PathCORE-T: 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

Co-author e-mail addresses:
Kathleen M. Chen: kchen@flatironinstitute.org
Jie Tan: tj8901nm@gmail.com
Gregory P. Way: gregway@mail.med.upenn.edu
Georgia Doing: Georgia.Doing.GR@dartmouth.edu
Deborah A. Hogan: Deborah.A.Hogan@dartmouth.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 PathCORE-T framework by implementing existing methods to identify pathway-pathway transcriptional relationships evident across a broad data compendium. PathCORE-T is applied to the output of feature construction algorithms; it identifies pairs of pathways observed in features more than expected by chance as *functionally co-occurring*. We demonstrate PathCORE-T by analyzing an existing eADAGE model of a microbial compendium and building and analyzing NMF features from the TCGA dataset of 33 cancer types. The PathCORE-T framework includes a demonstration web interface, with source code, that users can launch to (1) visualize the network and (2) review the expression levels of associated genes in the original data. PathCORE-T creates and displays the network of globally co-occurring pathways based on features observed in a machine learning analysis of gene expression data.

Conclusions: The PathCORE-T framework identifies transcriptionally co-occurring pathways from the results of unsupervised analysis of gene expression data and visualizes the relationships between pathways as a network. PathCORE-T recapitulated previously described pathway-pathway relationships and suggested experimentally testable additional hypotheses that remain to be explored.

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 [3, 4] have laid the foundation for computational methods that consider entire compendia of gene expression data to extract biological patterns [5]. 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 [6,7].

Feature extraction methods are a class of unsupervised algorithms that can reveal unannotated biological processes from genomic data [7]. Each feature can be defined by a subset of influential genes, and these genes suggest the biological or technical pattern captured by the feature. These features, like pathways, are often considered individually [7,8]. When examined in the context of knowledgebases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [9], most features are significantly enriched for more than one biological gene set [7]. 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-T (identifying ***path***way ***co***-occurrence ***re***lationships in ***t***ranscriptomic data) software, which implements existing methods that jointly consider features and gene sets to map pathways with shared transcriptional responses.

PathCORE-T offers a data-driven approach for identifying and visualizing transcriptional pathway-pathway relationships. In this case, relationships are drawn based on the sets of pathways, annotated in a resource of gene sets, occurring within constructed features. Because PathCORE-T starts from a feature extraction model, the number of samples in the compendium used for model generation and the fraction of samples needed to observe a specific biological or technical pattern is expected to vary by feature extraction method. Pathways must be perturbed in a sufficient fraction of experiments in the data compendium to be captured by any such method. To avoid discovering relationships between pathways that share many genes—which could more easily be discovered by directly comparing pathway membership—we implement an optional pre-processing step that corrects for genes shared between gene sets, which Donato et al. refer to as pathway crosstalk [10]. Donato et al.'s correction method, maximum impact estimation, has not previously been implemented in open source software. We have released our implementation of maximum impact estimation as its own Python package (PyPI package name: crosstalk-correction) so that it can be used independently of PathCORE-T. Applying this correction in PathCORE-T

software allows a user to examine relationships between gene sets based on how genes are expressed as opposed to which genes are shared.

We apply PathCORE-T to a microbial and a cancer expression dataset, each analyzed using different feature extraction methods, to demonstrate its broad applicability. 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) [7]. We provide a live demo of the PathCORE-T web application for this network: users can click on edges in the network to review the expression levels of associated genes in the original compendium (https://pathcore-demo.herokuapp.com/PAO1). To show its use outside of the microbial space, we also demonstrate PathCORE-T analysis of Pathway Interaction Database (PID)-annotated [11] non-negative matrix factorization (NMF) features [12, 13] extracted from The Cancer Genome Atlas's (TCGA) pan-cancer dataset of 33 different tumor types (doi:10.5281/zenodo.56735) [14].

In addition to visualizing the results of these two applications, the PathCORE-T web interface (https://pathcore-demo.herokuapp.com/) links to the documentation and source code for our implementation and example usage of PathCORE-T. Methods implemented in PathCORE-T are written in Python and pip-installable (PyPI package name: PathCORE-T). Examples of how to use these methods are provided in the PathCORE-T analysis repository (https://github.com/greenelab/PathCORE-T-analysis). In addition to scripts that reproduce the eADAGE and NMF analyses described in this paper, the PathCORE-T-analysis repository includes a Jupyter notebook (https://goo.gl/VuzN12) with step-by-step descriptions for the complete PathCORE-T framework.

Related work

Our approach diverges from other algorithms that we identified in the literature in its intent: PathCORE-T finds pathway pairs within a biological system that are overrepresented in features constructed from diverse transcriptomic data. This complements other work that developed models specific to a single condition or disease. Approaches designed to capture pathway-pathway interactions from gene expression experiments for disease-specific, case-control studies have been published [15,16]. For example, Pham et al. developed Latent Pathway Identification Analysis to find pathways that exert latent influences on transcriptionally altered genes [17]. Under this approach, the transcriptional response profiles for a binary condition (disease/normal), in conjunction with pathways specified in the KEGG and functions in Gene Ontology (GO) [18], are used to construct a pathway-pathway network where key pathways are identified by their network centrality scores [17]. 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 [19]. These methods captured pathway relationships associated with a particular disease state.

Global networks identify relationships between pathways that are not disease- or condition-specific. One such network, detailed by Li et al., relied on publicly available protein interaction data to determine pathway-pathway interactions [20]. 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 [20, 21]. Pita-Juárez et al. created the Pathway Coexpression Network (PCxN) as a tool to discover pathways correlated with a pathway of interest [22]. They estimated correlations between pathways based on the expression of their underlying genes (as annotated in MSigDB) across a curated compendium of microarray data [22]. Software like PathCORE-T that generates global networks of pathway relationships from unsupervised feature analysis models built using transcriptomics data has not yet been published.

The intention of PathCORE-T 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 [20–21, 23–25]. For example, Glass and Girvan described another network structure that relates functional terms in GO based on shared gene annotations [26]. In contrast with this approach, PathCORE-T

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.

Implementation

PathCORE-T 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-T implements a permutation test for evaluating and removing edges—pathway-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-T), and examples of how to use the methods in PathCORE-T are provided in the PathCORE-T-analysis repository (https://github.com/greenelab/PathCORE-T-analysis). We recommend that those interested in using the PathCORE-T software consult the documentation and scripts in PathCORE-T-analysis. Each of the functions in PathCORE-T 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-T requires the following inputs:

1) **A weight matrix** that connects each gene to each feature. We expect that this results from the application of a feature construction algorithm to a compendium of gene expression data. 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 x k, where n is the number of genes in the compendium and k is the number of features constructed. In principal component analysis (PCA), this is the loadings matrix [27]; in independent component analysis (ICA), it is the unmixing matrix [28]; 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 $A \sim WH$ [12]. In addition to the scripts we provide for the eADAGE and NMF examples in the PathCORE-T analysis repository, we include a Jupyter notebook (https://goo.gl/VuzN12) that demonstrates how a weight matrix can be constructed by applying ICA to the *P. aeruginosa* gene compendium.
2) **Gene signature rule(s)**. To construct a pathway co-occurrence network, the weight matrix must be processed into gene signatures by applying threshold(s) to the gene weights in each feature—we refer to these as gene signature rules. 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-T 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 is defined by a set of genes (e.g. KEGG pathways, PID pathways, GO biological processes). We provide the files for the *P. aeruginosa* KEGG pathway definitions and the Nature-NCI PID pathway definitions in the PathCORE-T analysis repository (https://github.com/greenelab/PathCORE-T-analysis/tree/master/data).

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. [7] that uses an ensemble of neural networks (an ensemble of ADAGE models) to capture biological patterns embedded in the expression compendium. We use models from Tan et al. [7]. In that work, Tan et al. evaluated multiple eADAGE model sizes to identify that *k*=300 features was an appropriate size for the current *P. aeruginosa* compendium. The authors also compared eADAGE to two other commonly used feature construction approaches, PCA and ICA [7]. Tan et al. produced 10 eADAGE models that each extracted *k*=300 features from the compendium of genome-scale *P. aeruginosa* data. Because PathCORE-T 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-T 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* x *k*, where *n* genes in the compendium are assigned positive or negative gene weights, according to a standard normal distribution, for each of the *k* features. 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 [7]. Tan et al.'s gene signature rules are specified as an input to the PathCORE-T analysis as well.

NMF

We also constructed an NMF model for the TCGA pan-cancer dataset. Given an NMF approximation of $A \sim WH$ [12], where *A* is the input expression dataset of size *n* x *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* x *k*) and *H* clusters on genes (size *k* x *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-T 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.

The selection of *k*=300 for the NMF model allowed us to make the eADAGE-based and NMF-based case studies roughly parallel. We verified that 300 components was appropriate by evaluating the percentage of variance explained by PCA applied to the TCGA dataset. In general, the principal components explained very little variance—the first principal component only explained 11% of the variance. At 300 components, the proportion of variance explained was 81%.

