pathway cannot be considered significant. The full tables for all samples and pathways in all 3 data sets are provided in the Supplement [1].

	1	2	3	4	5	6
ER-induced	0.666(–)	**0.000**(+)	0.47(+)	0.184(+)	0.48(+)	**0.002**(+)
RANK	**0.004**(–)	0.15(–)	0.11(–)	0.832(+)	**0.002**(–)	**0.000**(–)
IL-1	**0.008**(–)	**0.026**(–)	**0.014**(–)	0.356(+)	**0.012**(–)	**0.002**(–)
TNF-α	**0.002**(–)	0.132(–)	0.232(–)	0.43(+)	**0.004**(–)	**0.000**(–)
EGF network	0.292(–)	0.3(–)	0.484(–)	0.94(–)	0.176(–)	**0.000**(–)
EDAR	**0.006**(–)	0.144(–)	0.102(–)	0.834(+)	**0.002**(–)	**0.002**(–)
ER	0.968(+)	**0.000**(+)	0.316(+)	0.748(+)	0.864(–)	**0.002**(+)
ER-induced (v)	0.484(–)	**0.022**(+)	0.148(+)	0.322(+)	0.286(+)	0.386(+)
fMLP	**0.02**(–)	0.052(–)	**0.000**(–)	0.688(+)	**0.036**(–)	**0.038**(–)
TLR4	0.182(–)	0.272(–)	0.304(–)	0.784(+)	0.166(–)	**0.000**(–)
Insulin	0.86(–)	**0.024**(–)	0.414(–)	0.68(+)	0.828(+)	**0.016**(–)
TGF−β network	0.082(+)	0.72(+)	0.858(–)	0.332(–)	0.766(–)	0.29(–)
ER (v)	0.972(–)	**0.01**(+)	0.056(+)	0.1(+)	0.342(–)	0.094(+)
c-Kit	0.286(–)	**0.022**(–)	0.314(–)	0.82(–)	**0.05**(–)	0.24(–)
IL-10	0.38(–)	0.578(+)	0.796(–)	0.79(+)	0.204(–)	0.806(+)
TPO	0.386(–)	0.746(+)	0.512(–)	0.814(+)	0.618(–)	0.766(–)
ER-repressed (v)	0.368(+)	0.052(+)	0.062(+)	**0.038**(+)	**0.004**(–)	0.06(+)
VEGF	0.646(–)	0.342(+)	0.42(–)	0.988(–)	0.882(–)	0.848(–)
IFN	0.398(–)	0.742(+)	0.472(–)	0.774(+)	0.596(–)	0.744(–)
PRL	0.424(–)	0.774(+)	0.458(–)	0.794(+)	0.622(–)	0.756(–)
PDGF	0.966(+)	0.498(–)	0.608(–)	0.52(–)	0.752(+)	0.382(–)
Oncostatin M	0.298(–)	0.322(–)	0.194(–)	0.172(+)	0.746(+)	**0.002**(–)
T-cell antigen receptor	0.196(–)	0.1(–)	**0.006**(–)	0.928(+)	0.65(–)	0.36(+)
IL-12 - STAT4	0.356(+)	0.41(+)	0.87(+)	0.604(+)	0.234(–)	0.968(+)
B-cell antigen receptor	0.19(–)	0.094(–)	**0.006**(–)	0.936(+)	0.654(–)	0.306(+)
ER-repressed	0.53(+)	0.472(+)	0.428(+)	0.17(–)	0.074(–)	**0.006**(+)
IL-2 - STAT5	0.22(–)	0.884(–)	0.468(–)	0.798(–)	0.34(+)	**0.036**(–)
IL-4 - STAT6	0.83(–)	0.166(–)	0.91(–)	0.326(–)	0.106(+)	0.166(–)
Wnt	0.862(–)	0.086(–)	0.102(–)	0.588(–)	0.378(–)	0.52(–)

Table 3: The individual sample pathway activity *p*-values and sign for each pathway and six of the van 't Veer breast cancer samples. Bold face indicates significant (*i.e.* *p*-value $\leq$ 0.05) pathway activity in the sample. ER notation as in Table 2.

Table 4 shows the number of samples that are active and inactive at the 5% level, for every pathway in the van 't Veer data set. The table also contains the family-wise *p*-value, defined in Methods, which gives the probability of observing at least this total number of significant samples for a pathway. The family-wise *p*-value assumes that the samples are independent, which is only approximately true since the mean expression value of a gene across all samples is zero. Corresponding tables for the two other data sets can be found in the Supplement [1]. We note that the most significant pathways according to this measure are mostly the same as with the correlation based scores, although the *p*-values are numerically different.

