NIH Public Access
Author Manuscript
J Biomed Inform. Author manuscript; available in PMC 2011 December 1.

Published in final edited form as:
J Biomed Inform. 2010 December ; 43(6): 932–944. doi:10.1016/j.jbi.2010.07.001.

Independent component analysis: mining microarray data for fundamental human gene expression modules

Jesse M. Engreitz[1], **Bernie J. Daigle Jr.**[2], **Jonathan J. Marshall**[1], and **Russ B. Altman**[1,2,*]
[1]Department of Bioengineering, Stanford University, Stanford, CA 94305, USA
[2]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA

Abstract

As public microarray repositories rapidly accumulate gene expression data, these resources contain increasingly valuable information about cellular processes in human biology. This presents a unique opportunity for intelligent data mining methods to extract information about the transcriptional modules underlying these biological processes. Modeling cellular gene expression as a combination of functional modules, we use independent component analysis (ICA) to derive 423 fundamental components of human biology from a 9,395-array compendium of heterogeneous expression data. Annotation using the Gene Ontology (GO) suggests that while some of these components represent known biological modules, others may describe biology not well characterized by existing manually-curated ontologies. In order to understand the biological functions represented by these modules, we investigate the mechanism of the preclinical anticancer drug parthenolide (PTL) by analyzing the differential expression of our fundamental components. Our method correctly identifies known pathways and predicts that N-glycan biosynthesis and T-cell receptor signaling may contribute to PTL response. The fundamental gene modules we describe have the potential to provide pathway-level insight into new gene expression datasets.

Keywords

microarrays; independent component analysis; data mining; parthenolide; gene modules

1 Introduction

The wide use of high-throughput DNA microarray technology promises to provide an increasingly detailed view of the human transcriptome in many different contexts. As experimentalists continue to sample a variety of biological conditions and cell types, public gene expression repositories grow exponentially: the Gene Expression Omnibus (GEO), the largest of these databases, now contains over 400,000 individual microarrays

© 2010 Elsevier Inc. All rights reserved.

[*]**Send proofs to:** Dr. Russ B. Altman, 318 Campus Drive S172, MC: 5444, Stanford, CA 94305-5444, Tel: (650) 725-3394, Fax: (650) 723-8544, russ.altman@stanford.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Availability
The fundamental components, gene lists for our fundamental gene expression modules, and the R implementation of our method are available at http://www.simtk.org/home/fcanalysis.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

(www.ncbi.nlm.nih.gov/geo/). Although biomedical researchers typically design microarray experiments to explore specific biological contexts, meta-analyses that integrate data from multiple experiments have the potential to reveal relationships that are not accessible through any individual dataset [1–10]. One critical question in the post-genomic era is the identification of sets of functionally related genes that correspond to a particular biological process. While curated databases such as the Gene Ontology (GO) group genes with related functions, these resources are necessarily incomplete and prone to human error [11]. Furthermore, GO categories do not necessarily correspond to co-regulated transcriptional units [12]. Computational methods based on co-expression, therefore, have the potential to enhance the identification of genes that contribute to the same biological processes [1,7,13]. Given the wide sampling of diverse cellular conditions, a meta-analysis of human gene expression data should yield information about many of the key pathways and processes in human biology.

Methods that extract gene expression modules (subsets of genes that are co-regulated across subsets of conditions) provide a means for identifying functionally related genes from microarray data [13–21]. Popular methods for this analysis, such as hierarchical clustering, k-means clustering, and self-organizing maps (SOMs), identify disjoint sets of genes that are co-expressed throughout a particular dataset. Because these approaches require that genes belong to only one transcriptional module, however, they do not capture an important feature of gene regulation: it is common for the same proteins to serve different functions depending on biological context [22]. Furthermore, simple clustering approaches may be inappropriate for large-scale gene expression meta-analysis because they capture global signals in the compendium and exclude more specific signals [23]. These challenges have led to methods like biclustering that simultaneously cluster subsets of genes over subsets of conditions [23–25].

For a more biologically relevant model for gene expression, researchers often turn to matrix decomposition methods [19,25–32]. Gene expression data lends itself to representation as a matrix, with columns corresponding to genes and rows corresponding to experiments. Like biclustering, matrix decomposition offers an improvement over traditional clustering methods by modeling microarray data as a combination of biological modules that can share genes. Furthermore, matrix decomposition provides a generative model that allows for unified analysis of multiple datasets. One common method, principal component analysis (PCA), provides a representation of microarray data in terms of a set of linearly uncorrelated axes [26,27]. Each axis, or principal component, explains the maximal variance not represented by the previous, orthogonal components. While PCA can identify interesting biological information, its linear transformation involves only second order statistics; like clustering methods based on co-expression, PCA may miss more complex relationships between genes [31]. Furthermore, the application of PCA to biological data assumes that gene expression follows a multivariate normal distribution. Recent studies demonstrate that microarray gene expression measurements typically follow a super-Gaussian distribution [33,34], suggesting that PCA applies a model that does not accurately reflect our knowledge of biology.

Independent component analysis (ICA), in contrast, provides a more biologically plausible model for gene expression data by assuming non-Gaussian data distributions. A blind source separation algorithm, ICA models observations as a linear combination of latent feature variables, or components, which are chosen to be as *statistically independent* as possible. For microarray data, observations consist of microarray gene expression measurements, and independent components are interpreted to be transcriptional modules that often correspond to specific biological processes [28]. ICA has proven successful in a variety of biological inquiries, including identifying oscillating regulatory modules in yeast cell cycle data [28],

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

investigating tumor-related pathways [35–39], classifying disease datasets [40,41], characterizing transcriptional regulators [42,43], identifying disease-specific biomarkers [44], and examining response to bacterial infection [45]. Furthermore, ICA outperforms PCA and other unsupervised methods in identifying co-regulated and biologically relevant gene modules in diverse datasets [31,39]. Although ICA produces stochastic component estimates with different initial conditions, clustering and averaging the results from multiple runs can yield robust independent component estimates [38,45]. While previous applications of ICA to microarray data have used at most hundreds of experiments, no large-scale meta-analysis to identify fundamental gene modules has been attempted.

Compared with resources such as the Gene Ontology, ICA provides a data-driven method for exploring functional relationships and grouping genes into transcriptional modules. We apply ICA to a large microarray compendium initially comprised of 9,395 microarrays representing a diverse set of experimental conditions and cell types. We identify 423 fundamental components (FCs) of human biology, and show that these components yield gene expression modules with coherent functions. To evaluate the biological relevance of our fundamental components, we develop a method to perform differential expression analysis in the feature-space described by our FCs.

Using this technique, we assess the ability of our method to identify known mechanisms of parthenolide (PTL), a preclinical drug under investigation for its ability to selectively induce apoptosis in multiple cancer types [46–50]. Known PTL effects can be divided into two intracellular signals. First, PTL induces oxidative stress, evidenced by increased levels of reactive oxygen species [51] and activation of c-Jun N-terminal kinase (JNK) [52]. Second, PTL inhibits inflammatory responses via inhibition of STAT3 [53] and the transcription factor NF-κB [54,55]. PTL treatment leads to apoptosis in various cancer cell lines [46,51], and activation of p53 has been associated with the AML-specific apoptosis mechanism [46]. We show that independent components derived from a diverse compendium offer module-level insight into transcriptional response that cannot be gleaned from the PTL dataset alone.

2 Methods

2.1 Creation of the human gene expression compendium

We downloaded our microarray compendium from the Gene Expression Omnibus (GEO) [56], selecting all GEO Series (GSEs) run on the Affymetrix Human Genome U133 Plus 2.0 array (GEO accession GPL570) available on May 28, 2008. We filtered these arrays to remove samples that represented species other than human, and we removed GSEs with missing data so that imputation was not necessary. Normalized microarray data are often uploaded to GEO, but we downloaded only unprocessed CEL files in order to standardize the normalization procedures across all arrays in the compendium. After applying these filters, the resulting dataset consisted of 298 GEO Series comprised of 9,395 arrays.

We applied a two-step normalization pipeline as previously described [10]. For each series, we aggregated and normalized probe level information using robust multi-array average (RMA) [57], transformed each expression value using log base 2, and removed technical bias resulting from variation in hybridization conditions and starting material using the R package *bias* 0.0.3 [58]. This within-series normalization identified probe or arrays outliers within each dataset. We mapped probes to genes using the Bioconductor annotation package *hgu133plus2* 1.16.0 [59], and calculated expression values for each gene by averaging the values of probesets measuring the same gene. Finally, we performed quantile normalization [60] on the entire compendium using the *limma* R package (version 2.18.2) [61] in order to reconcile broader differences between datatsets and ensure that all arrays were on the same

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

scale prior to applying ICA. This produced a compendium comprising 20,099 genes and 9,395 arrays.

To reduce the contributions of over-represented conditions, and to eliminate rare signals that the compendium did not sample sufficiently, we grouped similar arrays using hierarchical clustering. First, we centered and scaled the expression values within an array, and centered the expression values for each gene. We calculated pairwise correlations between arrays using Spearman's rank correlation coefficient, *r*, a metric robust to outliers that performs well with microarray data [62,63]. We defined pairwise distances, *d*, by

$$d_{x,y}=1-|r_{x,y}|, \quad (1)$$

where *x* and *y* are the gene expression measurements of individual microarrays. Using these distances, we applied agglomerative hierarchical clustering to the compendium using average linkage. To filter the resulting clusters, we required that nodes pass two cutoffs. First, in order to cap the heterogeneity of arrays within a cluster, we empirically limited the maximum pairwise intra-cluster distance, d_{max}, to 0.3. Second, in order to eliminate conditions or signals that did not have sufficient representation within the compendium, we required a minimum cluster size, c_{min}, of 5. We generated representative nodes for clusters that passed these cutoffs by using average linkage to calculate centroids. We called these representative cluster nodes "meta-samples," and combined them to create a meta-compendium that reduced the dimensionality of the data. We examined the sensitivity of the meta-compendium and independent components extracted to the parameters d_{max} and c_{min} and found that our results were robust with different parameters (data not shown).

2.2 Iterated FastICA

ICA identifies latent variables, or independent components, that provide a new set of basis vectors for a dataset. In other words, ICA performs dimension reduction by describing the data in the feature-space identified by independent components. As opposed to gene modules, which typically consist of unordered subsets of genes, components contain scores for each gene in the domain. To extract a robust set of independent components from the meta-compendium, we applied iterated ICA and averaged the results. Here we briefly describe the ICA model; for a more thorough review of the procedural details of the application of ICA to microarray data, see Kong et al. [64].

