

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

2
3 Kathleen M. Chen¹, Jie Tan², Gregory P. Way¹, Georgia Doing³, Deborah A. Hogan³, Casey S. Greene^{1,*}

4
5 ¹ Department of Systems Pharmacology and Translational Therapeutics. Perelman School of Medicine.
6 University of Pennsylvania. Philadelphia PA. 19104

7 ² Department of Molecular and Systems Biology. Geisel School of Medicine at Dartmouth. Hanover NH.
8 03755

9 ³ Department of Microbiology and Immunology. Geisel School of Medicine at Dartmouth. Hanover NH.
10 03755

11
12 * To whom correspondence should be addressed.

13 Casey Greene

14 University of Pennsylvania

15 3400 Civic Center Blvd.

16 Philadelphia PA 10104

17 Phone: 215-573-2991

18 Email: csgreene@upenn.edu

19 Abstract

20
21 **Background:** Investigators often interpret genome-wide data by analyzing the expression levels of genes
22 within pathways. While this within-pathway analysis is routine, the products of any one pathway can affect
23 the activity of other pathways. Past efforts to identify relationships between biological processes have
24 evaluated overlap in knowledge bases or evaluated changes that occur after specific treatments.
25 Individual experiments can highlight condition-specific pathway-pathway interactions; however,
26 constructing a complete network of such relationships across many conditions requires analyzing results
27 from many studies.

28 **Results:** We developed the PathCORE software to predict global pathway-pathway interactions, i.e.
29 those evident across a broad data compendium. PathCORE starts with the results of robust feature
30 construction algorithms, which are now being developed and applied to transcriptomic data. PathCORE
31 identifies pathways grouped together in features more than expected by chance as *functionally co-*
32 *occurring*. We performed example analyses using PathCORE for a microbial compendium for which
33 eADAGE features were already available and a TCGA dataset of 33 cancer types that we analyzed via
34 NMF. PathCORE recapitulated previously described pathway-pathway interactions and suggested
35 additional edges with biological plausibility that still remain to be explored. The software also identifies
36 genes associated with each relationship and includes a user-installable web interface where users can (1)
37 visualize the resulting network and (2) review the expression levels of associated genes in the original
38 data, which helps biologists using the PathCORE software design experiments to test the relationships
39 that were identified.

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

44 **Keywords:** gene expression; unsupervised feature construction; crosstalk; unsupervised, pathway
45 interactions

46 Background

47
48 The number of publicly available genome-wide datasets is growing rapidly [1]. High-throughput
49 sequencing technologies that measure gene expression quickly with high accuracy and low cost continue
50 to enable this growth [2]. Expanding public data repositories have laid the foundation for computational
51 methods that consider entire compendia of gene expression data to extract biological patterns [3]. These
52 patterns may be difficult to detect in measurements from a single experiment. Unsupervised approaches,
53 which identify important signals in the data without relying on prior knowledge, may discover new
54 expression modules [4, 5].
55

56 Feature extraction methods are a class of unsupervised algorithms that can reveal unannotated
57 biological processes from genomic data [5]. Features can be constructed as representative “meta-genes”:
58 each feature has a set of influential genes, and these genes suggest the biological or technical pattern
59 captured by the feature. However, these features are often designed to be independent or may be
60 considered in isolation [6, 7]. When examined in the context of knowledgebases such as the Kyoto
61 Encyclopedia of Genes and Genomes (KEGG) [8], most features are significantly enriched for more than
62 one pathway [5]; it follows then that such features can be described by a set of functionally related
63 pathways. We introduce the PathCORE (**pathway co-occurrence relationships**) software, an approach for
64 connecting features learned from the data to known biological gene sets, e.g. pathways from KEGG or
65 other databases.

66
67 PathCORE offers a data-driven approach for predicting and visualizing global pathway-pathway
68 interactions. Interactions are drawn based on the sets of pathways, annotated in a resource of gene sets,
69 occurring within constructed features. To avoid simply discovering relationships between gene sets that
70 share many genes, PathCORE incorporates an optional pre-processing step that corrects for a situation
71 Donato et al. refer to as pathway crosstalk [9]. Donato et al. recognized that pathways with shared genes
72 were often discovered together due to overlapping genes in gene sets. We implement Donato et al.’s
73 maximum impact estimation in a Python package separate from, but used in, PathCORE (PyPI: crosstalk-
74 correction). With this correction, the PathCORE software allows a user to examine how pathways
75 influence each other in a biological system based on how genes are expressed as opposed to which
76 genes are shared.

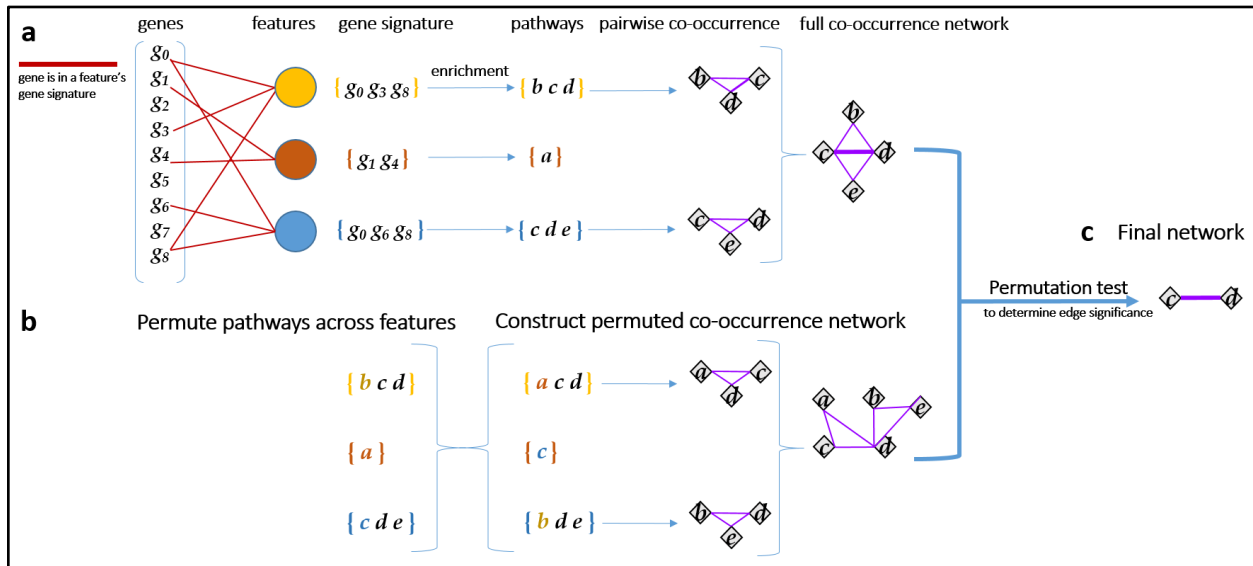
77
78 We demonstrate PathCORE by applying the software to both a microbial and a cancer
79 expression dataset. Briefly, for the microbial analysis we created a network of KEGG pathways from
80 recently described ensemble Analysis using Denoising Autoencoders for Gene Expression (eADAGE)
81 models trained on a compendium of *Pseudomonas aeruginosa* (*P. aeruginosa*) gene expression data [5].
82 We provide a live demo of the PathCORE web application for the *P. aeruginosa* KEGG network at
83 pathcore-demo.herokuapp.com/PAO1. PathCORE can be used with other feature construction
84 approaches as well. For example, we perform PathCORE analysis of the same *P. aeruginosa*
85 compendium using non-negative matrix factorization (NMF) on expression datasets [10, 11]. We also
86 demonstrate PathCORE’s use in large cancer genomics data by creating a Pathway Interaction Database
87 (PID) [12] pathway-pathway network of NMF features extracted from a The Cancer Genome Atlas
88 (TCGA) pan-cancer dataset of 33 different tumor types [13]. In these applications, PathCORE
89 successfully discerns biologically important pathway-pathway interactions from the constructed features.

90 Related work

91 Most published approaches that capture pathway-pathway interactions from gene expression
92 experiments were designed for disease-specific, case-control studies [14, 15]. Pham et al. developed
93 Latent Pathway Identification Analysis to find pathways that exert latent influences on transcriptionally
94 altered genes [16]. Under this approach, the transcriptional response profiles for a binary condition
95 (disease/normal), in conjunction with the pathway specified in the KEGG and functions in Gene Ontology
96 (GO), are used to construct a pathway-pathway network where key pathways are identified by their
97 network centrality scores [16, 17]. Similarly, Pan et al. measured the betweenness centrality of pathways
98 in disease-specific genetic interaction and coexpression networks to identify those most likely to be
99 associated with bladder cancer risk [18]. These methods captured pathway relationships associated with
100 a particular disease state. Our approach diverges from such studies in its intent: PathCORE finds
101 pathway relationships within a biological system that are discernable in features constructed from diverse
102 transcriptomic data--not necessarily specific to any one condition or disease.

