Multimodel Pathway Enrichment Methods for Functional Evaluation of Expression Regulation

ARTICLE *in* JOURNAL OF PROTEOME RESEARCH · APRIL 2012

Impact Factor: 4.25 · DOI: 10.1021/pr300038b · Source: PubMed

Paolo Cifani
Memorial Sloan-Kettering Cancer Center

Ann-Sofie Albrekt
Lund University

Malin Lindstedt
Lund University

Anders Heyden
Lund University

Journal of proteome research

Article

pubs.acs.org/jpr

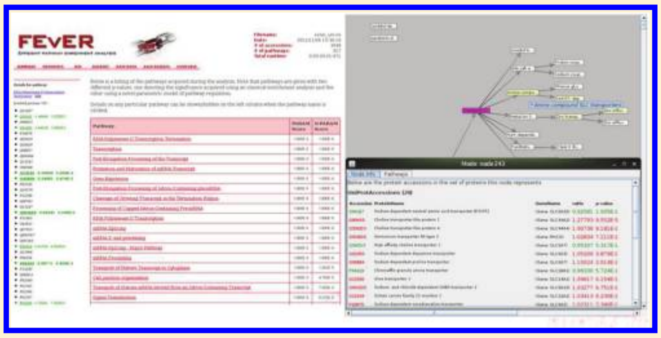
Multimodel Pathway Enrichment Methods for Functional Evaluation of Expression Regulation

Ufuk Kirik,[†] Paolo Cifani,[†] Ann-Sofie Albrekt,[†] Malin Lindstedt,[†] Anders Heyden,[‡] and Fredrik Levander*,[†]

[†]Department of Immunotechnology, Lund University Biomedical Centre D13, SE-221 84 Lund, Sweden

[‡]Centre for Mathematical Sciences, Lund University Box 118, SE-22100, Lund, Sweden

S *Supporting Information*



ABSTRACT: Functional analysis of quantitative expression data is becoming common practice within the proteomics and transcriptomics fields; however, a gold standard for this type of analysis has yet not emerged. To grasp the systemic changes in biological systems, efficient and robust methods are needed for data analysis following expression regulation experiments. We discuss several conceptual and practical challenges potentially hindering the emergence of such methods and present a novel method, called FEvER, that utilizes two enrichment models in parallel. We also present analysis of three disparate differential expression data sets using our method and compare our results to other established methods. With many useful features such as pathway hierarchy overview, we believe the FEvER method and its software implementation will provide a useful tool for peers in the field of proteomics. Furthermore, we show that the method is also applicable to other types of expression data.

KEYWORDS: *pathway enrichment, pathway analysis, functional analysis, systems biology, quantitative proteomics, expression data analysis*

■ INTRODUCTION

With the state-of-the-art instruments, current methods in proteomics provide thousands of proteins that are differentially expressed between samples, resulting in a tremendous swell in the depth and amount of data produced within the field. In particular, mass spectrometry (MS)-based methods have come to the point where the relative[1−3] or absolute[4] expression levels of several thousand proteins can be measured in complex samples. While paving the way for a whole new set of possibilities in extending our understanding of biological systems, these recent developments in the field bring a number of challenges to be conquered by the proteomics community.

One of these challenges is the growing need for thorough and effective methods for interpretation of the exponentially increasing experimental data. This need is particularly apparent in comparative experiments where differential expressions of genes or proteins are studied. In the context of differential gene expression, methods focusing on individual entities that show the most significant change were shown to be insufficient for comprehensive understanding of the underlying mechanisms.[5] Consequently, methods for systematic analysis of comparative expression data have begun to emerge. However, the task of identifying pathways and networks that have been subject to systemic regulation have several conceptual and practical challenges that need to be tackled.

One such challenge is based on the fact that there are significant differences in the quality of the underlying data based on the acquisition method. In comparison to transcriptomics experiments where mRNA from all genes of a genome can be covered, data sets in MS-based proteomics are still rather incomplete since only a portion of the proteome can be quantitatively measured due to the dynamic range problem associated with stochastic sampling. Furthermore, it has been argued that while the sampling process is shown to be biased toward highly abundant proteins, each replicate analysis will not necessarily sample the same portion of the proteome and thus complicates the further analysis of the results.[6−9] Moreover, distinctive quantification of proteins from detected peptides is often not possible, which leads to measured intensities being associated with groups of proteins that cannot be distinguished from one another, instead of individual proteins. Due to these properties of proteomics data, methods devised within the transcriptomics field are not necessarily suitable for MS-based proteomics data out-of-the-box.

Additionally, the lack of a clear definition of the notion of a pathway is a very significant conceptual challenge in efficient, optimized and robust functional analysis of expression data. Functional networks are defined in different ways that are not always compatible with one another. It is plausible that the problem lies, at least partly, in the fact that pathways refer to an abstract concept instead of a physical entity, which makes the

Received: January 13, 2012
Published: April 3, 2012

dx.doi.org/10.1021/pr300038b | *J. Proteome Res.* 2012, 11, 2955−2967

notion subject to interpretation. Consequently, the definition of what a pathway is and what is contained within a pathway varies depending on the database it originates from. Pathway representations originating from different sources describing the same biological knowledge might have significant differences in content, making both data integration (due to redundancies and data formats used) and results validation (due to discrepancies) complicated problems to solve.

Existing Methods

Several approaches for systemic analysis of proteomic data have emerged during the past few years; both as adaptations of existing methods for microarray data, and novel methods for proteomics data. One approach is using adaptations of the now established gene set enrichment analysis (GSEA)[5] to protein expression data.[10] Many other implementations of such over-representation analysis have been described in the literature since the GSEA method was published, and a review of many of these methods has been published by Nam and Kim 2008.[11] The major difference between methods implementing this type of over-representation analysis is in the choice of the statistical tests used to deliver the significance of the calculated enrichment scores.

Besides adaptations of methods developed for gene expression data, a number of different methods and software tools that aim for systemic analysis and visualization of MS-based proteomics data have been mentioned in the literature recently.[12−16] These methods vary significantly in their scope, ultimate goal, and form of implementation. Moreover, several more integrative efforts have emerged, one such effort is Reactome Pathway Analysis,[17] a web-based tool developed for the Reactome pathway database. Additionally there are commercial solutions for such pathway analysis, for example, Ingenuity Pathway Analysis (IPA)[18] and MetaCore.[19]

Despite the remarkable efforts from research groups worldwide, uncertainty remains regarding a standard, widely accepted method for pathway level analysis of expression data. The methods and tools mentioned above, and many others described in the literature, vary fundamentally in their scope, degree of software implementation and the resources they exploit. Likewise they require different levels of expertise and familiarity with different software environments in order for users to fully grasp and utilize the potential of the methods. An additional drawback of commercial software solutions is the lack of flexibility and transparency; individuals do not have the liberty to look into the source code, study the algorithms used and modify them as needed, since these products are not open source.

Here we describe a novel method based on combining two fundamentally different models with complementary statistical tests for effective analysis of pathway enrichment associated with comparative expression proteomics data. We show the potential of this method with an in-house implementation, which addresses several of the challenges mentioned above, by analysis of three independent data sets. The implementation also provides a user-friendly interface for quick and easy overview of analysis results.

■ EXPERIMENTAL PROCEDURES

The computational method we introduce is built around two main models for pathway enrichment; one based on a parametric evaluation of proteins that are involved in pathways, and the other based on nonparametric assessment of over-represented pathways based on differential expression data, see Figure 1.

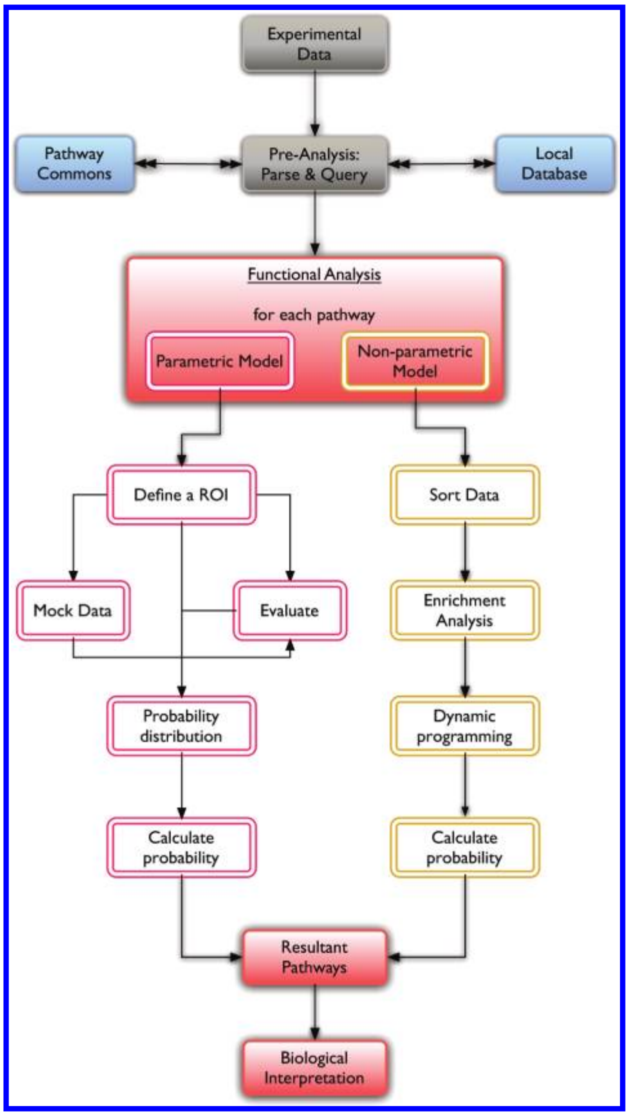


Figure 1. FEvER analysis workflow. Comparative expression data is parsed and queried using PathwayCommons database. To minimize the query time and the load put on remote servers, our implementation of FEvER method uses a local cache database, which stores previous queries made to PathwayCommons. The functional evaluation of the pathways is done using two separate pathway enrichment models, one that utilizes a user-defined region of interest (ROI) and one that is an adaptation of gene set enrichment analysis to proteomic data. Both models return probability measures as scores, and the results are then assembled into reports for interpretation.

