PLOS | COMPUTATIONAL BIOLOGY

Differential Expression Analysis for Pathways

Winston A. Haynes[1,2,3]*, Roger Higdon[1,2,4], Larissa Stanberry[1,2,4], Dwayne Collins[3], Eugene Kolker[1,2,4,5]

1 Bioinformatics & High-Throughput Analysis Laboratory, Seattle Children's Research Institute, Seattle, Washington, United States of America, **2** Data-Enabled Life Sciences Alliance International (DELSA Global), Seattle, Washington, United States of America, **3** Department of Mathematics and Computer Science, Hendrix College, Conway, Arkansas, United States of America, **4** Seattle Children's, Predictive Analytics, Seattle, Washington, United States of America, **5** Departments of Biomedical Informatics & Medical Education and Pediatrics, University of Washington School of Medicine, Seattle, Washington, United States of America

Abstract

Life science technologies generate a deluge of data that hold the keys to unlocking the secrets of important biological functions and disease mechanisms. We present DEAP, Differential Expression Analysis for Pathways, which capitalizes on information about biological pathways to identify important regulatory patterns from differential expression data. DEAP makes significant improvements over existing approaches by including information about pathway structure and discovering the most differentially expressed portion of the pathway. On simulated data, DEAP significantly outperformed traditional methods: with high differential expression, DEAP increased power by two orders of magnitude; with very low differential expression, DEAP doubled the power. DEAP performance was illustrated on two different gene and protein expression studies. DEAP discovered fourteen important pathways related to chronic obstructive pulmonary disease and interferon treatment that existing approaches omitted. On the interferon study, DEAP guided focus towards a four protein path within the 26 protein Notch signalling pathway.

Citation: Haynes WA, Higdon R, Stanberry L, Collins D, Kolker E (2013) Differential Expression Analysis for Pathways. PLoS Comput Biol 9(3): e1002967. doi:10.1371/journal.pcbi.1002967

Editor: Richard Bonneau, New York University, United States of America

Received June 29, 2012; **Accepted** January 18, 2013; **Published** March 14, 2013

Funding: Research reported in this publication was supported by the National Science Foundation under Division of Biological Infrastructure award 0969929, National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health under awards U01-DK-089571 and U01-DK-072473, The Robert B. McMillen Foundation award, and DELSA award from The Gordon and Betty Moore Foundation to EK. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Science Foundation, National Institutes of Health, The McMillen Foundation, or The Moore Foundation. 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: winston.haynes@seattlechildrens.org

Introduction

High throughput technologies, such as next generation sequencing, microarrays, mass spectrometry proteomics, and metabolomics, are capable of evaluating the expression levels of thousands of genes, proteins, or metabolites in an individual run. As a result, the life sciences are experiencing a massive influx of data, exponentially increasing the size of databases [1–3]. Currently, databases contain millions of data sets from transcriptomics and thousands of from proteomics [4–10]. Differential expression analysis, the comparison of expression across conditions, has become the primary tool for finding biomarkers, drug targets, and candidates for further research. Typically, gene expression data have been analyzed on a gene-by-gene basis, without regard for complex interactions and association mechanisms. Ignoring the underlying biological structure diminishes the power of analysis, obscuring the presence of important biological signals.

Biological Pathways

Genes and proteins can be grouped into different categories on the basis of many traits: sequence, function, interactions, etc.. Grouping genes by biological pathway is often the most relevant approach to biologists. For this study, we represent biological pathways as directed graphs, where the nodes are biological compounds and the edges represent their regulatory relationships, either catalytic or inhibitory. A catalytic edge exists when expression of the parent node increases expression of the child node (i.e. A_3 is a parent to child A_4 with a catalytic edge, *Figure 1*). In an inhibitory relationship, expression of the parent node decreases expression of the child node (i.e. A_1 is a parent to child A_4 with an inhibitory edge, *Figure 1*). Further, we define a path as a connected subset of the pathway (i.e. $A_3A_4A_7$ is a path, $A_1A_2A_3$ is not, *Figure 1*). We use the term path to signify either a simple path or a simple cycle, where the term simple implies no repeated nodes.

While biological pathways have long been known, recent experimental data and computational advances have elucidated many previously uncharacterized mechanisms. Repositories contain information about thousands of biological pathways, with each pathway containing up to several hundred proteins [11–14]. Identifying the handful of pathways most relevant to a particular data set is an important challenge. The primary assumption of this paper is that biologically relevant pathways are characterized by co-regulated differential expression of their paths.

Gene Set Analysis

Currently, the most popular approach to connect expression data to pathways is through gene set analysis. Gene set analysis methods consider sets of genes simultaneously as opposed to the gene-by-gene basis commonly used in differential expression analysis. One of the most prominent set-based methods is Gene Set Enrichment Analysis (GSEA), where the identified genes are ranked based on expression values [15,16]. Significance of

Author Summary

The data deluge represents a growing challenge for life sciences. Within this sea of data surely lie many secrets to understanding important biological and medical systems. To quantify important patterns in this data, we present DEAP (Differential Expression Analysis for Pathways). DEAP amalgamates information about biological pathway structure and differential expression to identify important patterns of regulation. On both simulated and biological data, we show that DEAP is able to identify key mechanisms while making significant improvements over existing methodologies. For example, on the interferon study, DEAP uniquely identified both the interferon gamma signalling pathway and the JAK STAT signalling pathway.

enriched gene sets is determined from a maximum running sum, which is calculated for each gene set by simultaneously walking down the ranked gene list and incrementing or decrementing the score on the basis of set membership. Other approaches calculate set based scores through different metrics and distributions [17–21]. Some of these methods compare gene sets relative to others (known as enrichment analysis or competitive approaches) while others compare individual gene sets across conditions without regard for other sets (known as self -contained approaches) [22].

The major limitation of set-based approaches in their application to pathway datasets is that they neglect the graph structure of the pathway. For example, in *Figure 1*, sporadic patterns of expression in nodes $A_1..A_8$ would prevent identification of significant differential expression by set analysis. Considering the additional information contained in the edges, it becomes clear that $A_3A_4A_7$ represents a path with similar differential expression from reactants to products. Consequently, $A_3A_4A_7$ represents a differentially expressed path and may possess biological significance, but is unlikely to be identified as such by set based approaches.

Pathway Analysis

We define pathway analysis as any approach which identifies patterns of differential expression in a data set by considering pathway structure. In pathway analysis, researchers are generally interested in identifying pathways associated with a biological condition and determining the components of those pathways that explain the association. Thus, hypothesis testing can be viewed as a two-step procedure: first, test an entire pathway for differential expression; second, identify the path providing the greatest contribution to that differential expression. Recent approaches to pathway analysis test the generic hypothesis of a pathway differential expression without identifying specific paths [23–31].

One of the most popular methods for pathway analysis, signalling pathway impact analysis (SPIA, *Table 1*) combines a set analysis score with a cumulative pathway score [23,24]. The pathway score is calculated by summing all edges in the graph. Catalytic and inhibitory relationships are considered by using a multiplier on the expression values. While this score takes into consideration the graph structure of pathways, it includes all possible paths, rather than just differentially expressed paths. For example, in *Figure 1*, the SPIA score would be based on the combination of path scores for $A_3A_4A_7$, $A_1A_4A_7$, $A_2A_5A_7$, $A_3A_6A_8$, and $A_3A_6A_7$ and the set score for $A_1..A_8$.

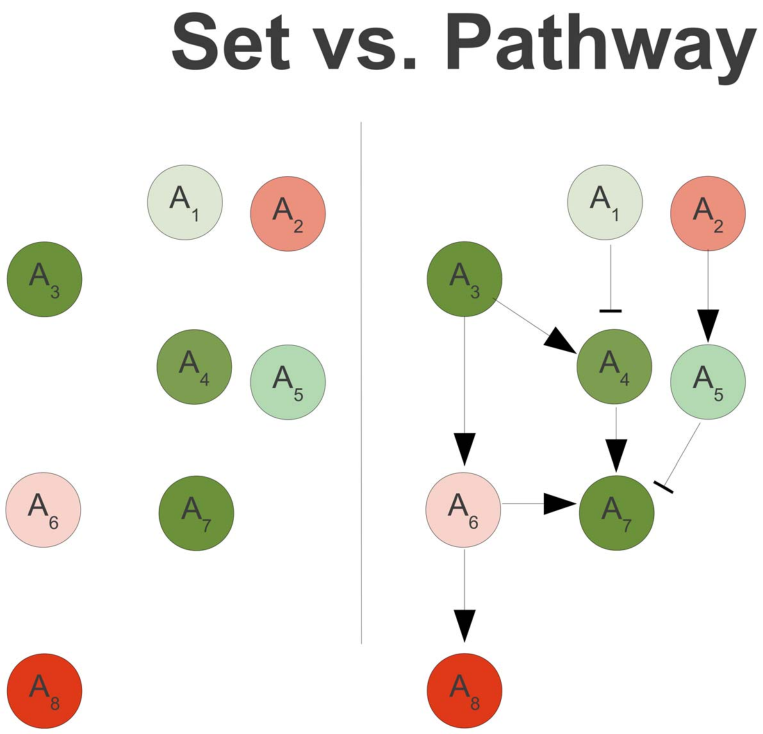

Figure 1. Set *vs.* pathway. Coloration from green to red represent differential expression levels, where dark green corresponds to high over expression and dark red indicates severe under expression. Edges with arrows and bars represent catalytic and inhibitory relationships, respectively. Considering $A_1..A_8$ as one set results in inconclusive patterns of gene expression. By considering pathway relationships, $A_3A_4A_7$ is recognized as a path of differentially expressed genes.
doi:10.1371/journal.pcbi.1002967.g001

Table 1. Term definitions.

Term	Definition
COPD	Chronic obstructive pulmonary disease
DEAP	Differential expression analysis for pathways. The approach presented in this paper.
False discovery rate	A statistical measure in multiple hypothesis testing which controls for the number of falsely rejected null hypothesis.
False positive rate	See Type I error rate.
GSEA	Gene set enrichment analysis [16]
High differential expression	In text, low differential expression refers specifically to simulations with $\mu = 1$
KEGG	Kyoto Encyclopedia of Genes and Genomes [13]. A pathways database.
Low differential expression	In text, low differential expression refers specifically to simulations with $\mu = 0.25$
MOPED	Model organism protein expression database [10] http://moped.proteinspire.org
p-value	Probability of obtaining the test statistic by random chance.
PANTHER	Protein analysis through evolutionary relationships [11]. A pathways database.
Path	A subset of a pathway which is connected by biochemical interactions.
Pathway	A series of biochemical interactions used in biological systems to perform biological functions.
Power	The frequency of occurrence of true positives. Equivalent to one minus the type II error rate (false negative rate).
Reactome	A pathways database [14]
SPIA	Signaling pathway impact analysis [24]
SPIRE	Systematic proteomics investigative research environment [40] http://www.proteinspire.org
Type I error rate	The frequency of occurrence of type I errors, false positives.
μ	Pathway effect. The average of 'on' genes within a pathway.

doi:10.1371/journal.pcbi.1002967.t001

Protein interaction permutation analysis, designed for siRNA experiments, calculates the significance of the number of interactions in a network for which both genes are "hits" [25]. Recently, *Zhao et al.* introduced an approach that includes pathway structure in the analysis of genome wide association studies [26]. However, neither of these methods are directly applicable to expression data. Other pathway analysis approaches calculate set enrichment scores, but weight gene products based on their correlation with neighboring genes in the pathway [27,28]. Alternatively, other approaches integrate omics data over pathways, but encode all expression data as -1, 0, or 1, limiting the information utilized from experimentation [29]. A mixed linear model presents an advanced approach to the hypothesis test, but is limited to acyclic models and implementation remains complex [30]. Like SPIA, all of these approaches account for the pathway structure as a whole, rather than identifying differentially expressed paths. To our knowledge, popular commercial pathway tools (i.e. Ingenuity Pathway Analysis, BioBase, GeneGo, Metacore, Ariadne) currently offer no methods that directly incorporate pathway analysis.