103
104
105 Fewer publications to date have focused on the construction of a general pathway-pathway
106 interaction network. Those that did determined the absence or presence of a pathway-pathway interaction
107 based on shared genes between gene sets, protein-protein interactions or other curated knowledgebases
108 [19-22]. A function-based method of constructing a global network, detailed by Li et al., relied on publicly
109 available protein interaction data to determine pathway-pathway interactions [21]. Two pathways were
110 connected in the network if the number of protein interactions between the pair was significant with
111 respect to the computed background distribution. Networks of this kind rely on databases of interactions,

112 though they can be subsequently used for pathway-centric analyses of transcriptomic data [21, 23]. Glass
 113 and Girvan described another network structure that relates functional terms in GO based on shared
 114 gene annotations [24]. In contrast with this approach, PathCORE specifically removes gene overlap in
 115 pathway definitions before they are used to build a network. Our software reports pathway-pathway
 116 connections from global gene expression patterns, as opposed to protein-protein interactions, while
 117 controlling for the fact that some pathways share genes.
 118



119 **Figure 1** The approach implemented in PathCORE to construct a pathway co-occurrence network from
 120 an expression compendium.
 121 (a) A user-selected feature extraction method is applied to expression data. Such methods assign each
 122 gene a weight, according to some distribution, that represents the gene's contribution to the feature. The
 123 set of genes that are considered highly representative of a feature's function is referred to as a feature's
 124 gene signature. The gene signature is user-defined and should be based on the weight distribution
 125 produced by the unsupervised method of choice. In the event that the weight distribution contains both
 126 positive and negative values, a user can specify criteria for both a positive and negative gene signature. A
 127 test of pathway enrichment is applied to identify corresponding sets of pathways from the gene
 128 signature(s) in a feature. We consider pathways significantly overrepresented in the same feature to
 129 co-occur. Pairwise co-occurrence relationships are used to build a network. Each edge is weighted by the
 130 number of features containing both pathways.
 131 (b) N permuted networks are generated to assess the statistical significance of a co-occurrence relation
 132 in the network. Two invariants are maintained during a permutation: (1) pathway side-specificity (positive
 133 and negative, when applicable) and (2) the number of distinct pathways in a feature's (side-specific) gene
 134 signature.
 135 (c) For each edge observed in the co-occurrence network, we compare its weight against the weight
 136 distribution generated from N (default: 10,000) permutations of the network. Edges with a q-value below
 137 alpha (default: 0.05) are kept in the final co-occurrence network.
 138
 139

140 Implementation

141 PathCORE identifies functional links between known pathways from the output of feature
 142 construction methods applied to gene expression data. The result is a network of pathway co-occurrence
 143 relationships that represents the grouping of biological processes or pathways within those features. We
 144 correct for gene overlap in the pathway annotations to avoid identifying co-occurrence relationships
 145 driven by shared genes. Additionally, PathCORE implements a permutation test for evaluating and
 146 removing edges—pathway relationships—in the resulting network that cannot be distinguished from a null
 147 model of random associations. Our software is written in Python and pip-installable (PyPI package name:
 148 pathcore). Each of the functions that we describe here can be used independently; however, we expect

149 most users to employ the complete approach for interpreting pathways shared in extracted features (Fig.
150 4).

151 152 Data organization

153 PathCORE requires the following inputs:

- 154 (1) A **weight matrix** that connects each gene to each feature. We expect that this will be generated
155 by applying a feature construction algorithm to a compendium of gene expression data. In
156 principal component analysis (PCA), this is the loadings matrix [25]; in independent component
157 analysis (ICA), the unmixing matrix [26]; in ADAGE or eADAGE it is termed the weight matrix [5,
158 7]; in NMF it is the matrix W , where the NMF approximation of the input dataset A is $A \sim WH$ [10].
159 The primary requirements are that features must contain the full set of genes in the compendium
160 and genes must have been assigned weights that quantify their contribution to a given feature.
161 Accordingly, a weight matrix will have the dimensions $n \times k$, where n is the number of genes in
162 the compendium and k is the number of features constructed.
- 163 (2) A **gene signature definition**. To construct a pathway co-occurrence network, the weight matrix
164 must be processed into gene signatures by applying a threshold to weights. Subsequent pathway
165 overrepresentation will be determined by the set(s) of genes within these signatures. These are
166 often the weights at the extremes of the distribution. How gene weights are distributed will
167 depend on the user's selected feature construction algorithm; because of this, a user must
168 specify the criterion for including a gene in a gene signature. PathCORE permits rules for a single
169 gene signature or both a positive and a negative gene signature. The use of 2 signatures may be
170 appropriate when the feature construction algorithm produces positive and negative weights, the
171 extremes of which both characterize a feature (e.g. PCA, ICA, ADAGE or eADAGE).
- 172 (3) A list of **pathway definitions**, where each pathway contains a set of genes (e.g. KEGG
173 pathways, PID pathways, GO biological processes).

174 175 Weight matrix construction and signature definition

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

178 179 *eADAGE*

180 eADAGE is an unsupervised feature construction algorithm developed by Tan et al. [5] that uses
181 an ensemble of neural networks (an ensemble of ADAGE models) to capture biological signatures
182 embedded in the expression compendium. By initializing eADAGE with different random seeds, Tan et al.
183 produced 10 eADAGE models that each extracted $k=300$ features from the compendium of genome-scale
184 *P. aeruginosa* data. Because PathCORE supports the aggregation of co-occurrence networks created
185 from different models on the same input data, we use all 10 of these models in the PathCORE analysis of
186 eADAGE models (doi:10.5281/zenodo.583172).

187
188 Tan et al. refers to the features constructed by eADAGE as nodes. They are represented as a
189 weight matrix of size $n \times k$, where n genes in the compendium are assigned positive or negative gene
190 weights, according to a standard normal distribution, for each feature k . Tan et al. determined that the
191 gene sets contributing the highest positive or highest negative weights (± 2.5 standard deviations) to a
192 feature described gene expression patterns across the compendium, and thus referred to the gene sets
193 as signatures. Because a feature's positive and negative gene signatures did not necessarily correspond
194 to the same biological process or function, Tan et al. analyzed each of these sets separately [5]. Tan et
195 al.'s gene signature rules are specified as an input to the PathCORE analysis as well.

196 197 *NMF*

198 We also constructed NMF models for the *P. aeruginosa* dataset and the TCGA pan-cancer dataset.
199 Given an NMF approximation of $A \sim WH$ [10], where A is the input expression dataset of size $n \times s$ (n
200 genes by s samples), NMF aims to find the optimal reconstruction of A by WH such that W clusters on
201 samples (size $n \times k$) and H clusters on genes (size $k \times s$). We set k to the desired number of features,
202 $k=300$, and use W as the input weight matrix for the PathCORE software. We found that the gene weight
203 distribution of an NMF feature is right-skewed and (as the name suggests) non-negative (Fig. S1). In this

204 case, we defined the gene signature to be the set of genes with weights 2.0 standard deviations above
205 the mean weight of each feature.

206 Construction of a pathway co-occurrence network

207 We employ a Fisher's exact test to determine the pathways significantly associated with each
208 gene signature. When considering significance of a particular pathway, the two categories of gene
209 classification are as follows: (1) presence or absence of the gene in the gene signature and (2) presence
210 or absence of the gene in the pathway definition. We specify a contingency table for each pathway and
211 calculate the p-value. After false discovery rate (FDR) correction, pathways with a q-value of less than
212 alpha (default: 0.05) are considered significantly enriched. Two pathways co-occur, or share an edge in
213 the pathway co-occurrence network, if they are both overrepresented in a gene signature. The number of
214 times such a pathway pair is present over all features corresponds to its edge weight in the pathway-
215 pathway network (Fig. 1a).