Enrichment Models

The foundation of the parametric model is based upon a number of central assumptions regarding expression regulation on a pathway level. The first of these assumptions is that it is statistically unlikely that a high number of proteins in a pathway are regulated significantly by pure chance. This assumption follows from fundamental combinatorics and is motivated in further detail in the Supporting Information (S1). Furthermore, a majority of the proteins that have pathway annotations are associated with multiple pathways, which leads to, what can be referred to as “pathway identification ambiguity”. On the basis

Table 1. Implemented Methods for Generation of Mock Data

	Methods for sampling mock ratio values
Empirical	Relies on creating an empirical ratio distribution from the experimental values using a method known as Variable Kernel Method together with Gaussian smoothing. Ratios for mock data are then sampled from this empirical distribution.
Logarithmized Gaussian	Assumes that a protein is just as likely to be up-regulated as down-regulated, and that fold change values follow a Gaussian distribution around 1. Thus ratios for the mock data are sampled by evaluating the exponential function of normal-distributed random variables.
Permutation	The same set of ratios is used for the mock data by permutation. Randomization relies entirely on the assumption that the permuted values are not correlated with the experimental data.

of this observation, the second assumption follows: pathways identified with proteins that do not have many pathway associations have less ambiguity and thus are easier to allocate to a biological function. The third, and the last, assumption of the model is that the expression regulation of a pathway is correlated tightly with the overall expression regulation of the proteins in that pathway. The support for such an assumption is rather weak, as there are many different factors to functional regulation such as the post-translational modifications (PTMs) and kinematic constants in the biochemical reactions encompassed within the pathway. In that sense, it is unrealistic to build an accurate model of pathway regulation by only considering the relative abundances of a number of proteins in that pathway. However the goal of the model is to point out pathways that are likely perturbed and of interest for further studies, since exact in vivo properties of proteins are, in most cases, not available. Nevertheless, we will be investigating ways to implement differential PTM analysis in newer versions of the FEvER method.

Following the three assumptions mentioned above, the parametric model evaluates pathways using a region of interest (ROI), defined by the user with a set of parameters prior to the start of the analysis. The ROI defines a primary focus region within the topology of the input data set and a metric by which expression regulation is evaluated. This focus region covers the proteins, which are considered biologically or statistically more significant than the rest of the data set, based on the metric defined. Consequently the proteins in the ROI contribute more toward a final enrichment score than the rest of proteins in the data set. The borders of the ROI are defined using thresholds for expression values of the proteins in the data set such as, abundance ratios and statistical significance of these ratios. What we mean by statistical significance here is essentially a measure of the variability of the abundance ratios, and it is assumed to be calculated from replicates prior to the pathway analysis and plays an important role in the following steps of the analysis. Furthermore, it is crucial to note that these *p*-values should be compensated for multiple hypotheses testing, in order to get the most reliable results.

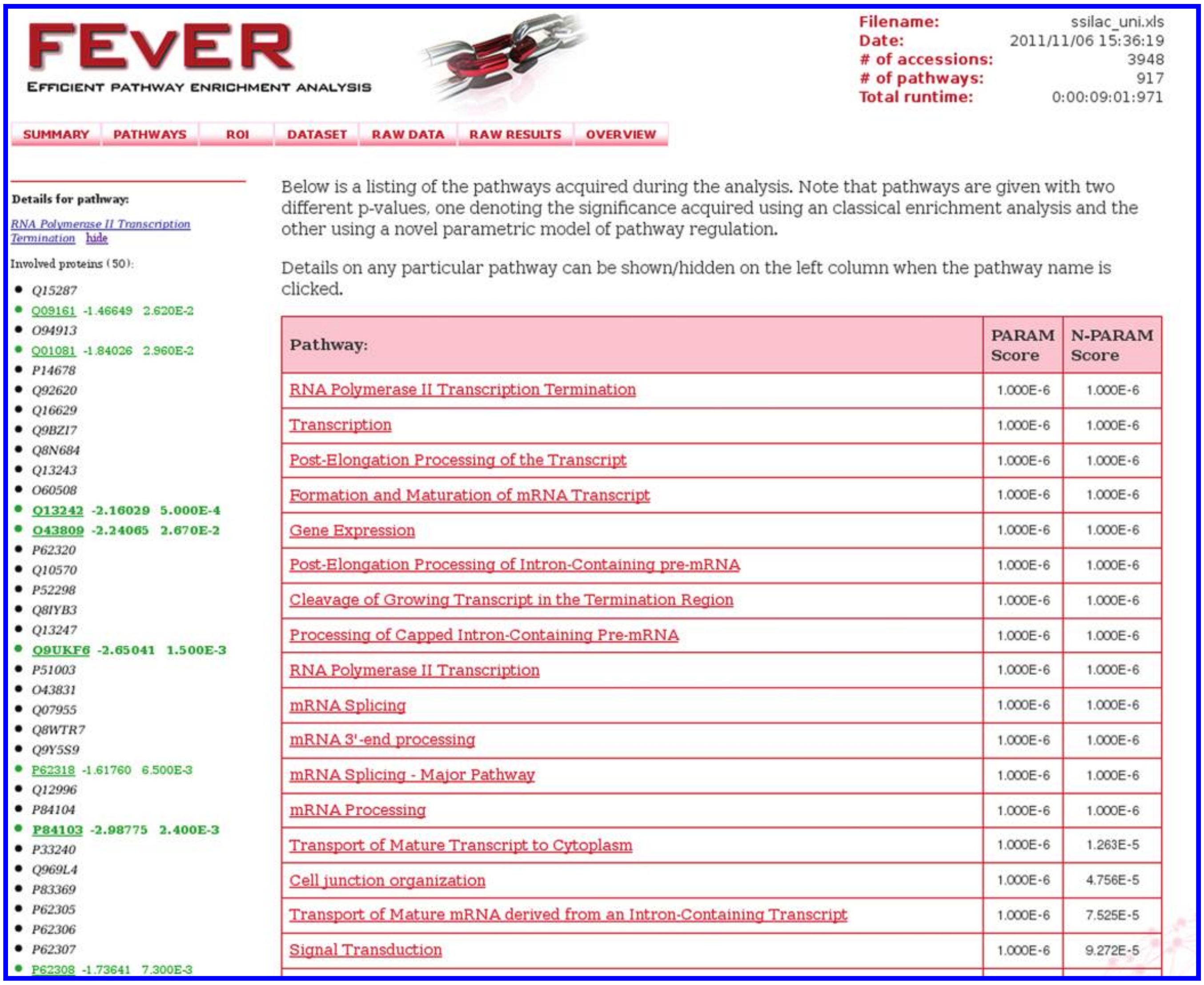
On the basis of the canonical pathway information acquired from public databases, the experimental data and the ROI defined by the user, an enrichment score is calculated (see Supporting Information S1). To assess the significance of this score, mock data is sampled using one of several different methods (see Table 1). The same pathway is re-evaluated extensively using the mock data sets, in a Monte Carlo manner giving an estimate of the probability of observing the same enrichment score, or better, as the final output of the model. In summary, this model is aimed to highlight pathways that are over-represented in the experimental data in the sense of containing proteins that have statistically reliable fold change values and few pathway associations.

The parametric model described above is a novel approach to use multiple different attributes of the proteins featured in the experimental results, which has several advantages over existing methods, as we show in the Results section. However, the model is based on a number of assumptions and requires the user to be involved in the analysis by defining a small set of parameters. To complement and, to a certain extent, even validate the parametric method, we coupled it with an established nonparametric method, a set-enrichment method initially described by Subramanian et al.[5] to run in parallel to the parametric model. This kind of enrichment analysis is well established within the field and is explained multiple times in literature with different implementations.[5,11,20−22]

Significance Testing

One aspect of enrichment analysis methods have been discussed thoroughly in literature, namely the null hypothesis testing.[11,21,23] It is essential to put forward a statistically and biologically valid null hypothesis, in order to draw sound conclusions following the hypothesis testing. Goeman and Buhlmann have proposed a classification of the existing enrichment methods, based on the statistical methods used in hypothesis testing.[21] According to this classification, the methods could in general be divided into two groups, those that are "self-contained", and those that are "competitive". This distinction is based on the formulation of the null hypotheses that are being tested in these methods. The difference in between the two types of hypotheses is that a competitive test compares the differential expression of a set to the complement of the set. In practical terms, this means testing whether the proteins of a pathway are expressed at most as often differentially expressed as the proteins that are not encompassed in that pathway. In contrast, a self-contained test is then defined to only consider the proteins in a pathway and test whether or not these proteins are differentially expressed. It has been discussed previously that self-contained and competitive tests are not objectively comparable in terms of statistical power and that they test different aspects of the data.[24]

In this context, the test used in the parametric model described above is closer to a self-contained test than a competitive one, as the proteins that are not contained within the pathway are not used in the model in any way, and do not contribute to the significance testing. In implementation of the nonparametric enrichment model, on the other hand, we have decided to use a competitive test. The motivation behind this decision was partly to complement the parametric model with a different type of test, and partly to assess the difference between the two approaches. Moreover instead of using a large number of permutations (a common practice among the available methods) to test the significance of the maximum enrichment score (MES) for a pathway MES_{obs}, we have implemented an algorithm, which calculates the exact probabilities of observing a MES equal to or greater than the MES_{obs}, using a technique called dynamic programming. The use of dynamic programming for calculation of significance calculations have been



Filename:	ssilac_uni.xls
Date:	2011/11/06 15:36:19
# of accessions:	3948
# of pathways:	917
Total runtime:	0:00:09:01:971

Details for pathway:

RNA Polymerase II Transcription Termination hide

Involved proteins (50):

- Q15287
- Q09161 -1.46649 2.620E-2
- O94913
- Q01081 -1.84026 2.960E-2
- P14678
- Q92620
- Q16629
- Q9BZI7
- Q8N684
- Q13243
- O60508
- **Q13242 -2.16029 5.000E-4**
- **O43809 -2.24065 2.670E-2**
- P62320
- Q10570
- P52298
- Q8IYB3
- Q13247
- **Q9UKF6 -2.65041 1.500E-3**
- P51003
- O43831
- Q07955
- Q8WTR7
- Q9Y5S9
- P62318 -1.61760 6.500E-3
- Q12996
- P84104
- **P84103 -2.98775 2.400E-3**
- P33240
- Q969L4
- P83369
- P62305
- P62306
- P62307
- P62308 -1.73641 7.300E-3

Below is a listing of the pathways acquired during the analysis. Note that pathways are given with two different p-values, one denoting the significance acquired using an classical enrichment analysis and the other using a novel parametric model of pathway regulation.