As an additional analysis, we determined the number of components (*k*=24) where each additional component explained less than 0.5% of the variance. We found that using a very small number of constructed features resulted in a substantial loss of power: PathCORE-T analysis with a single *k*=24 model yielded no significant edges after permutation test. However, PathCORE-T can be applied over multiple models as long as the feature construction method produces different solutions depending on random seed initialization. We performed 10 factorizations to generate an aggregate of 10 *k*=24-feature NMF models and found that the resulting co-occurrence network was denser (364 edges) than our *k*=300 factor network (119 edges). 65 edges were found in both networks. These shared edges had higher weights, on average, in both networks compared to edges unique to each network (https://goo.gl/vnDVNA).

Construction of a pathway co-occurrence network

We employ a Fisher's exact test [29] 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. For each pathway in the input list of pathway definitions, we specify a contingency table and calculate its p-value, which is corrected using the Benjamini—Hochberg [30] procedure to produce a feature-wise

false discovery rate (FDR). A pathway with an adjusted p-value that is less than the user-settable FDR significance cutoff, alpha (default: 0.05), is considered significantly enriched in a given gene signature. This cutoff value should be selected to most aid user interpretation of the model. The next step of PathCORE-T is to convert pathway-node relationships into pathway-pathway relationships. For this, we apply a subsequent permutation test over pathway-pathway edge weights that accounts for the frequency at which pathways are observed as associated with features. This permutation produces a p-value for each edge. Two pathways co-occur, or share an edge in the pathway co-occurrence network, if they are both overrepresented in a gene signature. The weight of each edge in the pathway-pathway graph corresponds to number of times such a pathway pair is present over all gene signatures in a model (Fig. 2a).

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). *N* is a user-settable parameter: the example PathCORE-T analyses we provide specify an *N* of 10,000. Increasing the value of *N* leads to more precise p-values, particularly for low p-values, but comes at the expense of additional computation time.

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 (by taking the sum of the background weight distribution for an edge and dividing it by *N*), we can divide each observed edge weight by its expected weight (dividing by 1 if the expected edge weight is 0 based on the *N* permutations) to get the edge's odds ratio. Edges in the final network are weighted by their odds ratios.

Gene overlap correction

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 [10]. 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 [10]. This provides a set of new pathway definitions that no longer share genes (Fig. 3c, e).

There was no existing open source implementation of this algorithm, so we implemented Donato et al.'s maximum impact estimation as a Python package (PyPI package name: crosstalk-correction). This software is separate from PathCORE-T because we expect that it may be useful in its own right for other analytical workflows, such as differential expression analysis. The procedure is written using NumPy functions and data structures, which allows for efficient implementation of array and matrix operations in Python [31].

In PathCORE-T, we used this software to resolve overlapping genes before pathway overrepresentation analysis. Overlap correction is applied to each feature of the model independently. This most closely matches the setting evaluated by the original authors of the method. Work on methods that resolves overlap by using information shared across features may provide opportunities for future enhancements but was deemed to be out of the scope of a software contribution.

With this step, the 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-T workflow by default; however, users interested in using PathCORE-T to find connections between overlapping gene sets can choose to disable the correction step.

PathCORE-T network visualization and support for experimental follow-up

The PathCORE-T 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. While we chose to represent pathway-pathway relationships as a network, 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-T 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 [32]. 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. The details we included in an edge-specific page (1) highlight 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) display 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-T network was built. k_G is the number of features that contain gene G (i.e. G is in k_G features' gene signatures), k_E the number of features that contain edge E (i.e. the two pathways connected by E are overrepresented in k_E features), and $k_{G\,\&\,E}$ the number of features that contain both gene G and edge E. The odds ratio is computed as follows:

$$observed = k_{G\,\&\,E} / k_G$$

$$expected = k_E / k$$

$$odds\ ratio = observed / expected$$

371 An odds ratio above 1 suggests that the gene is more likely to appear in features enriched for this pair of pathways.
372 In the web interface, we sort the genes by their odds ratio to highlight genes most observed with the co-
373 occurrence relationship.
374
375 The information specified in (2) requires an "expression score" for every sample. A sample expression
376 score is calculated using the genes we selected in goal (1): it is the average of the normalized gene expression
377 values weighted by the normalized gene odds ratio. Selection of the most and least expressed samples is based on
378 these scores. We use two heatmaps to show the (maximum of twenty) genes' expression values in each of the
379 fifteen most and least expressed samples (Fig. 4b).
380
381 For each sample in an edge page, a user can examine how the expression values of the edge's twenty
382 genes in that sample compare to those recorded for all other samples in the dataset that are from the same
383 experiment (Fig. 4c). Genes that display distinct expression patterns under a specific setting may be good
384 candidates for follow-up studies.
385
386 **Results**
387
388 PathCORE-T software
389 Unsupervised methods can identify previously undiscovered patterns in large collections of data.
390 PathCORE-T overlays curated knowledge after feature construction to help researchers interpret constructed
391 features in the context of existing knowledgebases. Specifically, PathCORE-T aims to clarify how expert-annotated
392 gene sets work together from a gene expression perspective. PathCORE-T starts from an unsupervised feature
393 construction model. Before applying the software, users should evaluate models to make sure that they capture
394 biological features in their dataset. Model evaluation can be performed in numerous ways depending on the
395 setting and potential assessments include consistency across biological replicates, reconstruction error given a
396 fixed dimensionality, and independent validation experiments. Tan et al. described several ways that models could
397 be evaluated [7]. Datasets will vary in terms of their amenability to analysis by different model-building strategies,
398 and researchers may wish to consult a recent review for more discussion of feature construction methods [33].
399
400 We implemented the methods contained in the PathCORE-T software in Python. The implementations of
401 the primary steps are pip-installable (PyPI package name: PathCORE-T), and examples of how to use the methods
402 in PathCORE-T are provided in the PathCORE-T-analysis repository (https://github.com/greenelab/PathCORE-T-
403 analysis). We also implemented an optional step, which corrects for overlapping genes between pathway
404 definitions, described by Donato et al. [10]. Though the algorithm had been described, no publicly available
405 implementation existed. We provide this overlap correction algorithm as a Python package (PyPI package name:
406 crosstalk-correction) available under the BSD 3-clause license. Each component of PathCORE-T can be used
407 independently of each other (Fig. 1c).
408
409 Here, we present analyses that can be produced by applying the full PathCORE-T pipeline to models
410 created from a transcriptomic compendium by an unsupervised feature construction algorithm. Input pathway
411 definitions are "overlap-corrected" for each feature before enrichment analysis. An overlap-corrected, weighted
412 pathway co-occurrence network is built by connecting the pairs of pathways that are overrepresented in features
413 of the model. Finally, we remove edges that cannot be distinguished from a null model of random associations
414 based on the results of a permutation test.
415
416 Case study: *P. aeruginosa* eADAGE models annotated with KEGG pathways
417 We used PathCORE-T to create a network of co-occurring pathways out of the expression signatures
418 extracted by eADAGE from a *P. aeruginosa* compendium [7]. For every feature, overlap correction was applied to
419 the *P. aeruginosa* KEGG pathway annotations and overlap-corrected annotations were used in the
420 overrepresentation analysis. PathCORE-T aggregates multiple networks by taking the union of the edges across all
421 networks and summing the weights of common pathway-pathway connections. We do this to emphasize the co-
422 occurrence relationships that are more stable [34]—that is, the relationships that appear across multiple models.
423 Finally, we removed edges in the aggregate network that were not significant after FDR correction when compared