Significance Testing

High-throughput data analysis typically falls into the category of $p>>n$ problems, where the number of genes or proteins, p, is considerably larger than the number of samples, n. Pathway and gene set analysis methods have the added complexity that gene expression within pathways is often highly correlated. Therefore, the statistical analysis approaches described above typically rely on random permutations of biological replicates in order to preserve expression correlation structure. However, the small sample size limits the number of possible permutations and, hence, the precision of *p*-value estimates. In addition, permutation tests are only applicable to simple experimental designs. Utilizing a random rotation approach circumvents these issues [32–34].

Proposed Solution: DEAP

In this study, we present a new pathway analysis method, Differential Expression Analysis for Pathways (DEAP). The primary assumption of DEAP is that patterns of differential expression in paths within a pathway are biologically meaningful. DEAP calculates the path within each pathway with the maximum absolute running sum score where catalytic/inhibitory edges are taken as positive/negative summands. To assess the statistical significance, we use a random rotation. Similar to other pathway analysis methods, DEAP tests a generic hypothesis of overall pathway differential expression. Contrary to current methods, DEAP identifies the most differentially expressed path to provide a refined focus for further biological exploration.

Result*s*

DEAP Algorithm

As illustrated in *Figure 2*, the DEAP algorithm begins by overlaying expression data onto the pathway graph (*Figure 2.1*). Every possible path from the graph is independently examined (*Figure 2.2*). A recursive function calculates the differential expression for each path by adding or subtracting all downstream nodes with catalytic or inhibitory relationships, respectively (*Figure 2.3*).

As an example, the score for the path containing all nodes in the inhibitory string in *Figure 3 (left)*, where green = +1 and red = -1, is calculated as:

$$\begin{aligned} & B_1-(B_2-(B_3-(B_4-(B_5-B_6)))) \\ & =1-(-1-(1-(-1-(1--1)))) \\ & =1-(-1-(1-(-1-(2))))=\ldots=6 \end{aligned} \tag{1}$$

The absolute value of the expression level is utilized as the DEAP score (*Figure 2.4*) to determine the path with maximal differential

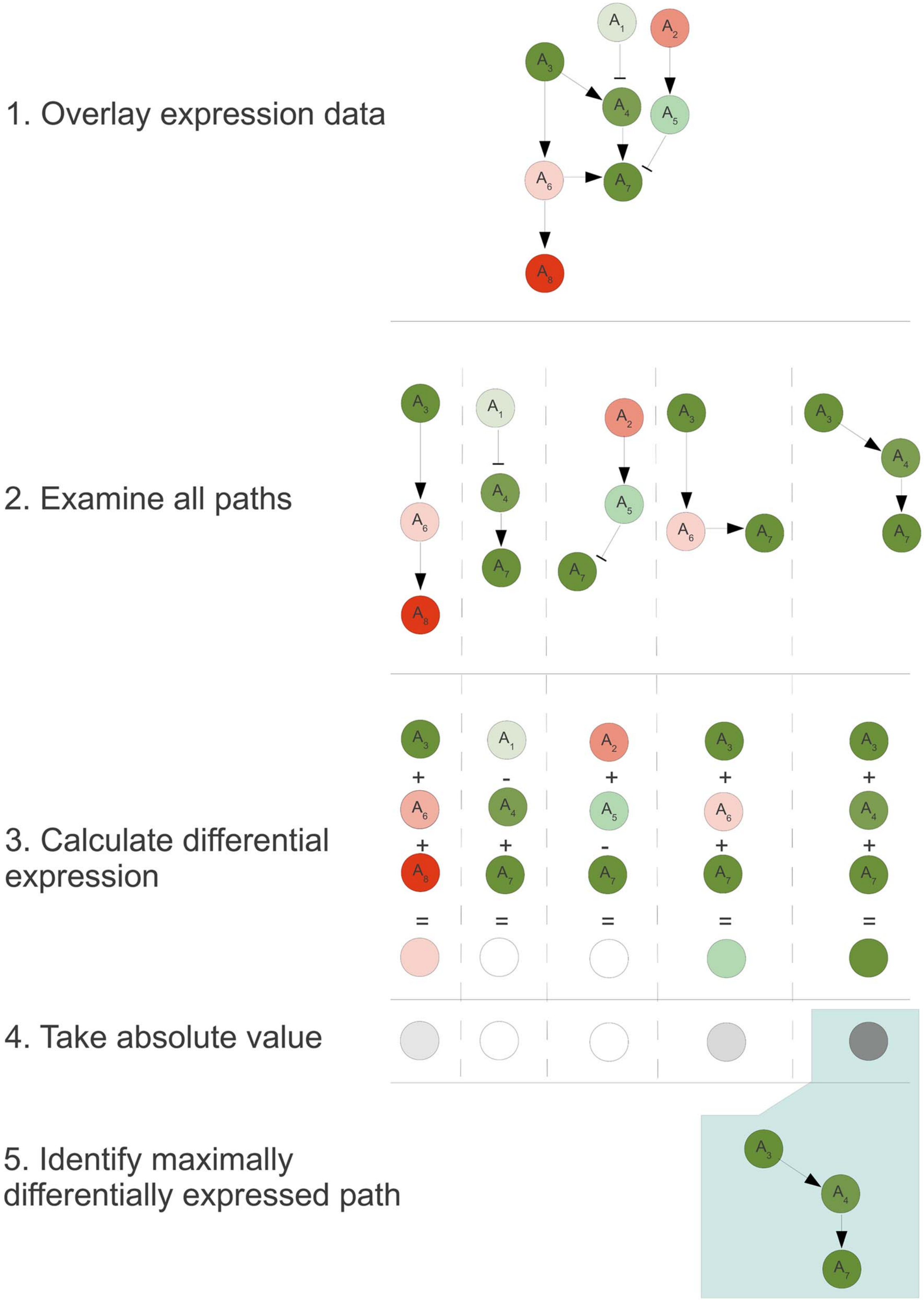

Figure 2. DEAP algorithm workflow. A visual representation of the DEAP algorithm workflow described in *Methods*.
doi:10.1371/journal.pcbi.1002967.g002

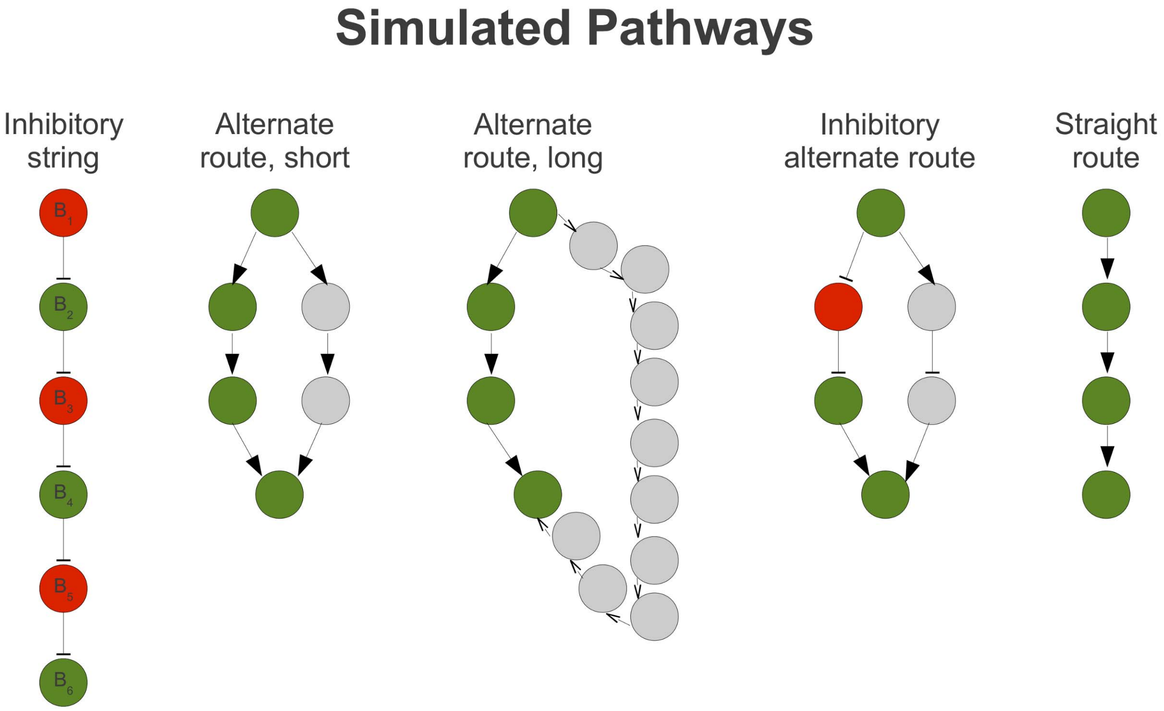

Figure 3. Simulated pathways. Five pathways designed to test analysis approaches are illustrated. Nodes labelled in green and red were built around distributions with μ of $+X$ and $-X$, respectively, where X represents a numerical value. Gray nodes represent data sampled from the standard normal distribution. Edges with arrows and bars represent catalytic and inhibitory relationships, respectively.
doi:10.1371/journal.pcbi.1002967.g003

expression (*Figure 2.5*). The DEAP algorithm returns both the maximum absolute value and the path associated with that maximum value. The algorithm is formalized in *Methods: DEAP Algorithm*.

DEAP scores for different pathways are not directly comparable due to size and structure differences among pathways. Thus, we employ a self-contained approach which individually assesses the significance of each pathway. Generating a null distribution is complicated by the low number of samples relative to gene identifications and the correlation of gene expression within pathways. Most existing approaches use permutation tests to preserve the correlation between genes; however, small sample size limits their effectiveness. We use random rotation to circumvent these issues [32–34]. Our random rotation implementation is applicable to a wide range of complex experimental designs with multiple conditions and replicates. The significance levels are adjusted for multiple comparisons using the false discovery rate method of Storey and Tibshirani [35].

For each pathway in the analysis, DEAP outputs its score, the corresponding p-value, and the path with the maximum absolute score (see examples in *Files S1, S2*). The open source implementation (licensed under the GNU Lesser General Public License v3.0) of this algorithm is available in Supplemental Materials (*File S3*).

DEAP Validation 1: Simulated Data on Simulated Pathways

Data from the five pathways illustrated in *Figure 3* were simulated as described in *Methods*. Algorithmic performance was measured in terms of power, the percentage of times each differentially expressed pathway was identified as significant ($p<0.05$), which is equivalent to one minus the type II error rate. The power of DEAP was compared to GSEA and SPIA, the two most popular gene set and pathway analysis methods, respectively. Comparative analysis of these methods included four key parameters: the overall effect (mean of 'on' genes, μ), variation in individual gene effects (σ^2_g), sample size (n), and type I error rate.