Details on any particular pathway can be shown/hidden on the left column when the pathway name is clicked.

Pathway:	PARAM Score	N-PARAM Score
RNA Polymerase II Transcription Termination	1.000E-6	1.000E-6
Transcription	1.000E-6	1.000E-6
Post-Elongation Processing of the Transcript	1.000E-6	1.000E-6
Formation and Maturation of mRNA Transcript	1.000E-6	1.000E-6
Gene Expression	1.000E-6	1.000E-6
Post-Elongation Processing of Intron-Containing pre-mRNA	1.000E-6	1.000E-6
Cleavage of Growing Transcript in the Termination Region	1.000E-6	1.000E-6
Processing of Capped Intron-Containing Pre-mRNA	1.000E-6	1.000E-6
RNA Polymerase II Transcription	1.000E-6	1.000E-6
mRNA Splicing	1.000E-6	1.000E-6
mRNA 3'-end processing	1.000E-6	1.000E-6
mRNA Splicing - Major Pathway	1.000E-6	1.000E-6
mRNA Processing	1.000E-6	1.000E-6
Transport of Mature Transcript to Cytoplasm	1.000E-6	1.263E-5
Cell junction organization	1.000E-6	4.756E-5
Transport of Mature mRNA derived from an Intron-Containing Transcript	1.000E-6	7.525E-5
Signal Transduction	1.000E-6	9.272E-5

Figure 2. End product of our implementation of the FEvER workflow is a report composed of a series of HTML pages that can be viewed in any web browser, on any platform. A screenshot of one of the pages of the report from one of our runs is displayed. We utilized Javascript language embedded in the reports to ensure the reports are as clean as possible and to provide additional information interactively to the user, when required. Details for each pathway, such as contained proteins and their expression values, can be shown/hidden with a simple mouse click, as shown in the figure. In the left column, the proteins that participate in the *RNA polymerase II transcription termination* pathway are shown; note that proteins are colored green/red depending on whether they are down-/up-regulated and written out in bold text if they fall within the region of interest set for the analysis.

described previously by Keller et al.[25] and a similar calculation model is used in GeneTrail.[26]

Implementation

We have designed a software tool, written in Java 6 environment, to test our main hypothesis, which using two complementary models for pathway enrichment leads to better results than a single one. The source code and binaries for this implementation, together with a brief user manual, are available at http://quantitativeproteomics.org/fever/. The software manages data handling tasks such as parsing experimental data, dealing with multiple identity problem for protein accessions and querying of pathway associations; evaluates the pathways associated with the proteins in the experimental data using the two models in parallel, and wraps the results in user-friendly reports for a quick and easy overview (Figure 2). These reports contain the resultant pathways with significance scores from the two models along with a consensus score (Supporting Information S1), as well as an overview of the input data set with links to other relevant data sources, such as UniProt, and also includes links for automated visualization of pathways using the ChiBE[27] tool with color-overlay to indicate fold change values for proteins.

The pathway knowledgebase is fragmented over numerous pathway databases, each containing only a portion. Considering that the these portions vary in biological scope, target organisms, size and data representation format, the task of integrating available data to base an analysis on becomes rather formidable. While several initiatives, for example, SBML[28] and BioPAX,[29] have been working on data representation standards to address this particular problem, the reality of the matter is still discouraging. In the implementation of this software tool, we have chosen to query Pathway Commons database[30] for pathway associations, which is a service that imports pathway data from multiple other databases into a single, freely available web resource. Furthermore, we have implemented a simple local data repository within the software to cache the results from previous analyses. An additional use for this local data repository is that it allows the possibility to integrate pathway information from other databases that are not featured in Pathway Commons.

In terms of performance, caching of previous query results eliminate the primary bottleneck in the process. In this scheme the first time FEvER is run, all featured proteins and pathways will be queried and cached. All successive searches will use the caches and add any “new” information whenever needed or when the database is updated. Furthermore, to increase efficacy of the analysis as well as decrease the process time, the software utilizes parallel processing making use of the modern multicore microprocessors. On an Intel Core i7 machine with 4 cores and 6 GB RAM, a typical proteomics data set with roughly 4000 proteins and 917 pathways featured, FEvER took 22 min to complete analysis with an empty cache. When the same data set was reanalyzed using the cached information the total process time was less than 11 min.

Handling Pathway Hierarchy

An additional challenge with the data management lies in the inherent hierarchy of pathways, which give rise to redundancy within the results, as discussed in the introduction. To get the most use from an enrichment analysis, this redundancy needs to be tackled. However, to our knowledge, this issue has not been addressed in the existing tools so far. We have, therefore, added a feature in our software tool to handle this hierarchy using a number of simple abstractions.

Pathways are usually defined as collections of biochemical reactions that occur in connection with one another toward a well-defined function. Pathway representation formats often have local resolution, that is, the cellular compartments the reactions take place in is usually described as well as the reactions themselves. A frequent scenario in many pathways is that a protein is involved in multiple reactions in different compartments. While this type of information is very useful in examination of a single pathway in detail, it cannot be used to evaluate potential expression regulation on a pathway level. This is due to the fact that the input contains merely a set of identified analytes with expression values. Since the score models evaluate the pathways based on the proteins contained within these pathways, it is possible to simplify pathways as distinctive sets of proteins. This abstraction allows set operations such as intersection and union to be executed on pathways. Using these set operations, one can define four different relations between pathways depending on their intersection and union with one another. Two pathways are deemed *disjoint* if they have no proteins in common, alternatively if all the proteins in one of the pathways are contained on the other one, the pathways are considered to have a *sub/super* relationship to one another. In the special case where the symmetric difference of the sets is empty, in other words neither set contains any proteins exclusively, the pathways are considered to be *equivalent*, with respect to the hierarchy.

Using this hierarchical relationship, our implementation converts the resultant pathways to nodes containing sets of proteins, and places these nodes in tree structures. The resultant “forest” of pathways is then visualized as a Java applet as a part of the reports, with a two-scale coloring scheme (corresponding to two scoring models) to represent how significant pathways are enriched in the experimental data set (Figure 3).

Test Data

In this study, we have used three data sets that reflect different types of quantitative expression data; one that is based on a comparative study of yeast in different growth media, a comparative analysis of human cancer samples in a, so-called, shotgun study and finally an mRNA expression study of the induced response by a sensitizing chemical in a myeloid human cell line.

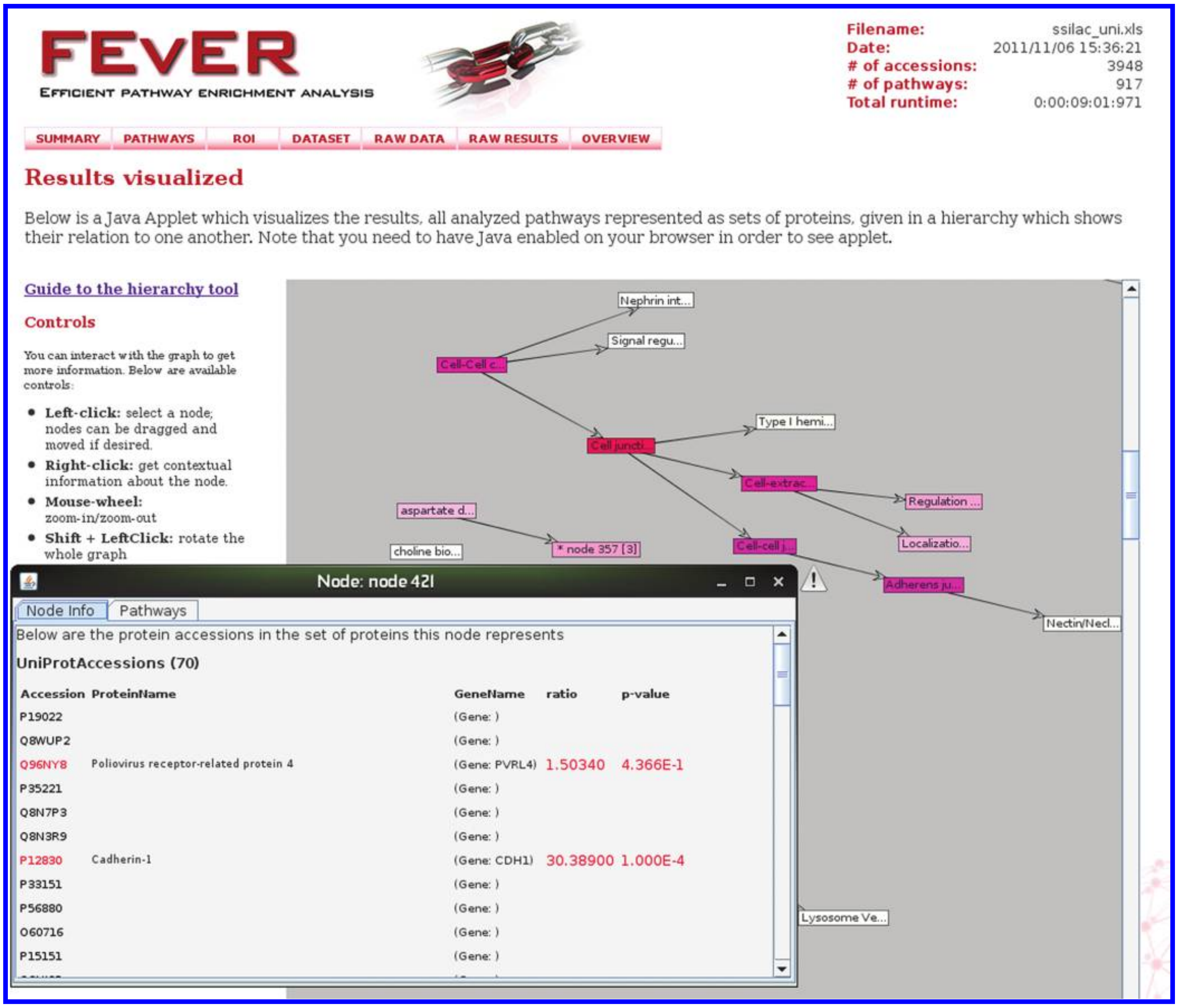
The yeast data set was from a comparison of batch cultures of *S. cerevisiae* grown in SILAC medium supplemented with 2% glucose or 3% ethanol (+0.5% glucose), with samples SCX-fractionated before LC−MS/MS and data analysis was done using MaxQuant,[31] as described by Olsson et al (Olsson et al., submitted, Tranche-hash: xNBoKsLK7paM+4ZEI3nSPo7TTVaZw4TrVyZtv22cgMiX8fW+km9Bky96GrXoduVjmWCBO1mmVSeEFKWvUZLUNlvuAzYAAAAAAABb6g==). Since the Pathway Commons database contained no pathways for yeast, but only interaction networks, we imported the YeastCyc database as the pathway data source. YeastCyc is a part of the BioCyc database collection[32] and as of 2011 contains 154 manually curated pathways, and the experimental protein data set consisted of 895 protein groups.