Let $\boldsymbol{X}$ represent the $m \times n$ matrix of microarray data measuring the expression of *n* genes in *m* experiments. ICA models this expression matrix as a linear combination of independent biological processes by decomposing $\boldsymbol{X}$ into a $k \times n$ source matrix $\boldsymbol{S}$ and a $m \times k$ mixing matrix $\boldsymbol{A}$ such that $\boldsymbol{X} = \boldsymbol{AS}$, where *k* is a user-supplied parameter no greater than the minimum of *m* and *n*. For our analyses, we set $k = m$ to extract the maximum number of components from the meta-compendium. These components, or rows of $\boldsymbol{S}$, are independent in the sense that the gene weights in each component reflect samplings of independent random variables. In the context of gene expression, this suggests that the sets of genes comprising the groups strongly contributing to each component have independent compositions. We performed ICA using the R interface to the FastICA algorithm (version 1.1–11), which attempts to maximize the non-Gaussianity of the component distributions [65]. For microarray data, these distributions are typically super-Gaussian: only a small number of genes contribute heavily to a specific biological process, while the majority of genes do not contribute [28]. We used the contrast function

$$g(u)=\log\cosh u \quad (2)$$

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

and allowed a convergence tolerance of 0.0001. To account for the stochastic nature of ICA, we ran the algorithm 20 times and pooled the independent component estimates in a method similar to the Icasso software [66]. We clustered component estimates with partitioning around medoids [67], a robust version of *k*-means, using the distance metric based on Spearman's correlation coefficient in Eq. (1). We defined the fundamental components (FCs) of our analysis as the *k* medoids that result from this clustering analysis. To visualize the independent component estimates, we used classical multi-dimensional scaling to represent the 20,099-dimension component matrix in two dimensions.

2.3 Functional annotation of fundamental components

To examine the biological processes associated with each component, we made use of two sources of information yielded by the ICA decomposition: the distribution of gene weights in a component (rows of $\boldsymbol{S}$), and the distribution of a component's expression in arrays in the compendium (columns of $\boldsymbol{A}$).

2.3.1 Annotation based on component gene weight distributions—The rows of the source matrix, $\boldsymbol{S}$, describe the contributions of individual genes to each component. Each component contains an expression value for each gene. ICA produces components with unit variance and zero mean, so the contributions of each gene are relative. We call the genes that significantly contribute to each component the *active genes*, and identified them by using a weight threshold [37] of three standard deviations from the mean. Similar to Lee and Batzoglou [31], we generated two sets of active genes for each component: one for genes with positive loadings and one for genes with negative loadings. Our initial observations, as well as those of Frigyei et al. [38], suggested that this division results in a greater number of significant annotations. For a given component, we called the module with more genes the *dominant module*. Since the signs of independent component expression values are arbitrary, we reoriented each component so that the loadings for the dominant module were positive.

To investigate the biological functions of our fundamental components, we used Alexa et al.'s *elim* algorithm [68] to identify enriched Biological Process GO categories in the active gene modules, assessing significance with Fisher's exact test. We annotated each gene cluster with GO categories with a Bonferroni-corrected p-value <0.05. To further characterize the biological processes associated with each component, we identified significant canonical pathways using Ingenuity Pathway Analysis (Ingenuity Systems, www.ingenuity.com). Ingenuity uses Fisher's exact test to calculate a p-value determining the probability that the association between genes in the dataset and the canonical pathway is explained by chance alone. We defined significant pathways as those with a p-value <0.05 after correcting for false discovery rate with the Benjamini-Hochberg method.

2.3.2 Annotation based on component expression in compendium—Independent components provide an orthogonal basis with which to describe the original data, so we modeled each original experiment as a linear combination of FCs. To generate a mixing matrix that describes the original arrays, we projected the full 9,395-array compendium into the feature-space defined by the FCs using Eq. (3).

$$A = XS^T \tag{3}$$

This resulted in an (array × component) mixing matrix, $\boldsymbol{A}$, whose columns describe the contributions of a component to each array in the compendium. A high absolute weight in this matrix for component *i* and array *j* indicates that the gene expression pattern present in component *i* plays a significant role in defining the expression profile of array *j*. Because component expression forms complex distributions [37], we annotated each FC with the 100

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

individual arrays in the compendium that have the highest expression of that component. In other words, we found the 100 arrays with the highest scores in each column of ***A***, which we called the *active arrays*. The active arrays define the context in which a gene module is expressed. We annotated each gene module with the GEO Series containing the active arrays, ordering the GSEs by the number of active arrays contributed by each series.

2.4 Differential expression analysis of parthenolide

From the full compendium, we selected and normalized GSE7538, a dataset consisting of 12 pairs of PTL-treated or untreated primary acute myelogenous leukemia (AML) specimens [8]. We projected this dataset onto the basis vectors defined by our fundamental components (Equation 3) to determine the levels of component expression in each array. Analogous to the use of differential expression analysis to identify differentially expressed (DE) genes of interest, we used the quantitative expression levels of components to identify DE components. Specifically, we applied the linear modeling and empirical Bayes methods of the *limma* 2.18.2 R package [61] to identify differentially expressed components between treated and untreated AML samples. To compare this approach with existing DE gene identification methods, we used the moderated t-statistic from *limma* to identify DE genes from the original dataset. We also applied ICA to the PTL dataset alone, using the methodology described in Section 2.2, to identify the maximum possible number of independent components (24). Using only the data from GSE7538, we again used linear modeling to identify DE components. For each comparison, we adjusted the p-values for differential expression using the Benjamini-Hochberg correction. We examined genes with p-values below 0.05 and components with p-values below 0.001. We also restricted our search to components whose differential expression had the same sign for all 12 treated/untreated pairs; while this strategy might eliminate true signal, it also removes inconsistent patterns that may represent heterogeneity within the tumors.

2.5 Identification of novel regulators of parthenolide response

To identify novel regulators of parthenolide response, we used the dominant gene modules from DE components to create protein interactions networks with Ingenuity Pathway Analysis (Ingenuity Systems, www.ingenuity.com). Ingenuity overlaid active genes onto a global molecular network developed from information contained in Ingenuity's Knowledge Base, and it algorithmically generated networks based on gene connectivity. We ranked genes in order of connectivity, and examined the highly-connected hub genes. To generate a list of known PTL-associated genes, we used the Pharmspresso text-mining tool to search the biomedical literature [69]. Pharmspresso searched all 322 PubMed articles with "parthenolide" in the title or abstract looking for co-occurrences of the term "parthenolide" with gene names in the same sentence, and obtained a list of 363 known PTL-associated genes.

3 Results

3.1 Identification of fundamental components of human biology

Our starting compendium consisted of 20,099 genes measured by 9,395 heterogeneous arrays, representing 298 GEO Series (GSEs) that spanned a wide range of human biology. While a majority of these data series contributed fewer than fifteen arrays, we noticed that the five largest GSEs constituted over one third of the total arrays in the compendium. For instance, GSE2109 contributed 1,974 primary tumor samples from various sources and GSE8052 contributed 404 lymphoblastoid samples from an asthma study. A compendium-wide analysis using Spearman's rank correlation as a similarity metric revealed strong correlations within datasets and between experiments performed with similar cell types and conditions (Supplementary Figure 1). We reasoned that the disproportionate representation

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

of some transcriptional patterns might cause certain biological conditions to dominate the ICA results. At the same time, we did not wish for rare gene expression patterns caused by extreme experimental perturbations or technical errors to skew the compendium. Therefore, to reduce the contributions of over-represented conditions and to eliminate cellular signatures that were not replicated within the compendium, we applied a filter based on hierarchical clustering to the original compendium (Figure 1). The resulting meta-compendium consisted of 423 meta-samples that included contributions from 5,818 of the original arrays.

We next applied iterated ICA to the meta-compendium to identify 423 fundamental components that span a wide range of human biology. Since ICA produces components in an arbitrary order, we sorted the components based on the amount of variance of the meta-compendium they explain (Supplementary Figure 2). Unlike PCA, ICA does not use variance explained as a criterion to select components. Still, ordering components based on this metric may provide biologically relevant information [28]: fundamental components that explain a large percentage of the variance in a compendium represent processes that are active in many biological or experimental conditions, while those that explain a relatively small percentage of the variance represent relatively infrequent transcriptional patterns that are active only under certain conditions. In this ordering scheme, the first independent component explained 3.2% of the variance in the meta-compendium. Over multiple runs, ICA reliably reproduced the ten components that explain the most variance (Figure 2a). To verify that low-ranking components also clustered together, we examined a sampling of components that explained varying fractions of the variance in the data (Figure 2b). Signals appeared robust even for the less well-represented components. Henceforth, we refer to components by their rank in this ordering. That is, Fundamental Component 1 (FC-1) explains the most variance in the meta-compendium, while FC-423 explains the least.

To investigate whether our method of meta-compendium creation succeeded in reducing the contributions of over-represented conditions, we compared the percentage of variance explained by each FC in the meta-compendium to that in the original compendium (Supplementary Figure 3). Some FCs explain more variance in the meta-compendium than in the original compendium; these FCs may represent relatively rare but reproducible biology that was up-weighted in the meta-compendium. Many FCs, however, explain 1.5- to 3-fold less variance in the meta-compendium, and may represent biology that was over-sampled in the original compendium. We examined the active array clusters for two representative FCs, 62 and 268, selected for their high reduction in variance explained and their relative separation in the component ordering (Supplementary Figure 3). FC-62 explains 1.26% of the variance in the full compendium but only 0.31% of the variance in the meta-compendium, and the 100 arrays with the highest expression of the component all are part of GSE8052, the 404-array asthma study mentioned previously (Supplementary Figure 1). Similarly, the 100 arrays with the highest expression for FC-268 consist entirely of brain samples from a single GEO series. Since FC-268 explains relatively little variance in either compendium, it likely represents an over-sampled gene expression signature that does not play a large role in most biological conditions.

