PathCORE: identifying and visualizing globally co-occurring pathways in large transcriptomic compendia

Kathleen M. Chen[1], Jie Tan[2], Gregory P. Way[1], Georgia Doing[3], Deborah A. Hogan[3], Casey S. Greene[1,]*

[1] Department of Systems Pharmacology and Translational Therapeutics. Perelman School of Medicine. University of Pennsylvania. Philadelphia PA. 19104
[2] Department of Molecular and Systems Biology. Geisel School of Medicine at Dartmouth. Hanover NH. 03755
[3] Department of Microbiology and Immunology. Geisel School of Medicine at Dartmouth. Hanover NH. 03755

[*] To whom correspondence should be addressed.
Casey Greene
University of Pennsylvania
3400 Civic Center Blvd.
Philadelphia PA 10104
Phone: 215-573-2991
Email: csgreene@upenn.edu

Abstract

Background: Investigators often interpret genome-wide data by analyzing the expression levels of genes within pathways. While this within-pathway analysis is routine, the products of any one pathway can affect the activity of other pathways. Past efforts to identify relationships between biological processes have evaluated overlap in knowledge bases or evaluated changes that occur after specific treatments. Individual experiments can highlight condition-specific pathway-pathway relationships; however, constructing a complete network of such relationships across many conditions requires analyzing results from many studies.

Results: We developed the PathCORE software to identify global pathway-pathway relationships, i.e. those evident across a broad data compendium. PathCORE starts with the results of robust feature construction algorithms, which are now being developed and applied to transcriptomic data. PathCORE identifies pathways grouped together in features more than expected by chance as functionally *co-occurring*. We performed example analyses using PathCORE for a microbial compendium for which eADAGE features were already available and a TCGA dataset of 33 cancer types that we analyzed via NMF. PathCORE recapitulated previously described pathway-pathway relationships and suggested additional relationships with biological plausibility that remain to be explored. The software also identifies genes associated with each relationship and includes demonstration source code for a web interface that users can modify to (1) visualize the network produced by applying the PathCORE software to their specific inputs and (2) review the expression levels of associated genes in the original data. The web interface can help biologists design experiments to test the relationships that were identified using PathCORE.

Conclusions: PathCORE is a hypothesis generation tool that identifies co-occurring pathways from the results of unsupervised analysis of the growing body of gene expression data and visualizes the relationships between pathways as a network. Software that steps beyond within-pathway relationships to between-pathway relationships can reveal levels of organization that have been less frequently considered.

Keywords: gene expression; unsupervised feature construction; crosstalk; pathway interactions

Background

The number of publicly available genome-wide datasets is growing rapidly [1]. High-throughput sequencing technologies that measure gene expression quickly with high accuracy and low cost continue to enable this growth [2]. Expanding public data repositories have laid the foundation for computational methods that consider entire compendia of gene expression data to extract biological patterns [3]. These patterns may be difficult to detect in measurements from a single experiment. Unsupervised approaches, which identify important signals in the data without being constrained to previously-described patterns,

may discover new expression modules and thus will complement supervised methods, particularly for exploratory analyses [4,5].

Feature extraction methods are a class of unsupervised algorithms that can reveal unannotated biological processes from genomic data [5]. Features can be constructed so that they preserve the genes in the dataset: each feature is defined by a subset of influential genes, and these genes suggest the biological or technical pattern captured by the feature. However, these features are often designed to be independent or may be considered in isolation [6,7]. When examined in the context of knowledgebases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [8], most features are significantly enriched for more than one biological gene set [5]. In this work, we refer to such a gene set by the colloquial term, pathway. It follows then that such features can be described by sets of functionally related pathways. We introduce the PathCORE (**path**way **co**-occurrence **re**lationships) software, an approach for connecting features learned from the data to known biological gene sets, e.g. pathways from KEGG or other databases.

PathCORE offers a data-driven approach for identifying and visualizing global pathway-pathway relationships. Identifying transcriptional relationships at the pathway level is a powerful strategy for developing hypotheses to disentangle complex biological phenomena and mechanistically describe cellular processes. Many cell regulatory processes are context-dependent: using large amounts of data to identify global correlations can guide the design of future experiments. In this case, relationships are drawn based on the sets of pathways, annotated in a resource of gene sets, occurring within constructed features. This means that the pathways must be perturbed in a sufficient fraction of the compendium to be captured by the model. The number of samples in the compendium used for model generation and the fraction of samples needed to observe a specific signature is expected to vary for different machine learning methods. To avoid simply discovering relationships between gene sets that share many genes, PathCORE also incorporates an optional pre-processing step that corrects for a situation Donato et al. refer to as pathway crosstalk [9]. Donato et al. recognized that pathways with shared genes were often discovered together due to overlapping genes in gene sets. We implement Donato et al.'s maximum impact estimation in a Python package separate from, but used in, PathCORE (PyPI: crosstalk-correction). With this correction, the PathCORE software allows a user to examine how pathways influence each other in a biological system based on how genes are expressed as opposed to which genes are shared.

We demonstrate PathCORE by applying the software to both a microbial and a cancer expression dataset. Briefly, for the microbial analysis we created a network of KEGG pathways from recently described ensemble Analysis using Denoising Autoencoders for Gene Expression (eADAGE) models trained on a compendium of *Pseudomonas aeruginosa* (*P. aeruginosa*) gene expression data (doi:10.5281/zenodo.583694) [5]. The pathway-pathway relationships in the eADAGE-based *P. aeruginosa* KEGG network can be viewed at https://pathcore-demo.herokuapp.com/PAO1. We provide a live demo of the PathCORE web application for this network: users can click on edges in the network to review the expression levels of associated genes in original compendium. PathCORE can be used with other feature construction approaches as well. For example, we perform PathCORE analysis of the same *P. aeruginosa* compendium using non-negative matrix factorization (NMF) on expression datasets [10,11]. We also demonstrate PathCORE's use in large cancer genomics data by creating a Pathway Interaction Database (PID) [12] pathway-pathway network of NMF features extracted from a The Cancer Genome Atlas (TCGA) pan-cancer dataset of 33 different tumor types (doi:10.5281/zenodo.56735) [13]. The network is visualized at https://pathcore-demo.herokuapp.com/TCGA; however, a demo application for the resulting PID network is not available. Note that PathCORE does not attempt to validate the quality of the underlying model; users should evaluate models to make sure that they capture biological features in their dataset before applying PathCORE. We describe several ways by which models can be evaluated in Tan et al. [5]. Datasets will vary in terms of their amenability to analysis by different model-building strategies. We assess the PathCORE software by its ability to group biologically related pathways known to interact in part through transcriptional co-regulation. PathCORE discerns biologically important pathway-pathway relationships using the features constructed (1) from both the *P.* aeruginosa gene expression compendium and the TCGA pan-cancer dataset and (2) by different feature construction approaches (eADAGE and NMF).

Related work

Our approach diverges from other algorithms that we identified in the literature in its intent: PathCORE finds pathway relationships within a biological system that are discernable in features constructed from diverse transcriptomic data, and this complements other work that developed models specific to a single condition or disease. Software to generate global networks of pathway relationships from transcriptomics data has not yet been published though approaches designed to capture pathway-pathway interactions from gene expression experiments for disease-specific, case-control studies have been published [14,15]. For example, Pham et al. developed Latent Pathway Identification Analysis to find pathways that exert latent influences on transcriptionally altered genes [16]. Under this approach, the transcriptional response profiles for a binary condition (disease/normal), in conjunction with the pathway specified in the KEGG and functions in Gene Ontology (GO) [17], are used to construct a pathway-pathway network where key pathways are identified by their network centrality scores [16]. Similarly, Pan et al. measured the betweenness centrality of pathways in disease-specific genetic interaction and coexpression networks to identify those most likely to be associated with bladder cancer risk [18]. These methods captured pathway relationships associated with a particular disease state. A global network, detailed by Li et al., relied on publicly available protein interaction data to determine pathway-pathway interactions [19]. Two pathways were connected in the network if the number of protein interactions between the pair was significant with respect to the computed background distribution. Such approaches rely on databases of interactions, though the interactions identified can be subsequently used for pathway-centric analyses of transcriptomic data [19,20].

The intention of PathCORE is to work from transcriptomic data in ways that do not give undue preference to combinations of pathways that share genes. Other methods have sought to consider shared genes between gene sets, protein-protein interactions or other curated knowledgebases to define pathway-pathway interactions [19,21–23]. For example, Glass and Girvan described another network structure that relates functional terms in GO based on shared gene annotations [24]. In contrast with this approach, PathCORE specifically removes gene overlap in pathway definitions before they are used to build a network. Our software reports pathway-pathway connections overrepresented in gene expression patterns extracted from a large transcriptomic compendium while controlling for the fact that some pathways share genes.