The tumor data is taken from a study in which Geiger et al describe the use of a blend of cancer cell lines as an internal standard in stable isotope labeling (SILAC)[33] shotgun-MS experiments.[34] One of the experiments in this study describes the use of the internal standard for a quantitative comparative study of ductal versus lobular breast cancer samples. The data set published, as Supporting Information,[34] contains 4318 protein groups that are identified and quantified. We have calculated the ratio of the means of the three replicate ratios, and applied a *t*-test to calculate the necessary *p*-values.

The transcriptomics data set used in this study is a small portion of a large study recently described by Johansson et al.,[35] investigating the changes in the transcriptome of a human cell line (MUTZ-3), following 24 h stimulation by both sensitizing and nonsensitizing chemicals as well as vehicle controls. For our purpose, we have narrowed down the scope and considered the data comparing one particular sensitizing agent, 1-chloro-2,4-dinitrobenzene (DNCB), versus a pool of 20 samples from nonsensitizing agents, in triplicates. This data set constitutes a significant stress test for our method since there are approximately 22000 entities in the data set for which a regulatory expression value is measured, compared to a typical proteomics experiment where only a fraction of the proteome can be measured.

■ RESULTS AND DISCUSSION

To put our method and software to test, we have used a series of data sets. In doing so, we demonstrate its applicability in different situations with varying types of data. A pivotal challenge that needs to be addressed to validate pathway-based methods is the fact that pathways are abstract and human-defined, in contrast to natural entities that can be measured experimentally. In addition, the abstract nature of pathways, the combined knowledgebase on pathway regulation, is still rather incomplete. As a consequence, there are no definitive tests to check whether or not the models give the right results. Put in more formal terms, when working with pathway level consequences of expression regulation, the assessment of false



Filename:	ssilac_uni.xls
Date:	2011/11/06 15:36:21
# of accessions:	3948
# of pathways:	917
Total runtime:	0:00:09:01:971

Results visualized

Below is a Java Applet which visualizes the results, all analyzed pathways represented as sets of proteins, given in a hierarchy which shows their relation to one another. Note that you need to have Java enabled on your browser in order to see applet.

Guide to the hierarchy tool

Controls

You can interact with the graph to get more information. Below are available controls:

- **Left-click:** select a node; nodes can be dragged and moved if desired.
- **Right-click:** get contextual information about the node.
- **Mouse-wheel:** zoom-in/zoom-out
- **Shift + LeftClick:** rotate the whole graph

Node: node 421

Below are the protein accessions in the set of proteins this node represents

UniProtAccessions (70)

Accession	ProteinName	GeneName	ratio	p-value
P19022		(Gene:)		
Q8WUP2		(Gene:)		
Q96NY8	Poliovirus receptor-related protein 4	(Gene: PVRL4)	1.50340	4.366E-1
P35221		(Gene:)		
Q8N7P3		(Gene:)		
Q8N3R9		(Gene:)		
P12830	Cadherin-1	(Gene: CDH1)	30.38900	1.000E-4
P33151		(Gene:)		
P56880		(Gene:)		
O60716		(Gene:)		
P15151		(Gene:)		

Figure 3. Using the hierarchy overview tool, it is possible to navigate the results generated by the analysis in an interactive manner. Resultant pathways are placed, as nodes, in a collection of hierarchical tree-like structures based on the proteins that are involved in each pathway. If multiple pathways contain the exact same set of proteins, they are placed in the same node, as the score models will not be able to discriminate them from one another. Each node contains information about the pathway(s) the node represents, such as the database it originates from and the scores acquired from the two models as well as the proteins that are contained within the pathway(s). To provide a quick overview of the results of the analysis, the nodes are painted using a two-variable (hue and intensity) coloring scheme. Each model gives signal on a particular color (magenta or yellow), while the total intensity of the color indicates how significant the scores are. In this color scheme, bright red nodes represent the absolutely most significant pathways, while white nodes represent pathways that have rated very insignificant with both models.

positives and false negatives becomes complicated, especially in distinguishing what is "clearly false" from what might be yet unknown or poorly understood.

Initial Validation

One way to address this issue is to start validating the results by looking at the true positives. In other words, we aimed to validate whether the FEvER method could highlight known systemic changes in well-studied, relatively simple experimental setups. For this purpose, we have analyzed a data set holding a quantitative proteomic comparison of yeast grown in glucose or ethanol. We expected to get a very clear set of changes in metabolic activity in yeast, due to the different growth media. Indeed, we got prominently high scores for pathways within the carbon metabolism, in particular top scores for the *TCA cycle* and *glyoxylate cycle* pathways, as well as aerobic respiration and electron transport chain pathway (Table 2). Proteins that are involved in these pathways are very significantly up-regulated in yeast grown in ethanol, which correlate well with previous results on similar studies.[36,37]

Furthermore, we observed a recognizable signal from *gluconeogenesis* pathway by the parametric model. This finding is not particularly surprising as the yeast cells in the ethanol-rich environment would need to use the carbons in the medium to synthesize necessary carbohydrates, in a reverse glycolytic manner.[37] Looking at individual proteins, we observe expression values supporting the previous statements, for example, alcohol dehydrogenase II (ADH2) was up-regulated more than 40 times in yeast cells growing in ethanol. ADH2 is shown to be a crucial element in catalyzing oxidation of ethanol to

Table 2. *p*-Values of the 10 Most Significant Pathways from a Comparative Analysis of Functional Regulation of the Yeast Proteome on Different Carbon Sources (Glucose or Ethanol)[a]

pathway name	parametric model	nonparam model
Aerobic respiration, electron transport chain	1.000×10^{-6}	1.000×10^{-6}
Superpathway of TCA cycle and glyoxylate cycle	1.000×10^{-6}	1.000×10^{-6}
TCA cycle, aerobic respiration	1.000×10^{-6}	1.000×10^{-6}
Glyoxylate cycle	1.000×10^{-6}	5.176×10^{-4}
Gluconeogenesis	1.958×10^{-3}	1.176×10^{-1}
Sucrose degradation	4.887×10^{-3}	7.282×10^{-2}
Removal of superxide radicals	7.921×10^{-3}	2.491×10^{-1}
2-ketogluterate dehydrogenase complex	9.019×10^{-3}	8.017×10^{-2}
Leucine degradation	9.138×10^{-3}	5.447×10^{-1}
Superpathway of glucose fermentation	1.005×10^{-2}	8.520×10^{-2}

[a] Note that 1×10^{-6} is the lowest possible p-value.

acetaldehyde, which leads to utilization of ethanol as the source of carbon.[38−40] We could thus conclude that the FEvER tool readily distinguished relevant pathways in this data set and that the parametric model was successful in identifying the *gluconeogenesis* pathway as significant, which the established nonparametric model did not.

Stress Test

As a second benchmark test, we picked a data set that represents the qualities that one would expect to have in a typical quantitative proteomics study such as multiple identifications per expression ratio, missing values in the form of identifications without ratios or ratios without any identification and containing a few thousand proteins in total. Pathways consisting of only high-abundance proteins are typically well covered by the experimental data, while some pathways only have one or few proteins identified and quantified. Pathway coverage is an important variable since to draw conclusions regarding expression regulation on pathway level it is necessary to have data on the protein level. In that sense, the fewer entities observed in the experiment, the less power conclusions will have following the pathway analysis, and this was one of the facts that were accounted for when generating the parametric model.