424 to the background distributions generated from 10,000 permutations of the network. Used in this way, PathCORE-
425 T software allowed for exploratory analysis of an existing well-validated model.
426
427 The eADAGE co-occurrence network that resulted from our exploratory analysis contained a number of
428 pathway-pathway relationships that have been previously characterized by other means (Fig. 5). Three glucose
429 catabolism processes co-occur in the network: glycolysis, pentose phosphate, and the Entner-Doudoroff pathway
430 (Fig. 5a). We also found a cluster relating organophosphate and inorganic phosphate transport- and metabolism-
431 related processes (Fig. 5b). Notably, phosphate uptake and acquisition genes were directly connected to the *hxc*
432 genes that encode a type II secretion system. Studies in *P. aeruginosa* clinical isolates demonstrated that the Hxc
433 secretion system was responsible for the secretion of alkaline phosphatases, which are phosphate scavenging
434 enzymes [35,36] and the phosphate binding DING protein [37]. Furthermore, alkaline phosphatases, DING and the
435 *hxc* genes are regulated by the transcription factor PhoB which is most active in response to phosphate limitation.
436 The identification of this relationship by PathCORE-T as a global pattern suggested the role of type II secretion and
437 phosphate limitation seen in a limited number of isolates may be generalizable to broader *P. aeruginosa* biology.
438 As shown in Fig. 5c, we also identified linkages between two pathways involved in the catabolism of sulfur-
439 containing molecules, taurine and methionine, and the general sulfur metabolism process. Other connections
440 between pathways involved in the transport of iron (ferrienterobactin binding) [38] and zinc (the *znu* uptake
441 system [39]) were identified (Fig. 5d). Interestingly, genes identified in the edge between the zinc transport and
442 MacAB-TolC pathways included the *pvd* genes involved in pyoverdine biosynthesis and regulation, a putative
443 periplasmic metal binding protein, as well as other components of an ABC transporter (genes PA2407, PA2408, and
444 PA2409 at https://goo.gl/bfqOk8) [40]. PathCORE-T suggested a relationship between zinc and iron pathways in *P.*
445 *aeruginosa* transcriptional data though such a relationship has not yet been described. Structural analysis of the
446 iron-responsive regulator Fur found that it also productively binds zinc in *E. coli* and *Bacillus subtilis* providing a
447 mechanism by which these pathways may be linked [41, 42].
448
449 The network constructed using the PathCORE-T framework had 203 edges between 89 pathways. For
450 comparison, we constructed a KEGG pathway-pathway network where edges were drawn between pathways with
451 significant gene overlap (FDR-corrected hypergeometric test < 0.05). The overlap-based network had 406 edges
452 between 158 pathways. Only 35 of the edges in the PathCORE-T network were between pathways that shared
453 genes, with an average Jaccard Index of only 0.035. The network constructed using PathCORE-T (with overlap-
454 correction applied by default) captured pathway co-occurrences not driven by shared genes between pathways.
455
456 Case study: TCGA's pan-cancer compendium analyzed by NMF with PID pathways
457 PathCORE-T is not specific to a certain dataset, organism, or feature construction method. We
458 constructed a 300-feature NMF model of TCGA pan-cancer gene expression data, which is comprised of 33
459 different cancer-types from various organ sites and applied the PathCORE-T software to those features. We chose
460 NMF because it has been used in previous studies to identify biologically relevant patterns in transcriptomic data
461 [12] and by many studies to derive molecular subtypes [43–45]. The 300 NMF features were analyzed using
462 overlap-corrected PID pathways, a collection of 196 human cell signaling pathways with a particular focus on
463 processes relevant to cancer [11].
464
465 PathCORE-T detected modules of co-occurring pathways that were consistent with our current
466 understanding of cancer-related interactions (Fig. 6). Because cancer-relevant pathways were used, it was not
467 surprising that cancer-relevant pathways appeared. However, the edges between those pathways were also
468 encouraging. For example, a module composed of a FoxM1 transcription factor network, an E2F transcription
469 factor network, Aurora B kinase signaling, ATR signaling, PLK1 signaling, and members of the Fanconi anemia DNA
470 damage response pathway was densely connected (Fig 6a). The connections in this module recapitulated well
471 known cancer hallmarks including cellular proliferation pathways and markers of genome instability, such as the
472 activation of DNA damage response pathways [46]. We found that several pairwise pathway co-occurrences
473 corresponded with previously reported pathway-pathway interactions [47–49]. We also observed a hub of
474 pathways interacting with Wnt signaling, among them the regulation of nuclear Beta-catenin signaling, FGF
475 signaling, and BMP signaling (Fig. 6b). The Wnt and BMP pathways are functionally integrated in biological
476 processes that contribute to cancer progression [50]. Additionally, Wnt/Beta-catenin signaling is a well-studied

477 regulatory system, and the effects of mutations in Wnt pathway components on this system have been linked to
478 tumorigenesis [51]. Wnt/Beta-catenin and FGF together influence the directional migration of cancer cell clusters
479 [52].
480
481 Two modules in the network related to angiogenesis, or the formation of new blood vessels (Fig. 6c, d).
482 Tumors transmit signals that stimulate angiogenesis because a blood supply provides the necessary oxygen and
483 nutrients for their growth and proliferation. One module relates angiogenesis factors to cell proliferation. This
484 module connected the following pathways: PDGFR-beta signaling, FAK-mediated signaling events, VEGFR1 and
485 VEGFR2-mediated signaling events, nuclear SMAD2/3 signaling regulation, and RB1 regulation (Fig. 6c). These
486 functional connections are known to be involved in tumor proliferation [53–55]. The other module indicated a
487 relationship by which the VEGF pathway interacts with the S1P3 pathway through Beta3 integrins (Fig. 6d). *S1P3* is
488 a known regulator of angiogenesis [56], and has been demonstrated to be associated with treatment-resistant
489 breast cancer and poor survival [57]. Moreover, this interaction module has been observed to promote endothelial
490 cell migration in human umbilical veins [58]. Taken together, this independent module may suggest a distinct
491 angiogenesis process activated in more aggressive and metastatic tumors that is disrupted and regulated by
492 alternative mechanisms [59].
493
494 Finally, PathCORE-T revealed a large, densely connected module of immune related pathways (Fig. 6e).
495 We found that this module contains many co-occurrence relationships that align with immune system processes.
496 One such example is the well characterized interaction cycle formed by T-Cell Receptor (TCR) signaling in naïve
497 CD4+ T cells and IL-12/IL-4 mediated signaling events [60–62]. At the same time, PathCORE-T identifies additional
498 immune-related relationships. We observed a cycle between the three transcription factor networks: ATF-2, AP-1,
499 and CaN-regulated NFAT-dependent transcription. These pathways can take on different, often opposing,
500 functions depending on the tissue and subcellular context. For example, ATF-2 can be an oncogene in one context
501 (e.g. melanoma) and a tumor suppressor in another (e.g. breast cancer) [63]. AP-1, comprised of Jun/Fos proteins,
502 is associated with both tumorigenesis and tumor suppression due to its roles in cell survival, proliferation, and cell
503 death [64]. Moreover, NFAT in complex with AP-1 regulates immune cell differentiation, but dysregulation of NFAT
504 signaling can lead to malignant growth and tumor metastasis [65]. The functional association observed between
505 the ATF-2, AP-1, and NFAT cycle together within the immunity module might suggest that dysregulation within this
506 cycle has profound consequences for immune cell processes and may trigger variable oncogenic processes.
507
508 Just as we did for the eADAGE-based *P. aeruginosa* KEGG pathways case study, we constructed a network
509 only from PID pathways with significant gene overlap. The network constructed using PathCORE-T and NMF
510 features had 119 edges between 57 pathways. The overlap-based network was much denser: it had 3826 edges
511 between 196 pathways. This suggested a high degree of overlap between PID pathways. For the PathCORE-T NMF
512 co-occurrence network, 96 of the edges were between pathways that shared genes. However, the average Jaccard
513 Index for these pathway pairs remained low, at 0.058.
514
515 **Conclusions**
516
517 Unsupervised analyses of genome-scale datasets that summarize key patterns in the data have the
518 potential to improve our understanding of how a biological system operates via complex interactions between
519 molecular processes. Feature construction algorithms capture coordinated changes in the expression of many
520 genes as features. The genes that contribute most to each feature co-vary. However, interpreting the features
521 generated by unsupervised approaches remains challenging. PathCORE-T creates a network of globally co-
522 occurring pathways based on features created from the analysis of a compendium of gene expression data.
523 Networks modeling the relationships between curated processes in a biological system offer a means for
524 developing new hypotheses about which pathways influence each other and when. Our framework provides a
525 data-driven characterization of the biological system at the pathway-level by identifying pairs of pathways that are
526 overrepresented across many features.
527
528 PathCORE-T connects the features extracted from data to curated resources. It is important to note that
529 PathCORE-T will only be able to identify pathways that occur in features of the underlying model, which means

