Unsupervised extraction of stable expression signatures from public compendia with eADAGE

Jie Tan[1,¶], Georgia Doing[2,¶], Kimberley A. Lewis[2], Courtney E. Price[2], Kathleen M. Chen[3], Kyle C. Cady[4,5], Barret Perchuk[4,5], Michael T. Laub[4,5], Deborah A. Hogan[2], Casey S. Greene[3,6,7,*]

1. Department of Molecular and Systems Biology, Geisel School of Medicine at Dartmouth, Hanover, NH, USA 03755
2. Department of Microbiology and Immunology, Geisel School of Medicine at Dartmouth, Hanover, NH, USA 03755
3. Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Philadelphia, PA, USA 19104
4. Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA.
5. Howard Hughes Medical Institute, Cambridge, MA, 02139, USA.
6. Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA 19104
7. Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA 19104

*To whom correspondence should be addressed: csgreene@mail.med.upenn.edu
¶ These authors contributed equally to this work.

Running title: Extracting signatures from public data

Abstract

Cross experiment comparisons in public data compendia are challenged by unmatched conditions and technical noise. The ADAGE method, which performs unsupervised integration with neural networks, can effectively identify biological patterns, but because ADAGE models, like many neural networks, are over-parameterized, different ADAGE models perform equally well. To enhance model robustness and better build signatures consistent with biological pathways, we developed an ensemble ADAGE (eADAGE) that integrated stable signatures across models. We applied eADAGE to a *Pseudomonas aeruginosa* compendium containing experiments performed in 78 media. eADAGE revealed a phosphate starvation response controlled by PhoB. While we expected PhoB activity in limiting phosphate conditions, our analyses found PhoB activity in other media with moderate phosphate and predicted that a second stimulus provided by the sensor kinase, KinB, is required for PhoB activation in this setting. We validated this relationship using both targeted and unbiased genetic approaches. eADAGE, which captures stable biological patterns, enables cross-experiment comparisons that can highlight measured but undiscovered relationships.

Keywords

denoising autoencoders/ensemble modeling/gene expression/*Pseudomonas aeruginosa*/PhoB crosstalk

Introduction

Available gene expression data are outstripping our knowledge about the organisms that we're measuring. Ideally each organism's data reveals the principles underlying gene regulation and consequent pathway activity changes in every condition in which gene expression is measured. Extracting this information requires new algorithms, but many commonly used algorithms are supervised. These algorithms require curated pathway knowledge to work effectively, and in many species such resources are biased in various ways (Gillis and Pavlidis, 2013; Greene and Troyanskaya, 2012; Schnoes et al., 2013). Annotation transfer can help, but such function assignments remain challenging for many biological processes (Jiang et al., 2016). An unsupervised method that doesn't rely on annotation transfer would bypass the challenges of both annotation transfer and biased knowledge.

Along with our wealth of data, abundant computational resources can now power deep unsupervised applications of neural networks, which are powerful methods for unsupervised feature learning (Bengio et al., 2013). In a neural network, input variables are provided to one or more layers of "neurons". Each neuron (also called node) has an activation function that determines whether or not it turns on given some input. The entire network is trained, which consists of adjusting the edge weights that each node provides to each other, by grading the quality of the output for some task. Denoising autoencoders (DAs), a type of unsupervised neural networks, are trained to remove noise that is intentionally added to the input data (Vincent et al., 2008). Masking noise, in which a fraction of the inputs are set to zero, is commonly used (Vincent et al., 2010) and successful denoising autoencoders must learn

dependency structure between the input variables. Adding artificial noise helps a DA to learn features that are robust to partial corruption of input data. This approach has properties that make it particularly suitable for gene expression data (Tan et al., 2015). First, the sigmoid activation function produces features that tend to be on or off, which helps to describe biological processes, e.g. transcription factor activation, with threshold effects. Second, the algorithm is robust to noise. We previously observed that a one-layer DA-based method, ADAGE (analysis using denoising autoencoders of gene expression), was more robust than linear approaches such as ICA or PCA in the context of public data, which employ heterogeneous experimental designs, lack shared controls, and provide limited metadata (Tan et al., 2016b).

Neural networks have many edge weights that must be fit during training. Given some gene expression dataset, there are many different DAs that could reconstruct the data equally well. In a technical sense we would say that the objective functions of neural networks are typically non-convex and trained through stochastic gradient descent. When we train multiple models, each represents a local minimum. Yu recently emphasized the importance of patterns that are stable across statistical models in the process of discovery (Yu, 2013). While run-to-run variability obscures some biological features within individual models, stable patterns across neural networks may clearly resolve biological pathways. To directly target stability, we introduce an unsupervised modeling procedure inspired by consensus clustering (Monti et al., 2003). Consensus clustering has become a standard part of clustering applications for biological datasets. Our approach builds an ensemble neural network that captures stable features and improves model robustness.

To apply the neural network approach to compendium-wide analyses, we first sought to create a comprehensive model in which biological pathways were successfully learned from gene expression data. We adapted ADAGE (Tan et al., 2016b) to capture pathways more specifically by increasing the number of nodes (model size) that reflect potential pathways from 50 to 300, a size that our analyses indicate the current public data compendium can support. We then built its ensemble version (eADAGE) and compared it with ADAGE, PCA, and ICA. While it is impossible to specify *a priori* the number of true biological pathways that exhibit gene expression signatures, we observed that eADAGE models produced gene expression signatures that corresponded to more biological pathways. This indicates that this method more effectively identifies biological signatures from noisy public data. While ADAGE models reveal biological features perturbed within an experiment, the more robust eADAGE models also enable analyses that cut across an organism's gene expression compendium.

To assess the utility of the eADAGE model in making predictions of biological activity, we applied it to the analysis of the *Pseudomonas aeruginosa* gene expression compendium which included 1051 samples grown in 78 distinct medium conditions, 128 distinct strains and isolates, and dozens of different environmental parameters. After grouping samples by medium type, we searched for eADAGE-defined signatures that differed between medium types. This cross-compendium analysis identified five media that elicited a response to low-phosphate mediated by the transcriptional regulator PhoB, and only one of these five media was specifically defined

as a condition with low phosphate. While PhoB is known to respond to low phosphate through its interaction with PhoR in low concentrations (Wanner and Chang, 1987), our analyses indicated that PhoB is also active at moderate phosphate concentrations. Specifically, in media with moderate phosphate concentrations, the eADAGE model predicted a previously undiscovered role for KinB in the activation of PhoB, and our molecular analyses of *P. aeruginosa* confirmed this prediction. Analysis of a collection of *P. aeruginosa* mutants defective in kinases validated the specificity of the KinB-PhoB relationship.

In summary, eADAGE more precisely and robustly captures biological processes and pathways from gene expression data than other unsupervised approaches. The signatures learned by eADAGE support functional gene set analyses without manual pathway annotation. The signatures are robust enough to enable biologists to identify not only differentially active signatures within one experiment, but also cross-compendium patterns that reveal undiscovered regulatory mechanisms captured within existing public data.

Results

eADAGE: ensemble modeling improves the model breadth, depth, and robustness

ADAGE is a neural network model. Each gene is connected to each node through a weighted edge (Figure 1A). We define a gene signature learned by an ADAGE model as a set of genes that contribute the highest positive or highest negative weights to a specific node (Figure 1B, see methods for detail). Therefore, one node results in two gene signatures, one on each high weight side. The positive and negative signatures derived from the same node do not necessarily compose inversely regulated processes (Figure S1), so we use them independently.

ADAGE models of the same size capture different pathways. This occurs because each ADAGE model is initialized with random weights, and the training processes are sensitive to initial conditions. eADAGE, in which we built an ensemble version of individual ADAGE models, took advantage of this variation to enhance model robustness. Each eADAGE model integrated nodes from 100 individual ADAGE models (Figure 2A). To unite nodes, we applied consensus clustering on nodes' weight vectors because the weight vector captures both the genes that contribute to a node and their magnitude. Our previous ADAGE analyses showed that genes contributing high weights characterized each node's biological significance, so we designed a weighted Pearson correlation to incorporate gene weights in building eADAGE models (see methods). We compared eADAGE to two baseline methods: individual ADAGE models and corADAGE, which combined nodes with an unweighted Pearson correlation. For direct comparison, the model sizes of ADAGE, eADAGE, and corADAGE were all fixed to 300 nodes, which we found to be appropriate for the current *P. aeruginosa* expression compendium through both data-driven and knowledge-driven heuristics (see supplemental information).

While ADAGE models are constructed without the use of any curated information such as KEGG (Kanehisa and Goto, 2000) and GO (Ashburner et al., 2000), we evaluate models by the extent to which they cover the pathways and processes defined in these resources to see how they capture existing biology. For each method, we determined the number of KEGG pathways