We used one of the data sets from a study by Geiger et al.,[34] a differential expression study of ductal breast cancer samples compared to lobular samples. Geiger and colleagues do not go into a comprehensive investigation of systemic regulation in this paper, however they do note that they have observed more focal adhesion proteins, as well as glycolytic proteins, in ductal sample compared to lobular. Furthermore they note that strong changes in expression of certain proteins such as E-cadherin, β1-integrin and pyruvate kinases are indicative of cell adhesion and glycolytic changes in ductal tumors; while overexpression of Cdc2 and MCM proteins in lobular tumor samples point to DNA replication and cell cycle pathways.[34]

At first glance, FEvER analysis of this data set highlights a number of large functional networks such as gene expression, metabolism and signaling transduction (Figure 2, Supplementary S8, Supporting Information). Further investigation of the reports however reveal further information, in detail, on which branches of these networks have scored particularly higher scores. In the case of networks related to gene expression, mRNA related pathways show very high scores due to very significant expression differences within the *Transcription* superpathway. In particular, the *RNA polymerase II (pol II) transcription* pathway scored the highest possible significance

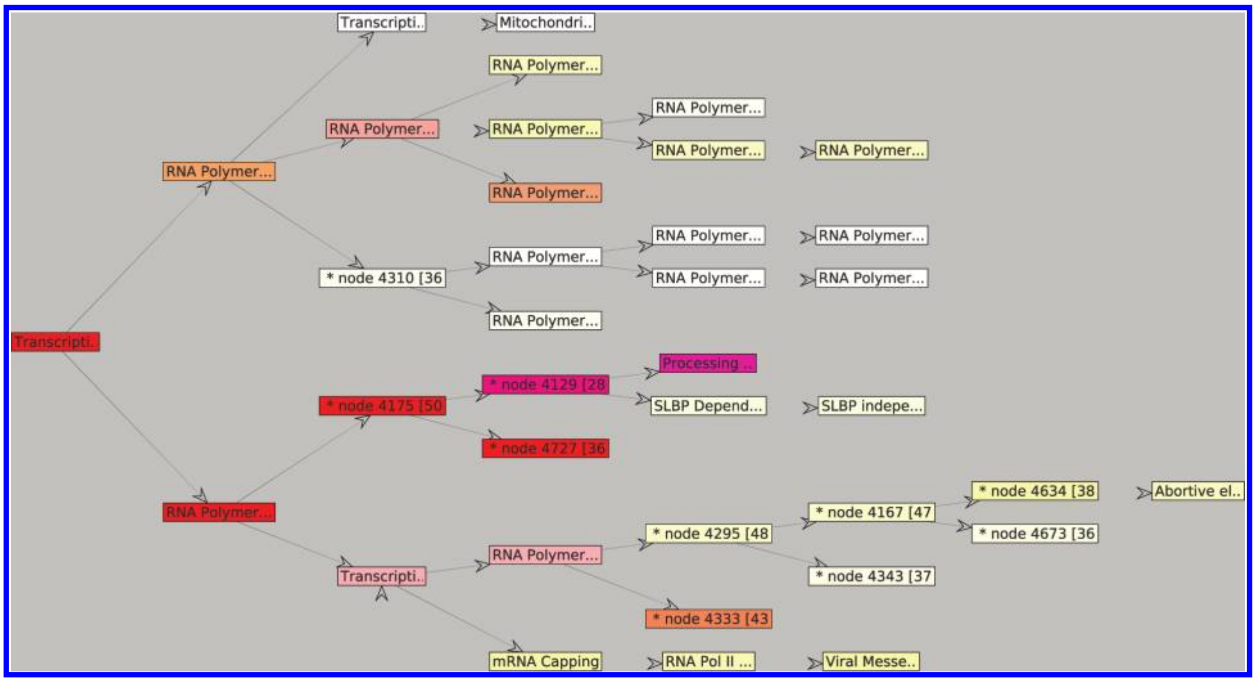


Figure 4. One of the top scoring superpathways in the ductal-lobular data set is the *Transcription* pathway, which also includes subpathways. This pathway has two main branches, namely, one covering the activity related to RNA polymerase I (pol I), RNA polymerase III (pol III) and mitochondrial transcription, and another covering RNA polymerase II (pol II) transcription network. The pathway hierarchy below shows much higher scores on the pol II branch. However, there is some signal on pol I transcription, although much less than pol II, still below $p < 0.01$ level. Mitochondrial transcription and pol III pathways do not show any noteworthy regulation in this data set.

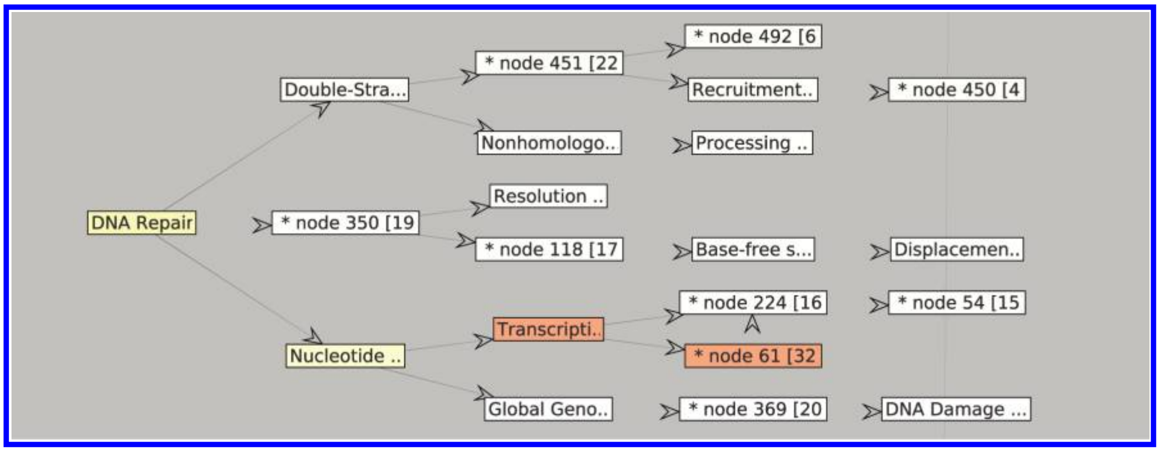


Figure 5. Pathway hierarchy overview tool reveals a very specific signal within the pathway hierarchy of DNA repair superpathway. The topmost branch here represents the nucleotide excision repair (NER) pathway, which has two branches transcription-coupled NER (TC-NER), and global genomic NER (GG-NER). We observe that most regulated proteins are active in TC-NER, which has two child nodes containing two pathways each. Again, only one branch contains the regulated proteins, which in this case is *node 61* represents the set of 32 proteins which are involved in pathways "formation of TC-NER repair complex" and "dual incision reaction in TC-NER". The other branch of TC-NER, *node 224*, on the other hand contains 16 proteins that are involved in gap-filling DNA-repair synthesis and ligation in both TC-NER and GG-NER. Out of the 16 proteins in this node, only one (Replication factor C subunit 4) was found in the expression data, which was not regulated significantly enough to fall in the ROI.

score with both models. RNA polymerase I (pol I) transcription activity is also highlighted, under the same superpathway, albeit with less significant score. Moreover, there is also very clear signal from the *formation and maturation of mRNA transcript* pathway, which encompasses pathways *mRNA splicing* and *transport of the mature transcript into the cytoplasm*, further down in the pathway hierarchy, see Figure 4. The FEvER tool thus enables pinpointing the relevant subpathways in the case of overlapping pathways.

Interestingly the two score models have a clear disagreement regarding the *Translation* superpathway. The nonparametric model rated the differential expression of the proteins involved translation pathway to be of utmost significance ($p < 1 \times 10^{-6}$), while the parametric model did not yield a particularly favorable score for this pathway ($p < 0.5311$), or any of its the subpathways. This contrast between how these two different models rate this major pathway is due partly to the different statistical tests used and partly to the fact that the nonparametric score does not use the differential expression ratios as effectively as the parametric model. In simplest terms, the statistical tests used in overrepresentation analyses typically measure the probability of observing N, or more, elements of a set by pure chance. In the context of expression data, conventional enrichment analysis methods evaluate to what extent the elements of a functional set, that is, a pathway, are observed in the experimental data and whether or not they are clustered together based on the expression values. This approach has shortcomings, however, as it does not use the available expression values to full extent and do not consider the variability of the abundance values or ratios. This drawback shows itself in the case of large pathways with many identified proteins, which have modest to marginal expression values such as the *Translation* superpathway in this data set. Out of the 78 proteins in this pathway none were within the ROI, which was set to include proteins that were regulated 2 folds with p-value less than 0.05. Considering that 309 proteins (7.8% of total) were within the ROI in this analysis, we believe that the chosen boundary for the ROI is not too strict for this data set. The parametric model we introduce, on the other hand, does a good job in utilizing the relevant experimental information in the evaluation and give a better assessment of potential systemic regulation.

Another major pathway, which shows very specific signal, is *Cell-cell communication* pathway, under which the signal is very specifically originating from *Cell junction organization* pathway. This pathway has two branches that got very high scores, *Cell-cell junction organization* and *Cell-extracellular matrix interactions* pathways. This particular finding shows good correlation with the preliminary statements by Geiger and colleagues. Additionally, multiple facets of the metabolism are also highlighted; several pathways under the *metabolism of lipids and lipoproteins* superpathway showed considerable scores in the analysis, like the superpathways covering the metabolism nucleotides and metabolism of amino acids and derivatives. The parametric model highlights several other small and specific metabolic pathways, such as *cholesterol biosynthesis*, *creatine-phosphate biosynthesis*, and *glutamine biosynthesis*. Previous studies indicate that glutamine metabolism may play a significant role in cancer, as glutamine deficiency has shown to lead to MYC-dependent apoptosis.[41]

Some of the other highlighted pathways include *Signaling by GPCR* under *Signaling transduction*, and *G2/M checkpoints* under *Cell cycle checkpoints* major pathways, as well as the *Apoptotic cleavage of cellular proteins* pathway all of which are flagged primarily by the parametric model as likely candidates of regulated pathways. *Transcription-coupled nucleotide excision repair* pathway, on the other hand, is rated significant below $p < 0.01$ by both methods, see Figure 5. In summary, the FEvER method successfully finds pathways that were found by the other analysis tools but adds on another layer of resolution by revealing specific subpathways that are the likely targets of systemic regulation.