Regardless of the level of differential expression, DEAP was consistently more powerful than were other approaches (*Figure 4*). For small μ values (low differential expression), the power of DEAP was approximately twice that of GSEA and SPIA, demonstrating improved sensitivity. For $\mu = 1$ (high differential expression), DEAP had an increase in power over both GSEA and SPIA of two orders of magnitude. At $\mu = 1.25$, the performance of SPIA improved substantially, approaching that of DEAP on all pathways except the long alternate route where SPIA was confounded by noise (*Figure S1*). Across the board, GSEA performed poorly because GSEA did not consider pathway structure and is dependent on comparisons to other pathways.

Sample size and within-gene variance also have significant effects on the performance of the algorithms. As sample size (n) grew, the power of DEAP relative to other approaches increased, particularly in pathways containing inhibitory edges (*Figure 5*). As variance (σ^2_g) increased, DEAP exhibited minor increases in power (*Figure 6*). Further, DEAP consistently outperformed GSEA and SPIA as variance increased.

To estimate the type I error rate, we simulated random data under the null hypothesis ($\mu = 0$, $\sigma^2_g = 0$, $n = 10$). The plots in *Figure 7* displays type I error rates with respect to the nominal values. SPIA was notably more conservative for every pathway structure. The performance of both GSEA and DEAP was on target; however, DEAP was more conservative on pathways with inhibitory edges (*Figure S4*).

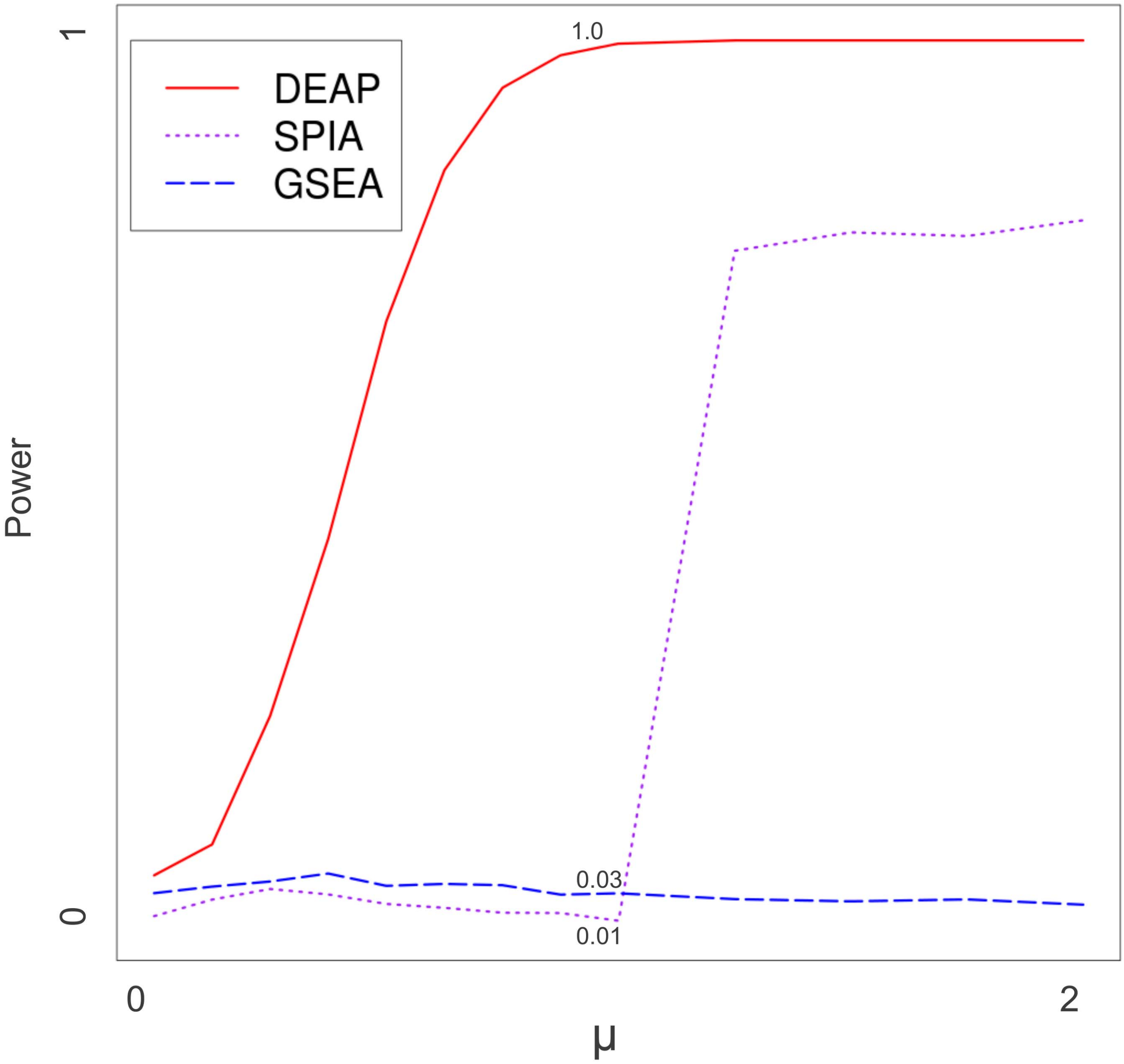

Figure 4. Power curve, variable pathway effect. Performance of GSEA, SPIA, and DEAP are compared as pathway effect (μ) changes. Specific values are indicated at $\mu = 1$. Power (y-axis) is the ratio of simulations, out of 5000 (5 pathways, 1000 simulations each), which were identified as significant ($p < 0.05$). Constants were $\sigma^2_g = 0$ and sample size = 10.
doi:10.1371/journal.pcbi.1002967.g004

An additional advantage of DEAP is the ability to identify the maximally differentially expressed path of the pathway. For the simulated data with $\mu = 1$ and $\mu = 2$, DEAP identified the entire differentially expressed path 99% and 100% of the time, respectively. For example, the long alternate route contains 14 proteins, but DEAP identified the differentially expressed region that contains only four, substantially reducing the search space.

In addition to comparing DEAP to GSEA and SPIA, we compared DEAP to several modifications of the DEAP algorithm, which were altered as follows: scores normalized by pathway length; all weights set to +1; and sum taken across the entire pathway. We also compared DEAP to a set-based implementation with rotation. DEAP had substantially higher power than all four approaches (*Table S1* and *Figures S1, S2, S3, S4*).

DEAP Validation 2: Simulated Data on Biological Pathways

While simulated pathways provide easily controllable examples to validate DEAP as an appropriate test of the hypothesis, biological pathways bring increased complexity from which the signal must be detected. To validate DEAP on more realistic pathway structures, we simulated activity on biological pathways from the KEGG and Reactome databases [13,14].

In the case of KEGG [13], we simulated data on the TGF-ß signaling pathway to indicate activity in the TGF- ß receptors leading to cell cycle arrest (*Figure 8*). In terms of sensitivity to the pathway effect (μ), variance (σ^2_g), and sample size, DEAP outperformed both GSEA and SPIA on the TGF-ß signaling pathway (*Figure 8*). Notably, increased variance diminishes the

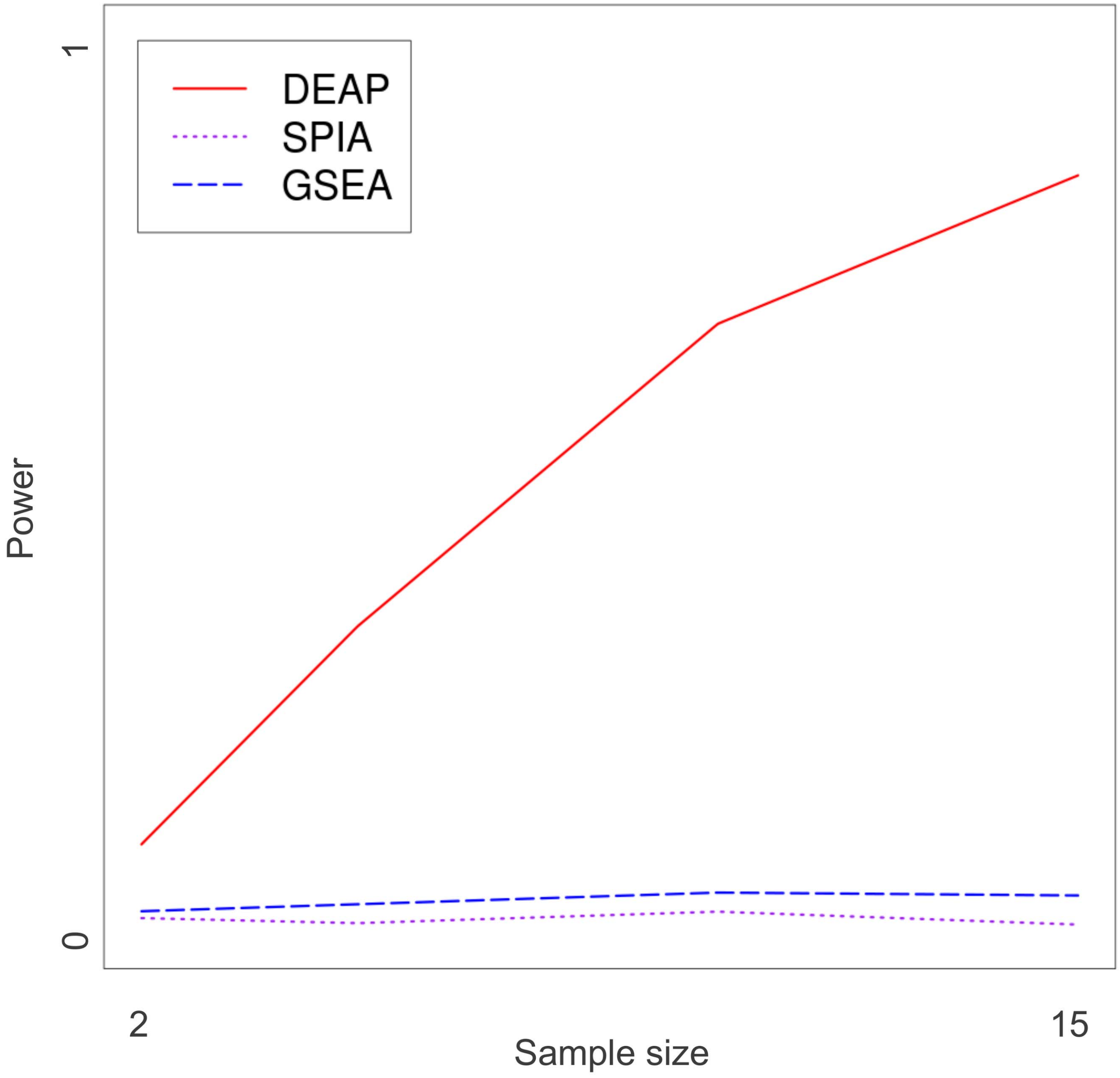

Figure 5. Power curve, variable gene variance. Performance of GSEA, SPIA, and DEAP are compared as gene variance (σ^2_g) changes. Power (*y*-axis) is the ratio of simulations, out of 5000 (5 pathways, 1000 simulations each), which were identified as significant ($p<0.05$). Constants were $\mu=0.5$ and sample size = 10.
doi:10.1371/journal.pcbi.1002967.g005

power of SPIA, but does not affect DEAP, reflecting its ability to identify signal in the noisy environments common in biological experimentation.