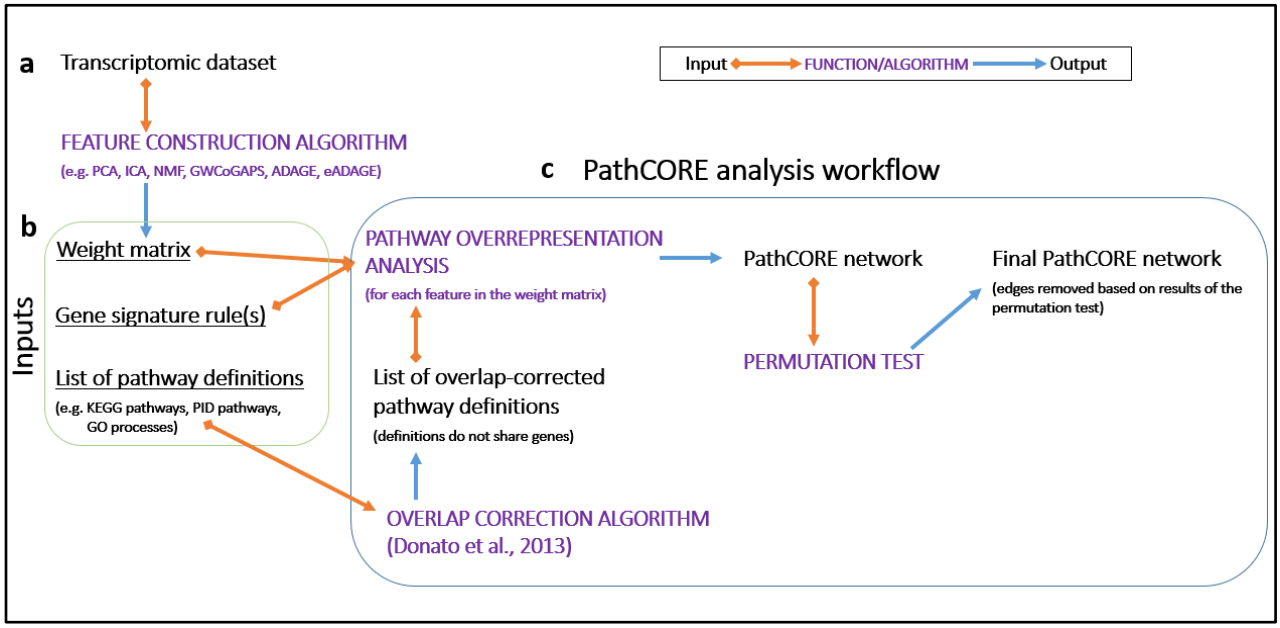

Figure 1 The PathCORE software analysis workflow.
(**a**) A user applies a feature construction algorithm to a transcriptomic dataset of genes-by-samples. The features constructed must preserve the genes in the dataset and assign weights to each of these genes according to some distribution. (**b**) Inputs required to run the complete PathCORE analysis workflow. The

constructed features are stored in a weight matrix. The user-defined gene signature rules--up to 2 for both a positive and negative gene signature--should be based on the algorithm's specified feature weight distribution. A list of pathway definitions will be used to interpret the features and build a pathway co-occurrence network. (**c**) Methods in the PathCORE analysis workflow (indicated using purple font) can be employed independently of each other so long as the necessary input(s) are provided.

Implementation

PathCORE identifies functional links between known pathways from the output of feature construction methods applied to gene expression data (Fig. 1a, b). The result is a network of pathway co-occurrence relationships that represents the grouping of biological gene sets within those features. We correct for gene overlap in the pathway annotations to avoid identifying co-occurrence relationships driven by shared genes. Additionally, PathCORE implements a permutation test for evaluating and removing edges—pathway relationships—in the resulting network that cannot be distinguished from a null model of random associations. Though we refer to the relationships in the network as co-occurrences, it is important to note that the final network displays co-occurrences that have been filtered based on this permutation test (Fig. 1c). Our software is written in Python and pip-installable (PyPI package name: pathcore), and examples of how to use the methods in PathCORE are provided in the PathCORE-analysis repository (https://github.com/greenelab/PathCORE-analysis). We recommend that those interested in using the PathCORE software consult the documentation and scripts in PathCORE-analysis. Each of the functions in PathCORE that we describe here can be used independently; however, we expect most users to employ the complete approach for interpreting pathways shared in extracted features (Fig. 1).

Data organization

PathCORE requires the following inputs:

1) A **weight matrix** that connects each gene to each feature. We expect that this will be generated by applying a feature construction algorithm to a compendium of gene expression data. In principal component analysis (PCA), this is the loadings matrix [25]; in independent component analysis (ICA), it is the unmixing matrix [26]; in ADAGE or eADAGE it is termed the weight matrix [5,7]; in NMF it is the matrix W, where the NMF approximation of the input dataset A is $A \sim WH$ [10].The primary requirements are that features must contain the full set of genes in the compendium and genes must have been assigned weights that quantify their contribution to a given feature. Accordingly, a weight matrix will have the dimensions $n \times k$, where n is the number of genes in the compendium and k is the number of features constructed.
2) A **gene signature definition**. To construct a pathway co-occurrence graph, the weight matrix must be processed into gene signatures by applying threshold(s) to the gene weights in each feature. Subsequent pathway overrepresentation will be determined by the set of genes that makes up a feature's gene signature. These are often the weights at the extremes of the distribution. How gene weights are distributed will depend on the user's selected feature construction algorithm; because of this, a user must specify criterion for including a gene in a gene signature. PathCORE permits rules for a single gene signature or both a positive and a negative gene signature. The use of 2 signatures may be appropriate when the feature construction algorithm produces positive and negative weights, the extremes of which both characterize a feature (e.g. PCA, ICA, ADAGE or eADAGE). Because a feature can have more than one gene signature, we maintain a distinction between a feature and a feature's gene signature(s).
3) A list of **pathway definitions**, where each pathway contains a set of genes (e.g. KEGG pathways, PID pathways, GO biological processes). We provide the files for the *P. aeruginosa* KEGG pathway definitions (for the eADAGE-based analysis of the *P. aeruginosa* gene compendium, https://goo.gl/hm4amF) and the Nature-NCI PID pathway definitions (for the NMF-based analysis of the TCGA dataset, https://goo.gl/n5GZTA) in the PathCORE analysis repository (https://github.com/greenelab/PathCORE-analysis).

Weight matrix construction and signature definition

In practice, users can obtain a weight matrix from many different methods. For the purposes of this paper, we demonstrate generality by constructing weight matrices via eADAGE and NMF.

eADAGE

eADAGE is an unsupervised feature construction algorithm developed by Tan et al. [5] that uses an ensemble of neural networks (an ensemble of ADAGE models) to capture biological signatures embedded in the expression compendium. Using data-driven and knowledge-driven heuristics, Tan et al. determined that a model of k=300 features was an appropriate size for the current *P. aeruginosa* compendium. By initializing eADAGE with different random seeds, Tan et al. produced 10 eADAGE models that each extracted k=300 features from the compendium of genome-scale *P. aeruginosa* data. Because PathCORE supports the aggregation of co-occurrence networks created from different models on the same input data, we use all 10 of these models in the PathCORE analysis of eADAGE models (doi:10.5281/zenodo.583172).

Tan et al. refers to the features constructed by eADAGE as nodes. They are represented as a weight matrix of size $n \times k$, where n genes in the compendium are assigned positive or negative gene weights, according to a standard normal distribution, for each feature k. Tan et al. determined that the gene sets contributing the highest positive or highest negative weights (+/- 2.5 standard deviations) to a feature described gene expression patterns across the compendium, and thus referred to the gene sets as signatures. Because a feature's positive and negative gene signatures did not necessarily correspond to the same biological processes or functions, Tan et al. analyzed each of these sets separately [5]. Tan et al.'s gene signature rules are specified as an input to the PathCORE analysis as well.

NMF

We also constructed NMF models for the *P. aeruginosa* dataset and the TCGA pan-cancer dataset. Given an NMF approximation of $A \sim WH$ [10], where A is the input expression dataset of size $n \times s$ (n genes by s samples), NMF aims to find the optimal reconstruction of A by WH such that W clusters on samples (size $n \times k$) and H clusters on genes (size $k \times s$). In order to match the number of features constructed in each eADAGE model by Tan et al., we set k, the desired number of features, to be 300 and used W as the input weight matrix for the PathCORE software. We found that the gene weight distribution of an NMF feature is right-skewed and (as the name suggests) non-negative (Fig. S1). In this case, we defined a single gene signature rule: an NMF feature's gene signature is the set of genes with weights 2.0 standard deviations above the mean weight of the feature.