Comparison with Existing Tools

To assess the utility of our method compared to some of the existing methods we analyzed the ductal/lobular data set with two established tools for pathway analysis the pathway analysis tools by Reactome and Ingenuity Pathway Analysis. The results from Reactome pathway analysis can be directly compared to the FEvER results, as the Reactome database is one of the sources for Pathway Commons, and thus all pathways from Reactome will be included in FEvER analyses. Since the underlying canonical pathway data is at least partially the same, the end results primarily reflect the differences between methods. A similarly direct comparison of results is not possible with IPA, however, as the IPA knowledgebase is not publicly available.

Investigating pathway level changes in a data set using Reactome pathway analysis can be done in multiple ways, as described in a recent publication.[17] Following this tutorial, we performed an over-representation analysis, as well as an expression analysis. The over-representation analysis tool uses a hypergeometric test statistic and reports unadjusted probabilities. The overrepresentation analysis highlights major pathways such as *metabolism* (3.41×10^{-31}); *gene expression* (5.72×10^{-20}), in particular the branch containing the *translation* pathway (3.01×10^{-23}) and its subpathways, and *metabolism of proteins* (3.13×10^{-20}), in particular *metabolism of mRNA* (6.37×10^{-22}). This clear domination of superpathways, such as metabolism and translation, is based on the fact that the data set features a large number of, what can be called, housekeeping proteins that are marginally regulated (within 1–1.5× fold change). On the lower end of the list, in other words, among the least overrepresented pathways, are the *Transcription* superpathway (0.0214), *DNA repair* (0.19) and *cell–cell communication* pathways (0.41), see Supporting Information S2 and S3. The expression analysis run returns a sortable table that lists all pathways represented in the data set, expression data can then be overlaid on visualized pathway graphs.

IPA platform offers a number of different types of analyses for different types of data. In our case, we used version 9.0 of IPA and ran a "core analysis", which highlight top biological functions and most regulated canonical pathways, and many other similar ranking lists. The results highlight many findings, one of which is the top scoring molecular and cellular functions. *RNA post-translational modification and protein synthesis* top this list, followed by *cellular assembly and organization*, *cellular function and maintenance*, and *cellular growth and proliferation*. The top canonical pathways, on the other hand, are given in Table 3. Furthermore, IPA highlights several networks based on the identified entities in the data. Among the top scoring networks in IPA analysis are *RNA post-translational modification, protein synthesis* and *cellular assembly* highlighted as the top three. A full set of findings from IPA is available in Supporting Information S4–S7.

Table 3. Top Canonical Pathways from Ingenuity Pathway Analysis (IPA)

IPA – top canonical pathways	*p*-value	ratio
EIF2 signaling	8.89×10^{-31}	89/222 (0.401)
Protein ubiquitination pathway	$3.00 \times 10^{-2}0$	86/274 (0.314)
Regulation of eIF4 and p70S6K signaling	2.07×10^{-17}	58/179 (0.324)
Valine, Leucine and Isoleucine degradation	2.05×10^{-13}	32/108 (0.296)
Clathrin-mediated endocytosis signaling	1.62×10^{-12}	52/172 (0.302)

The enrichment model used by Reactome pathway analysis tool is very similar to the nonparametric model featured in FEvER. Similar to what we have observed with the nonparametric model, the hypergeometric over-representation analysis is biased toward larger pathways that are represented in the data set with many proteins, however not necessarily with high fold change or reliable *p*-values. Moreover, the parametric model used in the FEvER method allows detection of pathways that would otherwise be missed with conventional enrichment models used by other methods. IPA results, on the other hand, are not directly comparable to FEvER results, since IPA splits the pathways into different domains. However, the top ranking pathways pointed out by IPA, as described above, are in agreement with those that both score models in FEvER identifies as significant. In contrast, the two score models in FEvER yields more sensitive analysis, and in combination with the hierarchical pathway view allows for fast pinpointing of the most relevant pathways.

Scalability

With the analysis of the first two data sets yielding promising results, we decided to move on to a secondary stress test, this time checking the robustness and flexibility of the method and the software by exposing it to a more complete expression data set. While proteomics experiments today typically results in partial proteome coverage, microarray experiments allow for almost complete analysis of the transcriptome. We foresee that the coverage of proteomics experiments will continue to increase, and therefore chose to analyze a transcriptomics data set using the tool to test how it rates in terms of scalability. The data set is from a transcriptome comparison of a myeloid human cell line resembling dendritic cells (DCs) exposed to a sensitizing chemical, DNCB, compared to cells exposed to nonsensitizing chemicals.

FEvER analysis revealed very significant, large-scale expression changes in a series of superpathways, among those most clearly within the cell cycle. Immune system and signal transduction also showed high degree changes. With the help of the pathway hierarchy overview feature of the FEvER software, we investigated whether or not the changes in these major pathways were concentrated on particular branches. In the case of mitotic cell cycle, we observe that the entire hierarchy shows very significance scores, with the exception of the G2 phase, see Figure 6. Likewise, cell cycle checkpoint pathways also show high degrees of regulation, in particular the *G2/M checkpoints* pathway. *DNA repair* superpathway, too, has very high scores, specifically double-strand break repair, and nucleotide excision repair pathways appear to be regulated entirely. A closer look reveals the fact that approximately 37% of the analytes in *Double-strand break repair* pathway, defined by Reactome pathway database, were down-regulated within the chosen ROI, indicating major regulation of this particular pathway.

Our results correlate well with previous findings on pathway-level effects of skin sensitizers. Using the DAVID tool,[42] Ku and colleagues have recently highlighted a number of pathways such as DNA replication, cell cycle regulation and pyrimidine metabolism in skin-draining lymph nodes using the local lymph node assay (LLNA).[43] These pathways were indeed highlighted with very high significance scores from both models; *DNA replication* ($p < 1 \times 10^{-6}$), *cell cycle checkpoints* ($p < 1 \times 10^{-6}$), *pyrimidine metabolism* ($p < 0.001$). Additionally *purine metabolism* ($p < 0.001$), particularly *purine ribonucleoside monophosphate biosynthesis* ($p < 1 \times 10^{-5}$) and *telomere maintenance* ($p < 1 \times 10^{-6}$) were highlighted by both score models. While the exact mechanisms of telomerase activity in human skin is still unknown, previous studies have linked differential telomerase activity to different types of stress conditions.[44,45]

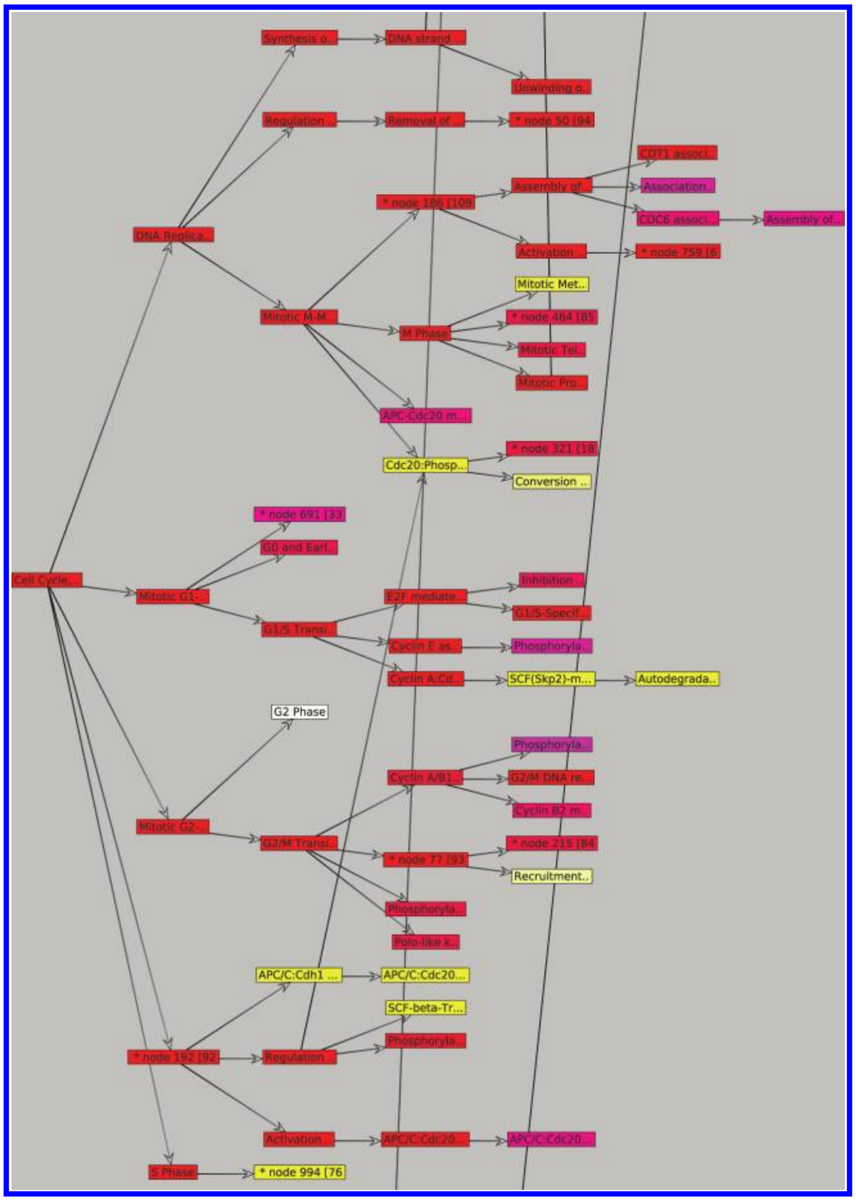


Figure 6. Mitotic cell cycle pathway hierarchy from the results of the DNCB-stimulation data set. A very strong and widespread effect of the sensitizing agent is observed on this superpathway. We can note that the only child nodes in this hierarchy that do not show significant scores with either model are the nodes containing "*G2 phase*" pathway and the "*amplification of signal from kinetochores*" pathway, which is hierarchically under the "mitotic spindle checkpoint" pathway.