In the case of Reactome [14], we simulated data on the post-transcriptional silencing by small RNAs pathway from to indicate RNA cleavage (*Figure 9*). DEAP had superior performance over GSEA and SPIA in terms of all tested variables: pathway effect (μ), variance (σ^2_g), and sample size (*Figure 9*).

In both sets of simulated data on real biological pathways, the type I error estimate was conservative for DEAP, GSEA, and SPIA (*Figure S5*). In addition to DEAP, GSEA, and SPIA, we applied the four alternative formulations of DEAP to both sets of biological pathways and noted the consistently strong performance of DEAP (*Figures S6, S7*).

DEAP Validation 3: Biological Data on Biological Pathways

To verify that the simulated data effects are biologically relevant, we also applied DEAP to two sets of biological data on biological pathways. The experimental data are from a transcriptomic study of interferon [36,37] and a proteomic study of chronic obstructive pulmonary disease (COPD). We applied DEAP, GSEA, and SPIA to identify differentially expressed pathways from the PANTHER database [11]. Pathway associations with the phenotypes were determined based on a literature review using Google Scholar (details in *Methods: Biological Data Validation*).

We analyzed a microarray expression data of cells of radio-insensitive tumors that had been treated with interferon [36,37]. DEAP identified six pathways with known literature associations

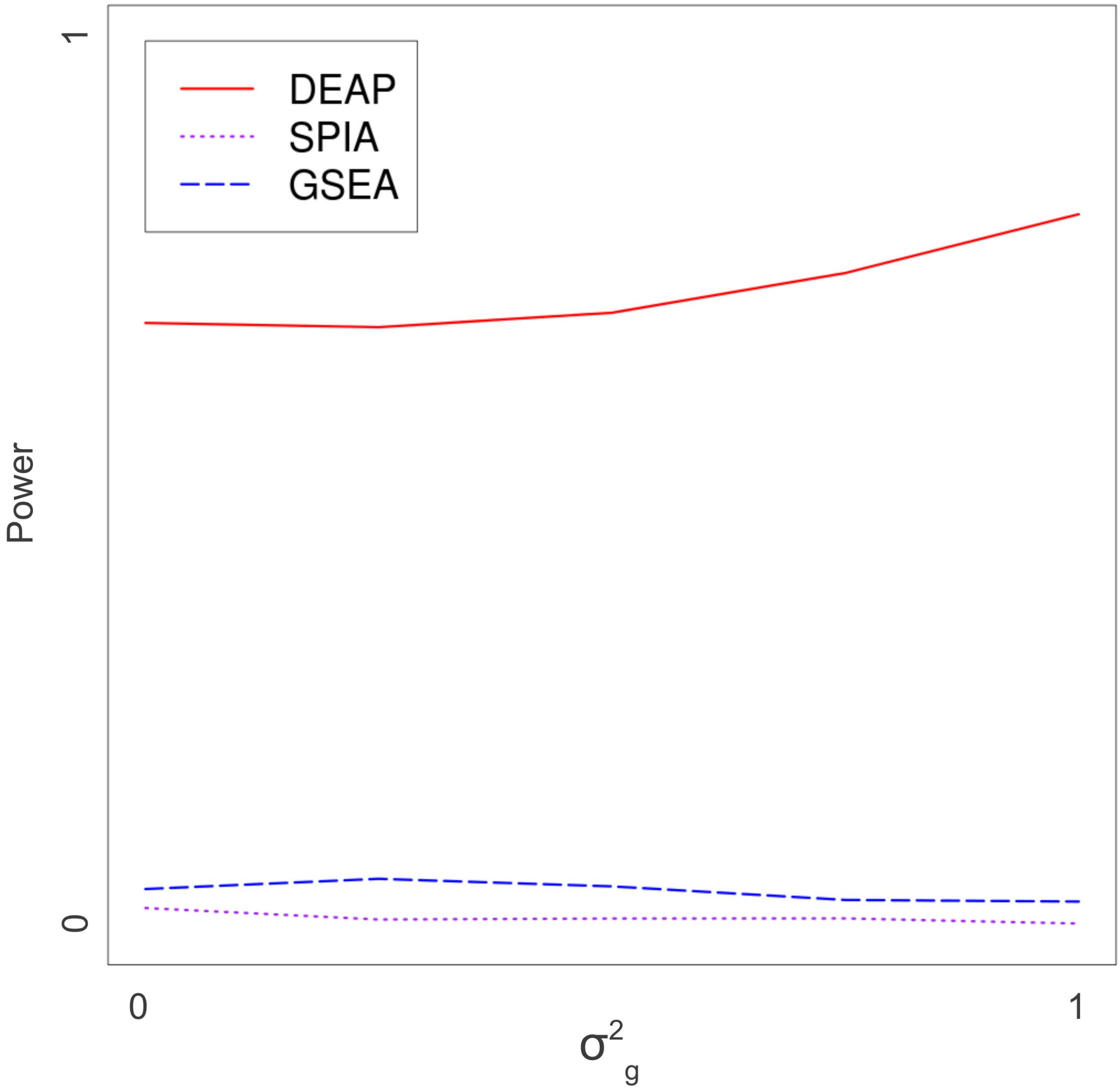

Figure 6. Power curve, variable sample size. Performance of GSEA, SPIA, and DEAP are compared as sample size changes. Power (*y*-axis) is the ratio of simulations, out of 5000 (5 pathways, 1000 simulations each), which were identified as significant ($p<0.05$). Constants were $\sigma^2_g=0$ and $\mu=0.5$.
doi:10.1371/journal.pcbi.1002967.g006

with interferon while GSEA identified five and SPIA identified none (*Table 2*). The two most clearly relevant pathways for this transcriptomics data set were interferon gamma signalling, as the cells had been stimulated with interferon; and JAK STAT signalling, the pathway being studied by the authors of the microarray study [36,37]. Unlike GSEA and SPIA, DEAP identified these pathways as significantly differentially expressed. The lack of overlap between the pathways identified by GSEA and DEAP is indicative of the different hypotheses being tested by these two approaches, with GSEA focusing on non-specific differential expression among pathway genes and DEAP focusing on differential expression among pathway connected genes. As such, these two approaches should be viewed as complementary approaches that can be simultaneously utilized to augment biological discovery.

Additionally, DEAP analysis of the interferon transcriptomics data uses path identification to reduce the search space for future experimentation. Consider the Notch signalling pathway, which contains 26 proteins and is known to be activated by interferon treatment [38]. GSEA and SPIA both did not identify Notch signalling as significantly differentially expressed due to generally sporadic expression patterns. However, DEAP analysis focused on consistent differential expression of 4 connected nodes and labelled Notch signalling as significantly differentially expressed (*Figure 10*). Without identifying the maximally differentially expressed path, the Notch signalling pathway would have been overlooked. Further, future experimentation can now focus on those four proteins exhibiting the most significant differential expression.

In order to illustrate DEAP on a different data type, we also analyzed a proteomics study which compared healthy smokers

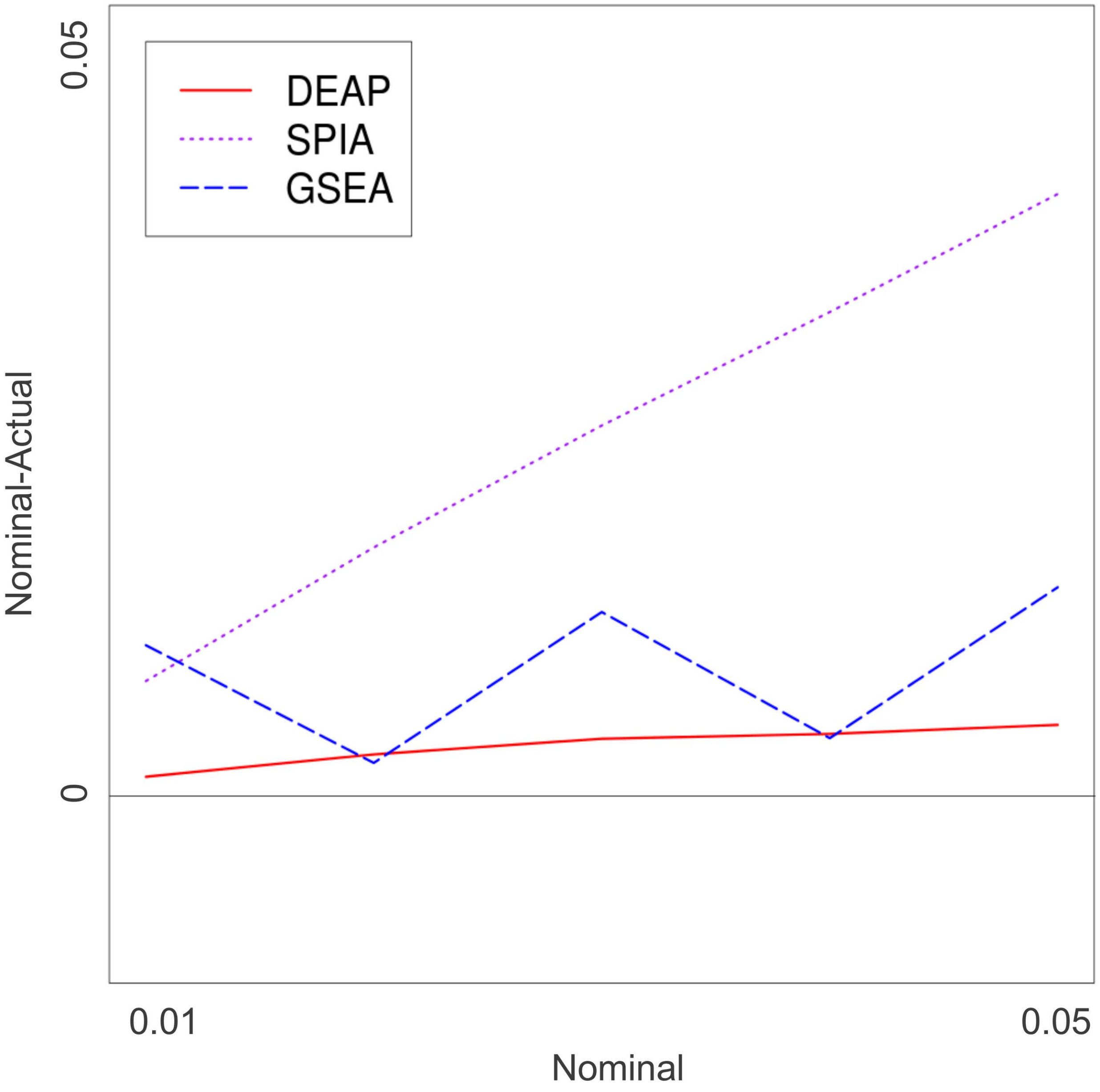

Figure 7. Type I error. The nominal value is plotted in the *x*-axis and the *y*-axis represents the nominal value minus the actual error rate. The line at Nominal-Actual = 0 represents cases where the the actual error rate perfectly corresponds with the nominal value. Values above and below this line correspond with under- and over-estimations of type I error, respectively.
doi:10.1371/journal.pcbi.1002967.g007

with patients diagnosed with COPD (*Methods: Biological data*, *Table 3*). On this data set, GSEA identified nine pathways, four of which had apparent associations with COPD. SPIA identified only one pathway with significant differential expression. DEAP identified 12 pathways and eight had literature-verified implications with COPD. Of notable clinical relevance to COPD is the inflammation mediated by chemokine and cytokine signalling pathway, which was identified only by DEAP [39].