155 significantly associated with at least one gene signature in a model, referred to as KEGG
156 coverage. eADAGE models exhibited greater KEGG coverage than those generated by other
157 methods (Figure 2B). Both corADAGE and eADAGE covered significantly more KEGG pathways
158 than ADAGE (t-test p-value of 1.04e-6 between corADAGE (n=10) and ADAGE (n=1000) and t-
159 test p-value of 1.41e-6 between eADAGE (n=10) and ADAGE (n=1000)). Moreover, eADAGE
160 models covered, on average, 10 more pathways than corADAGE (t-test p-value of 1.99e-3, n=10
161 for both groups). Genes that participate in multiple pathways can influence pathway
162 enrichment analysis, a factor termed pathway crosstalk (Donato et al., 2013). To control for this,
163 we performed crosstalk correction (Donato et al., 2013). After correction, the number of
164 covered pathways dropped by approximately half (Figure S2A), but eADAGE still covered
165 significantly more pathways than corADAGE (t-test p-value of 0.02) and ADAGE (t-test p-value
166 of 1.29e-05). We subsequently evaluated each method's coverage of GO biological processes
167 (GO-BP) and found consistent results (Figure S2B). eADAGE integrated multiple models to more
168 broadly capture pathway signals embedded in diverse gene expression compendia.
169
170 We next evaluated how specifically and completely signatures learned by the models capture
171 known biology. We use each gene signature's FDR corrected p-value for enrichment of a
172 KEGG/GO term as a combined measure, as it captures both the sensitivity and specificity. If a
173 pathway was significantly associated with multiple gene signatures in a model, we only
174 considered its most significant association. We found that 71% of KEGG and 79% of GO-BP
175 terms were more significantly enriched (had lower median p-values) in corADAGE models when
176 compared to individual ADAGE models. This increased to 87% for KEGG and 81% for GO-BP
177 terms in eADAGE models. We also directly compared eADAGE and corADAGE by this measure
178 and observed that 74% of KEGG and 61% of GO-BP terms were more significantly enriched in
179 eADAGE. We have found that different pathways were best captured at different model sizes
180 (Figure 2C). We next compared the 300-node eADAGE model to ADAGE models with different
181 number of nodes. Although the 300-node eADAGE models were constructed only from 300-
182 node ADAGE models, we found that 69% of KEGG and 69% of GO-BP terms were more
183 significantly enriched (i.e. lower median p-values) in eADAGE models than ADAGE models of
184 any size, including those with more nodes than the eADAGE models. Three example pathways
185 that are best captured either when model size is small, large, or in the middle are all well
186 captured in the 300-node eADAGE model (Figure 2C). These results demonstrate that eADAGE's
187 ensemble modeling procedure is effective in capturing consistent signals across models and
188 filtering out noise. Thus, eADAGE more completely and precisely captures the gene expression
189 signatures of biological pathways.
190
191 We designed eADAGE to provide a more robust analysis framework than individual ADAGE
192 models. To assess this, we examined the percentage of models that covered each pathway
193 (coverage rate) between ADAGE and eADAGE. The pathways covered by each individual ADAGE
194 model were highly variable. Most KEGG pathways were covered by less than half of individual
195 models but more than half of eADAGE models (Figure 2D), suggesting that eADAGE models
196 were more robust than individual ADAGE models. Subsequent evaluations of GO-BP were
197 consistent with this finding (Figure S2C). We excluded KEGG/GO terms always covered by both
198 individual ADAGE and eADAGE models and observed that 69% of the remaining KEGG and 71%

199 of the remaining GO terms were covered more frequently by eADAGE than ADAGE. This
200 suggests that their associations are stabilized via ensemble construction. In summary, these
201 comparisons of eADAGE and ADAGE reveal that not only are more pathways captured more
202 specifically, but also those that are captured are captured more consistently.
203
204 Principal component analysis (PCA) and independent component analysis (ICA) have been
205 previously used to extract biological features and build functional gene sets (Alter et al., 2000;
206 Chen et al., 2008; Engreitz et al., 2010; Frigyesi et al., 2006; Gong et al., 2007; Lutter et al., 2009;
207 Ma and Kosorok, 2009; Raychaudhuri et al., 2000, 2000; Roden et al., 2006). We performed PCA
208 and generated multiple ICA models from the same *P. aeruginosa* expression compendium and
209 evaluated their KEGG/GO term coverage following the same procedures used for eADAGE.
210 eADAGE substantially and significantly outperforms PCA in terms of pathway coverage (Figure
211 2E). Between eADAGE and ICA, we observed that eADAGE represented KEGG/GO terms more
212 precisely than ICA. Specifically, among terms significantly enriched in either approach, 68%
213 KEGG and 71% GO terms exhibited more significant enrichment in eADAGE. Increasing the
214 significance threshold for pathway coverage demonstrates the advantage of eADAGE (Figure 3D
215 and Figure S2D).
216
217 Pathway databases provide a means to compare unsupervised methods for signature discovery.
218 Not all pathways will be regulated at the transcriptional level, but those that are may be
219 extracted from gene expression data. The unsupervised eADAGE method revealed signatures
220 that corresponded to *P. aeruginosa* KEGG/GO terms better than PCA, ICA, ADAGE, and
221 corADAGE. It had higher pathway coverage (breadth), covered pathways more specifically
222 (depth), and more consistently (robustness) than existing methods.
223
224 **Elucidating functional signatures that are indicative of growth medium**
225
226 For biological evaluation, we built a single new eADAGE model with 300 nodes. The model's
227 weight matrix (Table S2) and all gene signatures (Table S3) are provided. For each signature, we
228 calculated its activity in each sample (see Methods, Table S4). A high activity indicates that the
229 majority of genes in the signature are highly expressed in the sample.
230
231 Analysis of differentially expressed genes is widely used to analyze single experiments, but
232 crosscutting signatures are required to reveal general response patterns from large-scale
233 compendia. Signature-based analyses can suggest mechanisms such as crosstalk and novel
234 regulatory networks. However, in order for this to be effective, these signatures must be robust
235 and comprehensive. By capturing biological pathways more completely and robustly, eADAGE
236 enables the analysis of signatures, including those that don't correspond to any KEGG pathway,
237 across the entire compendium of *P. aeruginosa.*
238
239 Gene expression experiments have been used to investigate a diverse set of questions about *P.*
240 *aeruginosa* biology, and these experiments have used many different media to emphasize
241 different phenotypes. Our manual annotation showed that 78 different base media were used
242 across the gene expression compendium (Table S1). While the compendium contains 125

243 different experiments, it is exceedingly rare for investigators to use multiple base lab media
244 within the same experiment. There were only two examples in the entire compendium (Table
245 S1). Other than LB, which is used in 43.6% (458/1051) of the samples in the compendium, most
246 media are only represented by a handful of samples.
247
248 To provide an illustrative example of cross-experiment analysis, we examined signature activity
249 across the six experiments in a base of M9 minimal medium (Miller, 1972), which used six
250 different carbon sources. Node147pos was highly active in phosphatidylcholine compared to all
251 other media (Figure 3A). This node was significantly enriched for the GO terms choline catabolic
252 process (FDR q-value of 2.9E-11) and glycine betaine catabolic process (FDR q-value of 4.6E-20).
253 Of all signatures, it had the largest overlap with the regulon of GbdR, the choline-responsive
254 transcription factor (Hampel et al., 2014) (FDR q-value of 2.5E-47), suggesting that choline
255 catabolism is active in this medium. Consistent with this, phosphatidylcholine, but not
256 palmitate, citrate, or glucose, serves as a source of choline for *P. aeruginosa* (Wargo et al.,
257 2011, 2009). Importantly, while Node147pos was differentially active within a single
258 experiment containing samples in phosphatidylcholine and palmitate (E-GEOD-7704), it was
259 also identifiable in comparisons to samples grown in M9 medium with different carbon sources
260 in experiments performed in different labs at different times. This illustrates how medium-
261 specific signatures can be identified without experiments designed to explicitly test the effect of
262 a specific medium component on gene expression.
263
264 **Distinct aspects of the response to low phosphate are captured among the most active**
265 **signatures**
266 To broadly examine signatures across all media, we calculated a medium activation score for
267 each signature-medium combination. This score reflected how a signature's activity in a
268 medium differed from its activity in all other samples (Figure S3, see methods for details). Table
269 S5 lists signatures with activation scores in a specific medium above a stringent threshold. A
270 signature could be active in multiple media (Figure S3), so we averaged their activation scores
271 when this occurred. Table S6 lists signatures that are most active in a group of media (a
272 complete list of signature-media group associations is in Table S7).
273
274 The two signatures with the highest pan-media activation scores were Node164pos and
275 Node108neg (Table S6). To evaluate the basis for the high activation scores, we examined their
276 underlying activities across all media (Node164pos is shown Figure 3A), and found that both
277 were highly active in King's A medium, Peptone medium, and NGM+<0.1mM phosphate
278 (NGMlowP), but not in NGM+25mM phosphate (NGMhighP). The activity differences between
279 NGMlowP and NGMhighP suggested that these signatures respond to phosphate levels. The
280 other two media (Peptone and King's A) in which Node164pos had high activity also had low
281 phosphate concentrations (0.4 mM) relative to other media. For example, commonly used LB
282 has a phosphate concentration of ~4.5 mM (Bertani, 2004) and many others have
283 concentrations above 20 mM.
284
285 KEGG pathway enrichment analysis of Node164pos genes showed enrichment in phosphate
286 acquisition related pathways (Table S6). One Node164pos gene encodes PhoB, a transcription

287 factor in the PhoR-PhoB two-component system that responds to low environmental
288 phosphate in *P. aeruginosa* (Bielecki et al., 2015; Blus-Kadosh et al., 2013; Santos-Beneit, 2015).
289 Further, Node164pos is the signature most enriched for a previously defined PhoB regulon (FDR
290 q-value of 8.1e-29 in hypergeometric test).
291
292 Expression levels of genes in Node164pos are higher in Peptone, King's A, and NGMlowP than
293 in NGMhighP (Figure 3B), including *phoA* which encodes alkaline phosphatase, an enzyme
294 whose activity can be monitored using a colorimetric assay. As expected, PhoA was activated
295 when phosphate concentrations were low (Figure 4A). Furthermore, PhoA activity was
296 dependent on PhoB and the PhoB-activating histidine kinase PhoR, consistent with published
297 work (Bielecki et al., 2015). Notably, PhoA activity was evident on King's A and Peptone (Figure
298 4B). Although King's A and Peptone are not considered to be phosphate-limited media, these
299 results provide striking evidence that they induced PhoB activity as predicted by Node164pos's
300 signature-medium relationship.
301
302 While Node108neg is not significantly associated with phosphate acquisition-related KEGG
303 pathways, it is enriched for the PhoB regulon (FDR q-value of 5.2e-9 in hypergeometric test,
304 Table S6) and shares over half of its thirty-two genes with Node164pos. Six of the seven PhoB-
305 regulated genes present in Node108neg are also regulated by TctD, a transcriptional repressor
306 described by Haussler and colleagues (Bielecki et al., 2015). Therefore, Node108neg primarily
307 represents genes that are both PhoB-activated and TctD-repressed. Subsequent analyses found
308 that Node108neg was the most differentially active signature between a Δ*tctD* strain and the
309 wild type in an RNAseq experiment (E-GEOD-64056). Importantly, eADAGE learned this TctD
310 regulon even though the expression compendium did not contain any samples of *tctD* mutants.
311 This demonstrates the utility of eADAGE in learning regulatory programs uncharacterized by
312 KEGG.
313
314 We evaluated whether the PhoB and TctD signals were also extracted by PCA, ICA, or ADAGE.
315 ICA and ADAGE captured signatures enriched of the PhoB regulon less than those of eADAGE
316 (Table S8). PCA captured a strong PhoB signal in its 19th principal component. However, it did
317 not learn the subtler TctD signal. In summary, the other methods were able to capture some of
318 this signature but in a manner that was less complete or failed to separate TctD.
319
320 **Cross-compendium analysis of Node164pos activity reveals a role for the histidine kinase KinB**
321 **in the regulation of PhoB**
322 Interestingly, Node164pos activity exhibited a wide spread in PIA medium, with six samples
323 having high activities and the other six having low activities (Figure 3A). All of the strains in
324 which Node164pos was low were from a study that used a PAO1 *kinB*::Gm^R mutant background
325 (Damron et al., 2012). The PIA-grown samples with high Node164pos activity used a PAO1
326 strain with *kinB* intact (Damron et al., 2013) leading us to propose that KinB may be a regulator
327 of PhoB on PIA. We confirmed that PhoA activity dependents on PhoB, PhoR, KinB on PIA
328 medium (Figure 4B) as illustrated by the fact that a screen of 63 histidine kinase in-frame
329 deletion mutants (Table S9) found only Δ*phoR* and Δ*kinB* had no PhoA activity on PIA, like the
330 *phoB* mutant. These kinases appear to regulate PhoB non-redundantly and to different extents