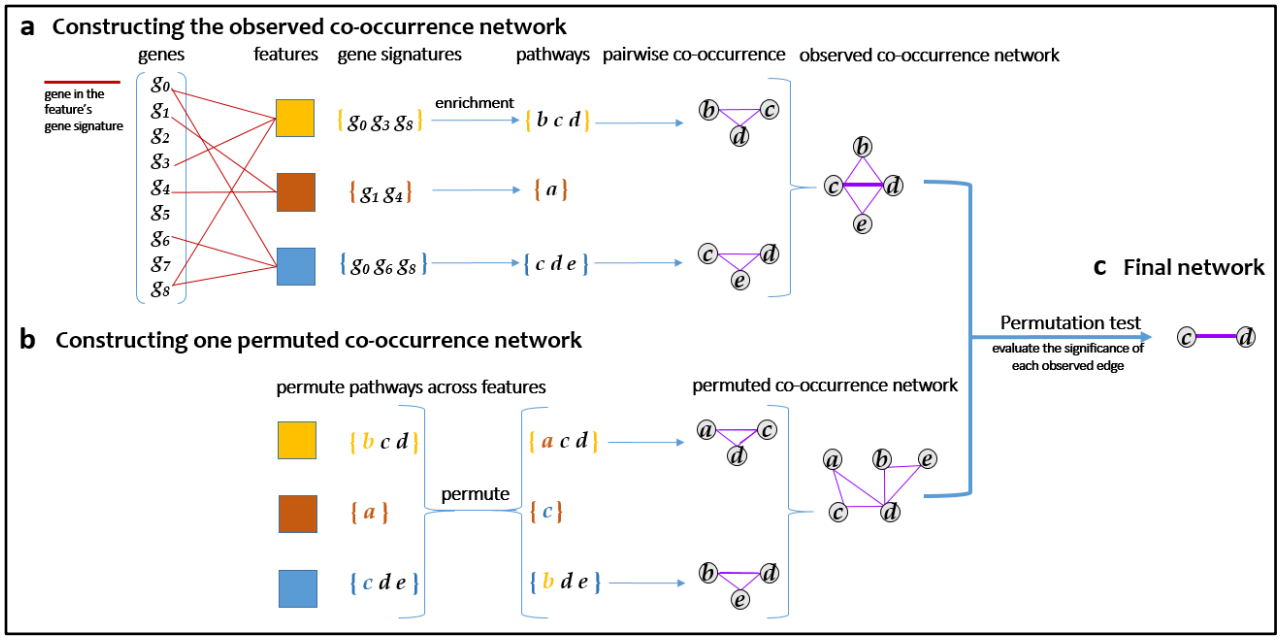

Figure 2 The approach implemented in PathCORE to construct a pathway co-occurrence network from an expression compendium.
(**a**) A user-selected feature extraction method is applied to expression data. Such methods assign each gene a weight, according to some distribution, that represents the gene's contribution to the feature. The set of genes that are considered highly representative of a feature's function is referred to as a feature's gene signature. The gene signature is user-defined and should be based on the weight distribution produced by the unsupervised method of choice. In the event that the weight distribution contains both positive and negative values, a user can specify criteria for both a positive and negative gene signature. A test of pathway enrichment is applied to identify corresponding sets of pathways from the gene signature(s) in a feature. We consider pathways significantly overrepresented in the same feature to *co-occur*. Pairwise co-occurrence relationships are used to build a network. Each edge is weighted by the number of features containing both pathways.
(**b**) *N* permuted networks are generated to assess the statistical significance of a co-occurrence relation in the graph. Here, we show the construction of one such permuted network. Two invariants are maintained during a permutation: (1) pathway side-specificity (if applicable, e.g. positive and negative gene signatures) and (2) the number of distinct pathways in a feature's gene signature.
(**c**) For each edge observed in the co-occurrence network, we compare its weight against the weight distribution generated from *N* (default: 10,000) permutations of the network to determine each edge's p-value. After correcting the p-value by the number of edges observed in the graph using the Benjamini—Hochberg procedure, only an edge with an adjusted p-value below alpha (default: 0.05) is kept in the final co-occurrence network.

<u>Construction of a pathway co-occurrence network</u>

We employ a Fisher's exact test to determine the pathways significantly associated with each gene signature. When considering significance of a particular pathway, the two categories of gene classification are as follows: (1) presence or absence of the gene in the gene signature and (2) presence or absence of the gene in the pathway definition. We specify a contingency table for each pathway in the input list of pathway definitions and calculate the p-value. For each gene signature over all pathways, a pathway's p-value is corrected using the Benjamini—Hochberg procedure to produce a feature-wise FDR. A pathway with an adjusted p-value that is less than the user-settable false discovery rate (FDR) significance cutoff, alpha (default: 0.05), is considered significantly enriched. PathCORE in practice applies a subsequent permutation test over pathway-pathway edge weights, so this cutoff value should be selected to most aid user interpretation of the model. Two pathways co-occur, or share an edge in the pathway co-occurrence network, if they are both overrepresented in a gene signature. The number of

times such a pathway pair is present over all gene signatures in a model corresponds to its edge weight in the pathway-pathway graph (Fig. 2a).

More specifically, construction of the pathway co-occurrence graph requires the construction of a bipartite graph between 2 sets of vertices: the model's gene signatures and the pathways of interest. A gene signature and a pathway share an edge if the pathway is overrepresented in the gene signature. If any path through a single signature exists between two pathways in the bipartite graph, they share an edge in the resulting pathway co-occurrence graph. The weight of the edge in the co-occurrence graph, before the permutation test is applied, is the number of paths between the two pathways in the intermediate bipartite graph.

Permutation test

The network that results from the preceding method is densely connected, and many edges may be spurious. To remove correlations that cannot be distinguished from random pathway associations, we define a statistical test that determines whether a pathway-pathway relationship appearing *x* times in a *k*-feature model is unexpected under the null hypothesis—the null hypothesis being that the relationship does not appear more often than it would in a random network. We create *N* weighted null networks, where each null network is constructed by permuting overrepresented pathways across the model's gene signatures while preserving the number of pathways for which each gene signature is enriched (Fig. 2b). In the case where we have positive and negative gene signatures, overrepresentation can be positive or negative. Because certain pathways may display bias toward one side—for example, a pathway may be overrepresented more often in features' positive gene signatures—we perform the permutation separately for each side. The *N* random networks produce the background weight distribution for every observed edge; significance can then be assessed by comparing the true (observed) edge weight against the null. The p-value for each edge *e* is calculated by summing the number of times a random network contained *e* at a weight greater than or equal to its observed edge weight and dividing this value by *N*. Following Benjamini—Hochberg FDR correction by the number of edges in the observed network, pathway-pathway relationships with adjusted p-values above alpha (user-settable default: 0.05) are removed from the network of co-occurring pathways (Fig. 2c). The threshold alpha value is a configurable parameter, and the user should select an FDR that best balances the costs and consequences of false positives. For highly exploratory analyses in which it may be helpful to have more speculative edges, this value can be raised. For analyses that require particularly stringent control, it can be lowered.

Because the expected weight of every edge can be determined from the *N* random networks (take the sum of the background weight distribution for an edge and divide it by *N*), we can divide the observed edge weights by their respective expected weights (divide by 1 if the expected edge weight is 0 based on the *N* permutations). Edges in the final network are weighted by their odds ratios.

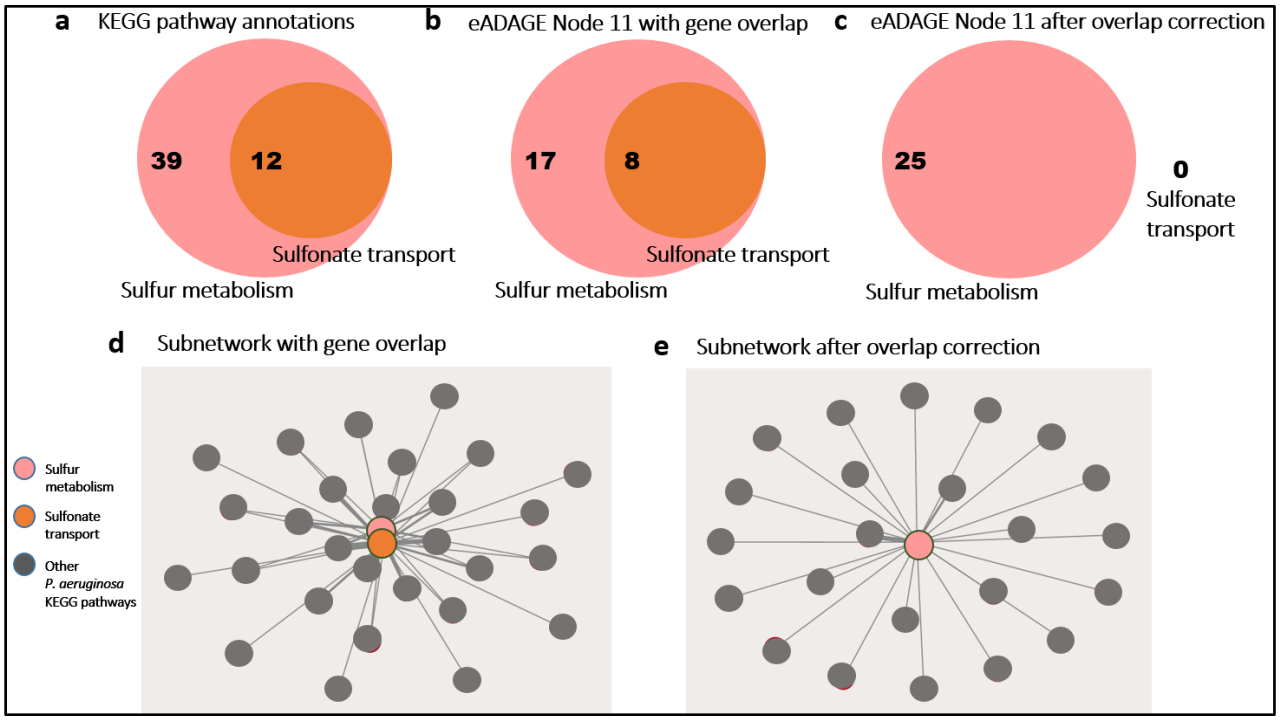

Figure 3 Correcting for gene overlap results in a sparser pathway co-occurrence network. (**a**) The KEGG pathway annotations for the sulfonate transport system are a subset of those for sulfur metabolism. 12 genes annotated to the sulfonate transport system are also annotated to sulfur metabolism. (**b**) Without applying the overlap correction procedure, 25 of the genes in the positive and negative gene signatures of the eADAGE feature "Node 11" are annotated to sulfur metabolism--of those, 8 genes are annotated to the sulfonate transport system as well. (**c**) All 8 of the overlapping genes are mapped to the sulfur metabolism pathway after overlap correction.
(**d**) A co-occurrence network built without applying the overlap correction procedure will report co-occurrence between the sulfonate transport system and sulfur metabolism, whereas (**e**) no such relation is identified after overlap correction.