217 Permutation test

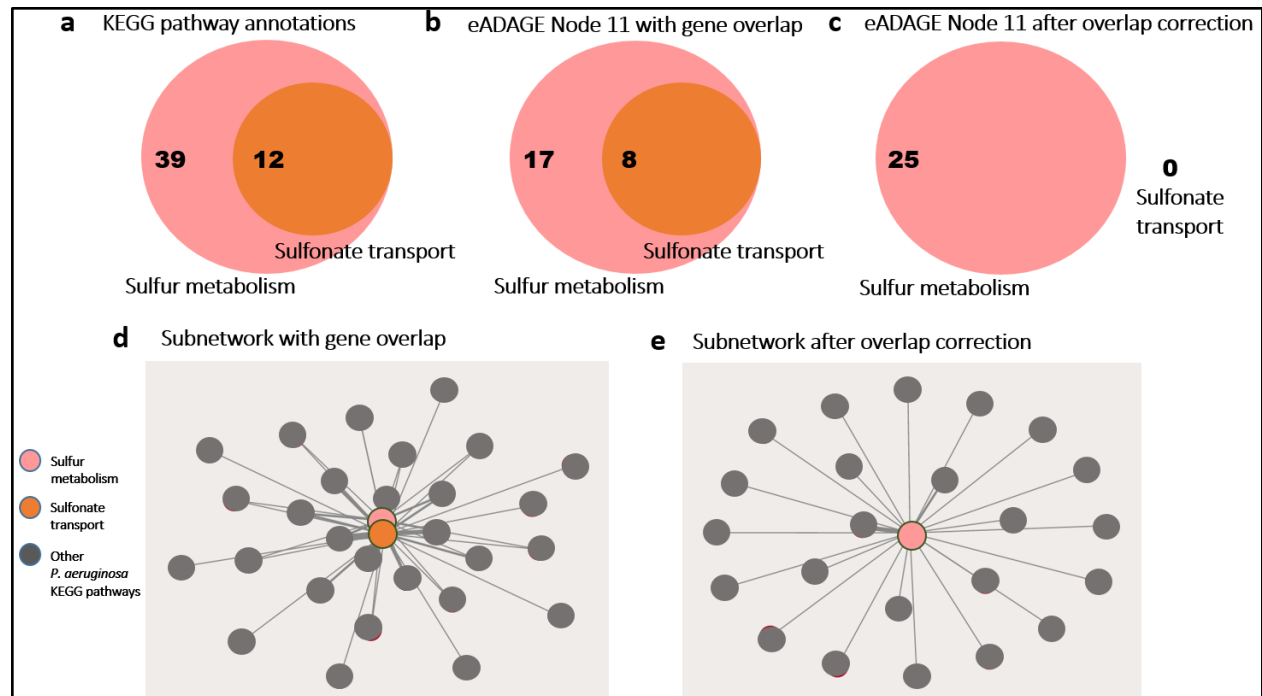
218 The network that results from the preceding method is densely connected, and many edges may
219 be spurious. To remove correlations that cannot be distinguished from random pathway associations, we
220 define a statistical test that determines whether a pathway-pathway relationship appearing x times in a k -
221 feature model is unexpected under the null hypothesis—the null hypothesis being that the relationship
222 does not appear more often than it would in a random network. We create N weighted null networks by
223 permuting overrepresented pathways across the model's features while preserving the number of
224 pathways for which each feature is enriched. In the case where we have positive and negative gene
225 signatures, overrepresentation can be positive or negative. Because certain pathways may display bias
226 toward one side—for example, a pathway may be overrepresented more often in features' positive gene
227 signatures—we perform the permutation separately for each side. The N random networks produce the
228 background weight distribution for every observed edge; significance can then be assessed by comparing
229 the true (observed) edge weight against the null (Fig. 1b). Pathway-pathway relationships with a q-value
230 above alpha (default: 0.05) are considered insignificant by this statistical test and are removed from the
231 network of co-occurring pathways (Fig. 1c).

232
233 Because we can derive the expected weight of every edge from the N random networks, we can
234 divide the observed edge weights by their respective expected weights (divide by 1 if the edge is not
235 present in any of the N permutations). Edges in the final network are weighted by their odds ratios.

237 Gene overlap correction

238 Pathways can co-occur because of shared genes (Fig. 2a, b, d). Though some approaches use
239 the overlap of genes to identify connected pathways, we sought to capture pairs of pathways that
240 persisted even when this overlap was removed. The phenomenon of observing enrichment of multiple
241 pathways due to gene overlap has been previously termed as “crosstalk,” and Donato et al. have
242 developed a method to correct for it [9]. Due to confusion around the term, we refer to this as *overlapping*
243 *genes* in this work, except where specifically referencing Donato et al. Their approach, called maximum
244 impact estimation, begins with a membership matrix indicating the original assignment of multiple genes
245 to multiple pathways. It uses expectation maximization to estimate the pathway in which a gene
246 contributes its greatest predicted impact (its maximum impact) and assigns the gene only to this pathway
247 [9]. This provides a set of new pathway definitions that no longer share genes (Fig. 2c, e).

248
249 We provide an implementation of Donato et al.'s maximum impact estimation as a Python
250 package separate from PathCORE so that it is available for any pathway analyses (PyPI package name:
251 crosstalk-correction). The procedure is written using NumPy functions and data structures, which allows
252 for efficient implementation of array and matrix operations in Python [28]. In PathCORE, overlapping
253 genes are addressed before pathway overrepresentation analysis so that the resulting pathway co-
254 occurrence network identifies interactions that are not driven by gene overlap. We incorporate this
255 correction into the PathCORE workflow by default; however, users can choose to disable it as well.



256
257
258
259
260
261
262
263
264
265
266
267

Figure 2 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.

PathCORE network visualization and support for experimental follow-up

As an optional step, a Flask application can be set up 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 [29]. 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. 3a).

We created a web interface for deeper examination of interactions 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

274
275
276
277
278
279
280
281
282
283
284
285
286
287
288

289 calculate the proportion of observed cases and divide by the expected proportion--equivalent to the
 290 frequency of the edge appearing in the model's features.

291
 292 Let K be the number of features from which the PathCORE network was built. K_G is the number of
 293 features that contain gene G (i.e. G is in feature K 's gene signature), K_E the number of features that
 294 contain edge E (i.e. the two pathways connected by E are overrepresented in feature K), and $K_{G \& E}$ the
 295 number of features that contain both gene G and edge E . The odds ratio is computed as follows:

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

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

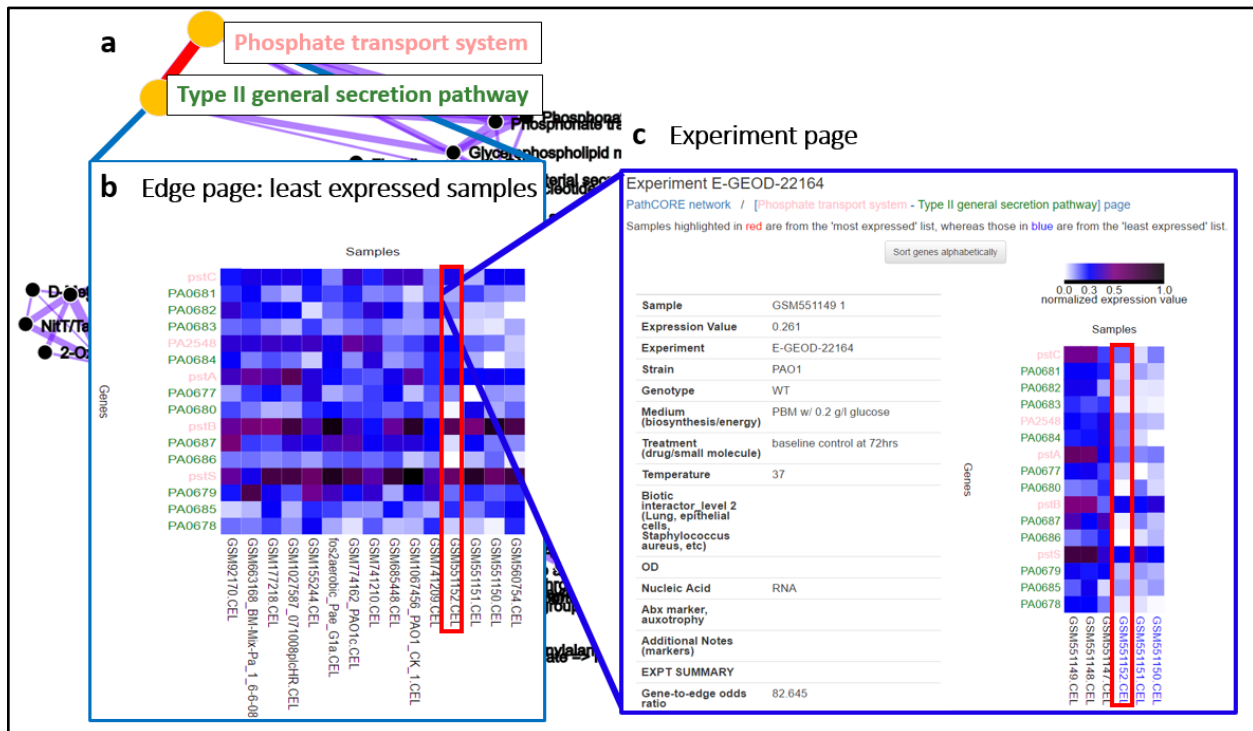
$$\text{Odds ratio} = \text{Observed} / \text{Expected}$$

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

303 The information specified in (2) requires an "expression score" for every sample. A sample
 304 expression score is calculated using the twenty genes we selected in goal (1): it is the average of the
 305 normalized gene expression values weighted by the normalized gene odds ratio. Selection of the most
 306 and least expressed samples is based on these scores. We use two heatmaps to show the twenty genes'
 307 expression values in each of the fifteen most and least expressed samples (Fig. 3b).

309 For each sample in an edge page, a user can examine how the expression values of the edge's
 310 twenty genes in that sample compare to those recorded for all other samples in the dataset that are from
 311 the same experiment (Fig. 3c). Genes that display distinct expression patterns under a specific setting
 312 may be good candidates for follow-up studies.