3.2 Functional annotation of fundamental components

To examine the biological processes and functions associated with each component, we made use of two sources of information yielded by ICA decomposition: the distribution of gene weights in a component, and the distribution of component expression in arrays in the compendium. First, to identify genes that contribute significantly to each component, we used a loading cut-off of 3.0, corresponding to three standard deviations above the zero mean. For each component we created two gene modules, one for significant genes with positive loadings and one for negative loadings. The mean module size was 155.95, with a

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

standard deviation of 73.57, although typically one module associated with a component was larger than the other; the mean difference between the sizes of a component's modules was 110.00, with a standard deviation of 80.57. For a particular component, we refer to the module with more genes as the dominant module. 14,932 out of 20,009 genes belong to at least one module. We then calculated GO category enrichment in the active gene sets. 292 out of the 423 FCs had enriched GO categories in at least one of the two active gene modules. Supplementary Table 1 shows the enriched GO categories for the dominant modules of the first ten components. We observed a weak correlation between the amount of variance explained by a FC and the total number of enriched GO categories found for its active genes (Figure 3). However, in accordance with Roden et al.'s observations with PCA [70], many FCs that explain less than 0.5% of variance in the compendium have strong GO annotation. This suggests that these FCs explain relatively infrequent but biologically coherent transcriptional patterns.

3.3 Differential expression analysis of parthenolide treatment response

To assess the utility of our fundamental components, we analyzed a dataset from the initial compendium (GEO accession GSE7538) using three methods: 1) standard differential gene expression analysis, 2) differential expression analysis based on independent components derived from the dataset alone, and 3) differential expression analysis using FCs derived from the full compendium of data. GSE7538 contains acute myelogenous leukemia (AML) $CD34^+$ cells from twelve patients treated with parthenolide (PTL), a sesquiterpene lactone with antiinflammatory and anti-cancer activities [46,47].

Linear modeling with empirical Bayes methods in gene-space identified 2574 differentially expressed (DE) genes at a significance level <0.05 after correcting for multiple hypothesis testing, and we annotated this set of genes with significant GO categories and Ingenuity Canonical Pathways. We obtained more enriched terms with higher scores when separating the genes into up-regulated and down-regulated sets, so we present those results here. The significant pathways identified from the genes up-regulated in PTL-treated $CD34^+$ AML cells included NRF2-mediated Oxidative Stress Response, Hypoxia Signaling in the Cardiovascular System, and Protein Ubiquitination Pathway (Table 1), all of which are consistent with oxidative stress-induced apoptosis caused by parthenolide. Differential expression analysis in gene-space also marked Glucocorticoid Receptor Signaling, which is related to the NF-κB signaling pathway [71,72].

Next we applied iterative ICA to the 24-experiment data series, identifying 24 independent components. We projected the data into this feature space, and identified one significant DE component that explains 31.6% of the variability between treated and untreated cells. This method predicted that the Protein Ubiquitination Pathway and NRF2-mediated Oxidative Stress Response are up-regulated with increased significance compared to the gene-space analysis (Table 1). GO enrichment analysis in this feature space did not yield as many enriched GO categories as the same analysis in gene-space (Supplementary Tables 2,3).

Finally we analyzed the differential expression in response to PTL treatment in terms of our fundamental components. We projected the original dataset into FC-space and identified 29 DE components with a p-value threshold of 0.001. To eliminate signals reflecting parthenolide response heterogeneity, we limited our results to the 19 components that were regulated in the same direction in all twelve paired samples (Figure 4). Unlike DE analysis using independent components from GSE7538 alone, no FC explained a large fraction of differential expression as compared to the others; rather, complex manipulations of a wide range of gene modules contributed to the expression changes between PTL-treated and untreated cells. In addition to GO enrichment and Ingenuity Pathway Analysis, we identified

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

for each DE component the experiments that contain arrays with the highest expression scores for that component.

Although neither gene-space nor experiment-specific feature-space analysis identified NF-κB as a significant pathway in PTL response, FC-space analysis identified NF-κB signaling as the top pathway associated with FC-66, the fundamental component with the greatest differential expression (Table 2). FC-66 was also enriched for JAK/Stat signaling and SAPK/JNK signaling (Supplementary Table 4), two pathways with known associations with PTL response [52,53]. Listing the experiments that highly express this component revealed that FC-66 plays an important role in leukemia (Table 2). In the second most over-expressed fundamental component, FC-362, NRF2-Mediated Oxidative Stress Response appeared at the top of the list of enriched Ingenuity pathways, echoing the findings in gene-space and experiment-specific component-space and matching previous experimental results [51]. Intriguingly, N-Glycan Biosynthesis appeared at the top of the list for the third most over-expressed fundamental component, FC-84. Subsets of glycan structures have specific functions that affect cell differentiation, growth and migration, and have been implicated with cancer development and metastasis [73]. Of the three most under-expressed FCs, only FC-62 offered descriptive GO categories and Ingenuity pathways. The dominant gene module for this component showed enrichment for EIF2 Signaling, and a closer examination of the gene list for this module revealed a preponderance of ribosomal proteins. This suggests that PTL modulates translational activity in the cell, perhaps as part of the stress response to the increase in reactive oxygen species.

Unexpectedly, FC-66 was over-expressed in PTL-treated samples, seemingly countering current belief that PTL inhibits NF-κB activation. To determine the reason for this apparent reversal, we examined the expression of genes in the Ingenuity NF-κB Signaling pathway for both FC-66 (Figure 5) and the original data from GSE7538 (Supplementary Figure 4). This analysis revealed that genes in the NF-κB pathway exhibit a wide range of transcriptional changes in response to parthenolide, regardless of whether the genes play inhibitory or stimulatory roles. In the original data, IκBα (NIKBIA), an inhibitor of the NF-κB complex, is down-regulated in PTL-treated samples, while other components of inhibitory complexes are up-regulated (IκBβ, IκBε). Similarly, subunits of the NF-κB complex itself show varied transcriptional responses. However, the direct target of PTL, IKKβ (IKBKB), had a negative loading in FC-66 and was down-regulated in PTL-treated compared to untreated cells, suggesting that transcription-level regulation and protein-level activation are coordinated for this key gene in parthenolide response.

To explore the biological validity of our method's predictions on the gene-level as well as the module-level, we created a protein interaction network from the dominant module of each differentially expressed FC. For each FC, we ranked genes by connectivity in this network, since highly-connected proteins, or *hub genes*, are likely more essential to biological function and survival [74–76] and potentially represent therapeutic targets [77]. We defined our hub genes as the most highly connected 1% of the 3278 genes active in at least one of the 19 DE components, yielding a list of 32 highly-connected genes (Table 3). To identify novel predictions, we compared this list of hub genes to our text-mined catalog of known PTL-associated genes. Fourteen of the 32 hub genes appear on this text-mined list, including key regulators of PTL response such as TNF, NFKB1, STAT4 and FOS.

Among the genes not on the Pharmspresso list, several appear to represent known genes that the text-mining algorithm did not detect. C/EBP-β (CEBPB), for instance, is a transcription factor that interacts with a known PTL-associated gene, C/EBP-α (CEBPA), whose loss of function leads to the development of AML [78]. Kawasaki et al. found that regulatory targets of C/EBP-β are differentially expressed in PTL-treated prostate cancer cells [79].

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

HGF, although not previously directly linked to PTL, depends on the STAT pathway for HGF/SF-Met mediated tumorigenesis [80], and so may be linked to the PTL response mechanism.

Other genes on this list may represent new genes and pathways involved in PTL response. CNBP and EIF3A both contribute to FC-62, whose dominant gene module is associated with protein translation. CNBP plays a role in controlling proliferation and cell survival, for instance by regulating the CT element of the c-*myc* protooncogene [81]. A group of highly-connected proteins – LCK, ZAP70, LAT, CD247 and CD3E – contribute to the T Cell Receptor Signaling pathway in FC-43 (Supplementary Table 4). Some of these genes also appear in the NF-κB Signaling pathway (Figure 5). Lewis and coworkers reported aberrant expression of T-cell markers in AML [82], and an independent microarray study implicated T-cell receptor signaling in AML stem cell function [83], suggesting that expression of this pathway may be relevant to AML response to PTL.

4 Discussion

To develop a data-driven view of the human transcriptome, the biological community needs analysis tools that leverage the growing repositories of gene expression data. We assembled a large cross-study compendium of human microarray data and extracted fundamental components of human biology using independent component analysis. These components described gene modules that have biologically coherent functions, and we used them to improve the differential expression analysis of a microarray dataset involving the preclinical drug parthenolide (PTL).

Because we identified gene modules through a meta-analysis of expression data from GEO, our analysis depends fundamentally on the quality of the primary microarray data. Variation in hybridization conditions and starting mRNA quantity and quality could introduce spurious signals whose origins are technical rather than biological. To address this issue, we applied the technical bias correction proposed by Eklund and Szallasi [58] to all GEO data. Our initial compendium also reflected a non-uniform sampling of human biology due to biases in the biological conditions commonly investigated (Supplementary Figure 1). In a study to mine functional gene relationships from a large microarray compendium, Huttenhower et al. applied Bayesian normalization to account for the over-representation of certain biological or experimental conditions [84]. Similarly, we wished to derive a broad but even sampling of human biology; we did not want our fundamental components to describe a disproportional number of pathways unique to cancer, for example. We applied a grouping method based on hierarchical clustering to create a meta-compendium that down-weighted the contributions from over-represented conditions and eliminated expression profiles that did not appear reproducibly in the compendium. In the process, we averaged microarray profiles within the same cluster, possibly removing information about some of the less well-represented transcriptional subunits present within datasets. However, we wanted to focus on modules active in multiple experiments throughout our collection, so we chose to retain well-represented transcriptional processes at the expense of less well-represented ones.

Additionally, this procedure for normalizing the weights of different biological conditions resulted in the removal of over 3,000 arrays. Of these arrays, approximately 1,900 of them consisted of tumor samples derived from a single compilation dataset (GEO accession GSE2109). We note, however, that even with this normalization, the scope of our results depends on the sampling of biology available.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

In our analysis, we used ICA to identify gene modules that vary together across our gene expression compendium. While Fehrmann et al. used principal component analysis (PCA) for a similar analysis [21], ICA provides a complementary method for analyzing gene expression data [31]. At the same time, ICA does not rely on prior knowledge for identification of gene modules; although prior knowledge may enhance the analysis of well-studied biological processes [16,29], we chose to use gene expression data alone to guide the formation of our gene modules. In this context, ICA has two major procedural limitations. First, the user must specify the number of components to extract. Like other studies using ICA, we searched for the maximum possible number of components (423, the number of samples in the meta-compendium), assuming that more biological processes exist than we can describe with this number of sources [31]. Second, ICA is a stochastic algorithm. To extract robust biological signals from the meta-compendium, we repeated ICA twenty times. In contrast to approaches that mitigate unstable signals by eliminating them entirely [37], we clustered the results to obtain more reliable component estimates [66] using partitioning around medoids. To maximize our ability to comprehensively model microarray data, we included all 423 components in downstream analyses. One final challenge associated with matrix decomposition methods involved identifying active genes. Although alternative methods exist [38], we followed the approach of Liebermeister [28] and used a gene loading threshold.