331 in PIA, as the Δ*phoR* mutant regained PhoA activity at later time points but the Δ*kinB* mutant
332 did not (Figure 4C).
333
334 Although the phosphate concentration of PIA (0.8mM) is lower than that of rich media such as
335 LB (~4.5mM), it is higher than that of Peptone and King's A (0.4mM). Therefore, we tested
336 whether a moderately low level of phosphate provokes KinB regulation of PhoA. Like in PIA, we
337 found that PhoA activity was evident at concentrations up to 0.5 mM phosphate in MOPS
338 medium in the wild type, but only at lower concentrations in the Δ*kinB* strain suggesting that
339 KinB plays a role at intermediate concentrations (Figure 4D). To our knowledge, KinB has not
340 been previously implicated in the activation of PhoB.
341
342 In summary, eADAGE effectively extracted biologically meaningful features, accurately
343 indicated their activity in multiple media spanning numerous independent experiments, and
344 revealed a novel regulatory mechanism. By summarizing gene-based expression information
345 into biologically relevant signatures, eADAGE greatly simplifies analyses that cut across large
346 gene expression compendia.
347

348 ## Discussion

349 Our eADAGE algorithm combines multiple ADAGE models into one ensemble model to address
350 model variability due to stochasticity and local minima. The algorithm is inspired by consensus
351 clustering, which reconciles the differences in cluster assignments in multiple runs. Comparable
352 approaches have also been applied for ICA, where researchers have used the centrotypes in
353 clustering multiple models as the final model (Himberg et al., 2004). The ICA centrotype
354 approach for ADAGE corresponds to corADAGE, and our comparison of eADAGE and corADAGE
355 shows that eADAGE not only covers more biological pathways, but also results in cleaner
356 representations of biological pathways. This direct comparison suggests that placing particular
357 emphasis on the genes most associated with a particular feature may be a useful property for
358 other unsupervised feature construction algorithms. While our results demonstrate that this
359 ensemble process can help improve the biological interpretability of neural networks, we do
360 not expect it to increase prediction accuracies in supervised learning problems.
361
362 eADAGE revealed patterns that were detectable from a large data compendium containing
363 experiments performed in 78 different media but that were not necessarily evident in individual
364 experiments. For example, one eADAGE signature revealed media in which *P. aeruginosa* had
365 high PhoB activity. PhoB is a global regulator, and understanding its state in different media can
366 provide important insight into medium-specific phenotypes. King's A and PIA, on which the
367 PhoB signature was active, are known to stimulate robust production of colorful secondary
368 metabolites (King et al., 1954) called phenazines. Separate studies have shown that PhoB can
369 influence phenazine levels (Jensen et al., 2006). Future studies will reveal whether or not the
370 low phosphate levels in these media contribute to this characteristic phenotype. We expect
371 that other signatures extracted from the compendium by eADAGE will serve as the basis for
372 additional work in which the patterns are not only examined but also validated.
373

We also uncovered a subtle aspect of the phosphate starvation response that depends on KinB, a histidine kinase not previously associated with PhoB. Bacterial two-component systems are often insulated from each other (Podgornaia and Laub, 2013). Though sensor kinase/response regulator cross-talk has been hypothesized as a mechanism of explaining the complexity of signaling networks (Fisher et al., 1995; Ninfa et al., 1988), it is challenging to find conditions where two kinases are needed for full response regulator activation (Verhamme et al., 2002). We propose that moderate levels of phosphate, like those in PIA, provide a niche for crosstalk: the activity of PhoR is low enough that the interaction with KinB is needed for full PhoB activity on this medium. Together, PhoR and KinB may enable a more sensitive and effective response to phosphate limitation. Alternatively, KinB may influence PhoB activity indirectly by regulating activities that affect PhoB levels, phosphorylation state, or protein-protein interactions. This relationship was not observed in experiments designed to perturb this process, which use high and very low phosphate concentrations. Instead, eADAGE analysis of *Pseudomonas aeruginosa* transcriptomic measurements across multiple experiments in different media were required to reveal this nuanced mechanism.

Existing public gene expression data compendia for more than one hundred organisms are of sufficient size to support eADAGE models (Greene et al., 2016). Cross-compendium analyses provide the opportunity to efficiently use existing data to identify regulatory patterns that are evident across multiple experiments, datasets, and labs. To tap this potential, we will require algorithms like eADAGE that robustly integrate these diverse datasets in a manner that is not tied to only aspects of biology that are well understood. Furthermore, while public compendia tend to be dominated by expression data, autoencoders have also been successfully applied to datasets based on large collections of electronic health record (Beaulieu-Jones et al., 2016; Miotto et al., 2016). Within the health records space, these methods are particularly effective at dealing with missing data (Beaulieu-Jones et al., 2016; Beaulieu-Jones and Moore, 2017). These features, along with their unsupervised nature, make DAs a promising approach for the integration of heterogeneous data types. We find that ensembles of DAs construct clearer features that more robustly capture biological processes. Ultimately, we expect unsupervised algorithms to be most helpful when they lead users to discover new underlying mechanisms, which require models that are accurate, robust, and interpretable.

Acknowledgements
This work was supported in part by a grant from the Gordon and Betty Moore Foundation (GBMF 4552) to CSG. This work was supported by National Institutes of Health (NIH) grant RO1-AI091702 to DAH. MTL is an investigator of the Howard Hughes Medical Institute. This work was supported by a pilot grant from the Cystic Fibrosis Foundation (STANTO15R0) to CSG and DAH. The authors would like to thank Gregory Way and René Zelaya for helpful code review. The authors also would like to thank Anastasia Baryshnikova for providing critical feedback on a preprint of this work.

Author contributions
JT, DAH and CSG conceived and designed the research. JT, GD and KMC performed computational analyses. GD, KAL and CEP performed molecular experiments. KC, BP and MTL

418 constructed and contributed the histidine kinase knock out collection. JT, GD, KMC, DAH and
419 CSG wrote the manuscript, and KAL, CEP, KMC, KD, BP and MTL provided critical feedback.
420
421 **Conflict of interest**
422 The authors have no conflicts of interest to report.
423

424 Reference

425 Alter, O., Brown, P.O., and Botstein, D. (2000). Singular value decomposition for genome-wide
426 expression data processing and modeling. Proc. Natl. Acad. Sci. U. S. A. 97, 10101–6.
427 Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K.,
428 Dwight, S.S., Eppig, J.T., et al. (2000). Gene Ontology: tool for the unification of biology.
429 Nat. Genet. 25, 25–29.
430 Beaulieu-Jones, B.K., Greene, C.S., and Pooled Resource Open-Access ALS Clinical Trials
431 Consortium (2016). Semi-supervised learning of the electronic health record for phenotype
432 stratification. J. Biomed. Inform. 64, 168–178.
433 Beaulieu-Jones, B.K., and Moore, J.H. (2017). MISSING DATA IMPUTATION IN THE ELECTRONIC
434 HEALTH RECORD USING DEEPLY LEARNED AUTOENCODERS, in: Biocomputing 2017.
435 WORLD SCIENTIFIC, pp. 207–218.
436 Bengio, Y., Courville, A., and Vincent, P. (2013). Representation Learning: A Review and New
437 Perspectives. IEEE Trans. Pattern Anal. Mach. Intell. 35, 1798–1828.
438 Bertani, G. (2004). Lysogeny at mid-twentieth century: P1, P2, and other experimental systems.
439 J. Bacteriol. 186, 595–600.
440 Bielecki, P., Jensen, V., Schulze, W., Gödeke, J., Strehmel, J., Eckweiler, D., Nicolai, T., Bielecka,
441 A., Wille, T., Gerlach, R.G., et al. (2015). Cross talk between the response regulators PhoB
442 and TctD allows for the integration of diverse environmental signals in *Pseudomonas*
443 *aeruginosa*. Nucleic Acids Res. 43, 6413–25.
444 Blus-Kadosh, I., Zilka, A., Yerushalmi, G., and Banin, E. (2013). The effect of *pstS* and *phoB* on
445 quorum sensing and swarming motility in *Pseudomonas aeruginosa*. PLoS One 8, e74444.
446 Chen, L., Xuan, J., Wang, C., Shih, I.-M., Wang, Y., Zhang, Z., Hoffman, E., Clarke, R., Devore, J.,
447 Peck, R., et al. (2008). Knowledge-guided multi-scale independent component analysis for
448 biomarker identification. BMC Bioinformatics 9, 416.
449 Damron, F.H., Barbier, M., McKenney, E.S., Schurr, M.J., and Goldberg, J.B. (2013). Genes
450 required for and effects of alginate overproduction induced by growth of *Pseudomonas*
451 *aeruginosa* on Pseudomonas isolation agar supplemented with ammonium metavanadate.
452 J. Bacteriol. 195, 4020–36.
453 Damron, F.H., Owings, J.P., Okkotsu, Y., Varga, J.J., Schurr, J.R., Goldberg, J.B., Schurr, M.J., and
454 Yu, H.D. (2012). Analysis of the *Pseudomonas aeruginosa* regulon controlled by the sensor
455 kinase KinB and sigma factor RpoN. J. Bacteriol. 194, 1317–30.
456 Donato, M., Xu, Z., Tomoiaga, A., Granneman, J.G., Mackenzie, R.G., Bao, R., Than, N.G.,
457 Westfall, P.H., Romero, R., and Draghici, S. (2013). Analysis and correction of crosstalk
458 effects in pathway analysis. Genome Res. 23, 1885–93.
459 Edgar, R. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data
460 repository. Nucleic Acids Res. 30, 207–210.
461 Engreitz, J.M., Daigle, B.J., Marshall, J.J., and Altman, R.B. (2010). Independent component