Pathway	+	-	family-wise *p*-value
ER-induced	30	36	5e-54
RANK	30	30	6e-46
IL-1	24	34	2e-43
TNF-α	29	29	2e-43
EGF network	30	28	2e-43
EDAR	29	29	2e-43
ER	29	29	2e-43
ER-induced (v)	22	34	7e-41
fMLP	24	31	1e-39
TLR4	27	27	2e-38
Insulin	22	22	5e-27
TGF$-\beta$ network	21	19	7e-23
ER (v)	15	22	5e-20
c-Kit	20	14	3e-17
IL-10	13	15	4e-12
TPO	11	15	1e-10
ER-repressed (v)	11	15	1e-10
VEGF	12	14	1e-10
IFN	11	14	7e-10
PRL	10	14	3e-09
PDGF	10	12	8e-08
Oncostatin M	7	14	4e-07
T-cell antigen receptor	11	8	6e-06
IL-12 - STAT4	7	10	8e-05
B-cell antigen receptor	10	6	2e-04
ER-repressed	7	5	1e-02
IL-2 - STAT5	5	7	1e-02
IL-4 - STAT6	6	3	1e-01
Wnt	4	4	2e-01

Table 4: This table shows the number of samples (out of 117) in the van 't Veer data set with pathway status active (+) or inactive (−) and with *Group Sample Score* *p*-value ≤ 0.05. Also shown are the corresponding family-wise *p*-values. ER notation as in Table 2.

Association between sample-wise pathway activity and clinical classifications

We analyzed the association between sample-wise pathway activity and clinical classifications using contingency tables. For every pathway and data set, we divided the samples into three groups: Samples where the pathway was active at a 5% significance level, samples where it was inactive at a 5% significance level, and insignificant samples referred to as undecided. For each data set, contingency tables of pathway activity versus clinical classifications were created, and χ^2 p-values were calculated.

In the data set of Golub *et al.*, the only available clinical classification is tumor type, *i.e.*, ALL or AML. Table 5 shows the contingencies for the Insulin pathway and the Il-1 pathway. Seven out of 29 pathways have contingency tables with a χ^2 p-value below 0.01.

	ALL	AML
Insulin pw(+)	1	5
Insulin pw(−)	15	0
Insulin pw(U)	11	6
p-value: 5e-04		

	ALL	AML
IL-1 pw(+)	0	6
IL-1 pw(−)	19	0
IL-1 pw(U)	8	5
p-value: 1e-05		

Table 5: Contingency tables for the ALL/AML status versus the Insulin and IL-1 pathways in the leukemia data set of Golub *et al.* [4]. Active, non-active and undecided pathways are denoted +, - and U respectively.

For the breast cancer data set of van 't Veer *et al.*, we investigated six clinical classifications: metastasis status (0), estrogen receptor status (20), progesterone receptor status (12), lymph node status (12), BRCA mutations (15) and histological grade (8). The numbers in parentheses refer to the number of significant contingency tables at the 0.01 level. The total number of pathways was again 29. For metastasis status, only 97 out of the 117 samples were labeled in the original data set, and this may contribute to the low degree of association between this clinical classification and pathway activity. However, similar results were obtained for the data set of Sotiriou *et al.*, which indicates that it may be difficult to obtain any association between the pathways analyzed in this work and breast cancer metastasis status. Table 6 shows the contingency between estrogen receptor status and the ER-induced pathway. As expected, there is a strong association between presence of the estrogen receptor protein, and the activity status of the ER-induced pathway. Somewhat more surprisingly, there are also strong associations between ER status and many other pathways. Similar results are obtained for the data set of Sotiriou *et al.*, but with fewer significant associations.

	ERp low	ERp med.	ERp high
ER-ind pw(+)	0	5	25
ER-ind pw(−)	31	3	3
ER-ind pw(U)	8	16	26
p-value: 2e-14			

	ERp low	ERp med.	ERp high
RANK pw(+)	25	2	2
RANK pw(−)	1	5	24
RANK pw(U)	13	17	28
p-value: 1e-11			

Table 6: Contingency tables of estrogen receptor protein (binned at three levels: 0, 5-50, 60-100) versus the ER-induced and RANK pathways in the breast cancer data set of van't Veer *et al.* [16]. Same notation as in Table 5.