313



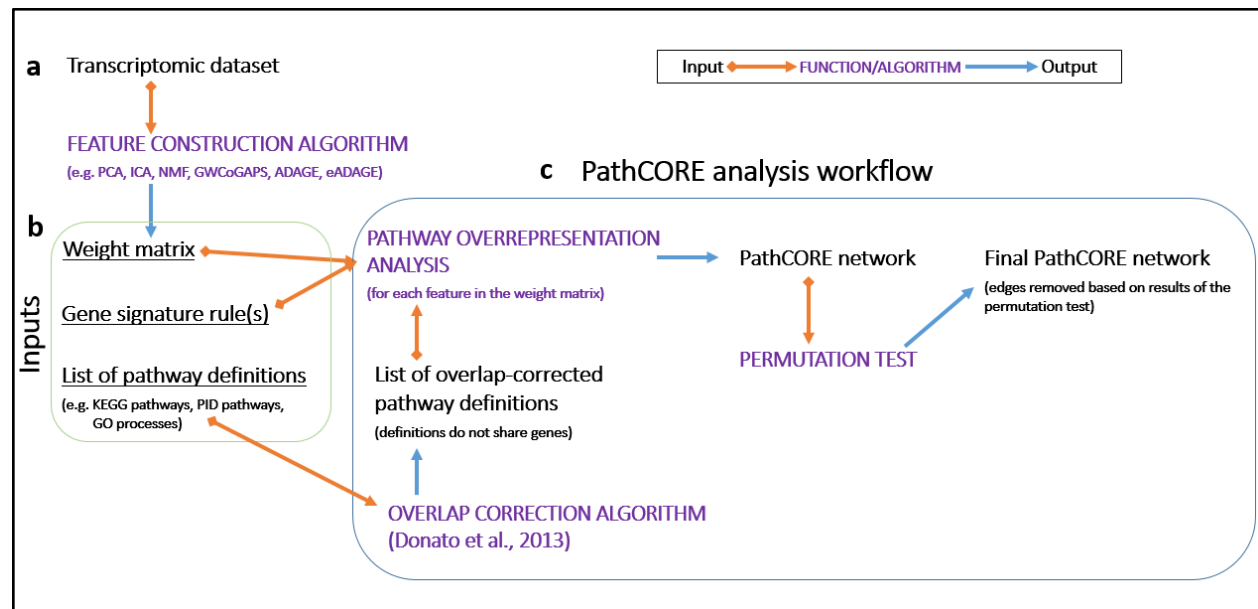
314 **Figure 3** A web application used to analyze pathway-pathway interactions in the eADAGE-based, *P.*
 315 *aeruginosa* KEGG network.
 316 (a) A user clicks on an edge (a pathway-pathway interaction) in the network visualization and (b) is
 317 directed to a page that displays expression data from the original transcriptomic dataset specific to the
 318 selected edge (goo.gl/Hs5A3e). The expression data is visualized as two heatmaps that indicate the
 319 fifteen most and fifteen least expressed samples corresponding to the edge. To select the "most" and
 320 "least" expressed samples, we assign each sample a summary "expression score." The expression score
 321

322 is based on the expression values of the genes (limited to the top twenty genes with an odds ratio above
323 1) annotated to one or both of the pathways. Here, we show the heatmap of least expressed samples
324 specific to the [Phosphate transport - Type II general secretion] relationship. (c) Clicking on a square in
325 the heatmap directs a user to an experiment page (goo.gl/KYNhwB) based on the sample corresponding
326 to that square. A user can use the experiment page to identify whether the expression values of genes
327 specific to an edge and a selected sample differ from those recorded in other samples of the experiment.
328 In this experiment page, the first three samples (labeled in black) are *P. aeruginosa* “baseline” replicates
329 grown for 72 h in drop-flow biofilm reactors. The following three samples (labeled in blue) are *P.*
330 *aeruginosa* grown for an additional 12 h (84 h total). Labels in blue indicate that the three 84 h replicates
331 are in the heatmap of least expressed samples displayed on the [Phosphate transport – Type II general
332 secretion] edge page.
333

334 Results

335 Interpreting features extracted by unsupervised clustering algorithms with PathCORE.

336 Networks modeling the relationships between curated processes in a biological system offer a
337 means for developing new hypotheses about which pathways influence each other and when. PathCORE
338 creates a network of globally co-occurring pathways based on features observed in a compendium of
339 gene expression data. Biological patterns in the data are extracted by a feature construction algorithm
340 such as PCA [30], ICA [6], NMF [10], GWCoGAPS [4], or eADAGE [5]. These algorithms capture sources
341 of variability in the data that induce coordinated changes in gene expression as features. The genes that
342 contribute the most to these features covary. This provides a data-driven categorization of the biological
343 system that can then be analyzed at the pathway-level by identifying annotated pathways
344 overrepresented in each feature.
345

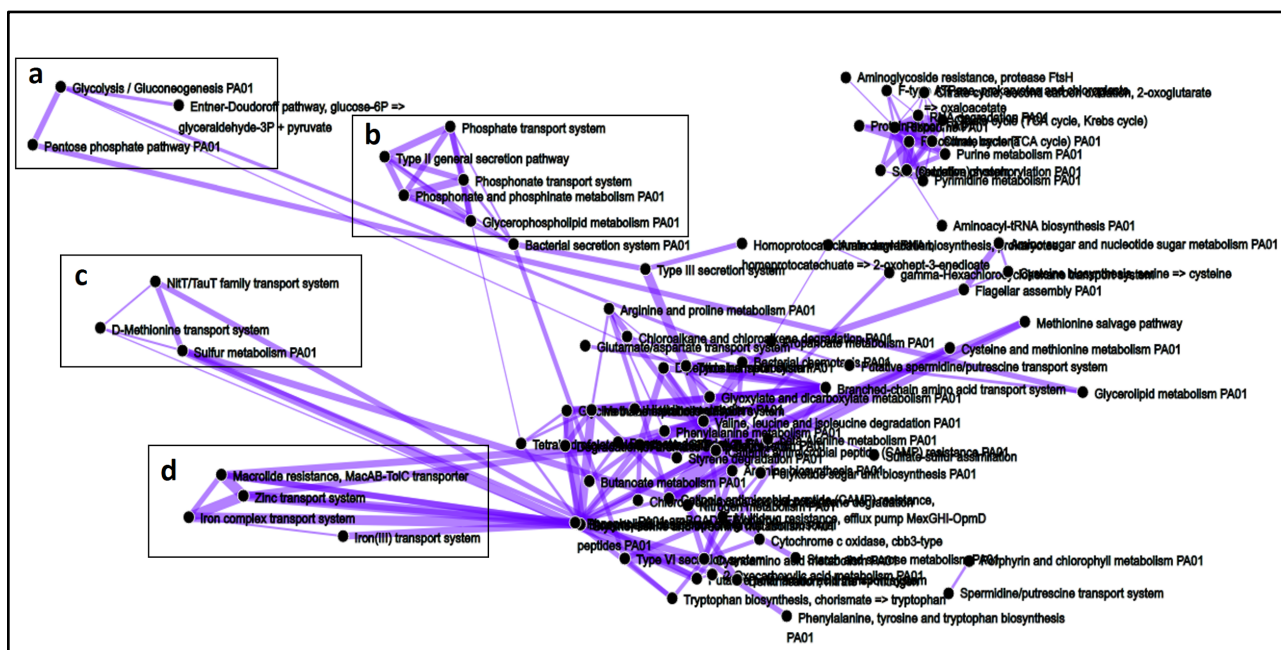


346
347 **Figure 4** The PathCORE software analysis workflow.

348 (a) A user applies a feature construction algorithm to a transcriptomic dataset of genes-by-samples. The
349 features constructed must preserve the genes in the dataset and assign weights to each of these genes
350 according to some distribution. (b) Inputs required to run the complete PathCORE analysis workflow. The
351 features constructed are stored in a weight matrix and the user-defined gene signature rules--up to 2 for
352 both a positive and negative gene signature--should be based on the algorithm's specified feature weight
353 distribution. A list of pathway definitions will be used to interpret the features constructed and build a
354 pathway-pathway co-occurrence network. (c) Methods in the PathCORE analysis workflow (capitalized
355 and in purple), can be employed independently of each other so long as the necessary input(s) are
356 provided.
357

358 The methods we implement in PathCORE can be used independently of each other (Fig. 4).
 359 Here, we present analyses that can be produced by applying the full PathCORE pipeline to models
 360 created from a transcriptomic compendium by an unsupervised feature construction algorithm. Input
 361 pathway definitions are “overlap-corrected” (correcting for gene overlap between definitions) for each
 362 feature before enrichment analysis. An overlap-corrected, weighted pathway co-occurrence network is
 363 built by connecting the pairs of pathways that are overrepresented in features of the model. Finally, we
 364 remove edges that cannot be distinguished from a null model of random associations based on the
 365 results of a permutation test.