Discussion

DEAP takes into account the graph structure of a pathway and determines the maximally expressed path. Pathway-centric analysis by DEAP is complementary to set-based analysis of other functional categories, as seen in both biological examples (*Tables 2–3*). Application of the random rotation approach allows for accurate assessment of statistical significance of the DEAP scores. On simulated data for simulated pathways, DEAP both increased power over existing approaches and accurately controlled the false positive rate. With high differential expression, this translated to a two-fold increase in the power of DEAP over GSEA and SPIA. On simulated data applied to real biological pathways, DEAP showed the strongest performance for all levels of pathway effect, variance, and sample size. Analysis of experimental transcriptomic and proteomic data indicates that DEAP identified important pathways related to a particular disease or condition where other approaches failed, specifically identifying six pathways related to interferon and eight related to COPD. Further, DEAP uniquely

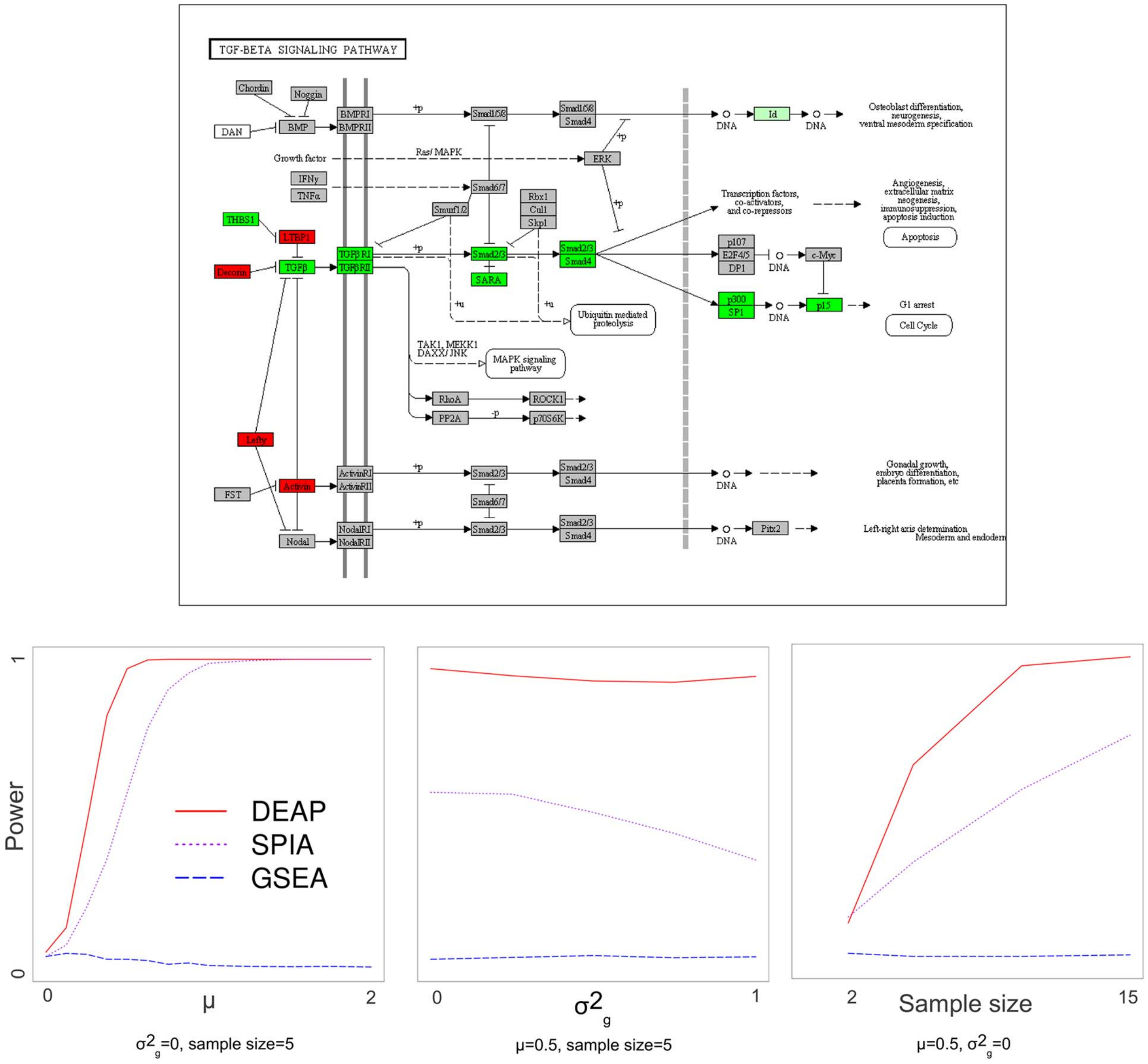

Figure 8. Simulated data on the TGFβ signalling pathway, power *vs.* pathway effect, variance, and sample size. At the top, the KEGG TGFβ signaling pathway is illustrated, with green, red, and grey nodes representing nodes whose simulated values were +μ, −μ, and 0, respectively [13]. The nodes are colored to indicate activity leading to G1 arrest in the cell cycle. At the bottom, power for detecting significant differential expression in this pathway is illustrated with respect to pathway effect, variance, and sample size. Figure adapted from http://www.genome.jp/kegg-bin/show_pathway?map04350 with permission from KEGG.
doi:10.1371/journal.pcbi.1002967.g008

identified the most expressed path of the pathway with 100% accuracy in simulated data.

Though we demonstrated DEAP on transcriptomics and proteomics studies, DEAP is widely applicable to other omics research areas (metabolomics, lipidomics, etc.) and expression technologies (next generation sequencing, RNAseq, etc.). This broad applicability extends from the flexible design of DEAP: the only required inputs are expression levels of biomolecules and corresponding pathways. Appropriate scaling of the expression levels is defined by the user. For instance, RNAseq data is very similar to spectral count proteomics data in that they are both count-based. Thus, RNAseq read counts can be used as input for DEAP in the same manner as peptide spectral counts. Further, RNA transcripts can be used in place of proteins.

To identify the most important pathways for further study, pathways can be ranked based on DEAP score significance. Specifically, future studies can be focused on the most differentially expressed paths within the pathways with the lowest false discovery rate, which can be especially beneficial when studying pathways that contain hundreds of biological compounds. Currently, DEAP is being integrated with our proteomics analysis pipeline SPIRE (http://proteinspire.org) and expression database MOPED (http://moped.proteinspire.org) [10,40] (*Table 1*). Application of DEAP to existing and future studies has the potential to discover meaningful biological patterns.

Methods

Simulated Data

Expression data (presumably on a log scale) for each gene in a pathway was simulated using a multivariate normal distribution defined in *Equation 2*:

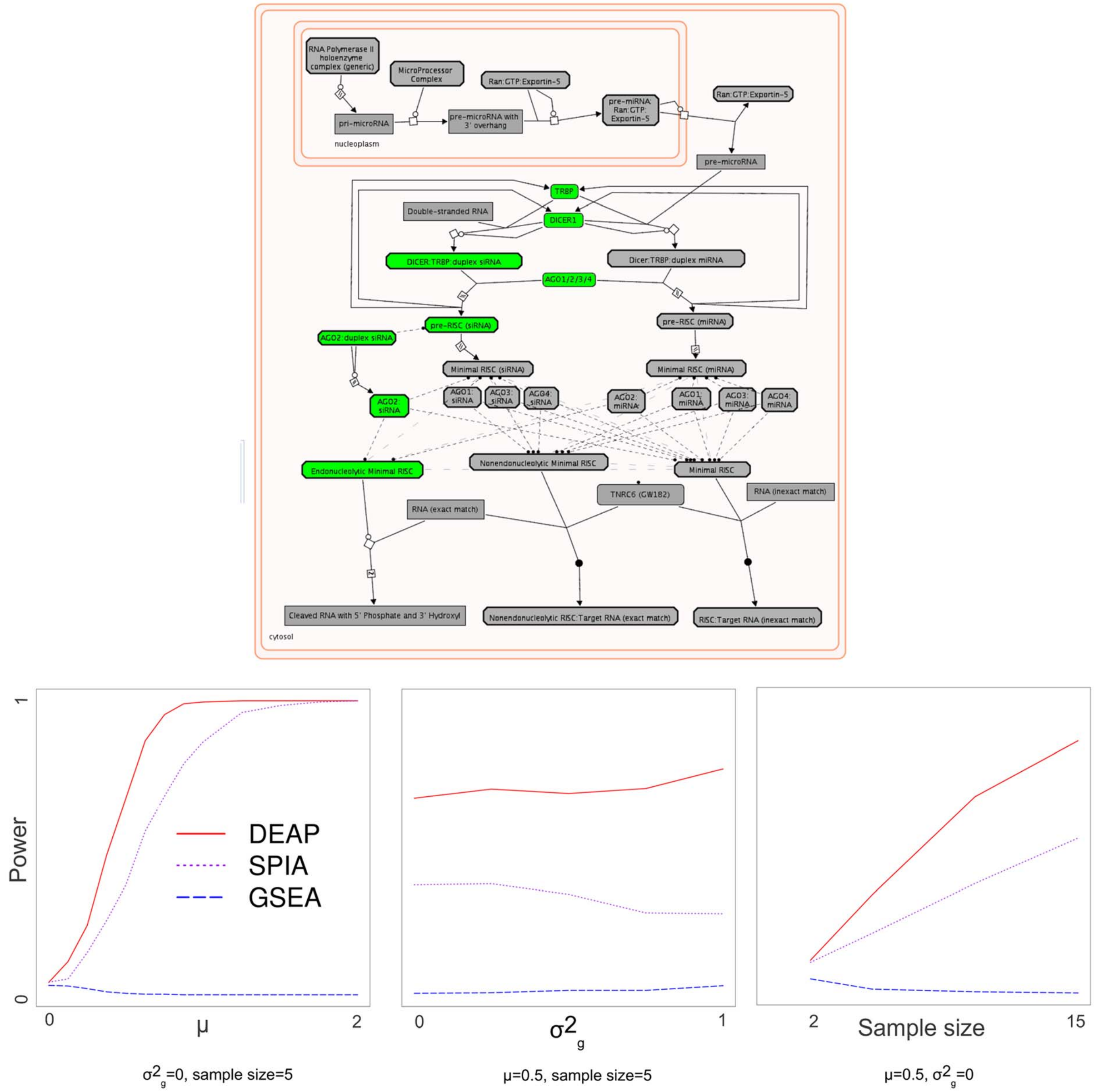

Figure 9. Simulated data on the post-transcriptional silencing by small RNAs, power *vs.* pathway effect, variance, and sample size. At the top, the Reactome post-transcriptional silencing by small RNAs pathway is illustrated, with green, red, and grey nodes representing nodes whose simulated values were +μ, −μ, and 0, respectively [14]. The nodes are colored to indicate activity leading to silencing by cleaved RNA with 5′ phosphate and 3′ hydroxyl. At the bottom, power for detecting significant differential expression in this pathway is illustrated with respect to pathway effect, variance, and sample size.
doi:10.1371/journal.pcbi.1002967.g009

$$E = d(\mu + g) + e \qquad (2)$$

In this equation d is the indicator of whether a gene is 'on' or 'off'. The value of d is 0 if the gene is 'off' and +1 if the gene is up-regulated and 'on' and −1 if the gene is down-regulated and 'on'. The value of d is determined by the predefined pathways. The variable μ is the mean of the absolute value of expression for 'on' genes and, therefore, represents the 'pathway effect'. The value of μ is held constant for each gene in the pathway and across replicate samples. The variable g is assumed to come from a normal distribution with mean 0 and variance σ^2_g. The variance σ^2_g measures how much individual gene expression deviates from the overall 'pathway effect', μ.