A general tendency of the contingency table analysis is illustrated in Table 7. Lowering the pathway activity *p*-value cutoff makes the association to clinical classifications more specific but less sensitive. The complete set of contingency tables for all three data sets can be found in the Supplement [1].

cutoff : 0.05	L+	L−
IL-12/STAT4 pw(+)	0	7
IL-12/STAT4 pw(−)	9	0
IL-12/STAT4 pw(U)	80	21
p-value: 3e-06		

cutoff: 0.1	L+	L−
IL-12/STAT4 pw(+)	2	8
IL-12/STAT4 pw(−)	13	0
IL-12/STAT4 pw(U)	74	20
p-value: 2e-05		

Table 7: Contingency table of lymph node infiltration status versus the IL-12/STAT4 pathway in the van 't Veer data set. The left and right tables are obtained with a pathway activity cutoff at 0.05 and 0.1 respectively. Same notation as in Table 5.

Conclusion and outlook

We have shown that downstream target genes of signal transduction pathways behave coherently in gene expression tumor data sets. First, we confirmed that downstream targets of transcription factors are correlated across samples. We then demonstrated that the same holds true for downstream targets of an entire pathway, even after discounting the correlations due to genes having a common transcription factor. The correlations for entire pathways were found to be more significant than those for individual transcription factors.

The presence of significant correlations confirms the expectation that gene expression is controlled by the activity of pathways. However, these correlations do not tell us in which samples a pathway is active or inactive. To reveal this, we devised the *Group Sample Score*. With this score we classified the samples into those where the pathway was significantly

active, significantly inactive or undecided, respectively. As seen in Table 4, the number of significant samples is, for most pathways, much higher than the random expectation.

In many cases, the active/inactive pathway status was highly correlated with independent clinical classifications. This confirms the relevance of pathways for understanding of the underlying biology. Furthermore, the activity status of one or more pathways may be used to subdivide the samples into groups with distinct biological characteristics. Such a subdivision is feasible if, for instance, tumors of a certain clinical diagnosis are an agglomerate of several subtypes.

The *Group Sample Score* is natural if a pathway either induces all its downstream targets, or represses them. However, in most pathways some downstream genes are induced, while others are repressed. To account for a mixture of induction and repression, one should include a sign, or more generally a weight, to each term in the sum. Such a weight might even depend on the type of tissue and the environment. Since this information was not readily available for the studied pathways, all genes were weighted equally. Nevertheless, we obtained significant results, indicative of a dominant trend among the downstream genes. For the estrogen receptor (ER) pathway, we did have information about the sign, but instead of introducing a more general score for this pathway alone, we split the ER pathway into two parts, with induced and repressed genes, respectively. In the breast cancer data set of van 't Veer *et al.* [16], there were 50 genes in the induced part and 27 in the repressed. As seen in Tables 2 and 4, the ER-induced pathway was highly significant, whereas the repressed pathway was not. The full pathway was also highly significant, although to a lesser extent. The significance of the full pathway is thus due to the induced genes, which constitute a majority of the downstream targets. The situation is similar for other pathways and data sets.

It should be stressed that correlations, and the pathway activity status observed in a sample, are only defined relative to the other samples in the same data set. If a pathway were active in all samples, it would not show up in our significance test. The status of a pathway, as we define it, is given by the downstream genes, and the connection to ligands, receptors and other pathway components cannot be inferred from this analysis.

Table 1 shows that the most significant transcription factor in the breast cancer data set of van 't Veer *et al.* is NF-$\varkappa$B. This transcription factor is also the most one in the leukemia data set of Golub *et al.*, whereas NF-$\varkappa$B1 is the most significant one in the data set of Sotirou *et al.*. Recently, NF-$\varkappa$B has been shown to be involved in the transformation from benign to malignant cells in inflammation-associated cancers. Pikarsky *et al.* [12] demonstrate this in a mouse model of human hepatocellular carcinoma, where the inflammatory mediator tumor-necrosis factor-α (TNF-α) is shown to play an important role as an activator of NF-$\varkappa$B. Greten *et al.* [6] find similar results in a mouse model of colitis-associated cancer.

Our current knowledge of pathways, and of downstream targets of transcription factors, is far from complete. However, we find that the results presented herein constitute a proof of concept for analyzing microarray gene expression in the context of signal transduction pathways.

Acknowledgments

This work was in part supported by the Swedish Foundation for Strategic Research (TB and MK), the Swedish Research Council (CP) and the Swedish National Research School for Bioinformatics and Genomics (CT).