To confirm the biological relevance of these gene modules, we devised a method to analyze differential expression in the feature-space defined by our fundamental components. This follows naturally from the typical assumption in work with ICA that the active genes in each component are differentially expressed [28,35]. Although several groups have used principal component features to enhance the detection of individual differentially expressed genes [85] or gene pathways from GO and KEGG [86], the opportunity to assess the differential expression of entire gene modules has not been adequately explored. We exploited the component loadings in the mixing matrix of the ICA decomposition and used the moderated *t* test [61] to assess the differential expression of globally-derived gene modules within a single dataset. Analysis in feature-space reduced the number of statistical tests performed, decreasing the need for multiple hypothesis testing correction and increasing the sensitivity of biological signal detection [87]. We empirically selected different p-value cutoffs for DE gene analysis ($p=0.05$) and DE component analysis ($p=0.001$) to optimize the performance of each method.

DE analysis in feature-space outperformed similar analysis in gene-space, identifying more known PTL-associated genes as significant (Figure 6). Additionally, the pathways enriched in differentially expressed FCs proved to be highly consistent with previously characterized PTL effects on cellular processes, identifying both NF-κB signaling and oxidative stress as major mechanisms involved in response to PTL treatment. Because we modeled a gene expression profile as a linear combination of components, however, we must interpret pathway predictions with care. An active gene in a DE component may not necessarily be a DE gene in gene-space analysis. In our investigation of the mechanism of parthenolide, FC-66 loadings of genes in the NF-κB pathway did not necessarily mirror the direction of regulation in differential gene expression analysis (Supplementary Figure 4). Complex gene networks such as the NF-κB signaling pathway contain proteins with both inhibitory and stimulatory effects, so the finding that FC-66 was over-expressed in PTL-treated cells did not necessarily indicate that the entire NF-κB pathway was both transcriptionally and functionally up-regulated. Rather, different genes responded to this stimulus according to their role in the pathway (Figure 5). As a further demonstration of this transcriptional complexity, we note that our method identified the hub gene Cox-2 (PTGS2) as differentially up-regulated in FC-406 and down-regulated in FC-101 (data not shown). While we were unable to determine the net direction in which PTL response influences

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Cox-2 expression, our result corroborates previous data suggesting that Cox-2 protein activity changes in response to PTL treatment [88,89]. This demonstrates the enhanced sensitivity of our approach, as DE analysis in gene-space did not identify Cox-2 as an important gene in PTL response.

After deriving gene modules using a data-driven approach, we used ontological annotations to provide a preliminary indication of the functional relationships between active genes. We found that for some of these gene modules, Ingenuity Pathway Analysis provided more descriptive relationships between genes, especially in the realm of signaling pathways (e.g. FC-66, Table 2). Our analysis of the mechanism of PTL, a drug that affects the NF-κB signaling pathway, particularly benefited from this form of annotation. In other cases, GO Biological Process categories labeled components where Ingenuity did not (e.g. FC-101, Table 2). Still, annotation of our fundamental components with GO categories left 131 components unlabeled, suggesting that GO does not completely describe some functional gene relationships. Indeed, some of the DE fundamental components identified in our analysis of PTL had neither descriptive GO categories nor significant Ingenuity Pathways. FC-208, for instance, was highly expressed in several cancer-related experiments, but was down-regulated in response to PTL (Table 2). This gene module may represent a cancer-specific transcriptional subunit that is suppressed by PTL. Similarly, FC-101 had vague GO annotations including “defense response,” but was highly expressed in three independent AML experiments (Table 2). The three genes with the highest weights in FC-101 consisted of DNTT, whose expression in AML cells correlates with outcome [90,91]; SOCS2, which is part of a gene signature prognostic for an unfavorable AML subtype [92]; and DDIT4, whose mRNA levels are induced in AML cell lines and normal $CD34^+$ cells during differentiation [93]. Thus our method detected a transcriptional subunit active in AML that is not adequately characterized by traditional functional analysis methods.

Since ICA defines components in a manner that allows a gene to participate in more than one transcriptional module, our approach also provides a mechanism for modeling context specificity [13]. As we highlighted using the example of Cox-2, our method defines the functional role of a gene in the context of other genes. Given the independence properties of components derived by ICA, we interpret these results as defining several specific contexts in which each gene operates. Furthermore, since we can identify experimental conditions that lead to the co-expression of sets of genes, we can match gene modules with their biological contexts. For example, the set of genes in the dominant module of FC-101 might be labeled as “up-regulated in AML” and “differentially expressed in PTL response.” This context awareness will prove invaluable as we continue to explore the functional relationships between genes at the genomic scale.

5 Conclusion

As opposed to a top-down resource such as the Gene Ontology (GO), we provide an entirely data-driven way to identify gene modules in a meta-analysis of expression data using ICA. These modules contain functionally coherent sets of genes and increase the descriptive power of differential expression analysis. The results of our analysis of a parthenolide-related dataset suggests that the fundamental components we have identified may serve as useful, reusable features; although we applied these modules to examine a microarray experiment from our original compendium, we can use the same components to investigate new expression data. In addition to differential expression analysis, these gene modules could serve as data-driven gene sets for Gene Set Enrichment Analysis [18] or as features for clustering and classifying diverse microarray expression data. Given the wide range of biological pathways represented in the compendium, these components could be applied to query a microarray database for experiments that modulate similar biological processes.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Finally, we can use this data-driven framework to annotate genes with context-specific functional relationships.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors wish to thank Craig Jordan for sharing his insights into parthenolide response; Yael Garten for identifying parthenolide-associated genes with Pharmspresso; and Ramon Felciano for supporting the work with Ingenuity Pathway Analysis.

Funding

This work was supported by NIH Grant 5U01GM061374-08 for the Pharmacogenomics Knowledge Base. Computational resources were provided by the Stanford Bio-X2 Cluster, supported by NSF award CNS-0619926. JME was supported in part by a Stanford Bio-X Undergraduate Research Award. BJD was supported by the Geraldine Jackson Fuhrman Stanford Graduate Fellowship and an HHMI predoctoral fellowship.

References

1. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M, Friend SH. Functional discovery via a compendium of expression profiles. Cell 2000 Jul;102(1):109–126. [PubMed: 10929718]
2. Rhodes DR, Barrette TR, Rubin MA, Ghosh D, Chinnaiyan AM. Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. Cancer Res 2002 Aug;62(15):4427–4433. [PubMed: 12154050]
3. Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N. Revealing modular organization in the yeast transcriptional network. Nat Genet 2002 Aug;31(4):370–377. [PubMed: 12134151]
4. Ramaswamy S, Ross KN, Lander ES, Golub TR. A molecular signature of metastasis in primary solid tumors. Nat Genet 2003 Jan;33(1):49–54. [PubMed: 12469122]
5. Xu L, Tan AC, Naiman DQ, Geman D, Winslow RL. Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. Bioinformatics 2005 Oct;21(20):3905–3911. [PubMed: 16131522]
6. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet J, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science 2006 Sep;313(5795):1929–1935. [PubMed: 17008526]
7. Hibbs MA, Hess DC, Myers CL, Huttenhower C, Li K, Troyanskaya OG. Exploring the functional landscape of gene expression: directed search of large microarray compendia. Bioinformatics 2007 Oct;23(20):2692–2699. [PubMed: 17724061]
8. Hassane DC, Guzman ML, Corbett C, Li X, Abboud R, Young F, Liesveld JL, Carroll M, Jordan CT. Discovery of agents that eradicate leukemia stem cells using an in silico screen of public gene expression data. Blood 2008 Jun;111(12):5654–5662. [PubMed: 18305216]
9. Dudley JT, Tibshirani R, Deshpande T, Butte AJ. Disease signatures are robust across tissues and experiments. Mol Syst Biol 2009;5:307. [PubMed: 19756046]
10. Daigle BJ Jr, Deng A, McLaughlin T, Cushman SW, Cam MC, Reaven G, Tsao PS, Altman RB. Using pre-existing microarray datasets to increase experimental power: application to insulin resistance. PLoS Comput Biol 2010;6(3):e1000718. [PubMed: 20361040]
11. Khatri P, Drăghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics 2005 Sep;21(18):3587–3595. [PubMed: 15994189]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

12. Ihmels J, Bergmann S, Berman J, Barkai N. Comparative gene expression analysis by differential clustering approach: application to the Candida albicans transcription program. PLoS Genet 2005 Sep;1(3):e39. [PubMed: 16470937]

13. Huang Y, Li H, Hu H, Yan X, Waterman MS, Huang H, Zhou XJ. Systematic discovery of functional modules and context-specific functional annotation of human genome. Bioinformatics 2007 Jul;23(13):222–229. [PubMed: 17110366]

14. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998 Dec;95(25):14863–14868. [PubMed: 9843981]

15. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 1998 Dec;9(12):3273–3297. [PubMed: 9843569]

16. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 2003 Jun;34(2):166–176. [PubMed: 12740579]

17. Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, Gifford DK. Computational discovery of gene modules and regulatory networks. Nat Biotechnol 2003 Nov;21(11):1337–1342. [PubMed: 14555958]

18. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005 Oct;102(43):15545–15550. [PubMed: 16199517]

19. Ihmels J, Bergmann S, Barkai N. Defining transcription modules using large-scale gene expression data. Bioinformatics 2004 Sep;20(13):1993–2003. [PubMed: 15044247]

20. Wang X, Dalkic E, Wu M, Chan C. Gene module level analysis: identification to networks and dynamics. Curr Opin Biotechnol 2008 Oct;19(5):482–491. [PubMed: 18725293]

21. Fehrmann RSN, de Jonge HJM, Ter Elst A, de Vries A, Crijns AGP, Weidenaar AC, Gerbens F, de Jong S, van der Zee AGJ, de Vries EGE, Kamps WA, Hofstra RMW, Te Meerman GJ, de Bont ESJM. A new perspective on transcriptional system regulation (TSR): towards TSR profiling. PLoS One 2008;3(2):e1656. [PubMed: 18297136]

22. Jeffery CJ. Moonlighting proteins: old proteins learning new tricks. Trends Genet 2003 Aug;19(8): 415–417. [PubMed: 12902157]

23. Madeira SC, Oliveira AL. Biclustering algorithms for biological data analysis: a survey. IEEE/ ACM Trans Comput Biol Bioinform 2004;1(1):24–45. [PubMed: 17048406]