Our results using the FEvER method not only support the findings from Ku et al, but also provide additional information utilizing the hierarchical structure of pathways. Besides the hierarchical overview, FEvER method also highlights additional pathways as potentially interesting targets for future studies. One such example is the *axon guidance* pathway, which is highlighted by both score models. On the first look this pathway appear irrelevant following stimulation with a sensitizing chemical. Further investigation using the hierarchy tool however reveals that the core of the regulation that project upward in the hierarchy originates from SEMA4D in *semaphorin signaling interactions*. SEMA4D (also known as CD100) has recently been associated with multiple roles within the immune system,[46,47] one of which is to participate in expression regulation of cell surface proteins such as CD80 and CD86 on DCs.[46,48]

Score Model Evaluation

Our results from the analysis of these three different data sets suggest that the parametric model is more specific in highlighting systemic changes, while being more sensitive to variations in the input data. The nonparametric model, in contrast, utilizes over-representation analysis based on combinatorics, which clearly prefers larger pathways with many identified proteins with much less regard for the expression values, compared to the parametric model. In terms of flexibility, the use of parameters allow for better support for different types of data, even though the default values of the parameters gave reliable results for the tested data sets. Additionally, the parametric score model is very modular and thus could be extended to include additional variables in the analysis, such as the PTM information, once this level of detail becomes widely integrated in canonical pathways.

The usage of parallel score models allows for both robust and sensitive analysis. The consensus of different models indicate significant changes that are valid regardless of specific assumptions. Contradicting scores from the models on the other hand, typically indicate either high levels of differential expression albeit not widespread over a pathway (high parametric score), or widespread low-level differential expression over a pathway (high nonparametric score). The latter case could be a symptom of a too-strict ROI for the data set, but could also indicate a systematic bias between the samples compared, such as unequal amounts of analytes. While two separate scores per pathway might be confusing at first, it allows for more in-depth investigation of results. Nevertheless, a meta-score in the range of (0,100] is implemented for initial ranking of the reported pathways and provides a convenient measure of the consensus of the score models (Supporting Information S1).

To investigate whether or not the score models are robust against perturbations in the input data, in the form of noise or missing values, we conducted a series of runs on a test data set with and without additional perturbations. Similarly we investigated whether or not the choice of ROI parameters introduce any additional artifacts to the analysis (Supporting Information S9). These simulations suggest that the models are robust against reasonable levels of variability in the underlying data, that is, small perturbations in the input data do not cause any large-scale changes in the outcome of the analysis. The ROI parameters display a more sophisticated effect on the outcome of the analysis. By increasing the threshold limits, in particular the fold change threshold, one can decrease the number of proteins in the data set that fall within the ROI. This will then lead for a more stringent evaluation of the pathways, yielding higher specificity at the cost of sensitivity. Consequently, looser thresholds will result in a broader ROI, which will have the opposite effect.

Another subject worth noting is the mock data sampling for the significance calculation of the parametric enrichment score. In our tests, we have noticed that the optimal method for generating mock data depends on the size of the input data set. The empirical distribution method is generally preferable to the other methods, since it does not make any assumptions for the distribution of the underlying data. However, there is a risk of under-sampling with this method during analyses of data sets with few analytes. In that case using permutation of expression values for randomization may be more accurate and this option is a feature already implemented in our software.

■ CONCLUSIONS

In this study, we have introduced a novel workflow that utilizes multiple models for pathway enrichment evaluation, and a publicly available software implementation of this workflow. The software not only implements the FEvER method workflow but also tackles a number of the well-known challenges in the field, such as multiple identification problem, pathway redundancies and pathway annotation querying, as well as hierarchical overview of pathways. We have also showed the utility of the method as well as the software implementation by analyzing three data sets with different properties.

Much like all other methods used within the field, the quality and reliability of results acquired from pathway enrichment analysis depends on the underlying data set. Furthermore, we believe that it is unrealistic to expect reliable results from a fully automated, single-button method that returns a definitive *yes/no* answer regarding pathways. We believe that for the most useful results, the users do indeed need to take part in the analysis, to fine-tune the variables to suit the underlying data in the best possible way. Despite the practical and conceptual challenges, we do envision a more unified analysis platform to be the future of functional proteomics. The FEvER method is developed to evaluate pathways using multiple scoring models in parallel and could be extended to cover more than two models. This design allows more robust conclusions to be inferred from the analysis, as the consensus of high-scores from different enrichment models indicates a spot of high significance without modeling bias. On the other hand, sensitive analysis can be achieved by targeting pathways that are highlighted by one specific model.

We believe both the method and software implementation, could serve the proteomics community by facilitating efficient and thorough analysis of functional evaluation of expression regulation. Implementation of multiple score models allow for in-depth pathway analysis, particularly with the parametric model enabling pinpointing of regulated pathways that are missed by established methods.

■ ASSOCIATED CONTENT

Supporting Information

Supplemental materials as described in text. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: Fredrik.Levander@immun.lth.se.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank ChiBE developers at BCBI, Bilket University, and staff at Pathway Commons for their support during the development of the software. Markus Ringnér, Erik Fredlund, Peter James and Robert Trevino are thanked for discussions regarding pathway resources and for sharing code for accessing online pathway resources (RT). Niclas Olsson, Peter James, Carl Borrebaeck, and Christer Wingren are thanked for giving prepublication access to the yeast data set. The work was funded by the Swedish Research Council through grant 2007-5188, BILS (Bioinformatics Infrastructure in Life Sciences), and the Swedish Foundation for Strategic Research (RBb08-0006).

■ ABBREVIATIONS

FEvER, Functional Evaluation of Expression regulation; GSEA, gene set enrichment analysis; PSEA, protein set enrichment analysis; KEGG, Kyoto Encyclopedia of Genes and Genomes; IPA, Ingenuity Pathway Analysis; GO, Gene Ontology; ROI, region of interest; SBML, systems biology mark-up language; BioPAX, Biological PAthway eXchange; ChiBE, Chisio BioPAX Editor; DAVID, Database for Annotation, Visualization and Integrated Discovery; PTM, post-translational modification; DC, dendritic cells; LLNA, local lymph node assay

■ REFERENCES

(1) Bousette, N.; Kislinger, T.; Fong, V.; Isserlin, R.; Hewel, J.; Emili, A.; Gramolini, A. Large-scale characterization and analysis of the murine cardiac proteome. *J. Proteome Res.* **2009**, *8* (4), 1887−1901.

(2) de Godoy, L. M. F.; Olsen, J. V.; Cox, J.; Nielsen, M. L.; Hubner, N. C.; Fröhlich, F.; Walther, T. C.; Mann, M. Comprehensive mass-

spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* **2008**, *455*, 1251−1254.

(3) Geiger, T.; Cox, J.; Mann, M. Proteomic Changes Resulting from Gene Copy Number Variations in Cancer Cells. *PLoS Genet.* **2010**, *6*, e1001090.

(4) Malmström, J.; Beck, M.; Schmidt, A.; Lange, V.; Deutsch, E. W.; Aebersold, R. Proteome-wide cellular protein concentrations of the human pathogen Leptospira interrogans. *Nature* **2009**, *460*, 762−765.

(5) Subramanian, A.; Tamayo, P.; Mootha, V. K.; Mukherjee, S.; Ebert, B. L.; Gillette, M. A.; Paulovich, A.; Pomeroy, S. L.; Golub, T. R.; Lander, E. S.; Mesirov, J. P. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 15545−15550.

(6) Picotti, P.; Aebersold, R.; Domon, B. The implications of proteolytic background for shotgun proteomics. *Mol. Cell. Proteomics* **2007**, *6*, 1589−1598.

(7) Wolf-Yadlin, A.; Hautaniemi, S.; Lauffenburger, D. A.; White, F. M. Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 5860−5865.

(8) Picotti, P.; Bodenmiller, B.; Mueller, L. N.; Domon, B.; Aebersold, R. Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. *Cell* **2009**, *138*, 795−806.

(9) Gstaiger, M.; Aebersold, R. Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nat. Rev. Genet.* **2009**, *10*, 617−627.

(10) Cha, S.; Imielinski, M. B.; Rejtar, T.; Richardson, E. A.; Thakur, D.; Sgroi, D. C.; Karger, B. L. In situ proteomic analysis of human breast cancer epithelial cells using laser capture microdissection: annotation by protein set enrichment analysis and gene ontology. *Mol. Cell. Proteomics* **2010**, *9*, 2529−2544.

(11) Nam, D.; Kim, S.-Y. Gene-set approach for expression pattern analysis. *Brief. Bioinform.* **2008**, *9*, 189−197.

(12) Gehlenborg, N.; Yan, W.; Lee, I. Y.; Yoo, H.; Nieselt, K.; Hwang, D.; Aebersold, R.; Hood, L. Prequips--an extensible software platform for integration, visualization and analysis of LC-MS/MS proteomics data. *Bioinformatics* **2009**, *25*, 682−683.

(13) Askenazi, M.; Li, S.; Singh, S.; Marto, J. A. Pathway Palette: A rich internet application for peptide-, protein- and network-oriented analysis of MS data. *Proteomics* **2010**, *10*, 1880−1885.

(14) Schramm, G.; Wiesberg, S.; Diessl, N.; Kranz, A.-L.; Sagulenko, V.; Oswald, M.; Reinelt, G.; Westermann, F.; Eils, R.; König, R. PathWave: discovering patterns of differentially regulated enzymes in metabolic pathways. *Bioinformatics* **2010**, *26*, 1225−1231.

(15) Zubarev, R. A.; Nielsen, M. L.; Fung, E. M.; Savitski, M. M.; Kel-Margoulis, O.; Wingender, E.; Kel, A. Identification of dominant signaling pathways from proteomics expression data. *J. Proteomics* **2008**, *71*, 89−96.

(16) Rigbolt, K. T. G.; Vanselow, J. T.; Blagoev, B. GProX, a user-friendly platform for bioinformatics analysis and visualization of quantitative proteomics data. *Mol. Cell. Proteomics* **2011**, *10*, No. O110.007450.

