Using Pathway Signatures as Means of Identifying Similarities among Microarray Experiments

Luca Beltrame[1,2], **Lisa Rizzetto**[1], **Raffaele Paola**[1], **Philippe Rocca-Serra**[3], **Luca Gambineri**[4], **Cristina Battaglia**[5], **Duccio Cavalieri**[1]*

1 Department of Pharmacology, University of Firenze, Firenze, Italy, **2** Institute for Biomedical Technologies, National Research Council, Milano, Italy, **3** European Bioinformatics Institute, Cambridge, United Kingdom, **4** Inspect.it, Arezzo, Italy, **5** Department of Science and Biomedical Technologies, University of Milano, Milano, Italy

Abstract

Widespread use of microarrays has generated large amounts of data, the interrogation of the public microarray repositories, identifying similarities between microarray experiments is now one of the major challenges. Approaches using defined group of genes, such as pathways and cellular networks (pathway analysis), have been proposed to improve the interpretation of microarray experiments. We propose a novel method to compare microarray experiments at the pathway level, this method consists of two steps: first, generate pathway signatures, a set of descriptors recapitulating the biologically meaningful pathways related to some clinical/biological variable of interest, second, use these signatures to interrogate microarray databases. We demonstrate that our approach provides more reliable results than with gene-based approaches. While gene-based approaches tend to suffer from bias generated by the analytical procedures employed, our pathway based method successfully groups together similar samples, independently of the experimental design. The results presented are potentially of great interest to improve the ability to query and compare experiments in public repositories of microarray data. As a matter of fact, this method can be used to retrieve data from public microarray databases and perform comparisons at the pathway level.

Citation: Beltrame L, Rizzetto L, Paola R, Rocca-Serra P, Gambineri L, et al. (2009) Using Pathway Signatures as Means of Identifying Similarities among Microarray Experiments. PLoS ONE 4(1): e4128. doi:10.1371/journal.pone.0004128

Editor: Dawn N. Albertson, Minnesota State University Mankato, United States of America

Received August 7, 2008; **Accepted** December 4, 2008; **Published** January 6, 2009

Funding: This work was supported by grants UE LSHB-CT-2004-512074, NUGO FOOD- CT-2004-506360 and MIUR PRIN 519MIUR060. LB was also supported by grant MIUR FIRB RBIP064CRT. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: duccio.cavalieri@unifi.it

Introduction

Since their first inception a decade ago, microarray studies have become widely used in the research community, thanks to their ability to assess the expression of thousands of genes in a single laboratory event. The belief that such wealth of genomic information the community could not afford to lose has led to the development of microarray standards [1,2] and databases including two major public microarray repositories, Gene Expression Omnibus (GEO) [3] and ArrayExpress [4], in the hope of enabling mining and exploration of newly acquired data space.

Identifying biologically meaningful information in relatively noisy data represents a significant tasks and so far breakthrough have been few and far between. Comparisons made on the level of gene lists obtained by different statistical methods or from different datasets hardly converge [5]. As a consequence, the usefulness of the vast amounts of data stored in public repositories is subject to debate. At the same time, it is becoming important to use more than a single data set when analyzing microarray data [6] and gather hundreds or thousands of samples to develop prognostic markers [7]. Reaching such a goal is difficult when information obtained from different experiments do not overlap. This is mainly because the data often has been generated with different microarray platforms, hybridization protocols, and the authors use different methods and different thresholds to calculate differentially expressed genes (DEGs) [8] . Finding ways to reliably compare different microarray data sets is therefore important to obtain biologically sound and reproducible information from different datasets.

In the past years collections of all the differentially expressed genes in a given condition that exclusively characterize that condition, have been proposed as “gene signatures” for a condition [9–11]. However, the reliability and reproducibility of such signatures has been questioned [12,13], because of the influence of the statistical assumptions used, or errors in the methodology.

The number of inconsistencies and discrepancies when microarray data sets are compared are often reduced when approaches that take into account biologically related sets of genes, rather than single entities, are used [14]. The first methods developed with this approach aimed at identifying significantly under- or over-represented terms in the Gene Ontology [15,16]. A second approach instead focuses in identifying significantly expressed gene sets (sometimes incorrectly referred as pathways or cellular networks) in a given condition using different statistical measurements: Z-score [17], Gene Set Enrichment Analysis [18], signed Fisher Exact Test [19], Global test [20] and impact analysis [21].

Recently we developed a bioinformatic environment called Eu.Gene [22] containing a repository of all the freely available biological pathways and different statistical methods dedicated to analyze expression datasets and assess for enrichment in biological pathways. EuGene relies on 2 components, (i) a database of consistently annotated pathways collected from a number of state of the art pathway resources and (ii) a module of computation implementing an array of tailored heuristics and statistical methods.

We developed a method allowing to assess similarity of samples in microarray databases. To this aim we used EuGene to generate "Pathway Signatures", recapitulating the biologically meaningful pathways related to some clinical/biological variable of interest, and used them in an analysis workflow to compare different microarray experiments.

Results

Selection of the data sets

We decided to use as reference experiments two distinct data sets: firstly, a data set of gene expression experiments on a model organism, *Saccharomyces cerevisiae* (*S. cerevisiae*) using the dataset from the Rosetta compendium of deletion mutants [23], and secondly 106 experiments (grouped in 7 sets) focused on the exposure of *Homo sapiens* dendritic cells (DCs) to different stimuli.

Following up the data set collection, data were pre-processed and subject to pathway analysis (see Methods). Then, we formulated two predictions basing on biological evidence against which the method would be tested. With regards to the yeast deletion mutants, we predicted that the *ssn6* and *tup1* mutants would be similar, as acting as part of the same protein complex. Regarding the dendritic cells data set, we performed a wet lab experiments stimulating DCs with cells of the yeast *S. cerevisiae* in exponential growth phase (analyzed in the same manner as the public data) and we predicted that this experiment would show similarity with the samples from GEO data set GSE6965 (gene expression profile of monocyte-derived DCs stimulated for 6 h with germinating germ tubes of *Aspergillus fumigatus*).

Statistical methods

We initially compared different statistical approaches for pathway signature generation. To address the issue of the most appropriate statistic, we evaluated both Fisher's Exact Test (FET) and Gene Set Enrichment Analysis (GSEA), two well-established methods for pathway analysis. As GSEA required samples to be divided into two distinct classes, it was not performed on the yeast data. On the other hand both GSEA and FET were used with the DC samples. FET produced results in the form of signed p-values, while GSEA produced signed enrichment scores (ES).

Yeast data was in the form of $\log_{10}$ ratios and thus, after a conversion to $\log_2$, was used directly for the Fisher's Exact Test.

In order to use dendritic cell data for the Fisher's Exact Test, we converted the Affymetrix absolute expression values into ratios (see Methods). When dealing with Affymetrix data, we had the choice of cross-normalizing all the data sets, or to perform normalization independently for each of them. Initially we compared the two approaches. We calculated the ratios from either the comparison between the treated samples against a common baseline (median expression of all control experiments) or by doing paired ratios between each treated sample and its corresponding control. Between those two approaches, the paired ratios performed better (data not shown) and was therefore used in all subsequent analyses both to prevent a "smoothing" effect of the differences between the data sets themselves, and to make sure that we could add more samples without having to normalize the data sets again.

In order to selecting the most representative FET threshold, FET thresholds were selected from the 2σ interval of the binomial distribution of the expression values.