462 analysis: mining microarray data for fundamental human gene expression modules. J.
463 Biomed. Inform. 43, 932–44.
464 Fisher, S.L., Jiang, W., Wanner, B.L., and Walsh, C.T. (1995). Cross-talk between the histidine
465 protein kinase VanS and the response regulator PhoB. Characterization and identification
466 of a VanS domain that inhibits activation of PhoB. J. Biol. Chem. 270, 23143–9.
467 Frigyesi, A., Veerla, S., Lindgren, D., Höglund, M., Quackenbush, J., Jutten, C., Herault, J.,
468 Chiappetta, P., Roubaud, M., Torrésani, B., et al. (2006). Independent component analysis
469 reveals new and biologically significant structures in microarray data. BMC Bioinformatics
470 7, 290.
471 Gillis, J., and Pavlidis, P. (2013). Assessing identity, redundancy and confounds in Gene Ontology
472 annotations over time. Bioinformatics 29, 476–82.
473 Gong, T., Xuan, J., Wang, C., Li, H., Hoffman, E., Clarke, R., and Wang, Y. (2007). Gene module
474 identification from microarray data using nonnegative independent component analysis.
475 Gene Regul. Syst. Bio. 1, 349–63.
476 Greene, C.S., Foster, J.A., Stanton, B.A., Hogan, D.A., and Bromberg, Y. (2016). Computational
477 Approaches to Study Microbes and Microbiomes. Pac Sym Biocomput 557–567.
478 Greene, C.S., and Troyanskaya, O.G. (2012). Accurate evaluation and analysis of functional
479 genomics data and methods. Ann. N. Y. Acad. Sci. 1260, 95–100.
480 Ha, D.-G., Richman, M.E., and O'Toole, G.A. (2014). Deletion mutant library for investigation of
481 functional outputs of cyclic diguanylate metabolism in *Pseudomonas aeruginosa* PA14.
482 Appl. Environ. Microbiol. 80, 3384–93.
483 Hampel, K.J., LaBauve, A.E., Meadows, J.A., Fitzsimmons, L.F., Nock, A.M., and Wargo, M.J.
484 (2014). Characterization of the GbdR regulon in Pseudomonas aeruginosa. J. Bacteriol. 196,
485 7–15.
486 Himberg, J., Hyvärinen, A., and Esposito, F. (2004). Validating the independent components of
487 neuroimaging time series via clustering and visualization. Neuroimage 22, 1214–1222.
488 Jensen, V., Lons, D., Zaoui, C., Bredenbruch, F., Meissner, A., Dieterich, G., Munch, R., and
489 Haussler, S. (2006). RhlR Expression in Pseudomonas aeruginosa Is Modulated by the
490 Pseudomonas Quinolone Signal via PhoB-Dependent and -Independent Pathways. J.
491 Bacteriol. 188, 8601–8606.
492 Jiang, Y., Oron, T.R., Clark, W.T., Bankapur, A.R., D'Andrea, D., Lepore, R., Funk, C.S., Kahanda, I.,
493 Verspoor, K.M., Ben-Hur, A., et al. (2016). An expanded evaluation of protein function
494 prediction methods shows an improvement in accuracy. Genome Biol. 17, 184.
495 Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic
496 Acids Res. 28, 27–30.
497 King, E.O., Ward, M.K., and Raney, D.E. (1954). Two simple media for the demonstration of
498 pyocyanin and fluorescin. J. Lab. Clin. Med. 44, 301–7.
499 Lundgren, B.R., Thornton, W., Dornan, M.H., Villegas-Peñaranda, L.R., Boddy, C.N., and Nomura,
500 C.T. (2013). Gene PA2449 is essential for glycine metabolism and pyocyanin biosynthesis in
501 *Pseudomonas aeruginosa* PAO1. J. Bacteriol. 195, 2087–100.
502 Lutter, D., Langmann, T., Ugocsai, P., Moehle, C., Seibold, E., Splettstoesser, W.D., Gruber, P.,
503 Lang, E.W., and Schmitz, G. (2009). Analyzing time-dependent microarray data using
504 independent component analysis derived expression modes from human macrophages
505 infected with *F. tularensis holartica*. J. Biomed. Inform. 42, 605–611.

506 Ma, S., and Kosorok, M.R. (2009). Identification of differential gene pathways with principal
507 component analysis. Bioinformatics 25, 882–9.
508 Miller, J.H. (1972). Experiments in molecular genetics. Cold Spring Harbor Laboratory.
509 Miotto, R., Li, L., Kidd, B.A., and Dudley, J.T. (2016). Deep Patient: An Unsupervised
510 Representation to Predict the Future of Patients from the Electronic Health Records. Sci.
511 Rep. 6, 26094.
512 Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus clustering: a resampling-
513 based method for class discovery and visualization of gene expression microarray data.
514 Mach. Learn. 52, 91–118.
515 Neidhardt, F.C., Bloch, P.L., and Smith, D.F. (1974). Culture medium for enterobacteria. J.
516 Bacteriol. 119, 736–47.
517 Ninfa, A.J., Ninfa, E.G., Lupas, A.N., Stock, A., Magasanik, B., and Stock, J. (1988). Crosstalk
518 between bacterial chemotaxis signal transduction proteins and regulators of transcription
519 of the Ntr regulon: evidence that nitrogen assimilation and chemotaxis are controlled by a
520 common phosphotransfer mechanism. Proc. Natl. Acad. Sci. U. S. A. 85, 5492–6.
521 Park, H.-S., and Jun, C.-H. (2009). A simple and fast algorithm for K-medoids clustering. Expert
522 Syst. Appl. 36, 3336–3341.
523 Piotrowski, M., Forster, T., Dobrezelecki, B., Sloan, T.M., Mitchell, L., Ghazal, P., Mewsissen, M.,
524 Petrou, S., Trew, A., and Hill, J. (2011). Optimisation and parallelisation of the partitioning
525 around medoids function in R, in: 2011 International Conference on High Performance
526 Computing & Simulation. IEEE, pp. 707–713.
527 Podgornaia, A.I., and Laub, M.T. (2013). Determinants of specificity in two-component signal
528 transduction. Curr. Opin. Microbiol. 16, 156–62.
529 Raychaudhuri, S., Stuart, J.M., and Altman, R.B. (2000). Principal components analysis to
530 summarize microarray experiments: application to sporulation time series. Pac. Symp.
531 Biocomput. 455–66.
532 Roden, J.C., King, B.W., Trout, D., Mortazavi, A., Wold, B.J., Hart, C.E., Tavazoie, S., Hughes, J.,
533 Campbell, M., Cho, R., et al. (2006). Mining gene expression data by interpreting principal
534 components. BMC Bioinformatics 7, 194.
535 Rustici, G., Kolesnikov, N., Brandizi, M., Burdett, T., Dylag, M., Emam, I., Farne, A., Hastings, E.,
536 Ison, J., Keays, M., et al. (2013). ArrayExpress update--trends in database growth and links
537 to data analysis tools. Nucleic Acids Res. 41, D987-90.
538 Santos-Beneit, F. (2015). The Pho regulon: a huge regulatory network in bacteria. Front.
539 Microbiol. 6, 402.
540 Schnoes, A.M., Ream, D.C., Thorman, A.W., Babbitt, P.C., and Friedberg, I. (2013). Biases in the
541 experimental annotations of protein function and their effect on our understanding of
542 protein function space. PLoS Comput. Biol. 9, e1003063.
543 Tan, J., Doing, G., Lewis, K.A., Price, C.E., Chen, K.M., Cady, K.C., Perchuk, B., Laub, M.T., Hogan,
544 D.A., and Greene, C.S. (2016a). eADAGE-1.0.0rc2. Zenodo.
545 Tan, J., Hammond, J.H., Hogan, D.A., and Greene, C.S. (2016b). ADAGE-Based Integration of
546 Publicly Available *Pseudomonas aeruginosa* Gene Expression Data with Denoising
547 Autoencoders Illuminates Microbe-Host Interactions. mSystems 1, e00025-15.
548 Tan, J., Ung, M., Cheng, C., and Greene, C.S. (2015). Unsupervised feature construction and
549 knowledge extraction from genome-wide assays of breast cancer with denoising