The value of g is randomly generated (although in many of the simulations is set to 0) for each gene in the pathway, but the same value is used for replicate samples. The variable e is assumed to come from a normal distribution with mean 0 and variance 1 and represent random variation in gene expression. The value of e is randomly generated for each combination of gene and sample. The simulations varied the values μ, σ^2_g, and the sample size (number of independent samples of pathway data). R scripts were used to generate the simulated data [41].

Five diverse pathways were specifically created to test the efficacy of identification by different scoring methods (*Figure 3*). Gray colored nodes had unaltered values from a standard normal distribution. Nodes labelled as green and red were sampled with μ

Table 2. Results from Interferon microarray data analysis using GSEA, SPIA, and DEAP [36,37].

Pathway	GSEA	SPIA	DEAP
Interferon gamma signaling			S
JAK STAT signaling [47]			S
PDGF signalling [48]			S
Notch signalling [38]			S
Interleukin signalling [49]			S
General transcription regulation [50]			S
Beta1 adrenergic receptor signalling			S
Histamine H1 receptor mediated signalling [51]	S		
Oxytocin receptor mediated signalling [52]	S		
Thyrotropin releasing hormone receptor signalling [53]	S		
Integrin signalling [54]	S		
Arginine biosynthesis [55]	S		
Parkinson disease	S		

An *S* indicates pathway differential expression significance of $p <= 0.05$. Italic text indicates previously discovered associations between the pathway and interferon. Non-italic text indicates no known associations between the pathway and interferon. Specific p-values are found in Table S5.
doi:10.1371/journal.pcbi.1002967.t002

values of +X and −X, respectively, where X was a positive number. Simulated data and pathways are available on Dryad: doi:10.5061/dryad.qh1pg.

Biological Data

Microarray data from a study of cells treated with interferon were acquired from the Gene Expression Omnibus (GDS3126) [6]. The sample was taken from radio-resistant tumors following treatment with a mixture of interferons [36,37]. It was hypothesized that interferon and biochemically-related pathways would be stimulated in this data set. The expression value was the logarithm of the case/control ratio. Though microarrays measure mRNA expression, the pathways represent information in terms of proteins. Therefore, the gene identifiers in the microarray data were mapped to UniProt protein identifiers using the UniProt website [42]. Handling the one-to-many relationship of genes and proteins is discussed below (see *Methods: DEAP*). When duplicate probes existed for the same gene, the expression value utilized for the gene was the arithmetic mean of these probes.

The COPD proteomics data can be found at PeptideAtlas (raw data) [7] and MOPED (processed data) [10] (moped.proteinspire.org). We analyzed data from CD4 and CD8 T-lymphocytes. The control patients were healthy smokers, with an average FEV1/FVC of 82.5%. Case patients had been medically diagnosed with COPD and had an average FEV1/FVC of 42.0%. A total of 10 cases and 10 controls were utilized in this analysis. Additional experimental details can be found associated with the PeptideAtlas accession numbers in *Table S3*. On MOPED, data is stored under the experimental name "steffan_copd." The tandem mass spectrometry data were analyzed through SPIRE with the parameters in *Table S4* [40]. Protein expression was measured by the number of peptide spectral matches identified for each

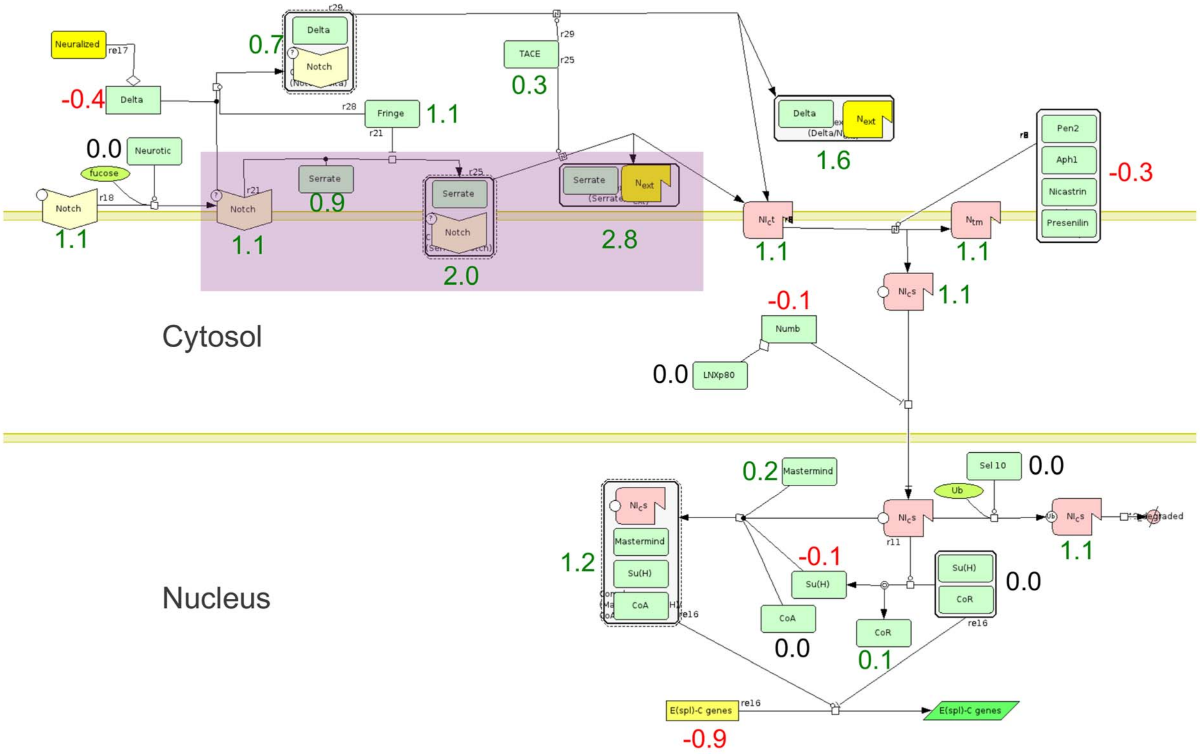

Figure 10. Maximally differentially expressed path identification. Maximally differentially expressed path identification by DEAP on the Notch signalling pathway. Pathway image is from PANTHER [11,46]. The path shaded in purple was identified by DEAP as the most differentially expressed. Numerical values are log-expression ratios from the Interferon microarray study [36,37].
doi:10.1371/journal.pcbi.1002967.g010

Table 3. Results from the COPD proteomics data analysis using GSEA, SPIA, and DEAP (*Methods: Biological data*).

Pathway	GSEA	SPIA	DEAP
Integrin signalling [56]		S	S
Interleukin signalling [57]			S
Inflammation mediated by chemokine and cytokine signalling [39]			M
Heme biosynthesis [58]	M		M
Ras pathway [59]			M
Plasminogen activating cascade [60]			M
De novo purine biosynthesis [61]			M
Ubiquitin proteasome [62]			M
Heterotrimeric G protein signalling, rod outer segment			M
PDGF signalling			M
De novo pyrimidine ribonucleotide biosynthesis*			M
De novo pyrimidine deoxyribonucleotide biosynthesis*			M
Muscarinic acetylcholine receptor signalling [63]	S		
Nicotonic acetylcholone receptor signalling [64]	S		
Blood coagulation [65]	M		
DNA replication	S		
Circadian clock system	S		
Asparagine and aspartate biosynthesis	M		
Salvage pyrimidine ribonucleotides*	M		
Arginine biosynthesis*	M		

An *S* and *M* indicate pathway differential expression significance of $p< = 0.05$ and marginal significance of $p< = 0.1$, respectively. Italic text indicates previously discovered associations between the pathway and COPD. Non-italic text indicates no known associations between the pathway and COPD.
*Arginine and pyrimidine are used therapeutically to treat COPD. Specific p-values are found in Table S6.
doi:10.1371/journal.pcbi.1002967.t003

protein normalized by the total number of spectra in the sample. For pathway analysis, we used the difference between the log normalized expression values.

Pathway data were downloaded from the PANTHER database [11]. A total of 165 pathways downloaded in SBML format from PANTHER pathway version 3.01. PANTHER pathways contain information about proteins, biochemicals, and other substrates. For the purposes of data interpretation, the pathways were broken into their protein components using an internally developed python script where connections of proteins through biochemical substrates were maintained as protein-protein interactions PANTHER's internal identifiers were mapped to UniProt identifiers. Ultimately, parsing of the PANTHER pathway database resulted in a graph structure in which each node represented a set of proteins that act as a set of reactants and/or products. Inhibitory or catalytic edges between two sets of proteins were determined as detailed in PANTHER.

Rotation Sampling

We used random rotation approach to estimate the null distribution of the test statistics and compute the p-values [32]. Rotation testing has been used recently in gene set analysis as an alternative to permutation and parametric tests [33,34]. Rotation tests have an advantage over permutation tests in that they produce reasonable results for small sample sizes and complex experimental designs. Rotation testing assumes that pathway and set data come from independent random samples of a multivariate normal distribution with mean zero under the null hypothesis. A rotation test is carried out by multiplying the original data by a random rotation matrix, calculating the test statistic, and repeating the procedure to generate a null distribution. Adjustments for an overall mean, covariates, or blocking factors are handled by performing the rotations of an orthogonal projection of the original data on to the residual space from a linear model and then transforming the rotated data back. A random rotation matrix was generated by first generating a matrix X of standard normal random variables and then taking the rotation matrix to be the orthogonal matrix Q from the QR decomposition of X. Scripts to carry out rotation testing were written using the R programming language and are available in *File S3*, released under the GNU Lesser General Public License v3.0. The user is able to input a custom design matrix which accounts for complex experimental designs with multiple conditions and replicates.

DEAP Algorithm

Given: a current edge, all other edges in graph, expression values for all proteins:

For single channel (unpaired) data, define $E(x)$ to be the difference between the logarithm of the arithmetic mean of expression values associated with protein x in the two conditions.

For two channel (paired) data, define $E(x)$ to be the arithmetic mean of the log expression ratio(s) associated with protein x.