367 PathCORE also offers support for users interested in experimentally verifying a pathway-pathway
 368 relationship (Fig. 3). We provide the code for setting up a web application where the network can be
 369 visualized and its edges analyzed using the original input information. A pathway-pathway edge page
 370 contains 2 heatmaps that display the samples in the compendium where the underlying genes are most
 371 and least expressed. When available, information about each sample can be included on the page so that
 372 users can refer to the conditions in which the expression patterns occurred (Fig. 3c).
 373



374 **Figure 5** eADAGE features constructed from publicly available *P. aeruginosa* expression data describe
 375 known KEGG pathway relationships.
 376 (a) The glycolysis/gluconeogenesis, pentose phosphate, and Entner-Doudoroff pathways share common
 377 functions related to glucose catabolism .
 378 (b) Organophosphate and inorganic phosphate transport- and metabolism-related processes frequently
 379 co-occur with bacterial secretion systems; in particular, we highlight the pairwise relationships between
 380 type II secretion and the phosphate-related processes.
 381 (c) Pathways involved in the catabolism of sulfur-containing molecules, taurine (NitT/TauT family
 382 transport) and methionine (D-Methionine transport), and the general sulfur metabolism process are
 383 functionally linked.
 384 (d) We observe pairwise relationships between zinc transport, iron transport, and the MacAB-ToIC
 385 transporter.
 386

387
 388 PathCORE identifies interactions between KEGG pathways in *P. aeruginosa* using features extracted
 389 from publicly available *P. aeruginosa* gene expression experiments

390 We used PathCORE to create a network of co-occurring pathways out of the expression
 391 signatures extracted from a *P. aeruginosa* compendium. For every feature, overlap correction was
 392 applied to the *P. aeruginosa* KEGG pathway annotations and overlap-corrected annotations were used in
 393 the overrepresentation analysis. PathCORE aggregates multiple networks by taking the union of the

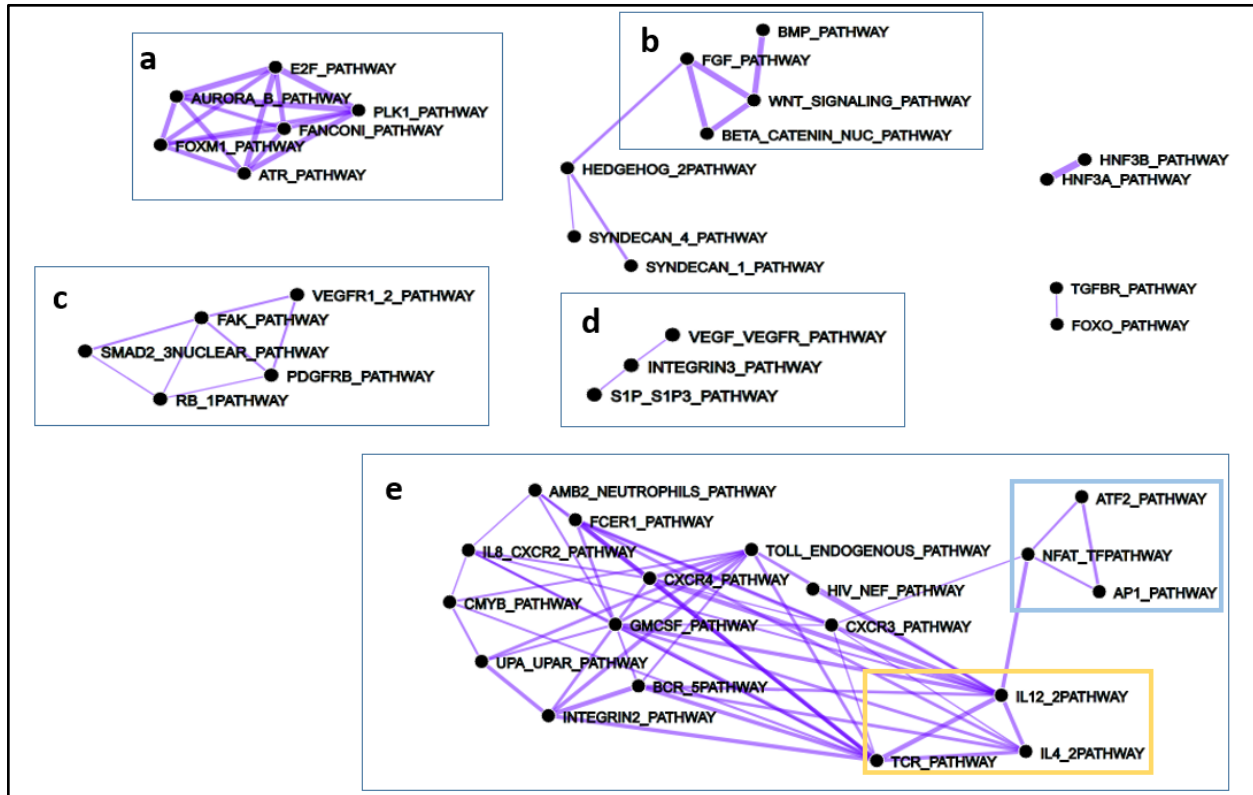
394 edges across all networks and summing the weights of common pathway-pathway connections. We do
395 this to emphasize the pathway-pathway co-occurrence relationships that are more stable [31] —that is,
396 the relationships that appear across multiple models. Finally, we removed edges in the aggregate
397 network that were not significant after FDR correction when compared to the background distributions
398 generated from 10,000 permutations of the network. We applied PathCORE to features built by both NMF
399 (Fig. S2) and eADAGE (discussed below). The pathway-pathway network generated by NMF [27] is
400 smaller than that generated by eADAGE; that is, there are fewer pathway-pathway connections in the
401 network. It is possible that this is due to a difference in the stability of the results from the two algorithms
402 or the comprehensiveness of the features extracted by each approach. eADAGE includes an ensemble
403 step that improves model consistency. The eADAGE authors also observed that models constructed by
404 this ensemble procedure also more comprehensively captured pathways than non-ensemble models [5].
405 The PathCORE analysis of a 300 feature NMF decomposition of the *P. aeruginosa* compendium
406 produced a KEGG network that is similar in size to the PID network (Fig. 6, S2).

407
408 The eADAGE co-occurrence network identifies a number of pathway-pathway interactions that
409 have been previously characterized (Fig. 5). This suggests that PathCORE can capture functional links
410 between biological pathways. Three glucose catabolism processes co-occur in the network: glycolysis,
411 pentose phosphate, and the Entner-Doudoroff pathway (Fig. 5a). We also found a cluster relating
412 organophosphate and inorganic phosphate transport- and metabolism-related processes (Fig. 5b).
413 Notably, phosphate uptake and acquisition genes are directly connected to the *hxc* genes that encode a
414 type II secretion system. This Hxc secretion system is responsible for the secretion of alkaline
415 phosphatases, which are phosphate scavenging enzymes [32, 33] and the phosphate binding DING
416 protein [34]. Furthermore, alkaline phosphatases, DING and the *hxc* genes are regulated by the
417 transcription factor PhoB which is most active in response to phosphate limitation. As shown in Fig. 5c,
418 we also identified linkages between two pathways involved in the catabolism of sulfur-containing
419 molecules, taurine and methionine, and the general sulfur metabolism process. Other connections
420 between pathways involved in the transport of iron (ferrienterobactin binding) [35] and zinc (the *znu*
421 uptake system [36]) were identified (Fig. 5d). Interestingly, genes identified in the edge between the zinc
422 transport and MacAB-TolC pathways include the *pvd* genes involved in pyoverdine biosynthesis and
423 regulation, a putative periplasmic metal binding protein, as well as other components of an ABC
424 transporter (genes PA2407, PA2408, and PA2409 at goo.gl/bfqOk8) [37].

425
426 We used the PathCORE web application for the eADAGE KEGG *P. aeruginosa* network
427 (pathcore-demo.herokuapp.com/PAO1) to analyze the connection between the phosphate transport
428 system and a type II general secretion system pathway (Fig. 3a, b; edge page at goo.gl/Hs5A3e). The
429 sixteen genes reported on the edge page all have odds ratios above 34; these genes are at least 34
430 times more likely to appear in the gene signatures in which both of these pathways are overrepresented.
431 Such genes may help to reveal the biological basis of the co-occurrence relationship. In this case, the
432 results suggest that there may be some overlap between machinery for transporting phosphate into the
433 cell and secreting substances out of the cell via type II secretion. Alternatively, the two processes may be
434 mechanistically separate but coregulated such that phosphate scavenging molecules may be secreted by
435 type II secretion coincidentally to aid in phosphate acquisition. The heatmap of the fifteen least expressed
436 samples shows that the *pstB* and *pstS* genes, annotated to the phosphate transport system, are
437 consistently expressed higher relative to the other genes in the edge for these samples. The *pstS*, *pstC*,
438 *pstA*, and *pstB* genes are proximal of the phosphate-specific transport (Pst) operon that encodes a high-
439 affinity orthophosphate transport system [38]. Future studies will examine whether deletion of the Pst
440 phosphate system impairs secretion by the type II secretion system.