<u>Gene overlap correction</u>

Pathways can co-occur because of shared genes (Fig. 3a, b, d). Though some approaches use the overlap of genes to identify connected pathways, we sought to capture pairs of pathways that persisted even when this overlap was removed. The phenomenon of observing enrichment of multiple pathways due to gene overlap has been previously termed as "crosstalk," and Donato et al. have developed a method to correct for it [9]. Due to confusion around the term, we refer to this as *overlapping genes* in this work, except where specifically referencing Donato et al. Their approach, called maximum impact estimation, begins with a membership matrix indicating the original assignment of multiple genes to multiple pathways. It uses expectation maximization to estimate the pathway in which a gene contributes its greatest predicted impact (its maximum impact) and assigns the gene only to this pathway [9]. This provides a set of new pathway definitions that no longer share genes (Fig. 3c, e).

We provide an implementation of Donato et al.'s maximum impact estimation as a Python package separate from PathCORE so that it is available for any pathway analyses (PyPI package name: crosstalk-correction). The procedure is written using NumPy functions and data structures, which allows for efficient implementation of array and matrix operations in Python [27]. In PathCORE, overlapping genes are resolved before pathway overrepresentation analysis so that the resulting pathway co-occurrence network identifies relationships that are not driven simply by the same genes being annotated to multiple pathways. Without this correction step, it is difficult to determine whether a co-occurrence relationship can be attributed to the features extracted from expression data or gene overlap in the two pathway annotations. We incorporate this correction into the PathCORE workflow by default; however,

users interested in using PathCORE to find connections between overlapping gene sets can choose to disable the correction step.

Figure 4 A web application used to analyze pathway-pathway relationships in the eADAGE-based, *P. aeruginosa* KEGG network.
(**a**) A user clicks on an edge (a pathway-pathway relationship) in the network visualization and (**b**) is directed to a page that displays expression data from the original transcriptomic dataset specific to the selected edge (https://goo.gl/Hs5A3e). The expression data is visualized as two heatmaps that indicate the fifteen most and fifteen least expressed samples corresponding to the edge. To select the "most" and "least" expressed samples, we assign each sample a summary "expression score." The expression score is based on the expression values of the genes (limited to the top twenty genes with an odds ratio above 1) annotated to one or both of the pathways. Here, we show the heatmap of least expressed samples specific to the [Phosphate transport - Type II general secretion] relationship. (**c**) Clicking on a square in the heatmap directs a user to an experiment page (https://goo.gl/KYNhwB) based on the sample corresponding to that square. A user can use the experiment page to identify whether the expression values of genes specific to an edge and a selected sample differ from those recorded in other samples of the experiment. In this experiment page, the first three samples (labeled in black) are *P. aeruginosa* "baseline" replicates grown for 72 h in drop-flow biofilm reactors. The following three samples (labeled in blue) are *P. aeruginosa* grown for an additional 12 h (84 h total). Labels in blue indicate that the three 84 h replicates are in the heatmap of least expressed samples displayed on the [Phosphate transport – Type II general secretion] edge page.

PathCORE network visualization and support for experimental follow-up

We chose to represent pathway-pathway relationships in a graph because it followed easily from the way our collaborators consider pathway co-occurrences. The PathCORE analysis workflow outputs a list of pathway-pathway relationships, or edges in a network visualization, as a text file. An example of the KEGG *P. aeruginosa* edges file is available for download on the demo application: http://pathcore-demo.herokuapp.com/quickview. Users can use this file output to visualize the identified relationships as an adjacency matrix or in any other format they choose.

As an optional step, users can set up a Flask application for each PathCORE network. Metadata gathered from the analysis are saved to TSV files, and we use a script to populate collections in a MongoDB database with this information. The co-occurrence network is rendered using the D3.js force-directed graph layout [28]. Users can select a pathway-pathway relationship in the network to view a new page containing details about the genes annotated to one or both pathways (Fig. 4a).

We created a web interface for deeper examination of relationships present in the pathway co-occurrence network. When presented with a visualization of the PathCORE network, our collaborators suggested that additional support for determining potential gene targets and experimental conditions would help them design experiments to validate novel relationships. The details we included in an edge-specific page address their suggestion by (1) highlighting up to twenty genes--annotated to either of the two pathways in the edge--contained in features that also contain this edge, after controlling for the total number of features that contain each gene, and (2) displaying the expression levels of these genes in each of the fifteen samples where they were most and least expressed. The quantity of information (twenty genes, thirty samples total) we choose to include in an edge page is intentionally limited so that users can review it in a reasonable amount of time.

To implement the functionality in (1), we computed an odds ratio for every gene annotated to one or both pathways in the edge. The odds ratio measures how often we observe a feature enriched for both the given gene and the edge of interest relative to how often we would expect to see this occurrence. We calculate the proportion of observed cases and divide by the expected proportion--equivalent to the frequency of the edge appearing in the model's features.

Let K be the number of features from which the PathCORE network was built. K_G is the number of features that contain gene G (i.e. G is in feature K's gene signature), K_E the number of features that contain edge E (i.e. the two pathways connected by E are overrepresented in feature K), and $K_{G \& E}$ the number of features that contain both gene G and edge E. The odds ratio is computed as follows:

$$\text{Observed} = K_{G \& E} / K_G$$

$$\text{Expected} = K_E / K$$

$$\text{Odds ratio} = \text{Observed} / \text{Expected}$$

An odds ratio above 1 suggests that the gene is more likely to appear in features enriched for this pair of pathways: we rank the genes by their odds ratio to highlight genes most observed with the co-occurrence relationship.

The information specified in (2) requires an "expression score" for every sample. A sample expression score is calculated using the genes we selected in goal (1): it is the average of the normalized gene expression values weighted by the normalized gene odds ratio. Selection of the most and least expressed samples is based on these scores. We use two heatmaps to show the (maximum of twenty) genes' expression values in each of the fifteen most and least expressed samples (Fig. 4b).

For each sample in an edge page, a user can examine how the expression values of the edge's twenty genes in that sample compare to those recorded for all other samples in the dataset that are from the same experiment (Fig. 4c). Genes that display distinct expression patterns under a specific setting may be good candidates for follow-up studies.

Results

<u>Interpreting features extracted by unsupervised clustering algorithms with PathCORE.</u>

Networks modeling the relationships between curated processes in a biological system offer a means for developing new hypotheses about which pathways influence each other and when. PathCORE creates a network of globally co-occurring pathways based on features observed in a compendium of gene expression data. Biological patterns in the data are extracted by a feature construction algorithm such as PCA [25], ICA [6], NMF [10], GWCoGAPS [4], or eADAGE [5]. These algorithms capture sources of variability in the data that induce coordinated changes in gene expression as features. The genes that contribute the most to these features covary. This provides a data-driven categorization of the biological system that can then be analyzed at the pathway-level by identifying annotated pathways overrepresented in each feature.

The methods we implement in PathCORE can be used independently of each other (Fig. 1c). Here, we present analyses that can be produced by applying the full PathCORE pipeline to models created from a transcriptomic compendium by an unsupervised feature construction algorithm. Input pathway definitions are “overlap-corrected” (correcting for gene overlap between definitions) for each feature before enrichment analysis. An overlap-corrected, weighted pathway co-occurrence network is built by connecting the pairs of pathways that are overrepresented in features of the model. Finally, we remove edges that cannot be distinguished from a null model of random associations based on the results of a permutation test.

PathCORE also offers support for users interested in experimentally verifying a pathway-pathway relationship (Fig. 4). We provide the code for setting up a web application where the network can be visualized and its edges analyzed using the original input information. A pathway-pathway edge page contains 2 heatmaps that display the samples in the compendium where the underlying genes are most and least expressed. When available, information about each sample can be included on the page so that users can refer to the conditions in which the expression patterns occurred (Fig. 4c).