The recursive function operates as follows:

1. Recursively examine all edges in the pathway set whose reactant node is the current edge's product node.

   a) If there are no such edges, set $max_{recursive}$ and $min_{recursive}$ as $\sum_{y \in products} E(y)$ where $y \in products$ refers to each protein, y, contained in the edge's products.

b) Otherwise, define $max_{recursive}$ and $min_{recursive}$ as the maximum and minimum scores, respectively, returned by the recursive function.

2. Assign max_{score} and min_{score} as the maximum and minimum, respectively of: $\sum_{z \in reactants} E(z) + T(edge) * max_{recursive}$ and $\sum_{z \in reactants} E(z) + T(edge) * min_{recursive}$ where $T(edge)$ is the multiplier associated with the edge type (−1 or 1 for inhibition or catalysis, respectively) and $z \in reactants$ refers to each protein, z, contained in the edge's reactants.
3. Return the maximum of $\{max_{score}, 0\}$ and the minimum of $\{min_{score}, 0\}$.

DEAP Statistics

In DEAP, the maximum order (by absolute value) path is used to test the null hypothesis about the expression of the entire pathway. This claim, that the expression of one path answers questions about the expression of the pathway, is justified on two levels.

On a biological level, significant fluctuations in activity do not require differential expression of an entire pathway. For example, in *Figure 1*, $A_3A_4A_7$ represents a path with similar expression levels that proceeds all the way from reactants to products, a pattern that seems to be significant.

From a logical perspective, consider a pathway, P, as the union of all paths of the pathway, $P_1, P_2, \ldots, P_K$. Each path is completely defined by its set of edges. Note that the k-paths are not entirely disjoint in the sense that some paths might share the nodes and the edges. However, we require each path to have a distinct set of edges. To test the hypothesis of a differentially expressed pathway requires testing whether any of the constituent paths is differentially expressed. This corresponds to testing the family of k-null hypothesis. To control the family wise error rate, we use a maximum order statistic, since the probability of making at least one incorrect decision under the null is equivalent to the probability of the maximum order statistic exceeding the threshold.

To approximate a null distribution of the test statistic, s^*, we performed n rotations of the data. For each rotation sample, we recompute the DEAP score, s_i. The p-value is calculated as a proportion of scores that are at least as extreme as the observed score, the proportion of simulated DEAP scores whose value are greater than or equal to the observed DEAP score:

$$p = \frac{\#(s_i \geq \mathrm{s}*)}{n}$$

DEAP Implementation

The DEAP algorithm was implemented to allow for efficient computation.

By maintaining global maximum and minimum values and updating their values as the recursive function proceeds, it is not necessary to examine all paths of the graph independently. Rather, we can initialize DEAP score calculations only at leaf edges, which have no upstream edges pointing to any proteins in their reactant set. To ensure that closed cycles are not missed, we track the edges which have been visited and examine additional edges until the difference of the complete edge set and the already visited edge set is empty. This greatly reduces the number of calculations per graph.

Once the recursive function has returned a maximum and minimum score for a particular edge, that score will remain constant regardless of the preceding edge except in the case of cycles (see paragraph below). Therefore, we use a dictionary mapping edges to maximum and minimum scores to prevent duplicative score calculations. After this implementation, score calculations that took several hours on particularly complex pathway structures completed in seconds.

In the case of cycles, scores may be dependent on the node of the cycle which is examined first. For these cycles, our current implementation represents a heuristic estimator rather than the exact optimal solution. Bidirectional edges are subject to this same limitation as they are equivalent to a two node cycle. Implementations that determined the exact optimal solution were prohibitively slow for practical application. Except in edge cases, the heuristic implementation will provide approximations of sufficient quality to identify significant patterns of differential expression.

Every DEAP score calculation is independent of other DEAP score calculations, so we set up processing for multi-threading. For example, on a 64-bit Intel Core i7-2720QM CPU with 8GB RAM, speed improvements of approximately 4-fold were noted for the score calculation process. Specific running time is highly dependent on expression data set size, experimental design, pathway complexity, and number of rotation testing iterations. Running DEAP on 90 simulated data files each with 10 samples, 1000 proteins, 1000 pathways, and performing 100 data rotations took 72 minutes when multi-threaded and 260 minutes when performed on a single thread.

The function tracks edges that have already been examined in a particular recursive cycle to prevent entrance into infinite loops in cyclical pathways. To control for duplicate protein identifiers, summations over the products and reactants were performed on the set of unique expression values rather than for every identifier. For example, if protein A and protein B both had expression levels of 1.743 and were both in the same protein set, then it was assumed they were the result of data duplication and 1.743 was only added to the score once. This duplication elimination was implemented primarily due to issues arising from redundant protein identifiers and potential mRNA translation into multiple proteins. For instance, the five UniProt identifiers for variants of Histone H3 (Q6NXT2, P68431, Q16695, Q71DI3, and P84243) are included in the same PANTHER pathway unit and share near identical protein sequences, so their proteomic and transcriptomic identification will be duplicated.

The algorithm was implemented in Python and is available in *File S3*, released under the GNU Lesser General Public License v3.0.

Biological Data Validation

Accuracy of pathway associations with experimental conditions were validated using a Google Scholar literature search. The literature search was performed by searching Google Scholar (http://scholar.google.com) for a combination of the pathway name and details of the experimental condition. We continued searching Google Scholar until satisfied that the association was confirmed or felt reasonably certain that there was not yet a literature confirmed association. Once a literature association was confirmed, the most pertinent reference was retained and cited in this manuscript.

Assumptions

The DEAP approach is based on the following fundamental assumptions:

1. The user provides expression values using an appropriately scaled metric that represent meaningful information. DEAP is independent from the calculation of individual gene expression values, with the stipulation that data be numeric, where positive values represent over-expression and negative values represent under-expression. For example, microarray expression data have been shown to be scale free in nature, so a logarithm scaled expression ratio was input to the DEAP algorithm.
2. Existing pathway knowledge is sufficient to make meaningful statements. Though pathways currently contain only a fraction of all proteins, this approach makes no attempt to expand that coverage [43].

Existing Approaches

The GSEAlm package for the R Project, available through BioConductor, was utilized to perform GSEA analysis [44]. Pathways were transformed into a gene set matrix and multi-sample expression data were loaded appropriately. Since GSEA performs test for up- and down-regulation independently, the minimum of these two values was taken and multiplied by two to adjust for a two-tail test.

SPIA analysis was performed using the SPIA package for the R Project, available through BioConductor [45]. To convert the pathways into the SPIA format, inhibitory and catalytic relationships were formatted into the inhibition and activation matrices, respectively. Since the SPIA implementation only allowed input of single expression ratios, the arithmetic mean of expression values for each protein was input into SPIA.

Supporting Information

Figure S1 Power vs. pathway effect for all 7 approaches for simulated data on simulated pathways.
(TIFF)

Figure S2 Power vs. sample variance for all 7 approaches for simulated data on simulated pathways.
(TIFF)

Figure S3 Power vs. sample size for all 7 approaches for simulated data on simulated pathways.
(TIFF)

Figure S4 Type I error for all 7 approaches for simulated data on simulated pathways.
(TIFF)

Figure S5 Type I error for simulated data on biological pathways.
(TIFF)

Figure S6 Power vs. pathway effect, sample size, and variance for all 7 approaches for simulated data on the KEGG TGFβ signalling pathway. Figure adapted from http://www.genome.jp/kegg-bin/show_pathway?map04350 with permission from KEGG
(TIFF)

Figure S7 Power vs. pathway effect, sample size, and variance for all 7 approaches for simulated data on the Reactome post-transcriptional silencing by small RNAs pathway.
(TIFF)

File S1 DEAP results on CF data.
(TXT)

File S2 DEAP results on COPD data.
(TXT)

File S3 Archive file of DEAP source code licensed under the GNU Lesser General Public License v3.0.
(ZIP)

Table S1 Provides a summary of approaches to pathway analysis, rationale for their inclusion, and summary of the results for simulated data.
(DOC)

Table S2 Decision justifications.
(DOC)

Table S3 PeptideAtlas accession numbers for COPD study.
(DOC)

Table S4 COPD search parameters using SPIRE.
(DOC)

Table S5 p-values for *Table 2: Results from Interferon microarray data analysis using GSEA, SPIA, and DEAP.*
(XLS)

Table S6 p-values for *Table 3: Results from the COPD proteomics data analysis using GSEA, SPIA, and DEAP.*
(XLS)

Acknowledgments

We sincerely thank Bill Broomall, Natali Kolker, Elizabeth Stewart, Chris Moss, Randy Salamon, Charles Smith, Dean Welch, and Greg Yandl from Seattle Children's and Carl Burch and Gabriel Ferrer from Hendrix College for their critical comments and insight into this research.

Author Contributions

Conceived and designed the experiments: WAH RH LS DC EK. Performed the experiments: WAH. Analyzed the data: WAH RH LS DC EK. Contributed reagents/materials/analysis tools: WAH RH LS. Wrote the paper: WAH RH LS DC EK.

References

1. Pennisi E (2011) Will Computers Crash Genomics? Science 331: 666–668. doi:10.1126/science.331.6018.666.
2. Science Staff (2011) Challenges and Opportunities. Science 331: 692–693. doi:10.1126/science.331.6018.692.
3. Gough NR, Yaffe MB (2011) Focus Issue: Conquering the Data Mountain. Science Signaling 4: eg2–eg2. doi:10.1126/scisignal.2001871.
4. Kolker E, Stewart E, Ozdemir V (2012) Opportunities and Challenges for the Life Sciences Community. OMICS: A Journal of Integrative Biology 16: 138–147. doi:10.1089/omi.2011.0152.
5. Ozdemir V, Rosenblatt DS, Warnich L, Srivastava S, Tadmouri GO, et al. (2011) Towards an Ecology of Collective Innovation: Human Variome Project (HVP), Rare Disease Consortium for Autosomal Loci (RaDiCAL) and Data-Enabled Life Sciences Alliance (DELSA). Current Pharmacogenomics and Personalized Medicine 9: 234–251.
6. Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30: 207–210.
7. Desiere F, Deutsch EW, Nesvizhskii AI, Mallick P, King NL, et al. (2005) Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. Genome Biol 6: R9. doi:10.1186/gb-2004-6-1-r9.
8. Vizcaíno JA, Côté R, Reisinger F, Foster JM, Mueller M, et al. (2009) A guide to the Proteomics Identifications Database proteomics data repository. Proteomics 9: 4276–4283. doi:10.1002/pmic.200900402.
9. Sherlock G, Hernandez-Boussard T, Kasarskis A, Binkley G, Matese JC, et al. (2001) The Stanford Microarray Database. Nucleic Acids Res 29: 152–155.
10. Kolker E, Higdon R, Haynes W, Welch D, Broomall W, et al. (2012) MOPED: Model Organism Protein Expression Database. Nucleic Acids Res 40: D1093–1099. doi:10.1093/nar/gkr1177.

11. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, et al. (2003) PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 13: 2129–2141. doi:10.1101/gr.772403.
12. Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, et al. (2008) The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res 36: D623–631. doi:10.1093/nar/gkm900.
13. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, et al. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 27: 29–34.
14. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, et al. (2005) Reactome: a knowledgebase of biological pathways. Nucleic Acids Res 33: D428–432. doi:10.1093/nar/gki072.
15. Mootha VK, Lindgren CM, Eriksson K-F, Subramanian A, Sihag S, et al. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 34: 267–273. doi:10.1038/ng1180.
16. 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 USA 102: 15545–15550. doi:10.1073/pnas.0506580102.
17. Kim S-Y, Volsky DJ (2005) PAGE: parametric analysis of gene set enrichment. BMC Bioinformatics 6: 144. doi:10.1186/1471-2105-6-144.
18. Jiang Z, Gentleman R (2007) Extensions to gene set enrichment. Bioinformatics 23: 306–313. doi:10.1093/bioinformatics/btl599.
19. Cha S, Imielinski MB, Rejtar T, Richardson EA, Thakur D, et al. (2010) 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 9: 2529–2544. doi:10.1074/mcp.M110.000398.
20. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, et al. (2005) Discovering statistically significant pathways in expression profiling studies. Proc Natl Acad Sci USA 102: 13544–13549. doi:10.1073/pnas.0506577102.
21. Rahnenführer J, Domingues FS, Maydt J, Lengauer T (2004) Calculating the Statistical Significance of Changes in Pathway Activity From Gene Expression Data. Statistical Applications in Genetics and Molecular Biology 3. Available:http://www.degruyter.com/view/j/sagmb.2004.3.1/sagmb.2004.3.1.1055/sagmb.2004.3.1.1055.xml. Accessed 11 December 2012.
22. Goeman JJ, Bühlmann P (2007) Analyzing gene expression data in terms of gene sets: methodological issues. Bioinformatics 23: 980–987. doi:10.1093/bioinformatics/btm051.
23. 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.
24. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, et al. (2009) A novel signaling pathway impact analysis. Bioinformatics 25: 75–82. doi:10.1093/bioinformatics/btn577.
25. Bankhead A 3rd, Sach I, Ni C, LeMeur N, Kruger M, et al. (2009) Knowledge based identification of essential signaling from genome-scale siRNA experiments. BMC Syst Biol 3: 80. doi:10.1186/1752-0509-3-80.
26. Zhao J, Gupta S, Seielstad M, Liu J, Thalamuthu A (2011) Pathway-based analysis using reduced gene subsets in genome-wide association studies. BMC Bioinformatics 12: 17. doi:10.1186/1471-2105-12-17.
27. Hung J-H, Whitfield TW, Yang T-H, Hu Z, Weng Z, et al. (2010) Identification of functional modules that correlate with phenotypic difference: the influence of network topology. Genome Biol 11: R23. doi:10.1186/gb-2010-11-2-r23.
28. Thomas R, Gohlke JM, Stopper GF, Parham FM, Portier CJ (2009) Choosing the right path: enhancement of biologically relevant sets of genes or proteins using pathway structure. Genome Biol 10: R44. doi:10.1186/gb-2009-10-4-r44.
29. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, et al. (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics 26: i237–i245. doi:10.1093/bioinformatics/btq182.
30. Shojaie A, Michailidis G (2009) Analysis of Gene Sets Based on the Underlying Regulatory Network. Journal of Computational Biology 16: 407–426. doi:10.1089/cmb.2008.0081.
31. Khatri P, Sirota M, Butte AJ (2012) Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLoS Computational Biology 8: e1002375. doi:10.1371/journal.pcbi.1002375.
32. Langsrud O (2005) Rotation tests. Statistics and Computing 15: 53–60. doi:10.1007/s11222-005-4789-5.
33. Wu D, Lim E, Vaillant F, Asselin-Labat M-L, Visvader JE, et al. (2010) ROAST: rotation gene set tests for complex microarray experiments. Bioinformatics 26: 2176–2182. doi:10.1093/bioinformatics/btq401.
34. Dørum G, Snipen L, Solheim M, Saebø S (2009) Rotation testing in gene set enrichment analysis for small direct comparison experiments. Stat Appl Genet Mol Biol 8: Article34. doi:10.2202/1544-6115.1418.
35. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci USA 100: 9440–9445. doi:10.1073/pnas.1530509100.
36. Khodarev NN, Minn AJ, Efimova EV, Darga TE, Labay E, et al. (2007) Signal transducer and activator of transcription 1 regulates both cytotoxic and prosurvival functions in tumor cells. Cancer Res 67: 9214–9220. doi:10.1158/0008-5472.CAN-07-1019.
37. Khodarev NN, Beckett M, Labay E, Darga T, Roizman B, et al. (2004) STAT1 is overexpressed in tumors selected for radioresistance and confers protection from radiation in transduced sensitive cells. Proc Natl Acad Sci USA 101: 1714–1719. doi:10.1073/pnas.0308102100.
38. Hu X, Ivashkiv LB (2009) Cross-regulation of Signaling Pathways by Interferon-γ: Implications for Immune Responses and Autoimmune Diseases. Immunity 31: 539–550. doi:10.1016/j.immuni.2009.09.002.
39. Fuke S, Betsuyaku T, Nasuhara Y, Morikawa T, Katoh H, et al. (2004) Chemokines in bronchiolar epithelium in the development of chronic obstructive pulmonary disease. Am J Respir Cell Mol Biol 31: 405–412. doi:10.1165/rcmb.2004-0131OC.
40. Kolker E, Higdon R, Welch D, Bauman A, Stewart E, et al. (2011) SPIRE: Systematic protein investigative research environment. J Proteomics 75: 122–126. doi:10.1016/j.jprot.2011.05.009.
41. R Development Core Team (n.d.) R: A language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing. Available:http://www.R-project.org/.
42. UniProt Consortium (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res 40: D71–75. doi:10.1093/nar/gkr981.
43. Wren JD (2009) A global meta-analysis of microarray expression data to predict unknown gene functions and estimate the literature-data divide. Bioinformatics 25: 1694–1701. doi:10.1093/bioinformatics/btp290.
44. Oron A, Gentleman R (n.d.) GSEAlm: Linear model toolset for Gene Set Enrichment Analysis. R project.
45. Tarca AL, Kathri P, Draghici S (2011) SPIA: Signaling Pathway Impact Analysis (SPIA) using combined evidence of pathway over-representation and unusual signaling perturbations. R project. Available:http://bioinformatics.oxfordjournals.org/cgi/reprint/btn577v1.
46. Huaiyu M, Thomas P (2009) PANTHER Pathway: An Ontology-Based Pathway Database Coupled with Data Analysis Tools. Protein Networks and Pathway Analysis. Methods in Molecular Biology 563: 123–140.
47. David M, Chen HE, Goelz S, Larner AC, Neel BG (1995) Differential regulation of the alpha/beta interferon-stimulated Jak/Stat pathway by the SH2 domain-containing tyrosine phosphatase SHPTP1. Mol Cell Biol 15: 7050–7058.
48. Suzuki H, Shibano K, Okane M, Kono I, Matsui Y, et al. (1989) Interferon-gamma modulates messenger RNA levels of c-sis (PDGF-B chain), PDGF-A chain, and IL-1 beta genes in human vascular endothelial cells. Am J Pathol 134: 35–43.
49. Gu Y (1997) Activation of Interferon-gamma Inducing Factor Mediated by Interleukin-1beta Converting Enzyme. Science 275: 206–209. doi:10.1126/science.275.5297.206.
50. Hartman SE (2005) Global changes in STAT target selection and transcription regulation upon interferon treatments. Genes & Development 19: 2953–2968. doi:10.1101/gad.1371305.
51. Krouwels FH, Hol BE, Lutter R, Bruinier B, Bast A, et al. (1998) Histamine affects interleukin-4, interleukin-5, and interferon-gamma production by human T cell clones from the airways and blood. Am J Respir Cell Mol Biol 18: 721–730.
52. Spencer TE (1996) Ovine interferon tau suppresses transcription of the estrogen receptor and oxytocin receptor genes in the ovine endometrium. Endocrinology 137: 1144–1147. doi:10.1210/en.137.3.1144.
53. Valyasevi RW (2001) Effect of Tumor Necrosis Factor-, Interferon-, and Transforming Growth Factor- on Adipogenesis and Expression of Thyrotropin Receptor in Human Orbital Preadipocyte Fibroblasts. Journal of Clinical Endocrinology & Metabolism 86: 903–908. doi:10.1210/jc.86.2.903.
54. Defilippi P, Truffa G, Stefanuto G, Altruda F, Silengo L, et al. (1991) Tumor necrosis factor alpha and interferon gamma modulate the expression of the vitronectin receptor (integrin beta 3) in human endothelial cells. J Biol Chem 266: 7638–7645.
55. Drapier J-C, Wietzerbin J, Hibbs JB (1988) Interferon-γ and tumor necrosis factor induce the L-arginine-dependent cytotoxic effector mechanism in murine macrophages*. European Journal of Immunology 18: 1587–1592. doi:10.1002/eji.1830181018.
56. Araya J, Cambier S, Markovics JA, Wolters P, Jablons D, et al. (2007) Squamous metaplasia amplifies pathologic epithelial-mesenchymal interactions in COPD patients. Journal of Clinical Investigation 117: 3551–3562. doi:10.1172/JCI32526.
57. Imaoka H, Hoshino T, Takei S, Kinoshita T, Okamoto M, et al. (2008) Interleukin-18 production and pulmonary function in COPD. Eur Respir J 31: 287–297. doi:10.1183/09031936.00019207.
58. Tsoumakidou M, Tzanakis N, Chrysofakis G, Siafakas NM (2005) Nitrosative stress, heme oxygenase-1 expression and airway inflammation during severe exacerbations of COPD. Chest 127: 1911–1918. doi:10.1378/chest.127.6.1911.
59. Anderson D, Hughes JA, Cebulska-Wasilewska A, Nizankowska E, Graca B (1998) Ras p21 protein levels in human plasma from patients with chronic obstructive pulmonary disease (COPD) compared with lung cancer patients and healthy controls. Mutat Res 403: 229–235.
60. Xiao W, Tong W, Ma D (2006) [Higher levels of urokinase plasminogen activator system components in the airways of chronic obstructive pulmonary disease patients]. Zhonghua Jie He He Hu Xi Za Zhi 29: 723–726.
61. Esther CR Jr, Lazaar AL, Bordonali E, Qaqish B, Boucher RC (2011) Elevated airway purines in COPD. Chest 140: 954–960. doi:10.1378/chest.10-2471.
62. Ottenheijm CAC, Heunks LMA, Li Y-P, Jin B, Minnaard R, et al. (2006) Activation of the ubiquitin-proteasome pathway in the diaphragm in chronic

obstructive pulmonary disease. Am J Respir Crit Care Med 174: 997–1002. doi:10.1164/rccm.200605-721OC.

63. Gosens R, Zaagsma J, Meurs H, Halayko AJ (2006) Muscarinic receptor signaling in the pathophysiology of asthma and COPD. Respir Res 7: 73. doi:10.1186/1465-9921-7-73.

64. Zhang J, Summah H, Zhu Y, Qu J-M (2011) Nicotinic acetylcholine receptor variants associated with susceptibility to chronic obstructive pulmonary disease: a meta-analysis. Respiratory Research 12: 158. doi:10.1186/1465-9921-12-158.

65. Undas A, Kaczmarek P, Sladek K, Stepien E, Skucha W, et al. (2009) Fibrin clot properties are altered in patients with chronic obstructive pulmonary disease. Beneficial effects of simvastatin treatment. Thromb Haemost 102: 1176–1182. doi:10.1160/TH09-02-0118.