441
442 To assess whether the relationships identified in the pathway analysis paralleled gene expression
443 patterns in the context of a published experiment, we looked across experiments to determine if the
444 genes contained in the edge were co-regulated. As an example of the types of relationships that we
445 observed, we present a single experiment, E-GEOD-22164 from Folsom et al. (Fig. 3c; experiment page
446 at goo.gl/KYNhwB), that contains data from two sample types with three replicates each [39]. One set of
447 samples, referred to as the baseline, is of *P. aeruginosa* grown for 72 h in drop-flow biofilm reactors. The
448 other is of *P. aeruginosa* grown for an additional 12 h (84 h total). We found that three of the samples with
449 the lowest expression of the genes within the shared edge were samples from the later timepoint. All six

450 of these samples (three replicates each) are displayed in the experiment page. The sixteen edge genes
 451 showed differential expression between the 72 h and 84 h timepoints. Particularly in the case of *pstA* and
 452 *pstC*, the baseline replicates had normalized expression values near the center of the range from the
 453 compendium whereas the 84 h samples had expression levels at the low end of the range. This suggests
 454 that genes involved in both phosphate transport and type II secretion are less expressed at 84 h
 455 compared to 72 h. Future studies will determine if this is due to a physiological change in the biofilm cells
 456 at the late time point such that phosphate demands were lower or different sources of phosphate become
 457 available.
 458



459 **Figure 6** PID pathway-pathway interactions discovered in NMF features constructed from the TCGA pan-
 460 cancer gene expression dataset.
 461

462 (a) Pathways in this module are responsible for cell cycle progression.
 463 (b) Wnt signaling interactions with nuclear Beta-catenin signaling, FGF signaling, and BMP signaling
 464 have all been linked to cancer progression.
 465 (c) Here, we observe functional links between pathways responsible for angiogenesis and those
 466 responsible for cell proliferation.
 467 (d) The VEGF-VEGFR pathway interacts with the S1P3 pathway through Beta3 integrins.
 468 (e) This module contains many interactions related to immune system processes. The interaction cycle
 469 formed by T-Cell Receptor (TCR) signaling in naïve CD4+ T cells and IL-12/IL-4 mediated signaling
 470 events, outlined in yellow, is one well-known example. The cycle in blue is formed by the ATF2, NFAT,
 471 and AP1 pathways; pairwise co-occurrence of these three transcription factor networks may suggest that
 472 dysregulation of any one of these pathways can trigger variable oncogenic processes in the immune
 473 system.
 474

475 PathCORE identifies interactions among PID pathways from the TCGA pan-cancer gene expression
 476 dataset.

477 PathCORE is not specific to a certain dataset or organism. We also constructed a 300-feature
 478 NMF model of TCGA pan-cancer gene expression, which is comprised of 33 different cancer-types from
 479 various organ sites, and applied the PathCORE software to those features. We chose NMF because it
 480 has been used in previous studies to identify biologically relevant patterns in transcriptomic data [10] and

481 by many studies to derive molecular subtypes [40-42]. The 300 NMF features were analyzed using
482 overlap-corrected PID pathways, a collection of 196 human cell signaling pathways with a particular focus
483 on processes relevant to cancer [12].
484

485 We found that PathCORE detects modules of co-occurring pathways consistent with our current
486 understanding of cancer-related interactions (Fig. 6). Importantly, because the connections were
487 constructed from many different cancer-types, these modules may represent pathway-pathway
488 interactions present in a large proportion of all tumors and may be good candidates for targeted
489 treatments.
490

491 For example, a module composed of a FoxM1 transcription factor network, an E2F transcription
492 factor network, Aurora B kinase signaling, ATR signaling, PLK1 signaling, and members of the Fanconi
493 anemia DNA damage response pathway are densely connected (Fig 6a). When two pathways share an
494 edge in the co-occurrence network, they are overrepresented together in one or more features. The
495 connections in this module recapitulate well known cancer hallmarks including cellular proliferation
496 pathways and markers of genome instability, such as the activation of DNA damage response pathways
497 [43]. We found that several pairwise pathway co-occurrences correspond with previously reported
498 pathway-pathway interactions [44-46]. We also observed a hub of pathways interacting with Wnt signaling
499 (Fig. 6b). In our network, pathways that co-occur with Wnt signaling include the regulation of nuclear
500 Beta-catenin signaling, FGF signaling, and BMP signaling. The Wnt and BMP pathways are functionally
501 integrated in biological processes that contribute to cancer progression [47]. Additionally, Wnt/Beta-
502 catenin signaling is a well-studied regulatory system, and the effects of mutations in Wnt pathway
503 components on this system have been linked to tumorigenesis [48]. Wnt/Beta-catenin and FGF together
504 influence the directional migration of cancer cell clusters [49].
505

506 Two modules in the network relate to angiogenesis, or the formation of new blood vessels (Fig.
507 6c, d). Tumors transmit signals that stimulate angiogenesis because a blood supply provides the
508 necessary oxygen and nutrients for their growth and proliferation. One module relates angiogenesis
509 factors to cell proliferation. This module connects the following pathways: PDGFR-beta signaling, FAK-
510 mediated signaling events, VEGFR1 and VEGFR2-mediated signaling events, nuclear SMAD2/3
511 signaling regulation, and RB1 regulation (Fig. 6c). These functional connections are known to be involved
512 in tumor proliferation [50-52]. The other module indicates a direct relationship by which the VEGF
513 pathway interacts with the S1P3 pathway through Beta3 integrins (Fig. 6d). *S1P3* is a known regulator of
514 angiogenesis [53], and has been demonstrated to be associated with treatment-resistant breast cancer
515 and poor survival [54]. Moreover, this interaction module has been observed to promote endothelial cell
516 migration in human umbilical veins [55]. Taken together, this independent module may suggest a distinct
517 angiogenesis process activated in more aggressive and metastatic tumors that is disrupted and regulated
518 by alternative mechanisms [56].
519

520 Finally, PathCORE revealed a large, densely connected module of immune related pathways
521 (Fig. 6e). We found that this module contains many interactions that align with immune system
522 processes. One such example is the well characterized interaction cycle formed by T-Cell Receptor
523 (TCR) signaling in naïve CD4+ T cells and IL-12/IL-4 mediated signaling events [57-59]. At the same
524 time, PathCORE predicts additional immune-related interactions. We observed a cycle between the three
525 transcription factor networks: ATF-2, AP-1, and CaN-regulated NFAT-dependent transcription. These
526 pathways can take on different, often opposing, functions depending on the tissue and subcellular
527 context. For example, ATF-2 can be an oncogene in one context (e.g. melanoma) and a tumor
528 suppressor in another (e.g. breast cancer) [60]. AP-1, comprised of Jun/Fos proteins, is associated with
529 both tumorigenesis and tumor suppression due to its roles in cell survival, proliferation, and cell death
530 [61]. Moreover, NFAT in complex with AP-1 regulates immune cell differentiation, but dysregulation of
531 NFAT signaling can lead to malignant growth and tumor metastasis [62]. The functional association
532 observed between the ATF-2, AP-1, and NFAT cycle together within the immunity module might suggest
533 that dysregulation within this cycle has profound consequences for immune cell processes and may
534 trigger variable oncogenic processes.
535

536 **Conclusions**

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

541
542 Gene set analyses can be heavily confounded by shared genes. Some pathways may be
543 observed together because they depend on each other, while others may simply contain some of the
544 same genes. In PathCORE, pathway annotations undergo a procedure called maximum impact
545 estimation, described in a publication by Donato et al., that maps each gene in each feature to the one
546 pathway in which it has the greatest estimated impact [9]. We provide this overlap correction algorithm as
547 a Python package (PyPI package name: crosstalk-correction) available under the BSD 3-Clause license.
548 Though the algorithm had been described, no publicly available implementation existed.

549
550 PathCORE includes software for analysis and visualization and can be broadly applied to
551 constructed features. We demonstrate PathCORE in two different contexts, analyses of the bacterium *P.*
552 *aeruginosa* and human pan-cancer datasets, using two different feature construction methods (eADAGE
553 and NMF). We provide a demonstration application containing the results of the eADAGE *P. aeruginosa*
554 analysis for researchers to explore. For each edge, users can explore heatmaps displaying the
555 expression levels of predicted driver genes in the original samples. This provides support for assessing
556 computationally-derived relationships in experimental follow-ups.

557
558 Unsupervised analyses of genome-scale datasets that summarize key patterns in the data have
559 the potential to improve our understanding of how a biological system operates via complex interactions
560 between molecular processes. However, interpreting the features generated by unsupervised approaches
561 is still challenging. PathCORE is a component of a software ecosystem that connects the features
562 extracted from data to curated resources. The specific niche that PathCORE aims to fill is in revealing to
563 researchers which gene sets most are most closely related to each other in machine learning-based
564 models of gene expression, which genes play a role in this co-occurrence, and which conditions drive this
565 interaction will help researchers most effectively employ these algorithms.