(17) Haw, R.; Hermjakob, H.; D'Eustachio, P.; Stein, L. Reactome pathway analysis to enrich biological discovery in proteomics data sets. *Proteomics* **2011**, *11*, 3598−3613.

(18) Ingenuity Pathway Analysis. Ingenuity Systems Inc., USA. http://www.ingenuity.com/products/pathways_analysis.html.

(19) MetaCore. GeneGo Inc., USA. http://www.genego.com/metacore.php.

(20) Mootha, V. K.; Lindgren, C. M.; Eriksson, K.-F.; Subramanian, A.; Sihag, S.; Lehar, J.; Puigserver, P.; Carlsson, E.; Ridderstråle, 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.; Groop, L. C. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **2003**, *34*, 267−273.

(21) Goeman, J. J.; Buhlmann, P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* **2007**, *23*, 980−987.

(22) Jiang, Z.; Gentleman, R. Extensions to gene set enrichment. *Bioinformatics* **2007**, *23*, 306−313.

(23) Fridley, B. L.; Jenkins, G. D.; Biernacka, J. M. Self-Contained Gene-Set Analysis of Expression Data: An Evaluation of Existing and Novel Methods. *PLoS ONE* **2010**, *5*, e12693.

(24) Emmert-Streib, F.; Glazko, G. V. Pathway Analysis of Expression Data: Deciphering Functional Building Blocks of Complex Diseases. *PLOS Comput. Biol.* **2011**, *7*, e1002053.

(25) Keller, A.; Backes, C.; Lenhof, H.-P. Computation of significance scores of unweighted Gene Set Enrichment Analyses. *BMC Bioinform.* **2007**, *8*, 290.

(26) Backes, C.; Keller, A.; Kuentzer, J.; Kneissl, B.; Comtesse, N.; Elnakady, Y. A.; Müller, R.; Meese, E.; Lenhof, H.-P. GeneTrail--advanced gene set enrichment analysis. *Nucleic Acids Res.* **2007**, *35*, W186−W192.

(27) Babur, Ö.; Dogrusoz, U.; Demir, E.; Sander, C. ChiBE: interactive visualization and manipulation of BioPAX pathway models. *Bioinformatics* **2010**, *26*, 429−431.

(28) Hucka, M.; Finney, A.; Sauro, H. M.; Bolouri, H.; Doyle, J. C.; Kitano, H.; Arkin, A. P.; Bornstein, B. J.; Bray, D.; Cornish-Bowden, A.; Cuellar, A. A.; Dronov, S.; Gilles, E. D.; Ginkel, M.; Gor, V.; Goryanin, I. I.; Hedley, W. J.; Hodgman, T. C.; Hofmeyr, J. H.; Hunter, P. J.; Juty, N. S.; Kasberger, J. L.; Kremling, A.; Kummer, U.; Le Novere, N.; Loew, L. M.; Lucio, D.; Mendes, P.; Minch, E.; Mjolsness, E. D.; Nakayama, Y.; Nelson, M. R.; Nielsen, P. F.; Sakurada, T.; Schaff, J. C.; Shapiro, B. E.; Shimizu, T. S.; Spence, H. D.; Stelling, J.; Takahashi, K.; Tomita, M.; Wagner, J.; Wang, J. SBML Forum. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **2003**, *19*, 524−531.

(29) Demir, E.; Cary, M. P.; Paley, S.; Fukuda, K.; Lemer, C.; Vastrik, I.; Wu, G.; D'Eustachio, P.; Schaefer, C.; Luciano, J.; Schacherer, F.; Martinez-Flores, I.; Hu, Z.; Jimenez-Jacinto, V.; Joshi-Tope, G.; Kandasamy, K.; Lopez-Fuentes, A. C.; Mi, H.; Pichler, E.; Rodchenkov, I.; Splendiani, A.; Tkachev, S.; Zucker, J.; Gopinath, G.; Rajasimha, H.; Ramakrishnan, R.; Shah, I.; Syed, M.; Anwar, N.; Babur, Ö.; Blinov, M.; Brauner, E.; Corwin, D.; Donaldson, S.; Gibbons, F.; Goldberg, R.; Hornbeck, P.; Luna, A.; Murray-Rust, P.; Neumann, E.; Reubenacker, O.; Samwald, M.; van Iersel, M.; Wimalaratne, S.; Allen, K.; Braun, B.; Whirl-Carrillo, M.; Cheung, K.-H.; Dahlquist, K.; Finney, A.; Gillespie, M.; Glass, E.; Gong, L.; Haw, R.; Honig, M.; Hubaut, O.; Kane, D.; Krupa, S.; Kutmon, M.; Leonard, J.; Marks, D.; Merberg, D.; Petri, V.; Pico, A.; Ravenscroft, D.; Ren, L.; Shah, N.; Sunshine, M.; Tang, R.; Whaley, R.; Letovksy, S.; Buetow, K. H.; Rzhetsky, A.; Schachter, V.; Sobral, B. S.; Dogrusoz, U.; McWeeney, S.; Aladjem, M.; Birney, E.; Collado-Vides, J.; Goto, S.; Hucka, M.; Novère, N. L.; Maltsev, N.; Pandey, A.; Thomas, P.; Wingender, E.; Karp, P. D.; Sander, C.; Bader, G. D. The BioPAX community standard for pathway data sharing. *Nat. Biotechnol.* **2010**, *28*, 935−942.

(30) Cerami, E. G.; Gross, B. E.; Demir, E.; Rodchenkov, I.; Babur, Ö.; Anwar, N.; Schultz, N.; Bader, G. D.; Sander, C. Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* **2011**, *39*, D685−D690.

(31) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26*, 1367−1372.

(32) Caspi, R.; Altman, T.; Dale, J. M.; Dreher, K.; Fulcher, C. A.; Gilham, F.; Kaipa, P.; Karthikeyan, A. S.; Kothari, A.; Krummenacker, M.; Latendresse, M.; Mueller, L. A.; Paley, S.; Popescu, L.; Pujar, A.; Shearer, A. G.; Zhang, P.; Karp, P. D. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **2010**, *38*, D473−D479.

(33) Ong, S.-E.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Steen, H.; Pandey, A.; Mann, M. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **2002**, *1*, 376−386.

(34) Geiger, T.; Cox, J.; Ostasiewicz, P.; Wisniewski, J. R.; Mann, M. Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nat. Methods* **2010**, *7*, 383−385.

(35) Johansson, H.; Lindstedt, M.; Albrekt, A.-S.; Borrebaeck, C. A. A genomic biomarker signature can predict skin sensitizers using a cell-based in vitro alternative to animal tests. *BMC Genomics* **2011**, *12*, 399.

(36) Kolkman, A.; Olsthoorn, M. M. A.; Heeremans, C. E. M.; Heck, A. J. R.; Slijper, M. Comparative proteome analysis of Saccharomyces cerevisiae grown in chemostat cultures limited for glucose or ethanol. *Mol. Cell. Proteomics* **2005**, *4*, 1−11.

(37) Futcher, B.; Latter, G. I.; Monardo, P.; McLaughlin, C. S.; Garrels, J. I. A sampling of the yeast proteome. *Mol. Cell. Biol.* **1999**, *19*, 7357−7368.

(38) Ciriacy, M. Genetics of alcohol dehydrogenase in Saccharomyces cerevisiae. II. Two loci controlling synthesis of the glucose-repressible ADH II. *Mol. Gen. Genet.* **1975**, *138*, 157−164.

(39) Russell, D. W.; Smith, M.; Williamson, V. M.; Young, E. T. Nucleotide sequence of the yeast alcohol dehydrogenase II gene. *J. Biol. Chem.* **1983**, *258*, 2674−2682.

(40) Wills, C.; Phelps, J. A technique for the isolation of yeast alcohol dehydrogenase mutants with altered substrate specificity. *Arch. Biochem. Biophys.* **1975**, *167*, 627−637.

(41) Yuneva, M.; Zamboni, N.; Oefner, P.; Sachidanandam, R.; Lazebnik, Y. Deficiency in glutamine but not glucose induces MYC-dependent apoptosis in human cells. *J. Cell Biol.* **2007**, *178*, 93−105.

(42) Huang, D. W.; Sherman, B. T.; Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **2009**, *4*, 44−57.

(43) Ku, H.-O.; Jeong, S.-H.; Kang, H.-G.; Son, S.-W.; Yun, S.-M.; Ryu, D.-Y. Pathway analysis of gene expression in local lymph nodes draining skin exposed to three different sensitizers. *J. Appl. Toxicol* **2011**.

(44) Taylor, R. S.; Ramirez, R. D.; Ogoshi, M.; Chaffins, M.; Piatyszek, M. A.; Shay, J. W. Detection of telomerase activity in malignant and nonmalignant skin conditions. *J. Invest. Dermatol.* **1996**, *106*, 759−765.

(45) Buckingham, E. M.; Klingelhutz, A. J. The role of telomeres in the ageing of human skin. *Exp. Dermatol.* **2011**, *20*, 297−302.

(46) Suzuki, K.; Kumanogoh, A.; Kikutani, H. Semaphorins and their receptors in immune cell interactions. *Nat. Immunol.* **2008**, *9*, 17−23.

(47) Nkyimbeng-Takwi, E.; Chapoval, S. P. Biology and function of neuroimmune semaphorins 4A and 4D. Immunol. *Immunol. Res.* **2011**, *50*, 10−21.

(48) Kumanogoh, A.; Suzuki, K.; Ch'ng, E.; Watanabe, C.; Marukawa, S.; Takegahara, N.; Ishida, I.; Sato, T.; Habu, S.; Yoshida, K.; Shi, W.; Kikutani, H. Requirement for the lymphocyte semaphorin, CD100, in the induction of antigen-specific T cells and the maturation of dendritic cells. *J. Immunol.* **2002**, *169*, 1175−1181.