When using FET, we addressed the question of how to use the collection of p-values for a defined set of pathways to generate a pathway signature with the aim of comparing different microarray experiments. Firstly, we corrected the p-values for multiple testing (see Methods) then our first attempt was to transform the p-values into a measure whose magnitude would express the degree of significance, and to that purpose, we transformed p-values into Pathway Enrichment Factors (PEFs). Secondly, we transformed p-values in signed binary enrichment factors (sBEFs) (see Methods): these factors categorized p-values into three classes (significantly up-regulated, significantly down-regulated, and not significant).

On the other hand, GSEA enrichment scores (ES) were used directly, without any modification.

We then evaluated the performance of these metrics (sBEF or enrichment score), as expressed in the ability of assessing similarity according to the expectation that biologically similar experiments should show identical or similar pathway profiles (defined as collections of PEFs, sBEFs or enrichment scores for the pathways used for testing).

Selection of methods to compare sets of p-values and/or enrichment scores

To find reliable ways to group samples together, we investigated a number of unsupervised hierarchical clustering techniques that would be used to group together the metrics we had selected. For this purpose, Unweighted Pair Group Method with Arithmetic mean (UPGMA) clustering, standard hierarchical clustering, and hierarchical clustering with support trees were used (Figures S1, S2, S3, S4, S5). In the last case we bootstrapped samples and pathways over 100 and 1000 iterations. Pathways and samples were clustered with different metrics: in particular, Euclidean distance, Pearson's correlation and Manhattan distance were used. We ranked the clustering methods based on the ability to group biologically similar samples together. We found hierarchical clustering with support trees to be the most reliable method, and that 100 iterations were sufficient to obtain reproducible results. Although all the distance metrics performed equally in grouping biologically similar samples, we found Euclidean distance to give the best visualization of the data.

Additionally, as sBEFs were categorical values, we also evaluated the similarity between samples by using a metric suited for this type of data (the Jaccard index).

Validation of the method

With regards to yeast data, the clustering of PEFs gave results that were agreeing with our prediction, as the *ssn6* and *tup1* deletion mutants clustered together when using Pearson's correlation as clustering metric (Figure 1). The same results were observed when using sBEFs, with the difference that they were independent from the clustering metric used (Figure 2; Figures S1, S2, S3). Pathways which shared a common profile among the two samples included up-regulated alternative glucose metabolism related pathways (activation of penthose-phosphate pathway, galactose metabolism and degradation, lactose metabolism) and down-regulated glucose metabolism pathways such as transcriptional activation of glucose metabolism genes. Standard hierarchical clustering and UPGMA clustering results were also in agreement with the prediction (Figures S4 and S5). On the other hand, the Jaccard index between the two samples was 0.14, a possible indication that such a method is less powerful than clustering to detect biologically relevant similarity.

The results obtained with the dendritic cells data also confirmed the predictions. wet lab experiments (analyzed in the same manner as the public data) and we predicted that this experiment would show similarity with the samples from GEO data set GSE6965. The Pathway Signatures of our ad-hoc sample produced

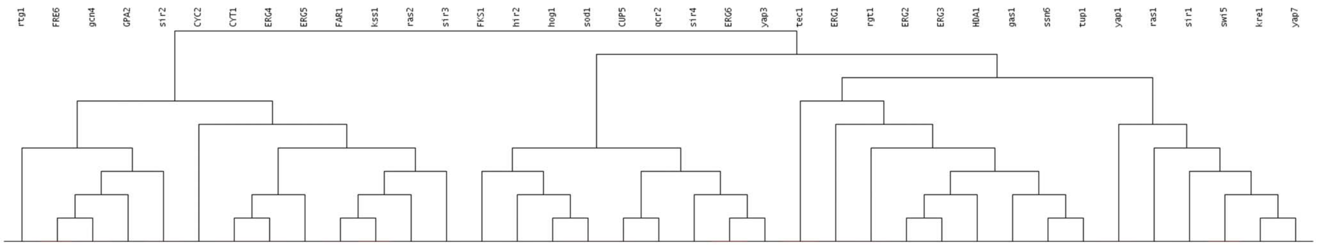

Figure 1. Clustering of yeast samples with PEFs using Pearson's correlation. Sample tree originated from the clustering of FET Pathway Enrichment Factors (PEFs) values with Pearson's correlation.
doi:10.1371/journal.pone.0004128.g001

stimulating DCs with cells of the yeast *S. cerevisiae* in exponential growth phase, clustered together with gene expression profile of monocyte-derived DCs stimulated for 6 h with germinating germ tubes of *A. fumigatus* (Figure 3; Figures S6, S7). The result was even more striking than the yeast data, both sBEFs and PEFs produced results that were independent from the clustering metric used. When using PEFs, the two samples belonged to the same sub-cluster (data not shown), while with sBEFs they belonged to the same sub-cluster and were the most closely associated. Pathways which shared a similar profile among the two samples included the Toll-like receptor signaling pathway, JAK-STAT signaling, and cytokine-cytokine receptor interaction (Figure 4). Since we expect the two stimuli to elicit signaling from the same pathogen associated molecular patterns (PAMPs) these results show that pathway-level signatures identify samples with similar biological characteristics. Standard hierarchical clustering and UPGMA clustering results were also in agreement with our prediction (Figures S8 and S9). The Jaccard index between our experiment and sample GSM160356 showed a good similarity (0.75) and the other replicate from data set GSE6965 showed even greater similarity (0.9). These findings show that FET pathway based metrics are a powerful tool to identify similarity between experiments.

On the other hand, when we clustered enrichment scores from GSEA on the dendritic cell data, results were not in agreement with the prediction (Figure 5): our experiment clustered with the GSE8658 data set (response to different ligands), while GSE6965 clustered with GSE7247 (an experiment on loading DCs with different antigens). These findings indicate that FET transformed p-values are more effective in finding similarities among different experiments when compared to GSEA enrichment scores.

Discussion

In this work we propose a novel method to compare microarray experiments at the pathway level, our work addressed three major tasks: (i) defined a metric measuring the probability of a set of pathways to be related to some clinical/biological variable of interest and the relative importance of that set in the context of the biological problem, (Pathway Signature) (ii) proposed a method using these pathway signatures to assess relative similarity of the experiments ; (iii) proved the validity of this method applying it to two different well defined biological problems on two independent collections of microarray experiments.

As a proof of concept we selected datasets representing well known stimuli in well known biological systems. One of the datasets was from *S. cerevisiae*, the most extensively studied model organism, with a well characterized genome where all the genes are represented on the array, widely used to test bioinformatics methods. We selected the Rosetta compendium of deletion mutants [23] as it measures the steady state response to a very precise alteration, the deletion of a given gene, and consequently is relatively simple to associate a phenotype to a particular pathway profile. As second dataset we decided to use data sets from *Homo sapiens* because the majority of the experiments stored in GEO and ArrayExpress are from human samples. We chose transcriptomic datasets from DCs as they have some clear advantages respect to other fields. Firstly, the cell type is well-defined, which enables the study of the alterations in gene expression following stimulation. Secondly, there is the possibility to perform prediction and hypothesis-driven functional genomics studies aiming at the reconstruction of the networks of molecular interactions characterizing specific DC differentiation programs.

The aim of our work was to propose a procedure to assess similarity between microarray experiments at the pathway level generating "Pathway signatures" (PS) for a set of experiments and use these signatures to interrogate microarray databases.