550 autoencoders. Pac. Symp. Biocomput. 20, 132–43.
551 Thompson, J.A., Tan, J., and Greene, C.S. (2016). Cross-platform normalization of microarray
552 and RNA-seq data for machine learning applications. PeerJ 4, e1621.
553 Verhamme, D.T., Arents, J.C., Postma, P.W., Crielaard, W., and Hellingwerf, K.J. (2002).
554 Investigation of in vivo cross-talk between key two-component systems of *Escherichia coli*.
555 Microbiology 148, 69–78.
556 Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing
557 robust features with denoising autoencoders, in: Proceedings of the 25th International
558 Conference on Machine Learning - ICML '08. ACM Press, New York, New York, USA, pp.
559 1096–1103.
560 Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. (2010). Stacked denoising
561 autoencoders: Learning useful representations in a deep network with a local denoising
562 criterion. J. Mach. Learn. Res. 11, 3371–3408.
563 Wanner, B.L., and Chang, B.D. (1987). The phoBR operon in Escherichia coli K-12. J. Bacteriol.
564 169, 5569–74.
565 Wargo, M.J., Gross, M.J., Rajamani, S., Allard, J.L., Lundblad, L.K.A., Allen, G.B., Vasil, M.L.,
566 Leclair, L.W., and Hogan, D.A. (2011). Hemolytic phospholipase C inhibition protects lung
567 function during Pseudomonas aeruginosa infection. Am. J. Respir. Crit. Care Med. 184,
568 345–54.
569 Wargo, M.J., Ho, T.C., Gross, M.J., Whittaker, L.A., and Hogan, D.A. (2009). GbdR regulates
570 Pseudomonas aeruginosa plcH and pchP transcription in response to choline catabolites.
571 Infect. Immun. 77, 1103–11.
572 Wilkerson, M.D., and Hayes, D.N. (2010). ConsensusClusterPlus: a class discovery tool with
573 confidence assessments and item tracking. Bioinformatics 26, 1572–3.
574 Yu, B. (2013). Stability. Bernoulli 19, 1484–1500.
575 Zaborin, A., Romanowski, K., Gerdes, S., Holbrook, C., Lepine, F., Long, J., Poroyko, V., Diggle,
576 S.P., Wilke, A., Righetti, K., et al. (2009). Red death in *Caenorhabditis elegans* caused by
577 *Pseudomonas aeruginosa* PAO1. Proc. Natl. Acad. Sci. U. S. A. 106, 6327–32.
578

579 ## Figure Legends

580 Figure 1: ADAGE model and signature definition.
581 A In an ADAGE model, every gene contributes a weight value to every node. The strength
582 of weight values is reflected by gene-node edge. Orange edges indicate high positive weight.
583 Blue edges indicate high negative weight. Dotted edges show low positive or negative weights.
584 B The distribution of a node's weight matrix (Node1 as an example) is roughly normally
585 distributed and centered at zero. Genes with weights higher than the positive high-weight (HW)
586 cutoff (GeneE and GeneA) form the gene signature Node1pos. Similarly, genes with weights
587 lower than the negative HW cutoff (GeneC) form the gene signature Node1neg.
588

589 **Figure 2: The construction and performance of eADAGE.**
590 A eADAGE construction workflow. 100 individual ADAGE models were built using the same
591 input dataset (step 1). Nodes from all models were extracted (step 2) and clustered based on
592 the similarities in their associated weight vectors (step 3). Nodes derived from different models
593 were rearranged by their clustering assignments (step 4). Weight vectors from nodes in the

594 same cluster were averaged and thus becoming the final weight vector of a newly constructed
595 node in an eADAGE model (step5).
596 B KEGG pathway coverage comparison between individual ADAGE and ensemble ADAGE.
597 eADAGE models (n=10) covers significantly more KEGG pathways than both corADAGE (n=10)
598 and ADAGE (n=1000).
599 C The enrichment significance of three example KEGG pathways in different models. The
600 three pathways show different trends as model size increases in individual ADAGE, however,
601 their median significance levels in eADAGE are comparable or better than all individual models
602 with different sizes. The grey dotted line indicates FDR q-value of 0.05 in pathway enrichment.
603 D The distribution of KEGG pathway coverage rate of ADAGE (n=1000) and eADAGE (n=10).
604 eADAGE models have larger proportion on the high coverage side than ADAGE models,
605 indicating pathways were captured more robustly in eADAGE.
606 E Comparison among PCA, ICA, and eADAGE in KEGG pathway coverage at different
607 significance levels. eADAGE outperforms PCA at all significance levels. eADAGE and ICA show
608 similar pathway coverage at the cutoff q-value = 0.05. However, ICA covers less pathways than
609 eADAGE as the significance cutoff becomes more stringent.
610
611 **Figure 3: eADAGE signatures that show medium-specific patterns.**
612 A Activity of Node147pos in M9-based media. Its activity is high in M9 with
613 phosphatidylcholine but low in other M9-based media.
614 A Activity of Node164pos in all media. NGM+<0.1phosphate, peptone, and King's A media
615 have evident elevation in Node164pos's activity. PIA medium show a wide range in
616 Node164pos's activity. All other media have very low activities.
617 B Expression heatmaps of genes in Node164pos across samples in NGM+<0.1phosphate,
618 peptone, King's A, and PIA media. Heatmap color range is determined by the Z-scored gene
619 expression of all samples in the compendium. These genes are highly expressed in all samples
620 grown on NGM + <0.1mM phosphate, peptone, King's A, and half of samples on PIA, but not
621 expressed in samples grown on NGM + 25mM phosphate.
622
623 **Figure 4: PhoA activity, as seen by the colorimetric BCIP assay in various media**
624 A PhoA activity, as seen by the blue-colored product of BCIP cleavage, is dependent on
625 low phosphate concentrations, *phoB, phoR* and, in NGM, *kinB*.
626 B PhoA is active in King's A, Peptone and PIA and is dependent on *phoB* and *phoR* on
627 King's A and peptone but dependent on *kinB* as well on PIA at 16 hours.
628 C PhoA is active in King's A, Peptone and PIA and is dependent on *phoB,* but no longer
629 *phoR,* while still dependent on *kinB* on PIA after 32 hours.
630 D PhoA activity is dependent on phosphate concentrations < 0.6 mM, *phoB, phoR* and *kinB*
631 as well at 0.5 mM phosphate in MOPS. Concentration 0.2 mM (not shown) mimics 0.1mM and
632 concentrations 0.7mM – 0.9mM (not shown) mimic 1.0 mM.
633
634 ## METHODS
635
636 **Data processing**

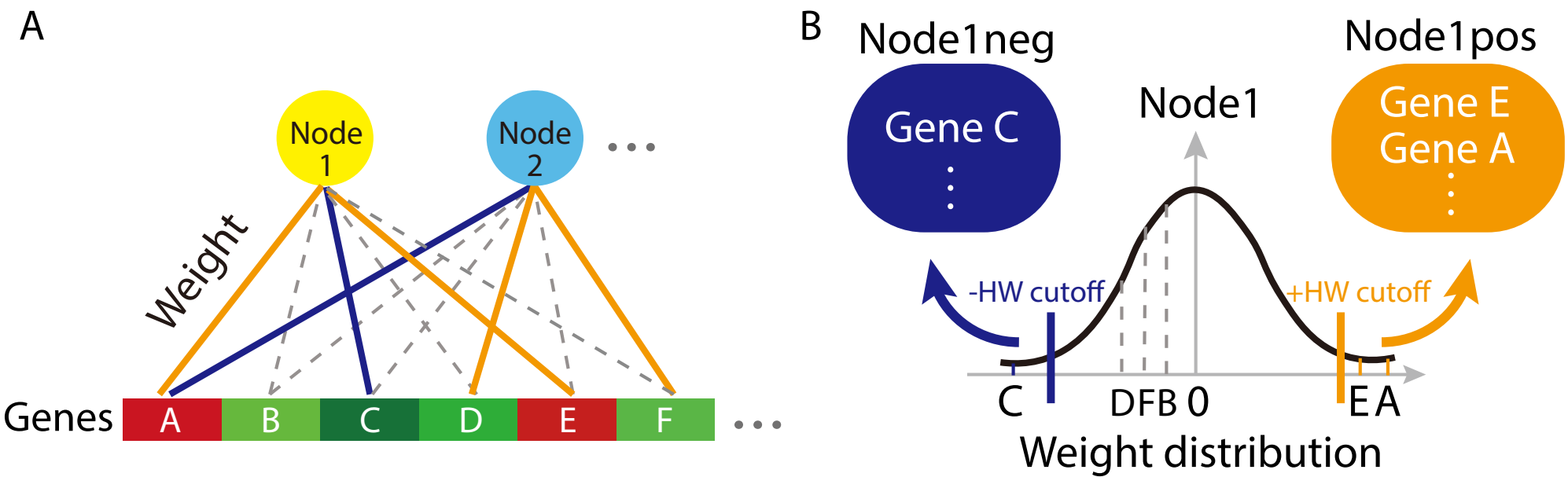
637 We followed the same procedures for data collection, processing, and normalization as (Tan et
638 al., 2016b) and updated the *P. aeruginosa* gene expression compendium to include all datasets
639 on GPL84 platform from the ArrayExpress database (Rustici et al., 2013) as of 31 July 2015. This
640 *P. aeruginosa* compendium contains 125 datasets with 1051 individual genome-wide assays.
641 Processed expression values of the Δ*tctD* RNAseq dataset were downloaded from ArrayExpress
642 (E-GEOD-64056) and normalized to the range of the compendium using TDM (Thompson et al.,
643 2016). We provide the *P. aeruginosa* expression compendium (Dataset S1) along with all the
644 code used in this paper (Tan et al., 2016a). The eADAGE repository is also tracked under version
645 control at https://bitbucket.org/greenelab/eadage.
646
647 **Construction of ADAGE models**
648 We constructed ADAGE models as described in (Tan et al., 2016b). To summarize the process
649 and outputs, we constructed a denoising autoencoder for the gene expression compendium.
650 Denoising autoencoders model the data in a lower dimension than the input space, and the
651 models are trained with random gene expression measurements set to zero. Thus an ADAGE
652 model must learn gene-gene dependencies to fill in this missing information. Once the ADAGE
653 model is trained, each node in the hidden layer contains a weight vector. These positive and
654 negative weights represent the strength of each gene's connection to that node.
655
656 **Gene signatures as sign-specific high-weight gene sets**
657 In previous work (Tan et al., 2016b) we defined high-weight (HW) genes as those in the
658 extremes of the weight distribution on the positive or negative side of a node. Here, we use a
659 more granular definition that accounts for sign specificity. Each node's gene weights are
660 approximately normal and centered at zero in ADAGE models (Tan et al., 2016b, 2015). We
661 defined positive HW genes as those that were more than 2.5 standard deviations from the
662 mean on the positive side, and negative HW genes as those that were more than 2.5 standard
663 deviations from the mean on the negative side. After this split, a model with n nodes provides
664 2n gene signatures. Because a node is simply named by the order that it occurs in a model, we
665 named two gene signatures derived from one node as "NodeXXpos" and "NodeXXneg".
666
667 **KEGG pathway and GO-BP term enrichment analysis**
668 To evaluate the biological relevance of gene signatures extracted by an ADAGE model, we
669 tested how they related to known KEGG pathways (Kanehisa and Goto, 2000). We tested a
670 signature's association with each KEGG pathway using hypergeometric test and corrected the
671 p-value by the number of KEGG pathways we tested following the Benjamini–Hochberg
672 procedure. We used a false discovery rate of 0.05 as the significance cutoff. The same
673 procedure was repeated using GO-BP terms. We downloaded biological process GO terms from
674 pseudomonas.com and only used manually curated terms. For KEGG and GO terms, we only
675 considered terms with more than 5 genes and less than 100 genes as meaningful pathways or
676 processes.
677
678 Genes can be annotated to multiple pathways. To control for this effect in our analysis, we also
679 performed a parallel analysis after applying crosstalk correction as described in (Donato et al.,
680 2013). This approach uses expectation maximization to map each gene to the pathway in which