24. Cheng Y, Church GM. Biclustering of expression data. Proc Int Conf Intell Syst Mol Biol 2000;8:93–103. [PubMed: 10977070]

25. Carmona-Saez P, Pascual-Marqui RD, Tirado F, Carazo JM, Pascual-Montano A. Biclustering of gene expression data by Non-smooth Non-negative Matrix Factorization. BMC Bioinformatics 2006;7:78. [PubMed: 16503973]

26. Raychaudhuri S, Stuart JM, Altman RB. Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac Symp Biocomput 2000:455–466. [PubMed: 10902193]

27. Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci U S A 2000 Aug;97(18):10101–10106. [PubMed: 10963673]

28. Liebermeister W. Linear modes of gene expression determined by independent component analysis. Bioinformatics 2002 Jan;18(1):51–60. [PubMed: 11836211]

29. Liao JC, Boscolo R, Yang Y, Tran LM, Sabatti C, Roychowdhury VP. Network component analysis: reconstruction of regulatory signals in biological systems. Proc Natl Acad Sci U S A 2003 Dec;100(26):15522–15527. [PubMed: 14673099]

30. Gong T, Xuan J, Wang C, Li H, Hoffman E, Clarke R, Wang Y. Gene Module Identification from Microarray Data Using Nonnegative Independent Component Analysis. Gene Regulation and Systems Biology 2007;1:349–363. [PubMed: 19936101]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

31. Lee S, Batzoglou S. Application of independent component analysis to microarrays. Genome Biol 2003;4(11):R76. [PubMed: 14611662]
32. Li Y, Adali T, Calhoun VD. Estimating the number of independent components for functional magnetic resonance imaging data. Hum Brain Mapp 2007 Nov;28(11):1251–1266. [PubMed: 17274023]
33. Purdom E, Holmes SP. Error distribution for gene expression data. Stat Appl Genet Mol Biol 2005;4 Article16.
34. Salas-Gonzalez D, Kuruoglu E, Ruiz D. A heavy-tailed empirical Bayes method for replicated microarray data. Computational Statistics & Data Analysis 2009 March;53(5):1535–1546.
35. Saidi SA, Holland CM, Kreil DP, MacKay DJC, Charnock-Jones DS, Print CG, Smith SK. Independent component analysis of microarray data in the study of endometrial cancer. Oncogene 2004 Aug;23(39):6677–6683. [PubMed: 15247901]
36. Martoglio A, Miskin JW, Smith SK, MacKay DJC. A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer. Bioinformatics 2002 Dec;18(12):1617–1624. [PubMed: 12490446]
37. Chiappetta P, Roubaud MC, Torrésani B. Blind source separation and the analysis of microarray data. J Comput Biol 2004;11(6):1090–1109. [PubMed: 15662200]
38. Frigyesi A, Veerla S, Lindgren D, Höglund M. Independent component analysis reveals new and biologically significant structures in micro array data. BMC Bioinformatics 2006;7:290. [PubMed: 16762055]
39. Teschendorff AE, Journée M, Absil PA, Sepulchre R, Caldas C. Elucidating the altered transcriptional programs in breast cancer using independent component analysis. PLoS Comput Biol 2007 Aug;3(8):e161. [PubMed: 17708679]
40. Zhang XW, Yap YL, Wei D, Chen F, Danchin A. Molecular diagnosis of human cancer type by gene expression profiles and independent component analysis. Eur J Hum Genet 2005 Dec;13(12): 1303–1311. [PubMed: 16205741]
41. Zheng C, Huang D, Kong X, Zhao X. Gene expression data classification using consensus independent component analysis. Genomics Proteomics Bioinformatics 2008 Jun;6(2):74–82. [PubMed: 18973863]
42. Li H, Zhan M. Unraveling transcriptional regulatory programs by integrative analysis of microarray and transcription factor binding data. Bioinformatics 2008 Sep;24(17):1874–1880. [PubMed: 18586698]
43. Capobianco E. Model validation for gene selection and regulation maps. Funct Integr Genomics 2008 May;8(2):87–99. [PubMed: 18064499]
44. Chen L, Xuan J, Wang C, Shih I, Wang Y, Zhang Z, Hoffman E, Clarke R. Knowledge-guided multi-scale independent component analysis for biomarker identification. BMC Bioinformatics 2008;9:416. [PubMed: 18837990]
45. Lutter D, Langmann T, Ugocsai P, Moehle C, Seibold E, Splettstoesser WD, Gruber P, Lang EW, Schmitz G. Analyzing time-dependent microarray data using independent component analysis derived expression modes from human macrophages infected with F. tularensis holartica. J Biomed Inform 2009 Aug;42(4):605–611. [PubMed: 19535009]
46. Guzman ML, Rossi RM, Karnischky L, Li X, Peterson DR, Howard DS, Jordan CT. The sesquiterpene lactone parthenolide induces apoptosis of human acute myelogenous leukemia stem and progenitor cells. Blood 2005 Jun;105(11):4163–4169. [PubMed: 15687234]
47. Wang W, Adachi M, Kawamura R, Sakamoto H, Hayashi T, Ishida T, Imai K, Shinomura Y. Parthenolide-induced apoptosis in multiple myeloma cells involves reactive oxygen species generation and cell sensitivity depends on catalase activity. Apoptosis 2006 Dec;11(12):2225–2235. [PubMed: 17051330]
48. Anderson KN, Bejcek BE. Parthenolide induces apoptosis in glioblastomas without affecting NF-kappaB. J Pharmacol Sci 2008 Feb;106(2):318–320. [PubMed: 18277052]
49. Liu Y, Lu W, Guo J, Du J, Li T, Wu J, Wang G, Wang J, Zhang X, Zhang Q. A potential target associated with both cancer and cancer stem cells: a combination therapy for eradication of breast cancer using vinorelbine stealthy liposomes plus parthenolide stealthy liposomes. J Control Release 2008 Jul;129(1):18–25. [PubMed: 18466993]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

50. Suvannasankha A, Crean CD, Shanmugam R, Farag SS, Abonour R, Boswell HS, Nakshatri H. Antimyeloma effects of a sesquiterpene lactone parthenolide. Clin Cancer Res 2008 Mar;14(6): 1814–1822. [PubMed: 18347184]
51. Wen J, You K, Lee S, Song C, Kim D. Oxidative stress-mediated apoptosis. The anticancer effect of the sesquiterpene lactone parthenolide. J Biol Chem 2002 Oct;277(41):38954–38964. [PubMed: 12151389]
52. Nakshatri H, Rice SE, Bhat-Nakshatri P. Antitumor agent parthenolide reverses resistance of breast cancer cells to tumor necrosis factor-related apoptosis-inducing ligand through sustained activation of c-Jun N-terminal kinase. Oncogene 2004 Sep;23(44):7330–7344. [PubMed: 15286701]
53. Sobota R, Szwed M, Kasza A, Bugno M, Kordula T. Parthenolide inhibits activation of signal transducers and activators of transcription (STATs) induced by cytokines of the IL-6 family. Biochem Biophys Res Commun 2000 Jan;267(1):329–333. [PubMed: 10623619]
54. Hehner SP, Hofmann TG, Dröge W, Schmitz ML. The antiinflammatory sesquiterpene lactone parthenolide inhibits NF-kappa B by targeting the I kappa B kinase complex. J Immunol 1999 Nov;163(10):5617–5623. [PubMed: 10553091]
55. Kwok BH, Koh B, Ndubuisi MI, Elofsson M, Crews CM. The anti-inflammatory natural product parthenolide from the medicinal herb Feverfew directly binds to and inhibits IkappaB kinase. Chem Biol 2001 Aug;8(8):759–766. [PubMed: 11514225]
56. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002 Jan;30(1):207–210. [PubMed: 11752295]
57. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003 Apr;4(2):249–264. [PubMed: 12925520]
58. Eklund AC, Szallasi Z. Correction of technical bias in clinical microarray data improves concordance with known biological information. Genome Biol 2008;9(2):R26. [PubMed: 18248669]
59. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004;5(10):R80. [PubMed: 15461798]
60. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003 Jan;19(2):185–193. [PubMed: 12538238]
61. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 2004;3 Article3.
62. Balasubramaniyan R, Hüllermeier E, Weskamp N, Kämper J. Clustering of gene expression data using a local shape-based similarity measure. Bioinformatics 2005 Apr;21(7):1069–1077. [PubMed: 15513997]
63. Yona G, Dirks W, Rahman S, Lin DM. Effective similarity measures for expression profiles. Bioinformatics 2006 Jul;22(13):1616–1622. [PubMed: 16595558]
64. Kong W, Vanderburg CR, Gunshin H, Rogers JT, Huang X. A review of independent component analysis application to microarray gene expression data. Biotechniques 2008 Nov;45(5):501–520. [PubMed: 19007336]
65. Hyvarinen A. Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans Neural Netw 1999;10(3):626–634. [PubMed: 18252563]
66. Himberg J, Hyvärinen A, Esposito F. Validating the independent components of neuroimaging time series via clustering and visualization. Neuroimage 2004 Jul;22(3):1214–1222. [PubMed: 15219593]
67. Kaufman, L.; Rousseeuw, PJ. Finding groups in data: an introduction to cluster analysis. Hoboken, N.J.: Wiley; 2005.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