The need for better methods to identify similarities in microarray data sets arises from the fact that although the analysis techniques have constantly improved over the past years, one of the biggest hurdles remains the comparability among distinct data sets produced by different researchers and laboratories, resulting in lists of genes which do not overlap, or overlap in a very limited fashion. A possible reason relies on the different assumptions on the data used by different statistical methods [8]. This is a strong limitation, because identifying biologically similar samples would increase the power and the reproducibility of a study, especially since in most studies the number of samples can be a limiting factor, which could be compensated using already published experiments.

Also, as the number of publicly available data sets increases (at the time of writing, GEO and Array Express host a total of 221,815 and 110,356 hybridizations, respectively), it is important to have reliable method to compare microarray data from different sources.

The improvements observed when comparing different experiments at the pathway level [14] is coherent with the assumption that genes never act alone in a biological system, but participate in a cascade of networks, an approach overlooked by gene-based analyses.

The selection of the statistical method used to measure enrichment is central to our approach.

Different metrics have been proposed to integrate the probability of alteration of a sector of the cellular network (pathway) and the relative importance of that pathway in the context of the biological problem, such as the probability vector [19], the impact factor (IF) [21]. Other methods have been devised for the identification of regulatory modules and their regulation program by integrating genome-wide location and expression data [24–27]. However, to our knowledge, these methods have not

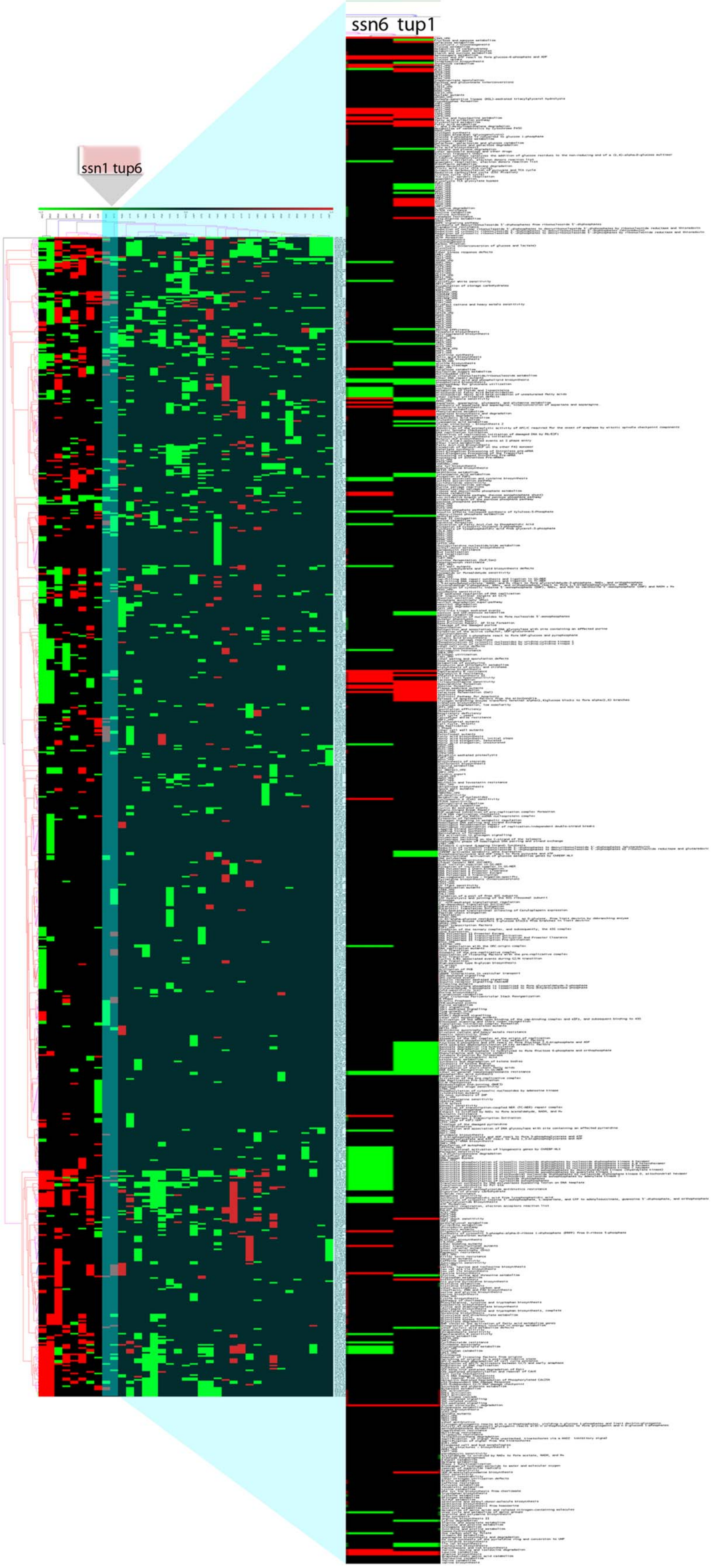

Figure 2. Clustering of signed Binary Enrichment Factors (sBEFs) obtained from the FET analysis on yeast samples. Colored spots indicate significant (p≤0.05) up- (red) or down- (green) regulation. The colors of the dendrogram indicate the percentages of the tree support (significance), from 50% (pink) to 100% (black). The inset shows the clustering of yeast mutants *ssn6* and *tup1*. The full figure is available as Figure S1.
doi:10.1371/journal.pone.0004128.g002

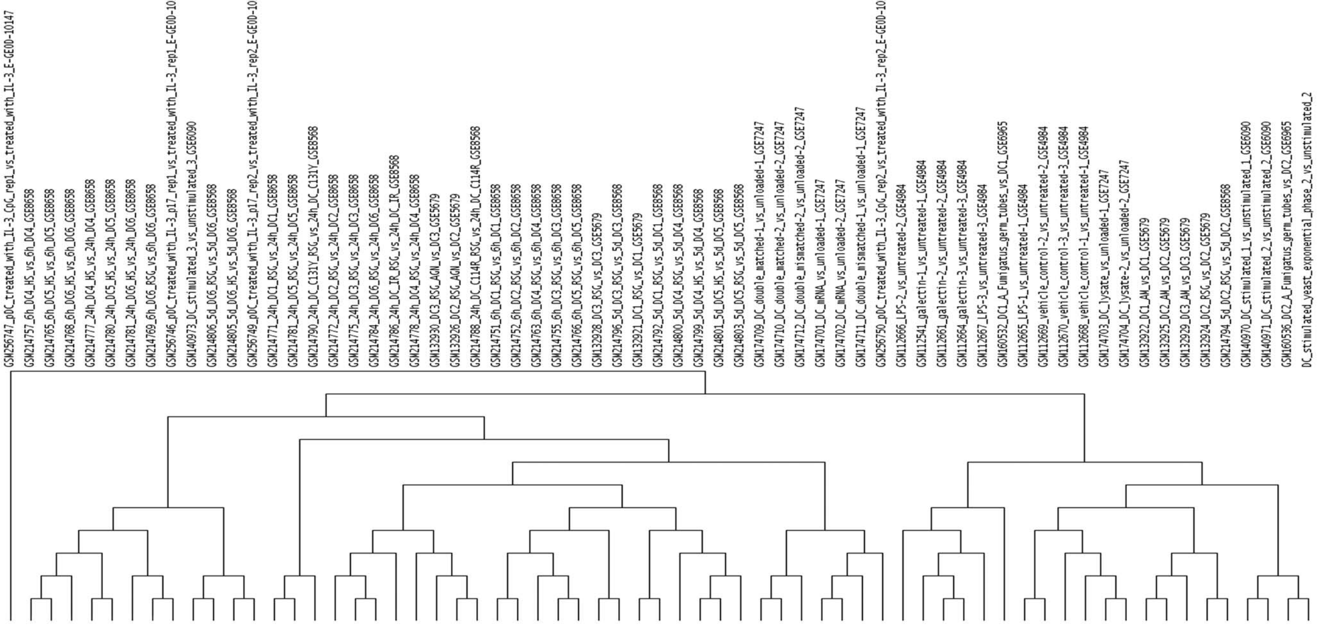

Figure 3. Sample clustering of Fisher's Exact Test samples on dendritic cell data. Sample tree originated from the clustering of FET signed Binary Enrichment Factors (sBEFs) values with Euclidean distance.
doi:10.1371/journal.pone.0004128.g003

been employed to compare a large number of experiments assembled from different microarray data sets.