566
567 **Project name:** PathCORE

568 **Project home page:** <https://pathcore-demo.herokuapp.com>

569 **Archived version:** <https://github.com/greenelab/PathCORE-analysis/releases/tag/v1.0> (links to
570 download .zip and .tar.gz files are provided here)

571 **Operating system:** Platform independent

572 **Programming language:** Python

573 **Other requirements:** Python 3 or higher

574 **License:** BSD 3-clause

575

576 **Declarations**

577

578 **Ethics approval and consent to participate**

579 Not applicable

580

581 **Consent for publication**

582 Not applicable

583

584 **Availability of data and materials**

585 Files for each of the PathCORE networks described in the results are provided in Supplementary
586 material.

587

588 *Data sets*

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

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

591

592 *Source code* (all links are from <https://github.com/>)

593 **PathCORE analysis:** ([greenelab/PathCORE-analysis/tree/v1.0](https://github.com/greenelab/PathCORE-analysis/tree/v1.0)) This repository contains all the scripts to
594 reproduce the analyses described in this paper. The Python scripts here should be used as a starting
595 point for new PathCORE analyses. Instructions for setting up a web application for a user's specific
596 PathCORE analysis are provided in this repository's README.

597
598 **Overlap correction:** ([kathyxchen/crosstalk-correction/tree/v1.0.4](https://github.com/kathyxchen/crosstalk-correction/tree/v1.0.4)) Donato et. al's procedure for overlap
599 correction [9] is a pip-installable Python package 'crosstalk-correction' that is separate from, but listed as
600 a dependency in, PathCORE. It is implemented using NumPy [28].

601
602 **PathCORE methods:** ([greenelab/PathCORE/tree/v1.0](https://github.com/greenelab/PathCORE/tree/v1.0)) The methods included in the PathCORE analysis
603 workflow (Fig. 4c) are provided as a pip-installable Python package 'pathcore'. It is implemented using
604 Pandas [62], SciPy (specifically, `scipy.stats`) [63], StatsModels [64], and the crosstalk-correction package.

605
606 **PathCORE demo application:** ([kathyxchen/PathCORE-demo/tree/v1.0](https://github.com/kathyxchen/PathCORE-demo/tree/v1.0)) The project home page, pathcore-
607 demo.herokuapp.com provides links to

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

614 **Competing interests**

615 The authors declare that they have no competing interests

616 **Funding**

617
618 KMC was funded by an undergraduate research grant from the Penn Institute for Biomedical Informatics.
619 CSG was funded in part by a grant from the Gordon and Betty Moore Foundation (GBMF 4552). JT, GD,
620 DAH, and CSG were funded in part by a pilot grant from the Cystic Fibrosis Foundation (STANTO15R0).
621 DAH was funded in part by R01-AI091702.

622 **Authors' contributions**

623
624 KMC implemented the software, performed the analyses, and drafted the manuscript. JT and GPW
625 contributed computational reagents. KMC, DAH, and CSG designed the project. KMC, JT, GPW, GD,
626 DAH, and CSG interpreted the results. JT, GPW, GD, DAH, and CSG provided critical feedback and
627 revisions on the manuscript. JT and GPW reviewed source code.

628 **Acknowledgements**

629
630 The authors would also like to thank Daniel Himmelstein, Kurt Wheeler, and René Zelaya for helping to
631 review the source code.

632 **References**

- 633
634 1. Greene CS, Foster JA, Stanton BA, Hogan DA, Bromberg Y: **Computational Approaches to Study
635 Microbes and Microbiomes**. Pacific Symposium on Biocomputing 2016, **21**:557.
- 636 2. Tatlow PJ, Piccolo SR: **A cloud-based workflow to quantify transcript-expression levels in public
637 cancer compendia**. Scientific Reports 2016, **6**.
- 638 3. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, Dylag M, Kurbatova N,
639 Brandizi M, Burdett T: **ArrayExpress update—simplifying data submissions**. Nucleic Acids Res 2014,
640 :gku1057.
- 641 4. Stein-O'Brien G, Carey J, Lee W, Considine M, Favorov A, Flam E, Guo T, Li L, Marchionni L,
642 Sherman T: **PatternMarkers and Genome-Wide CoGAPS Analysis in Parallel Sets (GWCoGAPS) for
643 data-driven detection of novel biomarkers via whole transcriptome Non-negative matrix
644 factorization (NMF)**. bioRxiv 2016, :083717.
- 645 5. Tan J, Doing G, Lewis KA, Price CE, Chen KM, Cady KC, Perchuk B, Laub MT, Hogan DA, Greene
646 CS: **System-wide automatic extraction of functional signatures in *Pseudomonas aeruginosa* with
647 eADAGE**. Cell Systems 2016, In Press; doi:10.1101/078659.
- 648

- 649 6. Engreitz JM, Daigle BJ, Marshall JJ, Altman RB: **Independent component analysis: mining**
650 **microarray data for fundamental human gene expression modules.** J Biomed Inform 2010,
651 **43(6):932-944.**
- 652 7. Tan J, Hammond JH, Hogan DA, Greene CS: **Adage-based integration of publicly available**
653 ***Pseudomonas aeruginosa* gene expression data with denoising autoencoders illuminates**
654 **microbe-host interactions.** mSystems 2016, **1(1):25.**
- 655 8. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** Nucleic Acids Res 2000,
656 **28(1):27-30.**
- 657 9. Donato M, Xu Z, Tomoiaga A, Granneman JG, MacKenzie RG, Bao R, Than NG, Westfall PH, Romero
658 R, Draghici S: **Analysis and correction of crosstalk effects in pathway analysis.** Genome Res 2013,
659 **23(11):1885-1893.**
- 660 10. Brunet J, Tamayo P, Golub TR, Mesirov JP: **Metagenes and molecular pattern discovery using**
661 **matrix factorization.** Proceedings of the national academy of sciences 2004, **101(12):4164-4169.**
- 662 11. Kim PM, Tidor B: **Subsystem identification through dimensionality reduction of large-scale**
663 **gene expression data.** Genome Res 2003, **13(7):1706-1718.**
- 664 12. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH: **PID: the Pathway**
665 **Interaction Database.** Nucleic Acids Res 2009, **37(suppl 1):D679.**
- 666 13. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, Shmulevich I, Sander C,
667 Stuart JM, Cancer Genome Atlas Research Network: **The Cancer Genome Atlas pan-cancer analysis**
668 **project.** Nat Genet 2013, **45(10):1113-1120.**
- 669 14. Visakh R, Nazeer KA: **Identifying epigenetically dysregulated pathways from pathway-pathway**
670 **interaction networks.** Comput Biol Med 2016, **76:160-167.**
- 671 15. Yang J, Luo R, Yan Y, Chen Y: **Differential pathway network analysis used to identify key**
672 **pathways associated with pediatric pneumonia.** Microb Pathog 2016, **101:50-55.**
- 673 16. Pham L, Christadore L, Schaus S, Kolaczyk ED: **Network-based prediction for sources of**
674 **transcriptional dysregulation using latent pathway identification analysis.** Proceedings of the
675 National Academy of Sciences 2011, **108(32):13347-13352.**
- 676 17. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS,
677 Eppig JT: **Gene Ontology: tool for the unification of biology.** Nat Genet 2000, **25(1):25-29.**
- 678 18. Anonymous *Proceedings of the Ecal*: Citeseer; 2013.
- 679 19. Bell L, Chowdhary R, Liu JS, Niu X, Zhang J: **Integrated bio-entity network: a system for**
680 **biological knowledge discovery.** PloS one 2011, **6(6):e21474.**
- 681 20. Chowbina SR, Wu X, Zhang F, Li PM, Pandey R, Kasamsetty HN, Chen JY: **HPD: an online**
682 **integrated human pathway database enabling systems biology studies.** BMC Bioinformatics 2009,
683 **10(11):S5.**
- 684 21. Li Y, Agarwal P, Rajagopalan D: **A global pathway crosstalk network.** Bioinformatics 2008,
685 **24(12):1442-1447.**
- 686 22. Wu X, Chen JY: **Molecular interaction networks: topological and functional characterizations.**
687 *Automation in Proteomics and Genomics: An Engineering Case-Based Approach* 2009, **145.**
- 688 23. de Anda-Juregui G, Meja-Pedroza RA, Espinal-Enriquez J, Hernandez-Lemus E: **Crosstalk events in**
689 **the estrogen signaling pathway may affect tamoxifen efficacy in breast cancer molecular**
690 **subtypes.** Computational biology and chemistry 2015, **59:42-54.**
- 691 24. Glass K, Girvan M: **Finding new order in biological functions from the network structure of gene**
692 **annotations.** PLoS Comput Biol 2015, **11(11):e1004565.**
- 693 25. Abdi H, Williams LJ: **Principal component analysis.** Wiley interdisciplinary reviews: computational
694 statistics 2010, **2(4):433-459.**
- 695 26. Stone JV: *Independent component analysis*: Wiley Online Library; 2004.
- 696 27. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P,
697 Weiss R, Dubourg V: **Scikit-learn: Machine learning in Python.** Journal of Machine Learning Research
698 2011, **12(Oct):2825-2830.**
- 699 28. Walt Svd, Colbert SC, Varoquaux G: **The NumPy array: a structure for efficient numerical**
700 **computation.** Computing in Science & Engineering 2011, **13(2):22-30.**
- 701 29. Bostock M: **D3.js.** Data Driven Documents 2012, **492.**
- 702 30. Yeung KY, Ruzzo WL: **Principal component analysis for clustering gene expression data.**
703 *Bioinformatics* 2001, **17(9):763-774.**
- 704 31. Yu B: **Stability.** Bernoulli 2013, **19(4):1484-1500.**