68. Alexa A, Rahnenführer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 2006 Jul;22(13):1600–1607. [PubMed: 16606683]
69. Garten Y, Altman RB. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. BMC Bioinformatics 2009;10 Suppl 2:S6. [PubMed: 19208194]
70. Roden JC, King BW, Trout D, Mortazavi A, Wold BJ, Hart CE. Mining gene expression data by interpreting principal components. BMC Bioinformatics 2006;7:194. [PubMed: 16600052]
71. Ray A, Prefontaine KE. Physical association and functional antagonism between the p65 subunit of transcription factor NF-kappa B and the glucocorticoid receptor. Proc Natl Acad Sci U S A 1994 Jan;91(2):752–756. [PubMed: 8290595]
72. Barnes PJ, Karin M. Nuclear factor-kappaB: a pivotal transcription factor in chronic inflammatory diseases. N Engl J Med 1997 Apr;336(15):1066–1071. [PubMed: 9091804]
73. Dennis JW, Granovsky M, Warren CE. Glycoprotein glycosylation and cancer progression. Biochim Biophys Acta 1999 Dec;1473(1):21–34. [PubMed: 10580127]
74. Jeong H, Mason SP, Barabási AL, Oltvai ZN. Lethality and centrality in protein networks. Nature 2001 May;411(6833):41–42. [PubMed: 11333967]
75. Tong AHY, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, Chen Y, Cheng X, Chua G, Friesen H, Goldberg DS, Haynes J, Humphries C, He G, Hussein S, Ke L, Krogan N, Li Z, Levinson JN, Lu H, Ménard P, Munyana C, Parsons AB, Ryan O, Tonikian R, Roberts T, Sdicu A, Shapiro J, Sheikh B, Suter B, Wong SL, Zhang LV, Zhu H, Burd CG, Munro S, Sander C, Rine J, Greenblatt J, Peter M, Bretscher A, Bell G, Roth FP, Brown GW, Andrews B, Bussey H, Boone C. Global mapping of the yeast genetic interaction network. Science 2004 Feb;303(5659):808–813. [PubMed: 14764870]
76. Carlson MRJ, Zhang B, Fang Z, Mischel PS, Horvath S, Nelson SF. Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. BMC Genomics 2006;7:40. [PubMed: 16515682]
77. Carter SL, Brechbühler CM, Griffin M, Bond AT. Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics 2004 Sep;20(14): 2242–2250. [PubMed: 15130938]
78. Mueller BU, Pabst T. C/EBPalpha and the pathophysiology of acute myeloid leukemia. Curr Opin Hematol 2006 Jan;13(1):7–14. [PubMed: 16319681]
79. Kawasaki BT, Hurt EM, Kalathur M, Duhagon MA, Milner JA, Kim YS, Farrar WL. Effects of the sesquiterpene lactone parthenolide on prostate tumor-initiating cells: An integrated molecular profiling approach. Prostate 2009 Jun;69(8):827–837. [PubMed: 19204913]
80. Zhang Y, Wang L, Jove R, Vande Woude GF. Requirement of Stat3 signaling for HGF/SF-Met mediated tumorigenesis. Oncogene 2002 Jan;21(2):217–226. [PubMed: 11803465]
81. Michelotti EF, Tomonaga T, Krutzsch H, Levens D. Cellular nucleic acid binding protein regulates the CT element of the human c-myc protooncogene. J Biol Chem 1995 Apr;270(16):9494–9499. [PubMed: 7721877]
82. Lewis RE, Cruse JM, Sanders CM, Webb RN, Suggs JL. Aberrant expression of T-cell markers in acute myeloid leukemia. Exp Mol Pathol 2007 Dec;83(3):462–463. [PubMed: 17927977]
83. Majeti R, Becker MW, Tian Q, Lee TM, Yan X, Liu R, Chiang J, Hood L, Clarke MF, Weissman IL. Dysregulated gene expression networks in human acute myelogenous leukemia stem cells. Proc Natl Acad Sci U S A 2009 Mar;106(9):3396–3401. [PubMed: 19218430]
84. Huttenhower C, Haley EM, Hibbs MA, Dumeaux V, Barrett DR, Coller HA, Troyanskaya OG. Exploring the human genome with functional maps. Genome Res 2009 Jun;19(6):1093–1106. [PubMed: 19246570]
85. Jonnalagadda S, Srinivasan R. Principal components analysis based methodology to identify differentially expressed genes in time-course microarray data. BMC Bioinformatics 2008;9:267. [PubMed: 18534040]
86. Ma S, Kosorok MR. Identification of differential gene pathways with principal component analysis. Bioinformatics 2009 Apr;25(7):882–889. [PubMed: 19223452]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

87. Qin H, Feng T, Harding SA, Tsai C, Zhang S. An efficient method to identify differentially expressed genes in microarray experiments. Bioinformatics 2008 Jul;24(14):1583–1589. [PubMed: 18453554]

88. Hwang D, Fischer NH, Jang BC, Tak H, Kim JK, Lee W. Inhibition of the expression of inducible cyclooxygenase and proinflammatory cytokines by sesquiterpene lactones in macrophages correlates with the inhibition of MAP kinases. Biochem Biophys Res Commun 1996 Sep;226(3): 810–818. [PubMed: 8831694]

89. Oka D, Nishimura K, Shiba M, Nakai Y, Arai Y, Nakayama M, Takayama H, Inoue H, Okuyama A, Nonomura N. Sesquiterpene lactone parthenolide suppresses tumor growth in a xenograft model of renal cell carcinoma by inhibiting the activation of NF-kappaB. Int J Cancer 2007 Jun; 120(12):2576–25781. [PubMed: 17290398]

90. Venditti A, Del Poeta G, Buccisano F, Tamburini A, Cox-Froncillo MC, Aronica G, Bruno A, Del Moro B, Epiceno AM, Battaglia A, Forte L, Postorino M, Cordero V, Santinelli S, Amadori S. Prognostic relevance of the expression of Tdt and CD7 in 335 cases of acute myeloid leukemia. Leukemia 1998 Jul;12(7):1056–1063. [PubMed: 9665190]

91. Huh YO, Smith TL, Collins P, Bueso-Ramos C, Albitar M, Kantarjian HM, Pierce SA, Freireich EJ. Terminal deoxynucleotidyl transferase expression in acute myelogenous leukemia and myelodysplasia as determined by flow cytometry. Leuk Lymphoma 2000 Apr;37(3–4):319–331. [PubMed: 10752983]

92. Bullinger L, Döhner K, Kranz R, Stirner C, Fröhling S, Scholl C, Kim YH, Schlenk RF, Tibshirani R, Döhner H, Pollack JR. An FLT3 gene-expression signature predicts clinical outcome in normal karyotype AML. Blood 2008 May;111(9):4490–4495. [PubMed: 18309032]

93. Gery S, Park DJ, Vuong PT, Virk RK, Muller CI, Hofmann W, Koeffler HP. RTP801 is a novel retinoic acid-responsive gene associated with myeloid differentiation. Exp Hematol 2007 Apr; 35(4):572–578. [PubMed: 17379067]

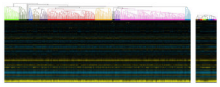

Figure 1.
Clustering procedure applied to subset of compendium. To reduce the contributions of highly replicated conditions in the compendium, we applied agglomerative hierarchical clustering using a distance metric based on Spearman's rank correlation to the full compendium, then selected and consolidated clusters with a maximum intra-cluster distance of 0.3 and a minimum size of 5. We selected 323 neuron samples that clustered together to graphically demonstrate this preprocessing step. Experiments from the original compendium (left) clustered together to form meta-samples (right) with corresponding colors. We excluded samples with black lines (left) from the meta-compendium because they did not satisfy our cluster selection criteria: they might represent extreme experimental perturbations or technical errors.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

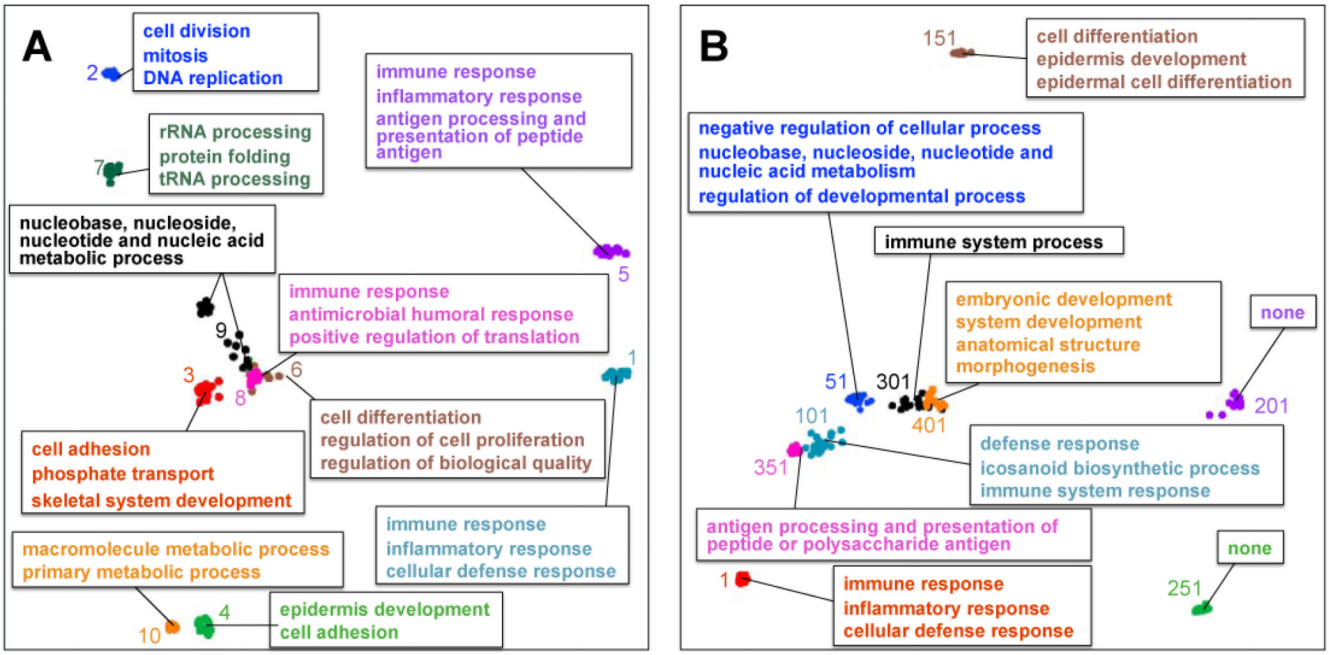

Figure 2.

Independent component estimates with GO Biological Process annotations. We applied ICA to the meta-compendium using 20 different random seeds to obtain 423 × 20 = 8460 component estimates. We clustered these estimates using partitioning around medoids and defined the 423 medoids as the fundamental components. We plotted a 2-D projection of the component estimates for **a)** the ten components that explain the most variance in the meta-compendium and **b)** a range of high- to low-ranking components that explain different amounts of variance in the meta-compendium. Colored numbers represent component ranks when ordered by variance explained.

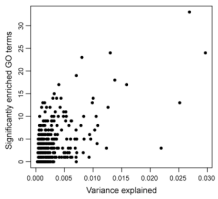

Figure 3.