FET p-values had to be transformed to improve interpretation: p-values express a probability, and the smaller they are, the more significant the result is, while from a conceptual point of view it is better to express pathway enrichment as a number that it is either categorical or the more significant the greater it is. That is the reason we first used the logarithm of reciprocal of the p-value (the Pathway Enrichment Factor), to express the measure in a scale that would avoid interpretation problems, the PEF was finally obtained multiplying the value for a sign representing the "direction" of a pathway, with the same approach described in [19]. Our results showed that clustering PEFs grouped samples according to their biological similarity. In yeast we observed as part of the same cluster transcriptional profiles of the deletion mutants of, *ssn6* and *tup*, genes that form a co repressor complex which is responsible for repressing a large number of *S. cerevisiae* genes, including glucose-responsive genes, DNA damage genes and oxygen utilization genes [28,29]. As *ssn6* and *tup1* act together, deletion of one or the other gene is expected to yield a similar behavior. On the yeast dataset the exact topology of the cluster was dependent on the metric used. This is quite understandable: sets of PEFs can be noisy, as there are all the possible ranges of p-values, and identifying which pathways were "significant" and which "not significant" was not always straightforward. This can influence the clustering, and so certain metrics prove to be more useful than others, for example Pearson's correlation was the metric which most correctly grouped *ssn6* and *tup1* in our yeast data. This result could be biologically relevant, as although ssn6 and tup1 are part of a complex, tup1 has also a function alone, and as a matter of fact, the pattern of significantly altered pathway (shown by their sBEFs) in the two deletion mutants exhibits differences. This biological difference could result in changes in the direction in which some pathways are affected by one deletion or by the other. Alternatively this result could reflect greater sensitivity to technical "noise" in the data of some of the metrics used.

Biclustering of PEFs on the dendritic cell data sets gave concordant and biologically relevant results, the responses elicited by stimulation with *S. cerevisiae* follow the same downstream signalling as the ones in response to the fungus *A. fumigatus* [30–33] their PEFs clustered one next to the other with all the metrics used.

On the other hand, sBEFs produced consistent results from a clustering point of view, both for the yeast and the DCs datasets, independently from the clustering method, thus making very easy the identification of similar datasets.

We selected sBEFs to produce the PS as they can be easily used as a "barcode" that can uniquely attached that sample, facilitating "querying" a database of PSs.

The observation that sBEFs, PEFs and PSs, identify similarity between samples that are biologically meaningful, indicates that the categorical transformation of the p-values does not affect the observed result, concluding that the observed similarity has a biological meaning rather than resulting from a manipulation of the data.

When testing different methods to compare experiments at the pathway level, we initially chose clustering because of its flexibility and robustness in representing biological data. The type of clustering used was biclustering with support trees over 100 or 1000 iterations [34] as implemented in TMeV [35], because the clustering results obtained with this method are more statistically sound as they do not depend on the original order of the genes and the pathways. Also, clustering has been applied successfully to gene expression studies [36–39]. From an ontological point of view the sBEFs can be considered as categorical phenotypes, thus clustering is a legitimate approach to classify "omics" data, as it has been used on datasets of both continuous and categorical data, such as human haplotypes in population studies [40,41] and phenotypic observations, such as lord Fisher's *Iris* data set [42], probably one of the most widely used data sets used in clustering and pattern recognition studies.

Yet since sBEFs are categorical rather than continuous, we investigated the use of an alternative metric (the Jaccard index),

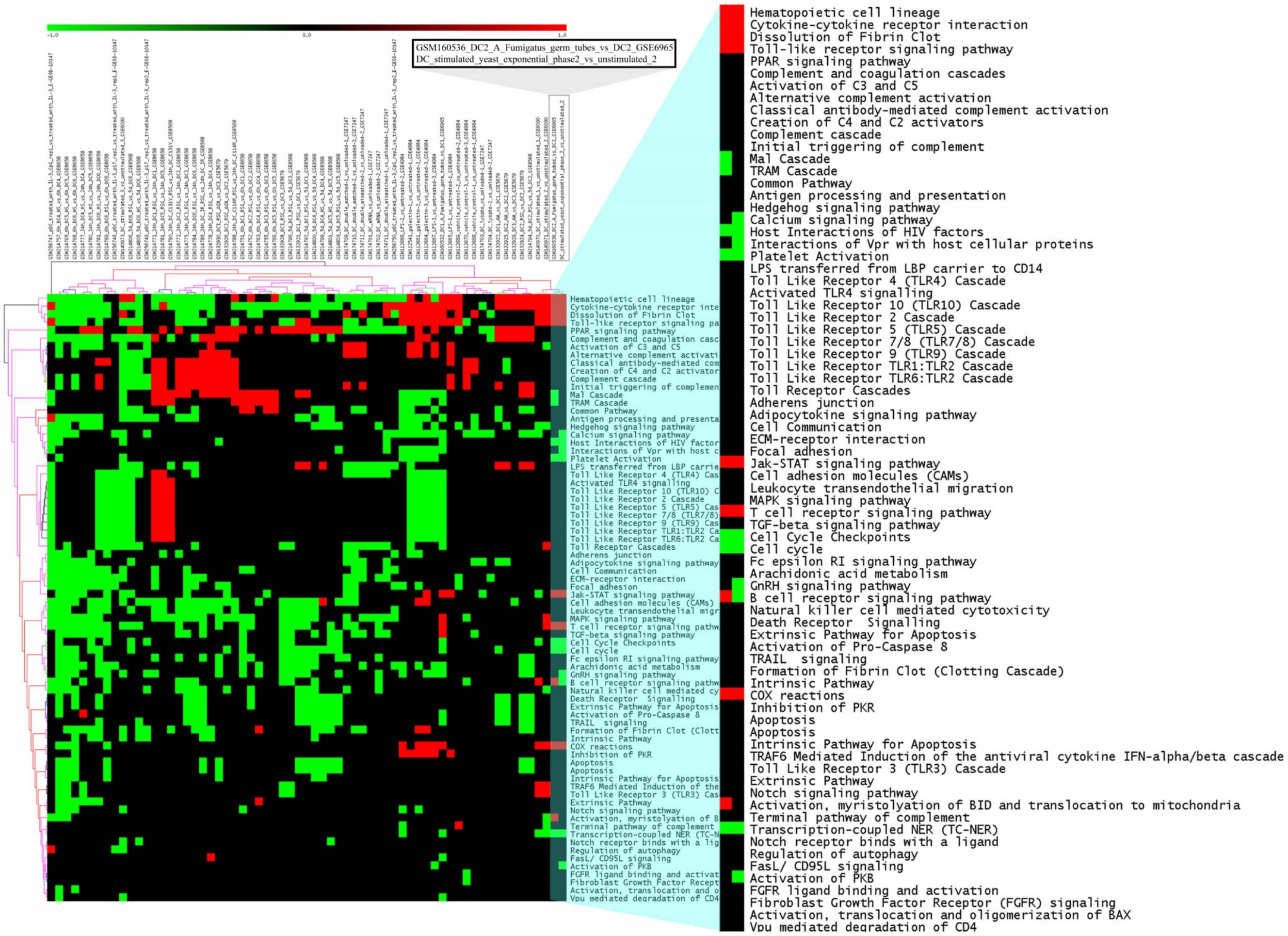

Figure 4. Clustering of signed Binary Enrichment Factors (sBEFs) obtained from the FET analysis on dendritic cell samples. Colored spots indicate significant ($p \leq 0.05$) up- (red) or down- (green) regulation. The colors of the dendrogram indicate the percentages of the tree support (significance), from 50% (pink) to 100% (black). The inset shows the clustering of sample GSM160356 from data set GSE6965 with the in-house dendritic cell experiment.