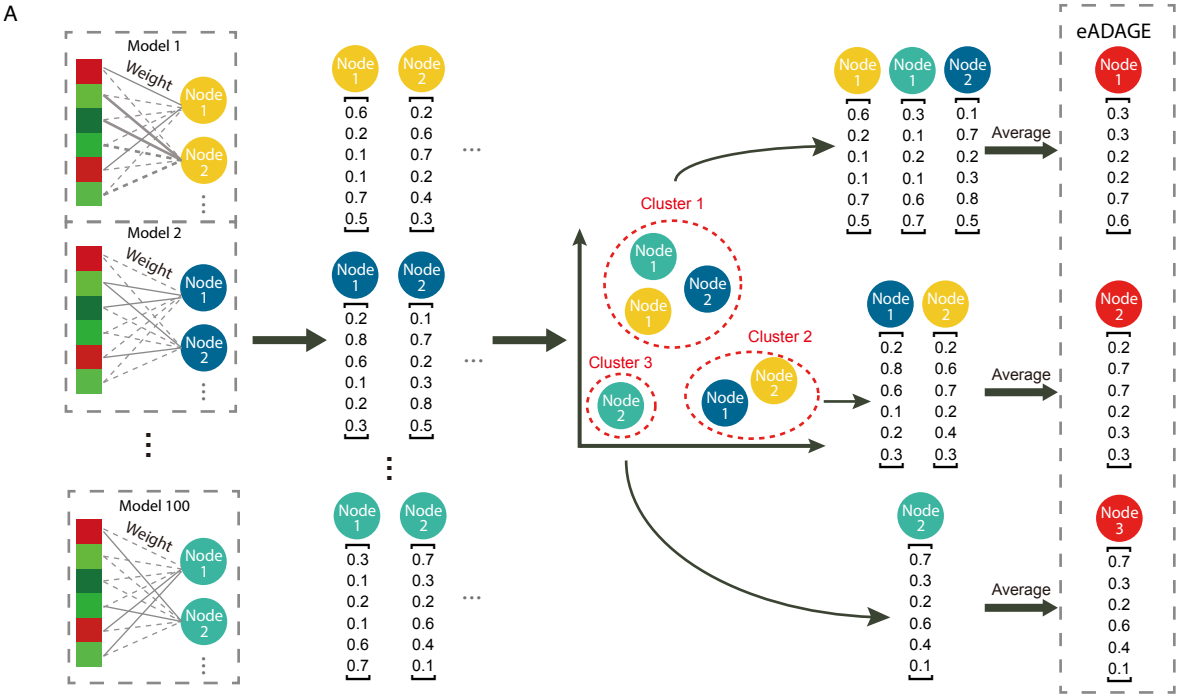
681 it has the greatest predicted impact. A gene-to-pathway membership matrix, defined using
682 KEGG pathway annotations, initially makes the assumption that each gene's role in all of its
683 assigned pathways remains constant independent of context. We then applied pathway
684 crosstalk correction using genes' weights for each node in the ADAGE model. We used the
685 expectation maximization algorithm to maximize the log-likelihood of observing the
686 membership matrix given each node's weight vector. This process inferred an underlying gene-
687 to-pathway impact matrix and iteratively estimated the probability that a particular gene g
688 contributed the greatest fraction of its impact to some pathway P. Upon convergence, we
689 assigned each gene to the pathway in which it had the maximum impact. The resulting pathway
690 definitions do not share genes. We then used these corrected definitions for an analysis parallel
691 to the KEGG process described above.
692
693 **Reconstruction error calculation**
694 The training objective of ADAGE is to, given a sample with added noise, return the originally
695 measured expression values. The error between the reconstructed data and the initial data is
696 the 'reconstruction error.' To summarize the difference over all genes we used cross-entropy
697 between the original sample and the reconstruction, which has been widely used with these
698 methods and in this domain (Tan et al., 2016b; Vincent et al., 2008). This matches the statistic
699 used during training of the model. To calculate reconstruction error for a model, we use the
700 mean reconstruction error across samples.
701
702 **Model size and sample size heuristics**
703 One important parameter of a denoising autoencoder model is the number of nodes in the
704 hidden layer, which we refer to as the model size. To evaluate the impact of model size and
705 choose the most appropriate size, we built 100 ADAGE models at each model size of 10, 50, 100,
706 200, 300, 500, 750, and 1000, using different random seeds. The random seed determines the
707 initialization values in the weight matrix and bias vectors in ADAGE construction, so different
708 random seeds will result in models that reach different local minima. Other training parameters
709 were set to the values previously identified as suitable for a gene expression compendium (Tan
710 et al., 2015). In total, 800 ADAGE models, i.e. 100 at each model size, were generated in the
711 model size evaluation experiment.
712
713 To evaluate the impact of sample size on the performance of ADAGE models, we randomly
714 generated subsets of the *P. aeruginosa* expression compendium with sample size of 100, 200,
715 500, and 800. We then trained 100 ADAGE models at each sample size, each with a different
716 combination of 10 different random subsets and 10 different random training initializations. To
717 evaluate each model, we randomly selected 200 samples not used during training as its testing
718 set. We performed this subsampling analysis at model size 50 and 300. In total, 800 ADAGE
719 models were built in the sample size evaluation experiment.
720
721 The impacts of model size and sample size on model selection were evaluated in the
722 supplement (Figure S4). For subsequent steps, we set the model size to 300 because it was the
723 size that was best supported in the current *P. aeruginosa* compendium by this evaluation.
724

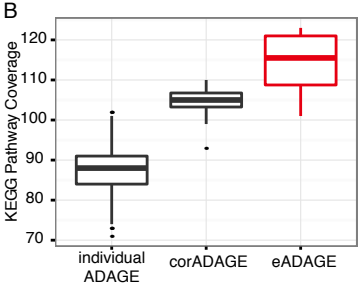
725 **Construction of eADAGE models**
726 We constructed ensemble ADAGE (eADAGE) models by combining many individual ADAGE
727 models in to a single model. For each eADAGE model we combined 100 individual ADAGE
728 models. The 100 models were trained with identical parameters but distinct random seeds. For
729 an eADAGE model of size 300, we trained 100 individual models with 300 nodes each, which
730 provided 30000 total nodes. Each node has a weight vector. We have previously observed that
731 high-weight genes provided the most information to each node (Tan et al., 2016b), so we
732 calculated a weighted Pearson correlation between each node's weight vectors. Our weighted
733 Pearson correlation used (|node1 weight|+|node2 weight|)/2 as the weight function for each
734 gene. We compared this to an unweighted Pearson correlation (corADAGE) as well a baseline
735 ADAGE model.
736
737 After calculating correlation (weighted for eADAGE and unweighted for corADAGE), we
738 converted the correlation to distance by calculating (1- correlation)/2. This provided a
739 30000*30000 distance matrix storing distances between every two nodes. We clustered this
740 distance matrix using the Partition Around Medoids (PAM) clustering algorithm (Park and Jun,
741 2009).We implemented clustering in R using the ConsensusClusterPlus package (Wilkerson and
742 Hayes, 2010) from Bioconductor with the ppam function from Sprint package to perform
743 parallel PAM (Piotrowski et al., 2011). We set the number of clusters to match the individual
744 ADAGE model (e.g. 300) allowing for direct comparison between the eADAGE and ADAGE
745 methods.
746
747 Clustering assigned each node to a cluster ranging from 1 to 300. We combined nodes assigned
748 to the same cluster by calculating the average of their weight vectors. These 300 averaged
749 vectors formed the weight matrix of the eADAGE model. Because the ensemble model is built
750 from the weight matrices of individual models, it does not have the parameters that form the
751 bias vectors. We built 10 eADAGE and 10 corADAGE models from 1000 ADAGE models with
752 each ensemble model built upon 100 different individual models. The individual eADAGE model
753 used for biological analysis in this work was constructed with random seed 123, which was
754 arbitrarily chosen before model construction and evaluation.
755
756 **PCA and ICA model construction**
757 We constructed PCA and ICA models and defined each model's weight matrix following the
758 same procedures in (Tan et al., 2016b). To compare with the 300-node eADAGE, we generated
759 models of matching size (300 components). For ICA, we evaluated 10 replicates. PCA provides a
760 single model. PCA and ICA models were evaluated through the KEGG pathway enrichment
761 analysis described above.
762
763 **Activity calculation for a gene signature**
764 We calculated a signature's activity for a specific sample as $A = W \cdot E/N$, in which W is a
765 vector of genes' absolute weights in that signature, E is a vector of genes' expression values
766 after zero-one normalization in that sample, and N is the number of genes. It can be viewed as
767 an averaged weighted sum of genes' expression levels for genes in the signature. We
768 normalized a signature's activity by the number of genes (N) in that signature, because different