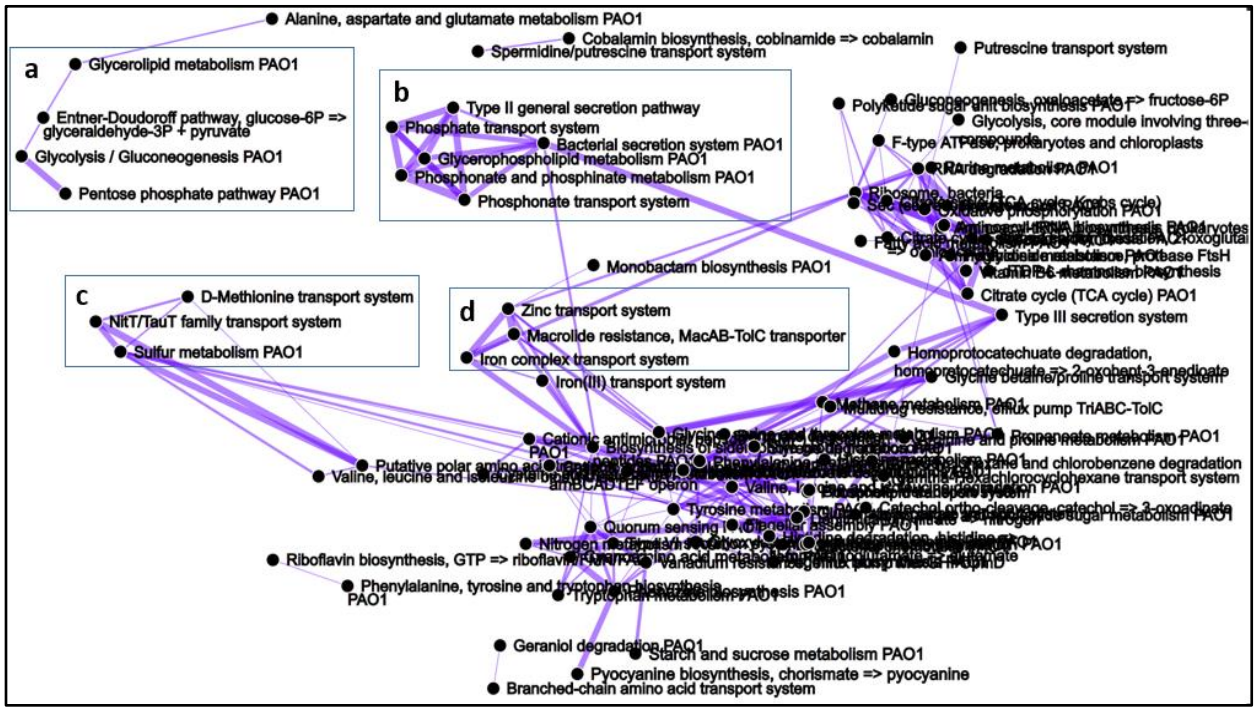

Figure 5 eADAGE features constructed from publicly available *P. aeruginosa* expression data describe known KEGG pathway-pathway relationships.
(**a**) The glycerolipid metabolism, Entner-Doudoroff, glycolysis/gluconeogenesis, and pentose phosphate, pathways share common functions related to glucose catabolism .
(**b**) Organophosphate and inorganic phosphate transport- and metabolism-related processes frequently co-occur with bacterial secretion systems. Here, we observe pairwise relationships between type II general secretion and phosphate-related processes.
(**c**) Pathways involved in the catabolism of sulfur-containing molecules--taurine (NitT/TauT family transport) and methionine (D-Methionine transport), and the general sulfur metabolism process--are functionally linked.
(**d**) The zinc transport, iron complex transport, and MacAB-TolC transporter systems are pairwise connected.
This network can be viewed at https://pathcore-demo.herokuapp.com/PAO1. The list of KEGG pathway-pathway relationships visualized in the network is available at the specified link (press Ctrl + L on the page) and as a supplemental file for this paper.

PathCORE identifies relationships between KEGG pathways in *P. aeruginosa* using features extracted from publicly available *P. aeruginosa* gene expression experiments

We used PathCORE to create a network of co-occurring pathways out of the expression signatures extracted from a *P. aeruginosa* compendium. For every feature, overlap correction was applied to the *P. aeruginosa* KEGG pathway annotations and overlap-corrected annotations were used in the overrepresentation analysis. PathCORE aggregates multiple networks by taking the union of the edges across all networks and summing the weights of common pathway-pathway connections. We do this to emphasize the co-occurrence relationships that are more stable [29]—that is, the relationships that appear across multiple models. Finally, we removed edges in the aggregate network that were not significant after FDR correction when compared to the background distributions generated from 10,000 permutations of the network. We applied PathCORE to features built by both NMF (Fig. S2) and eADAGE (discussed below). The pathway-pathway network generated by NMF [30] is smaller than that generated by eADAGE; that is, there are fewer pathway-pathway connections in the network. It is possible that this is due to a difference in the stability of the results from the two algorithms or the comprehensiveness of the features extracted by each approach. eADAGE includes an ensemble step that improves model consistency. The eADAGE authors also observed that models constructed by this ensemble procedure also more comprehensively captured pathways than non-ensemble models [5]. The PathCORE analysis of a 300-feature NMF decomposition of the *P. aeruginosa* compendium produced a KEGG network that is similar in size to the PID network (Fig. S2, 6).

The eADAGE co-occurrence network identifies a number of pathway-pathway relationships that have been previously characterized (Fig. 5). This suggests that PathCORE can capture functional links between biological pathways. Three glucose catabolism processes co-occur in the network: glycolysis, pentose phosphate, and the Entner-Doudoroff pathway (Fig. 5a). We also found a cluster relating organophosphate and inorganic phosphate transport- and metabolism-related processes (Fig. 5b). Notably, phosphate uptake and acquisition genes are directly connected to the *hxc* genes that encode a type II secretion system. Studies in *P. aeruginosa* clinical isolates demonstrated that the Hxc secretion system is responsible for the secretion of alkaline phosphatases, which are phosphate scavenging enzymes [31,32] and the phosphate binding DING protein [33]. Furthermore, alkaline phosphatases, DING and the *hxc* genes are regulated by the transcription factor PhoB which is most active in response to phosphate limitation. The identification of this relationship by PathCORE as a global pattern suggests the role of type II secretion and phosphate limitation seen in a limited number of isolates may be generalizable to broader *P. aeruginosa* biology. As shown in Fig. 5c, we also identified linkages between two pathways involved in the catabolism of sulfur-containing molecules, taurine and methionine, and the general sulfur metabolism process. Other connections between pathways involved in the transport of iron (ferrienterobactin binding) [34] and zinc (the *znu* uptake system [35]) were identified (Fig. 5d). Interestingly, genes identified in the edge between the zinc transport and MacAB-TolC pathways include the *pvd* genes involved in pyoverdine biosynthesis and regulation, a putative periplasmic metal binding protein, as well as other components of an ABC transporter (genes PA2407, PA2408, and PA2409 at https://goo.gl/bfqOk8) [36]. While there are intuitive similarities between the biochemical functions of zinc and iron transport, the identification of the two pathways as correlated in transcriptional data by PathCORE suggests a relationship in their regulation that may help characterize the coordinated cellular responses to these two distinct micronutrients.

We used the PathCORE web application for the eADAGE KEGG *P. aeruginosa* network (https://pathcore-demo.herokuapp.com/PAO1) to analyze the connection between the phosphate transport system and a type II general secretion system pathway (Fig. 3a, b; edge page at https://goo.gl/Hs5A3e). The sixteen genes reported on the edge page all have odds ratios above 34; these genes are at least 34 times more likely to appear in the gene signatures in which both of these pathways are overrepresented. Such genes may help to reveal the biological basis of the co-occurrence relationship. In this case, the results suggest that there may be some overlap between machinery for transporting phosphate into the cell and secreting substances out of the cell via type II secretion. Alternatively, the two processes may be mechanistically separate but coregulated such that phosphate scavenging molecules may be secreted by type II secretion coincidentally to aid in phosphate acquisition. The heatmap of the fifteen least expressed samples shows that the *pstB* and *pstS* genes, annotated to the phosphate transport system, are consistently expressed higher relative to the other genes in the edge

for these samples. The *pstS, pstC, pstA,* and *pstB* genes are proximal of the phosphate-specific transport (Pst) operon that encodes a high-affinity orthophosphate transport system [37]. Future studies will examine whether deletion of the Pst phosphate system impairs secretion by the type II secretion system.