doi:10.1371/journal.pone.0004128.g004

since biclustering may not be completely suited for this type of data. Our results showed an excellent agreement with the clustering for DC data, but some inconsistencies with the yeast data, in agreement with the clustering of the PEFs for this dataset. This could reflect some biological instances, or alternatively could be related to noise deriving from technological issues, as the yeast dataset is a none year old two color array. Another possibility is that measures like the Jaccard index may not be the most robust to capture the inherent complexity of "omics" data. In any case, our future work will be aimed at improving the grouping metrics for sBEFs.

Results with GSEA on the dendritic cell data did not satisfy our prediction, due to the different numbers of experiments present in the different classes and the difficulty to classify one data set in a class or another. In fact, the grouping observed with FET scores on DC data was absent when using GSEA data. There are two possible reasons to this inconsistency. First of all, GSEA compares two groups (treated versus untreated) as opposed to single-sample analysis performed with FET. Thus, the effect of inter-donor variability (all data sets have at least two biological replicates for each sample) is noticeable and is not corrected by the pathway approach. The second reason is the heterogeneity of the samples themselves, that are subject to different treatments. As a result, GSEA rounds everything to the lowest common denominator, presenting an "average" profile where individual differences are smoothed out. Also, the number of control and treated samples in each data set is different (this is usually the norm when comparing different data sets), so there is an imbalance among the various GSEA analyses.

We can conclude that the ability to investigate at the single sample makes the Fisher's Exact Test conceptually more appropriate to search for similarities among experiments in a microarray databases, that contain a number of hybridizations that should be interrogated without necessarily specifying the membership to a data set or another. Lastly, GSEA requires dividing the samples in two distinct phenotypic classes in order to operate, and division in classes is not necessarily straightforward. As a matter of fact, such a type of analysis is not suitable for two-color microarrays (which present data as a ratio between treatment and control), and therefore we were not able to use it with our yeast data set.

Overall our results show that using pathway signatures in conjunction with hierarchical clustering with support trees is a powerful and useful technique to compare experiments produced by different people and laboratories with greater power than with the traditional analysis techniques. The results are also easier to

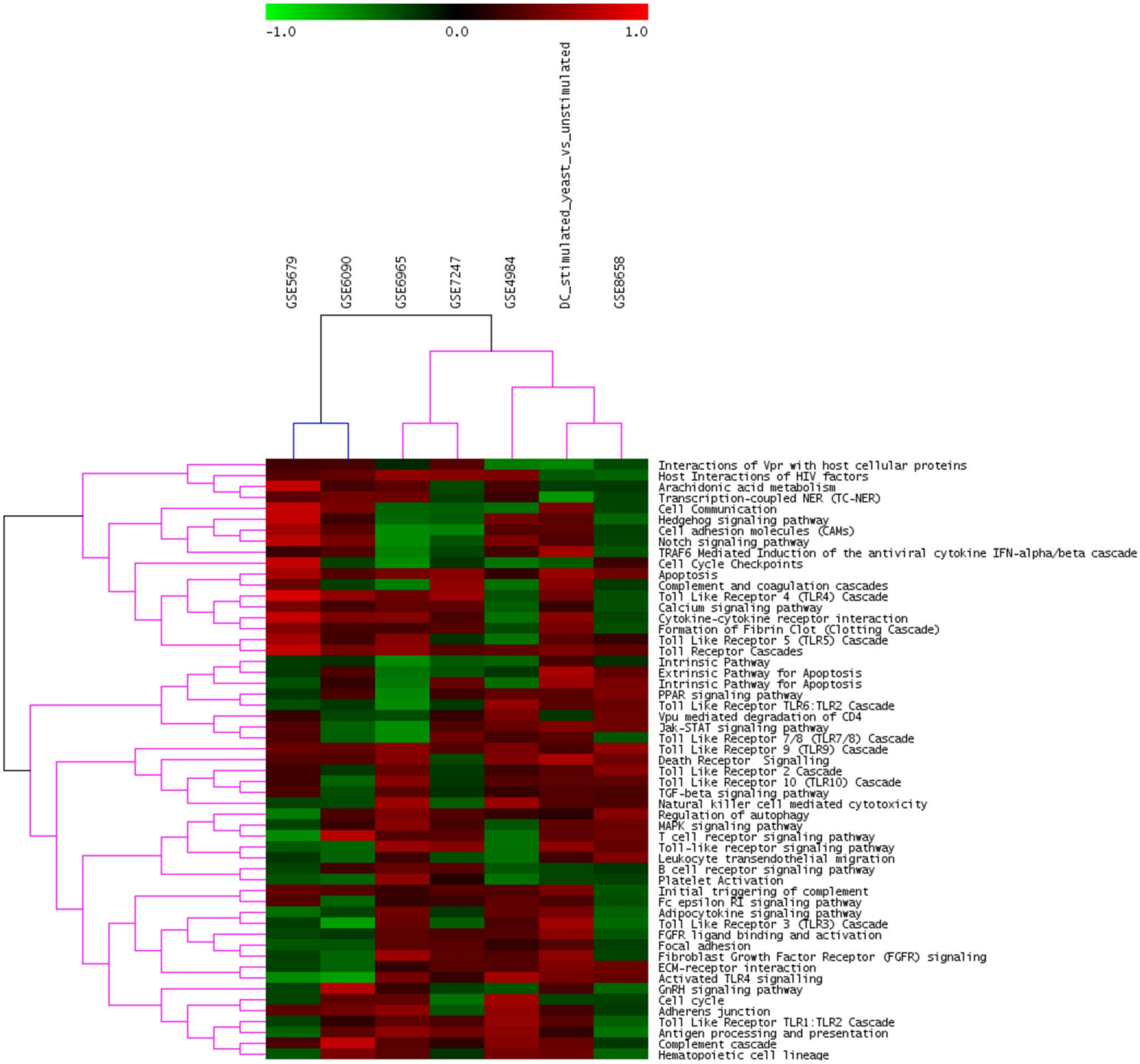

Figure 5. Clustering of Enchriment scores obtained from GSEA run on the dendritic cells data sets. Red indicates up-regulation, while green signifies down-regulation. The colors of the dendrogram indicate the percentages of the tree support (significance), from 50% (pink) to 100% (black).
doi:10.1371/journal.pone.0004128.g005

interpret and discover biologically meaningful implications, making this approach an ideal candidate to analyze data from different sources. PS generated using sBEFs can be useful as "barcodes" to classify experiments in microarray databases and clustering of sBEFs can be a useful way to query experiments in databases according to their similarity at a pathway level. Thus, we propose to store PSs as an additional experimental annotation in the microarray databases and implement methods using pathway signatures to query experiments in public databases and concurrent analyses of subsets of experiments.