Correlation between variance explained and enriched GO categories. We analyzed the active genes in the two modules associated with each FC and plotted the total number of enriched GO categories for a component versus the fraction of variance explained. Some FCs that explain a small percentage of variance have many enriched GO categories, suggesting that they represent rare but relevant expression profiles.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

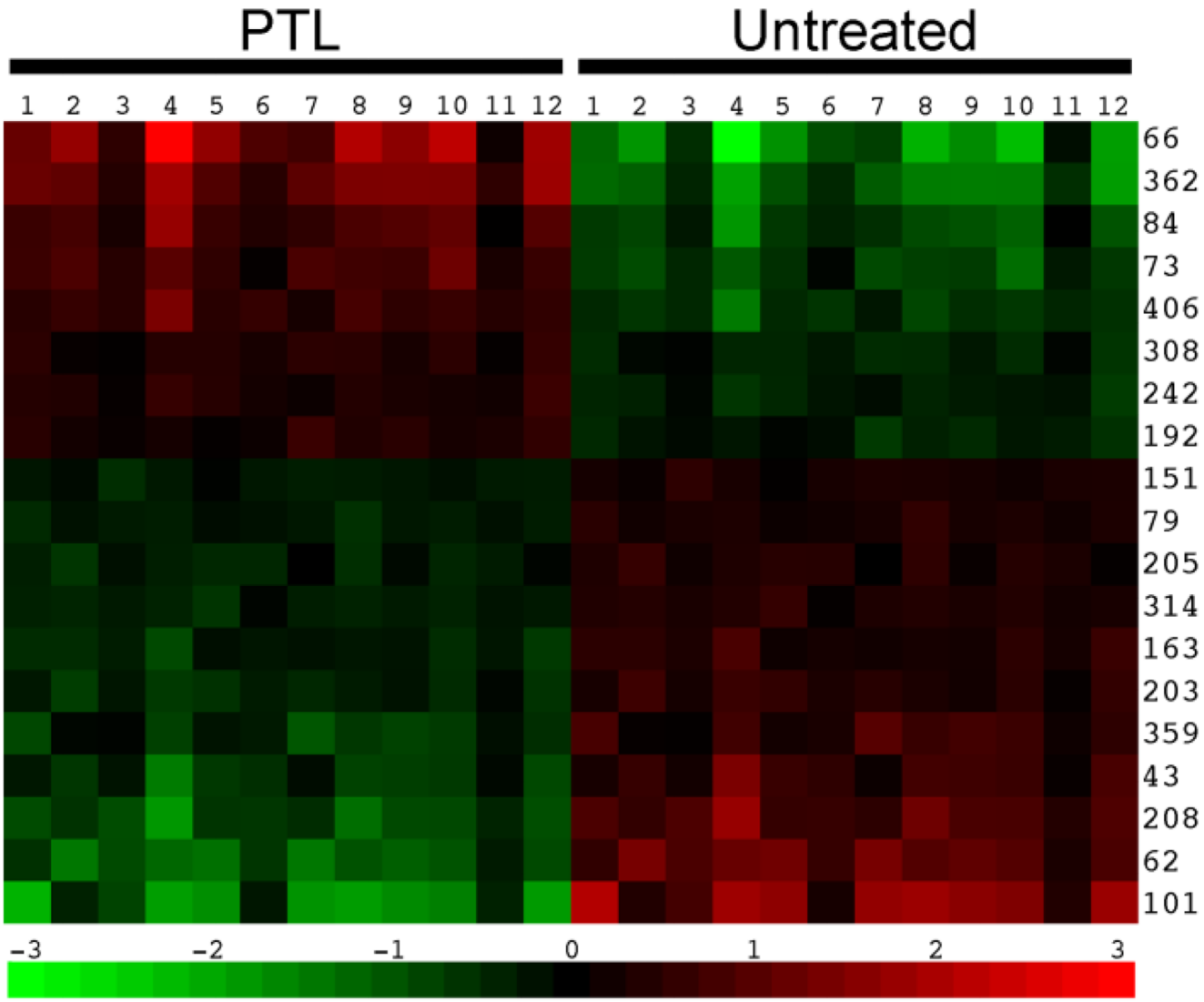

Figure 4.
Nineteen differentially expressed fundamental components between PTL-treated and untreated AML CD34^{+} cells, sorted by relative fold change. Expression units are arbitrary due to normalization and scaling.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

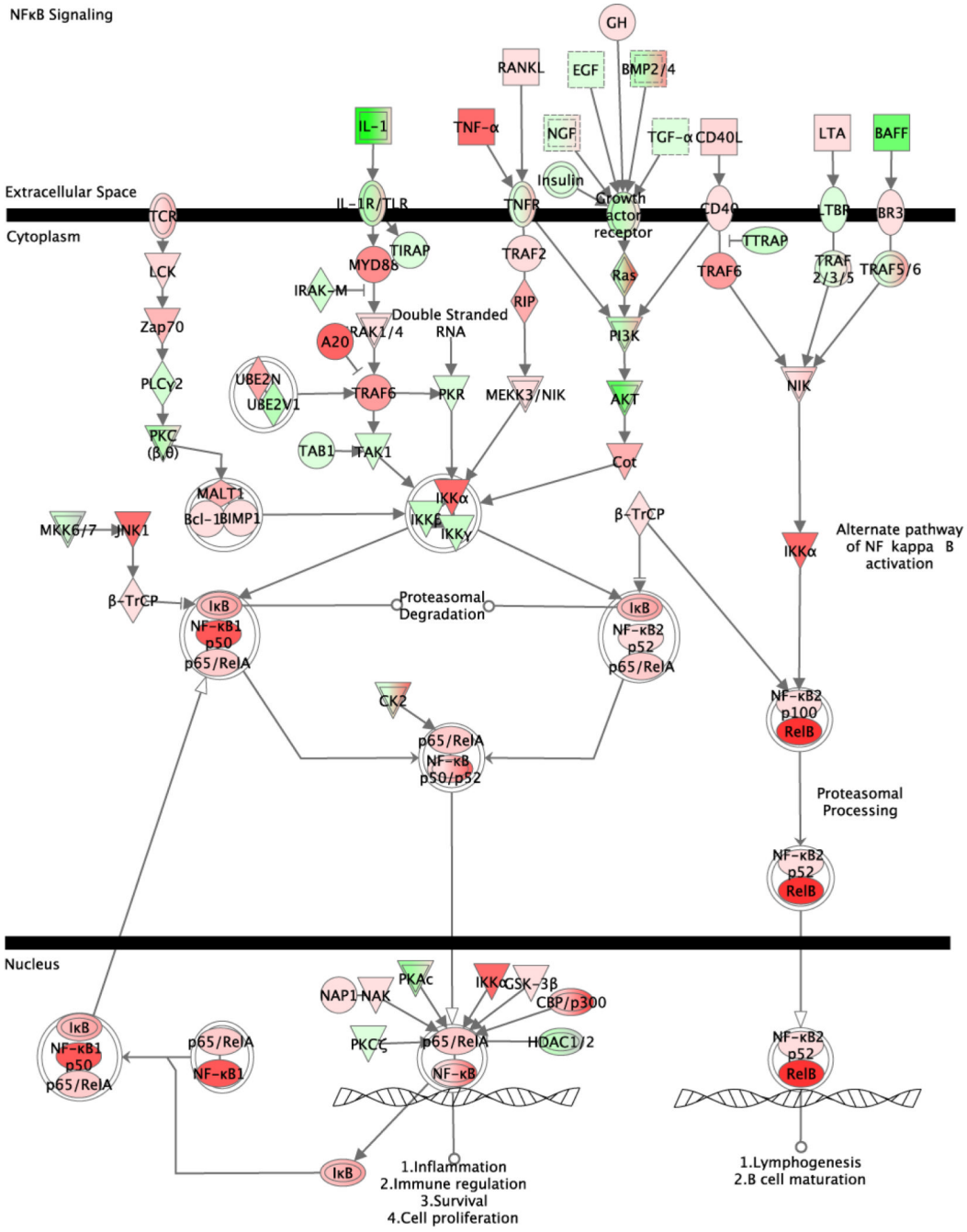

Figure 5.

FC-66 expression of the Ingenuity NF-κB Signaling canonical pathway. Ingenuity software overlaid gene expression loadings from FC-66 onto a pre-designed pathway. Red = over-expressed in PTL-treated versus untreated, green = under-expressed.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