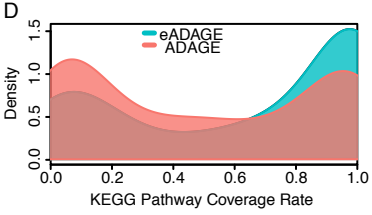
769 signatures have different number of genes. We use gene's absolute weight value in activity
770 calculation to keep activity positive. In this way, a high activity indicates that majority of genes
771 in the signature are highly expressed in the sample and a low activity indicates that majority of
772 genes in the signature are lowly expressed in the sample.
773
774 **Media annotation of the *P. aeruginosa* compendium**
775 A team of *P. aeruginosa* biologists annotated the media for all samples in the compendium by
776 referring to information associated with each sample in the ArrayExpress (Rustici et al., 2013)
777 and/or GEO (Edgar, 2002) databases and along with the original publication, if reported. Each
778 sample was annotated by two curators separately. Conflicting annotations, if they occurred,
779 were resolved by a third curator. The media annotation for all samples in the compendium
780 were provided in Table S1.
781
782 **Identification of signatures activated across media**
783 We calculated an activation score to identify gene signatures with dramatically elevated or
784 reduced activity in a specific medium. We grouped samples by their medium annotation. For
785 each gene signature and medium combination, we calculated the absolute difference between
786 the mean activity of the signature for samples in that medium as well as the mean activity
787 across the remainder of samples in the compendium. We divided this difference in the means
788 by the range of activity for all samples across the compendium. This score captures the
789 proportion by which the mean activity in a medium differs relative to the total difference across
790 the compendium. We termed this ratio the activation score.
791
792 To identify the most specifically active signatures for each medium, we constructed a table for
793 all pairs with an activation score greater than or equal to 0.4 (Table S5). This was highly
794 stringent: it captured only the top 2.4% of the potential signature-medium pairs (Figure S9). To
795 identify pan-media signatures, we limited signatures to those that were active in multiple
796 media (greater or equal to 0.4) and averaged their activation scores (Table S7). These signatures
797 exhibit parallel patterns for multiple media across multiple distinct experiments.
798
799 **Definition of the PhoB regulon**
800 A PhoB regulon for the PAO1 genome was adapted from the PhoB regulon of PA14 in (Bielecki
801 et al., 2015) in order to be comparable to models built with PAO1 genome. Of the 187 genes in
802 the PA14 regulon, 160 were in the PAO1 reference genome (www.pseudomonas.com).
803
804 **Strains and Media**
805 Strains used were WT, Δ*phoB* (DH2633, O'Toole lab collection), Δ*phoR* (DH2516) and Δ*kinB*
806 (DH2517), all in the PA14 background. All strains were maintained on LB with 1.5% agar and
807 grown at 37 °C. For cross-media and phosphate concentration comparisons, BCIP assays (see
808 methods below) were conducted on different base media with 1.5% agar (Fisher): King's A
809 (Pancreatic Digest of Gelatin (Difco) 20g/L; $MgCl_2$ 1.4g/L; K_2SO_4 10g/L; Glycerol 10ml/L) (King *et*
810 *al*, 1954), LB (Tryptone (Fisher) 10g/L; Yeast Extract (Fisher) 5g/L; NaCl 5g/L) (Bertani, 2004),
811 MOPS (morpholinepropanesulfonic acid 40mM; Glucose 20 ml/L; K_2SO_4 2.67mM; K_2HPO_2 0mM,
812 25mM or 0.1 – 1 mM) (Neidhardt et al., 1974), NGM (Pancreatic Digest of Gelatin 2.5g/L;

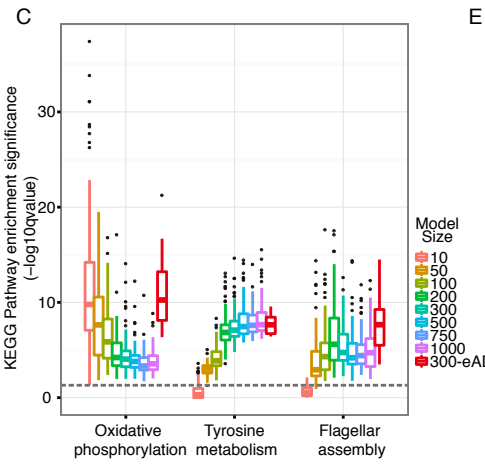
813 Cholesterol 5mg/L; NaCl 3g/L; $MgSO_4$ 1mM; $CaCl_2$ 1mM; KCl 25mM; Potassium Phosphate
814 buffer pH6 0 or 25 mM) (Zaborin et al., 2009), Peptone (Pancreatic Digest of Gelatin 10g/L;
815 $MgSO_4$ 1.5g/L; K_2SO_4 10g/L) (Lundgren et al., 2013), Pseudomonas Isolation Agar (PIA, prepared
816 as per instructions, BioWorld).
817
818 **BCIP assay**
819 Various media were supplemented with 5-bromo-4-chloro-3-indolyl phosphate (BCIP) DMF
820 solution to a final concentration of 60 µg/mL. BCIP assay plates were inoculated with 5 µl of
821 overnight *P. aeruginosa* culture in LB broth. Colonies were grown for 16 hours at 37 °C then
822 matured at room temperature until imaging. Images were collected 16 and 32 hours post
823 inoculation.
824
825 **Screen of a histidine kinase mutant collection**
826 Molecular techniques to construct the histidine kinase (HK) knock out collection were carried
827 out as previously described (Ha et al., 2014). For each strain in the HK collection, a BCIP assay
828 was performed on PIA. Plates were struck with an overnight *P. aeruginosa* culture concentrated
829 two-fold by centrifugation. Plates were incubated at 37 °C 12-16 hours and matured at room
830 temperature for an additional 12-16 hours alkaline phosphatase activity was determined
831 qualitatively, based on blue color.
832

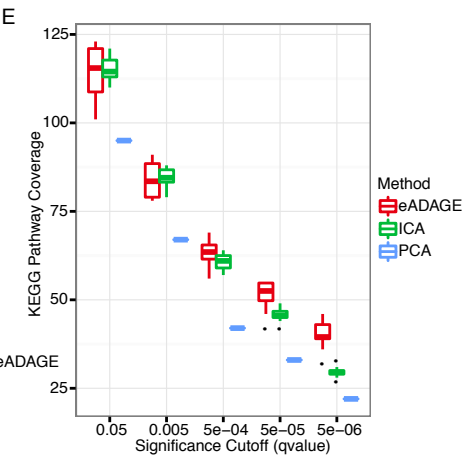

A
Node 1
Node 2
Weight
Genes
A
B
C
D
E
F
B
Node1neg
Gene C
Node1
Node1pos
Gene E
Gene A
-HW cutoff
+HW cutoff
C
DFB 0
E A
Weight distribution

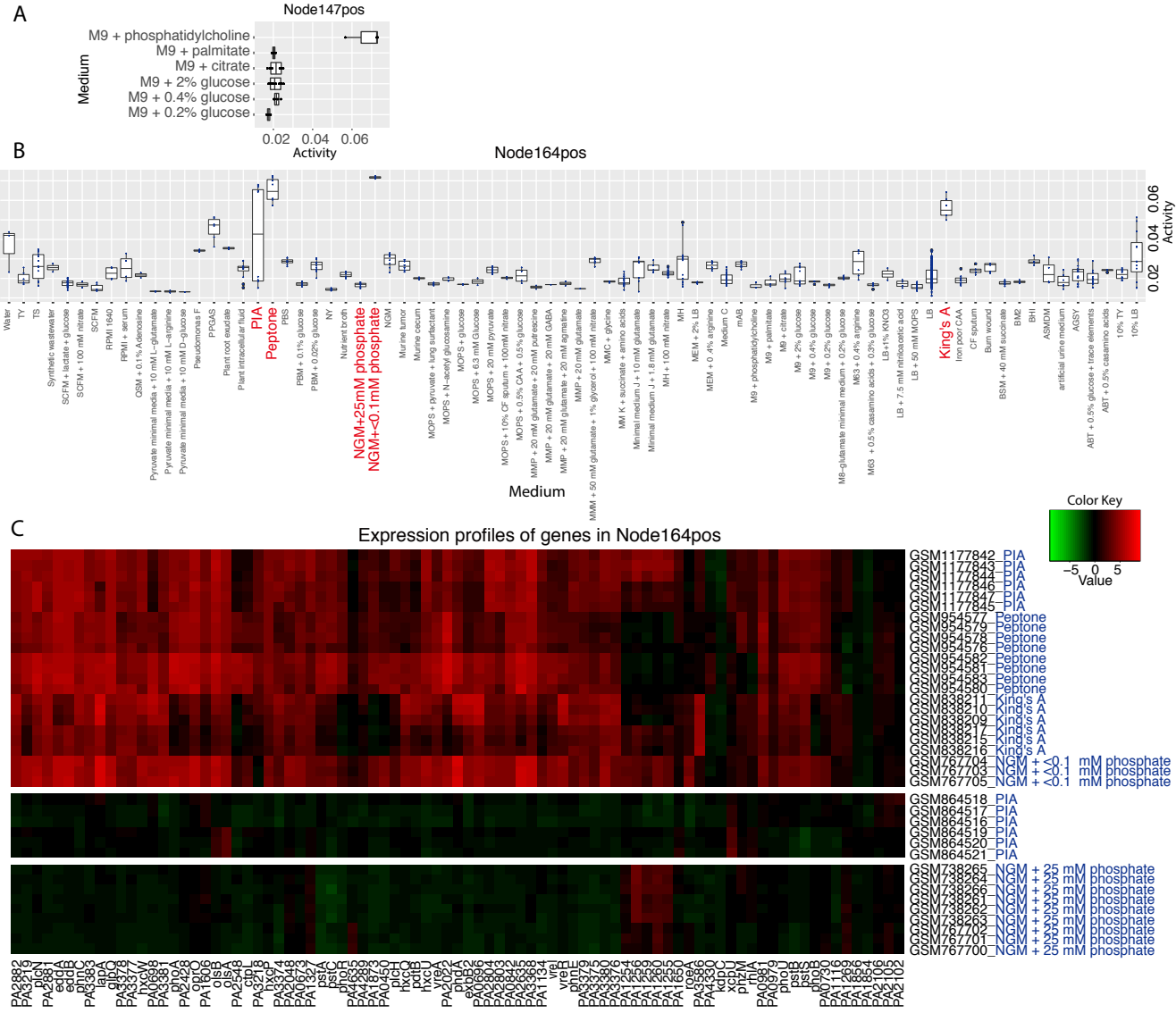

A
Model 1
Weight
Node 1
Node 2
Model 2
Weight
Node 1
Node 2
Model 100
Weight
Node 1
Node 2
Node 1
Node 2
0.6 0.2
0.2 0.6
0.1 0.7
0.1 0.2
0.7 0.4
0.5 0.3
0.2 0.1
0.8 0.7
0.6 0.2
0.1 0.3
0.2 0.8
0.3 0.5
0.3 0.7
0.1 0.3
0.2 0.2
0.1 0.6
0.6 0.4
0.7 0.1
Cluster 1
Cluster 2
Cluster 3
Node 1
Node 2
Node 3
0.6 0.3 0.1
0.2 0.1 0.7
0.1 0.2 0.2
0.1 0.1 0.3
0.7 0.6 0.8
0.5 0.7 0.5
0.2 0.2
0.8 0.6
0.6 0.7
0.1 0.2
0.2 0.4
0.3 0.3
0.7
0.3
0.2
0.6
0.4
0.1
Average
Average
Average
eADAGE
0.3
0.3
0.2
0.2
0.7
0.6
0.2
0.7
0.7
0.2
0.3
0.3
0.7
0.3
0.2
0.6
0.4
0.1