530 these pathways must be transcriptionally perturbed in at least some subset of the compendium. Models should be
531 evaluated before analysis with PathCORE-T. The network resulting from PathCORE-T can help to identify global
532 pathway-pathway relationships—a baseline network—that complements existing work to identify interactions
533 between pathways in the context of a specific disease. The specific niche that PathCORE-T framework aims to fill is
534 in revealing to researchers which gene sets are most closely related to each other in machine learning-based
535 models of gene expression, which genes play a role in this co-occurrence, and which conditions drive this
536 relationship.
537
538 **Project name:** PathCORE-T
539 **Project home page:** https://pathcore-demo.herokuapp.com
540 **Archived version:** https://github.com/greenelab/PathCORE-T-analysis/releases/tag/v1.2.0 (links to download .zip
541 and .tar.gz files are provided here)
542 **Operating system:** Platform independent
543 **Programming language:** Python
544 **Other requirements:** Python 3 or higher
545 **License:** BSD 3-clause
546
547 **Declarations**
548 **Ethics approval and consent to participate**
549 Not applicable
550
551 **Consent for publication**
552 Not applicable
553
554 **Availability of data and materials**
555 Files for each of the PathCORE-T networks described in the results are provided in Supplementary material.
556
557 *Data sets*
558 *P. aeruginosa* eADAGE models: doi:10.5281/zenodo.583172
559
560 *P. aeruginosa* gene expression compendium: doi:10.5281/zenodo.583694
561 The normalized gene expression compendium provided in this Zenodo record contains datasets on the
562 GPL84 from ArrayExpress as of 31 July 2015. It includes 1,051 samples grown in 78 distinct medium conditions, 128
563 distinct strains and isolates, and dozens of different environmental parameters [7]. Tan et al. compiled this dataset
564 and used it to construct the 10 eADAGE models from which we generate the eADAGE-based *P. aeruginosa* KEGG
565 network [7]. We use this same data compendium to generate the NMF *P. aeruginosa* model. The script used to
566 download and process those datasets into the compendium is available at https://goo.gl/YjOEQl.
567
568 TCGA pan-cancer dataset: doi:10.5281/zenodo.56735
569 The pan-cancer dataset was compiled using data from all 33 TCGA cohorts. It was generated by the TCGA
570 Research Network: http://cancergenome.nih.gov/ [14]. RNA-seq data was downloaded on 8 March 2016 from
571 UCSC Xena (https://xenabrowser.net/datapages/). Gene expression was measured using the Illumina HiSeq
572 technology. More information and the latest version of the dataset can be found at
573 https://xenabrowser.net/datapages/?dataset=TCGA.PANCAN.sampleMap/HiSeqV2&host=https://tcga.xenahubs.n
574 et.
575
576 *Source code*
577 PathCORE-T analysis: (https://github.com/greenelab/PathCORE-T-analysis/tree/v1.2.0) This repository contains all
578 the scripts to reproduce the analyses described in this paper. The Python scripts here should be used as a starting
579 point for new PathCORE-T analyses. Instructions for setting up a web application for a user's specific PathCORE-T
580 analysis are provided in this repository's README.
581

582 Overlap correction: (https://github.com/kathyxchen/crosstalk-correction/tree/v1.0.4) Donato et. al's procedure
583 for overlap correction [10] is a pip-installable Python package 'crosstalk-correction' that is separate from, but listed
584 as a dependency in, PathCORE-T. It is implemented using NumPy [31].
585
586 PathCORE-T methods: (https://github.com/greenelab/PathCORE-T/tree/v1.0.2) The methods included in the
587 PathCORE-T analysis workflow (Fig. 4c) are provided as a pip-installable Python package 'pathcore'. It is
588 implemented using Pandas [66], SciPy (specifically, scipy.stats) [67], StatsModels [68], and the crosstalk-correction
589 package.
590
591 PathCORE-T demo application: (https://github.com/kathyxchen/PathCORE-T-demo/tree/v2.1) The project home
592 page, https://pathcore-demo.herokuapp.com provides links to
593 (1) The web application for the eADAGE-based KEGG *P. aeruginosa* described in the first case study.
594 (2) A view of the NMF-based PID pathway co-occurrence network described in the second case study.
595 (3) A quick view page where users can temporarily load and visualize their own network file (generated from
596 the PathCORE-T analysis).
597
598 **Competing interests**
599 The authors declare that they have no competing interests
600
601 **Funding**
602 KMC was funded by an undergraduate research grant from the Penn Institute for Biomedical Informatics. CSG was
603 funded in part by a grant from the Gordon and Betty Moore Foundation (GBMF 4552). JT, GD, DAH, and CSG were
604 funded in part by a pilot grant from the Cystic Fibrosis Foundation (STANTO15R0). DAH was funded in part by R01-
605 AI091702.
606
607 **Authors' contributions**
608 KMC implemented the software, performed the analyses, and drafted the manuscript. JT and GPW contributed
609 computational reagents. KMC, DAH, and CSG designed the project. KMC, JT, GPW, GD, DAH, and CSG interpreted
610 the results. JT, GPW, GD, DAH, and CSG provided critical feedback and revisions on the manuscript. JT and GPW
611 reviewed source code.
612
613 **Acknowledgements**
614 The authors would also like to thank Daniel Himmelstein, Dongbo Hu, Kurt Wheeler, and René Zelaya for helping to
615 review the source code.
616
617 **References**
618 1. Greene CS, Foster JA, Stanton BA, Hogan DA, Bromberg Y. Computational Approaches to Study Microbes and
619 Microbiomes. Pac. Symp. Biocomput. 2016;21:557.

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

622 3. Barrett, T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional
623 genomics data sets—update. Nucleic Acids Res. 2012;41:D991-5.

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

626 5. Tan J, Hammond JH, Hogan DA, Greene CS. Adage-based integration of publicly available Pseudomonas
627 aeruginosa gene expression data with denoising autoencoders illuminates microbe-host interactions. mSystems.
628 2016;1:25.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

658 21. Anda-Juregui G de, Meja-Pedroza RA, Espinal-Enrquez J, Hernndez-Lemus E. Crosstalk events in the estrogen
659 signaling pathway may affect tamoxifen efficacy in breast cancer molecular subtypes. Comput. Biol. Chem.
660 2015;59:42–54.

661 22. Pita-Juárez Y, Altschuler G, Kariotis S, Wei W, Koler K, Green C, et al. The Pathway Coexpression Network:
662 Revealing pathway relationships. PLoS Comput Biol. 2018;14:e1006042.

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

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

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

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

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

672 28. Stone JV. Independent component analysis. Wiley Online Library. 2004.

673 29. Upton GJG. Fisher's Exact Test. J. Royal Stat. Soc. 1992;155:395-402.
674
675 30. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple
676 Testing. J. Royal Stat. Soc. 1995;57:289-300.
677
678 31. Walt S van der, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation.
679 Comput. Sci. Eng. 2011;13:22–30.

680 32. Bostock M. D3.js. Data Driven Doc. 2012;492.

681 33. Stein-O'Brien GL, Arora R, Culhane AC, Favorov A, Greene C, Goff LA, et al. Enter the matrix: Interpreting
682 unsupervised feature learning with matrix decomposition to discover hidden knowledge in high-throughput omics
683 data. bioRxiv. 2017;196915.
684
685 34. Yu B. Stability. Bernoulli. 2013;19:1484–500.

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

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

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

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

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

696 40. Winsor GL, Lam DK, Fleming L, Lo R, Whiteside MD, Nancy YY, et al. *Pseudomonas* Genome Database: improved
697 comparative analysis and population genomics capability for *Pseudomonas* genomes. Nucleic Acids Res.
698 2010;gkq869.

699 41. Althaus EW, Outten CE, Olson KE, Cao H, O'Halloran TV. The ferric uptake regulation (Fur) repressor is a zinc
700 metalloprotein. Biochemistry. 1999;38:6559-69.

701 42. Ma Z, Faulkner MJ, Helmann JD. Origins of specificity and cross-talk in metal ion sensing by *Bacillus subtilis* Fur.
702 Mol. Microbiol. 2012;86:1144-55.

703 43. Bailey P, Chang DK, Nones K, Johns AL, Patch A-M, Gingras M-C, et al. Genomic analyses identify molecular
704 subtypes of pancreatic cancer. Nature. 2016;531:47-52.

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

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

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

709 47. Moldovan G-L, D'Andrea AD. How the Fanconi Anemia pathway guards the genome. Annu. Rev. Genet.
710 2009;43:223–49.

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

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

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

717 51. MacDonald BT, Tamai K, He X. Wnt/β-catenin signaling: components, mechanisms, and diseases. Dev. Cell.
718 2009;17:9–26.

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

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

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

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

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

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

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

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

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

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

740 62. Vacaflores A, Chapman NM, Harty JT, Richer MJ, Houtman JC. Exposure of Human CD4 T Cells to IL-12 Results in
741 Enhanced TCR-Induced Cytokine Production, Altered TCR Signaling, and Increased Oxidative Metabolism. PloS One.
742 2016;11:e0157175.

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

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

745 65. Mller MR, Rao A. NFAT, immunity and cancer: a transcription factor comes of age. Nat. Rev. Immunol.
746 2010;10:645–56.