To assess whether the relationships identified in the pathway analysis paralleled gene expression patterns in the context of a published experiment, we looked across experiments to determine if the genes contained in the edge were co-regulated. As an example of the types of relationships that we observed, we present a single experiment, E-GEOD-22164 from Folsom et al. (Fig. 3c; experiment page at https://goo.gl/KYNhwB), that contains data from two sample types with three replicates each [38]. One set of samples, referred to as the baseline, is of *P. aeruginosa* grown for 72 h in drop-flow biofilm reactors. The other is of *P. aeruginosa* grown for an additional 12 h (84 h total). We found that three of the samples with the lowest expression of the genes within the shared edge were samples from the later timepoint. All six of these samples (three replicates each) are displayed in the experiment page. The sixteen edge genes showed differential expression between the 72 h and 84 h timepoints. Particularly in the case of *pstA* and *pstC*, the baseline replicates had normalized expression values near the center of the range from the compendium whereas the 84 h samples had expression levels at the low end of the range. This suggests that genes involved in both phosphate transport and type II secretion are less expressed at 84 h compared to 72 h. Future studies will determine if this is due to a physiological change in the biofilm cells at the late time point such that phosphate demands were lower or different sources of phosphate become available.

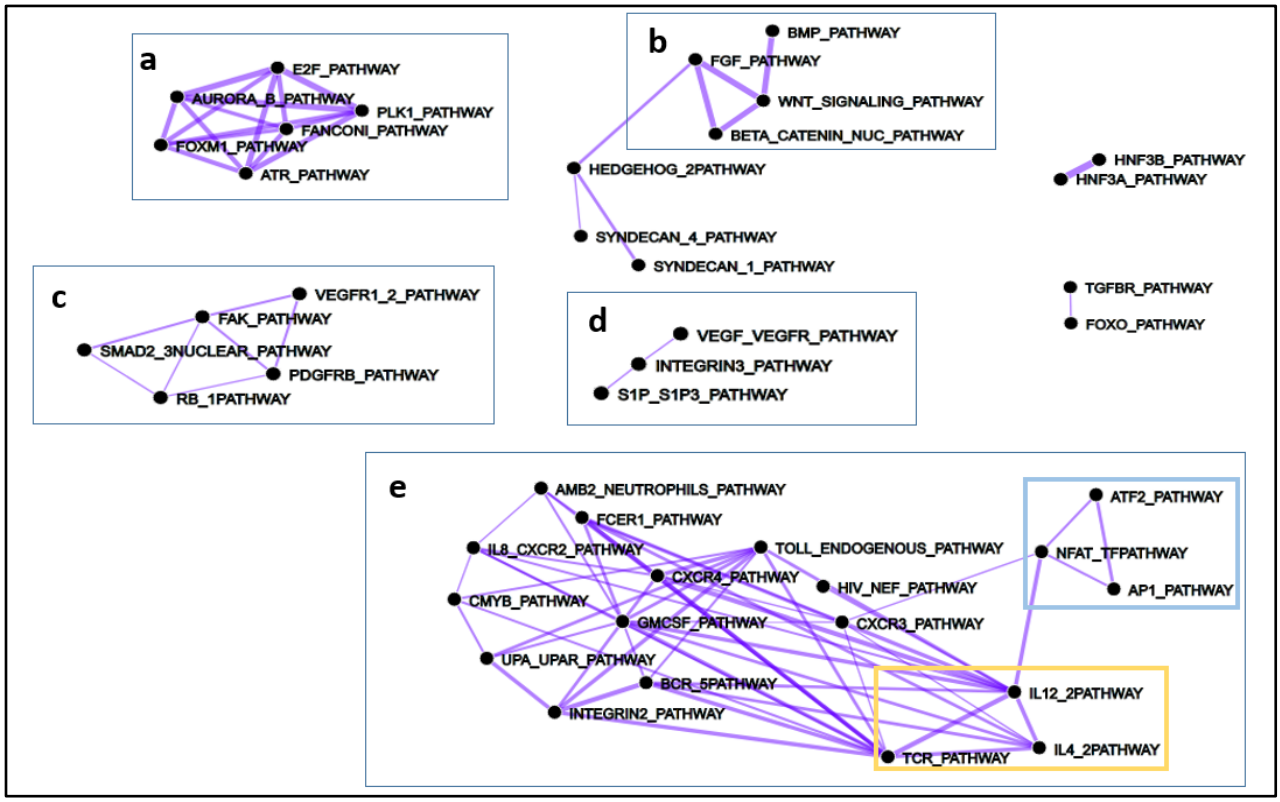

Figure 6 PID pathway-pathway relationships discovered in NMF features constructed from the TCGA pan-cancer gene expression dataset.
(**a**) Pathways in this module are responsible for cell cycle progression.
(**b**) Wnt signaling interactions with nuclear Beta-catenin signaling, FGF signaling, and BMP signaling have all been linked to cancer progression.
(**c**) Here, we observe functional links between pathways responsible for angiogenesis and those responsible for cell proliferation.
(**d**) The VEGF-VEGFR pathway interacts with the S1P3 pathway through Beta3 integrins.
(**e**) This module contains many relationships related to immune system processes. The interaction cycle

formed by T-Cell Receptor (TCR) signaling in naïve CD4+ T cells and IL-12/IL-4 mediated signaling events, outlined in yellow, is one well-known example. The cycle in blue is formed by the ATF2, NFAT, and AP1 pathways; pairwise co-occurrence of these three transcription factor networks may suggest that dysregulation of any one of these pathways can trigger variable oncogenic processes in the immune system.
This network can be viewed at https://pathcore-demo.herokuapp.com/TCGA. The list of PID pathway-pathway relationships visualized in the network is available at the specified link (press Ctrl + L on the page) and as a supplemental file for this paper.

PathCORE identifies relationships among PID pathways from the TCGA pan-cancer gene expression dataset.

PathCORE is not specific to a certain dataset or organism. We also constructed a 300-feature NMF model of TCGA pan-cancer gene expression, which is comprised of 33 different cancer-types from various organ sites, and applied the PathCORE software to those features. We chose NMF because it has been used in previous studies to identify biologically relevant patterns in transcriptomic data [10] and by many studies to derive molecular subtypes [39–41]. The 300 NMF features were analyzed using overlap-corrected PID pathways, a collection of 196 human cell signaling pathways with a particular focus on processes relevant to cancer [12].

We found that PathCORE detects modules of co-occurring pathways consistent with our current understanding of cancer-related interactions (Fig. 6). Importantly, because the connections were constructed from many different cancer-types, these modules may represent pathway-pathway interactions present in a large proportion of all tumors and may be good candidates for targeted treatments.

For example, a module composed of a FoxM1 transcription factor network, an E2F transcription factor network, Aurora B kinase signaling, ATR signaling, PLK1 signaling, and members of the Fanconi anemia DNA damage response pathway are densely connected (Fig 6a). When two pathways share an edge in the co-occurrence network, they are overrepresented together in one or more features. The connections in this module recapitulate well known cancer hallmarks including cellular proliferation pathways and markers of genome instability, such as the activation of DNA damage response pathways [42]. We found that several pairwise pathway co-occurrences correspond with previously reported pathway-pathway interactions [43–45]. We also observed a hub of pathways interacting with Wnt signaling (Fig. 6b). In our network, pathways that co-occur with Wnt signaling include the regulation of nuclear Beta-catenin signaling, FGF signaling, and BMP signaling. The Wnt and BMP pathways are functionally integrated in biological processes that contribute to cancer progression [46]. Additionally, Wnt/Beta-catenin signaling is a well-studied regulatory system, and the effects of mutations in Wnt pathway components on this system have been linked to tumorigenesis [47]. Wnt/Beta-catenin and FGF together influence the directional migration of cancer cell clusters [48].

Two modules in the network relate to angiogenesis, or the formation of new blood vessels (Fig. 6c, d). Tumors transmit signals that stimulate angiogenesis because a blood supply provides the necessary oxygen and nutrients for their growth and proliferation. One module relates angiogenesis factors to cell proliferation. This module connects the following pathways: PDGFR-beta signaling, FAK-mediated signaling events, VEGFR1 and VEGFR2-mediated signaling events, nuclear SMAD2/3 signaling regulation, and RB1 regulation (Fig. 6c). These functional connections are known to be involved in tumor proliferation [49–51]. The other module indicates a direct relationship by which the VEGF pathway interacts with the S1P3 pathway through Beta3 integrins (Fig. 6d). *S1P3* is a known regulator of angiogenesis [52], and has been demonstrated to be associated with treatment-resistant breast cancer and poor survival [53]. Moreover, this interaction module has been observed to promote endothelial cell migration in human umbilical veins [54]. Taken together, this independent module may suggest a distinct angiogenesis process activated in more aggressive and metastatic tumors that is disrupted and regulated by alternative mechanisms [55].

Finally, PathCORE revealed a large, densely connected module of immune related pathways (Fig. 6e). We found that this module contains many co-occurrence relationships that align with immune

system processes. One such example is the well characterized interaction cycle formed by T-Cell Receptor (TCR) signaling in naïve CD4+ T cells and IL-12/IL-4 mediated signaling events [56–58]. At the same time, PathCORE identifies additional immune-related relationships. We observed a cycle between the three transcription factor networks: ATF-2, AP-1, and CaN-regulated NFAT-dependent transcription. These pathways can take on different, often opposing, functions depending on the tissue and subcellular context. For example, ATF-2 can be an oncogene in one context (e.g. melanoma) and a tumor suppressor in another (e.g. breast cancer) [59]. AP-1, comprised of Jun/Fos proteins, is associated with both tumorigenesis and tumor suppression due to its roles in cell survival, proliferation, and cell death [60]. Moreover, NFAT in complex with AP-1 regulates immune cell differentiation, but dysregulation of NFAT signaling can lead to malignant growth and tumor metastasis [61]. The functional association observed between the ATF-2, AP-1, and NFAT cycle together within the immunity module might suggest that dysregulation within this cycle has profound consequences for immune cell processes and may trigger variable oncogenic processes.