Despite the effectiveness of the method, there are still some drawbacks that will need to be addressed in the future. First of all, Fisher's Exact Test result depend on the lower and upper cut-offs for expression, and at the current time they are defined by the user. We were well aware of the limitations of this method but also aware of its robustness if the appropriate threshold is used, in agreement with the findings from Bussemaker and Boorsma, that proved that the Fisher's Exact Test outperforms other metrics when appropriately selecting the threshold [43]. Thus, we calculated the appropriate cut-offs automatically, basing on the 2σ interval of the binomial distribution of the expression values [44], and we will implement this method directly in a future version of Eu.Gene. Secondly, both FET and GSEA make the assumption that genes in a pathway are independent from each other, which is clearly not the case in real biological system. An alternative system with the potential to dramatically improve the results of pathway analysis has been recently proposed [21], which keeps track of the interdependence of genes. However, its use is currently limited to signal transduction pathways from KEGG and requires an accurate description of causality among the events, which is not the case for the vast majority of the pathways present in public resources. We have partially incorporated this idea of direction by calculating signs attached to the p-values. We plan to increase the sophistication of the method, implementing proper directionality of the events using curated pathway sets with a consensus on the number, type and order of the connections between the members of the pathway. With the ability to include directionality, the application of methods for module network analysis using improved PSs instead of genes [45] holds the promise to unravel the hierarchical structure in the control of the cellular pathways and reconstruct the modular structure determining changes in a transcriptional profile.

Methods

Data sets

Yeast data sets. A selection of 37 experiments on yeast deletion mutants were obtained from the Rosetta compendium [23] . Prior to analysis, data (in the form of $\log_{10}$ ratios) were transformed into $\log_2$ ratios.

Dendritic cell data sets. 106 microarray samples, corresponding to 6 data sets, were retrieved from the Gene Expression Omnibus (GEO) database (Table 1). During the data selection process, only samples hybridized on Affymetrix GeneChip HG-U133 Plus 2.0 were selected. Experimental and sample data were then downloaded using a script written in R (http://www.r-project.org) which made use of the Bioconductor package GEOquery (http://www.bioconductor.org). Raw data (CEL files) were then downloaded and extracted using custom Bash shell scripts on the Linux operating system.

In-house stimulated dendritic cell samples. Peripheral blood mononucleated cells (PBMCs) were isolated from buffy coats by density gradient centrifugation using Biocoll (Biochrom). CD14+ monocytes were isolated from PBMCs by positive selection using MACS anti-CD14 microbeads and a Midi-MACS® magnetic cell sorting device (Miltenyi Biotec, Bergisch- Gladbach,Germany). Cells were cultured in RPMI 1640 medium (Gibco BRL) supplemented with 2 mM L-glutamine (Sigma), 1% (v/v) non-essential amino acids, 100 mM sodium pyruvate, 50 U/ml of penicillin and 50 mg/ml of streptomycin (Gibco BRL) containing 10% (v/v) FCS (Hyclone). Differentiation of monocytes into dendritic cells was promoted by addition of granulocyte-macrophage colony-stimulating factor (GM-CSF 1000 U/ml, Chemicon)and recombinant IL-4 (1000 U/ml, R&D Systems) for 5 days. $3*10^6$ DCs were either not stimulated or stimulated with yeast grown in exponential phase. Cells were harvested after 4 hr of stimulation.

RNA extraction was done with TRIzol reagent (Invitrogen). Sample pre-processing and biotin labeling were performed using the Affymetrix GeneChip® cDNA Synthesis Kit and IVT Labeling Kit (Affymetrix) according to the manufacturer's protocols. Microarray were then hybridized on Affymetrix GeneChip® HG-U133A 2.0 microarrays, and scanned according to the manufacturer's instructions on a GeneChip® Scanner 3000 (Affymetrix).

Extraction, hybridization and scanning were performed by the Genopolis consortium (University of Milano-Bicocca, Italy). Microarray data have been submitted to GEO (accession number GSE13901).

Pre-processing and normalization

CEL files were pre-processed and normalized with the Robust Multi-array Average (RMA) [46] procedure, in order to obtain

Table 1. Dendritic data set summary.

GEO ID	Description	Type	Sample no.	Reference	Year
GSE8658	PPARg regulated gene expression in human dendritic cells	Ligand response	63	[49]	2007
GSE7247	Dendritic Cells Compare the Similarity of Endogenous and Exogenous Antigens	Loading methodology comparison	10	[50]	2006
GSE4984	Monocyte Derived Dendritic Cell Maturation	Dendritic cell maturation	12	[51]	2006
GSE6090	DC-SIGN initiates an immature dendritic cell phenotype triggering Rho activation that is used by HIV-1	Receptor signalling	6	[52]	2007
GSE6965	Gene expression profiling of human dendritic cells after infection with A. fumigatus	Dendritic cell activation	4	[53]	2008
GSE5679	Comparative gene expression profile of PPARg and RARa ligand treated human dendritic cells	Ligand response	11	[54]	2008

Dendritic cells data sets obtained from Gene Expression Omnibus (GEO).
doi:10.1371/journal.pone.0004128.t001

absolute-scale expression levels for all genes, using custom chip definition files (CDFs) to update the annotation of the GeneChips, removing duplicate probes and mapping all probes to single Entrez Gene IDs, as previously described by Dai et al. [47]. Each data set was normalized separately. The computation was performed with the RMAExpress program, version 1.0 beta 10 (http://rmaexpress.bmbolstad.com), on the Linux operating system.

Selection of the pathway set

To identify biologically meaningful changes in our data sets, we assembled two pathway sets, depending on the species. All pathways were defined as gene lists (containing genes associated with the pathway). The sets were prepared as follows:

- ***Saccharomyces cerevisiae* data:** All the pathways present in Eu.Gene (n = 852)
- **Dendritic cell data**: Pathways containing less than three elements were excluded. Out of the remaining pathways, we built a selection curated by experts from the DC-THERA European network of excellence, where pathways were classified by their level of immunological relevance. Out of the 1038 human pathways available , 80 satisfying the experts' criteria were selected to be part of the set.

Pathway analysis

Eu.Gene [22] was used to perform the analysis on the selected pathway set. The analysis algorithms used were the Fisher's Exact Test (FET) [19] and Gene Set Enrichment Analysis (GSEA) [18].

Fisher's Exact Test

Prior to analysis, the Affymetrix data was converted into $\log_2$ ratios using the following procedure. For each data set, treated and untreated sample data were separated, and the expression level of each gene in each treated sample was divided by the expression level of the same gene in the corresponding paired untreated sample, and the resulting ratio was $\log_2$ transformed, obtaining a list of $\log_2$ ratios for each gene in each treated sample. The transformed expression data, corresponding to 66 ratios, were then used for Fisher's Exact Test.

Yeast data, once converted to $\log_2$ ratios, was instead used directly, without any other modifications.

The upper and lower cut-offs for FET were determined from the 2σ interval of the binomial distribution of expression values [44].The algorithm was then run using the hypergeometric distribution without approximation. Each pathway was associated to a signed p-value and a matrix of signed p-values for all samples was obtained for each data set. In order to provide a correction for multiple testing, p-values were adjusted following the procedure by Benjamini *et al.* [48].