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

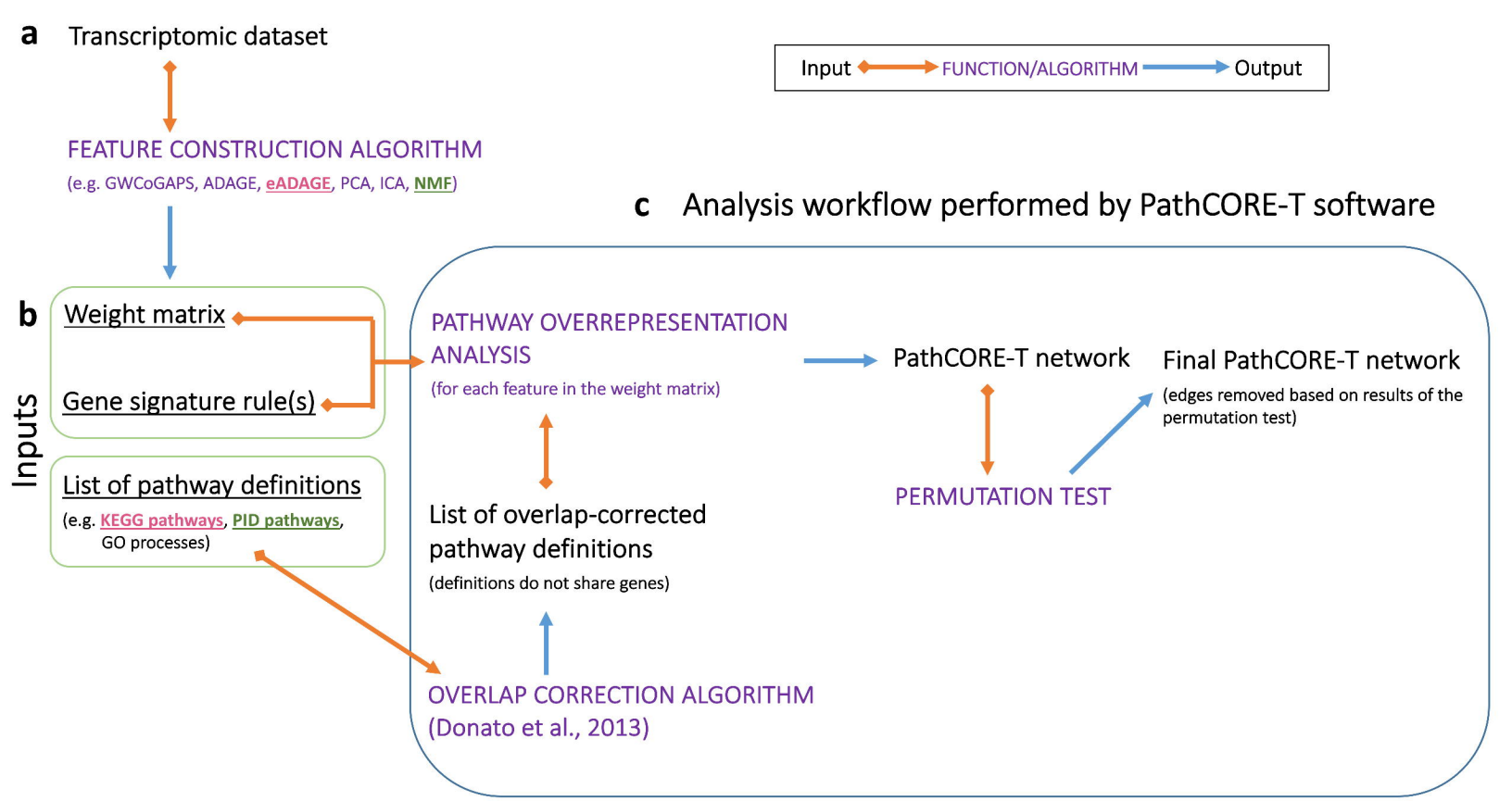
749 67. Jones E, Oliphant T, Peterson P. SciPy: open source scientific tools for Python. 2014;

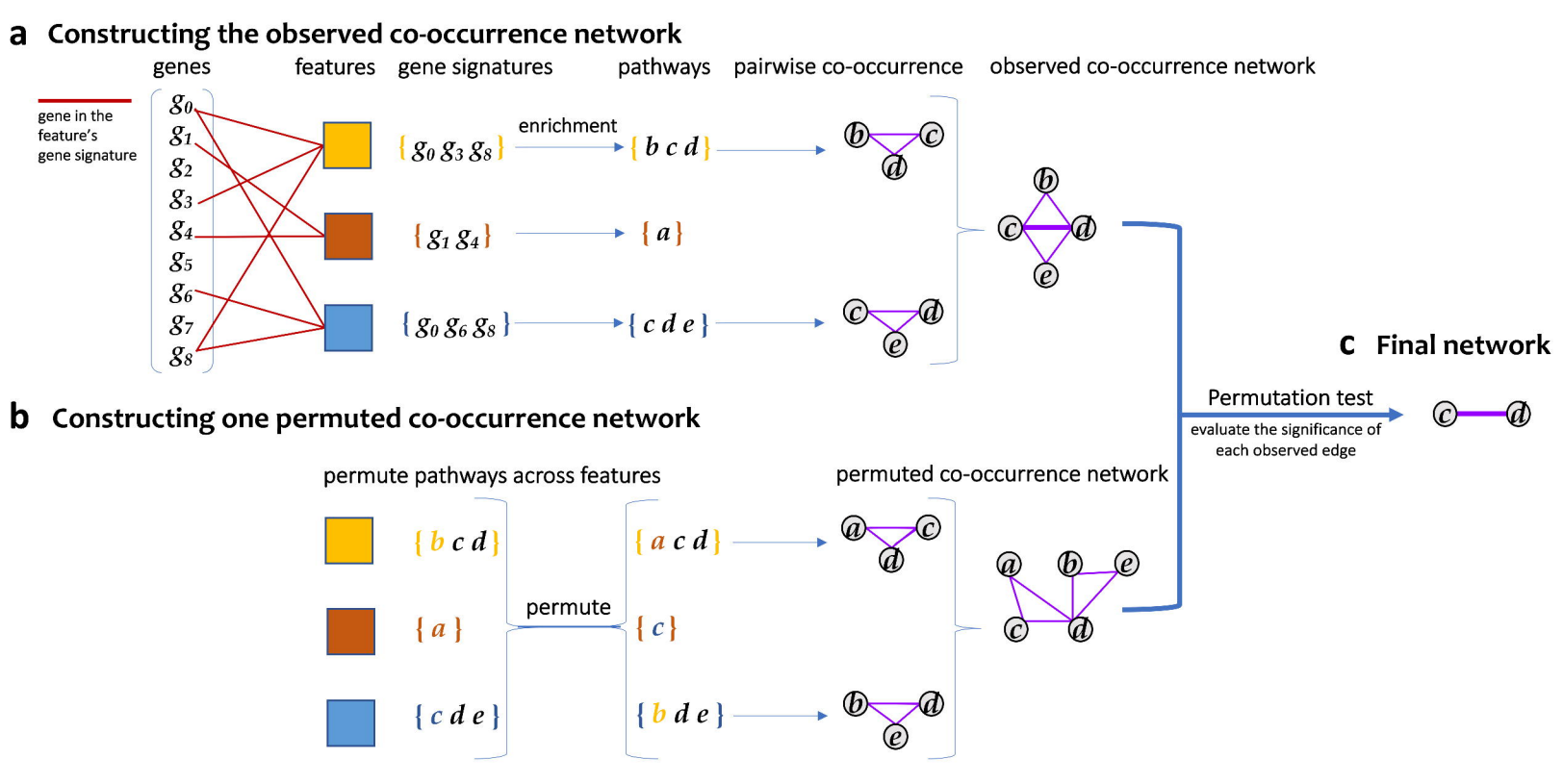
750 68. Seabold S, Perktold J. Statsmodels: Econometric and statistical modeling with python. Proc. 9th Python Sci.
751 Conf. 2010. p. 61.

752
753 **Figure 1** The PathCORE-T software analysis workflow.
754 **(a)** A user applies a feature construction algorithm to a transcriptomic dataset of genes-by-samples. The features
755 constructed must preserve the genes in the dataset and assign weights to each of these genes according to some
756 distribution. **(b)** Inputs required to run the complete PathCORE-T analysis workflow. The constructed features are
757 stored in a weight matrix. Based on how gene weights are distributed in the constructed features, a user defines
758 thresholds to select the set of genes most indicative of each feature's function—we refer to these user-defined
759 thresholds as gene signature rules. Finally, a list of pathway definitions will be used to interpret the features and
760 build a pathway co-occurrence network. **(c)** Methods in the PathCORE-T analysis workflow (indicated using purple
761 font) can be employed independently of each other so long as the necessary input(s) are provided. The 2 examples
762 we describe to demonstrate PathCORE-T software use the following inputs: (1) the weight matrix and gene
763 signature rules for eADAGE (applied to the *P. aeruginosa* gene compendium) and KEGG pathways and (2) the
764 weight matrix and gene signature rules for NMF (applied to the TCGA pan-cancer dataset) and PID pathways.
765
766 **Figure 2** The approach implemented in PathCORE-T to construct a pathway co-occurrence network from an
767 expression compendium.
768 **(a)** A user-selected feature extraction method is applied to expression data. Such methods assign each gene a
769 weight, according to some distribution, that represents the gene's contribution to the feature. The set of genes
770 that are considered highly representative of a feature's function is referred to as a feature's gene signature. The
771 gene signature is user-defined and should be based on the weight distribution produced by the unsupervised
772 method of choice. In the event that the weight distribution contains both positive and negative values, a user can
773 specify criteria for both a positive and negative gene signature. A test of pathway enrichment is applied to identify
774 corresponding sets of pathways from the gene signature(s) in a feature. We consider pathways significantly
775 overrepresented in the same feature to co-occur. Pairwise co-occurrence relationships are used to build a
776 network, where each edge is weighted by the number of features containing both pathways.
777 **(b)** *N* permuted networks are generated to assess the statistical significance of a co-occurrence relation in the
778 graph. Here, we show the construction of one such permuted network. Two invariants are maintained during a
779 permutation: (1) pathway side-specificity (if applicable, e.g. positive and negative gene signatures) and (2) the
780 number of distinct pathways in a feature's gene signature.
781 **(c)** For each edge observed in the co-occurrence network, we compare its weight against the weight distribution
782 generated from *N* (default: 10,000) permutations of the network to determine each edge's p-value. After