Conclusions

Unsupervised methods can identify previously undiscovered patterns in large collections of data. PathCORE overlays curated knowledge after feature construction to help researchers interpret constructed features in the context of existing knowledgebases. Specifically, PathCORE aims to clarify how expert-annotated gene sets work together from a gene expression perspective.

Gene set analyses can be heavily confounded by shared genes. Some pathways may be observed together because they depend on each other, while others may simply contain some of the same genes. In PathCORE, pathway annotations undergo a procedure called maximum impact estimation, described in a publication by Donato et al., that maps each gene in each feature to the one pathway in which it has the greatest estimated impact [9]. We provide this overlap correction algorithm as a Python package (PyPI package name: crosstalk-correction) available under the BSD 3-Clause license. Though the algorithm had been described, no publicly available implementation existed.

PathCORE includes software for analysis and visualization and can be broadly applied to constructed features. We demonstrate PathCORE in two different contexts, analyses of the bacterium *P. aeruginosa* and human pan-cancer datasets, using two different feature construction methods (eADAGE and NMF). We provide a demonstration application containing the results of the eADAGE *P. aeruginosa* analysis for researchers to explore. For each edge, users can explore heatmaps displaying the expression levels of potential driver genes in the original samples. This provides support for assessing computationally-derived relationships in experimental follow-ups.

Unsupervised analyses of genome-scale datasets that summarize key patterns in the data have the potential to improve our understanding of how a biological system operates via complex interactions between molecular processes. However, interpreting the features generated by unsupervised approaches is still challenging. PathCORE is a component of a software ecosystem that connects the features extracted from data to curated resources. It is important to note that PathCORE will only be able to identify pathways that occur in features of the underlying model. For these pathways to be mapped to features, they will need to be transcriptionally perturbed in at least some subset of the compendium. PathCORE is post-hoc analysis software that can be applied to models already known to capture biological patterns from gene expression compendia. A graph resulting from PathCORE can help to identify global pathway-pathway relationships—a baseline graph—that complements existing work to identify interactions between pathways in the context of a specific disease. The specific niche that PathCORE aims to fill is in revealing to researchers which gene sets are most closely related to each other in machine learning-based models of gene expression, which genes play a role in this co-occurrence, and which conditions drive this relationship.

Project name: PathCORE
Project home page: https://pathcore-demo.herokuapp.com
Archived version: https://github.com/greenelab/PathCORE-analysis/releases/tag/v1.1.0 (links to download .zip and .tar.gz files are provided here)
Operating system: Platform independent

Programming language: Python
Other requirements: Python 3 or higher
License: BSD 3-clause

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
Not applicable

Availability of data and materials
Files for each of the PathCORE networks described in the results are provided in Supplementary material.

Data sets

P. aeruginosa eADAGE models: doi:10.5281/zenodo.583172

P. aeruginosa gene expression compendium: doi:10.5281/zenodo.583694

The normalized gene expression compendium provided in this Zenodo record contains datasets on the GPL84 from ArrayExpress as of 31 July 2015. It includes 1,051 samples grown in 78 distinct medium conditions, 128 distinct strains and isolates, and dozens of different environmental parameters [5]. Tan et al. compiled this dataset and used it to construct the 10 eADAGE models from which we generate the eADAGE-based *P. aeruginosa* KEGG network [5]. We use this same data compendium to generate the NMF *P. aeruginosa* model. The script used to download and process those datasets into the compendium is available at https://goo.gl/YjOEQl.

TCGA pan-cancer dataset: doi:10.5281/zenodo.56735

The pan-cancer dataset was compiled using data from all 33 TCGA cohorts. It was generated by the TCGA Research Network: http://cancergenome.nih.gov/ [13]. RNA-seq data was downloaded on 8 March 2016 from UCSC Xena (https://xenabrowser.net/datapages/). Gene expression was measured using the Illumina HiSeq technology. More information and the latest version of the dataset can be found at https://xenabrowser.net/datapages/?dataset=TCGA.PANCAN.sampleMap/HiSeqV2&host=https://tcga.xenahubs.net.

Source code

PathCORE analysis: (https://github.com/greenelab/PathCORE-analysis/tree/v1.1.0) This repository contains all the scripts to reproduce the analyses described in this paper. The Python scripts here should be used as a starting point for new PathCORE analyses. Instructions for setting up a web application for a user's specific PathCORE analysis are provided in this repository's README.

Overlap correction: (https://github.com/kathyxchen/crosstalk-correction/tree/v1.0.4) Donato et. al's procedure for overlap correction [9] is a pip-installable Python package 'crosstalk-correction' that is separate from, but listed as a dependency in, PathCORE. It is implemented using NumPy [27].

PathCORE methods: (https://github.com/greenelab/PathCORE/tree/v1.0.2) The methods included in the PathCORE analysis workflow (Fig. 4c) are provided as a pip-installable Python package 'pathcore'. It is implemented using Pandas [62], SciPy (specifically, scipy.stats) [63], StatsModels [64], and the crosstalk-correction package.

PathCORE demo application: (https://github.com/kathyxchen/PathCORE-demo/tree/v2.0) The project home page, pathcore-demo.herokuapp.com provides links to

(1) The **web application** for the eADAGE-based KEGG *P. aeruginosa* described in the first case study.
(2) A **view** of the NMF-based PID pathway co-occurrence network described in the second case study.
(3) A **quick view page** where users can temporarily load and visualize their own network file (generated from the PathCORE analysis).

Competing interests

The authors declare that they have no competing interests

Funding

KMC was funded by an undergraduate research grant from the Penn Institute for Biomedical Informatics. CSG was funded in part by a grant from the Gordon and Betty Moore Foundation (GBMF 4552). JT, GD, DAH, and CSG were funded in part by a pilot grant from the Cystic Fibrosis Foundation (STANTO15R0). DAH was funded in part by R01-AI091702.

Authors' contributions

KMC implemented the software, performed the analyses, and drafted the manuscript. JT and GPW contributed computational reagents. KMC, DAH, and CSG designed the project. KMC, JT, GPW, GD, DAH, and CSG interpreted the results. JT, GPW, GD, DAH, and CSG provided critical feedback and revisions on the manuscript. JT and GPW reviewed source code.

Acknowledgements

The authors would also like to thank Daniel Himmelstein, Kurt Wheeler, and René Zelaya for helping to review the source code.

References

1. Greene CS, Foster JA, Stanton BA, Hogan DA, Bromberg Y. Computational Approaches to Study Microbes and Microbiomes. Pac. Symp. Biocomput. 2016;21:557.

2. Tatlow PJ, Piccolo SR. A cloud-based workflow to quantify transcript-expression levels in public cancer compendia. Sci. Rep. 2016;6.

3. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, et al. ArrayExpress update—simplifying data submissions. Nucleic Acids Res. 2014;gku1057.

4. Stein-O'Brien G, Carey J, Lee W, Considine M, Favorov A, Flam E, et al. PatternMarkers and Genome-Wide CoGAPS Analysis in Parallel Sets (GWCoGAPS) for data-driven detection of novel biomarkers via whole transcriptome Non-negative matrix factorization (NMF). bioRxiv. 2016;083717.

5. Tan J, Doing G, Lewis KA, Price CE, Chen KM, Cady KC, et al. Unsupervised Extraction of Stable Expression Signatures from Public Compendia with an Ensemble of Neural Networks. Cell Syst. 2017;5:63–71.e6.

6. Engreitz JM, Daigle BJ, Marshall JJ, Altman RB. Independent component analysis: mining microarray data for fundamental human gene expression modules. J. Biomed. Inform. 2010;43:932–44.

7. Tan J, Hammond JH, Hogan DA, Greene CS. Adage-based integration of publicly available *Pseudomonas aeruginosa* gene expression data with denoising autoencoders illuminates microbe-host interactions. mSystems. 2016;1:25.

8. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000;28:27–30.

9. Donato M, Xu Z, Tomoiaga A, Granneman JG, MacKenzie RG, Bao R, et al. Analysis and correction of crosstalk effects in pathway analysis. Genome Res. 2013;23:1885–93.

10. Brunet J-P, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proc. Natl. Acad. Sci. 2004;101:4164–9.

11. Kim PM, Tidor B. Subsystem identification through dimensionality reduction of large-scale gene expression data. Genome Res. 2003;13:1706–18.

12. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, et al. PID: the Pathway Interaction Database. Nucleic Acids Res. 2009;37:D679.

13. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. 2013;45:1113–20.

14. Visakh R, Nazeer KA. Identifying epigenetically dysregulated pathways from pathway–pathway interaction networks. Comput. Biol. Med. 2016;76:160–7.