Resulting p-values from the FET analysis were then transformed prior to clustering (see below).

Gene Set Enrichment Analysis

For each data set, samples were divided into treated ("Type-A") and untreated ("Type-B") groups and their expression measures (in absolute scale) were used directly for the analysis. GSEA was run on each data set separately, excluding pathways which contained less than 12 genes or more than 400 genes, to improve specificity and then significance of the obtained Enrichment Score was assessed by performing 1000 random permutations on the data. Enrichment scores from the analysis were then used for clustering.

p-value transformation (FET)

The FET p-value tables from Eu.Gene were then transformed in order to be used for hierarchical clustering. Two approaches were used. Firstly, p-values were transformed into Pathway Enrichment Factors (PEF) as follows:

$$\pm\log_2(1/p - \text{value})$$

where the sign is determined by the Sign of the p-value obtained from the FET analysis in Eu.Gene.

With the second approach, p-values were converted into signed Binary Enrichment Factors (sBEFs)

$$\begin{cases} \pm 1 & \text{if p-value} \leq 0.05 \text{ depending on Sign} \\ 0 & \text{if p-value} > 0 \text{ or Sign is not defined} \end{cases}$$

In the case of sBEFs, after transformation they were filtered to exclude non-significant pathways in all the samples (i.e., rows containing only zeroes in all samples).

Clustering

Hierarchical clustering was performed with the TIGR Multiexperiment Viewer (TMeV), version 4.1.1 [35]. Specifically, we used standard hierarchical clustering and clustering with support trees [34]: in the latter case pathways and samples were bootstrapped over 100 and 1000 iterations. Euclidean distance, Pearson's correlation and Manhattan distance were used as distance metrics, with average linkage clustering. Both pathways and samples were clustered.

Unweighted Pair Group Method with Arithmetic mean (UPGMA) clustering was performed using Eu.Gene's internal "Path-Blast" function, creating similarity trees between different experiments in each data set with a p-value threshold of 0.05.

Jaccard index calculation

From the sBEF matrix, the Jaccard similarity index was calculated for every possible pairing between the various sample columns. The computation was performed with the Python programming language, using the scipy-hcluster package (http://code.google.com/p/scipy-cluster/).

Supporting Information

Figure S1 Clustering of signed Binary Enrichment Factors using Euclidean distance using support trees on yeast data. Colored spots indicate significant ($p <= 0.05$) up- (red) or down- (green) regulation. The colors of the dendrogram indicate the percentages of the tree support (significance), from 50% (pink) to 100% (black). Found at: doi:10.1371/journal.pone.0004128.s001 (0.94 MB PNG)

Figure S2 Clustering of signed Binary Enrichment Factors using Pearson's Correlation using support trees on yeast data. Colored spots indicate significant ($p <= 0.05$) up- (red) or down- (green) regulation. The colors of the dendrogram indicate the percentages of the tree support (significance), from 50% (pink) to 100% (black). Found at: doi:10.1371/journal.pone.0004128.s002 (0.92 MB PNG)

Figure S3 Clustering of signed Binary Enrichment Factors using Manhattan Distance using support trees on yeast data. Colored spots indicate significant ($p <= 0.05$) up- (red) or down- (green) regulation. The colors of the dendrogram indicate the percentages of the tree support (significance), from 50% (pink) to 100% (black).

Found at: doi:10.1371/journal.pone.0004128.s003 (0.94 MB PNG)

Figure S4 Standard hierarchical clustering of signed Binary Enrichment Factors on yeast data. Colored spots indicate significant (p<=0.05) up- (red) or down- (green) regulation.
Found at: doi:10.1371/journal.pone.0004128.s004 (0.92 MB PNG)

Figure S5 UPGMA clustering of Fisher's Exact Test analysis results on yeast data.
Found at: doi:10.1371/journal.pone.0004128.s005 (0.16 MB JPG)

Figure S6 Clustering of signed Binary Enrichment Factors using Pearson's Correlation using support trees on dendritic cell data. Colored spots indicate significant (p<=0.05) up- (red) or down- (green) regulation. The colors of the dendrogram indicate the percentages of the tree support (significance), from 50% (pink) to 100% (black).
Found at: doi:10.1371/journal.pone.0004128.s006 (0.22 MB PNG)

Figure S7 Clustering of signed Binary Enrichment Factors using Manhattan Distance using support trees on dendritic cell data. Colored spots indicate significant (p<=0.05) up- (red) or down- (green) regulation. The colors of the dendrogram indicate the percentages of the tree support (significance), from 50% (pink) to 100% (black).
Found at: doi:10.1371/journal.pone.0004128.s007 (0.22 MB PNG)

Figure S8 Standard hierarchical clustering of signed Binary Enrichment Factors on dendritic cell data. Colored spots indicate significant (p<=0.05) up- (red) or down- (green) regulation.
Found at: doi:10.1371/journal.pone.0004128.s008 (0.20 MB PNG)

Figure S9 UPGMA clustering of Fisher's Exact Test analysis results on dendritic cell data.
Found at: doi:10.1371/journal.pone.0004128.s009 (0.63 MB JPG)

Acknowledgments

We would like to thank Chris Evelo, Bruce Conklin and Thomas Kelder for discussing the paper, and Chiara Romualdi for helpful discussion.

Author Contributions

Conceived and designed the experiments: LB PRS LG CB DC. Performed the experiments: LR. Analyzed the data: LB RP. Contributed reagents/materials/analysis tools: LR LG. Wrote the paper: LB DC.