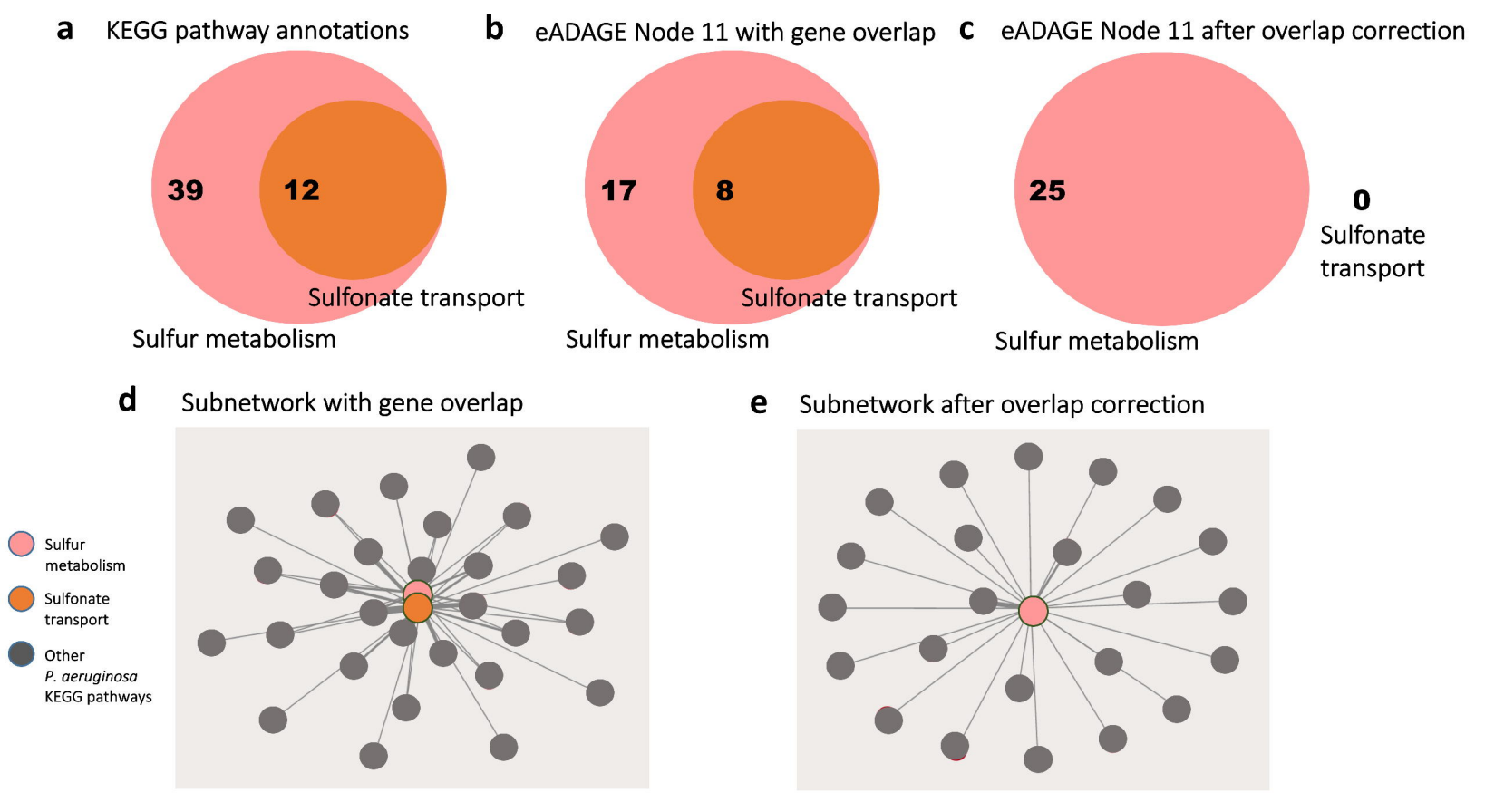
783 correcting the p-value by the number of edges observed in the graph using the Benjamini—Hochberg procedure,
784 only an edge with an adjusted p-value below alpha (default: 0.05) is kept in the final co-occurrence network.
785
786 **Figure 3** Correcting for gene overlap results in a sparser pathway co-occurrence network.
787 **(a)** The KEGG pathway annotations for the sulfonate transport system are a subset of those for sulfur metabolism.
788 12 genes annotated to the sulfonate transport system are also annotated to sulfur metabolism. **(b)** Without
789 applying the overlap correction procedure, 25 of the genes in the positive and negative gene signatures of the
790 eADAGE feature "Node 11" are annotated to sulfur metabolism--of those, 8 genes are annotated to the sulfonate
791 transport system as well. **(c)** All 8 of the overlapping genes are mapped to the sulfur metabolism pathway after
792 overlap correction.
793 **(d)** A co-occurrence network built without applying the overlap correction procedure will report co-occurrence
794 between the sulfonate transport system and sulfur metabolism, whereas **(e)** no such relation is identified after
795 overlap correction.
796
797 **Figure 4** A web application used to analyze pathway-pathway relationships in the eADAGE-based, *P. aeruginosa*
798 KEGG network.
799 **(a)** A user clicks on an edge (a pathway-pathway relationship) in the network visualization. **(b)** The user is directed
800 to a page that displays expression data from the original transcriptomic dataset specific to the selected edge
801 (https://goo.gl/Hs5A3e). The expression data is visualized as two heatmaps that indicate the fifteen most and
802 fifteen least expressed samples corresponding to the edge. To select the "most" and "least" expressed samples,
803 we assign each sample a summary "expression score." The expression score is based on the expression values of
804 the genes (limited to the top twenty genes with an odds ratio above 1) annotated to one or both of the pathways.
805 Here, we show the heatmap of least expressed samples specific to the [Phosphate transport - Type II general
806 secretion] relationship. **(c)** Clicking on a square in the heatmap directs a user to an experiment page
807 (https://goo.gl/KYNhwB) based on the sample corresponding to that square. A user can use the experiment page
808 to identify whether the expression values of genes specific to an edge and a selected sample differ from those
809 recorded in other samples of the experiment. In this experiment page, the first three samples (labeled in black) are
810 *P. aeruginosa* "baseline" replicates grown for 72 h in drop-flow biofilm reactors. The following three samples
811 (labeled in blue) are *P. aeruginosa* grown for an additional 12 h (84 h total). Labels in blue indicate that the three
812 84 h replicates are in the heatmap of least expressed samples displayed on the [Phosphate transport – Type II
813 general secretion] edge page.
814
815 **Figure 5** eADAGE features constructed from publicly available *P. aeruginosa* expression data describe known KEGG
816 pathway-pathway relationships.
817 (a) The glycerolipid metabolism, Entner-Doudoroff, glycolysis/gluconeogenesis, and pentose phosphate, pathways
818 share common functions related to glucose catabolism.
819 (b) Organophosphate and inorganic phosphate transport- and metabolism-related processes frequently co-occur
820 with bacterial secretion systems. Here, we observe pairwise relationships between type II general secretion and
821 phosphate-related processes.
822 (c) Pathways involved in the catabolism of sulfur-containing molecules--taurine (NitT/TauT family transport) and
823 methionine (D-Methionine transport), and the general sulfur metabolism process--are functionally linked.
824 (d) The zinc transport, iron complex transport, and MacAB-TolC transporter systems are pairwise connected.
825 The fully labeled network can be viewed at https://pathcore-demo.herokuapp.com/PAO1. The list of KEGG
826 pathway-pathway relationships visualized in the network is available at the specified link (Ctrl + L for list view) and
827 as a supplemental file for this paper.
828
829 **Figure 6** PID pathway-pathway relationships discovered in NMF features constructed from the TCGA pan-cancer
830 gene expression dataset.
831 (a) Pathways in this module are responsible for cell cycle progression.
832 (b) Wnt signaling interactions with nuclear Beta-catenin signaling, FGF signaling, and BMP signaling have all been
833 linked to cancer progression.
834 (c) Here, we observe functional links between pathways responsible for angiogenesis and those responsible for cell
835 proliferation.