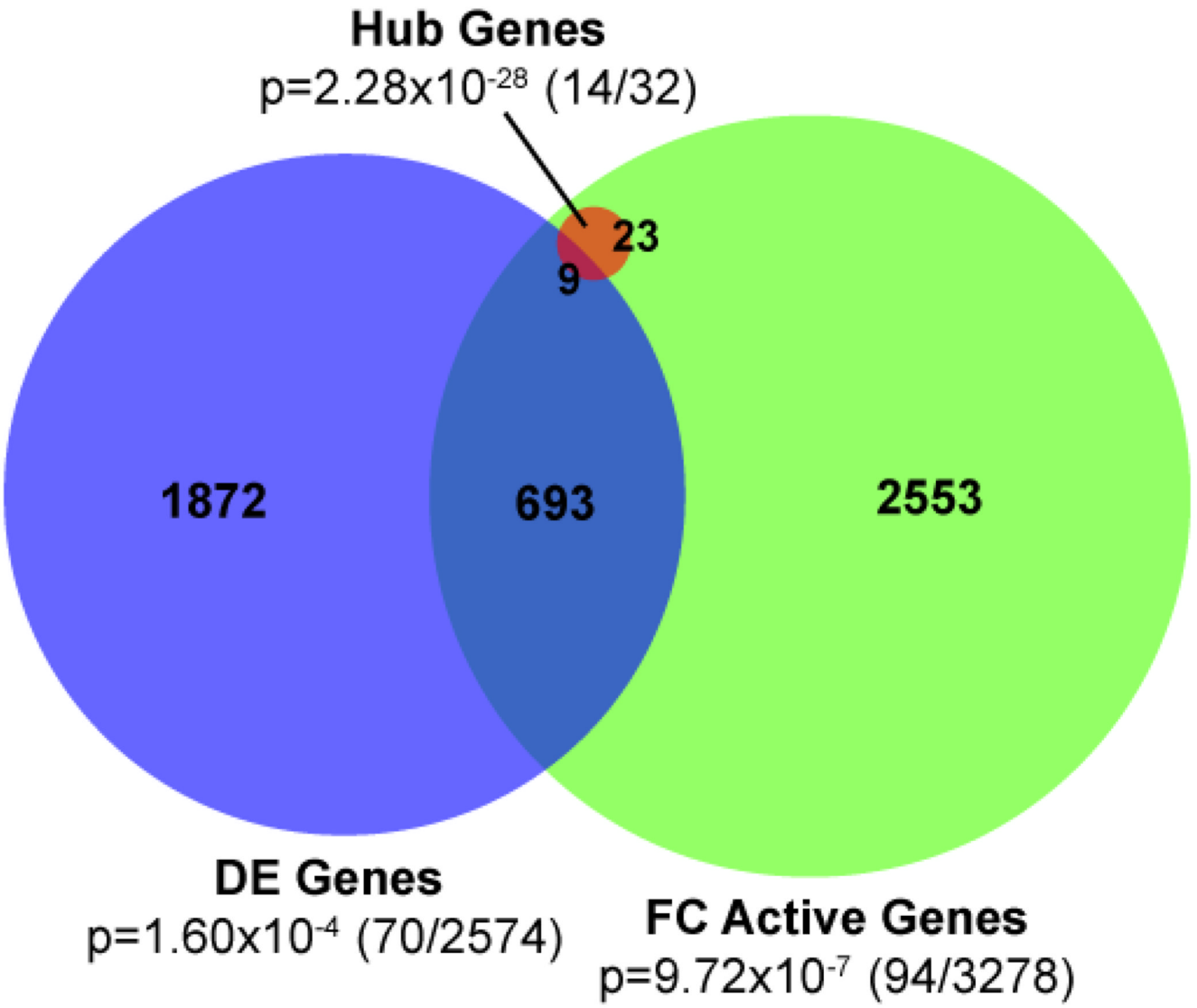

Figure 6.
Comparison of differential expression analysis of GSE7538 in gene- and fundamental component feature-space. We defined hub genes as the most highly-connected 1% of the 3,278 genes active in at least one of the differentially expressed FCs. P-values denote enrichment of known PTL-associated genes in each group, calculated using the hypergeometric test. Fractions represent the number of known PTL-associated genes and the total number of genes in each set (known/total).
Note: All figures except Figure 3 should be reproduced in color on the Web and on print.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 1

Enriched pathways for DE genes and DE experiment-specific independent components.*

Experiment	Annotation	P-Value
DE Gene Analysis	***Ingenuity Canonical Pathways (up-regulated)***	
	Protein Ubiquitination Pathway	8.51E-03
	NRF2-mediated Oxidative Stress Response	8.51E-03
	Hypoxia Signaling in the Cardiovascular System	8.51E-03
	Polyamine Regulation in Colon Cancer	1.51E-02
	Glucocorticoid Receptor Signaling	1.51E-02
	GO Categories (up-regulated)	
	organ development	4.87E-15
	multicellular organismal process	3.94E-09
	cell-cell signaling	1.18E-07
	system process	3.00E-07
	signal transduction	1.38E-05
	Ingenuity Canonical Pathways (down-regulated)	
	none	
	GO Categories (down-regulated)	
	generation of precursor metabolites and energy	1.87E-10
	cell surface receptor linked signal transduction	6.89E-09
	blood circulation	2.51E-08
	regulation of multicellular organismal process	1.33E-07
	positive regulation of biological process	3.59E-07
Experiment-spcific	***Ingenuity Canonical Pathways (up-regulated)***	
DE Component Analysis	NRF2-mediated Oxidative Stress Response	2.95E-09
	protein Ubiquitination Pathway	7.59E-03
	System	7.59E-03
	Inositol Metabolism	1.10E-02
	Polyamine Regulation in Colon Cancer	1.91E-02
	GO Categories (up-regulated)	
	response to unfolded protein	1.77E-12
	protein folding	4.63E-09

*P-values for Ingenuity Canonical Pathways have Benjamini-Hochberg correction applied. P-values for GO categories have Bonferroni correction applied.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 2

Annotations for six DE fundamental components with the highest absolute expression change.*

Component (Δexpr)**	Annotation	P-Value
66	***Ingenuity Canonical Pathways***	
(+3.11)	NF-κB Signaling	8.51E-03
	T Cell Receptor Signaling	8.51E-03
	Type I Diabetes Mellitus Signaling	8.51E-03
	IL-6 Signaling	1.51E-02
	Glucocorticoid Receptor Signaling	1.51E-02
	GO Categories	
	regulation of transcription, DNA-dependent	2.76E-06
	negative regulation of biological process	4.23E-02
	GEO experiments with high expression of this component	
GSE10609	the recurrent SET-NUP214 fusion as a new HOXA activation mechanism in pediatric T-ALL	
GSE7440	Early Response and Outcome in High-Risk Childhood Acute Lymphoblastic Leukemia: A Children's Oncology Group Study	
GSE10358	Discovery and validation of expression data for the Genomics of Acute Myeloid Leukemia Program at Washington University.	
GSE10792	Genome wide genotyping and gene expression data of childhood B-cell precursor ALL without known genetic aberration	
GSE7757	Robustness of gene expression signatures in leukemia: comparison of three distinct total RNA preparation procedures	
362	***Ingenuity Canonical Pathways***	
(+2.34)	NRF2-mediated Oxidative Stress Response	1.05E-04
	Complement System	1.23E-02
	Aryl Hydrocarbon Receptor Signaling	1.23E-02
	Glucocorticoid Receptor Signaling	2.40E-02
	Role of Pattern Recognition Receptors in Recognition of Bacteria and Viruses	2.40E-02
	GO Categories	
	response to unfolded protein	1.67E-15
	protein folding	2.86E-15
	organ development	4.48E-02
	GEO experiments with high expression of this component	
GSE7307	Human body index - transcriptional profiling	
GSE7621	Expression data of substantia nigra from postmortem human brain of Parkinson's disease patients (PD)	
GSE10315	Multipotent mesenchymal stromal cells: identification of pathways common to TGFβ3/BMP2-induced chondrogene	
GSE8977	Bone-marrow-derived mesenchymal stem cells promote breast cancer metastasis	
GSE2816	cMyb and vMyb in human monocytes	
84	***Ingenuity Canonical Pathways***	
(+1.53)	N-Glycan Biosynthesis	3.38E-10
	Endoplasmic Reticulum Stress Pathway	1.23E-04
	NRF2-mediated Oxidative Stress Response	1.23E-02
	Antigen Presentation Pathway	1.72E-02

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Component (Δexpr)**	Annotation	P-Value
	Lipid Antigen Presentation by CD1	3.03E-02
	GO Categories	
	protein folding	4.92E-10
	protein amino acid N-linked glycosylation	2.46E-07
	secretion	6.40E-05
	intracellular protein transport	2.26E-04
	cell redox homeostasis	2.90E-04
	GEO experiments with high expression of this component	
GSE10315	Multipotent mesenchymal stromal cells: identification of pathways common to TGFβ3/BMP2-induced chondrogene	
GSE6283	Specific transcriptional changes in human fetus with autosomal trisomies	
GSE6400	Cultured A549 lung cancer cells treated with actinomycin D and sapphyrin PCI-2050	
GSE6241	The effects of Serum Amyloid A on gene expression profile in HUVECs	
GSE7846	Differentially expressed genes in HEECs of eutopic endometrium of patients with endometriosis compared with control	
208	***Ingenuity Canonical Pathways***	
(−1.73)	none	
	GO Categories	
	none	
	GEO experiments with high expression of this component	
GSE2109	Expression Project for Oncology (expO)	
GSE10609	the recurrent SET-NUP214 fusion as a new HOXA activation mechanism in pediatric T-ALL	
GSE9891	Expression profile of 285 ovarian tumour samples	
GSE6532	Definition of clinically distinct molecular subtypes in estrogen receptor positive breast carcinomas using genomic grade	
GSE9151	Allergen induced gene expression of airway epithelial cells shows a possible role for TNF-α	
62	***Ingenuity Canonical Pathways***	
(−1.95)	EIF2 Signaling	1.91E-10
	Regulation of eIF4 and p70S6K Signaling	3.89E-08
	mTOR Signaling	3.72E-04
	GO Categories	
	translation	<1E-30
	translational elongation	3.99E-05
	regulation of translational initiation	4.18E-04
	ribonucleoprotein complex biogenesis	3.64E-03
	GEO experiments with high expression of this component	
GSE10358	Discovery and validation of expression data for the Genomics of Acute Myeloid Leukemia Program at Washington University.	
GSE10609	the recurrent SET-NUP214 fusion as a new HOXA activation mechanism in pediatric T-ALL	
GSE2109	Expression Project for Oncology (expO)	
GSE10792	Genome wide genotyping and gene expression data of childhood B-cell precursor ALL without known genetic aberrations	
GSE7307	Human body index - transcriptional profiling	
101	***Ingenuity Canonical Pathways***	

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Component (Δexpr)**	Annotation	P-Value
(−2.68)	none	
	GO Categories	
	defense response	2.22E-03
	icosanoid biosynthetic process	5.91E-03
	immune system process	8.37E-03
	GEO experiments with high expression of this component	
GSE10358	Discovery and validation of expression data for the Genomics of Acute Myeloid Leukemia Program at Washington University.	
GSE7757	Robustness of gene expression signatures in leukemia: comparison of three distinct total RNA preparation procedures.	
GSE8023	AML1-ETO transduced human cord blood cells, CD34 selected, compared to normal cord blood cells, CD34 selected	

* Only the top five annotations in each category are shown. For a full list, see Supplementary Table 4. P-values for Ingenuity Canonical Pathways have Benjamini-Hochberg correction applied. P-values for GO categories have Bonferroni correction applied.

** A positive expression change indicates that the component is up-regulated in PTL-treated compared to untreated.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 3

32 hub genes from DE fundamental components in PTL-treated compared to untreated AML CD34^{+} cells.*

Connections	Gene	FC(s)
47	XBP1	84
33	TNF	66
30	**LCK**	**43**
25,14,11	**CEBPB**	**192,362,406**
21	IL1B	101
18,13	CEBPA	101,406
16	**ZAP70**	**43**
15	**CNBP**	**62**
15	**CD40LG**	**43**
14,11	FOS	62,359
14	**IL1A**	**73**
14	IL10	406
14	**HGF**	**406**
13	**HSPA8**	**362**
13	**HLA-B**	**84**
12	HSPA5	84
12	NFKB1	66
11,10	EP300	66,406
11,9	STAT4	66,43
11	**HSP90AA1**	**362**
11	**LAT**	**43**
11	CD28	43
10	**IL15**	**406**
10	**CD247**	**43**
10	**CD3E**	**43**
9	PTGS2	101
9	**EIF3A**	**62**
9	**HSP90AB1**	**362**
9	**CD2**	**43**
9	**VEGFA**	**79**
9	MMP9	79
8	**CREM**	**66**

* **Bold** indicates a gene that does not appear on the text-mined list and thus represents a potentially novel prediction.