References

1. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, et al. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 29: 365–371. doi:10.1038/ng1201-365.
2. Ball CA, Sherlock G, Parkinson H, Rocca-Sera P, Brooksbank C, et al. (2002) Standards for microarray data. Science 298: 539.
3. Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30: 207–210.
4. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, et al. (2003) ArrayExpress–a public repository for microarray gene expression data at the EBI. Nucleic Acids Res 31: 68–71.
5. Ein-Dor L, Kela I, Getz G, Givol D, Domany E (2005) Outcome signature genes in breast cancer: is there a unique set? Bioinformatics 21: 171–178. doi:10.1093/bioinformatics/bth469.
6. Xu L, Tan AC, Winslow RL, Geman D (2008) Merging microarray data from separate breast cancer studies provides a robust prognostic test. BMC Bioinformatics 9: 125. doi:10.1186/1471-2105-9-125.
7. Ein-Dor L, Zuk O, Domany E (2006) Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc Natl Acad Sci U S A 103: 5923–5928. doi:10.1073/pnas.0601231103.
8. Allison DB, Cui X, Page GP, Sabripour M (2006) Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet 7: 55–65. doi:10.1038/nrg1749.
9. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AAM, et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415: 530–536. doi:10.1038/415530a.
10. Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, et al. (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. N Engl J Med 346: 1937–1947. doi:10.1056/NEJMoa012914.
11. Yu J, Sieuwerts A, Zhang Y, Martens J, Smid M, et al. (2007) Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer. BMC Cancer 7: 182. doi:10.1186/1471-2407-7-182.
12. Ransohoff DF (2004) Rules of evidence for cancer molecular-marker discovery and validation. Nature Reviews Cancer 4: 309–314. doi:doi:10.1038/nrc1322.
13. Simon R, Radmacher MD, Dobbin K, McShane LM (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 95: 14–18.
14. Manoli T, Gretz N, Gröne H, Kenzelmann M, Eils R, et al. (2006) Group testing for pathway analysis improves comparability of different microarray datasets. Bioinformatics 22: 2500–2506. doi:10.1093/bioinformatics/btl424.
15. Khatri P, Draghici S, Ostermeier GC, Krawetz SA (2002) Profiling gene expression using onto-express. Genomics 79: 266–270. doi:10.1006/geno.2002.6698.
16. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, et al. (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol 4: R28. doi:PMC154579.
17. Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, Lawlor SC, et al. (2003) MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome Biol 4: R7.
18. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102: 15545–15550. doi:10.1073/pnas.0506580102.
19. Grosu P, Townsend JP, Hartl DL, Cavalieri D (2002) Pathway Processor: a tool for integrating whole-genome expression results into metabolic networks. Genome Res 12: 1121–1126. doi:10.1101/gr.226602.
20. Goeman JJ, Geer SAVD, Kort FD, Houwelingen HCV (2003) A global test for groups of genes: testing association with a clinical outcome. Bioinformatics 20: 93–99.
21. Draghici S, Khatri P, Tarca AL, Amin K, Done A, et al. (2007) A systems biology approach for pathway level analysis. Genome Res 17: 1537–1545. doi:10.1101/gr.6202607.
22. Cavalieri D, Castagnini C, Toti S, Maciag K, Kelder T, et al. (2007) Eu.Gene Analyzer a tool for integrating gene expression data with pathway databases. Bioinformatics 1: doi:10.1093/bioinformatics/btm333.
23. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, et al. (2000) Functional Discovery via a Compendium of Expression Profiles. Cell 102: 109–126. doi:10.1016/S0092-8674(00)00015-5.
24. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM (1999) Systematic determination of genetic network architecture. Nat Genet 22: 281–285. doi:10.1038/10343.
25. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, et al. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 34: 166–176. doi:10.1038/ng1165.
26. Segal E, Friedman N, Koller D, Regev A (2004) A module map showing conditional activity of expression modules in cancer. Nat Genet 36: 1090–1098. doi:10.1038/ng1434.
27. Segal E, Friedman N, Kaminski N, Regev A, Koller D (2005) From signatures to models: understanding cancer using microarrays. Nat Genet 37 Suppl: S38–S45. doi:10.1038/ng1561.
28. Malavé TM, Dent SYR (2006) Transcriptional repression by Tup1-Ssn6. Biochem Cell Biol 84: 437–43. doi:o06-073.
29. Smith RL, Johnson AD (2000) Turning genes off by Ssn6-Tup1: a conserved system of transcriptional repression in eukaryotes. Trends Biochem Sci 25: 325–30. doi:10871883.
30. Buentke E, Scheynius A (S.d.) Dendritic cells and fungi. APMIS 111: 789–96. doi:12974780.
31. van de Veerdonk FL, Kullberg BJ, van der Meer JWM, Gow NAR, Netea MG (2008) Host-microbe interactions: innate pattern recognition of fungal pathogens. Curr Opin Microbiol 11: 305–12. doi:S1369-5274(08)00074-X.
32. Roeder A, Kirschning CJ, Rupec RA, Schaller M, Weindl G, et al. (2004) Toll-like receptors as key mediators in innate antifungal immunity. Med Mycol 42: 485–98. doi:15682636.
33. Tada H, Nemoto E, Shimauchi H, Watanabe T, Mikami T, et al. (2002) Saccharomyces cerevisiae- and Candida albicans-derived mannan induced

production of tumor necrosis factor alpha by human monocytes in a CD14- and Toll-like receptor 4-dependent manner. Microbiol Immunol 46: 503–12. doi:12222939.

34. Graur D, Li W (2000) Fundamentals of Molecular Evolution. SunderlandMA: Sinauer Associates. pp 209–210.
35. Saeed AI, Sharov V, White J, Li J, Liang W, et al. (2003) TM4: a free, open-source system for microarray data management and analysis. Biotechniques 34: 374–378.
36. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95: 14863–14868.
37. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, et al. (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: 6745–50. doi:PMC21986.
38. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286: 531–537.
39. Milano A, Pendergrass SA, Sargent JL, George LK, McCalmont TH, et al. (2008) Molecular Subsets in the Gene Expression Signatures of Scleroderma Skin. PLoS ONE 3: e2696. doi:10.1371/journal.pone.0002696.
40. Hoehe MR, Köpke K, Wendel B, Rohde K, Flachmeier C, et al. (2000) Sequence variability and candidate gene analysis in complex disease: association of mu opioid receptor gene variation with substance dependence. Hum Mol Genet 9: 2895–908. doi:11092766.
41. Durrant C, Zondervan KT, Cardon LR, Hunt S, Deloukas P, et al. (2004) Linkage Disequilibrium Mapping via Cladistic Analysis of Single-Nucleotide Polymorphism Haplotypes. Am J Hum Genet 75: 35–43.
42. Fisher R (1936) The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics 7: 179–188.
43. Bussemaker HJ, Ward LD, Boorsma A (2007) Dissecting complex transcriptional responses using pathway-level scores based on prior information. BMC Bioinformatics 8 Suppl 6: S6. doi:PMC1995543.
44. Segota I, Bartonicek N, Vlahovicek K (2008) MADNet: microarray database network web server. Nucleic Acids Res;doi:10.1093/nar/gkn289.
45. Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, et al. (2002) Revealing modular organization in the yeast transcriptional network. Nat Genet 31: 370–377. doi:10.1038/ng941.
46. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4: 249–264. doi:10.1093/biostatistics/4.2.249.
47. Dai M, Wang P, Boyd AD, Kostov G, Athey B, et al. (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res 33: e175. doi:10.1093/nar/gni179.
48. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I (2001) Controlling the false discovery rate in behavior genetics research. Behav Brain Res 125: 279–284.
49. Szatmari I, Töröcsik D, Agostini M, Nagy T, Gurnell M, et al. (2007) PPARgamma regulates the function of human dendritic cells primarily by altering lipid metabolism. Blood 110: 3271–3280. doi:10.1182/blood-2007-06-096222.
50. Decker WK, Xing D, Li S, Robinson SN, Yang H, et al. (2006) Double loading of dendritic cell MHC class I and MHC class II with an AML antigen repertoire enhances correlates of T-cell immunity in vitro via amplification of T-cell help. Vaccine 24: 3203–3216. doi:10.1016/j.vaccine.2006.01.029.
51. Fulcher JA, Hashimi ST, Levroney EL, Pang M, Gurney KB, et al. (2006) Galectin-1-matured human monocyte-derived dendritic cells have enhanced migration through extracellular matrix. J Immunol 177: 216–226.
52. Hodges A, Sharrocks K, Edelmann M, Baban D, Moris A, et al. (2007) Activation of the lectin DC-SIGN induces an immature dendritic cell phenotype triggering Rho-GTPase activity required for HIV-1 replication. Nat Immunol 8: 569–577. doi:10.1038/ni1470.
53. Mezger M, Kneitz S, Wozniok I, Kurzai O, Einsele H, et al. (2008) Proinflammatory response of immature human dendritic cells is mediated by dectin-1 after exposure to Aspergillus fumigatus germ tubes. J Infect Dis 197: 924–31. doi:10.1086/528694.
54. Szatmari I, Pap A, Rühl R, Ma J, Illarionov PA, et al. (2006) PPARgamma controls CD1d expression by turning on retinoic acid synthesis in developing human dendritic cells. J Exp Med 203: 2351–2362. doi:10.1084/jem.20060141.