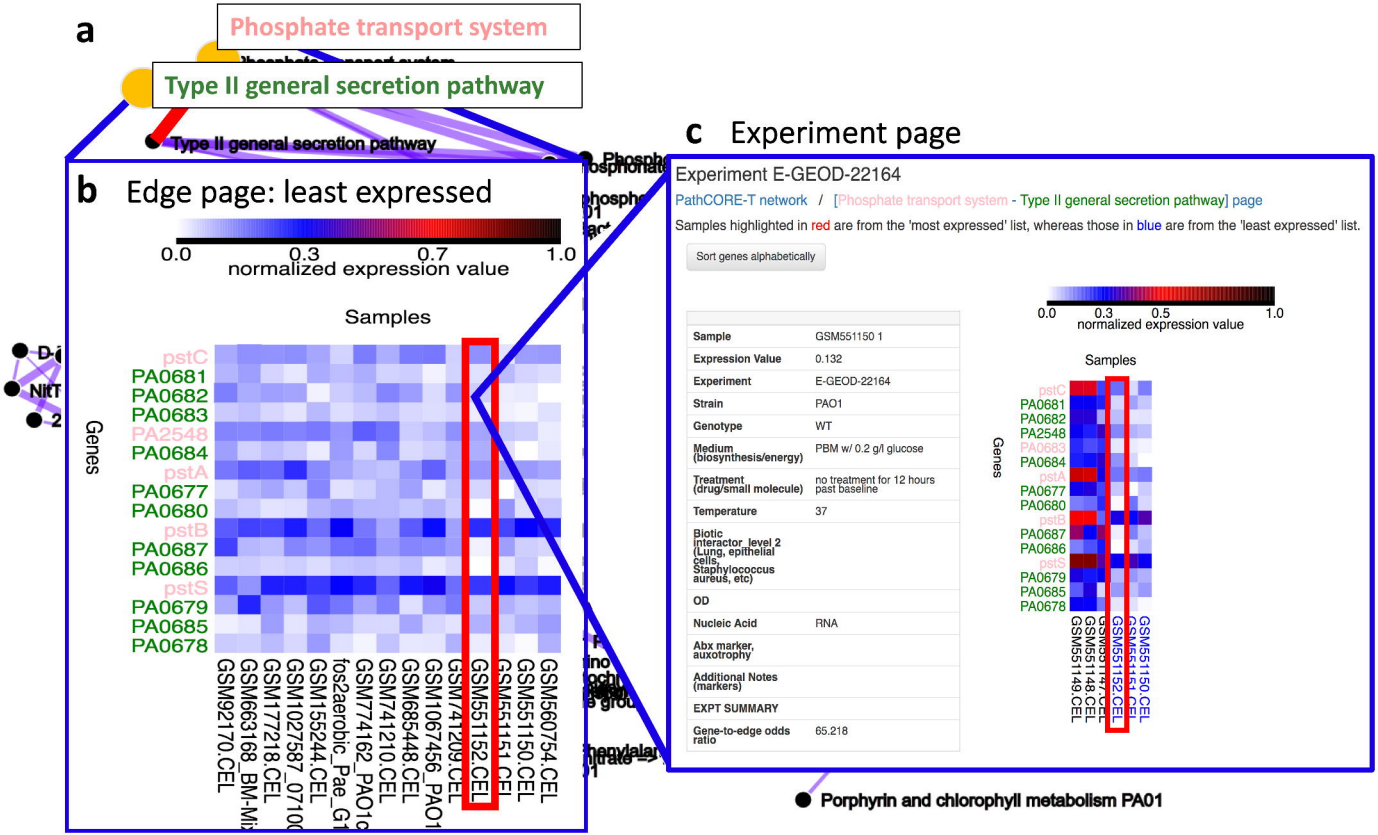
836 (d) The VEGF-VEGFR pathway interacts with the S1P3 pathway through Beta3 integrins.
837 (e) This module contains many relationships related to immune system processes. The interaction cycle formed by
838 T-Cell Receptor (TCR) signaling in naïve CD4+ T cells and IL-12/IL-4 mediated signaling events, outlined in yellow, is
839 one well-known example. The cycle in blue is formed by the ATF2, NFAT, and AP1 pathways; pairwise co-
840 occurrence of these three transcription factor networks may suggest that dysregulation of any one of these
841 pathways can trigger variable oncogenic processes in the immune system.
842 The list of PID pathway-pathway relationships visualized in the network is available as a supplemental file for this
843 paper.
844

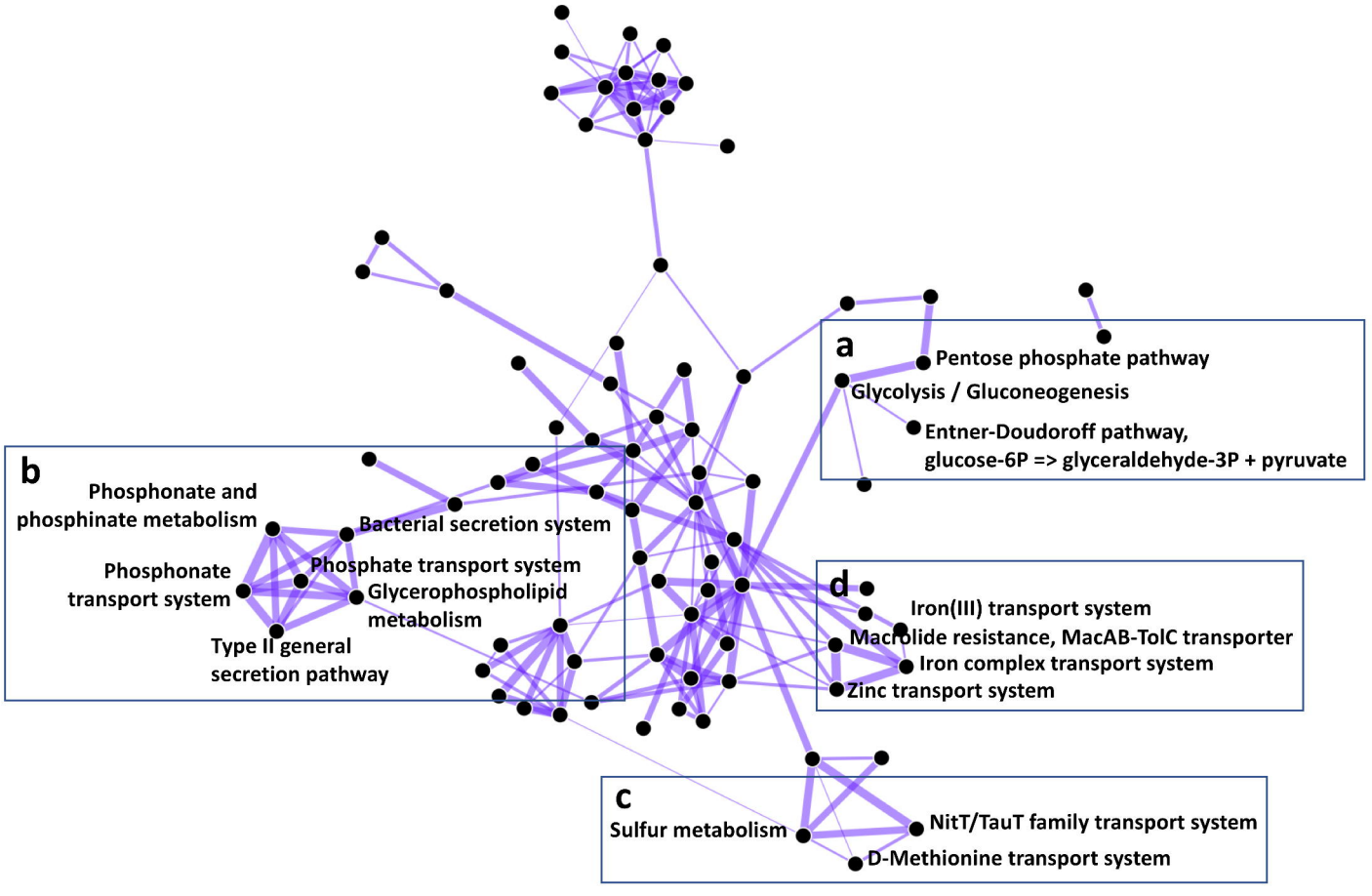

a
Transcriptomic dataset
FEATURE CONSTRUCTION ALGORITHM
(e.g. GWCoGAPS, ADAGE, eADAGE, PCA, ICA, NMF)
Input
FUNCTION/ALGORITHM
Output
b
Inputs
Weight matrix
Gene signature rule(s)
List of pathway definitions
(e.g. KEGG pathways, PID pathways, GO processes)
c
Analysis workflow performed by PathCORE-T software
PATHWAY OVERREPRESENTATION ANALYSIS
(for each feature in the weight matrix)
PathCORE-T network
Final PathCORE-T network
(edges removed based on results of the permutation test)
PERMUTATION TEST
List of overlap-corrected pathway definitions
(definitions do not share genes)
OVERLAP CORRECTION ALGORITHM
(Donato et al., 2013)