- 705 32. Liu X, Long D, You H, Yang D, Zhou S, Zhang S, Li M, He M, Xiong M, Wang X:
706 **Phosphatidylcholine affects the secretion of the alkaline phosphatase PhoA in *Pseudomonas***
707 **strains.** Microbiol Res 2016, **192**:21-29.
- 708 33. Ball G, Durand R, Lazdunski A, Filloux A: **A novel type II secretion system in *Pseudomonas***
709 ***aeruginosa*.** Mol Microbiol 2002, **43**(2):475-485.
- 710 34. Ball G, Viarre V, Garvis S, Voulhoux R, Filloux A: **Type II-dependent secretion of a *Pseudomonas***
711 ***aeruginosa* DING protein.** Res Microbiol 2012, **163**(6):457-469.
- 712 35. Stephens DL, Choe MD, Earhart CF: ***Escherichia coli* periplasmic protein FepB binds**
713 **ferrienterobactin.** Microbiology 1995, **141**(7):1647-1654.
- 714 36. Ellison ML, III JMF, Parrish W, Danell AS, Pesci EC: **The transcriptional regulator Np20 is the zinc**
715 **uptake regulator in *Pseudomonas aeruginosa*.** PloS one 2013, **8**(9):e75389.
- 716 37. Winsor GL, Lam DK, Fleming L, Lo R, Whiteside MD, Nancy YY, Hancock RE, Brinkman FS:
717 ***Pseudomonas* Genome Database: improved comparative analysis and population genomics**
718 **capability for *Pseudomonas* genomes.** Nucleic Acids Res 2010, :gkq869.
- 719 38. Rico-Jimnez M, Reyes-Darias JA, Ortega I, Pea AID, Morel B, Krell T: **Two different mechanisms**
720 **mediate chemotaxis to inorganic phosphate in *Pseudomonas aeruginosa*.** Scientific Reports 2016,
721 **6.**
- 722 39. Folsom JP, Richards L, Pitts B, Roe F, Ehrlich GD, Parker A, Mazurie A, Stewart PS: **Physiology of**
723 ***Pseudomonas aeruginosa* in biofilms as revealed by transcriptome analysis.** BMC microbiology
724 2010, **10**(1):294.
- 725 40. Bailey P, Chang DK, Nones K, Johns AL, Patch A, Gingras M, Miller DK, Christ AN, Bruxner TJ,
726 Quinn MC: **Genomic analyses identify molecular subtypes of pancreatic cancer.** Nature 2016, .
- 727 41. Cancer Genome Atlas Research Network: **Integrated genomic analyses of ovarian carcinoma.**
728 Nature 2011, **474**(7353):609-615.
- 729 42. Yuan Y, Van Allen EM, Omberg L, Wagle N, Amin-Mansour A, Sokolov A, Byers LA, Xu Y, Hess KR,
730 Diao L: **Assessing the clinical utility of cancer genomic and proteomic data across tumor types.**
731 Nat Biotechnol 2014, **32**(7):644-652.
- 732 43. Hanahan D, Weinberg RA: **Hallmarks of cancer: the next generation.** Cell 2011, **144**(5):646-674.
- 733 44. Moldovan G, D'Andrea AD: **How the fanconi anemia pathway guards the genome.** Annu Rev
734 Genet 2009, **43**:223-249.
- 735 45. Sadasivam S, DeCaprio JA: **The DREAM complex: master coordinator of cell cycle-dependent**
736 **gene expression.** Nature Reviews Cancer 2013, **13**(8):585-595.
- 737 46. Tomida J, Itaya A, Shigechi T, Unno J, Uchida E, Ikura M, Masuda Y, Matsuda S, Adachi J,
738 Kobayashi M: **A novel interplay between the Fanconi anemia core complex and ATR-ATRIP kinase**
739 **during DNA cross-link repair.** Nucleic Acids Res 2013, **41**(14):6930-6941.
- 740 47. Itasaki N, Hoppler S: **Crosstalk between Wnt and bone morphogenic protein signaling: a**
741 **turbulent relationship.** Developmental Dynamics 2010, **239**(1):16-33.
- 742 48. MacDonald BT, Tamai K, He X: **Wnt/ β -catenin signaling: components, mechanisms, and**
743 **diseases.** Developmental cell 2009, **17**(1):9-26.
- 744 49. Aman A, Piotrowski T: **Wnt/ β -catenin and Fgf signaling control collective cell migration by**
745 **restricting chemokine receptor expression.** Developmental cell 2008, **15**(5):749-761.
- 746 50. Petersen M, Pardali E, Van Der Horst G, Cheung H, Van Den Hoogen C, Van Der Pluijm G, Ten Dijke
747 P: **Smad2 and Smad3 have opposing roles in breast cancer bone metastasis by differentially**
748 **affecting tumor angiogenesis.** Oncogene 2010, **29**(9):1351-1361.
- 749 51. Yoon H, Dehart JP, Murphy JM, Lim SS: **Understanding the roles of FAK in cancer: inhibitors,**
750 **genetic models, and new insights.** Journal of Histochemistry & Cytochemistry 2015, **63**(2):114-128.
- 751 52. Chinnam M, Goodrich DW: **RB1, development, and cancer.** Curr Top Dev Biol 2011, **94**:129.
- 752 53. Takuwa Y, Du W, Qi X, Okamoto Y, Takuwa N, Yoshioka K: **Roles of sphingosine-1-phosphate**
753 **signaling in angiogenesis.** World journal of biological chemistry 2010, **1**(10):298-306.
- 754 54. Watson C, Long JS, Orange C, Tannahill CL, Mallon E, McGlynn LM, Pyne S, Pyne NJ, Edwards J:
755 **High expression of sphingosine 1-phosphate receptors, S1P 1 and S1P 3, sphingosine kinase 1,**
756 **and extracellular signal-regulated kinase-1/2 is associated with development of tamoxifen**
757 **resistance in estrogen receptor-positive breast cancer patients.** The American journal of pathology
758 2010, **177**(5):2205-2215.
- 759 55. Paik JH, Chae S, Lee M, Thangada S, Hla T: **Sphingosine 1-phosphate-induced endothelial cell**
760 **migration requires the expression of EDG-1 and EDG-3 receptors and Rho-dependent activation of**

- 761 **$\alpha\beta 3$ -and $\beta 1$ -containing integrins.** J Biol Chem 2001, **276**(15):11830-11837.
- 762 56. Serini G, Valdembri D, Bussolino F: **Integrins and angiogenesis: a sticky business.** Exp Cell Res
- 763 2006, **312**(5):651-658.
- 764 57. Brogdon JL, Leitenberg D, Bottomly K: **The potency of TCR signaling differentially regulates**
- 765 **NFATc/p activity and early IL-4 transcription in naive CD4 T cells.** The Journal of Immunology 2002,
- 766 **168**(8):3825-3832.
- 767 58. Hsieh C, Macatonia SE, Tripp CS, Wolf SF, O'Garra A, Murphy KM: **Development of TH1 CD4 T**
- 768 **Cells Through IL-12.** Science 1993, **260**:547.
- 769 59. Vacaflores A, Chapman NM, Harty JT, Richer MJ, Houtman JC: **Exposure of Human CD4 T Cells to**
- 770 **IL-12 Results in Enhanced TCR-Induced Cytokine Production, Altered TCR Signaling, and**
- 771 **Increased Oxidative Metabolism.** PloS one 2016, **11**(6):e0157175.
- 772 60. Lau E, Ze'ev AR: **ATF2—at the crossroad of nuclear and cytosolic functions.** J Cell Sci 2012,
- 773 **125**(12):2815-2824.
- 774 61. Shaulian E, Karin M: **AP-1 as a regulator of cell life and death.** Nat Cell Biol 2002, **4**(5):E136.
- 775 62. Miller MR, Rao A: **NFAT, immunity and cancer: a transcription factor comes of age.** Nature
- 776 Reviews Immunology 2010, **10**(9):645-656.