References

[1] Supplementary material is available at:
`http://www.thep.lu.se/pub/Preprints/04/lu_tp_04_32_supp.zip`

[2] G. F. Berriz, O. D. King, B. Bryant, C. Sander, and F. P. Roth. Characterizing gene sets with funcassociate. *Bioinformatics*, 19(18):2502–2504, Dec 2003.

[3] S. Draghici, P. Khatri, R. P. Martins, G. C. Ostermeier, and S. A. Krawetz. Global functional profiling of gene expression. *Genomics*, 81(2):98–104, Feb 2003.

[4] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, Oct 1999.

[5] N. R. Gough. Science's signal transduction knowledge environment: the connections maps database. *Ann N Y Acad Sci*, 971:585–587, Oct 2002.

[6] F. R. Greten, L. Eckmann, T. F. Greten, J. M. Park, Z. W. Li, L. J. Egan, M. F. Kagnoff, and M. Karin. Ikkbeta links inflammation and tumorigenesis in a mouse model of colitis-associated cancer. *Cell*, 118(3):285–96, Aug 2004.

[7] S. P. Gygi, Y. Rochon, B. R. Franza, and R. Aebersold. Correlation between protein and mRNA abundance in yeast. *Mol Cell Biol*, 19(3):1720–30, Mar 1999.

[8] M. Krull, N. Voss, C. Choi, S. Pistor, A. Potapov, and E. Wingender. Transpath: an integrated database on signal transduction and a tool for array analysis. *Nucleic Acids Res*, 31(1):97–100, Jan 2003.

[9] C. Y. Lin, A. Strom, V. B. Vega, S. L. Kong, A. L. Yeo, J. S. Thomsen, W. C. Chan, B. Doray, D. K. Bangarusamy, A. Ramasamy, L. A. Vergara, S. Tang, A. Chong, V. B. Bajic, L. D. Miller, J. A. Gustafsson, and E. T. Liu. Discovery of estrogen receptor alpha target genes and response elements in breast tumor cells. *Genome Biol*, 5(9):R66, 2004.

[10] V. K. Mootha, C. M. Lindgren, K. F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrale, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler, and L. C. Groop. Pgc-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, 34(3):267–73, Jul 2003.

[11] R. Pandey, R. K. Guru, and D. W. Mount. Pathway miner: extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data. *Bioinformatics*, 20(13):2156–2158, Sep 2004.

[12] E. Pikarsky, R. M. Porat, I. Stein, R. Abramovitch, S. Amit, S. Kasem, E. Gutkovich-Pyest, S. Urieli-Shoval, E. Galun, and Y. Ben-Neriah. Nf-kappab functions as a tumour promoter in inflammation-associated cancer. *Nature*, 431(7007):461–466, Sep 2004.

[13] J. Rahnenfuhrer, F. S. Domingues, J. Maydt, and T. Lengauer. Calculating the statistical significance of changes in pathway activity from gene expression data. *Statistical Applications in Genetics and Molecular Biology*, 3(1), 2004.

[14] M. D. Robinson, J. Grigull, N. Mohammad, and T. R. Hughes. Funspec: a web-based cluster interpreter for yeast. *BMC Bioinformatics*, 3(1):35, Nov 2002.

[15] C. Sotiriou, S. Y. Neo, L. M. McShane, E. L. Korn, P. M. Long, A. Jazaeri, P. Martiat, S. B. Fox, A. L. Harris, and E. T. Liu. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci U S A*, 100(18):10393–10398, Sep 2003.

[16] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536, Jan 2002.

[17] E. Wingender, X. Chen, E. Fricke, R. Geffers, R. Hehl, I. Liebich, M. Krull, V. Matys, H. Michael, R. Ohnhauser, M. Pruss, F. Schacherer, S. Thiele, and S. Urbach. The transfac system on gene expression regulation. *Nucleic Acids Res*, 29(1):281–283, Jan 2001.

[18] B. R. Zeeberg, W. Feng, G. Wang, M. D. Wang, A. T. Fojo, M. Sunshine, S. Narasimhan, D. W. Kane, W. C. Reinhold, S. Lababidi, K. J. Bussey, J. Riss, J. C. Barrett, and J. N. Weinstein. Gominer: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*, 4(4):R28, 2003.

[19] A. Zien, R. Kuffner, R. Zimmer, and T. Lengauer. Analysis of gene expression data with pathway scores. *Proc Int Conf Intell Syst Mol Biol*, 8:407–17, 2000.