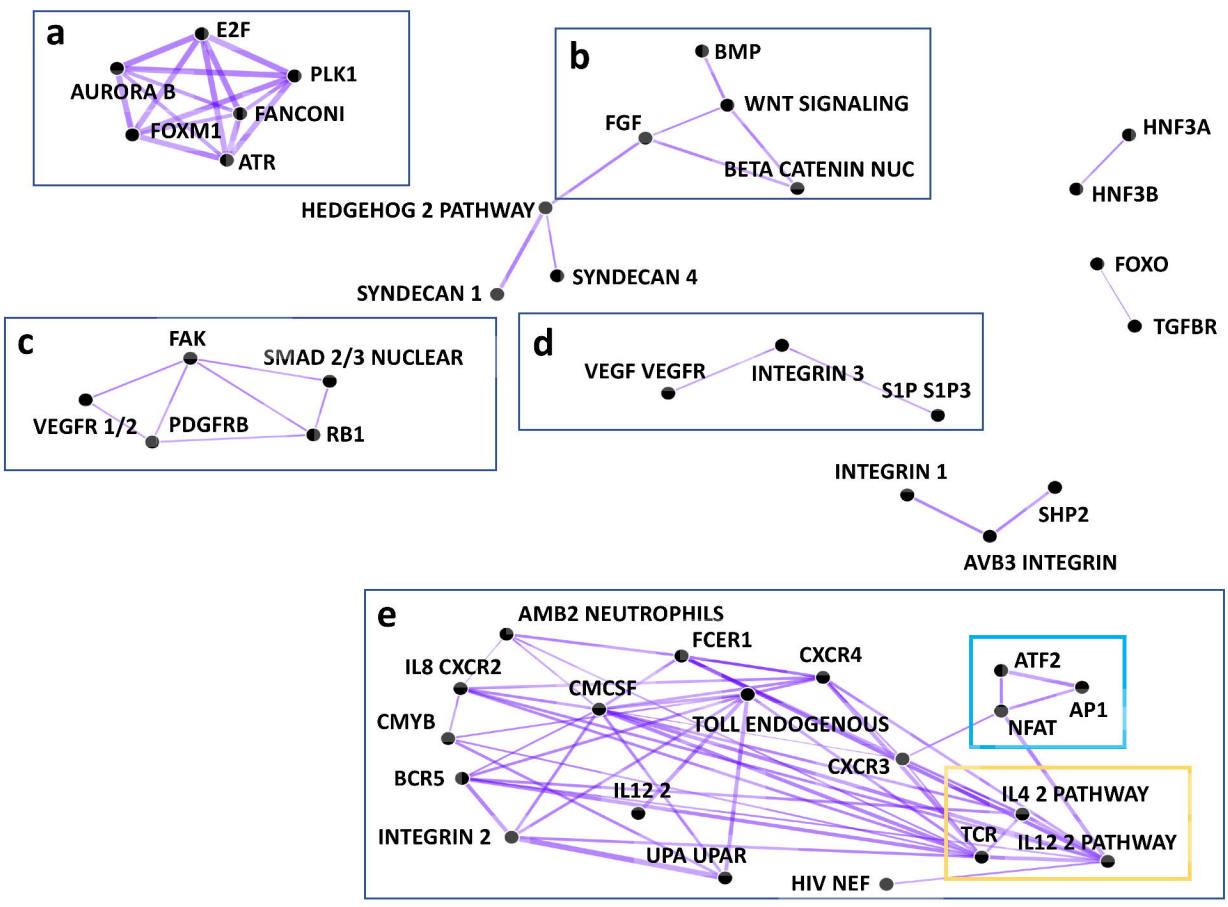


a Constructing the observed co-occurrence network
genes
features
gene signatures
pathways
pairwise co-occurrence
observed co-occurrence network
gene in the feature's gene signature
g_0 g_1 g_2 g_3 g_4 g_5 g_6 g_7 g_8
$\{g_0\ g_3\ g_8\}$
enrichment
$\{b\ c\ d\}$
$\{g_1\ g_4\}$
$\{a\}$
$\{g_0\ g_6\ g_8\}$
$\{c\ d\ e\}$
b Constructing one permuted co-occurrence network
permute pathways across features
permuted co-occurrence network
$\{b\ c\ d\}$
$\{a\}$
$\{c\ d\ e\}$
permute
$\{a\ c\ d\}$
$\{c\}$
$\{b\ d\ e\}$
c Final network
Permutation test
evaluate the significance of each observed edge


a
KEGG pathway annotations
39
12
Sulfonate transport
Sulfur metabolism
b
eADAGE Node 11 with gene overlap
17
8
Sulfonate transport
Sulfur metabolism
c
eADAGE Node 11 after overlap correction
25
0
Sulfonate transport
Sulfur metabolism
d
Subnetwork with gene overlap
e
Subnetwork after overlap correction
Sulfur metabolism
Sulfonate transport
Other P. aeruginosa KEGG pathways


a
Phosphate transport system
Type II general secretion pathway
Type II general secretion pathway
b
Edge page: least expressed
0.0
0.3
0.7
1.0
normalized expression value
Samples
Genes
pstC
PA0681
PA0682
PA0683
PA2548
PA0684
pstA
PA0677
PA0680
pstB
PA0687
PA0686
pstS
PA0679
PA0685
PA0678
GSM92170.CEL
GSM663168_BM-Mi
GSM177218.CEL
GSM1027587_0710
GSM155244.CEL
fos2aerobic_Pae_G1
GSM774162_PAO1c
GSM741210.CEL
GSM685448.CEL
GSM1067456_PAO1
GSM741209.CEL
GSM551152.CEL
GSM551151.CEL
GSM551150.CEL
GSM560754.CEL
c
Experiment page
Experiment E-GEOD-22164
PathCORE-T network / [Phosphate transport system - Type II general secretion pathway] page
Samples highlighted in red are from the 'most expressed' list, whereas those in blue are from the 'least expressed' list.
Sort genes alphabetically
Sample GSM551150 1
Expression Value 0.132
Experiment E-GEOD-22164
Strain PAO1
Genotype WT
Medium (biosynthesis/energy) PBM w/ 0.2 g/l glucose
Treatment (drug/small molecule) no treatment for 12 hours past baseline
Temperature 37
Biotic interactor_level 2 (Lung, epithelial cells, Staphylococcus aureus, etc)
OD
Nucleic Acid RNA
Abx marker, auxotrophy
Additional Notes (markers)
EXPT SUMMARY
Gene-to-edge odds ratio 65.218
0.0
0.3
0.5
1.0
normalized expression value
Samples
Genes
pstC
PA0681
PA0682
PA2548
PA0683
PA0684
pstA
PA0677
PA0680
pstB
PA0687
PA0686
pstS
PA0679
PA0685
PA0678
GSM551149.CEL
GSM551148.CEL
GSM551152.CEL
GSM551150.CEL
Porphyrin and chlorophyll metabolism PA01


a
Pentose phosphate pathway
Glycolysis / Gluconeogenesis
Entner-Doudoroff pathway,
glucose-6P => glyceraldehyde-3P + pyruvate
b
Phosphonate and
phosphinate metabolism
Bacterial secretion system
Phosphonate
transport system
Phosphate transport system
Glycerophospholipid
metabolism
Type II general
secretion pathway
d
Iron(III) transport system
Macrolide resistance, MacAB-TolC transporter
Iron complex transport system
Zinc transport system
c
Sulfur metabolism
NitT/TauT family transport system
D-Methionine transport system


a
E2F
AURORA B
PLK1
FANCONI
FOXM1
ATR
b
BMP
WNT SIGNALING
FGF
BETA CATENIN NUC
HEDGEHOG 2 PATHWAY
SYNDECAN 1
SYNDECAN 4
HNF3A
HNF3B
FOXO
TGFBR
c
FAK
SMAD 2/3 NUCLEAR
VEGFR 1/2
PDGFRB
RB1
d
VEGF VEGFR
INTEGRIN 3
S1P S1P3
INTEGRIN 1
SHP2
AVB3 INTEGRIN
e
AMB2 NEUTROPHILS
FCER1
CXCR4
IL8 CXCR2
CMCSF
CMYB
TOLL ENDOGENOUS
BCR5
IL12 2
INTEGRIN 2
UPA UPAR
HIV NEF
CXCR3
ATF2
AP1
NFAT
IL4 2 PATHWAY
TCR
IL12 2 PATHWAY