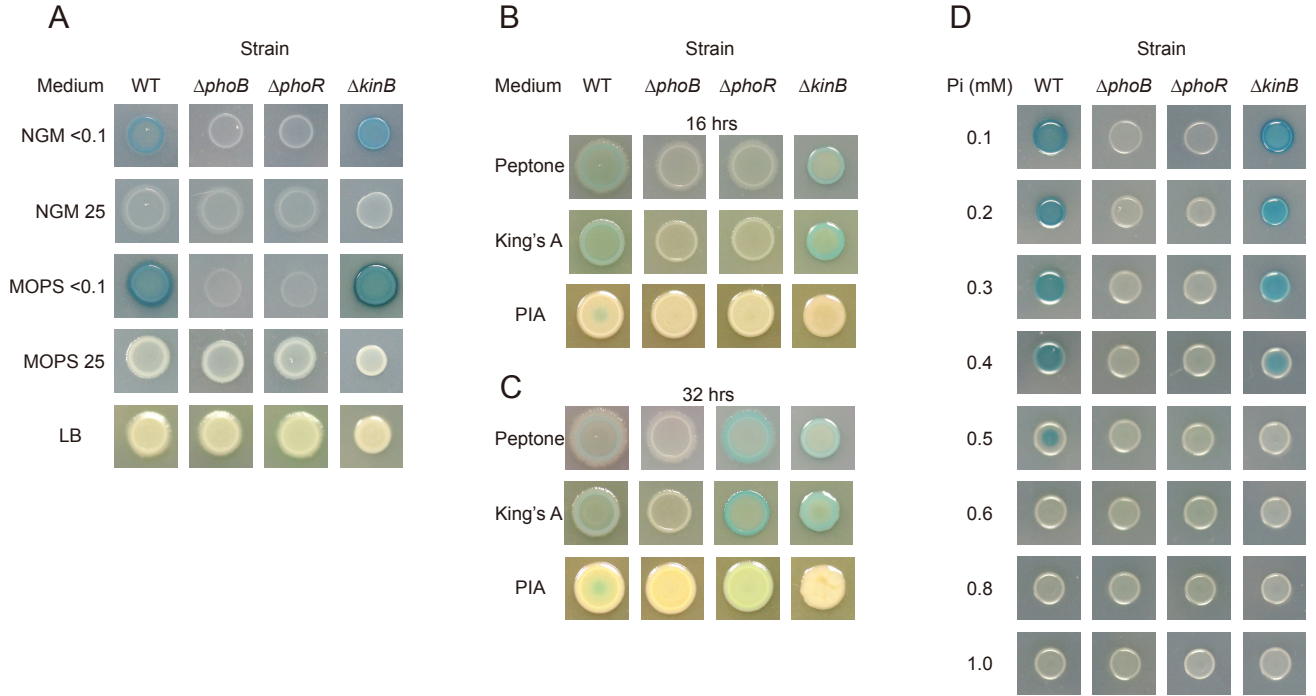

B
KEGG Pathway Coverage
120
110
100
90
80
70
individual ADAGE
corADAGE
eADAGE


D
Density
1.5
1.0
0.5
0.0
eADAGE
ADAGE
0.0
0.2
0.4
0.6
0.8
1.0
KEGG Pathway Coverage Rate


C
KEGG Pathway enrichment significance (−log10qvalue)
30
20
10
0
Oxidative phosphorylation
Tyrosine metabolism
Flagellar assembly
Model Size
10
50
100
200
300
500
750
1000
300-eADAGE


E
KEGG Pathway Coverage
125
100
75
50
25
0.05
0.005
5e−04
5e−05
5e−06
Significance Cutoff (qvalue)
Method
eADAGE
ICA
PCA


A
Node147pos
Medium
M9 + phosphatidylcholine
M9 + palmitate
M9 + citrate
M9 + 2% glucose
M9 + 0.4% glucose
M9 + 0.2% glucose
0.02 0.04 0.06
Activity
B
Node164pos
Activity
0.02 0.04 0.06
Water
TY
TS
Synthetic wastewater
SCFM + lactate + glucose
SCFM + 100 mM nitrate
SCFM
RPMI 1640
RPMI + serum
GSM + 0.1% Adenosine
Pyruvate minimal media + 10 mM L-glutamate
Pyruvate minimal media + 10 mM L-arginine
Pyruvate minimal media + 10 mM D-glucose
Pseudomonas F
PPGAS
Plant root exudate
Plant intracellular fluid
PIA
Peptone
PBS
PBM + 0.1% glucose
PBM + 0.02% glucose
NY
Nutrient broth
NGM+25mM phosphate
NGM+<0.1mM phosphate
NGM
Murine tumor
Murine cecum
MOPS + pyruvate + lung surfactant
MOPS + N-acetyl glucosamine
MOPS + glucose
MOPS + 6.3 mM Glucose
MOPS + 20 mM pyruvate
MOPS + 10% CF sputum + 100mM nitrate
MOPS + 0.5% CAA + 0.5% glucose
MMP + 20 mM glutamate + 20 mM putrescine
MMP + 20 mM glutamate + 20 mM GABA
MMP + 20 mM glutamate + 20 mM agmatine
MMP + 20 mM glutamate
MMM + 50 mM glutamate + 1% glycerol + 100 mM nitrate
MMC + glycine
MM K + succinate + amino acids
Minimal medium J + 10 mM glutamate
Minimal medium J + 1.8 mM glutamate
MH + 100 mM nitrate
MH
MEM + 2% LB
MEM + 0.4% arginine
Medium C
mAB
M9 + phosphatidylcholine
M9 + palmitate
M9 + citrate
M9 + 2% glucose
M9 + 0.4% glucose
M9 + 0.2% glucose
M8-glutamate minimal medium + 0.2% glucose
M63 + 0.4% arginine
M63 + 0.5% casamino acids + 0.3% glucose
LB+1% KNO3
LB + 7.5 mM nitriloacetic acid
LB + 50 mM MOPS
LB
King's A
Iron poor CAA
CF sputum
Burn wound
BSM + 40 mM succinate
BM2
BHI
ASM2M
artificial urine medium
AGSY
ABT + 0.5% glucose + trace elements
ABT + 0.5% casamino acids
10% TY
10% LB
Medium
C
Expression profiles of genes in Node164pos
Color Key
−5 0 5
Value
GSM1177842_PIA
GSM1177843_PIA
GSM1177844_PIA
GSM1177846_PIA
GSM1177847_PIA
GSM1177845_PIA
GSM954577_Peptone
GSM954579_Peptone
GSM954578_Peptone
GSM954576_Peptone
GSM954582_Peptone
GSM954581_Peptone
GSM954583_Peptone
GSM954580_Peptone
GSM838211_King's A
GSM838210_King's A
GSM838209_King's A
GSM838217_King's A
GSM838215_King's A
GSM838216_King's A
GSM767704_NGM + <0.1 mM phosphate
GSM767703_NGM + <0.1 mM phosphate
GSM767705_NGM + <0.1 mM phosphate
GSM864518_PIA
GSM864517_PIA
GSM864516_PIA
GSM864519_PIA
GSM864520_PIA
GSM864521_PIA
GSM738265_NGM + 25 mM phosphate
GSM738264_NGM + 25 mM phosphate
GSM738266_NGM + 25 mM phosphate
GSM738261_NGM + 25 mM phosphate
GSM738262_NGM + 25 mM phosphate
GSM738263_NGM + 25 mM phosphate
GSM767702_NGM + 25 mM phosphate
GSM767701_NGM + 25 mM phosphate
GSM767700_NGM + 25 mM phosphate
PA2882
PA3719
PA2881
PA2883
edcA
edcB
phnC
PA3383
tapA
PA3378
PA3372
hxcW
PA0698
PA3381
PA2423
opnC
PA1606
olsB
olsA
PA3218
PA3219
hxcP
PA3374
PA0048
PA0622
PA1922
pstA
pstC
phoR
PA4635
PA4633
PA1879
PA0450
plcH
hxcQ
pstB
hxcU
PA2022
phzA
exbB2
PA0696
PA2804
PA2843
PA0635
PA3368
PA1134
vgrI
pilG
PA3379
PA3375
PA3380
PA3376
PA1254
PA1259
PA1260
PA1255
PA1650
PA3366
PA3365
PA3330
kdpC
xcpU
phzM
PA0104
PA0979
phoU
pstB
pstS
phoB
PA1110
PA1269
PA1856
PA1854
PA2106
PA2102


A
Strain
Medium
WT
ΔphoB
ΔphoR
ΔkinB
NGM <0.1
NGM 25
MOPS <0.1
MOPS 25
LB
B
Strain
Medium
WT
ΔphoB
ΔphoR
ΔkinB
16 hrs
Peptone
King's A
PIA
C
32 hrs
Peptone
King's A
PIA
D
Strain
Pi (mM)
WT
ΔphoB
ΔphoR
ΔkinB
0.1
0.2
0.3
0.4
0.5
0.6
0.8
1.0