15. Yang J-B, Luo R, Yan Y, Chen Y. Differential pathway network analysis used to identify key pathways associated with pediatric pneumonia. Microb. Pathog. 2016;101:50–5.

16. Pham L, Christadore L, Schaus S, Kolaczyk ED. Network-based prediction for sources of transcriptional dysregulation using latent pathway identification analysis. Proc. Natl. Acad. Sci. 2011;108:13347–52.

17. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat. Genet. 2000;25:25–9.

18. Pan Q, Hu T, Andrew AS, Karagas MR, Moore JH. Bladder cancer specific pathway interaction networks. ECAL. Citeseer; 2013. p. 94–101.

19. Li Y, Agarwal P, Rajagopalan D. A global pathway crosstalk network. Bioinformatics. 2008;24:1442–7.

20. Anda-Juregui G de, Meja-Pedroza RA, Espinal-Enrquez J, Hernndez-Lemus E. Crosstalk events in the estrogen signaling pathway may affect tamoxifen efficacy in breast cancer molecular subtypes. Comput. Biol. Chem. 2015;59:42–54.

21. Bell L, Chowdhary R, Liu JS, Niu X, Zhang J. Integrated bio-entity network: a system for biological knowledge discovery. PloS One. 2011;6:e21474.

22. Chowbina SR, Wu X, Zhang F, Li PM, Pandey R, Kasamsetty HN, et al. HPD: an online integrated human pathway database enabling systems biology studies. BMC Bioinformatics. 2009;10:S5.

23. Wu X, Chen JY. Molecular interaction networks: topological and functional characterizations. Autom. Proteomics Genomics Eng. Case-Based Approach. 2009;145.

24. Glass K, Girvan M. Finding new order in biological functions from the network structure of gene annotations. PLoS Comput Biol. 2015;11:e1004565.

25. Abdi H, Williams LJ. Principal component analysis. Wiley Interdiscip. Rev. Comput. Stat. 2010;2:433–59.

26. Stone JV. Independent component analysis. Wiley Online Library; 2004.

27. Walt S van der, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation. Comput. Sci. Eng. 2011;13:22–30.

28. Bostock M. D3. js. Data Driven Doc. 2012;492.

29. Yu B. Stability. Bernoulli. 2013;19:1484–500.

30. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 2011;12:2825–30.

31. Liu X, Long D, You H, Yang D, Zhou S, Zhang S, et al. Phosphatidylcholine affects the secretion of the alkaline phosphatase PhoA in *Pseudomonas* strains. Microbiol. Res. 2016;192:21–9.

32. Ball G, Durand ric, Lazdunski A, Filloux A. A novel type II secretion system in *Pseudomonas aeruginosa*. Mol. Microbiol. 2002;43:475–85.

33. Ball G, Viarre V, Garvis S, Voulhoux R, Filloux A. Type II-dependent secretion of a *Pseudomonas aeruginosa* DING protein. Res. Microbiol. 2012;163:457–69.

34. Stephens DL, Choe MD, Earhart CF. *Escherichia coli* periplasmic protein FepB binds ferrienterobactin. Microbiology. 1995;141:1647–54.

35. Ellison ML, IIII JMF, Parrish W, Danell AS, Pesci EC. The transcriptional regulator Np20 is the zinc uptake regulator in *Pseudomonas aeruginosa*. PloS One. 2013;8:e75389.

36. Winsor GL, Lam DK, Fleming L, Lo R, Whiteside MD, Nancy YY, et al. *Pseudomonas* Genome Database: improved comparative analysis and population genomics capability for *Pseudomonas* genomes. Nucleic Acids Res. 2010;gkq869.

37. Rico-Jimnez M, Reyes-Darias JA, Ortega lvaro, Pea AID, Morel B, Krell T. Two different mechanisms mediate chemotaxis to inorganic phosphate in *Pseudomonas aeruginosa*. Sci. Rep. 2016;6.

38. Folsom JP, Richards L, Pitts B, Roe F, Ehrlich GD, Parker A, et al. Physiology of *Pseudomonas aeruginosa* in biofilms as revealed by transcriptome analysis. BMC Microbiol. 2010;10:294.

39. Bailey P, Chang DK, Nones K, Johns AL, Patch A-M, Gingras M-C, et al. Genomic analyses identify molecular subtypes of pancreatic cancer. Nature. 2016;

40. Network CGAR. Integrated genomic analyses of ovarian carcinoma. Nature. 2011;474:609–15.

41. Yuan Y, Allen EMV, Omberg L, Wagle N, Amin-Mansour A, Sokolov A, et al. Assessing the clinical utility of cancer genomic and proteomic data across tumor types. Nat. Biotechnol. 2014;32:644–52.

42. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144:646–74.

43. Moldovan G-L, D'Andrea AD. How the Fanconi Anemia pathway guards the genome. Annu. Rev. Genet. 2009;43:223–49.

44. Sadasivam S, DeCaprio JA. The DREAM complex: master coordinator of cell cycle-dependent gene expression. Nat. Rev. Cancer. 2013;13:585–95.

45. Tomida J, Itaya A, Shigechi T, Unno J, Uchida E, Ikura M, et al. A novel interplay between the Fanconi Anemia core complex and ATR-ATRIP kinase during DNA cross-link repair. Nucleic Acids Res. 2013;41:6930–41.

46. Itasaki N, Hoppler S. Crosstalk between Wnt and bone morphogenic protein signaling: a turbulent relationship. Dev. Dyn. 2010;239:16–33.

47. MacDonald BT, Tamai K, He X. Wnt/β-catenin signaling: components, mechanisms, and diseases. Dev. Cell. 2009;17:9–26.

48. Aman A, Piotrowski T. Wnt/β-catenin and Fgf signaling control collective cell migration by restricting chemokine receptor expression. Dev. Cell. 2008;15:749–61.

49. Petersen M, Pardali E, Horst GVD, Cheung H, Hoogen CVD, Pluijm GVD, et al. Smad2 and Smad3 have opposing roles in breast cancer bone metastasis by differentially affecting tumor angiogenesis. Oncogene. 2010;29:1351–61.

50. Yoon H, Dehart JP, Murphy JM, Lim S-TS. Understanding the roles of FAK in cancer: inhibitors, genetic models, and new insights. J. Histochem. Cytochem. 2015;63:114–28.

51. Chinnam M, Goodrich DW. RB1, development, and cancer. Curr. Top. Dev. Biol. 2011;94:129.

52. Takuwa Y, Du W, Qi X, Okamoto Y, Takuwa N, Yoshioka K. Roles of sphingosine-1-phosphate signaling in angiogenesis. World J. Biol. Chem. 2010;1:298–306.

53. Watson C, Long JS, Orange C, Tannahill CL, Mallon E, McGlynn LM, et al. High expression of sphingosine 1-phosphate receptors, S1P 1 and S1P 3, sphingosine kinase 1, and extracellular signal-regulated kinase-1/2 is associated with development of tamoxifen resistance in estrogen receptor-positive breast cancer patients. Am. J. Pathol. 2010;177:2205–15.

54. Paik JH, Chae S, Lee M-J, Thangada S, Hla T. Sphingosine 1-phosphate-induced endothelial cell migration requires the expression of EDG-1 and EDG-3 receptors and Rho-dependent activation of αvβ3- and β1-containing integrins. J. Biol. Chem. 2001;276:11830–7.

55. Serini G, Valdembri D, Bussolino F. Integrins and angiogenesis: a sticky business. Exp. Cell Res. 2006;312:651–8.

56. Brogdon JL, Leitenberg D, Bottomly K. The potency of TCR signaling differentially regulates NFATc/p activity and early IL-4 transcription in naive CD4 T cells. J. Immunol. 2002;168:3825–32.

57. Hsieh C-S, Macatonia SE, Tripp CS, Wolf SF, O'Garra A, Murphy KM. Development of TH1 CD4 T Cells Through IL-12. Science. 1993;260:547.

58. Vacaflores A, Chapman NM, Harty JT, Richer MJ, Houtman JC. Exposure of Human CD4 T Cells to IL-12 Results in Enhanced TCR-Induced Cytokine Production, Altered TCR Signaling, and Increased Oxidative Metabolism. PloS One. 2016;11:e0157175.

59. Lau E, Ze'ev AR. ATF2–at the crossroad of nuclear and cytosolic functions. J Cell Sci. 2012;125:2815–24.

60. Shaulian E, Karin M. AP-1 as a regulator of cell life and death. Nat. Cell Biol. 2002;4:E136.

61. Mller MR, Rao A. NFAT, immunity and cancer: a transcription factor comes of age. Nat. Rev. Immunol. 2010;10:645–56.

62. McKinney W. Data structures for statistical computing in python. Proc. 9th Python Sci. Conf. van der Voort S, Millman J; 2010. p. 51–6.

63. Jones E, Oliphant T, Peterson P. SciPy: open source scientific tools for Python. 2014;

64. Seabold S, Perktold J. Statsmodels: Econometric and statistical modeling with python. Proc. 9th Python Sci. Conf. 2010. p. 61.